Universiteit
Leiden
The Netherlands

Genetic and methodological aspects of statin-induced lipid response

Smit, R.A.J.

Citation
Smit, R. A. J. (2019, April 18). *Genetic and methodological aspects of statin-induced lipid response*. Retrieved from https://hdl.handle.net/1887/71733

Version:	Not Applicable (or Unknown)
License:	Leiden University Non-exclusive license
Downloaded from:	https://hdl.handle.net/1887/71733

Note: To cite this publication please use the final published version (if applicable).

Cover Page

Universiteit Leiden

The handle http://hdl.handle.net/1887/71733 holds various files of this Leiden University dissertation.

Author: Smit, R.A.J.
Title: Genetic and methodological aspects of statin-induced lipid response
Issue Date: 2019-04-18

CHAPTER 7

Survival bias in Mendelian randomization studies: a threat to causal inference

Roelof AJ Smit, Stella Trompet, Olaf M Dekkers, J Wouter Jukema, Saskia le Cessie

Submitted

ABSTRACT

It has been argued that survival bias may distort results in Mendelian randomization studies in older populations. Through simulations of a simple causal structure we investigate which factors influence the extent of this bias in the context of exposures which affect survival. We observed that selecting on survival may decrease instrument strength and will, for exposures with directionally concordant effects on survival (and the outcome), introduce bias towards the null for the instrument-outcome association if the true causal effect is not equal to null, and bias from the null if the true causal effect is null. Stronger selection effects and higher ages at study inclusion generally increased this bias when the true causal effect was not equal to null. Moreover, the impact of this bias may differ depending on the distribution of the exposures. The bias in the estimated exposure-outcome relation depended on whether Mendelian randomization estimation was conducted in the one- or two-sample setting. Finally, we discuss how survival bias may be detected in epidemiological cohorts, and which statistical approaches might help to alleviate this and other types of selection bias.

INTRODUCTION

An increasing number of studies are proposing genetic instruments to examine the causal effect of (typically modifiable) exposures on health states or disease. This approach is known as Mendelian randomization. The basic idea is that a genetic marker (polymorphism or haplotype) serves as a proxy for a particular exposure, under the assumption that the potential effect of the genetic marker on the outcome of interest is only through this exposure. Given the continued methodological developments in the field of Mendelian randomization (1), and that summary level statistics from genome-wide association studies (GWAS) are increasingly made publicly available, it is expected that this trend will continue for the foreseeable future. Considerable efforts are being made to facilitate and standardize this advance in Mendelian randomization studies (2-4).

Although often assumed to give a valid causal estimate in contexts where observational evidence might be biased due to residual confounding or reverse causation, Mendelian randomization studies can give biased results when analyses are performed in selected subgroups, as spurious associations may emerge when selection is performed on a common effect of two variables – "one of which is either the treatment or a cause of treatment, and the other is either the outcome or a cause of the outcome" (5). Formally known as collider-stratification bias in causal graph theory, this specific form of selection bias has been suggested to contribute to several counterintuitive phenomena in the clinical literature. These include observations that maternal smoking is associated with lower infant mortality amongst low birthweight infants (the 'birthweight paradox') (6, 7), that obesity is associated with greater survival in individuals with certain chronic diseases (the 'obesity paradox') (8), and that higher levels of serum cholesterol and blood pressure appear protective in the oldest old (9-12). The latter examples are thought to exemplify a subtype of selection bias, known as survival bias, caused by only recruiting or analyzing the non-random subset of the population who have survived long enough to be included.

It has been argued that in Mendelian randomization studies in older populations, survival bias may distort results (13, 14). While this issue has received limited attention in the literature, some researchers have recognized this potential source of bias. For example, Østergaard and colleagues noted that the protective associations of systolic blood pressure with Alzheimer's disease observed in their Mendelian randomization study might arise as a result of differential survival bias (15). Another notable discussion of survival bias followed the observation that variants known to increase BMI associated with a lower risk of Parkinson's disease (16), which contrasted with the null effect

observed in a large meta-analysis of cohort studies on the topic (17). We aimed to investigate the impact of survival bias on Mendelian randomization analyses through a simulation study. In this paper, we will describe which factors influence the extent of this bias. We will also discuss how to determine whether survival bias is present in epidemiological cohorts, and which (statistical) approaches may help to minimize or address this bias.

METHODS

Review of the theory

We define *X* as the exposure and *Y* as the outcome of interest (**Figure 1**). Drawing valid conclusions from a Mendelian randomization analysis requires using a genetic instrument *G* (e.g. a single-nucleotide polymorphism) that meets three key assumptions: i.) *G* explains variation in exposure *X*, ii.) *G* is independent of the (known and unknown) confounders *U* of the association between *X* and the outcome *Y*, and iii.) *G* is independent of *Y* given *X* and *U* (18). In addition, in order to obtain a point estimate of a causal estimate, a fourth assumption is required. This may either be the assumption of homogeneity, or the sometimes more plausible, alternative assumption of monotonicity (19). If these assumptions hold, a causal effect of *X* on *Y* can be reliably estimated, as the association between *G* and *Y* should be essentially free from reverse causality and residual confounding (20).

Consider the following example where we are interested in a causal effect of *X* (e.g. cholesterol) on a continuous outcome *Y* (e.g. cognitive test performance) (**Figure 2**). The inherent concept of Mendelian randomization, that alleles are randomly assigned at conception, would normally ensure that the association measure between *G* and *Y* can be attributed solely to the effect of the exposure *X* on *Y*. However, we must consider that for older populations the study population is restricted by design, including only those who have survived until a certain age.

In this situation, survival until study inclusion (*S*) is influenced by the exposure of interest *X* and a second exposure *R* (e.g. smoking) (**Figure 2A)**. For the purpose of simplicity we assume that these two exposures are uncorrelated in the unselected population. However, if we condition on a common effect of *X* and *R*, i.e. survival (*S*=1), we induce an association between *X* and *R*, and therefore also between *G* and *R*. More intuitively, if someone survives until study inclusion with risk factor *R* (i.e. smokes), they are less likely to also have high levels of risk factor *X* (i.e. hypercholesterolemia), and in extension less likely to have

