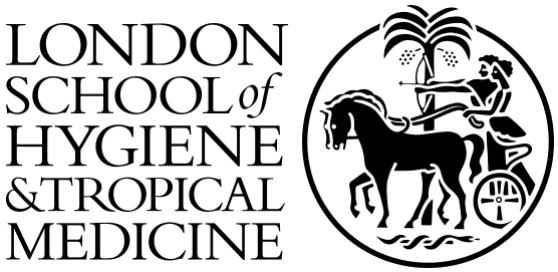


Comparison of variance estimators for meta-analysis of instrumental variable estimates.

A F Schmidt* [a], A D Hingorani [a], B J Jefferis [b], J White [c], R H H Groenwold [d], F Dudbridge [e], for the UCLEB consortium [f].

a. Institute of Cardiovascular Science, Faculty of Population Health, University College London, London WC1E 6BT, United Kingdom.
b. Department of Primary Care and Population Health, University College London, Rowland Hill Street, London NW3 2PF, United Kingdom.
c. UCL Genetics Institute, University College London, London WC1E 6BT, United Kingdom.
d. Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, P.O. Box 85500, 3508 GA Utrecht, the Netherlands.
e. Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London WC1E 7HT, United Kingdom.
f. See acknowledgement.

***** Contact: 0044 (0)20 3549 5625

E-mail address: amand.schmidt@ucl.ac.uk (A.F.Schmidt).

Running title: IV variance estimators

Word count text: 3887

Word count abstract: 196

Number of references: 32

Number of tables: 2

Number of figures: 4

(Web)appendix: 1

Key Messages

1. To increase power, Mendelian randomization studies frequently combine study results (two-stage meta-analysis) or study datasets (one-stage meta-analysis). When conducting a two-stage meta-analysis, different variance estimators may not only impact coverage or type 1 error rates but also point estimates.
2. In two-stage meta-analyses of weak instrument or rare diseases, resampling based variance estimators are expected to result in biased point estimates with coverage below 0.95. Two-stage meta-analyses using the delta-method are expected to perform better.
3. In the presence of between study heterogeneity, the delta-method applied at stage one of the meta-analysis will likely result in the least biased estimate with relatively good coverage.
4. In one-stage meta-analysis scenarios, point estimates are not influenced by the choice of variance estimator, and coverage is generally similar between the variance estimators. One-stage meta-analyses are however, still affected by the size and quality of the included studies.

Abstract

Background Mendelian randomization studies perform instrumental variable (IV) analysis using genetic IVs. Results of individual Mendelian randomization studies can be pooled through meta-analysis. We explored how different variance estimators influence the meta-analysed IV estimate.

Method Two versions of the delta method (IV before or after pooling), four bootstrap estimators, a jack-knife estimator and a heteroscedasticity-consistent (HC) variance estimator were compared using simulation. Two types of meta-analyses were compared, a two-stage meta-analysis, pooling results and a one-stage meta-analysis, pooling datasets.

Results Using a two-stage meta-analysis, coverage of the point estimate using bootstrapped estimators deviated from nominal levels at weak instrument settings and/or outcome probabilities ≤ 0.10. The jack-knife estimator was the least biased resampling method, the HC estimator often failed at outcome probabilities ≤ 0.50, and overall the delta method estimators were the least biased. In the presence of between study heterogeneity the delta method before meta-analysis performed best. Using a one-stage meta-analysis all methods performed equally well and better than two-stage meta-analysis of greater or equal size.

Conclusion In the presence of between study heterogeneity two-stage meta-analyses should preferentially use the delta method before meta-analysis. Weak instrument bias can be reduced by performing a one-stage meta-analysis.

Keywords Epidemiology methods; Mendelian Randomization Analysis; Statistics.

Introduction

Despite considerable effort, observational (i.e., nonrandomized) studies are sensitive to confounding bias and reverse causation(1-4). To overcome these problems, Mendelian randomization (MR) studies have been advocated, using one or multiple Single-Nucleotide Polymorphisms (SNPs) as an instrument in instrumental variable (IV) analyses(5;6).

In this type of Mendelian randomization study the effects of an IV on an intermediate phenotype and on an outcome are estimated, and combined to derive the causal effect of the intermediate on the outcome. This causal effect is unbiased if (amongst others) the following three assumptions hold: [1] the IV is associated with phenotype, [2] conditional on the phenotype and the (possibly unmeasured) confounders, the IV is independent of the outcome and [3] the IV is independent of confounders(7).

While the performance of the different IV point estimators has previously been explored(8;9), the performance of the different variance estimators remains unclear. This is especially important because, to increase precision, Mendelian randomization studies often meta-analyse results from multiple studies. Because of this, different variance estimators not only impact type-1 error rates and confidence intervals but may also lead to different point estimates.

Typically three types of meta-analysis can be defined: an aggregated meta-analysis combining study specific results; a two-stage individual patient data meta-analysis, in which an analysis script is designed and shared prospectively, before pooling study specific results; and third, a one-stage individual patient data meta-analysis sharing the actual datasets. Given the usually straightforward analyses in genetic epidemiology the difference between aggregated meta-analysis and two-stage individual patient data meta-analysis are often small, therefore here we only differentiate between two-stage meta-analyses and one-stage meta-analyses. A recent

review by Boef et.al.(10) showed that 47 out of 80 meta-analysis of Mendelian randomization performed a two-staged analysis; among those, 10 performed IV analysis within each study before combining, whereas 9 combined gene-phenotype and gene-outcome associations separately before performing IV analysis. We note that gene scores are also used as instruments(11), using aggregated results this can be implemented, for example, by meta-analysing aggregated results of the gene-biomarker and the gene-outcome relationships into two estimates(12) and applying the ratio estimator (see methods). Alternatively, when individual patient data is available, gene scores can be implemented using the "two-stage least squared like" estimator (TSLS, see methods).

In the present study we used simulations to compare multiple variance estimators. In addition, an empirical example on the effect of low-density lipoprotein cholesterol (LDL-C) on cardiovascular disease (CVD) is included.

Methods

Simulation set-up

Initially we focus on a two-stage meta-analysis where each study has information on a single SNP (Z), a continuous phenotype (X), and a dichotomous endpoint (Y). The goal is to estimate the causal [marginal] odds ratio (OR) of one unit increase in phenotype on the outcome.

Data-generating process

J studies were simulated, for the jth study a disease outcome, a phenotype and an IV were generated for n_j independent subjects, where $j = 1, \dots, J$. To increase readability the following notation is presented for one study, with the same process applied to all studies. The IV variable, Z, counts the number of minor alleles for the ith individual. Following a biallelic model,

genotypes were generated from two independent Bernoulli distributions resulting in the usual Hardy-Weinberg proportions:

$$Prob(Z = 0, Z = 1, Z = 2) = (q^2, 2pq, p^2).$$

Where p represents the probability of the rare allele and $q = 1 - p$ the probability of the major allele. Phenotype X was generated dependent on Z and an unobserved confounder C:

$$x_i = \alpha_0 + \alpha_1 z_i + \alpha_2 c_i + \varepsilon_i \quad with\ \varepsilon_i \sim N(0,1), c_i \sim N(0,1).$$

For the ith individual the probability of an event was generated based on X and C:

$$logit(Prob[y_i = 1|c_i, x_i]) = log\left(\frac{Prob[y_i = 1|c_i, x_i]}{1 - Prob[y_i = 1|c_i, x_i]}\right) = \delta_0 + \delta_1(\alpha_0 + \alpha_1 z_i + \alpha_2 c_i + \varepsilon_i) + \delta_2 c_i$$

$$= \delta_0 + \delta_1 x_i + \delta_2 c_i,$$

the event was sampled from a Bernoulli distribution:

$$y_i \sim Bernoulli(Prob[y_i = 1|c_i, x_i]).$$

Data analyses

Point estimators

Given that the confounder C is unobserved it is impossible to estimate the causal effect of the phenotype X on the outcome using regular methods such as logistic regression. Instead, SNP Z can be used to estimate the causal effect of the phenotype on the outcome. The ratio estimator

is a relatively straightforward estimator of the logarithm of the causal odds ratio (logOR), which is the estimand here.

$$\hat{\theta} = (\hat{\gamma}_1 - \hat{\delta}_3)/\hat{\alpha}_1 \quad [1]$$

Here $\hat{\gamma}_1$ represents the effect of the SNP on the outcome measured as the log(OR), $\hat{\delta}_3$ the log(OR) effect of the SNP on the outcome conditional on the phenotype and unmeasured confounders, and $\hat{\alpha}_1$ the mean difference effect of the SNP on the phenotype (estimated by fitting a linear regression of the type $x_i = \hat{\alpha}_0 + \hat{\alpha}_1 z_i + \varepsilon_i$ [equation 2]). If ***every*** confounding variable (C) ***was*** measured $\hat{\gamma}_1$ and $\hat{\delta}_3$, could be estimated by fitting the following (logistic regression) models $logit(Prob[y_i = 1|z_i]) = \hat{\gamma}_0 + \hat{\gamma}_1 z_i$ and $logit(Prob[y_i = 1|z_i, x_i, c_i]) = \hat{\delta}_0 + \hat{\delta}_1 x_i + \hat{\delta}_2 c_i + \hat{\delta}_3 z_i$. However, because it is never known if ***all*** confounders are measured (and correctly specified) this strategy is not feasible. Instead, following the exclusion restriction (assumption 3 above), we assume that $\hat{\delta}_3 = 0$, and equation 1 reduces to the ratio of $\hat{\gamma}_1$ and $\hat{\alpha}_3$. This ratio estimator is typically used when there is a single instrument or when a multi gene score is based on a meta-analysis of aggregated results(12).

Instead of the ratio estimator, the “two-stage least squares like” point estimator (TSLS), also referred to as the two-stage predictor substitution estimators(13), is used to estimate the IV effect using a (weighted) gene score (8).

$$logit(Prob[y_i = 1|\hat{x}_i]) = \hat{\beta}_0 + \hat{\theta}\hat{x}_i \quad [3]$$

Where $\hat{x}_i$ represents the fitted value of a linear model regressing x_i on z_i (i.e., the fitted values from a linear regression defined in equation 2).

Variance estimators

Following the usual research practice we will focus on a two-stage meta-analysis where in the second stage study specific results are pooled by the inverse of the variance(14). Because results are pooled by the inverse of the variance we initially focus on different variance estimators, excluding methods that directly estimate a confidence interval.

The **delta method**(15;16) (DM) has the closed form solution:

$$\hat{\sigma}^2_{DM} = \frac{\hat{\sigma}^2_{\hat{\gamma}_1}}{(\hat{\alpha}_1)^2} + \hat{\sigma}^2_{\hat{\alpha}_1}\frac{(\hat{\gamma}_1)^2}{(\hat{\alpha}_1)^4} - 2\hat{\sigma}^2_{\hat{\gamma}_1,\hat{\alpha}_1}\frac{\hat{\gamma}_1}{(\hat{\alpha}_1)^3} \quad [4]$$

Where $\hat{\sigma}^2_{\hat{\gamma}_1}$ represents the estimated variance in $\hat{\gamma}_1$, $\hat{\sigma}^2_{\hat{\alpha}_1}$ the variance in $\hat{\alpha}_1$ and $\hat{\sigma}^2_{\hat{\gamma}_1,\hat{\alpha}_1}$ the estimated covariance between $\hat{\gamma}_1$ and $\hat{\alpha}_1$. Often the delta method is applied to meta-analysis settings where $\hat{\sigma}^2_{\hat{\gamma}_1,\hat{\alpha}_1}$ is set to zero, resulting in a small over estimation of the variance; this was followed here. Two versions of the delta method were compared: (1) calculating the ratio estimator and the $\hat{\sigma}^2_{DM}$ in each study followed by meta-analysis of $\hat{\theta}$ **[DM1]**, and (2) calculating $\hat{\theta}$ using the ratio estimator and $\hat{\sigma}^2_{DM}$ after separately meta-analysing $\hat{\gamma}_1$and $\hat{\alpha}_1$ **[DM2]**.

Alternatively, by sampling with replacement from the observed sample, creating a resampled dataset of size n, and repeating this B times, a non-parametric bootstrapped distribution (17) can be constructed. This distribution can then be used to estimate the variance in the IV point estimate (**basic bootstrap [BB]**):

$$\hat{\sigma}^2_{Boot} = \frac{1}{B-1}\sum_{b=1}^{B}\left(\bar{\theta}^* - \hat{\theta}^*_b\right)^2 \quad [5]$$

With $\hat{\theta}_b^*$ the IV estimate estimated in the bth bootstrap sample and $\bar{\theta}^*$ the mean IV estimate over the B bootstrap samples; here $B = 1{,}000$.

All bootstrap variance estimators assume symmetry in bootstrap distribution, due to data sparseness however, extreme values of $\hat{\theta}^*$ may occur, overestimation the $\hat{\sigma}^2_{Boot}$. Straightforward solutions that are less sensitive to data sparseness, include a bootstrap stratified for the outcome **[outcome stratified; OS]** or stratified for the SNP status **[SNP stratified; SS]**. A more computer intensive solution is to perform a **double bootstrap [DB]**(17), where for every bth bootstrap sample, R new bootstrap samples of size n are taken using the bth bootstrap sample as the source population. For every bth bootstrap sample the variance is estimated, with the median of these estimates representing the **DB** IV variance estimate. In our simulations $R = 50$, and $B_{DB} = R * 5$. An **jack-knife [JK](17)** variance estimator can also be used:

$$\hat{\sigma}^2_{jack} = \frac{n-1}{n}\sum_{i=1}^{n}\left(\bar{\theta}_{jack} - \hat{\theta}_{-i}\right)^2$$

Here $\bar{\theta}_{jack}$ represents the mean IV estimate over the n jack-knife estimates and $\hat{\theta}_{-i}$ the IV estimate deleting the ith observation.

The previous variance estimators were all applied using the ratio estimator. The **robust sandwich [RB]** heteroscedasticity-consistent [HC] variance estimator can be used for the TSLS IV, in which the variance estimate $\hat{\sigma}^2_{\hat{x}y}$ for $\hat{\theta}$ (equation 3) is replaced by the RB estimate. Here we used HC1 and note that JK and RB estimators are related in the sense that, the JK approximates the HC3 estimator, which is a refinement of HC1(18). Note, that the HC estimators are implemented not to adjust for any heteroscedasticity, but merely to penalize the naive variance estimator which assumes that the $\hat{x}$ in equation 3 is measured without error.

Simulation scenarios

In all simulations $J = 10$ studies were generated, with n_j sampled from a uniform distribution [400, 3600] (see Table 1 for an overview of the simulation parameters). In **scenario I**, the minor allele frequency (p) was set to 0.50, 0.10, 0.05, 0.01, and 0.005. The probability of the outcome was 0.50. To (initially) prevent weak instrument bias (19) the SNP effect on the phenotype was set to $\alpha_1 = 0.50$, and the unmeasured confounder effect to $\alpha_2 = 0.50$. By fixing the SNP-phenotype association and decreasing p the explained variance due to the SNP decreases, as well as the F-statistic. For example, in scenario 1 the average F-statistic was 126, 46, 25, 6, and 5. To simulate a large amount of confounding the log(OR) of the unmeasured confounder effect on the outcome was set to $\delta_2 = 1.50$, and the phenotype log(OR) set to $\delta_1 = 0.00$ (i.e., no causal effect). In **scenario II**, p was set to 0.15 and the probability of the outcome was set to 0.10, 0.05, 0.02, and 0.01. **Scenarios III and IV** differed from II only with respect to $p = \{0.05, 0.01\}$.

