Original Article

Impact of Selection Bias on Estimation of Subsequent Event Risk

Yi-Juan Hu, PhD*; Amand F. Schmidt, PhD*; Frank Dudbridge, PhD; Michael V. Holmes, PhD; James M. Brophy, MD, PhD; Vinicius Tragante, PhD; Ziyi Li, MS; Peizhou Liao, PhD; Arshed A. Quyyumi, MD; Raymond O. McCubrey, MS; Benjamin D. Horne, PhD; Aroon D. Hingorani, PhD; Folkert W. Asselbergs, MD, PhD†; Riyaz S. Patel, MD†; Qi Long, PhD†; on behalf of the GENIUS-CHD Consortium‡

Background—Studies of recurrent or subsequent disease events may be susceptible to bias caused by selection of subjects who both experience and survive the primary indexing event. Currently, the magnitude of any selection bias, particularly for subsequent time-to-event analysis in genetic association studies, is unknown.

Methods and Results—We used empirically inspired simulation studies to explore the impact of selection bias on the marginal hazard ratio for risk of subsequent events among those with established coronary heart disease. The extent of selection bias was determined by the magnitudes of genetic and nongenetic effects on the indexing (first) coronary heart disease event. Unless the genetic hazard ratio was unrealistically large (>1.6 per allele) and assuming the sum of all nongenetic hazard ratios was <10, bias was usually <10% (downward toward the null). Despite the low bias, the probability that a confidence interval included the true effect decreased (undercoverage) with increasing sample size because of increasing precision. Importantly, false-positive rates were not affected by selection bias.

Conclusions—In most empirical settings, selection bias is expected to have a limited impact on genetic effect estimates of subsequent event risk. Nevertheless, because of undercoverage increasing with sample size, most confidence intervals will be over precise (not wide enough). When there is no effect modification by history of coronary heart disease, the false-positive rates of association tests will be close to nominal. **(*Circ Cardiovasc Genet*. 2017;10:e001616. DOI: 10.1161/CIRCGENETICS.116.001616.)**

Key Words: alleles ■ confidence intervals ■ genetic association studies ■ risk ■ sample size ■ selection bias

Advances in acute treatments and public health policies have shifted the balance of coronary heart disease (CHD) such that an increasing number of individuals are surviving a first clinical CHD event (eg, myocardial infarction [MI]) and living with established CHD.[1] In the United Kingdom and United States, these numbers are estimated to be 3 and 16 million, respectively.[2] These individuals are at very high risk of subsequent or recurrent coronary and cardiovascular events, which can be fatal, disabling, or require ongoing costly interventions.[2]

See Editorial by Dungan
See Clinical Perspective

Despite the extent of the problem, little is known about risk factors for subsequent CHD events in comparison to first CHD events. As a result, risk stratification in survivors is limited while secondary prevention advice beyond lipid management has remained largely unaltered over 3 decades.[3] More importantly novel therapies beyond lipid lowering, antiplatelets, and antihypertensives have been slow to emerge.

Received September 5, 2016; accepted July 7, 2017.

From the Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA (Y.-J.H., Z.L., P.L.); Groningen Research Institute of Pharmacy, University of Groningen, the Netherlands (A.F.S.); Institute of Cardiovascular Science and The Farr Institute, University College London, United Kingdom (A.F.S., A.D.H., F.W.A., R.P.); Department of Non-Communicable Disease Epidemiology, London School of Hygiene & Tropical Medicine, United Kingdom (F.D.); Department of Health Sciences, University of Leicester, United Kingdom (F.D.); Clinical Trial Service Unit & Epidemiological Studies Unit (CTSU), Nuffield Department of Population Health, University of Oxford, United Kingdom (M.V.H.); Medical Research Council Population Health Research Unit at the University of Oxford, United Kingdom (M.V.H.); Department of Medicine, McGill University, Montreal Quebec, Canada (J.M.B.); Division of Heart and Lungs, Department of Cardiology, University Medical Center Utrecht, The Netherlands (V.T., F.W.A.); Division of Cardiology, Department of Medicine, Emory Clinical Cardiovascular Research Institute, Emory University School of Medicine, Atlanta, GA (A.A.Q.); Intermountain Heart Institute, Intermountain Medical Center, Murray, UT (R.O.M., B.D.H.); Department of Biomedical Informatics, University of Utah, Salt Lake City (B.D.H.); and Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia (Q.L.).

*Drs Hu and Schmidt contributed equally as joint first authors.

†Drs Asselbergs, Patel, and Long contributed equally as joint senior authors.

‡A list of GENIUS-CHD Consortium members is given in the Appendix.

The Data Supplement is available at http://circgenetics.ahajournals.org/lookup/suppl/doi:10.1161/CIRCGENETICS.116.001616/-/DC1.

Correspondence to Yi-Juan Hu, PhD, Department of Biostatistics & Bioinformatics, Rollins School of Public Health, Emory University, 1518 Clifton Rd NE, Atlanta, GA 30322. E-mail yijuan.hu@emory.edu or Amand F. Schmidt, PhD, Institute of Cardiovascular Science & The Farr Institute, University College London, 222 Euston Rd, Room 206, London NW1 2DA, United Kingdom. E-mail amand.schmidt@ucl.ac.uk

The high residual risk in those with CHD suggests that the existence of other risk factors, such as those predisposing to rupture of atherosclerotic plaques rather than to the development and progression of atherosclerosis.[4] In this regard, identification of genetic variants associating with subsequent CHD events may offer the most promising approach to identifying relevant and novel molecular pathways, which may in turn be amenable to therapeutic modification.

A key reason for our knowledge deficit here is the lack of suitable resources to facilitate prospective study of genetic and nongenetic risk factors among individuals with established CHD. Few cohorts of CHD individuals exist relative to general population cohorts that are more common. In response, the GENIUS-CHD consortium (The Genetics of Subsequent Coronary Heart Disease)[5] has been developed, bringing together >60 prospective studies of >250 000 individuals with established CHD, including data on genes, biomarkers, and incidence of subsequent fatal and nonfatal events.

Despite such efforts, a methodological barrier to studying subsequent CHD events (eg, a second MI after a first nonfatal MI) is the problem of selection bias. Here, we consider 2 sources of selection bias, index event bias and survival bias. Index event bias occurs when selecting a subset of subjects based on the occurrence of an index event (eg, the first clinical event). This selection can induce correlations between previously independent risk factors among those selected,[6,7] which can lead to biased associations. To be more specific, those suffering a first event on the basis of exposure to a particularly strong risk factor may have lower levels of exposure to other individually weaker, independent risk factors. This then mitigates the risk of a subsequent event, despite ongoing exposure to the strong risk factor. A frequently cited example of index event bias is the association of patent foramen ovale with the first occurrence of cryptogenic stroke but not with stroke recurrence.[7] Index event bias may also contribute to the apparent protective effect of adiposity on risk of subsequent CHD events, the so-called obesity paradox.[8] Moreover, because subjects can only be included in a study after surviving up to the time of inclusion, survival bias may also inflate the bias further still. Thus, in the context of subsequent event studies for CHD, the impact of selection bias may be important because any bias caused by selecting individuals on an indexing event (ie, index event bias) is compounded by selecting surviving subjects (ie, survival bias).

The influence of these biases on estimates of genetic effects on subsequent CHD events is currently unknown. This is important because, contrary to most observational studies,[9] genetic studies are less prone to confounding bias,[10] thus leaving selection bias as the potentially major source of bias.[11] In this simulation study, we sought to quantify the magnitude of index event bias and survival bias on the associations of genetic and nongenetic exposures with time-to-event data as well as binary data in relation to subsequent CHD risk.

Methods

To quantify the impact of index event bias and survival bias, we simulated data of the type anticipated to be encountered in the GENIUS-CHD consortium.[5] We focus on the marginal (ie, unconditional) association of a genetic or nongenetic exposure of interest while averaging over all other covariates because (1) the primary analysis in the GENIUS-CHD consortium similarly focuses on marginal associations and (2) a comprehensive set of other risk factors may not be collected in all cohorts/sites to allow estimation of a uniform conditional association. More specifically, we focus on the estimators of marginal associations from logistic or Cox regression that do not correct for index event bias and survival bias. Refer to the study by Jiang et al,[12] for a detailed discussion on marginal and conditional associations.

Specifically, we simulate data with the aim of estimating the effect of a gene variant or a biomarker on subsequent CHD events when the first event can be either fatal or nonfatal. The term subsequent CHD events is used in preference to recurrent given that fatal events are not recurrent and also to capture the wide range of CHD events that may be of interest to investigators both individually (eg, subsequent MI, subsequent revascularization, subsequent heart failure admissions) and as composite end points. For the purposes of these simulations described below, we use MI as our exemplar indexing event and subsequent CHD event.

Thus, let D_1 denote the first event and S be the indicator of surviving the first event. Using the notation, we define 3 populations (Figure 1): population 1 the general population that was at risk of a first event, population 2 the subpopulation who had a first event, and population 3 the subpopulation who had a first event and survived. We study the index event bias alone using population 2, as well as the combined effect of index event bias and survival bias using population 3. In the remaining Methods section, we briefly outline the methods and defer technical details to the Data Supplement.

Scenario 1

We first consider the scenario depicted in the directed acyclic graph in Figure 2A. Here G denotes the genotype (coded as the number of minor alleles) at a single nucleotide polymorphism of interest, X denotes the combined effect of all the remaining (known and unknown) genetic and nongenetic exposures (eg, diet and exercise) that are assumed to be independent of G, and D_2 denotes the subsequent event. Note that we assume D_1 affects survival not directly but through G and X. We initially set the minor allele frequency (MAF) of G, π, to 0.3, which is the median MAF of discovered genetic variants for MI based on empirical GWAS data (Genome-Wide Association Study [CARDIoGRAMplusC4D Consortium[13]]). We simulated X to be normally distributed with mean 0 and SD 1. The first event D_1 is binary throughout and is generated from a logistic regression model

$$\log\{P(D_1 = 1)/P(D_1 = 0)\} = \alpha_0 + \alpha_G G + \alpha_X X, \qquad (1)$$

where α_0 is set to achieve an overall disease rate of c_1. We initially set c_1 to 0.2%, after the approximate incidence of MI in the general population[2]; in a later sensitivity analysis, we vary c_1 between 0.1% and 1% to capture the variable MI rates in different populations and conditions as well as different type of MI (eg, ST-segment–elevation and non–ST-segment–elevation infarcts). We manipulate exp (α_G), the odds ratio (OR) of G from 1 to 1.3, 1.6, 2, and 3. We also manipulate exp (α_X), the OR of X, from 3 and 5 to 10, where an OR of 10

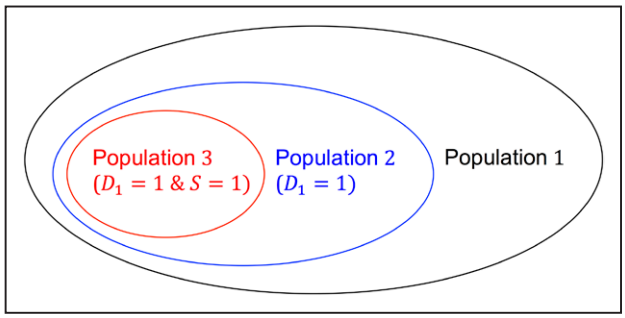


Figure 1. Three populations. D_1 denotes the first/index event. S is the indicator of surviving the first event. Population 1, general population; population 2, those with a first event (fatal and nonfatal cases); population 3, those with a nonfatal first event.