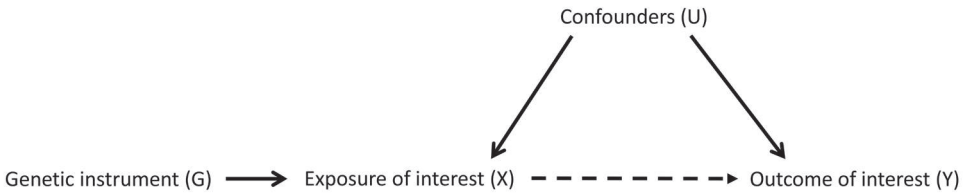


Figure 1. Schematic outline of the Mendelian randomization approach

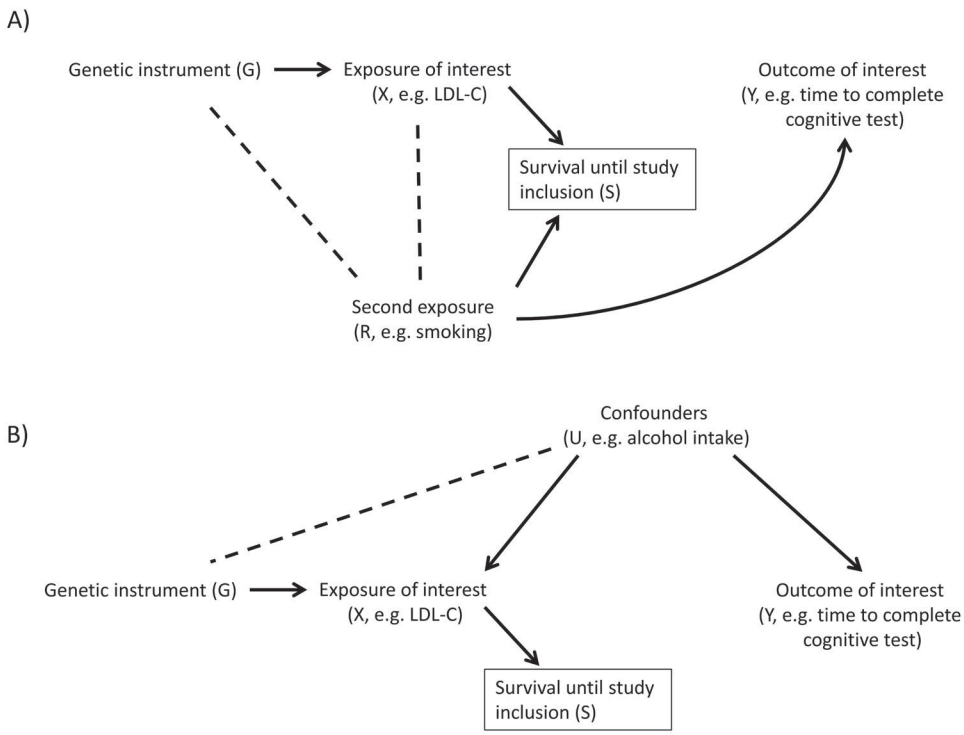


Figure 2. In the example of two exposures affecting probability of survival, conditioning on survival (*S*) may induce an association between previously uncorrelated risk factors *X* (and its genetic proxy *G*) and *R* (dashed lines shown in A). More intuitively, if you are a smoker and still alive at study inclusion, you are less likely to also have a high level of LDL-cholesterol (LDL-C), and vice versa. Additionally, conditioning on survival may induce an association between the genetic instrument G and any confounders *U* of the *X*-*Y* association (dashed line shown in B), even in the absence of risk factor *R*. In both situations, the association between the genetic instrument and the outcome of interest might thus become biased. Please note that while we did not include a line from *X* to *Y* in either causal structure, we also simulate scenarios where *X* does have a causal effect on *Y*. Adapted from Boef AG, et al. (13)

inherited trait *G* that causes hypercholesterolemia. We therefore expect that the previously uncorrelated, competing variables will become negatively associated when restricting the analyses to the 'survivors'. It follows that the estimated *G*-*Y* association can therefore no longer be solely attributed to the effect of exposure *X* on outcome *Y* (i.e. will become biased), as conditioning on the common effect of *X* and *R* has opened an indirect path from *G* to *Y* going through *R* (21).

In the example above we have assumed no confounders exist of the *X*-*Y* association. However, their presence may be problematic (**Figure 2B**). This is because restricting the analysis to survivors means that entry into the study becomes conditional upon the value of *X*. As *X* in turn depends both on the genetic instrument *G* and the confounder *U* (e.g. alcohol intake), *G* and *U* may become correlated.

In essence, the third assumption described at the start of this paragraph has become violated through survival bias in both examples. While our simulations will primarily focus on the causal structure shown in **Figure 2A**, we present simulations on **Figure 2B** and on a combination of these causal structures in the supplemental material.

Data generation

All simulation scenarios assume the basic causal structure shown in **Figure 2A**. All causal associations between variables are chosen such that an increase in cause will lead to an increase in the consequence, except for the effect on survival where higher values in exposure, confounder or risk factor correspond to lower survival times. In addition, our simulations assumed constant treatment effects. For each scenario we generated a dataset of 10 million observations with multiple randomly generated variables: a binary genetic instrument (*G*), a continuous exposure (*X*) influenced by *G*, a second exposure (*R*), a continuous outcome (*Y*) principally influenced by *R* and in later scenarios also by *X*, and finally an age of death influenced by both *X* and *R*. In secondary analyses we additionally generate a continuous confounder (*U*) with equal effects on *X* and *Y* (appendix). All simulations were performed separately for binary and continuously distributed *R*'s.

Details of data generation and parameters values are presented in **Table 1**. Of note, *X* was standardized to have a mean of 0 and standard deviation of 1. The effect of *G* on *X* was chosen such that the corresponding strength of the instrument, measured by the partial R^2, equaled 1, 5, 10, and 15%. In addition, while the per-unit effect size was the same for the two types of *R*, the different scales of measurement (dichotomous (e.g. presence or absence

of hypercholesterolemia) against per standard deviation increase) means their impact on other variables will differ.

Table 1. Parameters values and details of data generation

Parameter (scale)	Data generation and alternative values	Standard value
G (binary)	Prevalence of 25, 50, 75%	50%
X (continuous)	Normally distributed with mean 0 and var(X|G)=1, with varying contribution of G (and if applicable U)	
Variance of X explained by G	1, 5, 10, 15% of X	5%
U (continuous)	Normally distributed with μ=0, σ=1. Only included in scenarios in appendix	
Effect of U on X	Increase of 0.5 per one unit increase in U	None
R (binary)	Prevalence of 12.5, 25, 50%	25%
R (continuous)	Normally distributed with μ=0, σ=1	
Age of death	Gompertz distributed with baseline parameters $a=4.59053\times10^{-5}$ and $b=8.76978320\times10^{-2}$, with varying (additional) contribution of X and R	
Effects of X on age of death	HR of 1.1, 1.25, or 1.5 per one unit increase in X	HR 1.25
Effects of R on age of death	HR of 1.25, 1.5, 2, or 4 if R=1 (binary R) or per one unit increase in R (continuous R)	HR 1.5
S (binary)	Indicates whether age of death is larger than age at inclusion	
Y (continuous)	Normally distributed with mean 0 and variance(Y|X,R)=1, with varying contribution of X and R (and if applicable U)	
Effects of X on Y	Increase of 0, 0.5, 1, or 2 per one unit increase in X	0
Effects of R on Y	Increase of 0.25, 0.5, or 1 if R=1 (binary R) or per one unit increase in R (continuous R)	0.5
Effect of U on Y	Increase of 0.5 per one unit increase in U	None
Number of observations	10.000.000 in all scenarios	

S.D. denotes standard deviation.

To generate survival time we obtained the 2016 mortality data of the United States from the Human Mortality Database (22). Using the *MortalityLaws* R-package (23) we estimated the parameters of the Gompertz model (24) within

this real-world dataset (**eFigure 1**), which were subsequently used to generate survival times for our simulated population. Effects of both *X* and *R* on age of death were modelled as hazard ratios, with having higher levels of *X* and/or *R* translating into an earlier death (on average), and lower levels of *X* and/or *R* in a later death (on average). Subsequently, we considered different age boundaries for study inclusion, from 75-95 years, thereby steadily decreasing the number of surviving participants (*S*=1). We used R (version 3.4.1) for all data generation and analyses (25). Sample code is provided as supplemental material.

Effects on instrument strength

Firstly, we examined whether selecting on survival may influence the strength of instrument *G*, reflected by the squared correlation between *G* and *X* (R^2), which indicates how much variance of *X* is explained by *G*. Given that selecting on survival will yield smaller data sets, and that the F-statistic strongly depends on sample size, we did not consider the F-statistic as a measure of instrument strength (26). We chose different strengths of the instrument, while all other parameter values were kept fixed at a standard value given in **Table 1**. No effect of *X* on *Y* was assumed.

Effects on association between the genetic instrument G and exposure R

Secondly, we considered the effect of different parameters on the induced correlation between *G* and *R* within an increasingly selected population. Effects of changing the following parameters were considered:

i. variance of *X* explained by *G* (R^2);
ii. effects of *X* on age at death;
iii. effects of *R* on age at death;
iv. effects of *R* on *Y*;
v. prevalence of *G*;
vi. prevalence of *R* (for dichotomous *R*).

In each simulation, the other parameters were held at their standard values, and no effect of *X* on *Y* was assumed. Accompanying confidence intervals for the correlation between *G* and *R* were calculated using Fisher's z-transformation. (27)

Effects on association between genetic instrument G and outcome of interest Y

Thirdly, we examined how this induced correlation between *G* and *R* influences the *Y*~*G* association, estimated with linear regression. Different true effects of *X* on *Y* were assumed (**Table 1**). Other parameters were again held at their standard value.