All simulations were repeated 2,000 times and were performed with the statistical package R version 3.1.2 for Unix(20). The number of replications was chosen to ensure sufficient precision to detect small deviations from the nominal coverage rate of 0.95 (the 95% lower and upper bounds are 0.940 and 0.960)(21). Results were pooled using the inverse variance method following a fixed- or random-effects model where appropriate.

Performance metrics

Results were evaluated using the following metrics. Mean bias$\left(\overline{\log \mathrm{OR}} - \log[True\ OR]\right)$, with the first term representing the mean of the $log\widehat{OR}$; mean standard error [SE], empirical SE [ESE]; estimated by taking the standard deviation of the distribution of $log\widehat{OR}$. The root mean squared error $\left[RMSE = \sqrt{\left(\overline{\log \mathrm{OR}} - \log[True\ OR]\right)^2 + ESE^2}\right]$, coverage rate, defined as the proportion of

times the 95% confidence interval included the true OR, and the number of models that failed to return estimates.

Additional analyses

Obviously, the absolute performance of the methods depends on the mean sample size per study. To explore the performance in a larger sample size setting, a “medium” sized meta-analysis of 60,000 subjects was simulated by repeating scenario 1.

Instead of combining study results in a two-stage meta-analysis one can also combine datasets in a one-stage meta-analysis. This was explored by repeating scenario 1, concatenating the studies together in a single file and adjusting all analyses for study (i.e., bootstrapped by study or adding a study covariable). Given that results do not have to be pooled in a second stage we only report on a single DM estimator, and instead report on the bootstrap based **percentile confidence interval**(22); which directly estimates the confidence interval (instead of the variance).

In a third sensitivity analysis scenario 1 was repeated introducing between study variance of the gene-phenotype association. This was simulated by replacing α_0, α_1, and ε_i by $\alpha_{0j} \sim N(0.10, 1^2)$, $\alpha_{1j} \sim N(0.50, 1^2)$, and $\varepsilon_{ij} \sim N(0, \varsigma_j^2)$ with $\varsigma_j^2 \sim N(1.50, 0.3^2)$.

In a fourth sensitivity analysis we evaluated the performance of 1) using only the first term of the delta method (the Toby Johnson **[TJ]** method), and 2) replacing the asymptotic variance estimates, $\hat{\sigma}^2_{\gamma_1}$, and $\hat{\sigma}^2_{\alpha_1}$, in the delta method (using the first two terms) by bootstrapped estimates **[DM BB]**. Both methods were implemented by applying the algorithms before meta-analysis, and after meta-analysis (i.e., TJ1, TJ2, DM1 BB, and DM2 BB). Performance was evaluated in

scenario 1. Additionally, in a fifth sensitivity analysis, we explored performance for continuous outcomes; implemented by repeating scenario 1 using the parameters of scenario 1 as mean differences. See Appendix Figure 1 for a flowchart of the methods evaluated.

Results

Figure 1 depicts the performance of the IV variance estimators under different minor allele frequencies (MAF) or instrument strengths (F-statistic). Unless explicitly stated all results pertain to the two-stage meta-analysis. At a MAF of 0.50 pooled odds ratio (OR) estimates of all methods were unbiased, but differences between the estimators increased as MAF decreased to 0.005 (or F-statistic went towards zero). Coverage of both the DM estimators increased towards 1.00 as MAF decreased; the RMSE was equal for both DM estimators, and smaller than the RMSE of other methods (Figure 1). JK and RB coverage deteriorated towards 0.80 at lower MAFs. Coverage of the bootstrap methods decreased below 0.95 at a MAF of 0.10/F-statistic 25, recovering to 0.95 at lower MAFs using the BB, SS and DB methods. This unexpected behaviour in coverage was due to the bias in SE (i.e., difference between mean SE and ESE, see Figure 1, Appendix table 1) trailing behind the bias in OR. Generally the mean SE and ESE agreed well for the DM.

In scenarios II-IV the outcome incidence varied from 0.10 to 0.01 and the MAF was set to 0.15, 0.05, or 0.01 respectively (Appendix tables 2-4). At lower outcome probabilities bias in both DM1 and DM2 was similar, and lower than bias of other methods. For example, in scenario IV at an outcome probability of 0.05 the mean OR was 1.339 and 1.572 for DM1 and DM2, respectively. Coverage of DM1 and DM2 differed substantially at lower outcome probabilities, for example in scenario IV with an outcome probability of 0.01 coverage was 0.793 and 0.550 respectively. Differences between ESE and mean SE was similar however (DM1: -5.729, and DM2: -5.404), as were the RMSE estimates (DM1: 3.268, and DM2: 3.670). Coverage of the JK and bootstrap

methods was similar and decreased below 0.95 at lower outcome probabilities. RMSE was also similar for all resampling methods, but higher than the DM methods. RB estimates were by far the most biased, with the lowest coverage and highest RMSE; this coincided with frequent failure of this method to return estimates.

Repeating scenario 1 with a larger sample size (60,000 subjects), showed a comparable relative pattern as before (Figure 2, Appendix Table 5). Repeating scenario 1 using a one-stage meta-analysis (20,000 subjects) improved performance. There was no difference between the methods in, mean OR, bias, or RMSE (Appendix Table 6); even in extreme settings bias was low -0.016 (MAF of 0.005 or F-statistic of 4). Coverage (Figure 3) was generally close to 0.95 or slightly larger, and agreement between mean SE and ESE was generally good; only deviating at a MAF of 0.005 or an F-statistic of 4. A non-parametric bootstrap percentile confidence interval was evaluated, performing similar to other methods (coverage ≈ 0.95). Repeating scenario 1 with between study variance showed similar performance as in the original fixed effect scenario (Appendix Table 7), except for more conservative coverage rates, and DM2 being the most biased estimator at MAF ≤ 0.01, e.g., -0.257 mean bias at MAF 0.005, which coincided with a coverage rate of almost 1, a RMSE of 10.289. DM1 performed better than all other methods with a coverage of 0.981, and an RMSE of 0.127, at a MAF of 0.005.

The Toby Johnson **[TJ]** variance estimator performed comparably to the DM1 or DM2 in scenario I with only slightly lower coverage (Appendix Table 8). Implementing the delta method by replacing the asymptotic variance estimators with bootstrapped estimators **[DM BB]** performed similarly to the **BB** method (Appendix Table 8). Repeating scenario 1 with a continuous outcome revealed a comparable relative performance of the variance estimators (Appendix Table 9).

Results on the LDL-C effect on CVD.

Table 2 shows empirical results of two different IV's in a 6 study meta-analysis to estimate the effect of LDL-C on CVD, (see Appendix for a description of the data sources, and baseline data). Using SNP rs11591147 as an IV (mean F-statistic = 13.42) in a two-stage meta-analysis showed that the bootstrap methods had the largest standard errors and their point estimates not only disagreed with results from the remaining variance estimators but also between themselves. As expected, using a one-stage meta-analysis increased precision and decreased differences between methods, resulting in an IV estimate of 0.93 (95%CI 0.50;1.72). Results from the weak instrument rs2965101 (mean F-statistic = 1.34) revealed large differences between the bootstrap estimators and the remaining estimators; the minimal bootstrap SE estimate was 13.19, compared to an SE of 1.49 using DM2. Precision increased using a one-stage meta-analysis, however the bootstrapped SE were still comparatively large. Given that one-stage meta-analysis are analysed by a single analyst, it becomes practical to explore the bootstrap distributions (figure 4). After omitting a number of outliers the bootstrap became relatively symmetric and the SE estimates were: 1.27 (BB), 1.29 (OS), 1.33 (SS), 3.51 (DB). The large SE of the DB and its truncated distribution show that 50 times 250 repetitions were insufficient in this setting.

Discussion

This study showed that, depending on the strength of the IV and/or the outcome incidence, there is considerable difference in the performance of instrumental variable (IV) variance estimators in two-stage meta-analysis. The delta method (DM) showed the least amount of bias and the best coverage; with the delta method implemented before meta-analysis performing better in the presence of between study variance. Bootstrap and robust variance estimators (RB) produced extreme estimates in two-stage meta-analysis. Differences between methods were minimal using a one-stage meta-analysis; all providing unbiased estimates and appropriate coverage. An empirical example on the LDL-C effect on CVD incidence, confirmed that these issues also

occur in applied settings. Relative performance of the variance estimators was similar when using a continuous outcome instead of a binary endpoint.

At lower MAF/F-statistic values, or lower outcome probabilities the RB estimators often failed to converge, making it difficult to evaluate whether the underperformance of RB was due to the estimator itself or to informative failures. Looking at the JK (which failed in less than 1% of the simulations, and which is an approximation of the HC3 which is a refinement of the HC1 used in the RB), it seems that to some extent this underperformance of the RB may be explained by computational problems in the R sandwich package(23); this needs further study. Following the usual practice in applied Mendelian randomization analyses, the ratio and the TSLS point estimators were used. Additionally to the usual three IV assumptions, these point estimators also assume the phenotype to be normally distributed conditional on the SNP and confounders, and homogeneity of the phenotype (X) effect on the outcome (24). In our simulations these assumptions held, however in applied settings this is not necessarily the case, given that confounders are often unmeasured, these assumptions are also impossible to evaluate. Instead of making these assumptions, different estimators or estimands may be considered in empirical settings. For example, structural mean models or generalized method of moments point estimators, or the risk difference estimand(8;24), make fewer assumptions.

Our results underline the difficulty of using the observed F-statistic(7) as a measure of expected bias due to a weak instrument. We observed an increased performance in a one-stage meta-analysis with on average 20,000 subjects and a “weak” instrument (MAF 0.05, mean F-statistic 5.97), compared to a two-stage meta-analysis with on average 60,000 subjects and a “strong” instrument (MAF 0.05, mean F-statistic 15.98). When conducting a one-stage meta-analysis, results do not have to be pooled by the inverse of an estimated study specific variance. Therefore, in this scenario, point estimates, precision (ESE), and RMSEs were not influenced by

the choice of variance estimators. The choice of variance estimator did influence coverage, which was nevertheless markedly improved over a two-stage design.

The underperformance of the bootstrap estimators in the two-stage meta-analysis may come as a surprise to some; however, the improved performance (over e.g., a Wald based confidence interval) shown in the literature mostly holds for bootstrap confidence intervals such as the bias corrected and accelerated bootstrapped confidence interval(17;22;25). Because of the need for a variance estimate in the second stage of a two-stage meta-analysis the bootstrap can only be used to estimate the standard error of the IV estimate, which implicitly assumes symmetry of the bootstrap distribution(17;22;25). We did however evaluate the percentile method to directly estimate the confidence interval when we replicated scenario 1 using a one-stage meta-analysis. Results indeed showed proper coverage, however, this was similar to the increased performance of all other estimators. We evaluated a delta method estimator replacing the asymptotic variance estimates by bootstrapped variance estimates; this approach performed worse than the regular delta method (DM1 or DM2). These results show that even though the asymptotic approximations of$\hat{\sigma}^2_{\gamma_1}$ and $\hat{\sigma}^2_{\alpha_1}$ do not strictly hold these estimates are better approximations (in such situations) than bootstrapped alternatives.

The simulations presented here are naturally limited and the following points merit discussion. First, different simulation parameters will result in different absolute performance. Instead we focussed on relative (i.e., between methods) performance which we expect to be more robust. Second, by fixing the effect of the instrument (the SNP) on the phenotype, the instrument strength decreases with MAF, hence our results include analyses with F-statistics below 10. These are analyses, some might argue, an applied researcher would not perform due to violation of IV assumption 1. We showed, however, that despite the “weak” instrument, valid estimates can be derived. Third, while it seems logical to increase the number of bootstraps as

the data becomes sparser (or the IV becomes weaker), we kept the number fixed to preserve comparability between scenarios. Fourth, for simplicity we focussed on scenarios with a single SNP instrument, whereas, to prevent weak-instrument bias, most Mendelian randomization studies use multiple SNPs. Nevertheless, relevant information for these multiple SNP approaches can be found in our analyses by focussing on strong-instrument settings. Fifth, we only explored performance under the null [i.e., OR = 1] because, 1) coverage was often too low making comparisons in power pointless, and 2) we wished to prevent influence of non-collapsibility(26). Sixth, the small ORs observed in low frequency scenarios were most likely due to the outcome being constant for a certain allele number (i.e., perfect separation). In these settings penalized models, using for example a Firth(27;28) or Lasso(29) penalization, are expected to perform better(30). Finally, random effects or fixed effect analysis models were used depending on the simulation scenario including between study variance or not(31). In empirical analyses, the choice between random effects and fixed effect models typically depends on a heterogeneity measure(32). However, bias in point and variance estimates will influence the observed heterogeneity, resulting in different modelling choices depending on the performance of the estimator. This would make between methods comparisons difficult. Therefore, the choice of model was based on the true, rather than the observed, between study variance.

In conclusion, the choice of variance estimator in instrumental variable analyses using a two-stage meta-analysis is important. Simulations showed that the delta method applied at stage-one of the two-stage meta-analysis performed best. If resampling variance estimators are used, we suggest always checking study specific plots of these distributions for outliers. This is especially important if the outcome and/or SNPs are rare or if the instrument is weak. Out of all the resampling methods the jack-knife estimator performed best. However, in such a scenario an even better alternative, when possible, is to perform a one-stage meta-analysis making the choice of variance estimator less influential. If a one-stage design is used, resampling

techniques can be used to directly estimate confidence intervals for which methods exist that do not assume a symmetric distribution (e.g., the percentile method).

Conflict of interest statement

None of the authors of this paper has a financial or personal relationship with other people or organisations that could inappropriately influence or bias the content of the paper.

Acknowledgement

For the use of the UCLEB data we acknowledge the following researchers for their invaluable help in acquiring and preparing the data: Y. Ben-Shlomo, N Chaturvedi,J Engmann, A Hughes, S Humphries, E Hypponen, M Kivimaki, D Kuh, M Kumari, U Menon, R Morris, C Power, J Price, G Wannamethee, and P Whincup.

Author contributions

AFS and FD contributed to the idea and design of the study. AFS performed the analyses and drafted the manuscript. ADH, BJJ, RHHG, JPC, JW, and FD provided guidance during initial planning of the paper and during critical revision.

Guarantor

AFS had full access to all of the data and takes responsibility for the integrity of the data presented.