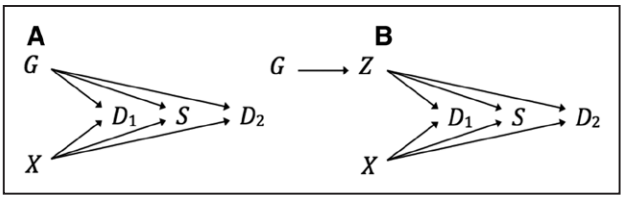


Figure 2. Directed acyclic graphs. **A**, Scenario 1: the genetic variant (*G*) associates with risk of first event (D_1), survival (*S*), and risk of subsequent event (D_2). **B**, Scenario 2: the genetic variant encodes a biomarker (*Z*) that associates with risk of first event, survival, and risk of subsequent event.

means that the total effects of all the possible protective and harmful genetic and nongenetic exposures (except *G*) sum up to 10, which is a plausible extreme of these influences. Similarly, the survival indicator *S* is binary and is generated from a logistic regression model

$$\log\{P(S=0)/P(S=1)\} = \gamma_0 + \gamma_G G + \gamma_X X, \quad (2)$$

where γ_0 is set to achieve an overall index event death rate of c_s. In empirical CHD data, c_s can be as high as 30% if all deaths[2] from the index MI (including those who get treated in hospital and those who die suddenly at home and never get to hospital) are counted; among those who get treated in hospital, c_s can be as low as 10%. Thus, we initially set c_s to 20%, a value between the 2 extremes. When D_2 represents time to subsequent event, it is generated from a proportional hazards model (assuming the baseline time-to-event follows an exponential distribution with rate parameter 2)

$$\lambda(T \mid G, X) = 2T\exp(\beta_G G + \beta_X X), \quad (3)$$

with the censoring rate of $(1-c_2)$. We initially set c_2, the incidence of subsequent CHD events, to 5%, which approximates the observational occurrence of subsequent MI.[2] When D_2 is binary, it is generated from a logistic regression model

$$\log\{P(D_2=1)/P(D_2=0)\} = \beta_0 + \beta_G G + \beta_X X, \quad (4)$$

where β_0 is set to achieve the occurrence of the subsequent MI of 5%. In all simulation studies, we set $\alpha_G=\gamma_G=\beta_G$ and $\alpha_X=\gamma_X=\beta_X$, that is, *G* has equal conditional effects on both initial fatal and nonfatal events as well as subsequent CHD events, and *X* also has equal conditional effects on the 3 outcomes. We use a sample size of 25 000, which represents the median sample size of >80 GWAS (Data Supplement). In all simulations, we estimate the marginal effect of *G* on D_2, which is the hazard ratio (HR) or OR of *G* in the standard Cox model or logistic regression model with *G* as the sole risk factor; we refer to it as the naive estimate.

Scenario 2

We also consider a mediation setting (Figure 2B) in which *G* influences D_1, *S*, and D_2 through a known biomarker (and through no other path), denoted as *Z*. We assume that 5% or 10% variance of *Z* is explained by *G*. To reflect the direct effect of *Z*, we replace $\alpha_G G$, $\gamma_G G$, and $\beta_G G$ in Equations (1)–(4) by $\alpha_Z Z$, $\gamma_Z Z$, and $\beta_Z Z$, respectively. Here, we focus on the estimates for the marginal *G* and D_2 association and the marginal *Z* and D_2 association using the standard Cox model or logistic regression model with *G* or *Z* as the sole risk factor; we again refer to them as the naive estimates.

Calculation of the True Marginal Association

To calibrate bias of the naive estimates for the marginal association (ie, HR or OR) of *G* on D_2 in scenario 1 and for the marginal associations of *G* on D_2 and *Z* on D_2 in scenario 2, we calculate the corresponding true marginal associations. This is achieved by the counterfactual method, in which we simulate the outcome in both the presence and the absence of the exposure *G* conditional on the distribution of *X* observed in the population of interest (ie, population 2 or 3; Data Supplement) and then we estimate the marginal associations in the same manner as described above.

Evaluation Metrics

The scenarios are evaluated using the following metrics. We assess the percentage bias for the naive estimates of marginal association against the true marginal association. We also assess the coverage of the 95% confidence interval (CI), which has an expected value of 0.95 for a well-behaved CI. In addition, we evaluate the type 1 error (ie, the proportion of falsely rejecting the null hypothesis of no association when there is no association) and power (ie, the proportion of rejecting the null hypothesis when there is an association) at the nominal significance level of 0.05. All results are based on 5000 replications of the scenarios.

Results

Figure 3 presents the results exploring selection bias in the time-to-event analysis of the *G* effect on subsequent CHD events (scenario 1). When the genetic exposure has no effect (ie, the HR of *G* is 1), there is also no selection bias in either populations 2 (who had a first event) or 3 (who had a first event and survived) and the type 1 error is correctly controlled at 0.05. When the genetic exposure has an effect, the bias in population 2 (index event bias alone) is generally <10% unless the HRs of both *G* and *X* become large (eg, 2 and 10, respectively). The bias in population 3 (cumulative effect of index event bias plus survival bias) is, as expected, larger than the bias in population 2, but still <10% unless the HR of *G* is >1.3; all biases described here and below are downward toward the null. Figure I in the Data Supplement illustrates, for 1 set of effect sizes of *G* and *X* that are used to simulate the outcomes, the true and naive estimates of the marginal effect size of *G* with populations 2 and 3. However, the CI may have poor coverage because of the large sample size and hence, small variance associated with the (biased) estimate of the HR of *G*. Additional details are presented in Table I in the Data Supplement.

In sensitivity analyses, we evaluated the bias as the overall disease rate in the general population c_1, rate of noncensored subsequent CHD events c_2, index event death rate c_s, and single nucleotide polymorphism MAF π varied. We observe from Figure II in the Data Supplement that the bias is generally insensitive to any of these parameters. To explore power and bias in other sample sizes, the simulation scenario 1 was repeated using a sample size of 1000, 5000, 10 000, and 50 000. The results in Figure 4 show that as the sample size increases, the bias stays similar. Meanwhile, power increases and coverage tends to fall below the nominal level, both owing to the shrunken variance for the (biased) estimate of HRs.

In Figures 5 and 6, we show the results of HR for a genetic exposure *G* and a phenotypic exposure *Z*, respectively, in scenario 2. The bias, because of index event bias alone or the cumulative effect of index event bias plus survival bias, is generally <10% when the HR of *G* is ≤1.3. The test of *Z* is more powerful than that of *G*. However, the bias in the latter test is smaller. More detailed results are provided in Tables II and III in the Data Supplement, which also reveal agreement between the empirical standard error and the mean of standard error estimates.

The results for OR estimates are presented in Tables IV through VI in the Data Supplement showing similar patterns as for the HR estimates. For the OR, we further compared power

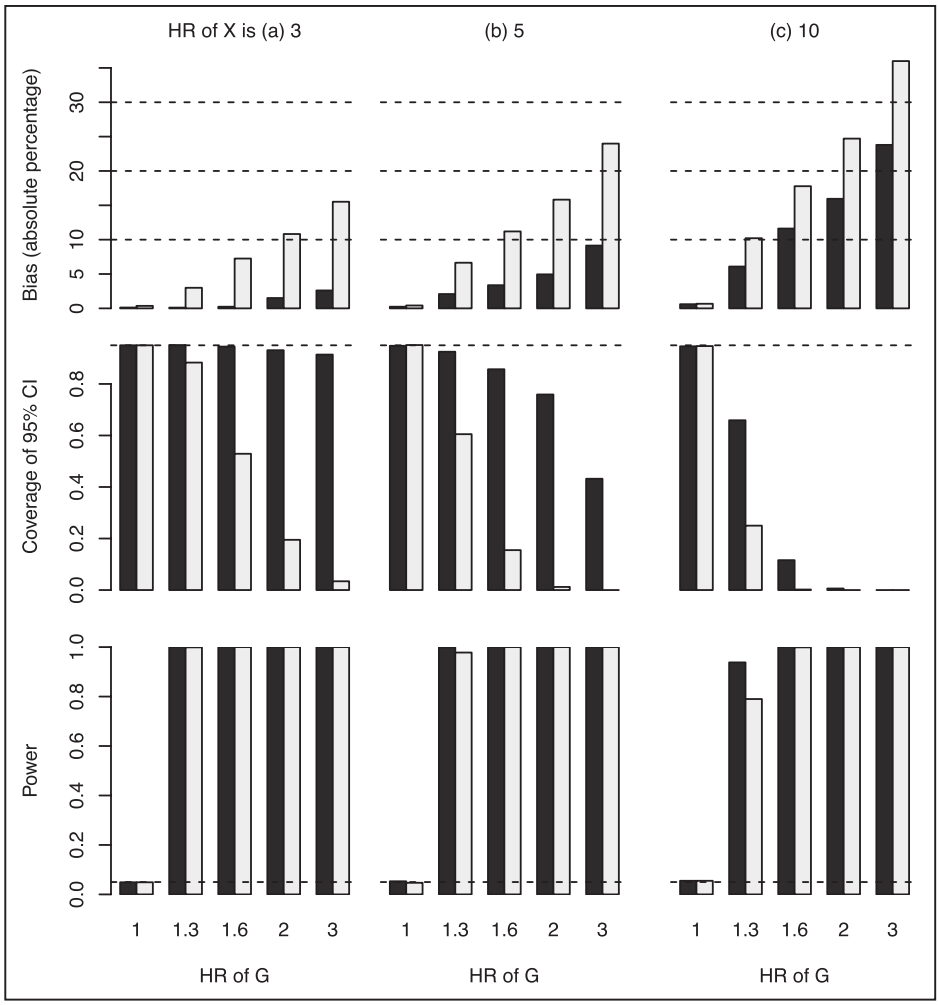


Figure 3. Results of the estimated hazard ratio (HR) for a genetic variant that associates with risk of first event, survival, and risk of a subsequent coronary heart disease event (scenario 1) Power under the HR of 1 for *G* means type 1 error. The black bars pertain to population 2 (selection of subjects with fatal or nonfatal first events) and the gray bars to population 3 (selection of subjects with nonfatal first events). **Middle**, The dashed line indicates the expected coverage of 0.95. **Lower**, The dashed line indicates the nominal significance level of 0.05. Sample size is set at 25 000. CI indicates confidence interval.

of rejecting the null-hypotheses between populations 1, 2, and 3. Under our simulation scheme that *G* and *X* have equal effects on both initial and subsequent CHD events, the power is higher in population 1 than in population 2 (eg, 100% versus 89.3% when the ORs of *G* and *X* are 1.3 and 10, respectively, in scenario 1). This difference in power is not only attributable to the difference of the true marginal OR but also the selection bias. The power is higher in population 2 than in population 3 (eg, 89.3% versus 76.7% when the ORs of *G* and *X* are 1.3 and 10, respectively, in scenario 1) because of the loss of high-risk subjects. The impact of selection bias on the observed MAF is increasing the MAF from 0.300 to 0.330 and 0.328 for populations 1, 2, and 3, respectively, in a realistically extreme case (the ORs for *G* and *X* are 1.3 and 10, respectively, in scenario 1).

To explore whether our findings apply to other designs, we repeated scenario 1 with a 1:1 case–control design. We showed in Table VII and Figure III in the Data Supplement that case–control studies are similarly affected by selection bias as cohort studies. For example, in an extreme case (the ORs for *G* and *X* are 3 and 10, respectively), bias was 9.59% in cohort studies versus 9.64% in case–control studies.

Discussion

The current simulation study, designed to mimic the scenarios encountered in studies of subsequent CHD events, such as those proposed by the GENIUS-CHD consortium, demonstrated that selection biases (ie, index event bias and survival bias) have little impact on gene-disease association estimates when the genetic risk factors have the modest effects observed in most studies. Typically, bias was greater when genetic risk factors had very large effects (ie, HR of $G \geq 2$). We confirmed that the type 1 error rate was unaffected, given that selection bias cannot occur when a gene has no effect on disease and assuming an absence of effect modification by history of disease. However, coverage probabilities of CIs could be considerably less than the nominal level, and they decreased to 0 with increasing sample sizes and selection bias pressure (ie, larger HRs of *G* and *X* on the occurrence of an indexing event). Given the agreement between the empirical standard error and the mean of standard error estimates, the observed undercoverage seems to be predominantly caused by bias in the point estimate.