Effects on instrumental variable (IV) estimators

Finally, we considered how the induced correlation between *G* and *R* might influence an IV-estimator. In its simplest form this estimator equals the ratio of regression coefficients, known as the Wald ratio (28), defined for our continuous outcome *Y* as

$$Wald\ ratio = \frac{coefficient\ of\ G\ in\ regression\ of\ outcome\ Y\ on\ G}{coefficient\ of\ G\ in\ regression\ of\ exposure\ X\ on\ G} = \frac{\hat{\beta}(Y \sim G)}{\hat{\beta}(X \sim G)}$$

The Wald ratio thus quantifies the causal effect of the exposure on the outcome and estimates the mean increase in outcome per unit increase in exposure. Increasingly, summarized data (coefficients and standard errors) from large genome-wide association study (GWAS) consortia are made publicly available, which enable researchers to perform two-sample Mendelian randomization even if their own study does not allow for estimation of both coefficients necessary to calculate the Wald ratio (29). These external datasets are generally more likely to have primarily included middle-aged participants (30-32), and thus less likely to be affected by survival bias. Therefore, under the assumption of no age-related effect modification, we not only considered the scenario where both coefficients are estimated with linear regression in the same increasingly selected dataset (i.e. 'internal' estimation), but also what happens if the association measure between *G* and *X* were to be taken from an external dataset not selected on survival (i.e. 'external' estimation, by taking the fixed value of our total population). Confidence intervals for the internally estimated Wald ratio were calculated using the *tsls* function from the *sem* R-package (33).

RESULTS

Instrument strength

As shown in the main plot of **Figure 3**, the variance explained in exposure *X* by *G* decreases when higher ages-at-inclusion are considered. The decline in R^2 between age 75 and 95 years is greater in absolute terms, but comparable in relative terms, for stronger genetic instruments. For example, for the instrument explaining 1% of variance in X in the unselected (i.e. entire) sample R^2 declined from 0.99% at 75 years to 0.90% at 95 years, set against a decline from 14.75% to 13.35% for the instrument originally explaining 15% of variance in X. Shown in the figure's insets are the **A)** the change in prevalence of *G* and **B)** the survival curve for the population from 75 years until 95 years. The prevalence of *G* was observed to decline from 0.49 at age 75 years to 0.46 at age 95. Furthermore, of the population alive at 75 years, 15.6% was still alive at 95 years. Results for the continuously distributed R were comparable (**eFigure 2**).

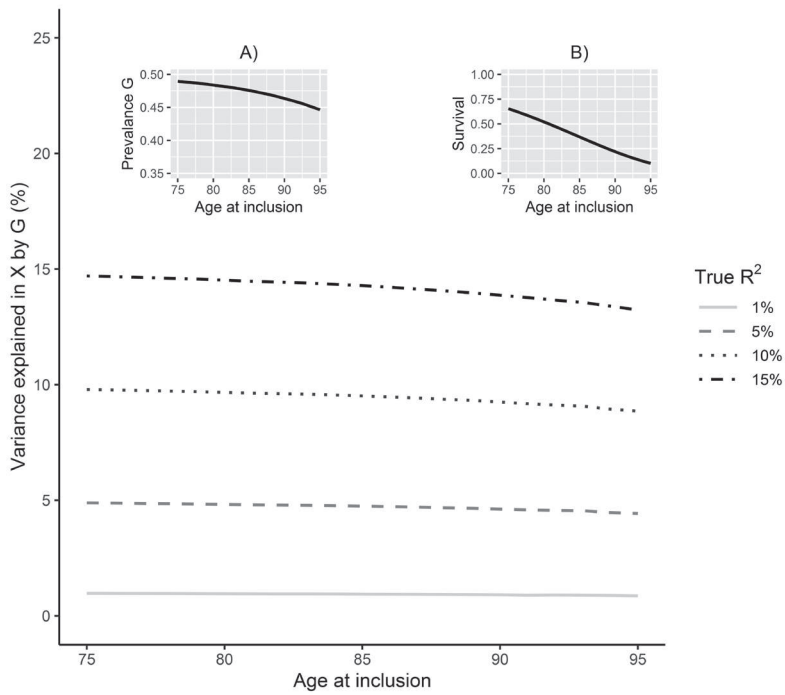


Figure 3. Variance explained in the exposure of interest *X* by its genetic proxy *G* for an increasingly selected population, when incorporating a binary *R*. Shown in the insets are A) the prevalence of *G* and B) the accompanying survival curve, both with the true (i.e. unselected) R^2 set at 5%.

Correlation between G and R

The induced negative correlation due to selection on age between the causally independent variables *G* and *R* across different simulation scenarios is shown in **Figure 4**. Keeping all other parameters constant, the correlation becomes more negative as **i.** the instrument is stronger (i.e. more variance in *X* is explained by *G)* (**4A,B**), **ii.** *X* has greater effects on age at death (**4C,D**), **iii.** *R* has greater effects on age at death (**4E,F**), and **iv.** as *R*'s prevalence becomes greater (dichotomous *R)* (**4K**). However, once the prevalence of *R* exceeds 0.5 the induced correlation between *G* and *R* decreases again. In contrast, the correlation remains constant for different effects of *R* on *Y* (**4G,H**), and is largely unchanged by changing the prevalence of G (**4I,J**). Of note, the association between age-at-inclusion and the induced *G~R* correlation attenuates at higher ages when the deleterious effect of *R* on *S* corresponds to an hazard ratio of 4 (**4E**), with the nadir of the curve occurring between 80 and 85 years of age. This specific example likely results from the rapid depletion of the *R*-carrying participant pool, an effect also visible but less extreme for the simulations incorporating a continuous *R* (**4F**).

Bias to Y ~ G association

Varying the true underlying effect of *X* on *Y* reveals how the association between *G* and *Y* is biased by selecting on *S*=1 (**Figure 5**). In cases where the true effect ≠ 0, a bias towards the null is seen, underestimating the true effect. While this bias is greater in absolute terms, in relative terms we observe a slight attenuation across different effects of *X* on *Y* when considering a dichotomous *R* (at 95 years: 12.3% underestimation for true effect of 0.5 (**5C**) versus 10.1% for true effect of 2 (**5E**). A different pattern was observed for the situation where the true effect of *X* on *Y* is null. In this case, where the statistical association between the genetic variant and the outcome of interest is completely due to bias, the resulting association becomes nominally negative (**5A**). The same, but slightly exaggerated pattern occurs for a continuously distributed *R*. The *Y~G* association namely moves away from the null to a considerably greater extent when the true effect of *X* on *Y* is null (**5B**), and greater attenuation of the bias towards the null occurs for greater effects of *X* of *Y* when the true effect is not equal to 0 (**5D,F**).

Bias to IV estimator

The IV estimator is influenced by survival bias, where magnitude and direction of the bias are dependent on **i.** whether the association measure between *G* and *X* is estimated within the same selected dataset as the association measure