Funding

AFS is funded by UCLH NIHR Biomedical Research Centre and is a UCL Springboard Population Health Sciences Fellow. FD is funded by the MRC (K006215). AFS is funded by a UCL springboard population science fellowship. The **UCLEB Consortium** is supported by a

British Heart Foundation Programme Grant (RG/10/12/28456). We acknowledge the **British Regional Heart Study** team for data collection. The British Regional Heart study is supported by a British Heart Foundation grants (RG/08/013/25942) and BHF (RG/13/16/30528). The British Heart Foundation had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. The WHII study is supported by grants from the Medical Research Council (G0902037; ID85374), British Heart Foundation (RG/07/008/23674), Stroke Association, National Heart Lung and Blood Institute (5RO1 HL036310), National Institute on Aging (5RO1AG13196) Agency for Health Care Policy Research (HS06516), and the John D. and Catherine T. MacArthur Foundation Research Networks on Successful Midlife Development and Socio-economic Status and Health. Samples from the **ELSA** DNA Repository (EDNAR), received support under a grant (AG1764406S1) awarded by the National Institute on Ageing (NIA). ELSA was developed by a team of researchers based at the National Centre for Social Research, University College London and the Institute of Fiscal Studies. The data were collected by the National Centre for Social Research. **CaPS** was funded by the Medical Research Council and undertaken by the former MRC Epidemiology Unit (South Wales). The DNA bank was established with funding from a MRC project grant. The data archive is maintained by the University of Bristol. **EAS** is funded by the British Heart Foundation (Programme Grant RG/98002), with Metabochip genotyping funded by a project grant from the Chief Scientist Office of Scotland (Project Grant CZB/4/672). **MRC** NSHD is funded by the UK Medical Research Council. The **WHII** study is supported by grants from the Medical Research Council (G0902037; ID85374), British Heart Foundation (RG/07/008/23674), Stroke Association, National Heart Lung and Blood Institute (5RO1 HL036310), National Institute on Aging (5RO1AG13196) Agency for Health Care Policy Research (HS06516), and the John D. and Catherine T. MacArthur Foundation Research Networks on Successful Midlife Development and Socio-economic Status and Health.

Prior postings and presentations

This study and its results were neither previously published. An abstract containing this work was presented at the 2015 Mendelian Randomization Conference: From Population Health to Pharmaceutical Developments.

Reference List

(1) Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 2000 Jun 22;342(25):1887-92.

(2) Schmidt AF, Rovers MM, Klungel OH, Hoes AW, Knol MJ, Nielen M, et al. Differences in interaction and subgroup-specific effects were observed between randomized and nonrandomized studies in three empirical examples. J Clin Epidemiol 2013 Jun;66(6):599-607.

(3) Vandenbroucke JP. When are observational studies as credible as randomised trials? Lancet 2004 May 22;363(9422):1728-31.

(4) Vandenbroucke JP. Why do the results of randomised and observational studies differ? BMJ 2011;343:d7020.

(5) Hernan MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? Epidemiology 2006 Jul;17(4):360-72.

(6) Lawlor DA, Harbord RM, Sterne JA, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Stat Med 2008 Apr 15;27(8):1133-63.

(7) Martens EP, Pestman WR, de BA, Belitser SV, Klungel OH. Instrumental variables: application and limitations. Epidemiology 2006 May;17(3):260-7.

(8) Palmer TM, Sterne JA, Harbord RM, Lawlor DA, Sheehan NA, Meng S, et al. Instrumental variable estimation of causal risk ratios and causal odds ratios in Mendelian randomization analyses. Am J Epidemiol 2011 Jun 15;173(12):1392-403.

(9) Rassen JA, Schneeweiss S, Glynn RJ, Mittleman MA, Brookhart MA. Instrumental variable analysis for estimation of treatment effects with dichotomous outcomes. Am J Epidemiol 2009 Feb 1;169(3):273-84.

(10) Boef AG, Dekkers OM, Le Cessie S. Mendelian randomization studies: a review of the approaches used and the quality of reporting. Int J Epidemiol 2015 Apr;44(2):496-511.

(11) Burgess S, Dudbridge F, Thompson SG. Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. Stat Med 2015 Dec 13.

(12) Johnson T. Efficient calculation for multi-SNP genetic risk scores. 2013. Technical report, The Comprehensive R Archive Network. 2016.
Ref Type: Online Source

(13) Terza JV, Basu A, Rathouz PJ. Two-Stage Residual Inclusion Estimation: Addressing Endogeneity in Health Econometric Modeling. J Health Econ 2008 May 4;27(3):531-43.

(14) Stewart LA, Parmar MK. Meta-analysis of the literature or of individual patient data: is there a difference? Lancet 1993 Feb 13;341(8842):418-22.

(15) Pierce BL, Burgess S. Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. Am J Epidemiol 2013 Oct 1;178(7):1177-84.

(16) Thomas DC, Lawlor DA, Thompson JR. Re: Estimation of bias in nongenetic observational studies using "Mendelian triangulation" by Bautista et al. Ann Epidemiol 2007 Jul;17(7):511-3.

(17) Davison AC, Hinkley DV. Bootstrap Methods and their Application. 1 ed. New York: Cambridge University Press; 1997.

(18) Long JS, Ervin LH. Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model. The American Statistician 2000;54(3):217-24.

(19) Burgess S, Thompson SG. Avoiding bias from weak instruments in Mendelian randomization studies. Int J Epidemiol 2011 Jun;40(3):755-64.

(20) R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2013.

(21) Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Stat Med 2006 Dec 30;25(24):4279-92.

(22) Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat Med 2000 May 15;19(9):1141-64.

(23) Zeileis A. Object-Oriented Computation of Sandwich Estimators. Journal of Statistical Software 2006;16(9):1-16.

(24) Didelez V, Meng S, Sheehan NA. Assumptions of IV Methods for Observational Epidemiology. Statistical Science 2010;25(1).

(25) Rizzo LM. Statistical Computing with R. 1 ed. Chapman & Hall/CRC; 2007.

(26) Burgess S. Identifying the odds ratio estimated by a two-stage instrumental variable analysis with a logistic regression model. Stat Med 2013 Nov 30;32(27):4726-47.

(27) Heinze G. A comparative investigation of methods for logistic regression with separated or nearly separated data. Stat Med 2006 Dec 30;25(24):4216-26.

(28) logistf: Firth's bias reduced logistic regression [computer program]. Version R package version 1.21 2013.

(29) James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning: with Applications in R. 1st ed. Springer; 2013.

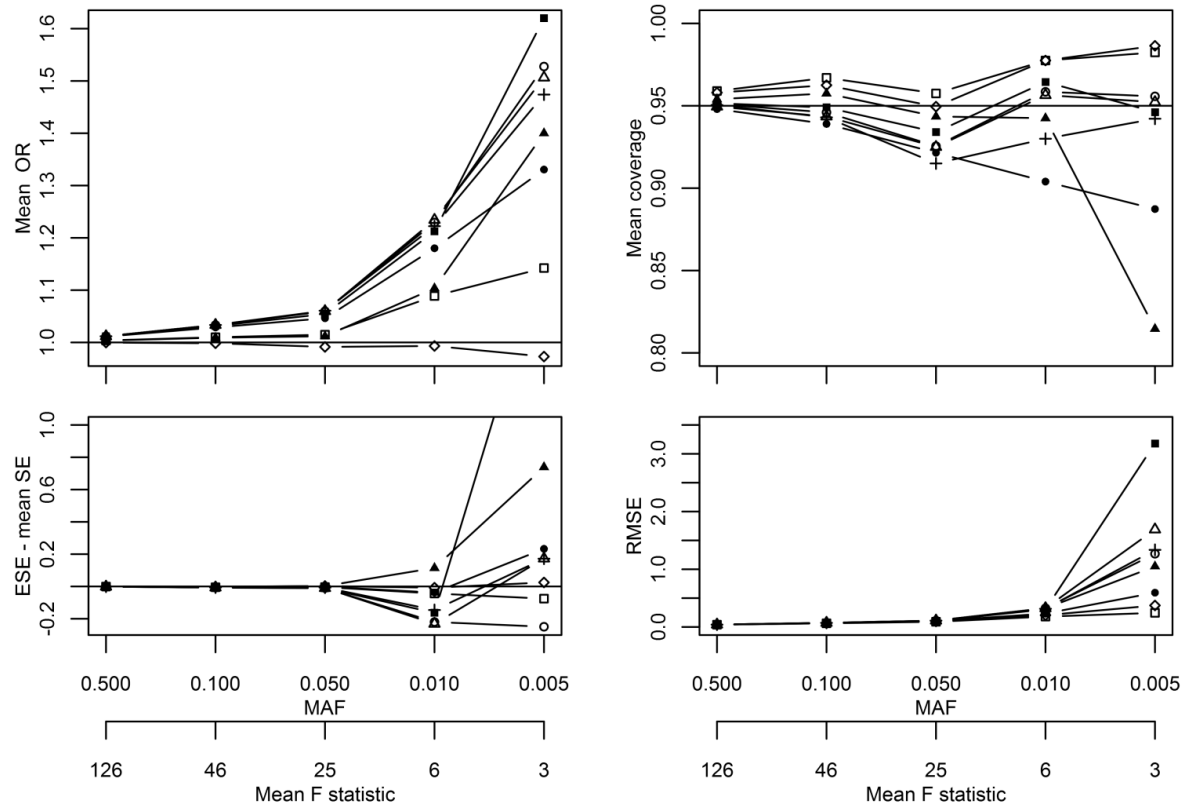
(30) Schmidt AF, Klungel OH, Groenwold RH. Adjusting for Confounding in Early Postlaunch Settings: Going Beyond Logistic Regression Models. Epidemiology 2016 Jan;27(1):133-42.

(31) Greenland S. Invited commentary: a critical look at some popular meta-analytic methods. Am J Epidemiol 1994 Aug 1;140(3):290-6.

(32) Schmidt AF, Groenwold RH, Knol MJ, Hoes AW, Nielen M, Roes KC, et al. Exploring interaction effects in small samples increases rates of false-positive and false-negative findings: results from a systematic review and simulation study. J Clin Epidemiol 2014 Jul;67(7):821-9.

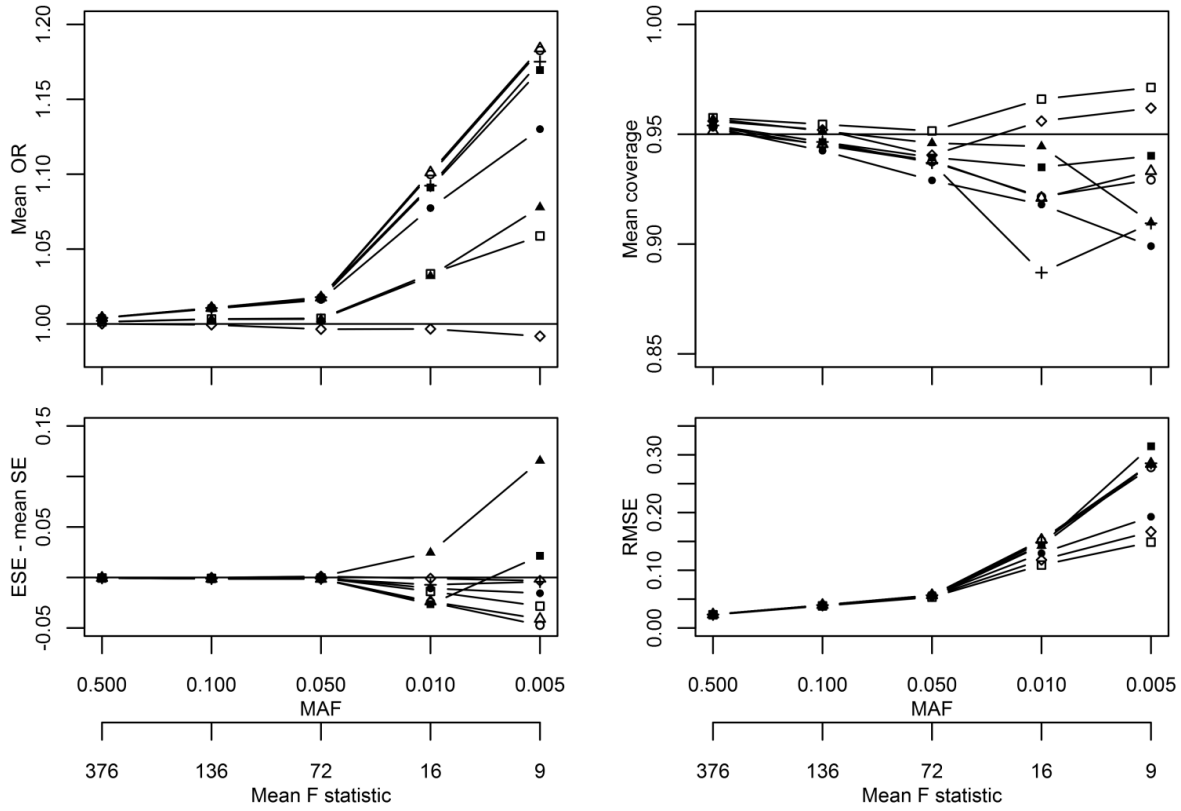
Figure captions

Figure 1 Simulation results from scenarios I comparing different IV estimators.*



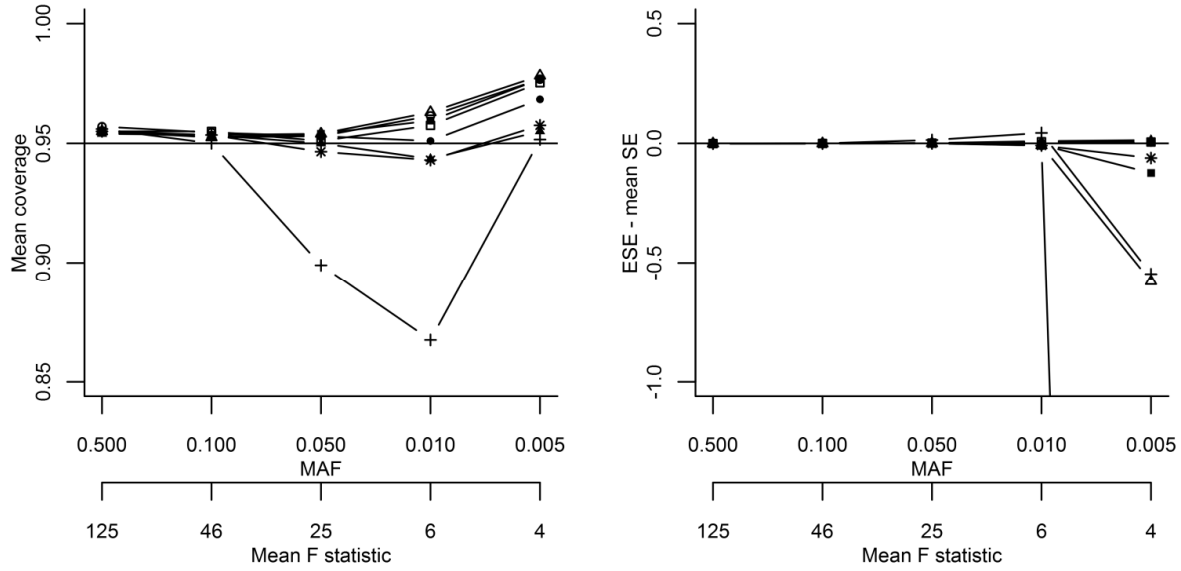
* solid line with a square symbol, delta method followed by meta-analysis [DM1]; solid line with a circle symbol, basic bootstrap [BB]; solid line with triangle symbol, outcome stratified bootstrap [OS]; solid line with a plus symbol, SNP stratified bootstrap [SS]; solid line with a filled out square symbol, double bootstrap [DB]; solid line with a filled out circle symbol, jackknife estimator [JK]; solid line with a filled out triangle symbol, robust variance estimator [RB]; solid line with a rhombus (diamond) symbol, meta-analysis followed by delta method [DM2]. The DB y-value of 2.071 is not depicted for a MAF of 0.005 on the bottom left graph.