Previously, methodological reports addressing the problem of selection bias in association studies have done so in the context of nongenetic or phenotypic exposures.[6,14–16] In this setting, Greenland[11] suggested that in most instances the magnitude of selection bias compared with confounding bias is modest. This was partially reiterated by Smits et al,[16] only finding an appreciable selection bias in scenarios where the effect on the first event was very large. However, with an increasing focus on the genetic context of subsequent CHD,[5] a more specific question

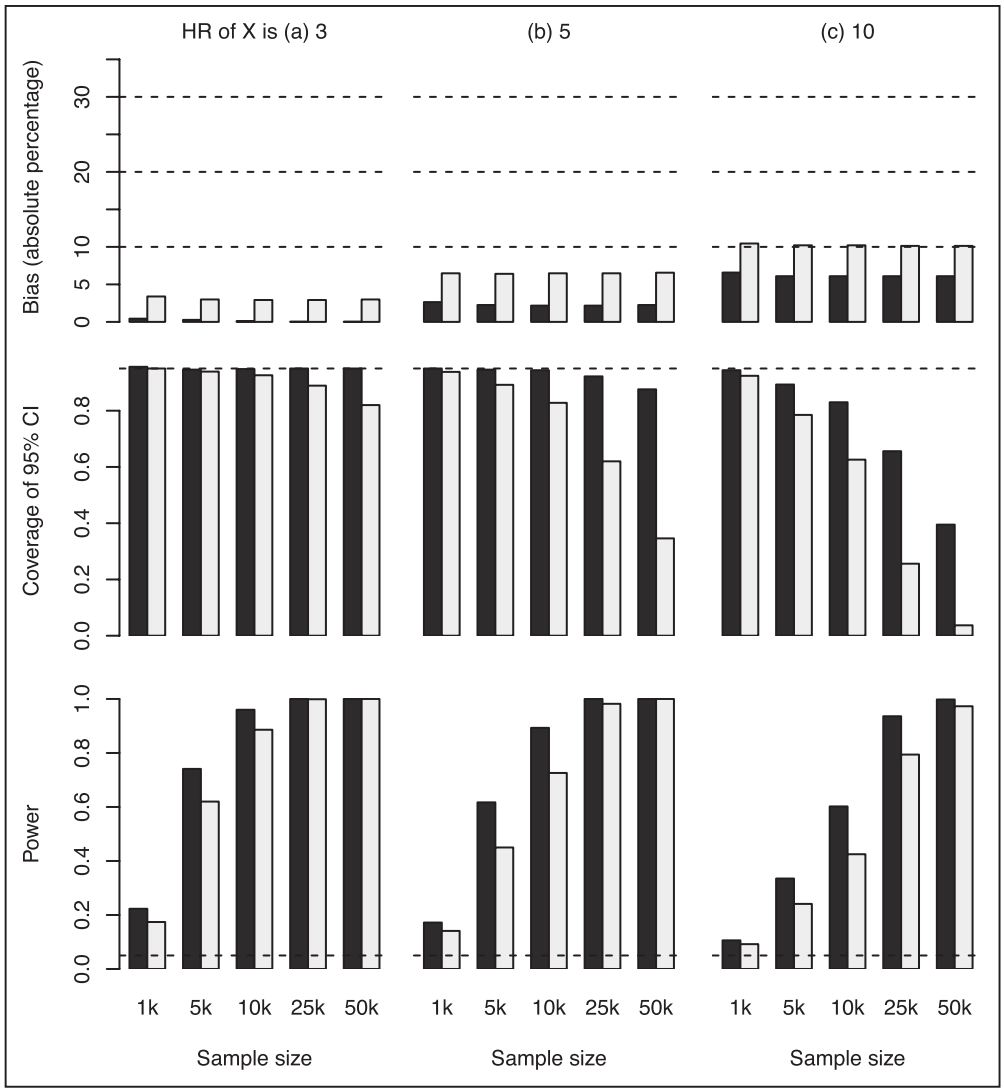


Figure 4. Results of the estimated hazard ratio (HR) for a genetic variant (scenario 1) with different sample sizes. The HR of *G* is set to 1.3. The black bars pertain to population 2 (selection of subjects with fatal or nonfatal first events) and the gray bars to population 3 (selection of subjects with nonfatal first events). **Middle**, The dashed line indicates the expected coverage of 0.95. **Lower**, The dashed line indicates the nominal significance level of 0.05. CI indicates confidence interval.

has arisen about the impact of selection bias in studying those who have been selected on and have survived a potentially fatal index event. While some studies have examined the impact of selection bias on effect estimates in case–control studies,[17,18] to our knowledge this question has not been addressed for time-to-event analysis of longitudinal cohort studies exploring associations with recurrent or subsequent CHD events.

Few studies have directly compared genetic risk of first versus subsequent CHD events to explore the comparability of these simulation studies to real examples. Our group, however, has previously compared the effects of the 9p21 risk variant on first incidence of CHD to subsequent CHD events, finding a more attenuated association for the latter: HR, 1.19 per risk allele with 95% CI (1.17–1.22) versus HR, 1.01 per risk allele with 95% CI (0.97–1.06).[19] Given that 9p21 has a small effect size (HR or OR≤1.3) in the unselected population, the observed 9p21 results for subsequent CHD events are unlikely to be solely attributable to index event bias or survival bias but possibly to other factors, such as risk-modifying therapies.

An important simplification of our simulation study was to focus on genetic and nongenetic exposures that are free of confounding bias. This may seem unrealistic, however, our focus was predominantly on selection bias in genetic exposures. Because the assortment of genetic variants at meiosis and conception occurs at random and is independent of other factors, one may expect the association of genes with an outcome to be affected less by confounding, especially when there is no population stratification. However, in real-life settings, selection bias and confounding bias are likely to both affect effect estimates of the association between environmental exposures and subsequent CHD events, making causal inference of such associations challenging.

Another simplification we made is the assumption that D_1 affects survival not directly but through *G* and *X*. This assumption does not necessarily agree with all biological mechanisms. However, and importantly so, this simplification does not change the simulation results. Given that D_1 is caused by both *X* and *G* (through *Z*), selection bias is induced by conditioning on a certain level of D_1, which results in a correlation between *X* and *G*. Allowing D_1 to be related to *S* will change the absolute number of survival but will not change the correlation between *X* and *G*, because D_1 itself is caused by these variables.

Our simulations involved a prospective cohort design, raising the question of whether they apply to other designs most notably

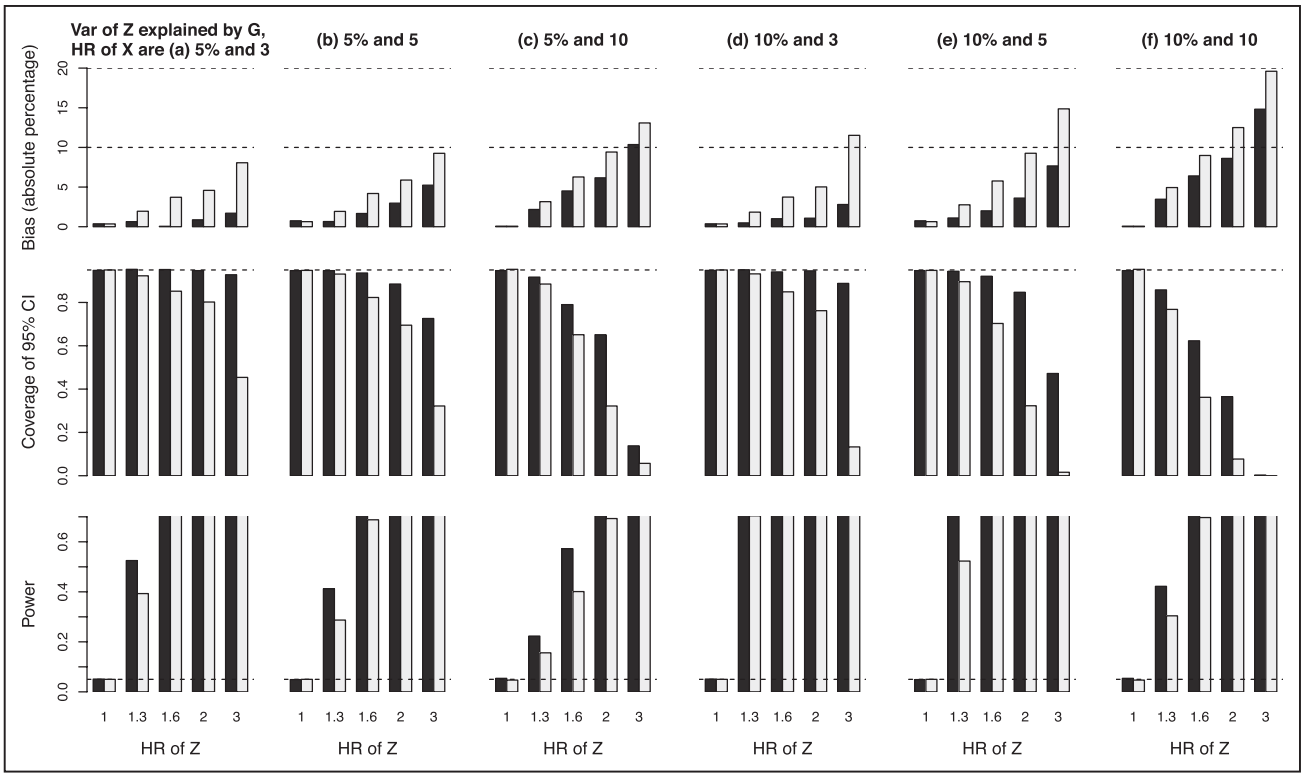


Figure 5. Results of the estimated hazard ratio (HR) for a genetic variant that encodes a biomarker that associates with risk of first event, survival, and risk of a subsequent CHD event (scenario 2) power under the HR of 1 for *Z* means type 1 error. The black bars pertain to populations 2 (selection of subjects with fatal or nonfatal first events) and the gray bars to population 3 (selection of subjects with nonfatal first events). **Middle**, The dashed line indicates the expected coverage of 0.95. **Lower**, The dashed line indicates the nominal significance level of 0.05. CI indicates confidence interval.

case–control studies. To provide some insight, we repeated scenario 1 with a 1:1 case–control design, and we showed that case–control studies are similarly affected by selection bias as cohort studies. Although cohort and case–controls studies are equally susceptible to selection bias of the type considered here (ie, selection bias caused by selecting on subjects surviving a first [CHD] event), it is well known that case–control studies may also be affected by other selection biases in the general population (ie, those who did not experience a CHD event). For example, in a retrospective case–control study, inclusion in the study may depend on the exposure status (eg, a drug), which results in selection bias. However, this is a different type of selection bias as discussed here, see for example van Rein et al,[20] for a discussion of this more generic form of selection bias.

In genetic association studies, another common source of bias is winner's curse, in which the disease risk of a newly identified genetic association is overestimated because of low statistical power for identifying the genetic association at a stringent genome-wide significance level. The bias from winner's curse differs from the index event/survival bias considered here in several ways. First, the former bias results from selecting estimates whose *P* values pass the stringent genome-wide significance level while the latter results from selecting a population stratum. Second, the former is related to statistical power and hence, sample size while the latter is not. Lastly, the former is biased upward whereas the latter is downward.

There are some limitations to our study. First, we recognize that part of these assessments could have been performed using analytic derivations instead of simulation studies. For example, Sperrin et al[21] presented an interesting analytic assessment of the obesity paradox, although our focus on time-to-event analyses, would have made a similar analytic solution as Sperrin et al difficult. Second, we focused primarily on the marginal effect estimate without adjusting for any covariates as explained earlier, although we accept that in some cases the conditional effect estimate may be of more interest.[22] Nonetheless, in the case of conditional effects, we would expect performance to improve if the covariates included are related to the outcome, in which case our simulations can be seen as a worst-case scenario of performance when none of the covariates related to the outcome are included. In particular, if the principal components for ancestry are included to account for population stratification, their correlations with the single nucleotide polymorphism of interest would diminish the selection bias because only the variability in the single nucleotide polymorphism that is unexplained by the principal components is subject to the selection bias. Finally, we have focused on the 5% nominal significance level and the 95% CI. Alternatively, a GWAS typically adopts a genome-wide significance level that is much <5% (eg, 5×10^{-8}). We have focused on 5% in our simulation studies (1) because the genome-wide significance level would require a substantial number of replicates and cause the simulation studies to become impractical, (2) because the type 1 error is unaffected by selection bias, the use of any significance level would not change our conclusions, and (3) although the GENIUS-CHD and similar consortiums are interested in high-throughput work,