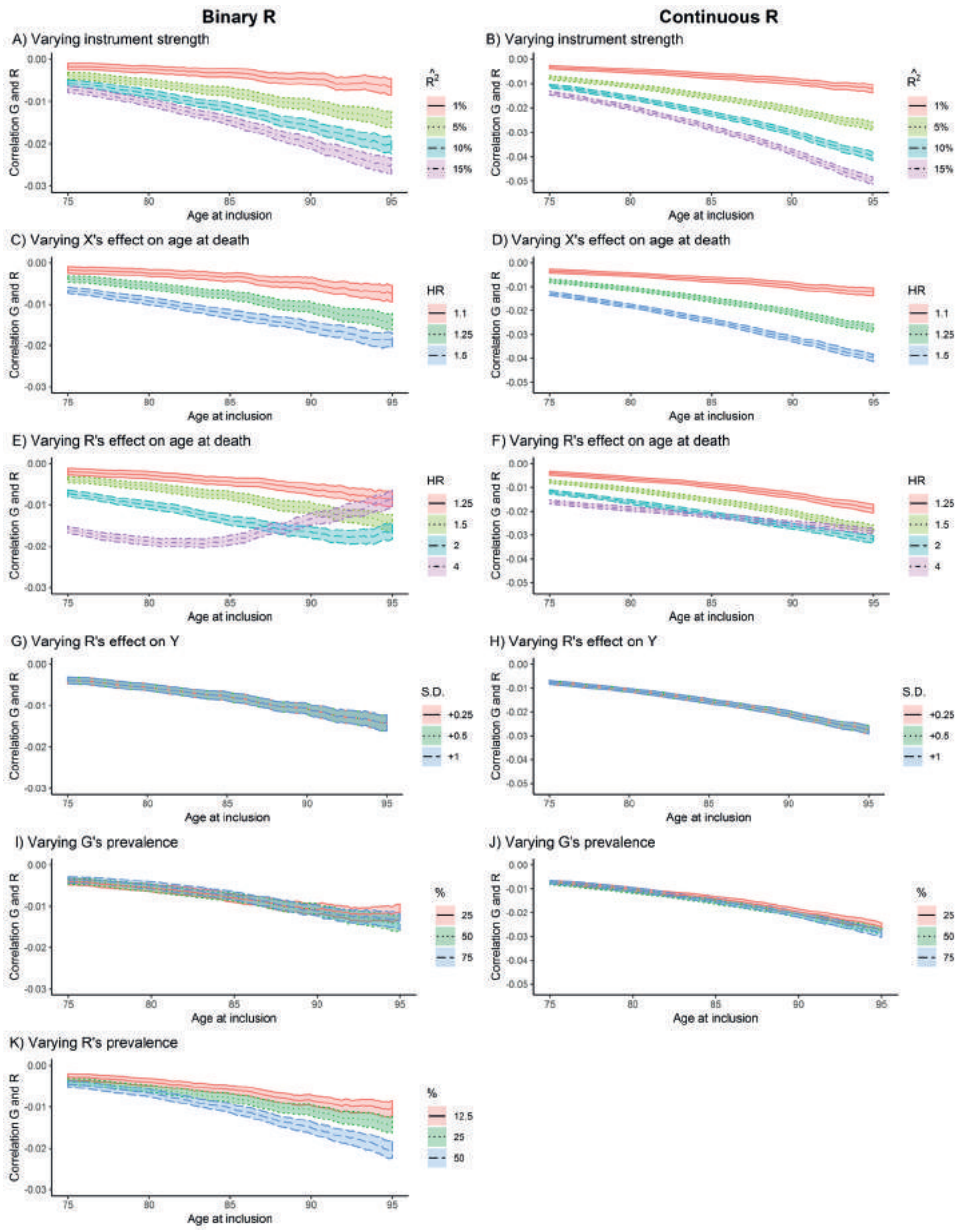


Figure 4. Effect of varying different parameters on the induced correlation (95% CI) between the genetic instrument *G* and the second exposure *R* for an increasingly selected population. Shown for binary (left column) and continuously (right column) distributed *R*. S.D. denotes standard deviation.

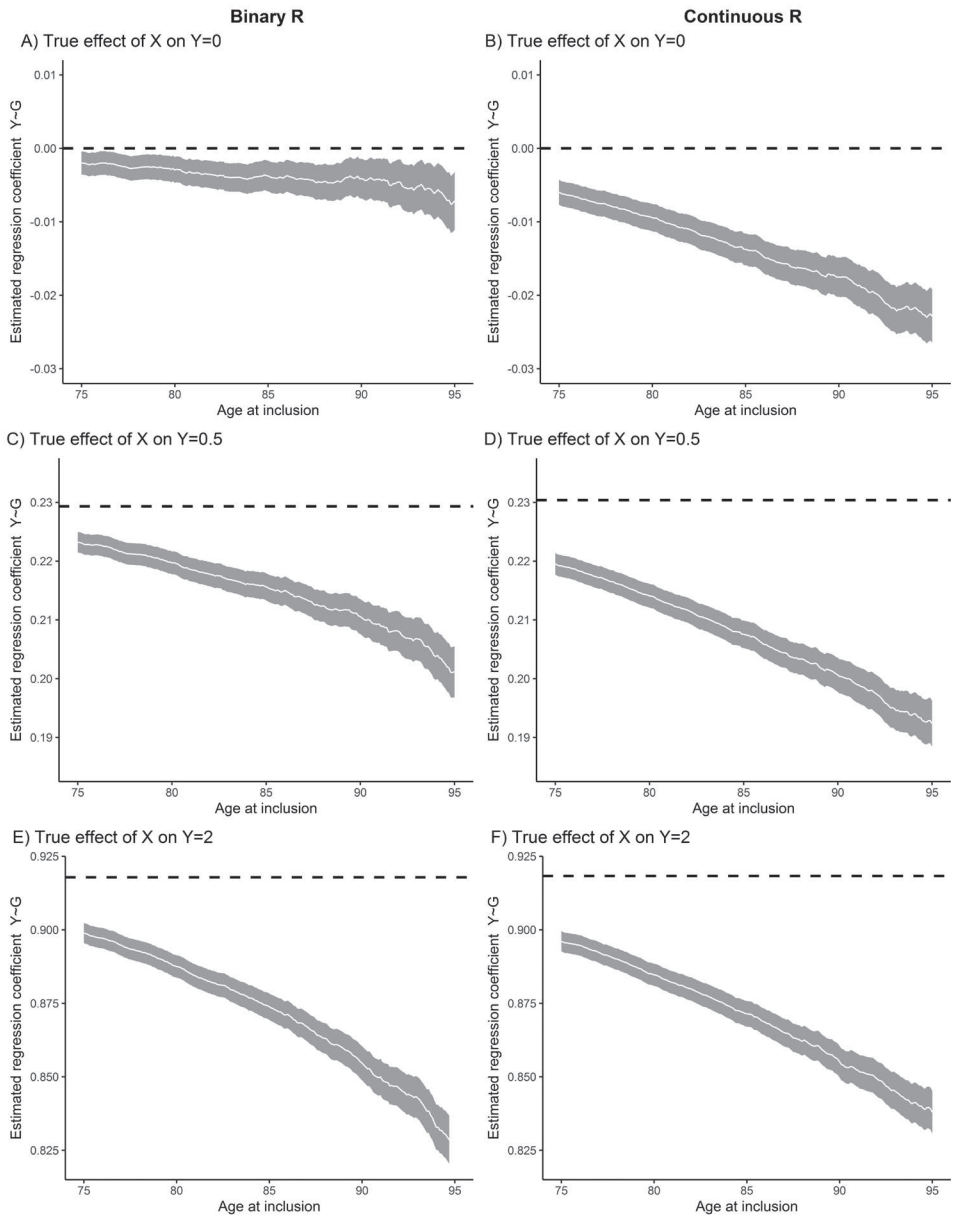


Figure 5. Effect of survival bias on the association between the genetic instrument G and the outcome of interest *Y*, for different true effects of exposure *X* on *Y*. Data are presented as regression coefficients (95% CI) estimated with linear regression. The true (i.e. unselected) regression coefficient for *G* on *Y* is shown as a dashed line in each plot. Shown for binary (left column) and continuously (right column) distributed *R*.