Figure 2 Sensitivity analysis repeating simulation I comparing different IV estimators with an average of 60,000 subjects.*



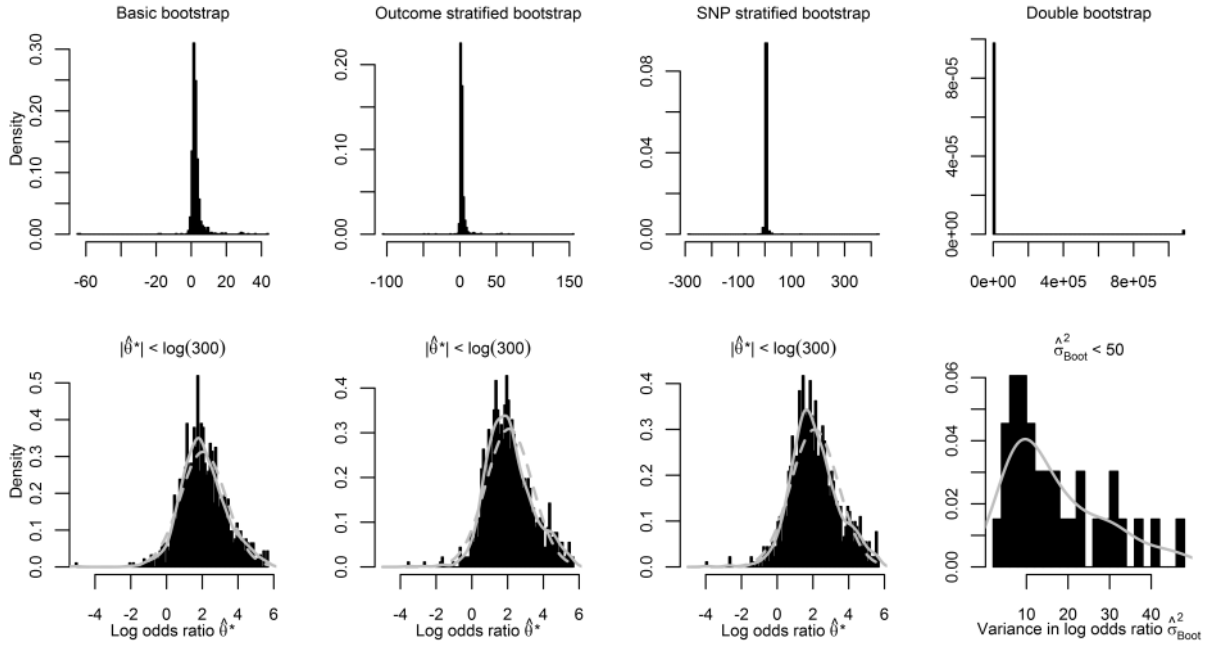
* solid line with a square symbol, delta method followed by meta-analysis [DM1]; solid line with a circle symbol, basic bootstrap [BB]; solid line with triangle symbol, outcome stratified bootstrap [OS]; solid line with a plus symbol, SNP stratified bootstrap [SS]; solid line with a filled out square symbol, double bootstrap [DB]; solid line with a filled out circle symbol, jackknife estimator [JK]; solid line with a filled out triangle symbol, robust variance estimator [RB]; solid line with a rhombus (diamond) symbol, meta-analysis followed by delta method [DM2].

Figure 3 Sensitivity analysis repeating simulation I comparing different IV estimators using a one stage meta-analysis design with an average of 20,000 subjects.*



* solid line with a square symbol, delta method followed by meta-analysis [DM1]; solid line with a circle symbol, basic bootstrap [BB]; solid line with triangle symbol, outcome stratified bootstrap [OS]; solid line with a plus symbol, SNP stratified bootstrap [SS]; solid line with a filled out square symbol, double bootstrap [DB]; solid line with a filled out circle symbol, jackknife estimator [JK]; solid line with a filled out triangle symbol, robust variance estimator [RB]; solid line with a star symbol, bootstrapped percentile method. The BB y-value of -13.463 is not depicted for a MAF of 0.005 on the right graph.

Figure 4 Bootstrap distributions for IV rs2965101 for the relation of LDL-C and CVD.*



* Solid grey lines indicate the non-parametric density (only presented in the second row), with dashed grey lines indicating the expected density given a normal distribution (not presented for the double bootstrap).

Table 1 Simulation scenarios assessing performance of different variance estimators for an instrumental variance analysis*.

Parameters	Scenario I	Scenario II	Scenario III	Scenario IV
Number of studies J	10	10	10	10
Sample size sample from a uniform distribution $U(a,b)$	(400, 3600)	(400, 3600)	(400, 3600)	(400, 3600)
Minor allele frequency p	{0.50, 0.10, 0.05,0.01, 0.005}	**0.15**	**0.05**	**0.01**
Effect of SNP on the phenotype α_1	0.50	0.50	0.50	0.50
Effect of unobserved confounder on the phenotype α_2	1.00	1.00	1.00	1.00
Intercept α_0	0.10	0.10	0.10	0.10
Log(OR) of the phenotype effect on the outcome δ_1	0.00	0.00	0.00	0.00
Log(OR) of the unobserved confounder effect on the outcome δ_2	1.50	1.50	1.50	1.50
Probability of the outcome	0.50	**{0.10, 0.05, 0.02, 0.01}**	{0.10, 0.05, 0.02, 0.01}	{0.10, 0.05, 0.02, 0.01}
Ln(odds) outcome intercept δ_0	0.00	**{-2.20, -2.94, -3.89, -4.60}**	{-2.20, -2.94, -3.89, -4.60}	{-2.20, -2.94, -3.89, -4.60}

* Changes from the previous scenario (on the left) are presented in bold. Alpha's represent mean differences, beta's the natural logarithm of the odds ratio.

Table 2 Instrumental variable analysis of the LDL-C effect on CVD using instrument rs11591147 and rs2965101*.

	Fixed effect 2 stage meta-analysis		Random effects 2 stage meta-analysis		Fixed effect 1 stage meta-analysis		Heterogeneity statistics£	
	Odds ratio (95%CI)	SE	Odds ratio (95%CI)	SE	Odds ratio (95%CI)	SE	$\dot{X}^2$(p-value)	T^2
Crude LDL-C association	1.06(1.01;1.11)	0.03	1.10(0.96;1.25)	0.07	1.06(1.01;1.11)	0.03	33.25(0.00)	0.02
rs11591147 IV LDL-C estimates								
DM before MA [DM1]	0.94(0.50;1.75)	0.32	094(0.501.75)	0.32	0.93(0.50;1.72)	0.31	4.88(0.43)	0.00
Basic bootstrap [BB]	1.24(0.48;3.18)	0.48	1.24(0.48;3.18)	0.48	0.93(0.49;1.76)	0.33	0.98(0.98)	0.00
Outcome stratified bootstrap [OS]	1.44(0.49;4.18)	0.55	1.44(0.49;4.18)	0.55	0.93(0.49;1.78)	0.33	0.08(1.00)	0.00
SNP stratified bootstrap [SS]	0.89(0.30;2.64)	0.55	0.89(0.30;2.64)	0.55	0.93(0.50;1.72)	0.31	0.38(1.00)	0.00
Double bootstrap [DB]	1.05(0.38;2.85)	0.51	1.05(0.38;2.85)	0.51	0.93(0.50;1.72)	0.31	1.58(0.93)	0.00
Jackknife [JK]	0.90(0.45;1.81)	0.35	0.90(0.45;1.81)	0.35	0.93(0.51;1.69)	0.31	4.05(0.58)	0.00
Robust HC1 [RB]	0.82(0.45;1.51)	0.31	0.81(0.41;1.60)	0.35	0.93(0.50;1.74)	0.32	5.85(0.33)	0.11
DM after MA [DM2]&	0.87(0.46;1.65)	0.33	0.85(0.40;1.80)	0.38	NA	NA	7.47(0.19)/6.21 (0.29)	0.01/0.03
Percentile Method$	NA	NA	NA	0.00	0.93(0.49;1.78)	NA	NA	NA
rs2965101 IV LDL-C estimates								
DM before MA [DM1]	1.55(0.35;17.90)	1.25	1.55(0.13;17.90)	1.25	8.16(0.50;132.64)	1.42	3.11(0.66)	0.00
Basic bootstrap [BB]	0.61(0.00;$2*10^{21}$)	25.35	0.61(0.00;$2*10^{21}$)	25.35	8.16(0.00;$9*10^{4}$)	4.77	0.01(1.00)	0.00
Outcome stratified bootstrap [OS]	4.61(0.00;$5*10^{30}$)	35.32	4.61(0.00;$5*10^{30}$)	35.32	8.16(0.00;$6*10^{7}$)	8.07	0.01(1.00)	0.00
SNP stratified bootstrap [SS]	6.67(0.00;10^{29})	33.29	6.67(0.00;10^{29})	33.29	8.16(0.00;$4*10^{15}$)	17.21	0.00(1.00)	0.00
Double bootstrap [DB]	1.55(0.00;$3*10^{11}$)	13.19	1.55(0.00;$3*10^{11}$)	13.19	8.16(0.00;10^{5})	4.93	0.03(1.00)	0.00
Jackknife [JK]	1.56(0.13;18.04)	1.25	1.56(0.13;18.04)	1.25	8.16(0.70;95.04)	1.25	3.13(0.65)	0.00
Robust HC1 [RB]	3.03(0.47;19.47)	0.95	2.72(0.20;37.48)	1.34	8.16(0.91;72.85)	1.12	8.11(0.13)	3.86
DM after MA [DM2]&	8.52(0.46;157.69)	1.49	9.01(0.36;223.27)	1.64	NA	NA	2.64(0.76)/6.14 (0.29)	0.00/0.00
Percentile Method$	NA	NA	NA	NA	8.16(0.88;10^{5})	NA	NA	NA

* The mean F-statistics of two-stage designed IPDMAs were 13.42, and 1.34 for rs11591147 and rs2965101, respectively. The F-statistics of one-stage designed IPDMA were 500.07, and 485.53 rs11591147 and rs2965101, respectively. The explained variance due to the instruments (measured as the squared spearman correlation coefficient) were 0. $70*10^{-2}$ and 0. $64*10^{-4}$. £ The heterogeneity statistics were determined for the fixed effect two-stage meta-analysis, tau-squared was calculated using the methods of moments estimator, chi-squared test statistic and p-value were based on the Q-test. $ The percentile method is only available for the one-stage design. & For DM2 the heterogeneity statistics represent the heterogeneity in $\hat{\gamma}_1$ and $\hat{\gamma}_3$, see equation 2. DM = delta method; MA = meta-analysis; SE = standard error.

Data sources for the empirical example of the LDL-C effect on CVD.

To empirically compare performance of the different estimators, see main text, we used SNP rs11591147 in the PCSK9 gene and SNP rs2965101 in the BCL3 gene as instruments to estimate the causal effect of LDL-C on CVD. Data were used from 6 studies in the UCLEB consortium(1) (overall n = 11581 with minimal n = 764, and maximum n = 3041; overall CVD events = 2050), British Regional Heat Study (BRHS)(2), Caerphilly Prospective Study (CaPS)(3), Edinburgh Artery Study (EAS)(4), English Longitudinal Study of Ageing (ELSA)(5), MRC National Survey of Health and Development (MRC46)(6), and Whitehall-II (WHII)(7). Between study heterogeneity was measured using the Q-test (8) and the method of moments estimator of the tau-squared (9). These two instruments were chosen because of a lack of pleiotropy (Appendix 1 figure 1), small correlation (r < 0.01), their different frequency (rs11591147 average p = 0.02, min 0.02; max 0.02; rs2965101 average p = 0.32, min 0.31; max 0.33), and different magnitudes of association with LDL-C (Spearman correlations of -0.082 and -0.008 for rs11591147 and rs2965101 with LDL-C).

Appendix table 1 Simulation results for scenario I assessing performance of different instrumental variable variance estimators under different levels of MAF with an outcome probability of 0.50 *.

	MAF = 0.500	MAF = 0.100	MAF = 0.050	MAF = 0.010	MAF = 0.005
Mean odds ratio (truth=1.000)					
Crude	1.778	1.826	1.840	1.851	1.853
DM before MA [DM1]	1.004	1.010	1.015	1.089	1.142
Basic bootstrap [BB]	1.012	1.033	1.060	1.229	1.528
Outcome stratified bootstrap [OS]	1.012	1.033	1.060	1.234	1.506
SNP stratified bootstrap [SS]	1.012	1.033	1.061	1.222	1.474
Double bootstrap [DB]	1.011	1.031	1.054	1.212	1.620
Jackknife [JK]	1.011	1.029	1.046	1.180	1.331
Robust HC1 [RB]	1.004	1.009	1.012	1.104	1.400
DM after MA [DM2]	1.000	0.998	0.991	0.993	0.973
Mean bias					
Crude	0.576	0.602	0.610	0.616	0.617
DM before MA [DM1]	0.004	0.010	0.015	0.085	0.133
Basic bootstrap [BB]	0.012	0.033	0.058	0.206	0.424
Outcome stratified bootstrap [OS]	0.012	0.033	0.058	0.211	0.410
SNP stratified bootstrap [SS]	0.012	0.033	0.059	0.201	0.388
Double bootstrap [DB]	0.011	0.031	0.053	0.193	0.482
Jackknife [JK]	0.011	0.028	0.045	0.166	0.286
Robust HC1 [RB]	0.004	0.009	0.012	0.099	0.336
DM after MA [DM2]	0.000	-0.002	-0.009	-0.007	-0.027
Coverage					
Crude	0.000	0.000	0.000	0.000	0.000
DM before MA [DM1]	0.959	0.967	0.958	0.978	0.982
Basic bootstrap [BB]	0.951	0.946	0.926	0.959	0.956
Outcome stratified bootstrap [OS]	0.950	0.944	0.925	0.957	0.952
SNP stratified bootstrap [SS]	0.951	0.943	0.915	0.930	0.942
Double bootstrap [DB]	0.952	0.949	0.934	0.965	0.946
Jackknife [JK]	0.948	0.939	0.922	0.904	0.887
Robust HC1 [RB]	0.954	0.958	0.944	0.943	0.815
DM after MA [DM2]	0.958	0.963	0.950	0.978	0.986
Mean SE					
Crude	0.012	0.012	0.013	0.013	0.013
DM before MA [DM1]	0.040	0.068	0.093	0.202	0.281
Basic bootstrap [BB]	0.041	0.071	0.104	0.472	1.445
Outcome stratified bootstrap [OS]	0.041	0.071	0.104	0.479	1.466
SNP stratified bootstrap [SS]	0.041	0.070	0.101	0.389	1.106
Double bootstrap [DB]	0.041	0.071	0.104	0.400	1.072
Jackknife [JK]	0.040	0.067	0.093	0.201	0.288
Robust HC1 [RB]	0.040	0.067	0.092	0.191	0.256
DM after MA [DM2]	0.040	0.068	0.094	0.212	0.347
ESE					
Crude	0.012	0.012	0.013	0.013	0.013
DM before MA [DM1]	0.040	0.063	0.089	0.159	0.205
Basic bootstrap [BB]	0.040	0.064	0.094	0.252	1.196
Outcome stratified bootstrap [OS]	0.040	0.064	0.094	0.248	1.639
SNP stratified bootstrap [SS]	0.040	0.064	0.094	0.244	1.278
Double bootstrap [DB]	0.040	0.064	0.093	0.237	3.143
Jackknife [JK]	0.040	0.063	0.091	0.164	0.522
Robust HC1 [RB]	0.040	0.065	0.094	0.306	0.995
DM after MA [DM2]	0.040	0.066	0.096	0.204	0.372
RMSE					
Crude	0.576	0.602	0.610	0.616	0.617
DM before MA [DM1]	0.040	0.063	0.091	0.180	0.245
Basic bootstrap [BB]	0.041	0.072	0.110	0.325	1.269
Outcome stratified bootstrap [OS]	0.041	0.072	0.111	0.325	1.689
SNP stratified bootstrap [SS]	0.041	0.072	0.111	0.316	1.336
Double bootstrap [DB]	0.041	0.071	0.107	0.305	3.179
Jackknife [JK]	0.041	0.069	0.102	0.233	0.595
Robust HC1 [RB]	0.040	0.065	0.095	0.322	1.050
DM after MA [DM2]	0.040	0.066	0.097	0.205	0.373
Number of failed models					
Crude	0	0	0	0	0
DM before MA [DM1]	0	0	0	0	12
Basic bootstrap [BB]	0	0	0	0	12
Outcome stratified bootstrap [OS]	0	0	0	0	12
SNP stratified bootstrap [SS]	0	0	0	0	12
Double bootstrap [DB]	0	0	0	0	12
Jackknife [JK]	0	0	0	0	12
Robust HC1 [RB]	0	0	0	0	15
DM after MA [DM2]	0	0	0	0	12
ESE – mean SE					
Crude	0.000	0.000	0.000	0.000	0.000
DM before MA [DM1]	-0.001	-0.005	-0.004	-0.044	-0.075