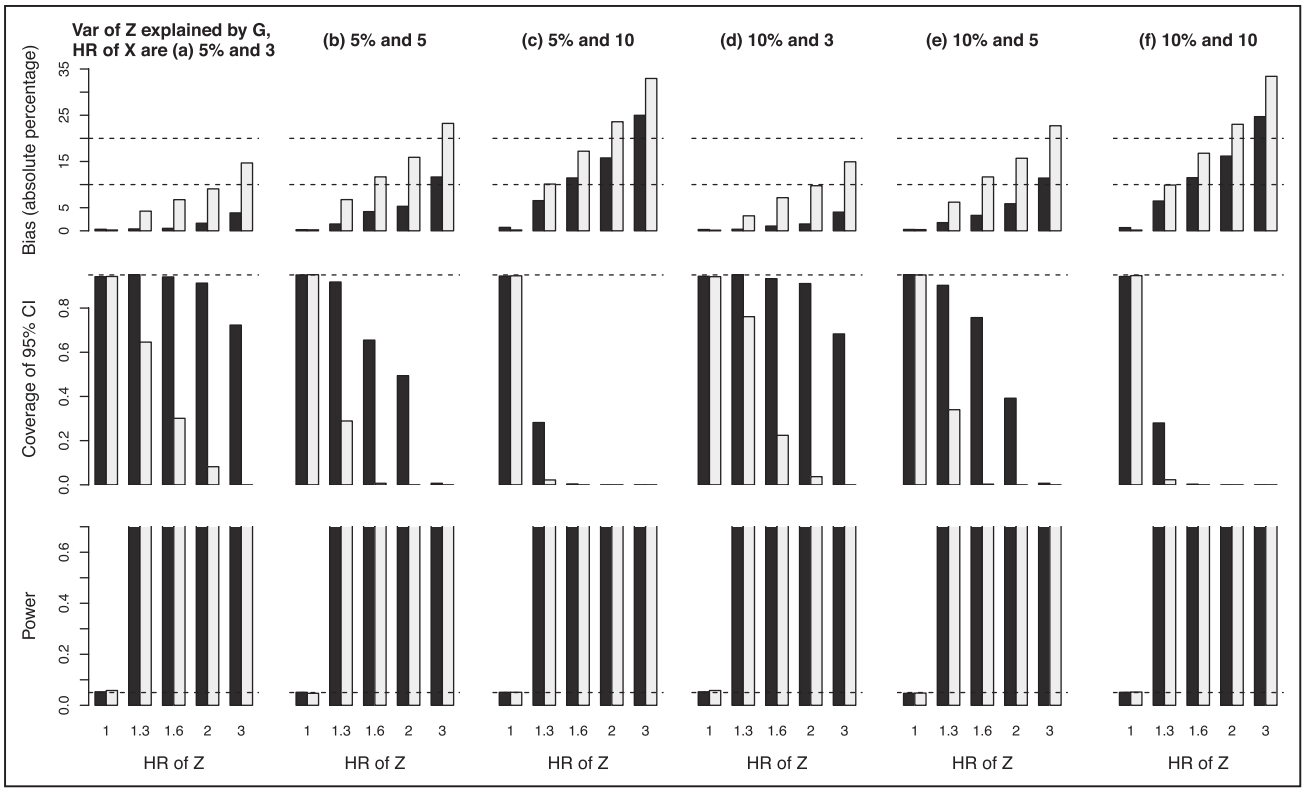


Figure 6. Results of the estimated hazard ratio (HR) for a nongenetic biomarker that associates with risk of first event, survival, and risk of a subsequent CHD event (scenario 2) power under the HR of 1 for *Z* means type 1 error. The black bars pertain to populations 2 (selection of subjects with fatal or nonfatal first events) and the gray bars to population 3 (selection of subjects with nonfatal first events). **Middle**, The dashed line indicates the expected coverage of 0.95. **Lower**, The dashed line indicate 1.00 and 0.05 (the nominal significance level). CI indicates confidence interval.

considerable effort is invested in performing Mendelian randomization (ie, instrumental variable) analyses which typically uses the 5% nominal significance level.

In conclusion, bias caused by selecting subjects with a history of disease is relatively small in genetic association studies for subsequent events, such as those for recurrent or subsequent CHD. Importantly, unless the associations are modified by the presence or absence of the first event, the type 1 error rate remains unaffected. Alternatively, the problem of selection bias may be absent entirely if the causes of the first disease event do not influence disease progression. These findings support the methodological validity of seeking common genetic variants for risk of subsequent events for CHD and potentially other diseases where recurrence and progression is clinically relevant. However, while tests are valid, researchers should be aware that despite the likely low degree of bias, the probability that the CIs include the true effect decreases with increasing sample size, resulting in coverage often (much) lower than the nominal level (eg, 95%).

Sources of Funding

This study was supported by the National Institutes of Health (R01GM116065 and R03AI111396 to Dr Hu, R03CA173770, R03CA183006, and R21NS091630 to Dr Long); University College London (UCL) Hospitals National Institute for Health Research (NIHR) Biomedical Research Centre (BRC10200 to Dr Schmidt and Dr Hingorani [Dr Hingorani is NIHR Senior Investigator], BRC169529 to Dr Asselbergs); UCL Springboard Population Health Sciences fellowship (to Dr Schmidt); Medical Research Council (MR/K006215/1 to Dr Dudbridge); a Dekker scholarship of Netherlands Heart Foundation (Junior Staff Member 2014T001 to Dr Asselbergs); and a British Heart Foundation Intermediate Fellowship (FS/14/76/30933 to Dr Patel).

Appendix

GENIUS-CHD Consortium

Axel Åkerblom, Ale Algra, Hooman Allayee, Peter Almgren, Jeffrey L. Anderson, Maria G. Andreassi, Chiara V. Anselmi, Diego Ardissino, Benoit J. Arsenault, Christie M. Ballantyne, Ekaterina V. Baranova, Hassan Behloui, Thomas O. Bergmeijer, Connie R. Bezzina, Eythor Bjornsson, Simon C. Body, Bram Boeckx, Eric (H.) Boersma, Eric Boerwinkle, Peter Bogaty, Peter S. Braund, Lutz P. Breitling, Hermann Brenner, Carlo Briguori, Jasper J. Brugts, Ralph Burkhardt, Vicky A. Cameron, John F. Carlquist, Clara Carpeggiani, Kathryn F. Carruthers, Gavino Casu, Gianluigi Condorelli, Sharon Cresci, Nicolas Danchin, Ulf de Faire, John Deanfield, Graciela Delgado, Panos Deloukas, Kenan Direk, Robert N. Doughty, Heinz Drexel, Nubia E. Duarte, Marie-Pierre Dubé, Line Dufresne, James C. Engert, Niclas Eriksson, Natalie Fitzpatrick, Luisa Foco, Ian Ford, Keith A.A. Fox, Bruna Gigante, Crystel M. Gijsberts, Domenico Girelli, Yan Gong, Daniel F. Gudbjartsson, Emil Hagström, Jaana Hartiala, Stanley L. Hazen, Claes Held, Anna Helgadottir, Harry Hemingway, Mahyar Heydarpour, Imo E. Hoefer, Kees Hovingh, Jaroslav A. Hubacek, Stefan James, Julie A. Johnson, J. Wouter Jukema, Marcin P. Kaczor, Karol A. Kaminski, Jiri Kettner, Marek Kiliszek, Marcus Kleber, Olaf H. Klungel, Daniel Kofink, Mika Kohonen, Salma Kotti, Pekka Kuukasjärvi, Bo Lagerqvist, Diether Lambrechts, Chim C. Lang, Jari O. Laurikka, Karin Leander, Vei-Vei Lee, Terho Lehtimäki, Andreas Leiherer, Petra A. Lenzini, Daniel Levin, Daniel Lindholm, Marja-Liisa Lokki, Paulo A. Lotufo, Leo-Pekka Lyytikäinen, B. Khan Mahmoodi, Anke

H. Maitland-van der Zee, Nicola Martinelli, Winfried März, Nicola Marziliano, Ruth McPherson, Olle Melander, Ute Mons, Jochen D. Muehlschlegel, Joseph B. Muhlestein, Cristopher P. Nelson, Chris Newton Cheh, Oliviero Olivieri, Grzegorz Opolski, Colin N. Palmer, Guillaume Pare, Gerard Pasterkamp, Carl J. Pepine, Witold Pepinski, Alexandre C. Pereira, Anna P. Pilbrow, Louise Pilote, Jan Pitha, Rafal Ploski, A. Mark Richards, Christoph H. Saely, Nilesh J. Samani, Ayman Samman-Tahhan, Marek Sanak, Pratik B. Sandesara, Naveed Sattar, Markus Scholz, Agneta Siegbahn, Tabassome Simon, Juha Sinisalo, J. Gustav Smith, John A. Spertus, Kari Stefansson, Alexandre F.R. Stewart, David J. Stott, Wojciech Szczeklik, Anna Szpakowicz, Michael W.T. Tanck, Wilson H. Tang, Jean-Claude Tardif, Jur M. ten Berg, Andrej Teren, George Thanassoulis, Joachim Thiery, Gudmundur Thorgeirsson, Gudmar Thorleifsson, Unnur Thorsteinsdottir, Adam Timmis, Stella Trompet, Frans van de Werf, Yolanda van der Graaf, Pim van der Haarst, Sander W. van der Laan, Ragnar O. Vilmundarson, Salim S. Virani, Frank L.J. Visseren, Efthymia Vlachopoulou, Lars Wallentin, Johannes Waltenberger, Els Wauters, Arthur A.M. Wilde

Disclosures

None.

References

1. Capewell S, Allender S, Critchley J, Lloyd-Williams F, O'Flaherty M, Rayner M, et al. *Modelling the UK Burden of Cardiovascular Disease to 2020: A Research Report for the Cardio & Vascular Coalition and the British Heart Foundation*. London, United Kingdom: Cardio & Vascular Coalition and the British Heart Foundation; 2008.
2. Mozaffarian D, Benjamin EJ, Go AS, Arnett DK, Blaha MJ, Cushman M, et al; American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics–2015 update: a report from the American Heart Association. *Circulation*. 2015;131:e29–e322. doi: 10.1161/CIR.0000000000000152.
3. Piepoli MF, Hoes AW, Agewall S, Albus C, Brotons C, Catapano AL, et al; Authors/Task Force Members. 2016 European Guidelines on cardiovascular disease prevention in clinical practice: the Sixth Joint Task Force of the European Society of Cardiology and other societies on cardiovascular disease prevention in clinical practice (constituted by representatives of 10 societies and by invited experts)developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation (EACPR). *Eur Heart J*. 2016;37:2315–2381. doi: 10.1093/eurheartj/ehw106.
4. Reilly MP, Li M, He J, Ferguson JF, Stylianou IM, Mehta NN, et al; Myocardial Infarction Genetics Consortium; Wellcome Trust Case Control Consortium. Identification of ADAMTS7 as a novel locus for coronary atherosclerosis and association of ABO with myocardial infarction in the presence of coronary atherosclerosis: two genome-wide association studies. *Lancet*. 2011;377:383–392. doi: 10.1016/S0140-6736(10)61996-4.
5. Patel RS, Asselbergs FW. The GENIUS-CHD consortium. *Eur Heart J*. 2015;36:2674–2676.
6. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004;15:615–625.
7. Dahabreh IJ, Kent DM. Index event bias as an explanation for the paradoxes of recurrence risk research. *JAMA*. 2011;305:822–823. doi: 10.1001/jama.2011.163.
8. Banack HR, Kaufman JS. Does selection bias explain the obesity paradox among individuals with cardiovascular disease? *Ann Epidemiol*. 2015;25:342–349. doi: 10.1016/j.annepidem.2015.02.008.
9. 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;66:599–607. doi: 10.1016/j.jclinepi.2012.08.008.
10. Hingorani A, Humphries S. Nature's randomised trials. *Lancet*. 2005;366:1906–1908. doi: 10.1016/S0140-6736(05)67767-7.
11. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*. 2003;14:300–306.
12. Jiang H, Kulkarni PM, Wang Y, Mallinckrodt CH. Nonparametric covariate adjustment in estimating hazard ratios. *Pharm Stat*. 2016;15:46–53. doi: 10.1002/pst.1725.
13. The CARDIoGRAMplusC4D Consortium. A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet*. 2015;47:1121–1130. doi: 10.1038/ng.3396.
14. Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, Richardson D, et al. Illustrating bias due to conditioning on a collider. *Int J Epidemiol*. 2010;39:417–420. doi: 10.1093/ije/dyp334.
15. Flanders WD, Eldridge RC, McClellan W. A nearly unavoidable mechanism for collider bias with index-event studies. *Epidemiology*. 2014;25:762–764. doi: 10.1097/EDE.0000000000000131.
16. Smits LJ, van Kuijk SM, Leffers P, Peeters LL, Prins MH, Sep SJ. Index event bias-a numerical example. *J Clin Epidemiol*. 2013;66:192–196. doi: 10.1016/j.jclinepi.2012.06.023.
17. Anderson CD, Nalls MA, Biffi A, Rost NS, Greenberg SM, Singleton AB, et al. The effect of survival bias on case-control genetic association studies of highly lethal diseases. *Circ Cardiovasc Genet*. 2011;4:188–196. doi: 10.1161/CIRCGENETICS.110.957928.
18. Dungan JR, Qin X, Horne BD, Carlquist JF, Singh A, Hurdle M, et al. Case-only survival analysis reveals unique effects of genotype, sex, and coronary disease severity on survivorship. *PLoS One*. 2016;11:e0154856. doi: 10.1371/journal.pone.0154856.
19. Patel RS, Asselbergs FW, Quyyumi AA, Palmer TM, Finan CI, Tragante V, et al. Genetic variants at chromosome 9p21 and risk of first versus subsequent coronary heart disease events: a systematic review and meta-analysis. *J Am Coll Cardiol*. 2014;63:2234–2245. doi: 10.1016/j.jacc.2014.01.065.
20. van Rein N, Cannegieter SC, Rosendaal FR, Reitsma PH, Lijfering WM. Suspected survivor bias in case-control studies: stratify on survival time and use a negative control. *J Clin Epidemiol*. 2014;67:232–235. doi: 10.1016/j.jclinepi.2013.05.011.
21. Sperrin M, Candlish J, Badrick E, Renehan A, Buchan I. Collider bias is only a partial explanation for the obesity paradox. *Epidemiology*. 2016;27:525–530. doi: 10.1097/EDE.0000000000000493.
22. Groenwold RHH, Moons KGM, Peelen LM, Knol MJ, Hoes AW. Reporting of treatment effects from randomized trials: a plea for multivariable risk ratios. *Contemp Clin Trials*. 2011;32:399–402.