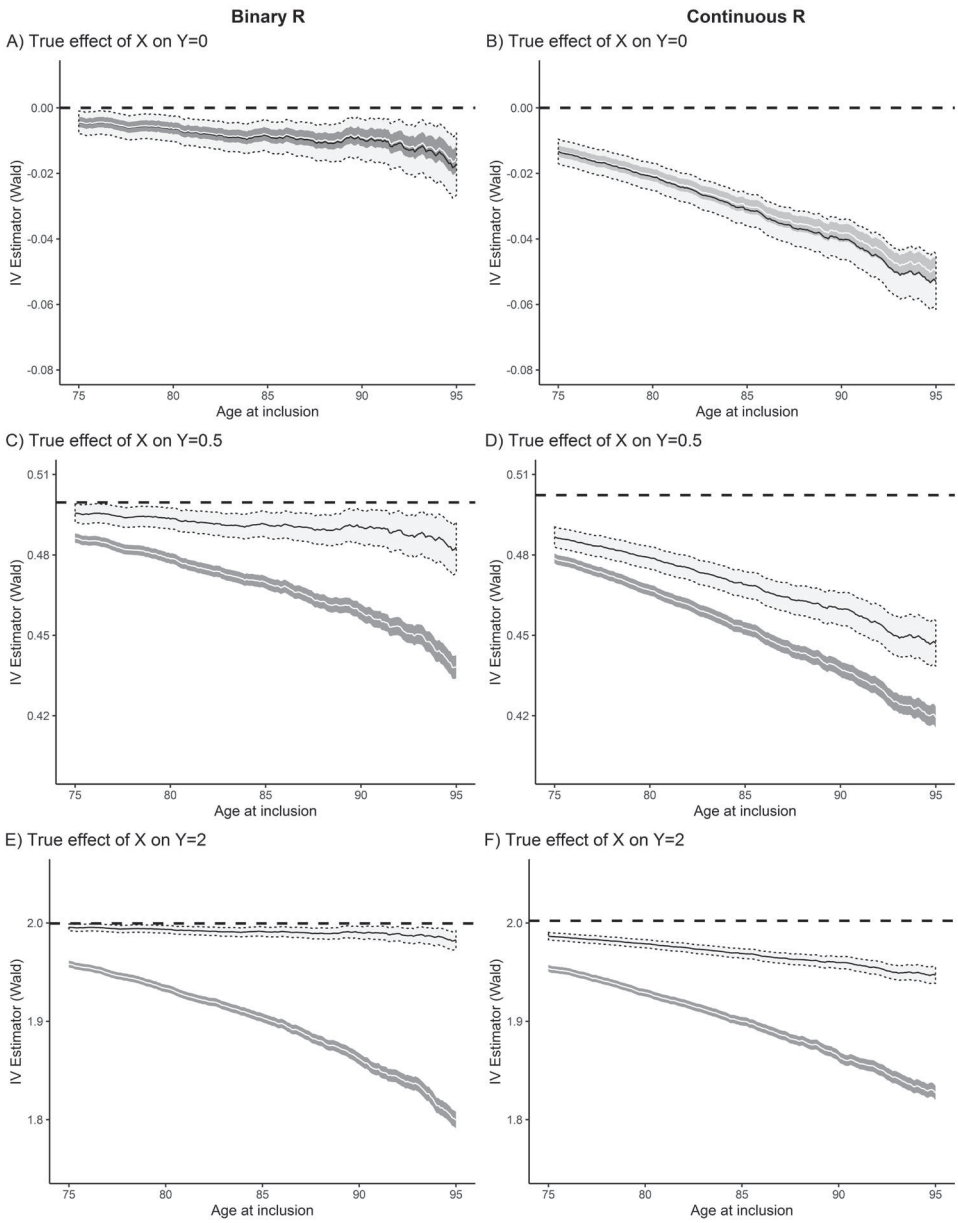


Figure 6. Wald ratios (95% CI) based on internally (white ribbon) versus externally (grey ribbon) estimated association between exposure *X* and the outcome *Y*, for different true effects of exposure *X* on *Y*. Shown for binary (left column) and continuously (right column) distributed *R*. Dashed lines denote the true (i.e. unselected) Wald ratio, which equals the true causal effect of *X* on *Y*.

between *G* and *Y* was, or within an external source not selected on age, and **ii.** whether the true effect of *X* on *Y* is null or not (**Figure 6**). When both the numerator (*Y~G*) and denominator (*X~G*) of the Wald ratio (i.e. our causal effect estimate) are taken from the same selected dataset, we observe that they will be similarly biased. Taking the ratio of these two therefore seemingly cancels out much of the bias to the IV estimator, compared to the situation where only the numerator is taken from a population selected on survival. In this latter situation, the relative degree of the bias equals that seen for the association measure between *G* and *Y*. The two IV estimators diverge more strongly as stronger true effects of *X* on *Y* are considered. This is more clearly observed when a dichotomous *R* is considered. For a continuously distributed *R*, selection bias partially persists for the internally estimated IV estimator (**6B,D,F**).

Alternative causal structures

Simulation results for the causal structure depicted under **Figure 2B**, and for the combination of **2A** and **2B**, did not show markedly different results (**eFigures 4-10**).

DISCUSSION

In this paper we show that previously uncorrelated, competing risk factors may become associated due to selection on survival, consequently biasing estimates from Mendelian randomization studies. More specifically we observed that, if the effect of the exposure of interest on the outcome of interest is genuinely non-null and the selection-related exposures have directionally concordant effects on the outcome, the association measure between genetic proxies of that exposure and the outcome will become biased towards the null. Of further importance is the observation that as the population size decreases instrument strength also weakens, as measured by R^2. The combination of a smaller population size with a weaker instrument strength will be detrimental to statistical power in hypothesis testing. It should be noted here that the decrease in instrument strength not just results from the decreasing prevalence of the genetic instrument, but also due to the genetic instrument becoming associated with the random noise contributing to the exposure (**eFigure 3**). We additionally observed that the induced correlation between *G* and *R* is greater for stronger instruments. However, as bias amplification is smaller for stronger instruments, we expect that instrument strength will not substantially affect the degree of bias of either *Y~G* or IV estimators.

A fundamental assumption in inferring causality using Mendelian randomization is that the genetic instrument should not independently associate with traits of aetiological significance to the outcome other than the exposure of interest. In the simple causal structure considered in our simulations, we observe that this assumption is violated by selection on survival. While we solely explored scenarios with one genetic instrument, this problem will also occur for any combination of genetic instruments for exposures which jointly influence the probability of surviving until study inclusion. In essence, quasi-pleiotropic effects are induced by conditioning on survival till study inclusion. More specifically, given that these pleiotropic effects are unlikely to average to zero across a combination of genetic instruments proxying the same exposure, survival bias is equivalent to introducing directional pleiotropy into Mendelian randomization analyses. To our knowledge it has not been examined whether robust analysis methods specifically aimed at correcting for bias due to unbalanced directional pleiotropy, such as MR Egger regression (34), would be able to cope with this problem. Of particular interest would be whether sets of polygenic instruments, whose individual metabolic pathways to the intermediate phenotype may differ, might be differentially affected by survival bias.

While our simulations specifically examined age-related selection, researchers with data on (younger) populations selected on alternative characteristics (e.g. disease status) will similarly have to consider the possible influence of selection bias in genetic analyses, including genome-wide association testing (35-37). This also holds for investigations within increasingly popular mega-biobanks such as population-based UK Biobank and the Million Veterans Program, both of which have had relatively limited response rates (38-40). Alternative causal structures which might give rise to selection bias in Mendelian randomization studies have been presented elsewhere (40).

There exist several ways for researchers to substantiate the claim that survival bias may be present in their study population, most of which require individual level data. One approach is to examine the associations between the genetic instrument(s) and confounders of the association between exposure and outcome of interest (*X* and *Y*), and/or with variables upon which the population was selected. A key point here is that no association should be present in younger, less-selected populations. Theoretically, if no trends across age are found, it is unlikely that the genetic variant significantly influences mortality. However, this approach will generally only be feasible if large-scale data across different age groups is available on a variety of phenotypic traits, or if the population is strongly enriched or depleted for the trait of interest (35). Leveraging summary

statistics from genome-wide testing performed in large-scale population-based studies may make it possible to differentiate between survival bias-induced associations and alternative pleiotropic mechanisms. Alternatively, researchers can examine whether the strength of the instrument (i.e. the explained variance in the exposure of interest by the instrument) is significantly lower in older than that reported in younger populations. In extension, allele frequencies of high-risk variants are likely to decline in an age-dependent manner, as observed in our simulations, as individuals with a substantially deleterious genetic predisposition will gradually be phased out of the population. This is in line with previously described observations of a large-scale genetic risk score for low-density lipoprotein cholesterol decreasing with increasing age (41). However, it should be noted there does not exist a failsafe method of ruling out survival bias, nor were the above approaches developed for the IV-context under bias amplification. In addition, these methods assume that cohort effects are not present, with younger and older populations coming from the same source population.