Basic bootstrap [BB]	-0.001	-0.007	-0.010	-0.220	-0.250
Outcome stratified bootstrap [OS]	-0.001	-0.007	-0.010	-0.231	0.173
SNP stratified bootstrap [SS]	-0.001	-0.007	-0.007	-0.146	0.172
Double bootstrap [DB]	-0.001	-0.008	-0.011	-0.163	2.071
Jackknife [JK]	-0.001	-0.004	-0.002	-0.038	0.233
Robust HC1 [RB]	0.000	-0.002	0.003	0.115	0.739
DM after MA [DM2]	0.000	-0.002	0.003	-0.008	0.025

* MAF = minor allele frequency; DM = delta method; MA = meta-analysis; SNP = single nucleotide polymorphism ; SE = standard error; ESE = empirical standard error; RMSE = square root of the mean squared error. The crude model regresses the log(odds) of the dichotomous outcome on the continuous phenotype. The mean F-statistics for the IV-phenotype association are: 126.42, 45.97, 24.67, 5.98, 3.47

Appendix table 2 Simulation results for scenario II assessing performance of different instrumental variable variance estimators under different probabilities for the outcome with the MAF fixed at 0.15. *

	Prob(y = 1) = 0.1	Prob(y = 1) = 0.05	Prob(y = 1) = 0.02	Prob(y = 1) = 0.01
Mean odds ratio (truth=1.000)				
Crude	1.269	1.272	1.273	1.272
DM before MA [DM1]	1.010	1.022	1.047	1.105
Basic bootstrap [BB]	1.018	1.034	1.099	1.290
Outcome stratified bootstrap [OS]	1.018	1.034	1.099	1.298
SNP stratified bootstrap [SS]	1.018	1.035	1.099	1.293
Double bootstrap [DB]	1.017	1.033	1.091	1.272
Jackknife [JK]	1.017	1.030	1.066	1.147
Robust HC1 [RB]	1.010	1.023	0.937	0.494
DM after MA [DM2]	1.008	1.019	1.048	1.113
Mean bias				
Crude	0.238	0.241	0.241	0.240
DM before MA [DM1]	0.010	0.021	0.046	0.100
Basic bootstrap [BB]	0.018	0.034	0.095	0.254
Outcome stratified bootstrap [OS]	0.018	0.033	0.094	0.261
SNP stratified bootstrap [SS]	0.018	0.034	0.095	0.257
Double bootstrap [DB]	0.017	0.032	0.087	0.241
Jackknife [JK]	0.016	0.029	0.064	0.137
Robust HC1 [RB]	0.010	0.023	-0.065	-0.706
DM after MA [DM2]	0.008	0.019	0.047	0.107
Coverage				
Crude	0.000	0.000	0.000	0.002
DM before MA [DM1]	0.965	0.958	0.958	0.945
Basic bootstrap [BB]	0.964	0.958	0.949	0.913
Outcome stratified bootstrap [OS]	0.964	0.956	0.952	0.894
SNP stratified bootstrap [SS]	0.963	0.958	0.952	0.910
Double bootstrap [DB]	0.965	0.962	0.960	0.925
Jackknife [JK]	0.961	0.953	0.954	0.936
Robust HC1 [RB]	0.953	0.948	0.905	0.720
DM after MA [DM2]	0.959	0.949	0.949	0.937
Mean SE				
Crude	0.016	0.022	0.034	0.047
DM before MA [DM1]	0.091	0.124	0.193	0.272
Basic bootstrap [BB]	0.095	0.130	0.217	0.381
Outcome stratified bootstrap [OS]	0.094	0.129	0.211	0.345
SNP stratified bootstrap [SS]	0.094	0.130	0.216	0.380
Double bootstrap [DB]	0.095	0.132	0.219	0.374
Jackknife [JK]	0.092	0.125	0.199	0.289
Robust HC1 [RB]	0.090	0.122	0.189	0.258
DM after MA [DM2]	0.091	0.124	0.194	0.273
ESE				
Crude	0.016	0.021	0.034	0.048
DM before MA [DM1]	0.087	0.118	0.182	0.259
Basic bootstrap [BB]	0.087	0.120	0.192	0.344
Outcome stratified bootstrap [OS]	0.087	0.120	0.187	0.305
SNP stratified bootstrap [SS]	0.087	0.120	0.191	0.333
Double bootstrap [DB]	0.087	0.120	0.191	0.324
Jackknife [JK]	0.087	0.119	0.184	0.271
Robust HC1 [RB]	0.089	0.122	0.848	2.308
DM after MA [DM2]	0.090	0.123	0.188	0.265
RMSE				
Crude	0.239	0.242	0.244	0.245
DM before MA [DM1]	0.087	0.120	0.188	0.277
Basic bootstrap [BB]	0.089	0.125	0.214	0.428
Outcome stratified bootstrap [OS]	0.089	0.124	0.210	0.402
SNP stratified bootstrap [SS]	0.089	0.124	0.213	0.420
Double bootstrap [DB]	0.089	0.124	0.210	0.404
Jackknife [JK]	0.088	0.123	0.195	0.303
Robust HC1 [RB]	0.089	0.124	0.850	2.413
DM after MA [DM2]	0.090	0.124	0.194	0.286
Number of failed models				
Crude	0	0	0	0
DM before MA [DM1]	0	0	0	0
Basic bootstrap [BB]	0	0	0	0
Outcome stratified bootstrap [OS]	0	0	0	0
SNP stratified bootstrap [SS]	0	0	0	0
Double bootstrap [DB]	0	0	0	0
Jackknife [JK]	0	0	0	0
Robust HC1 [RB]	0	0	6	39
DM after MA [DM2]	0	0	0	0
ESE – mean SE				
Crude	0.000	-0.001	0.000	0.001
DM before MA [DM1]	-0.005	-0.006	-0.011	-0.014

Basic bootstrap [BB]	-0.007	-0.010	-0.024	-0.037
Outcome stratified bootstrap [OS]	-0.007	-0.010	-0.024	-0.039
SNP stratified bootstrap [SS]	-0.007	-0.010	-0.025	-0.047
Double bootstrap [DB]	-0.008	-0.011	-0.028	-0.050
Jackknife [JK]	-0.005	-0.006	-0.015	-0.019
Robust HC1 [RB]	-0.001	-0.001	0.659	2.050
DM after MA [DM2]	-0.002	-0.001	-0.006	-0.008

* MAF = minor allele frequency; DM = delta method; MA = meta-analysis; SNP = single nucleotide polymorphism ; SE = standard error; ESE = empirical standard error; RMSE = square root of the mean squared error. The crude model regresses the log(odds) of the dichotomous outcome on the continuous phenotype. The mean F-statistic for the IV-phenotype association is 64.62.

Appendix table 3 Simulation results for scenario III assessing performance of different instrumental variable variance estimators under different probabilities for the outcome with the MAF fixed at 0.05. *

	Prob(y = 1) = 0.10	Prob(y = 1) = 0.05	Prob(y = 1) = 0.02	Prob(y = 1) = 0.01
Mean odds ratio (truth=1.000)				
Crude	1.276	1.278	1.279	1.277
DM before MA [DM1]	1.037	1.062	1.195	1.412
Basic bootstrap [BB]	1.063	1.126	1.540	0.206
Outcome stratified bootstrap [OS]	1.062	1.125	1.551	0.304
SNP stratified bootstrap [SS]	1.065	1.131	1.547	0.178
Double bootstrap [DB]	1.062	1.120	1.490	0.158
Jackknife [JK]	1.055	1.093	1.277	1.238
Robust HC1 [RB]	1.025	0.810	0.038	<0.001
DM after MA [DM2]	1.034	1.068	1.227	1.490
Mean bias				
Crude	0.244	0.245	0.246	0.244
DM before MA [DM1]	0.036	0.060	0.178	0.345
Basic bootstrap [BB]	0.061	0.119	0.432	-1.581
Outcome stratified bootstrap [OS]	0.060	0.118	0.439	-1.191
SNP stratified bootstrap [SS]	0.063	0.123	0.436	-1.727
Double bootstrap [DB]	0.060	0.114	0.399	-1.846
Jackknife [JK]	0.054	0.089	0.245	0.213
Robust HC1 [RB]	0.024	-0.210	-3.272	-10.835
DM after MA [DM2]	0.033	0.066	0.205	0.399
Coverage				
Crude	0.000	0.000	0.000	0.002
DM before MA [DM1]	0.963	0.958	0.921	0.899
Basic bootstrap [BB]	0.965	0.958	0.902	0.812
Outcome stratified bootstrap [OS]	0.962	0.965	0.896	0.818
SNP stratified bootstrap [SS]	0.954	0.950	0.888	0.785
Double bootstrap [DB]	0.964	0.967	0.918	0.803
Jackknife [JK]	0.952	0.949	0.908	0.893
Robust HC1 [RB]	0.938	0.885	0.530	0.147
DM after MA [DM2]	0.953	0.939	0.892	0.860
Mean SE				
Crude	0.016	0.022	0.034	0.048
DM before MA [DM1]	0.150	0.205	0.317	0.459
Basic bootstrap [BB]	0.171	0.252	0.640	1.965
Outcome stratified bootstrap [OS]	0.171	0.249	0.595	1.740
SNP stratified bootstrap [SS]	0.166	0.243	0.610	1.857
Double bootstrap [DB]	0.174	0.259	0.649	1.968
Jackknife [JK]	0.153	0.213	0.355	0.578
Robust HC1 [RB]	0.147	0.199	0.286	0.325
DM after MA [DM2]	0.151	0.206	0.319	0.461
ESE				
Crude	0.016	0.022	0.033	0.047
DM before MA [DM1]	0.136	0.185	0.281	0.378
Basic bootstrap [BB]	0.145	0.201	0.440	3.669
Outcome stratified bootstrap [OS]	0.145	0.199	0.424	3.356
SNP stratified bootstrap [SS]	0.145	0.197	0.435	3.773
Double bootstrap [DB]	0.147	0.203	0.453	3.646
Jackknife [JK]	0.138	0.187	0.296	0.676
Robust HC1 [RB]	0.289	1.520	5.414	8.208
DM after MA [DM2]	0.146	0.198	0.304	0.410
RMSE				
Crude	0.244	0.246	0.248	0.249
DM before MA [DM1]	0.141	0.194	0.333	0.512
Basic bootstrap [BB]	0.158	0.233	0.617	3.995
Outcome stratified bootstrap [OS]	0.157	0.231	0.611	3.561
SNP stratified bootstrap [SS]	0.158	0.233	0.616	4.150
Double bootstrap [DB]	0.158	0.232	0.603	4.086
Jackknife [JK]	0.148	0.207	0.385	0.709
Robust HC1 [RB]	0.290	1.535	6.326	13.593
DM after MA [DM2]	0.150	0.208	0.366	0.572
Number of failed models				
Crude	0	0	0	0
DM before MA [DM1]	0	0	0	0
Basic bootstrap [BB]	0	0	0	0
Outcome stratified bootstrap [OS]	0	0	0	0
SNP stratified bootstrap [SS]	0	0	0	0
Double bootstrap [DB]	0	0	0	0
Jackknife [JK]	0	0	0	0
Robust HC1 [RB]	2	8	123	383
DM after MA [DM2]	0	0	0	0
ESE – mean SE				
Crude	-0.001	0.000	-0.001	-0.001
DM before MA [DM1]	-0.014	-0.020	-0.036	-0.080

Basic bootstrap [BB]	-0.026	-0.051	-0.200	1.704
Outcome stratified bootstrap [OS]	-0.025	-0.050	-0.170	1.616
SNP stratified bootstrap [SS]	-0.021	-0.046	-0.175	1.916
Double bootstrap [DB]	-0.028	-0.056	-0.197	1.678
Jackknife [JK]	-0.014	-0.027	-0.059	0.099
Robust HC1 [RB]	0.142	1.321	5.128	7.883
DM after MA [DM2]	-0.005	-0.008	-0.016	-0.051

* MAF = minor allele frequency; DM = delta method; MA = meta-analysis; SNP = single nucleotide polymorphism ; SE = standard error; ESE = empirical standard error; RMSE = square root of the mean squared error. The crude model regresses the log(odds) of the dichotomous outcome on the continuous phenotype. The mean F-statistic for the IV-phenotype association is 24.66.