CLINICAL PERSPECTIVE

As a community, we have made great advances in reducing mortality from acute coronary heart disease (CHD). Consequently, however, more people are surviving and living with established CHD and remaining at significant risk of future cardiac events and death. Attention is now turning to discover genetic variants and biomarkers that confer risk of CHD progression in the hope that novel findings might uncover mechanistic insights into recurrent or progressive heart disease, such as processes promoting plaque vulnerability or rupture. Such insights could potentially lead to development of new drugs, influencing mechanisms beyond lipid and thrombosis pathways. The GENIUS-CHD Consortium has been developed to specifically address the task of identifying risk variants and markers of subsequent events among those with established CHD. However, studying those with established disease brings its own challenges. Selection of individuals who both experience and survive a primary index CHD event leads to bias which may impact all findings. In this article, using empirically inspired simulation studies, we demonstrate that unless the observed genetic effects are very large, association findings will be minimally impacted by selection bias (attenuated to the null), although with increasing sample size confidence intervals are less likely to include the true effect. As common genetic variants have small effect sizes, we thus expect genetic association studies of disease progression or recurrence to yield relatively unbiased estimates. This study and findings therefore have implications for CHD, as well as any other condition where disease progression is important.

Circulation
Genomic and Precision Medicine

Impact of Selection Bias on Estimation of Subsequent Event Risk

Yi-Juan Hu, Amand F. Schmidt, Frank Dudbridge, Michael V. Holmes, James M. Brophy, Vinicius Tragante, Ziyi Li, Peizhou Liao, Arshed A. Quyyumi, Raymond O. McCubrey, Benjamin D. Horne, Aroon D. Hingorani, Folkert W. Asselbergs, Riyaz S. Patel and Qi Long

Circ Cardiovasc Genet. 2017;10:
doi: 10.1161/CIRCGENETICS.116.001616
Circulation: Cardiovascular Genetics is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231

Print ISSN: 1942-325X. Online ISSN: 1942-3268

SUPPLEMENTAL MATERIAL

Methods

Scenario 1

When D_2 is the time to event, it includes two variables (T, δ), where T is the time to event and δ is the indicator of non-censoring. We generate T from the proportional hazards regression model (3), where $\lambda(T|G, X)$ is the hazard function with risk factors of G and X and $2T$ is an arbitrary baseline hazard function. We generate δ from the uniform distribution $U(0, \tau)$, where τ is chosen to yield the fixed proportion c_2 of non-censored subsequent events.

Scenario 2

We simulate Z with a normal distribution of mean ηG and variance 1. We set η to 0.354 and 0.514 to represent that 5% and 10% variance of Z is explained by G.

Calculation of the true marginal association

The true marginal association (i.e., HR or OR) of G on D_2 in scenario 1 can be calculated by the following counterfactual method. First, we generate a sample of the general population with values on G, X, and D_1; the sample size is large, here 200,000,000 (so that the "new dataset" defined below will be sufficiently large for estimating the true marginal association with ignorable variation). We then obtain the subset with $D_1 = 1$ when the target population is population 2, or the subset with $D_1 = 1$ and $S = 1$ when the target population is population 3. In either case, we denote the subset by $\mathcal{A}$. Let N be the sample size of $\mathcal{A}$. Second, we create a new dataset that consists of three sets of samples, each of size N. We generate the first set of samples by taking all subjects in $\mathcal{A}$, replacing their values of G by 0, and generating D_2 based on the original values of X and the new values of G. We generate the second and third sets of samples

similarly except that we set all G's to be 1 and 2, respectively. Finally, we calculate the association estimate using the new values of G and D_2 in the new dataset. This way, we obtain the true association estimate in the target population that is characterized by the observed distribution of X in the same population. The marginal effect of G on D_2 or Z on D_2 in scenario 2 is estimated in a similar manner as described above, except that new values of Z are generated after setting G to values of 0, 1, and 2.

Evaluation metrics

The percentage bias is $(\bar{\theta} - \hat{\theta}_{true})/\hat{\theta}_{true} \times 100$, where $\hat{\theta}_{true}$ is the true marginal association estimated by the counterfactual method and $\bar{\theta}$ is the average of the naïve estimates $\hat{\theta}_i$ across B replicates, i.e., $\bar{\theta} = \sum_{i=1}^{B} \hat{\theta}_i / B$. The coverage of the 95% confidence interval (CI) is the proportion of times the 95% CI $\hat{\theta}_i \pm Z_{0.975} SE(\hat{\theta}_i)$ includes $\hat{\theta}_{true}$, where $SE(\hat{\theta}_i)$ is the estimated standard error for $\hat{\theta}_i$ within each replicate and $Z_{0.975}$ is the 0.975 quantile of the standard normal distribution. We set the number of replicates B to 5000 so that the estimated coverage of the 95% CI should fall between 0.944 and 0.956 with 95% probability.

Table S1. Results of the estimated hazard ratio (HR) for a genetic variant that associates with risk of first event, survival, and risk of a subsequent CHD event (scenario 1)

Conditional HR		Marginal HR															
		Population 2								Population 3							
X on D_2	G on D_2	True		Naive estimator					Bias	True		Naive estimator					Bias
$\exp\beta_X$	$\exp\beta_G$	HR	SE	HR	SE	SEE	COV	POW	%	HR	SE	HR	SE	SEE	COV	POW	%
3	1	1	0.005	1	0.044	0.044	0.95	0.049	0.12	1	0.006	1	0.044	0.043	0.95	0.049	0.36
	1.3	1.27	0.006	1.27	0.052	0.051	0.951	1	-0.11	1.27	0.007	1.23	0.05	0.05	0.883	0.999	-2.99
	1.6	1.54	0.008	1.53	0.061	0.061	0.945	1	-0.24	1.56	0.009	1.45	0.058	0.058	0.529	1	-7.25
	2	1.91	0.01	1.88	0.078	0.077	0.931	1	-1.5	1.93	0.012	1.72	0.07	0.07	0.195	1	-10.82
	3	2.79	0.017	2.71	0.127	0.128	0.914	1	-2.61	2.81	0.019	2.37	0.103	0.105	0.034	1	-15.52
5	1	1	0.005	1	0.044	0.043	0.948	0.053	-0.24	1	0.005	1	0.043	0.043	0.951	0.047	-0.42
	1.3	1.25	0.006	1.23	0.049	0.05	0.925	0.999	-2.08	1.26	0.007	1.18	0.049	0.048	0.605	0.978	-6.64
	1.6	1.49	0.008	1.44	0.058	0.057	0.857	1	-3.37	1.51	0.009	1.34	0.054	0.054	0.155	1	-11.2
	2	1.8	0.009	1.71	0.068	0.069	0.759	1	-4.95	1.84	0.011	1.54	0.062	0.062	0.012	1	-15.83
	3	2.56	0.015	2.32	0.105	0.105	0.432	1	-9.14	2.62	0.018	1.99	0.085	0.085	0	1	-23.98
10	1	0.99	0.005	1	0.043	0.043	0.946	0.055	0.61	0.99	0.005	1	0.043	0.043	0.947	0.055	0.67
	1.3	1.23	0.006	1.15	0.046	0.046	0.659	0.938	-6.09	1.25	0.007	1.12	0.046	0.046	0.25	0.79	-10.21
	1.6	1.46	0.007	1.29	0.051	0.05	0.116	1	-11.62	1.49	0.008	1.23	0.049	0.049	0.002	0.999	-17.78
	2	1.72	0.009	1.45	0.056	0.056	0.006	1	-15.95	1.79	0.011	1.35	0.053	0.053	0	1	-24.71
	3	2.33	0.012	1.78	0.071	0.071	0	1	-23.79	2.51	0.016	1.6	0.064	0.065	0	1	-35.99

The naive estimator was obtained from Cox regression. The true value was obtained from the counterfactual method. SE is the empirical standard error. SEE is the mean of standard error estimates. COV is the coverage of the 95% CI. POW is the power when $\exp\beta_G \neq 1$ and the type 1 error when $\exp\beta_G = 1$. Bias (%) is the difference between the estimated marginal HR and the true marginal HR, divided by the true marginal HR.

Table S2. Results of the estimated hazard ratio (HR) for a genetic variant that encodes a biomarker that associates with risk of first event, survival, and risk of a subsequent CHD event (scenario 2)

	Conditional OR		Marginal OR														
			Population 2								Population 3						
			True		Naive estimator					Bias	True		Naive estimator					Bias
G on Z	X on D_2	Z on D_2																
η	$\exp\beta_X$	$\exp\beta_Z$	OR	SE	OR	SE	SEE	COV	POW	%	OR	SE	OR	SE	SEE	COV	POW	%
0.354	3	1	1	0.005	1	0.044	0.043	0.948	0.051	-0.36	1	0.006	1	0.043	0.044	0.95	0.05	-0.35
		1.3	1.08	0.005	1.09	0.045	0.046	0.953	0.525	0.64	1.09	0.006	1.07	0.046	0.046	0.923	0.393	-1.95
		1.6	1.16	0.006	1.16	0.048	0.048	0.952	0.949	0.05	1.18	0.007	1.13	0.048	0.047	0.852	0.849	-3.71
		2	1.25	0.006	1.24	0.051	0.05	0.947	0.999	-0.87	1.25	0.007	1.19	0.049	0.049	0.802	0.991	-4.58
		3	1.39	0.007	1.36	0.054	0.054	0.928	1	-1.7	1.4	0.008	1.28	0.052	0.052	0.454	1	-8.07
	5	1	1.01	0.005	1	0.043	0.043	0.947	0.049	-0.75	0.99	0.005	1	0.043	0.043	0.948	0.05	0.64
		1.3	1.08	0.005	1.07	0.045	0.045	0.947	0.412	-0.65	1.08	0.006	1.06	0.045	0.045	0.931	0.287	-1.93
		1.6	1.15	0.006	1.13	0.046	0.047	0.936	0.861	-1.67	1.16	0.006	1.11	0.046	0.046	0.823	0.688	-4.19
		2	1.23	0.006	1.2	0.048	0.049	0.885	0.992	-2.97	1.23	0.007	1.16	0.047	0.047	0.695	0.938	-5.88
		3	1.36	0.007	1.29	0.052	0.051	0.726	1	-5.24	1.35	0.008	1.23	0.048	0.049	0.322	0.999	-9.26
	10	1	1	0.005	1	0.043	0.043	0.947	0.054	-0.07	1	0.005	1	0.043	0.043	0.953	0.047	0.06
		1.3	1.07	0.005	1.05	0.044	0.044	0.917	0.223	-2.17	1.07	0.006	1.04	0.044	0.044	0.885	0.156	-3.16
		1.6	1.14	0.006	1.09	0.045	0.045	0.79	0.572	-4.51	1.15	0.006	1.07	0.044	0.044	0.651	0.401	-6.27
		2	1.21	0.006	1.13	0.045	0.045	0.651	0.865	-6.16	1.22	0.007	1.1	0.045	0.045	0.322	0.693	-9.42
		3	1.33	0.007	1.19	0.043	0.043	0.138	0.998	-10.36	1.33	0.007	1.16	0.047	0.046	0.057	0.95	-13.08
0.514	3	1	1	0.005	1	0.044	0.043	0.948	0.051	-0.36	1	0.006	1	0.043	0.044	0.95	0.05	-0.35
		1.3	1.14	0.006	1.13	0.047	0.047	0.951	0.832	-0.48	1.13	0.006	1.11	0.046	0.047	0.932	0.704	-1.84
		1.6	1.25	0.006	1.24	0.051	0.05	0.941	1	-1.01	1.25	0.007	1.2	0.049	0.049	0.849	0.995	-3.74
		2	1.38	0.007	1.36	0.054	0.055	0.946	1	-1.08	1.36	0.007	1.29	0.052	0.052	0.762	1	-5.02
		3	1.61	0.008	1.56	0.062	0.062	0.888	1	-2.8	1.62	0.009	1.44	0.056	0.057	0.133	1	-11.52
	5	1	1.01	0.005	1	0.043	0.043	0.947	0.049	-0.75	0.99	0.005	1	0.043	0.043	0.948	0.05	0.64
		1.3	1.12	0.005	1.11	0.046	0.046	0.944	0.704	-1.1	1.12	0.006	1.09	0.046	0.046	0.896	0.523	-2.76
		1.6	1.23	0.006	1.2	0.049	0.049	0.921	0.994	-2	1.23	0.007	1.16	0.049	0.048	0.703	0.944	-5.77
		2	1.34	0.007	1.3	0.052	0.052	0.847	1	-3.61	1.36	0.008	1.23	0.049	0.05	0.323	0.999	-9.26
		3	1.57	0.008	1.45	0.057	0.057	0.472	1	-7.66	1.58	0.009	1.35	0.053	0.054	0.016	1	-14.86
	10	1	1	0.005	1	0.043	0.043	0.947	0.054	-0.07	1	0.005	1	0.043	0.043	0.953	0.047	0.06
		1.3	1.11	0.005	1.07	0.045	0.044	0.858	0.422	-3.46	1.12	0.006	1.06	0.045	0.044	0.768	0.304	-4.94
		1.6	1.21	0.006	1.13	0.045	0.046	0.623	0.882	-6.4	1.22	0.007	1.11	0.045	0.045	0.362	0.697	-8.98
		2	1.31	0.006	1.19	0.047	0.047	0.365	0.995	-8.62	1.32	0.007	1.16	0.046	0.046	0.077	0.95	-12.5
		3	1.51	0.007	1.29	0.043	0.044	0.003	1	-14.82	1.53	0.009	1.23	0.048	0.048	0	1	-19.6

The naïve estimator was obtained from Cox regression. The true value was obtained from the counterfactual method. SE is the empirical standard error. SEE is the mean of standard error estimates. COV is the coverage of the 95% CI. POW is the power when $\exp\beta_Z \neq 1$ and the type 1 error when $\exp\beta_Z = 1$. Bias (%) is the difference between the estimated marginal HR and the true marginal HR, divided by the true marginal HR. η = 0.354 and 0.514 represent that 5% and 10% variance of Z is explained by G, respectively.