Recent work by Canan and colleagues suggests that for the causal structure under investigation in our simulations, selection bias may be corrected via inverse probability weighting (14). In general, we expect that if the selection gradient solely depends on measured variables which are available for the entire original study population (i.e. also for those individuals who are not selected in the study sample), and assuming a constant treatment effect, both inverse probability weighting and multiple imputation could be suitable solutions for selection bias. If data are only available for the selected individuals, but a sufficient set of selection-related variables are precisely measured, then inclusion of these selection-related variables in multivariable regression models may resolve the bias if the models are well-specified. The value of representative cohorts with little selection (e.g. birth cohorts) cannot be overstated in this context (40, 42), though genotyping genetically informative family members may hold promise as well (43). Alternative strategies have been proposed in the context of hazard models (44-46), which may fare better when selection depends on (partially) unobserved variables. In addition, methods of using covariate balance to detect dependent censoring in longitudinal studies exist, though these approaches have not been extended to IV-analysis where bias amplification may occur (47, 48).

We must acknowledge several limitations of our study. In our simulations we made a number of assumptions, due to which caution must be taken in making generalizations. These include that exposures but also genotypes had constant effects during life, ignoring possible antagonistic pleiotropy (49), and that

survival bias would similarly affect different components of the causal structure (e.g. both the numerator and denominator of the Wald ratio). In addition, we solely considered one commonly occurring genetic instrument and uncorrelated exposures with directionally concordant effects on survival (and the outcome of interest). *R* could however be considered a combined vector for many possible competing causes of death before study inclusion. Furthermore, we did not consider a binary outcome of interest, to avoid the issue of non-collapsibility, and restricted our investigations to a linear instrument-exposure association. We also did not examine the effect of possible effect modification between the two exposures, which might lead to stronger correlations between the genetic instrument and exposure *R* and therefore increased bias (50). These choices were aimed at examining the basic underpinnings of survival bias in the context of Mendelian randomization studies, in absence of real-world complexities.

In conclusion, using a simple causal structure we were able to demonstrate that survival bias may lead to biased estimates in Mendelian randomization studies. It will be of interest to examine more detailed simulations in the future, using greater numbers of instruments and exposures to derive bias formulas (as others have done for collider bias in binary variable structures (51)), ideally coupled with comparing the performance of the possible correction methods for survival bias described above. Finally, future work should explore the implications of using different instrumental variable assumptions such as monotonicity, instead of the assumption of homogenous treatment effects of our simulations.

REFERENCES

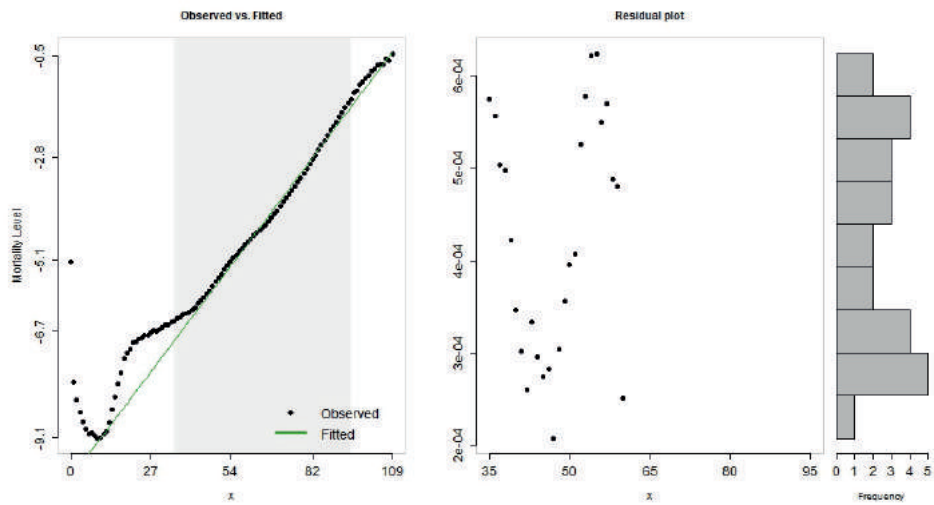
1. Zheng J, Baird D, Borges MC, et al. Recent Developments in Mendelian Randomization Studies. Curr Epidemiol Rep. 2017;4:330-345.
2. Boef AG, Dekkers OM, Le Cessie S, et al. Reporting Instrumental Variable Analyses. Epidemiology. 2013;24:937-8.
3. Hemani G, Zheng J, Elsworth B, et al. The MR-Base platform supports systematic causal inference across the human phenome. eLife. 2018;7:pii: e34408.
4. Haycock PC, Burgess S, Wade KH, et al. Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. Am J Clin Nutr. 2016;103:965-78.
5. Hernán MA, Robins JM. Causal Inference. Boca Raton: Chapman & Hall/CRC, forthcoming. 2018.
6. Yerushalmy J. The relationship of parents' cigarette smoking to outcome of pregnancy--implications as to the problem of inferring causation from observed associations. Am J Epidemiol. 1971;93:443-56.
7. VanderWeele TJ. Commentary: Resolutions of the birthweight paradox: competing explanations and analytical insights. Int J Epidemiol. 2014;43:1368-73.
8. Sperrin M, Candlish J, Badrick E, et al. Collider Bias Is Only a Partial Explanation for the Obesity Paradox. Epidemiology. 2016;27:525-30.
9. Schatz IJ, Masaki K, Yano K, et al. Cholesterol and all-cause mortality in elderly people from the Honolulu Heart Program: a cohort study. The Lancet. 2001;358:351-355.
10. Oates DJ, Berlowitz DR, Glickman ME, et al. Blood pressure and survival in the oldest old. J Am Geriatr Soc. 2007;55:383-8.
11. Molander L, Lövheim H, Norman T, et al. Lower systolic blood pressure is associated with greater mortality in people aged 85 and older. J Am Geriatr Soc. 2008;56:1853-9.
12. Weverling-Rijnsburger AWE, Blauw GJ, Lagaay AM, et al. Total cholesterol and risk of mortality in the oldest old. The Lancet. 1997;350:1119-1123.
13. Boef AG, Le Cessie S, Dekkers OM. Mendelian randomization studies in the elderly. Epidemiology. 2015;26:e15-16.
14. Canan C, Lesko C, Lau B. Instrumental Variable Analyses and Selection Bias. Epidemiology. 2017;28:396-398.
15. Østergaard SD, Mukherjee S, Sharp SJ, et al. Associations between Potentially Modifiable Risk Factors and Alzheimer Disease: A Mendelian Randomization Study. PLoS Med. 2015;12:e1001841.
16. Noyce AJ, Kia DA, Hemani G, et al. Estimating the causal influence of body mass index on risk of Parkinson disease: A Mendelian randomisation study. PLoS Med. 2017;14:e1002314.
17. Wang YL, Wang YT, Li JF, et al. Body Mass Index and Risk of Parkinson's Disease: A Dose-Response Meta-Analysis of Prospective Studies. PLoS One. 2015;10:e0131778.