Appendix table 4 Simulation results for scenario IV assessing performance of different instrumental variable variance estimators under different probabilities for the outcome with the MAF fixed at 0.01. *

	Prob(y = 1) = 0.10	Prob(y = 1) = 0.05	Prob(y = 1) = 0.02	Prob(y = 1) = 0.01
Mean odds ratio (truth=1.000)				
Crude	1.279	1.280	1.284	1.281
DM before MA [DM1]	1.141	1.339	2.181	3.699
Basic bootstrap [BB]	1.239	0.772	0.013	<0.001
Outcome stratified bootstrap [OS]	1.236	0.778	0.014	<0.001
SNP stratified bootstrap [SS]	1.230	0.741	0.011	<0.001
Double bootstrap [DB]	1.245	0.537	0.008	<0.001
Jackknife [JK]	1.247	1.447	0.171	<0.001
Robust HC1 [RB]	0.074	<0.001	<0.001	<0.001
DM after MA [DM2]	1.204	1.572	3.021	5.324
Mean bias				
Crude	0.246	0.247	0.250	0.247
DM before MA [DM1]	0.132	0.292	0.780	1.308
Basic bootstrap [BB]	0.214	-0.259	-4.324	-9.525
Outcome stratified bootstrap [OS]	0.212	-0.251	-4.301	-9.540
SNP stratified bootstrap [SS]	0.207	-0.300	-4.502	-10.020
Double bootstrap [DB]	0.219	-0.622	-4.813	-9.912
Jackknife [JK]	0.220	0.370	-1.765	-9.437
Robust HC1 [RB]	-2.605	-9.740	-19.690	-20.864
DM after MA [DM2]	0.186	0.452	1.106	1.672
Coverage				
Crude	0.000	0.000	0.000	0.004
DM before MA [DM1]	0.968	0.942	0.867	0.793
Basic bootstrap [BB]	0.984	0.961	0.795	0.573
Outcome stratified bootstrap [OS]	0.985	0.965	0.809	0.572
SNP stratified bootstrap [SS]	0.967	0.936	0.732	0.486
Double bootstrap [DB]	0.986	0.951	0.744	0.493
Jackknife [JK]	0.943	0.922	0.793	0.386
Robust HC1 [RB]	0.524	0.165	0.003	0.000
DM after MA [DM2]	0.929	0.859	0.690	0.550
Mean SE				
Crude	0.017	0.022	0.034	0.048
DM before MA [DM1]	0.326	0.456	0.905	8.753
Basic bootstrap [BB]	1.397	3.769	7.178	8.861
Outcome stratified bootstrap [OS]	1.397	3.736	7.026	8.997
SNP stratified bootstrap [SS]	1.000	2.712	5.678	7.035
Double bootstrap [DB]	1.358	3.384	5.464	6.171
Jackknife [JK]	0.368	0.599	1.476	2.516
Robust HC1 [RB]	0.282	0.288	0.214	0.201
DM after MA [DM2]	0.346	0.483	0.776	8.670
ESE				
Crude	0.017	0.023	0.035	0.050
DM before MA [DM1]	0.260	0.345	0.515	2.994
Basic bootstrap [BB]	0.578	2.312	6.154	7.351
Outcome stratified bootstrap [OS]	0.645	2.282	5.888	7.380
SNP stratified bootstrap [SS]	0.636	2.370	6.072	7.316
Double bootstrap [DB]	0.844	4.171	5.500	6.673
Jackknife [JK]	0.294	0.619	4.938	8.682
Robust HC1 [RB]	4.862	7.871	6.122	4.493
DM after MA [DM2]	0.324	0.442	0.604	3.266
RMSE				
Crude	0.246	0.248	0.253	0.252
DM before MA [DM1]	0.292	0.452	0.934	3.268
Basic bootstrap [BB]	0.616	2.326	7.521	12.032
Outcome stratified bootstrap [OS]	0.679	2.296	7.292	12.061
SNP stratified bootstrap [SS]	0.669	2.389	7.559	12.406
Double bootstrap [DB]	0.872	4.217	7.309	11.949
Jackknife [JK]	0.367	0.721	5.244	12.823
Robust HC1 [RB]	5.516	12.523	20.620	21.342
DM after MA [DM2]	0.374	0.632	1.260	3.670
Number of failed models				
Crude	0	0	0	0
DM before MA [DM1]	0	0	0	0
Basic bootstrap [BB]	0	0	0	0
Outcome stratified bootstrap [OS]	0	0	0	0
SNP stratified bootstrap [SS]	0	0	0	0
Double bootstrap [DB]	0	0	0	0
Jackknife [JK]	0	0	0	0
Robust HC1 [RB]	93	365	782	1305
DM after MA [DM2]	0	0	0	0
ESE – mean SE				
Crude	0.000	0.000	0.000	0.002
DM before MA [DM1]	-0.066	-0.111	-0.390	-5.759

Basic bootstrap [BB]	-0.820	-1.457	-1.024	-1.510
Outcome stratified bootstrap [OS]	-0.752	-1.454	-1.138	-1.617
SNP stratified bootstrap [SS]	-0.363	-0.342	0.394	0.281
Double bootstrap [DB]	-0.514	0.787	0.036	0.502
Jackknife [JK]	-0.074	0.020	3.462	6.167
Robust HC1 [RB]	4.580	7.583	5.909	4.292
DM after MA [DM2]	-0.022	-0.041	-0.171	-5.404

* MAF = minor allele frequency; DM = delta method; MA = meta-analysis; SNP = single nucleotide polymorphism ; SE = standard error; ESE = empirical standard error; RMSE = square root of the mean squared error. The crude model regresses the log(odds) of the dichotomous outcome on the continuous phenotype. The mean F-statistic for the IV-phenotype association is 5.95.

Appendix table 5 Sensitivity analysis repeating simulation scenario 1 with an increased mean sample size of 60,000 subjects. *

	MAF = 0.500	MAF = 0.100	MAF = 0.050	MAF = 0.010	MAF = 0.005
Mean odds ratio (truth=1.000)					
Crude	1.779	1.826	1.839	1.852	1.853
DM before MA [DM1]	1.001	1.003	1.004	1.034	1.059
Basic bootstrap [BB]	1.004	1.011	1.018	1.100	1.183
Outcome stratified bootstrap [OS]	1.004	1.011	1.018	1.101	1.184
SNP stratified bootstrap [SS]	1.004	1.011	1.018	1.092	1.175
Double bootstrap [DB]	1.004	1.010	1.017	1.091	1.169
Jackknife [JK]	1.004	1.010	1.016	1.077	1.130
Robust HC1 [RB]	1.001	1.003	1.003	1.032	1.078
DM after MA [DM2]	1.000	0.999	0.996	0.997	0.992
Mean bias					
Crude	0.576	0.602	0.609	0.616	0.617
DM before MA [DM1]	0.001	0.003	0.004	0.033	0.057
Basic bootstrap [BB]	0.004	0.011	0.017	0.095	0.168
Outcome stratified bootstrap [OS]	0.004	0.011	0.017	0.096	0.169
SNP stratified bootstrap [SS]	0.004	0.011	0.018	0.088	0.161
Double bootstrap [DB]	0.004	0.010	0.017	0.087	0.157
Jackknife [JK]	0.004	0.010	0.016	0.075	0.122
Robust HC1 [RB]	0.001	0.003	0.003	0.031	0.075
DM after MA [DM2]	0.000	-0.001	-0.004	-0.003	-0.008
Coverage					
Crude	0.000	0.000	0.000	0.000	0.000
DM before MA [DM1]	0.958	0.955	0.952	0.966	0.971
Basic bootstrap [BB]	0.954	0.945	0.938	0.922	0.929
Outcome stratified bootstrap [OS]	0.952	0.946	0.938	0.921	0.933
SNP stratified bootstrap [SS]	0.954	0.947	0.937	0.887	0.909
Double bootstrap [DB]	0.954	0.947	0.939	0.935	0.940
Jackknife [JK]	0.953	0.943	0.929	0.918	0.899
Robust HC1 [RB]	0.957	0.952	0.946	0.944	0.910
DM after MA [DM2]	0.956	0.952	0.941	0.956	0.962
ESE					
Crude	0.007	0.007	0.007	0.007	0.007
DM before MA [DM1]	0.023	0.039	0.054	0.118	0.166
Basic bootstrap [BB]	0.023	0.040	0.055	0.142	0.269
Outcome stratified bootstrap [OS]	0.023	0.040	0.055	0.142	0.270
SNP stratified bootstrap [SS]	0.023	0.039	0.055	0.126	0.240
Double bootstrap [DB]	0.023	0.040	0.055	0.141	0.252
Jackknife [JK]	0.023	0.039	0.054	0.117	0.165
Robust HC1 [RB]	0.023	0.039	0.053	0.115	0.159
DM after MA [DM2]	0.023	0.039	0.054	0.119	0.170
Empirical SE					
Crude	0.007	0.007	0.007	0.007	0.007
DM before MA [DM1]	0.023	0.038	0.053	0.104	0.137
Basic bootstrap [BB]	0.023	0.038	0.054	0.117	0.222
Outcome stratified bootstrap [OS]	0.023	0.038	0.054	0.118	0.229
SNP stratified bootstrap [SS]	0.023	0.038	0.054	0.119	0.235
Double bootstrap [DB]	0.023	0.038	0.054	0.114	0.273
Jackknife [JK]	0.023	0.038	0.054	0.106	0.149
Robust HC1 [RB]	0.023	0.039	0.054	0.139	0.274
DM after MA [DM2]	0.023	0.039	0.055	0.118	0.167
RMSE					
Crude	0.576	0.602	0.609	0.616	0.617
DM before MA [DM1]	0.023	0.038	0.053	0.109	0.149
Basic bootstrap [BB]	0.023	0.040	0.057	0.151	0.278
Outcome stratified bootstrap [OS]	0.023	0.040	0.057	0.153	0.284
SNP stratified bootstrap [SS]	0.023	0.040	0.057	0.148	0.285
Double bootstrap [DB]	0.023	0.040	0.057	0.144	0.315
Jackknife [JK]	0.023	0.039	0.056	0.130	0.193
Robust HC1 [RB]	0.023	0.039	0.054	0.143	0.284
DM after MA [DM2]	0.023	0.039	0.055	0.118	0.167
Number of failed models					
Crude	0	0	0	0	0
DM before MA [DM1]	0	0	0	0	4
Basic bootstrap [BB]	0	0	0	0	4
Outcome stratified bootstrap [OS]	0	0	0	0	4
SNP stratified bootstrap [SS]	0	0	0	0	4
Double bootstrap [DB]	0	0	1	0	4
Jackknife [JK]	0	0	0	0	4
Robust HC1 [RB]	0	0	0	1	7
DM after MA [DM2]	0	0	0	0	4
ESE – mean SE					
Crude	0.000	0.000	0.000	0.000	0.000
DM before MA [DM1]	-0.001	-0.001	-0.001	-0.014	-0.028

Basic bootstrap [BB]	-0.001	-0.001	-0.001	-0.025	-0.047
Outcome stratified bootstrap [OS]	-0.001	-0.001	-0.001	-0.024	-0.041
SNP stratified bootstrap [SS]	0.000	-0.001	-0.001	-0.007	-0.004
Double bootstrap [DB]	-0.001	-0.001	-0.001	-0.026	0.021
Jackknife [JK]	0.000	-0.001	0.000	-0.011	-0.016
Robust HC1 [RB]	0.000	0.000	0.001	0.024	0.116
DM after MA [DM2]	0.000	0.000	0.001	-0.001	-0.003

* MAF = minor allele frequency; DM = delta method; MA = meta-analysis; SNP = single nucleotide polymorphism ; SE = standard error; ESE = empirical standard error; RMSE = square root of the mean squared error. The crude model regresses the log(odds) of the dichotomous outcome on the continuous phenotype. The mean F-statistics for the IV-phenotype association are 375.98, 136.00, 72.20, 15.98, and 8.55.

Appendix table 6 Sensitivity analysis repeating simulation scenario 1 with a mean sample size of 20,000 subjects using a one stage meta-analysis design. *

	MAF = 0.500	MAF = 0.100	MAF = 0.050	MAF = 0.010	MAF = 0.005
Mean odds ratio (truth=1.000)					
Crude	1.780	1.827	1.840	1.853	1.855
Delta method [DM]	1.000	1.000	0.997	0.991	0.984
Basic bootstrap [BB]	1.000	1.000	0.997	0.991	0.984
Outcome stratified bootstrap [OS]	1.000	1.000	0.997	0.991	0.984
SNP stratified bootstrap [SS]	1.000	1.000	0.997	0.991	0.984
Double bootstrap [DB]	1.000	1.000	0.997	0.991	0.984
Jackknife [JK]	1.000	1.000	0.997	0.991	0.984
Robust HC1 [RB]	1.000	1.000	0.997	0.991	0.984
Percentile Method	1.000	1.000	0.997	0.991	0.984
Mean bias					
Crude	0.576	0.603	0.610	0.617	0.618
Delta method [DM]	0.000	0.000	-0.003	-0.009	-0.016
Basic bootstrap [BB]	0.000	0.000	-0.003	-0.009	-0.016
Outcome stratified bootstrap [OS]	0.000	0.000	-0.003	-0.009	-0.016
SNP stratified bootstrap [SS]	0.000	0.000	-0.003	-0.009	-0.016
Double bootstrap [DB]	0.000	0.000	-0.003	-0.009	-0.016
Jackknife [JK]	0.000	0.000	-0.003	-0.009	-0.016
Robust HC1 [RB]	0.000	0.000	-0.003	-0.009	-0.016
Percentile Method	0.000	0.000	-0.003	-0.009	-0.016
Coverage					
Crude	0.000	0.000	0.000	0.000	0.000
Delta method [DM]	0.955	0.955	0.951	0.958	0.975
Basic bootstrap [BB]	0.957	0.955	0.953	0.962	0.977
Outcome stratified bootstrap [OS]	0.955	0.953	0.954	0.963	0.978
SNP stratified bootstrap [SS]	0.956	0.950	0.899	0.868	0.952
Double bootstrap [DB]	0.955	0.954	0.954	0.960	0.977
Jackknife [JK]	0.956	0.953	0.953	0.951	0.968
Robust HC1 [RB]	0.954	0.954	0.950	0.944	0.955
Percentile Method	0.955	0.954	0.947	0.943	0.958
Mean SE					
Crude	0.012	0.012	0.012	0.013	0.013
Delta method [DM]	0.041	0.068	0.093	0.211	0.307
Basic bootstrap [BB]	0.041	0.068	0.094	0.227	13.775
Outcome stratified bootstrap [OS]	0.041	0.068	0.094	0.227	0.891
SNP stratified bootstrap [SS]	0.041	0.067	0.080	0.174	0.859
Double bootstrap [DB]	0.041	0.068	0.094	0.224	0.436
Jackknife [JK]	0.041	0.068	0.093	0.213	0.310
Robust HC1 [RB]	0.041	0.068	0.093	0.209	0.298
Percentile Method	NA	NA	NA	NA	NA
ESE					
Crude	0.012	0.013	0.013	0.013	0.013
Delta method [DM]	0.040	0.067	0.094	0.218	0.312
Basic bootstrap [BB]	0.040	0.067	0.094	0.218	0.312
Outcome stratified bootstrap [OS]	0.040	0.067	0.094	0.218	0.312
SNP stratified bootstrap [SS]	0.040	0.067	0.094	0.218	0.312
Double bootstrap [DB]	0.040	0.067	0.094	0.218	0.312
Jackknife [JK]	0.040	0.067	0.094	0.218	0.312
Robust HC1 [RB]	0.040	0.067	0.094	0.218	0.312
Percentile Method	0.040	0.067	0.094	0.218	0.312
RMSE					
Crude	0.577	0.603	0.610	0.617	0.618
Delta method [DM]	0.040	0.067	0.094	0.218	0.313
Basic bootstrap [BB]	0.040	0.067	0.094	0.218	0.313
Outcome stratified bootstrap [OS]	0.040	0.067	0.094	0.218	0.313
SNP stratified bootstrap [SS]	0.040	0.067	0.094	0.218	0.313
Double bootstrap [DB]	0.040	0.067	0.094	0.218	0.313
Jackknife [JK]	0.040	0.067	0.094	0.218	0.313
Robust HC1 [RB]	0.040	0.067	0.094	0.218	0.313
Percentile Method	0.040	0.067	0.094	0.218	0.313
Number of failed models					
Crude	0	0	0	0	0
Delta method [DM]	0	0	0	0	0
Basic bootstrap [BB]	0	0	0	0	0
Outcome stratified bootstrap [OS]	0	0	0	0	0
SNP stratified bootstrap [SS]	0	0	0	0	0
Double bootstrap [DB]	0	0	0	0	0
Jackknife [JK]	0	0	0	0	0
Robust HC1 [RB]	0	0	0	0	0
Percentile Method	0	0	0	0	0
ESE – mean SE					
Crude	0.000	0.000	0.000	0.000	0.000
Delta method [DM]	-0.001	0.000	0.001	0.007	0.006