Table S3. Results of the estimated hazard ratio (HR) for a non-genetic biomarker that associates with risk of first event, survival, and risk of a subsequent CHD event (scenario 2)

	Conditional OR		Marginal OR														
			Population 2								Population 3						
			True		Naive estimator					Bias	True		Naive estimator					Bias
G on Z	X on D_2	Z on D_2																
η	$\exp\beta_X$	$\exp\beta_Z$	OR	SE	OR	SE	SEE	COV	POW	%	OR	SE	OR	SE	SEE	COV	POW	%
0.354	3	1	1	0.004	1	0.028	0.027	0.942	0.053	0.34	1	0.004	1	0.028	0.027	0.943	0.058	0.18
		1.3	1.26	0.005	1.27	0.034	0.035	0.951	1	0.41	1.28	0.005	1.23	0.034	0.034	0.646	1	-4.25
		1.6	1.54	0.006	1.53	0.042	0.042	0.941	1	-0.53	1.55	0.007	1.45	0.04	0.04	0.301	1	-6.74
		2	1.89	0.007	1.86	0.051	0.052	0.913	1	-1.64	1.88	0.008	1.71	0.048	0.049	0.082	1	-9.07
		3	2.67	0.011	2.56	0.073	0.074	0.723	1	-3.87	2.7	0.012	2.3	0.067	0.069	0	1	-14.68
	5	1	1	0.004	1	0.027	0.027	0.95	0.051	-0.25	1	0.004	1	0.027	0.027	0.951	0.047	0.22
		1.3	1.25	0.005	1.23	0.034	0.034	0.918	1	-1.48	1.26	0.005	1.18	0.033	0.032	0.289	1	-6.75
		1.6	1.5	0.006	1.43	0.04	0.039	0.655	1	-4.16	1.52	0.006	1.34	0.037	0.037	0.007	1	-11.67
		2	1.77	0.007	1.68	0.046	0.046	0.494	1	-5.3	1.82	0.008	1.53	0.042	0.043	0	1	-15.9
		3	2.46	0.01	2.18	0.06	0.062	0.007	1	-11.65	2.52	0.011	1.93	0.055	0.056	0	1	-23.24
	10	1	1.01	0.004	1	0.027	0.027	0.945	0.051	-0.74	1	0.004	1	0.027	0.027	0.946	0.051	-0.18
		1.3	1.23	0.005	1.15	0.031	0.031	0.282	0.999	-6.53	1.25	0.005	1.12	0.03	0.03	0.022	0.987	-10.12
		1.6	1.45	0.006	1.28	0.034	0.035	0.004	1	-11.44	1.48	0.006	1.23	0.033	0.033	0	1	-17.22
		2	1.7	0.006	1.43	0.038	0.039	0	1	-15.77	1.75	0.007	1.34	0.037	0.037	0	1	-23.59
		3	2.28	0.009	1.71	0.043	0.044	0	1	-24.99	2.34	0.01	1.57	0.042	0.044	0	1	-32.96
0.514	3	1	1	0.004	1	0.027	0.027	0.944	0.053	0.28	1	0.004	1	0.027	0.027	0.942	0.058	0.13
		1.3	1.27	0.005	1.27	0.034	0.034	0.951	1	-0.34	1.27	0.005	1.23	0.033	0.033	0.761	1	-3.24
		1.6	1.55	0.006	1.53	0.041	0.041	0.933	1	-1.03	1.56	0.006	1.45	0.039	0.039	0.224	1	-7.17
		2	1.88	0.007	1.85	0.05	0.05	0.911	1	-1.48	1.9	0.008	1.71	0.046	0.047	0.037	1	-9.76
		3	2.66	0.01	2.55	0.071	0.071	0.683	1	-4.04	2.7	0.012	2.29	0.066	0.067	0	1	-14.93
	5	1	1	0.004	1	0.026	0.027	0.951	0.047	-0.29	1	0.004	1	0.027	0.027	0.95	0.048	0.26
		1.3	1.25	0.004	1.23	0.032	0.033	0.903	1	-1.77	1.26	0.005	1.18	0.031	0.032	0.34	1	-6.2
		1.6	1.48	0.005	1.43	0.037	0.038	0.757	1	-3.34	1.52	0.006	1.34	0.036	0.036	0.003	1	-11.66
		2	1.78	0.007	1.68	0.045	0.045	0.392	1	-5.86	1.82	0.008	1.53	0.041	0.042	0	1	-15.69
		3	2.44	0.009	2.17	0.057	0.059	0.007	1	-11.41	2.49	0.011	1.92	0.053	0.055	0	1	-22.73
	10	1	1.01	0.004	1	0.026	0.026	0.943	0.051	-0.69	1	0.004	1	0.026	0.026	0.947	0.052	-0.16
		1.3	1.23	0.004	1.15	0.031	0.03	0.28	1	-6.45	1.24	0.005	1.12	0.029	0.03	0.023	0.993	-9.9
		1.6	1.45	0.005	1.28	0.033	0.033	0.003	1	-11.49	1.47	0.006	1.22	0.032	0.032	0	1	-16.78
		2	1.7	0.006	1.43	0.037	0.037	0	1	-16.16	1.74	0.007	1.34	0.035	0.036	0	1	-23.04
		3	2.25	0.008	1.69	0.039	0.04	0	1	-24.7	2.35	0.01	1.57	0.042	0.043	0	1	-33.41

The naïve estimator was obtained from Cox regression. The true value was obtained from the counterfactual method. SE is the empirical standard error. SEE is the mean of standard error estimates. COV is the coverage of the 95% CI. POW is the power when $\exp\beta_Z \neq 1$ and the type 1 error when $\exp\beta_Z = 1$. Bias (%) is the difference between the estimated marginal HR and the true marginal HR, divided by the true marginal HR. η = 0.354 and 0.514 represent that 5% and 10% variance of Z is explained by G, respectively.

Table S4. Results of the estimated odds ratio (OR) for a genetic variant that associates with risk of first event, survival, and risk of a subsequent CHD event (scenario 1)

Conditional OR		Marginal OR																	
		Population 1		Population 2								Population 3							
X on D_2	G on D_2	Naive estimator		True		Naive estimator					Bias	True		Naive estimator					Bias
$\exp\beta_X$	$\exp\beta_G$	OR	POW	OR	SE	OR	SE	SEE	COV	POW	%	OR	SE	OR	SE	SEE	COV	POW	%
3	1	1	0.053	1	0.003	1	0.045	0.045	0.948	0.05	-0.57	1	0.003	1	0.045	0.045	0.949	0.051	-0.26
	1.3	1.3	1	1.25	0.003	1.24	0.052	0.052	0.952	0.999	-0.22	1.26	0.004	1.21	0.052	0.051	0.869	0.993	-3.43
	1.6	1.6	1	1.48	0.004	1.47	0.061	0.061	0.949	1	-0.29	1.5	0.005	1.41	0.058	0.059	0.673	1	-6.14
	2	1.99	1	1.77	0.004	1.76	0.074	0.073	0.945	1	-0.73	1.8	0.005	1.65	0.068	0.069	0.44	1	-8.34
	3	2.98	1	2.45	0.006	2.41	0.108	0.109	0.94	1	-1.54	2.5	0.007	2.2	0.096	0.097	0.175	1	-12.04
5	1	1	0.047	1	0.002	1	0.045	0.045	0.946	0.053	-0.08	1	0.003	1	0.045	0.045	0.951	0.049	0.17
	1.3	1.29	1	1.2	0.003	1.19	0.05	0.05	0.95	0.982	-1.03	1.21	0.003	1.16	0.049	0.049	0.825	0.928	-4.29
	1.6	1.58	1	1.38	0.003	1.36	0.056	0.057	0.941	1	-1.42	1.41	0.004	1.3	0.054	0.054	0.534	1	-7.47
	2	1.97	1	1.62	0.004	1.58	0.065	0.066	0.921	1	-2.06	1.65	0.004	1.48	0.062	0.061	0.228	1	-10.7
	3	2.92	1	2.17	0.005	2.09	0.091	0.09	0.855	1	-3.61	2.23	0.006	1.88	0.08	0.08	0.025	1	-15.62
10	1	1	0.055	1	0.003	1	0.045	0.045	0.945	0.055	-0.19	1	0.003	1	0.045	0.045	0.949	0.049	-0.39
	1.3	1.27	1	1.18	0.003	1.15	0.05	0.049	0.91	0.893	-2.4	1.18	0.003	1.12	0.048	0.048	0.783	0.767	-4.94
	1.6	1.54	1	1.35	0.004	1.29	0.055	0.054	0.804	1	-4.56	1.36	0.004	1.24	0.053	0.052	0.424	0.998	-8.79
	2	1.89	1	1.55	0.005	1.46	0.06	0.061	0.693	1	-5.87	1.56	0.005	1.38	0.057	0.057	0.142	1	-11.88
	3	2.73	1	2.06	0.007	1.86	0.078	0.078	0.328	1	-9.59	2.08	0.007	1.7	0.071	0.071	0.002	1	-18.43

The naïve estimator was obtained from logistic regression. The true value was obtained from the counterfactual method. SE is the empirical standard error. SEE is the mean of standard error estimates. COV is the coverage of the 95% CI. POW is the power when $\exp\beta_G \neq 1$ and the type 1 error when $\exp\beta_G = 1$. Bias (%) is the difference between the estimated marginal OR and the true marginal OR, divided by the true marginal OR.