18. Greenland S. An introduction to instrumental variables for epidemiologists. Int J Epidemiol. 2000;29:722-9.
19. Labrecque J, Swanson SA. Understanding the Assumptions Underlying Instrumental Variable Analyses: a Brief Review of Falsification Strategies and Related Tools. Curr Epidemiol Rep. 2018;5:214-220.
20. Ebrahim S, Davey Smith G. Mendelian randomization: can genetic epidemiology help redress the failures of observational epidemiology? Hum Genet. 2008;123:15-33.
21. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999;10:37-48.
22. Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org or www.humanmortality.de (data downloaded on July 30 2018).
23. Pascariu MD. MortalityLaws: Parametric Mortality Models, Life Tables and HMD. 2018 R package version 1.5.0. https://github.com/mpascariu/MortalityLaws.
24. Gompertz B. On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies. Philosophical Transactions of the Royal Society of London. 1825;115:513-585.
25. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2016.
26. Stock J, Yogo M. Testing for Weak Instruments in Linear IV Regression. Cambridge, 2005.
27. Fisher RA. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika. 1915;10:507-521.
28. Wald A. The fitting of straight lines if both variables are subject to error. Ann Math Stat. 1940;11:284-300.
29. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. Genet Epidemiol. 2013;37:658-65.
30. Locke AE, Kahali B, Berndt SI, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015;518:197-206.
31. Willer CJ, Schmidt EM, Sengupta S, et al. Discovery and refinement of loci associated with lipid levels. Nat Genet. 2013;45:1274-83.
32. Ehret GB, Munroe PB, Rice KM, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature. 2011;478:103-9.
33. Fox J, Byrnes J, Culbertson M, et al. "Package 'sem'," in http://cran.r-project.org/web/packages/sem/sem.pdf. 2018.
34. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int J Epidemiol. 2015;44:512-25.
35. Yaghootkar H, Bancks MP, Jones SE, et al. Quantifying the extent to which index event biases influence large genetic association studies. Hum Mol Genet. 2017;26:1018-1030.

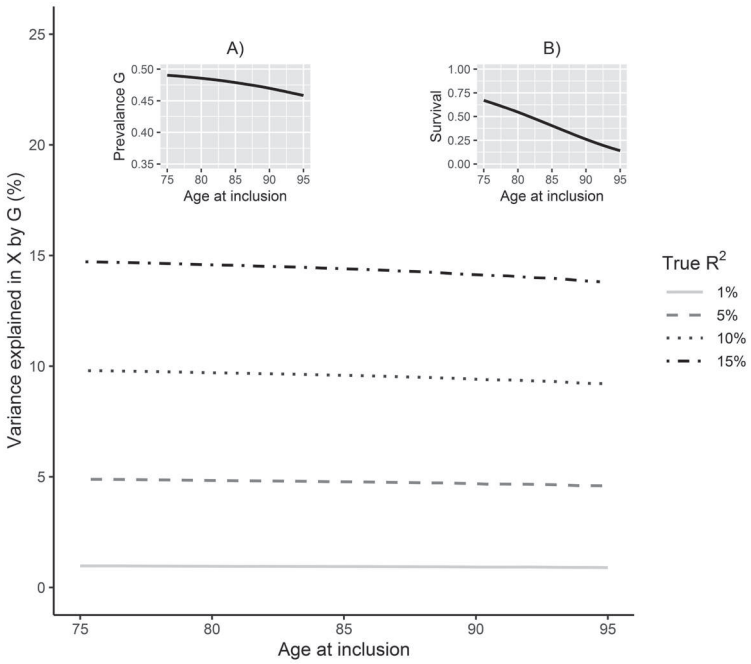
36. Anderson CD, Nalls MA, Biffi A, et al. The effect of survival bias on case-control genetic association studies of highly lethal diseases. Circ Cardiovasc Genet. 2011;4:188-96.
37. Hu YJ, Schmidt AF, Dudbridge F, et al. Impact of Selection Bias on Estimation of Subsequent Event Risk. Circ Cardiovasc Genet. 2017;10.
38. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants with the General Population. Am J Epidemiol. 2017;186:1026-1034.
39. Gaziano JM, Concato J, Brophy M, et al. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. J Clin Epidemiol. 2016;70:214-23.
40. Munafo MR, Tilling K, Taylor AE, et al. Collider scope: when selection bias can substantially influence observed associations. Int J Epidemiol. 2017;47:226-235.
41. Postmus I, Deelen J, Sedaghat S, et al. LDL cholesterol still a problem in old age? A Mendelian randomization study. Int J Epidemiol. 2015;44:604-12.
42. Domingue BW, Belsky DW, Harrati A, et al. Mortality selection in a genetic sample and implications for association studies. Int J Epidemiol. 2017;46:1285-1294.
43. Chen L, Weinberg CR, Chen J. Using family members to augment genetic case-control studies of a life-threatening disease. Stat Med. 2016;35:2815-30.
44. Vansteelandt S, Dukes O, Martinussen T. Survivor bias in Mendelian randomization analysis. Biostatistics. 2018;19:426-443
45. Stensrud MJ. Handling survival bias in proportional hazards models: A frailty approach. arXiv:1701.06014v4 [stat.ME] (preprint). 2017.
46. Carlin CS, Solid CA. An approach to addressing selection bias in survival analysis. Stat Med. 2014;33:4073-86.
47. Zubizarreta JR. Stable Weights that Balance Covariates for Estimation with Incomplete Outcome Data. Journal of the American Statistical Association. 2015;110:910-922.
48. Jackson JW. Diagnostics for Confounding of Time-varying and Other Joint Exposures. Epidemiology. 2016;27:859-69.
49. Toupance B, Godelle B, Gouyon PH, et al. A model for antagonistic pleiotropic gene action for mortality and advanced age. Am J Hum Genet. 1998;62:1525-34.
50. Hernán MA, Hernández-Díaz S, Robins JM. A Structural Approach to Selection Bias. Epidemiology. 2004;15:615-625.
51. Nguyen TQ, Dafoe A, Ogburn EL. The Magnitude and Direction of Collider Bias for Binary Variables. arXiv:1609.00606 [stat.ME] (preprint). 2017.

Supplemental Material

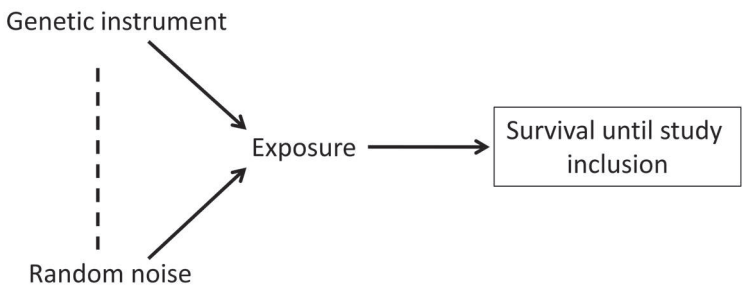
eFigure 1. Results of fitting Gompertz-model using mortality-data (USA 2016)
eFigure 2. Variance explained in X by G for continuously distributed R, for causal structure presented under Figure 2A (main text)
eFigure 3. Genetic instruments and noise in X
eFigure 4: Causal structure also presented in Figure 2B (main text)
eFigures 5-6: Results of simulations for causal structure shown in eFigure 4
eFigure 7: Causal structure combining those presented in Figure 2 (main text)
eFigures 8-10: Results of simulations for causal structure shown in eFigure 7



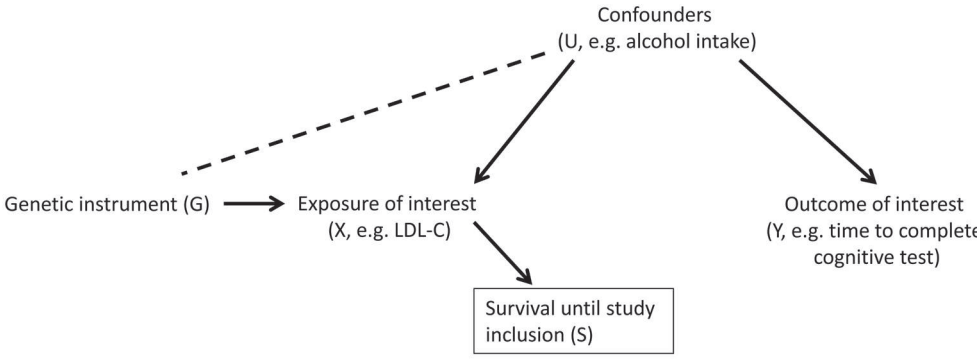
eFigure 1. Results from fitting the Gompertz-model ($a*e^{b*age}$) onto the 2016 mortality data of the United States (age range 35-95) obtained from the Human Mortality Database (www.mortality.org), using the *MortalityLaws* R-package. Estimated model parameters: a, 0.0000459053; b, 0.0876978320.