Basic bootstrap [BB]	-0.001	-0.001	0.000	-0.009	-13.463
Outcome stratified bootstrap [OS]	-0.001	-0.001	0.000	-0.009	-0.578
SNP stratified bootstrap [SS]	-0.001	0.000	0.014	0.044	-0.547
Double bootstrap [DB]	-0.001	-0.001	0.000	-0.006	-0.124
Jackknife [JK]	-0.001	0.000	0.001	0.005	0.002
Robust HC1 [RB]	-0.001	0.000	0.001	0.009	0.014
Percentile Method	-0.001	0.000	0.001	0.007	0.006
	-0.001	-0.001	0.000	-0.009	-0.061

The basic bootstrap percentile method does not estimate a standard error. * MAF = minor allele frequency; DM = delta method; MA = meta-analysis; SNP = single nucleotide polymorphism ; SE = standard error; ESE = empirical standard error; RMSE = square root of the mean squared error. The crude model regresses the log(odds) of the dichotomous outcome on the continuous phenotype. The mean F-statistics for the IV-phenotype association 125.33, 45.95, 24.84, 5.93, and 3.55.

Appendix table 7 Sensitivity analysis repeating simulation scenario 1 with between study variance, a mean sample size of 20,000 subjects, and using a two stage meta-analysis design. *

	MAF = 0.500	MAF = 0.100	MAF = 0.050	MAF = 0.010#	MAF = 0.005
Mean odds ratio (truth=1.000)					
Crude	1.375	1.418	1.437	1.453	1.454
DM before MA [DM1]	1.001	1.002	1.006	1.019	1.033
Basic bootstrap [BB]	1.003	1.006	1.012	1.007	1.050
Outcome stratified bootstrap [OS]	1.002	1.006	1.012	1.035	1.140
SNP stratified bootstrap [SS]	0.905	1.006	1.011	1.023	1.069
Double bootstrap [DB]	1.025	1.016	1.005	1.044	1.020
Jackknife [JK]	1.004	1.007	1.014	1.042	1.066
Robust HC1 [RB]	1.002	1.002	1.006	1.021	1.041
DM after MA [DM2]	0.978	0.996	0.996	$15.21*10^{9}$	0.773
Mean bias					
Crude	0.318	0.349	0.363	0.374	0.374
DM before MA [DM1]	0.001	0.002	0.006	0.019	0.033
Basic bootstrap [BB]	0.003	0.006	0.012	0.007	0.049
Outcome stratified bootstrap [OS]	0.002	0.006	0.012	0.035	0.131
SNP stratified bootstrap [SS]	-0.100	0.006	0.011	0.022	0.067
Double bootstrap [DB]	0.025	0.016	0.005	0.043	0.019
Jackknife [JK]	0.004	0.006	0.013	0.041	0.064
Robust HC1 [RB]	0.002	0.002	0.006	0.021	0.040
DM after MA [DM2]	-0.023	-0.004	-0.004	23.445	-0.257
Coverage					
Crude	0.000	0.000	0.000	0.000	0.000
DM before MA [DM1]	0.966	0.955	0.963	0.973	0.981
Basic bootstrap [BB]	0.962	0.951	0.955	0.971	0.983
Outcome stratified bootstrap [OS]	0.963	0.950	0.957	0.971	0.985
SNP stratified bootstrap [SS]	0.960	0.949	0.953	0.957	0.981
Double bootstrap [DB]	0.962	0.949	0.960	0.973	0.984
Jackknife [JK]	0.964	0.950	0.957	0.957	0.964
Robust HC1 [RB]	0.971	0.958	0.962	0.964	0.955
DM after MA [DM2]	0.999	0.996	0.997	0.999	0.998
Mean SE					
Crude	0.031	0.032	0.032	0.034	0.034
DM before MA [DM1]	0.022	0.036	0.049	0.102	0.142
Basic bootstrap [BB]	0.022	0.036	0.050	0.148	0.426
Outcome stratified bootstrap [OS]	0.023	0.035	0.049	0.152	0.481
SNP stratified bootstrap [SS]	0.024	0.036	0.049	0.132	0.336
Double bootstrap [DB]	0.024	0.037	0.051	0.158	0.474
Jackknife [JK]	0.022	0.036	0.049	0.106	0.172
Robust HC1 [RB]	0.023	0.038	0.054	0.124	0.281
DM after MA [DM2]	24.865	5.416	17.452	$19.60*10^{5}$	43.80*10
ESE					
Crude	0.037	0.037	0.038	0.039	0.040
DM before MA [DM1]	0.020	0.034	0.047	0.094	0.123
Basic bootstrap [BB]	0.021	0.038	0.048	1.347	3.026
Outcome stratified bootstrap [OS]	0.054	0.035	0.048	0.146	5.466
SNP stratified bootstrap [SS]	4.683	0.037	0.078	0.503	1.780
Double bootstrap [DB]	1.341	0.459	0.278	0.810	5.307
Jackknife [JK]	0.020	0.035	0.048	0.111	0.224
Robust HC1 [RB]	0.021	0.036	0.051	0.236	0.890
DM after MA [DM2]	1.681	0.972	1.338	$10.51*10^{2}$	10.286
RMSE					
Crude	0.320	0.351	0.365	0.376	0.377
DM before MA [DM1]	0.020	0.034	0.047	0.096	0.127
Basic bootstrap [BB]	0.021	0.038	0.050	1.347	3.027
Outcome stratified bootstrap [OS]	0.054	0.035	0.049	0.150	5.467
SNP stratified bootstrap [SS]	4.684	0.038	0.079	0.503	1.782
Double bootstrap [DB]	1.341	0.459	0.278	0.811	5.307
Jackknife [JK]	0.021	0.035	0.050	0.118	0.233
Robust HC1 [RB]	0.021	0.036	0.051	0.237	0.891
DM after MA [DM2]	1.681	0.972	1.338	$10.51*10^{10}$	10.289
Number of failed models					
Crude	0	0	0	0	0
DM before MA [DM1]	0	0	0	0	5
Basic bootstrap [BB]	0	0	0	0	5
Outcome stratified bootstrap [OS]	0	0	0	0	5
SNP stratified bootstrap [SS]	0	0	0	0	5
Double bootstrap [DB]	0	0	0	0	5
Jackknife [JK]	0	0	0	0	5
Robust HC1 [RB]	0	0	0	2	28
DM after MA [DM2]	0	0	0	0	5
ESE – mean SE					
Crude	0.005	0.006	0.006	0.005	0.007
DM before MA [DM1]	-0.002	-0.001	-0.002	-0.008	-0.019

Basic bootstrap [BB]	-0.001	0.002	-0.002	1.199	2.600
Outcome stratified bootstrap [OS]	0.031	-0.001	-0.002	-0.006	4.985
SNP stratified bootstrap [SS]	4.659	0.001	0.029	0.371	1.444
Double bootstrap [DB]	1.316	0.422	0.227	0.652	4.833
Jackknife [JK]	-0.002	-0.001	-0.001	0.005	0.052
Robust HC1 [RB]	-0.002	-0.002	-0.003	0.112	0.609
DM after MA [DM2]	-23.184	-4.444	-16.113	$-195.978*10^{4}$	-427.721

The large deviation of the DM2 method seen at a MAF of 0.010 is due to an single estimated log odds ratio of 46988.78, excluding this value results in a mean OR, mean bias, and empirical SE of 0.952, -0.049, 29.869 respectively. * MAF = minor allele frequency; DM = delta method; MA = meta-analysis; SNP = single nucleotide polymorphism ; SE = standard error; ESE = empirical standard error; RMSE = square root of the mean squared error. The crude model regresses the log(odds) of the dichotomous outcome on the continuous phenotype. The mean F-statistics for the IV-phenotype association are 408.31, 147.15, 77.54, 17.16, 9.02.

Appendix table 8 Sensitivity analysis repeating simulation scenario 1 with additional variance estimators. *

	MAF = 0.500	MAF = 0.100	MAF = 0.050	MAF = 0.010#	MAF = 0.005
Mean odds ratio (truth=1.000)					
Crude	1.778	1.825	1.839	1.850	1.854
TJ before MA [DM2]	1.003	1.010	1.018	1.066	1.126
TJ after MA [TJ2]	0.999	0.998	0.998	0.978	0.966
DM with BB before MA [DM1 BB]	1.002	1.009	1.013	1.106	1.238
DM with BB after MA [DM2 BB]	0.999	0.998	0.998	0.976	1.177
Mean bias					
Crude	0.576	0.602	0.609	0.615	0.617
TJ before MA [DM2]	0.003	0.010	0.018	0.064	0.119
TJ after MA [TJ2]	-0.001	-0.002	-0.002	-0.022	-0.035
DM with BB before MA [DM1 BB]	0.002	0.009	0.013	0.101	0.213
DM with BB after MA [DM2 BB]	-0.001	-0.002	-0.002	-0.024	0.163
Coverage					
Crude	0.000	0.000	0.000	0.000	0.000
TJ before MA [DM2]	0.950	0.960	0.949	0.942	0.940
TJ after MA [TJ2]	0.949	0.957	0.949	0.946	0.967
DM with BB before MA [DM1 BB]	0.954	0.966	0.962	0.969	0.889
DM with BB after MA [DM2 BB]	0.947	0.961	0.953	0.952	0.859
Mean SE					
Crude	0.012	0.012	0.012	0.013	0.013
TJ before MA [DM2]	0.040	0.067	0.091	0.192	0.261
TJ after MA [TJ2]	0.041	0.067	0.093	0.210	0.313
DM with BB before MA [DM1 BB]	0.041	0.105	0.151	238.758	28.277
DM with BB after MA [DM2 BB]	0.041	0.068	0.094	0.228	0.757
ESE					
Crude	0.012	0.013	0.013	0.012	0.013
TJ before MA [DM2]	0.040	0.065	0.092	0.190	0.243
TJ after MA [TJ2]	0.040	0.067	0.094	0.215	0.311
DM with BB before MA [DM1 BB]	0.040	0.085	0.108	3.274	2.183
DM with BB after MA [DM2 BB]	0.040	0.067	0.094	0.297	4.730
RMSE					
Crude	0.576	0.602	0.610	0.616	0.617
TJ before MA [DM2]	0.040	0.066	0.093	0.201	0.270
TJ after MA [TJ2]	0.040	0.067	0.094	0.216	0.313
DM with BB before MA [DM1 BB]	0.041	0.086	0.108	3.275	2.193
DM with BB after MA [DM2 BB]	0.040	0.067	0.094	0.298	4.733
Number of failed models					
Crude	0	0	0	0	0
TJ before MA [DM2]	0	0	0	0	22
TJ after MA [TJ2]	0	0	0	0	22
DM with BB before MA [DM1 BB]	0	0	0	0	22
DM with BB after MA [DM2 BB]	0	0	0	0	22
ESE – mean SE					
Crude	0.000	0.000	0.000	0.000	0.000
TJ before MA [DM2]	0.000	-0.001	0.000	-0.002	-0.018
TJ after MA [TJ2]	0.000	-0.001	0.000	0.005	-0.001
DM with BB before MA [DM1 BB]	-0.001	-0.020	-0.043	-235.485	-26.094
DM with BB after MA [DM2 BB]	0.000	-0.001	0.000	0.069	3.973

* MAF = minor allele frequency; TJ = Toby Johnson; MA = meta-analysis; BB = basic bootstrap; SE = standard error; ESE = empirical standard error; RMSE = square root of the mean squared error. The crude model regresses the log(odds) of the dichotomous outcome on the continuous phenotype. The mean F-statistics for the IV-phenotype association are 125.45, 46.14, 24.76, 5.99, 3.45.