Table S5. Results of the estimated odds ratio (OR) for a genetic variant that encodes a biomarker that associates with risk of first event, survival, and risk of a subsequent CHD event (scenario 2)

	Conditional OR		Marginal OR																
			Population 1		Population 2								Population 3						
G on Z	X on D_2	Z on D_2	Naive estimator		True		Naive estimator					Bias	True		Naive estimator					Bias
η	$\exp\beta_X$	$\exp\beta_Z$	OR	POW	OR	SE	OR	SE	SEE	COV	POW	%	OR	SE	OR	SE	SEE	COV	POW	%
0.354	3	1	1	0.051	1	0.003	1	0.044	0.045	0.948	0.052	-0.16	1	0.003	1	0.045	0.045	0.952	0.049	-0.24
		1.3	1.1	0.552	1.08	0.003	1.08	0.047	0.047	0.953	0.412	-0.09	1.09	0.003	1.07	0.046	0.047	0.938	0.338	-1.6
		1.6	1.18	0.963	1.15	0.003	1.14	0.049	0.049	0.947	0.866	-0.68	1.15	0.004	1.12	0.049	0.048	0.921	0.768	-2.08
		2	1.27	1	1.21	0.003	1.2	0.05	0.051	0.948	0.993	-0.89	1.22	0.004	1.18	0.05	0.05	0.843	0.966	-3.98
		3	1.47	1	1.32	0.003	1.3	0.054	0.054	0.932	1	-1.73	1.33	0.004	1.25	0.052	0.053	0.68	1	-6.08
	5	1	1	0.049	1	0.002	1	0.044	0.045	0.954	0.047	-0.19	1	0.003	1	0.044	0.045	0.952	0.048	-0.18
		1.3	1.1	0.545	1.07	0.002	1.06	0.046	0.047	0.956	0.284	-0.58	1.07	0.003	1.05	0.045	0.046	0.932	0.215	-1.96
		1.6	1.18	0.958	1.12	0.002	1.11	0.048	0.048	0.947	0.694	-0.56	1.13	0.003	1.09	0.047	0.047	0.899	0.553	-2.92
		2	1.27	1	1.18	0.003	1.17	0.05	0.05	0.943	0.948	-1.19	1.19	0.003	1.14	0.049	0.049	0.833	0.848	-4.23
		3	1.45	1	1.28	0.003	1.26	0.052	0.053	0.936	1	-1.6	1.29	0.003	1.21	0.051	0.051	0.702	0.995	-5.92
	10	1	1	0.051	1	0.003	1	0.044	0.045	0.95	0.049	-0.34	1	0.003	1	0.044	0.045	0.956	0.042	0.17
		1.3	1.09	0.487	1.06	0.003	1.05	0.045	0.046	0.948	0.2	-0.84	1.06	0.003	1.04	0.045	0.046	0.934	0.154	-1.94
		1.6	1.16	0.938	1.11	0.003	1.09	0.047	0.047	0.937	0.508	-1.84	1.11	0.003	1.08	0.046	0.047	0.89	0.385	-3.04
		2	1.25	0.999	1.16	0.003	1.14	0.048	0.049	0.922	0.842	-2.27	1.16	0.003	1.11	0.048	0.048	0.828	0.7	-4.31
		3	1.42	1	1.26	0.003	1.22	0.052	0.052	0.896	0.997	-2.92	1.26	0.004	1.18	0.051	0.05	0.71	0.976	-5.78
0.514	3	1	1	0.051	1	0.003	1	0.044	0.045	0.948	0.052	-0.16	1	0.003	1	0.045	0.045	0.952	0.049	-0.24
		1.3	1.14	0.86	1.12	0.003	1.12	0.048	0.048	0.954	0.723	-0.28	1.13	0.004	1.1	0.047	0.048	0.925	0.616	-2.13
		1.6	1.27	1	1.22	0.003	1.21	0.052	0.051	0.944	0.995	-0.53	1.23	0.004	1.18	0.051	0.05	0.864	0.972	-3.5
		2	1.42	1	1.32	0.003	1.31	0.055	0.055	0.939	1	-1.13	1.33	0.004	1.26	0.053	0.053	0.766	1	-5.04
		3	1.74	1	1.5	0.004	1.46	0.06	0.061	0.905	1	-2.5	1.51	0.004	1.38	0.057	0.058	0.448	1	-8.37
	5	1	1	0.049	1	0.002	1	0.044	0.045	0.954	0.047	-0.19	1	0.003	1	0.044	0.045	0.952	0.048	-0.18
		1.3	1.14	0.842	1.1	0.002	1.09	0.048	0.047	0.949	0.518	-0.66	1.1	0.003	1.08	0.047	0.047	0.92	0.402	-2.34
		1.6	1.27	1	1.18	0.003	1.17	0.05	0.05	0.944	0.952	-1.2	1.19	0.003	1.14	0.049	0.049	0.828	0.852	-4.12
		2	1.42	1	1.27	0.003	1.25	0.053	0.053	0.935	1	-1.59	1.28	0.003	1.21	0.052	0.051	0.703	0.993	-5.83
		3	1.72	1	1.42	0.003	1.4	0.058	0.058	0.923	1	-1.96	1.44	0.004	1.32	0.054	0.055	0.501	1	-7.85
	10	1	1	0.051	1	0.003	1	0.044	0.045	0.95	0.049	-0.34	1	0.003	1	0.044	0.045	0.956	0.042	0.17
		1.3	1.13	0.807	1.09	0.003	1.07	0.047	0.047	0.94	0.369	-1.46	1.09	0.003	1.06	0.047	0.046	0.911	0.267	-2.58
		1.6	1.25	1	1.16	0.003	1.14	0.049	0.049	0.914	0.84	-2.42	1.16	0.003	1.11	0.048	0.048	0.815	0.684	-4.52
		2	1.38	1	1.24	0.003	1.2	0.051	0.051	0.876	0.994	-3.34	1.25	0.004	1.17	0.05	0.05	0.669	0.957	-6.32
		3	1.65	1	1.39	0.004	1.34	0.055	0.056	0.839	1	-3.96	1.41	0.004	1.28	0.054	0.054	0.394	1	-9.01

The naïve estimator was obtained from logistic regression. The true value was obtained from the counterfactual method. SE is the empirical standard error. SEE is the mean of standard error estimates. COV is the coverage of the 95% CI. POW is the power when $\exp\beta_Z \neq 1$ and the type 1 error when $\exp\beta_Z = 1$. Bias (%) is the difference between the estimated marginal OR and the true marginal OR, divided by the true marginal OR. η = 0.354 and 0.514 represent that 5% and 10% variance of Z is explained by G, respectively.

Table S6. Results of the estimated odds ratio (OR) for a non-genetic biomarker that associates with risk of first event, survival, and risk of a subsequent CHD event (scenario 2)

	Conditional OR		Marginal OR																	
			Population 1		Population 2								Population 3							
G on Z	X on D_2	Z on D_2	Naive estimator		True		Naive estimator					Bias	True		Naive estimator					Bias
η	$\exp\beta_X$	$\exp\beta_Z$	OR	POW	OR	SE	OR	SE	SEE	COV	POW	%	OR	SE	OR	SE	SEE	COV	POW	%
0.354	3	1	1	0.055	1	0.002	1	0.029	0.028	0.945	0.056	0.19	1	0.003	1	0.028	0.028	0.951	0.05	-0.09
		1.3	1.3	1	1.24	0.003	1.24	0.035	0.035	0.952	1	0.1	1.26	0.003	1.21	0.034	0.035	0.784	1	-3.32
		1.6	1.6	1	1.49	0.003	1.47	0.042	0.042	0.932	1	-0.95	1.5	0.004	1.41	0.04	0.041	0.439	1	-5.93
		2	1.99	1	1.78	0.004	1.75	0.051	0.052	0.913	1	-1.72	1.82	0.005	1.65	0.049	0.05	0.097	1	-9.36
		3	2.97	1	2.49	0.006	2.41	0.077	0.078	0.83	1	-3.18	2.55	0.007	2.2	0.07	0.072	0.006	1	-13.8
	5	1	1	0.043	1	0.002	1	0.028	0.028	0.956	0.045	0.18	1	0.002	1	0.028	0.028	0.951	0.048	0.06
		1.3	1.29	1	1.2	0.002	1.19	0.033	0.034	0.943	1	-0.84	1.21	0.002	1.16	0.033	0.033	0.662	0.999	-4.32
		1.6	1.58	1	1.38	0.002	1.36	0.04	0.039	0.916	1	-1.45	1.42	0.003	1.3	0.038	0.038	0.153	1	-8.19
		2	1.97	1	1.62	0.003	1.58	0.047	0.047	0.871	1	-2.35	1.67	0.004	1.48	0.043	0.044	0.014	1	-11.41
		3	2.89	1	2.18	0.005	2.1	0.066	0.066	0.77	1	-3.75	2.24	0.005	1.89	0.059	0.059	0	1	-15.56
	10	1	1	0.054	1	0.002	1	0.029	0.028	0.948	0.052	0.03	1	0.002	1	0.029	0.028	0.947	0.052	-0.24
		1.3	1.27	1	1.18	0.002	1.15	0.033	0.033	0.867	0.998	-2.23	1.18	0.003	1.12	0.033	0.032	0.58	0.985	-4.84
		1.6	1.54	1	1.34	0.003	1.28	0.037	0.037	0.69	1	-4.09	1.36	0.003	1.24	0.035	0.035	0.105	1	-8.86
		2	1.88	1	1.54	0.003	1.45	0.042	0.042	0.482	1	-5.6	1.56	0.004	1.37	0.039	0.04	0.006	1	-11.97
		3	2.69	1	1.99	0.005	1.86	0.056	0.056	0.368	1	-6.72	2.03	0.005	1.7	0.051	0.052	0	1	-16.41
0.514	3	1	1	0.056	1	0.002	1	0.028	0.028	0.944	0.056	0.15	1	0.003	1	0.027	0.028	0.951	0.05	-0.09
		1.3	1.3	1	1.24	0.003	1.25	0.034	0.034	0.951	1	0.08	1.26	0.003	1.21	0.033	0.034	0.761	1	-3.47
		1.6	1.6	1	1.48	0.003	1.47	0.041	0.041	0.943	1	-0.75	1.5	0.004	1.41	0.04	0.04	0.377	1	-6.16
		2	1.99	1	1.78	0.004	1.75	0.05	0.051	0.903	1	-1.79	1.81	0.004	1.65	0.047	0.048	0.11	1	-8.89
		3	2.96	1	2.48	0.006	2.41	0.076	0.076	0.852	1	-2.8	2.52	0.007	2.2	0.07	0.071	0.011	1	-12.85
	5	1	1	0.045	1	0.002	1	0.027	0.028	0.954	0.045	0.14	1	0.002	1	0.027	0.028	0.95	0.049	0.03
		1.3	1.29	1	1.2	0.002	1.19	0.033	0.033	0.939	1	-0.87	1.21	0.002	1.16	0.032	0.032	0.635	1	-4.39
		1.6	1.58	1	1.38	0.002	1.36	0.038	0.038	0.916	1	-1.47	1.41	0.003	1.3	0.036	0.036	0.171	1	-7.82
		2	1.96	1	1.63	0.003	1.58	0.045	0.045	0.84	1	-2.7	1.66	0.003	1.48	0.042	0.042	0.019	1	-11.19
		3	2.89	1	2.19	0.004	2.11	0.064	0.064	0.762	1	-3.66	2.24	0.005	1.9	0.058	0.058	0	1	-15.35
	10	1	1	0.053	1	0.002	1	0.028	0.028	0.949	0.051	0	1	0.002	1	0.028	0.028	0.949	0.051	-0.2
		1.3	1.27	1	1.18	0.002	1.15	0.032	0.032	0.864	0.999	-2.29	1.18	0.003	1.12	0.031	0.031	0.565	0.988	-4.82
		1.6	1.54	1	1.34	0.003	1.28	0.036	0.036	0.715	1	-3.79	1.35	0.003	1.24	0.034	0.034	0.102	1	-8.51
		2	1.88	1	1.54	0.003	1.46	0.041	0.041	0.473	1	-5.54	1.56	0.004	1.38	0.039	0.039	0.006	1	-11.93
		3	2.69	1	2.01	0.005	1.87	0.056	0.055	0.334	1	-6.81	2.04	0.005	1.71	0.051	0.051	0	1	-15.93

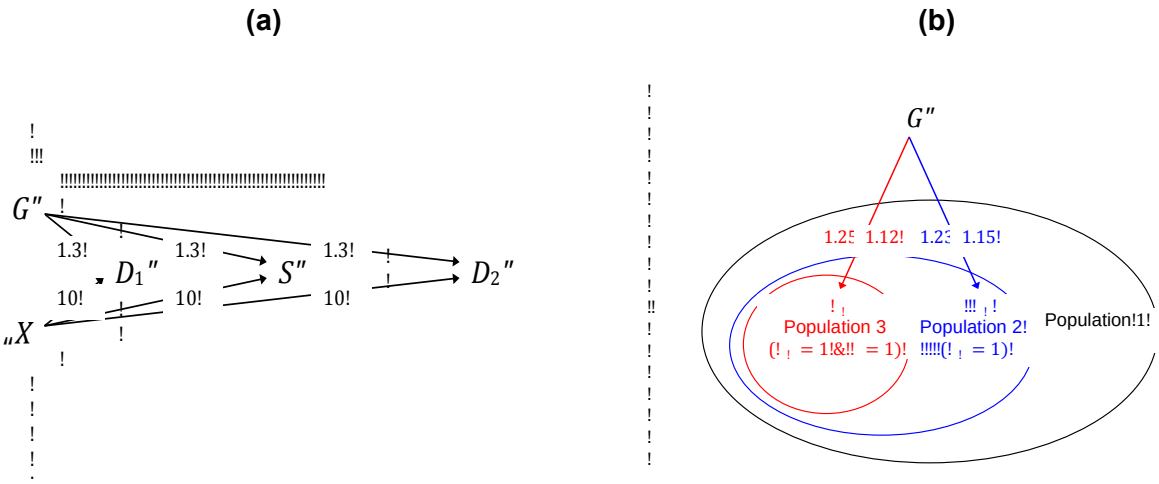
The naïve estimator was obtained from logistic regression. The true value was obtained from the counterfactual method. SE is the empirical standard error. SEE is the mean of standard error estimates. COV is the coverage of the 95% CI. POW is the power when $\exp\beta_Z \neq 1$ and the type 1 error when $\exp\beta_Z = 1$. Bias (%) is the difference between the estimated marginal OR and the true marginal OR, divided by the true marginal OR. η = 0.354 and 0.514 represent that 5% and 10% variance of Z is explained by G, respectively.

Table S7. Results of the estimated odds ratio (OR) for a genetic variant (scenario 1) in case-control studies

Conditional OR		Marginal OR													
		Population 1		Population 2						Population 3					
X on D_2	G on D_2	Naive estimator		True		Naive estimator			Bias	True		Naive estimator			Bias
$\exp\beta_X$	$\exp\beta_G$	OR	POW	OR	SE	OR	COV	POW	%	OR	SE	OR	COV	POW	%
3	1	1	0.05	1	0.003	1	0.954	0.043	-0.37	1	0.003	1	0.955	0.046	-0.16
	1.3	1.3	1	1.25	0.003	1.24	0.954	1	-0.3	1.26	0.004	1.21	0.517	1	-3.51
	1.6	1.6	1	1.48	0.004	1.47	0.951	1	-0.29	1.5	0.005	1.41	0.073	1	-6.07
	2	1.99	1	1.77	0.004	1.76	0.931	1	-0.62	1.8	0.005	1.65	0.006	1	-8.12
	3	2.98	1	2.45	0.006	2.42	0.913	1	-1.17	2.5	0.007	2.22	0	1	-11.48
5	1	1	0.052	1	0.002	1	0.952	0.047	-0.08	1	0.003	1	0.949	0.048	0.27
	1.3	1.29	1	1.2	0.003	1.19	0.924	1	-0.87	1.21	0.003	1.16	0.372	1	-4.21
	1.6	1.58	1	1.38	0.003	1.36	0.872	1	-1.42	1.41	0.004	1.3	0.013	1	-7.47
	2	1.97	1	1.62	0.004	1.58	0.798	1	-2.06	1.65	0.004	1.48	0	1	-10.58
	3	2.93	1	2.17	0.005	2.09	0.52	1	-3.56	2.23	0.006	1.88	0	1	-15.44
10	1	1	0.057	1	0.003	1	0.953	0.046	-0.29	1	0.003	1	0.951	0.044	-0.39
	1.3	1.27	1	1.18	0.003	1.15	0.739	1	-2.4	1.18	0.003	1.12	0.248	1	-4.86
	1.6	1.54	1	1.35	0.004	1.29	0.301	1	-4.49	1.36	0.004	1.24	0.001	1	-8.71
	2	1.89	1	1.55	0.005	1.46	0.088	1	-5.93	1.56	0.005	1.38	0	1	-11.95
	3	2.75	1	2.06	0.007	1.86	0	1	-9.64	2.08	0.007	1.7	0	1	-18.29

The case-control design with 12500 cases and 12500 controls was adopted for populations 1, 2, and 3. The naïve estimator was obtained from logistic regression. The true value was obtained from the counterfactual method. SE is the empirical standard error. SEE is the mean of standard error estimates. COV is the coverage of the 95% CI. POW is the power when $\exp\beta_G \neq 1$ and the type 1 error when $\exp\beta_G = 1$. Bias (%) is the difference between the estimated marginal OR and the true marginal OR, divided by the true marginal OR.

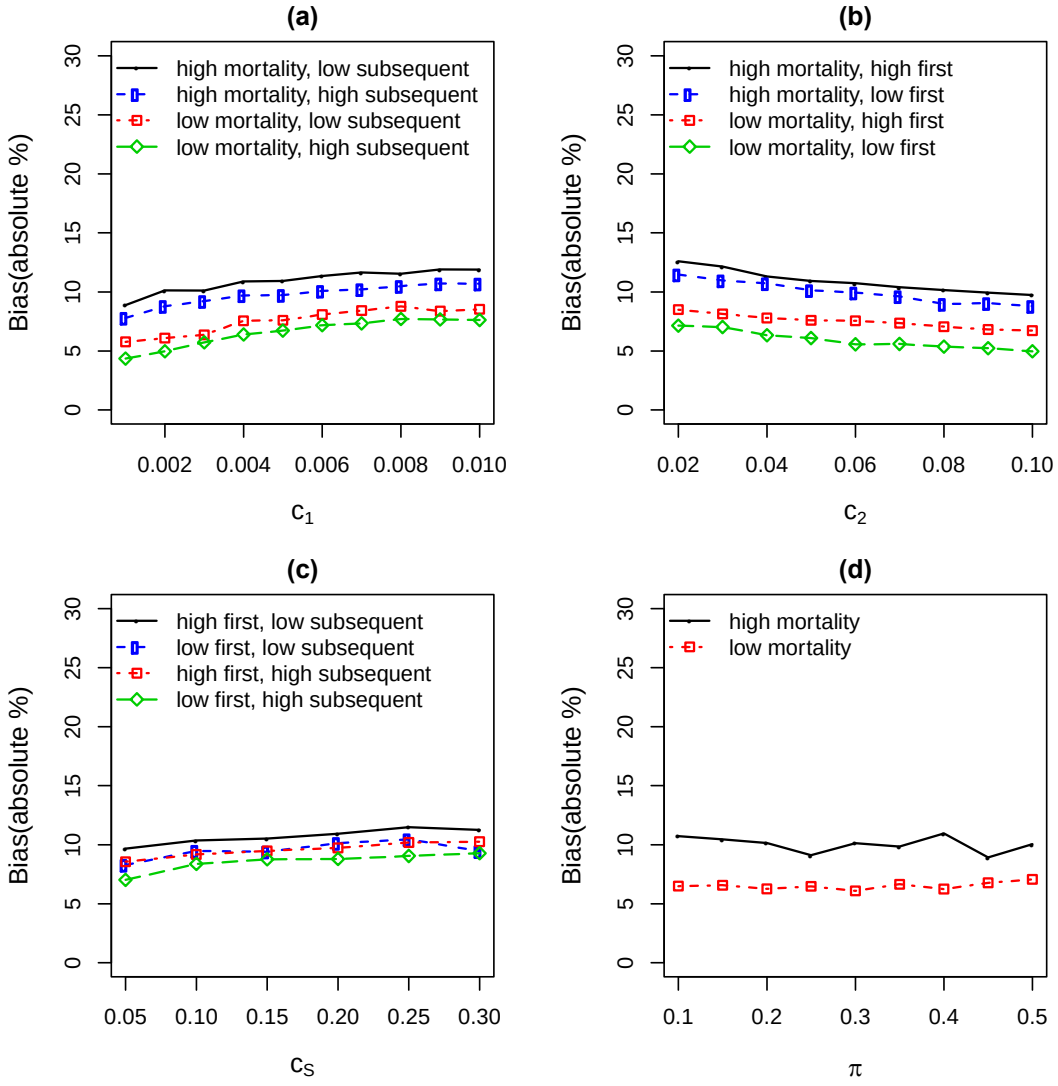
Figure S1. Illustration of different effect sizes



(a) The outcomes D_1, S, and D_2 are generated by the logistic regression models (1) and (2) and the proportional hazards model (3), respectively, with a conditional effect of 1.3 for the SNP of interest (G) and a conditional effect of 10 for all remaining factors (X) in all three models.

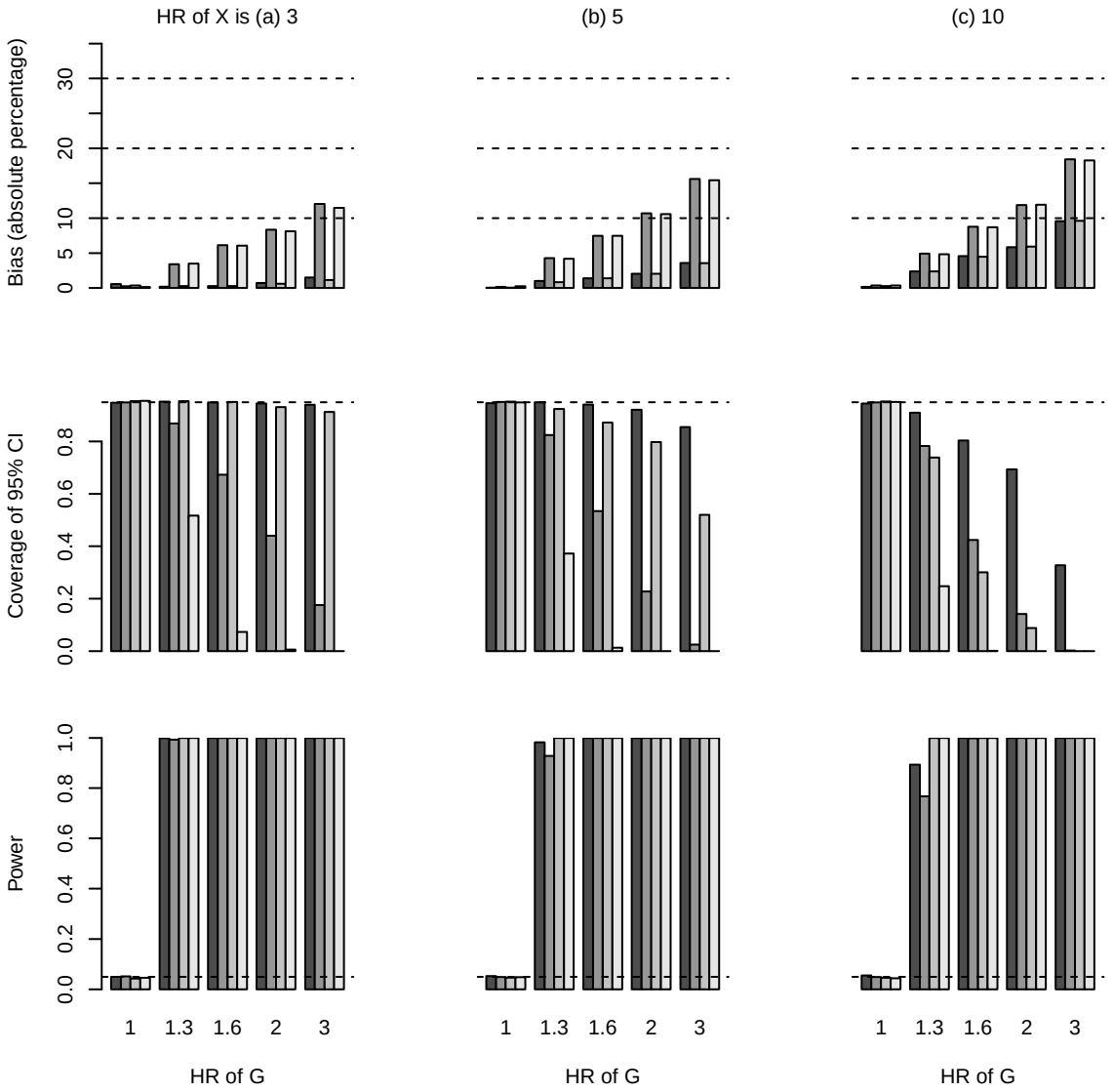
(b) Conditioning on having the first CHD event ($D_1 = 1$, population 2), the true marginal hazard ratio (HR) is estimated to be 1.23 by the counterfactual method and the naïve marginal HR estimate is 1.15 by the standard Cox model, resulting in a bias of 6.09% towards the null; conditioning on both having and surviving the first CHD event ($D_1 = 1$ and $S = 1$, population 3), the true and naïve marginal HR estimates are 1.25 and 1.12, respectively, resulting in a bias of 10.21% towards the null.

Figure S2. Bias as the overall disease rate in the general population (c_1), rate of non-censored subsequent events (c_2), death rate (c_S), and SNP minor allele frequency (π) vary



All results were obtained under scenario 1 (where the genetic variant associates with risk of first event, survival, and risk of subsequent event) when D_2 is time to event and the hazard ratios for G and X are 1.3 and 10, respectively. "high mortality": $c_S = 0.2$. "low mortality": $c_S = 0.0$. "high first": $c_1 = 0.005$. "low first": $c_1 = 0.002$. "high subsequent": $c_2 = 0.1$. "low subsequent": $c_2 = 0.05$. For (a)-(c), $\pi = 0.3$. For (d), $c_1 = 0.002$ and $c_2 = 0.05$.

Figure S3. Results of the estimated odds ratio (OR) for a genetic variant (scenario 1) in cohort and case-control studies



The sample size is 25000 for both study designs. The four bars (from left to right) at each HR of G pertain to population 2 (selection of subjects with fatal or non-fatal first events) in cohort studies, population 3 (selection of subjects with non-fatal first events) in cohort studies, population 2 in case-control studies, and population 3 in case-control studies. The dashed line in the middle panel indicates the expected coverage of 0.95. The dashed line in the lower panel indicates the nominal significance level of 0.05. Power under the HR of 1 for G means type 1 error.

Appendix

GENIUS-CHD Consortium full co-author list with affiliations to be supplied