Appendix table 9 Sensitivity analysis repeating simulation scenario 1 with using a continuous outcome. *

	MAF = 0.500	MAF = 0.100	MAF = 0.050	MAF = 0.010	MAF = 0.005
Mean, mean difference (truth=0.000)					
Crude	0.706	0.741	0.749	0.748	0.749
DM before MA [DM1]	0.005	0.031	0.191	0.113	0.191
Basic bootstrap [BB]	0.016	0.090	0.386	0.277	0.386
SNP stratified bootstrap [SS]	0.016	0.091	0.382	0.265	0.382
Double bootstrap [DB]	0.016	0.086	0.341	0.266	0.341
Jackknife [JK]	0.015	0.074	0.342	0.224	0.342
Robust HC1 [RB]	0.005	0.027	0.223	0.108	0.223
DM after MA [DM2]	-0.001	-0.002	-0.033	-0.021	-0.033
Mean bias					
Crude	0.706	0.734	0.741	0.748	0.749
DM before MA [DM1]	0.005	0.014	0.031	0.113	0.191
Basic bootstrap [BB]	0.016	0.046	0.090	0.277	0.386
SNP stratified bootstrap [SS]	0.016	0.046	0.091	0.265	0.382
Double bootstrap [DB]	0.016	0.044	0.086	0.266	0.341
Jackknife [JK]	0.015	0.040	0.074	0.224	0.342
Robust HC1 [RB]	0.005	0.013	0.027	0.108	0.223
DM after MA [DM2]	-0.001	-0.003	-0.002	-0.021	-0.033
Coverage					
Crude	0.000	0.000	0.000	0.000	0.000
DM before MA [DM1]	0.949	0.953	0.947	0.943	0.923
Basic bootstrap [BB]	0.921	0.885	0.831	0.820	0.834
SNP stratified bootstrap [SS]	0.923	0.878	0.816	0.778	0.776
Double bootstrap [DB]	0.922	0.885	0.836	0.807	0.759
Jackknife [JK]	0.921	0.884	0.836	0.684	0.602
Robust HC1 [RB]	0.946	0.944	0.932	0.873	0.728
DM after MA [DM2]	0.952	0.947	0.956	0.958	0.974
Mean SE					
Crude	0.007	0.007	0.007	0.007	0.007
DM before MA [DM1]	0.036	0.061	0.083	0.181	0.246
Basic bootstrap [BB]	0.037	0.063	0.091	0.392	0.934
SNP stratified bootstrap [SS]	0.037	0.063	0.089	0.325	0.743
Double bootstrap [DB]	0.037	0.063	0.090	0.306	0.530
Jackknife [JK]	0.036	0.060	0.081	0.164	0.212
Robust HC1 [RB]	0.036	0.060	0.081	0.164	0.200
DM after MA [DM2]	0.036	0.061	0.084	0.190	0.281
ESE					
Crude	0.007	0.007	0.007	0.007	0.007
DM before MA [DM1]	0.036	0.058	0.075	0.140	0.180
Basic bootstrap [BB]	0.037	0.060	0.082	0.220	0.420
SNP stratified bootstrap [SS]	0.037	0.060	0.082	0.219	0.496
Double bootstrap [DB]	0.037	0.060	0.081	0.199	2.100
Jackknife [JK]	0.037	0.059	0.079	0.151	0.240
Robust HC1 [RB]	0.037	0.060	0.080	0.167	0.322
DM after MA [DM2]	0.037	0.061	0.083	0.194	0.300
RMSE					
Crude	0.706	0.734	0.741	0.748	0.749
DM before MA [DM1]	0.037	0.060	0.081	0.180	0.263
Basic bootstrap [BB]	0.040	0.075	0.122	0.354	0.570
SNP stratified bootstrap [SS]	0.040	0.075	0.123	0.344	0.627
Double bootstrap [DB]	0.040	0.074	0.118	0.333	2.127
Jackknife [JK]	0.040	0.071	0.108	0.270	0.418
Robust HC1 [RB]	0.037	0.061	0.085	0.199	0.392
DM after MA [DM2]	0.037	0.061	0.083	0.195	0.302
Number of failed models					
Crude	0	0	0	0	0
DM before MA [DM1]	0	0	0	0	12
Basic bootstrap [BB]	0	0	0	0	12
SNP stratified bootstrap [SS]	0	0	0	0	12
Double bootstrap [DB]	0	0	0	0	12
Jackknife [JK]	0	0	0	0	12
Robust HC1 [RB]	0	0	0	0	12
DM after MA [DM2]	0	0	0	0	12
ESE – mean SE					
Crude	0.000	0.000	0.000	0.000	0.000
DM before MA [DM1]	0.000	-0.003	-0.008	-0.041	-0.066
Basic bootstrap [BB]	0.000	-0.004	-0.009	-0.172	-0.514
SNP stratified bootstrap [SS]	0.000	-0.003	-0.006	-0.106	-0.246
Double bootstrap [DB]	0.000	-0.003	-0.009	-0.107	1.570
Jackknife [JK]	0.001	-0.001	-0.002	-0.014	0.027
Robust HC1 [RB]	0.001	0.000	-0.001	0.004	0.121
DM after MA [DM2]	0.001	0.001	0.000	0.004	0.019

* MAF = minor allele frequency; DM = delta method; MA = meta-analysis; SNP = single nucleotide polymorphism ; SE = standard error; ESE = empirical standard error; RMSE = square root of the mean squared error. The crude model regresses the log(odds) of

the dichotomous outcome on the continuous phenotype. The mean F-statistics for the IV-phenotype association are: 125.63, 46.04, 24.93, 5.94, 3.49

Appendix table 10 Baseline characteristics of a 6 study IPDMA using SNPs rs11591147 and rs2965101 in an instrumental variables analysis of the LDL-C effect on CVD*.

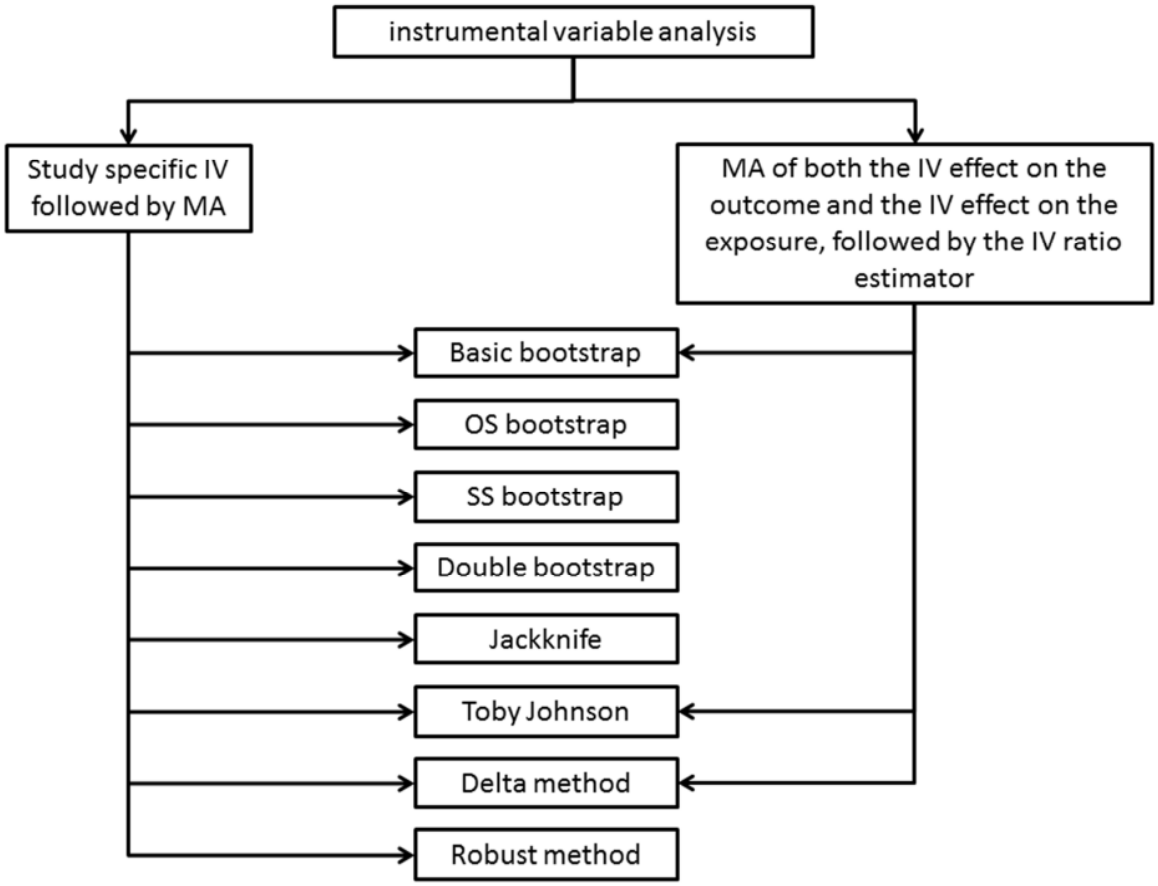
	BRHS		CaPS		EAS		ELSA	
	Mean(sd)	n	Mean(sd)	n	Mean(sd)	n	Mean(sd)	n
CVD	0.34	802	0.17	182	0.67	510	0.86	1624
Men	1.00	2342	1.00	1087	0.48	370	0.53	993
Age (years)	68.91(5.62)	2342	56.77(4.46)	1065	64.51(5.64)	764	73.69(9.44)	1883
Systolic blood pressure (mm Hg)	144.17(19.95)	2340	145.78(22.40)	1061	143.42(23.76)	763	139.01(19.66)	1662
Diastolic blood pressure (mm Hg)	81.88(12.86)	2340	84.56(12.01)	1061	77.46(12.20)	761	72.97(11.43)	1662
Cholesterol (mmol/L)	6.37(1.04)	2331	5.63(1.00)	1031	7.11(1.34)	763	5.71(1.28)	1873
HDL-C (mmol/L)	1.15(0.25)	2245	1.03(0.25)	1031	1.45(0.37)	760	1.49(0.39)	1872
LDL-C (mmol/L)	3.89(1.00)	2277	3.75(0.90)	1006	5.35(1.24)	760	3.43(1.06)	1835
Triglycerides (mmol/L)	2.06(1.23)	1500	1.92(1.14)	1031	1.53(0.87)	763	1.80(1.11)	1873
rs11591147 (n rare alleles)								
0	0.97	2261	0.97	1050	0.97	739	0.97	1819
1	0.03	81	0.03	37	0.11	81	0.06	121
2	0.00	0	0.00	0	0.00	0	0.00	0
rs2965101 (n rare alleles)								
0	0.46	1073	0.46	498	0.46	351	0.48	906
1	0.44	1023	0.44	477	0.44	334	0.42	795
2	0.10	245	0.10	112	0.03	25	0.03	64
Total sample size		2342		1087		764		1883

Appendix table 10 continued.

	MRC46		WHII		Total	
	Mean(sd)	n	Mean(sd)	n	Mean(sd)	n
CVD	0.06	144	0.14	409	0.18	2624
Men	0.50	1231	0.23	713	0.56	8422
Age (years)	53.00(0.00)	2464	48.94(5.98)	3041	59.63(11.21)	11559
Systolic blood pressure (mm Hg)	136.22(20.01)	2425	120.37(13.01)	3034	135.40(21.20)	11285
Diastolic blood pressure (mm Hg)	84.55(12.17)	2425	79.63(9.11)	3034	80.49(12.08)	11283
Cholesterol (mmol/L)	6.09(1.07)	2314	6.44(1.13)	3040	6.21(1.20)	11352
HDL-C (mmol/L)	1.67(0.52)	2149	1.41(0.40)	3023	1.39(0.44)	11087
LDL-C (mmol/L)	3.52(0.97)	2139	4.37(1.01)	2980	3.96(1.14)	11004
Triglycerides (mmol/L)	2.16(1.51)	2310	1.44(1.15)	3041	1.80(1.26)	10517
rs11591147 (n rare alleles)						
0	0.96	2368	0.97	2947	0.97	11184
1	0.04	94	0.03	92	0.03	393
2	0.00	2	0.00	2	0.00	4
rs11206510 (n rare alleles)						
0	0.47	1154	0.45	1362	0.46	5327
1	0.44	1074	0.45	1373	0.44	5093
2	0.10	236	0.10	306	0.10	1160
Total sample size		2464		3041		11581

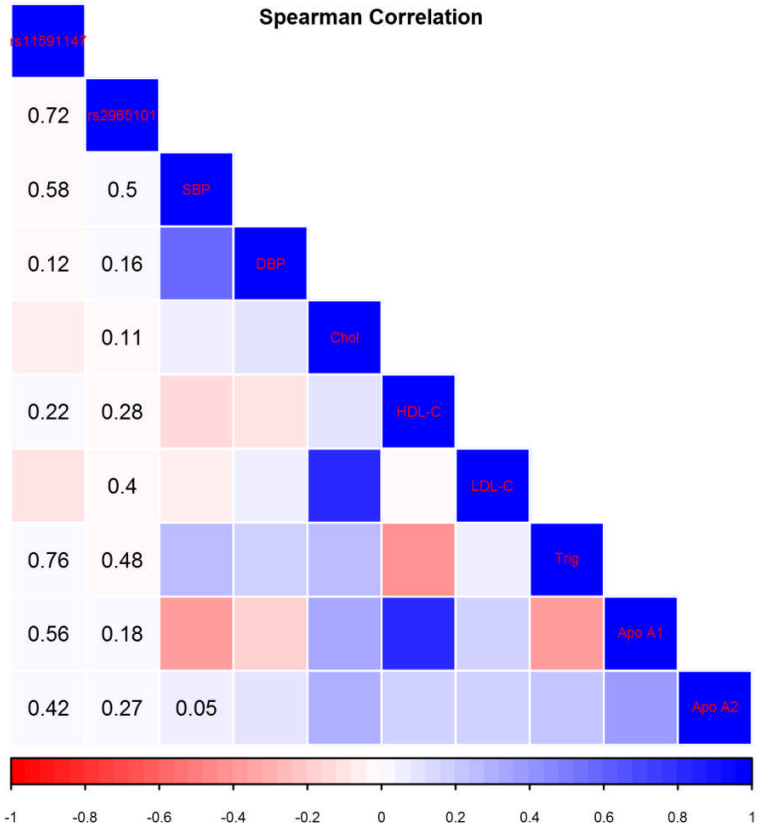
* The baseline numbers are based on complete data on CVD, and SNPS.

Appendix figure 1 Flowchart of the implementation of the different variance estimators in a simulation study of a two-stage meta-analysis of an instrumental variable analysis*.



*MA; meta-analysis, SS; SNP stratified, OS; outcome stratified.

Appendix Figure 2 Spearman pairwise correlation matrix for PCSK9 SNPs rs13465, rs6511720, and multiple phenotypes; with p-values for non-significant associations depicted (alpha = 0.05).



Reference List

(1) Shah T, Engmann J, Dale C, Shah S, White J, Giambartolomei C, et al. Population genomics of cardiometabolic traits: design of the University College London-London School of Hygiene and Tropical Medicine-Edinburgh-Bristol (UCLEB) Consortium. PLoS One 2013;8(8):e71345.

(2) Shaper AG, Pocock SJ, Walker M, Cohen NM, Wale CJ, Thomson AG. British Regional Heart Study: cardiovascular risk factors in middle-aged men in 24 towns. Br Med J (Clin Res Ed) 1981 Jul 18;283(6285):179-86.

(3) Bainton D, Miller NE, Bolton CH, Yarnell JWG, Sweetnam PM, Baker IA, et al. Plasma triglyceride and high density lipoprotein cholesterol as predictors of ischaemic heart disease in British men: The Caerphilly and Speedwell Collaborative Heart Disease Studies. Br Heart J 1992 Jul;68(1):60-6.

(4) Fowkes FG, Housley E, Cawood EH, Macintyre CC, Ruckley CV, Prescott RJ. Edinburgh Artery Study: prevalence of asymptomatic and symptomatic peripheral arterial disease in the general population. Int J Epidemiol 1991 Jun;20(2):384-92.

(5) Marmot MG, Banks J, Blundell R, Lessof C, Nazroo J. Health, wealth and lifestyles of the older population in England: ELSA 2002. 2003.

(6) Kuh D, Pierce M, Adams J, Deanfield J, Ekelund U, Friberg P, et al. Cohort profile: updating the cohort profile for the MRC National Survey of Health and Development: a new clinic-based data collection for ageing research. Int J Epidemiol 2011 Feb;40(1):e1-e9.

(7) Marmot MG, Davey Smith G, Stansfeld S, Patel C, North F, Head J, et al. Health inequalities among British civil servants: the Whitehall II study. Lancet 1991 Jun 8;337(8754):1387-93.

(8) Schmidt AF, Groenwold RH, Knol MJ, Hoes AW, Nielen M, Roes KC, et al. Exploring interaction effects in small samples increases rates of false-positive and false-negative findings: results from a systematic review and simulation study. J Clin Epidemiol 2014 Jul;67(7):821-9.

(9) Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002 Jun 15;21(11):1539-58.