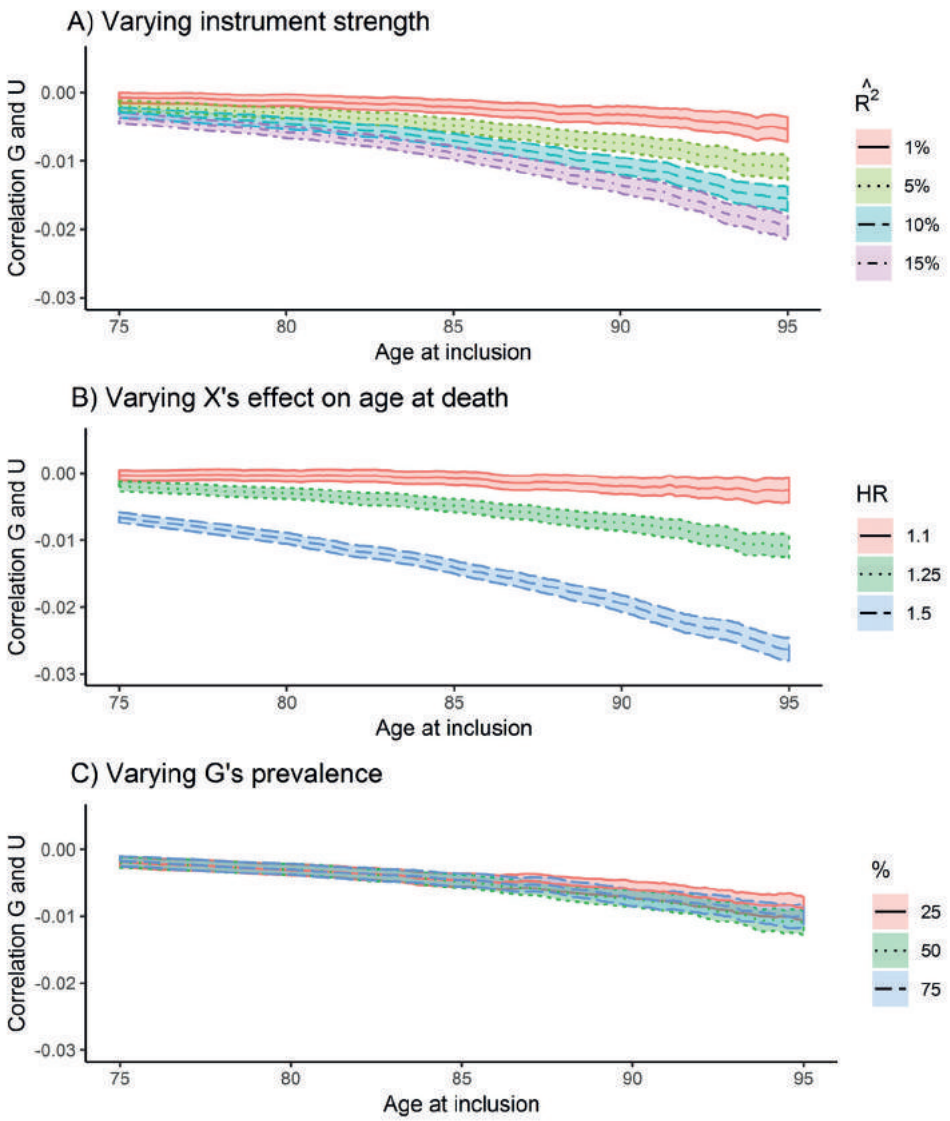
eFigure 2. Variance explained in the exposure of interest *X* by its genetic proxy *G* for an increasingly selected population, for a continuously distributed *R*. Shown in the insets are A) the prevalence of *G* and B) the accompanying survival curve, both with the true (i.e. unselected) R^2 set at 5%.



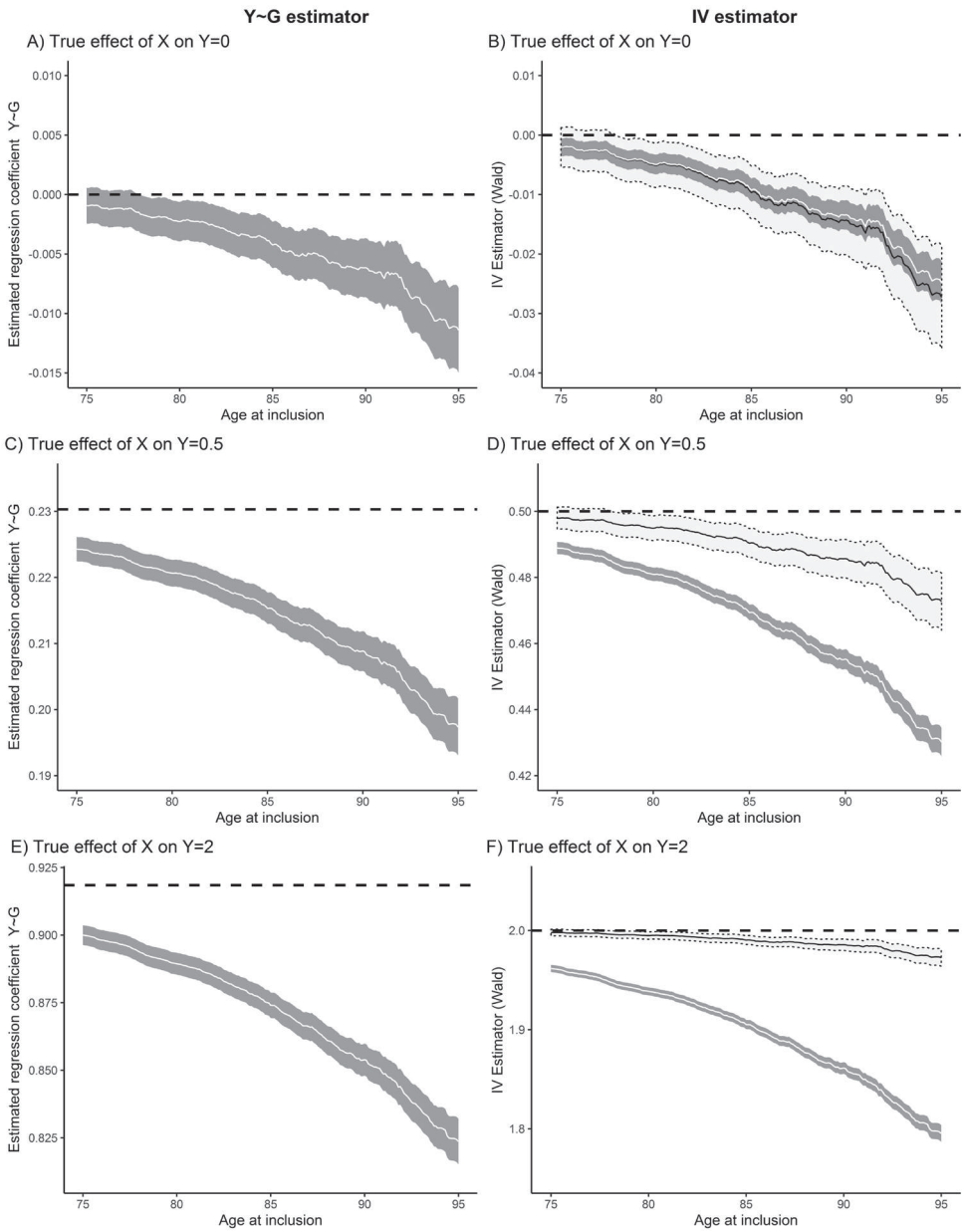
eFigure 3. Genetic instruments for exposures which affect the likelihood of surviving until study inclusion will become weaker if only due to becoming increasingly associated with the random noise in the exposure.



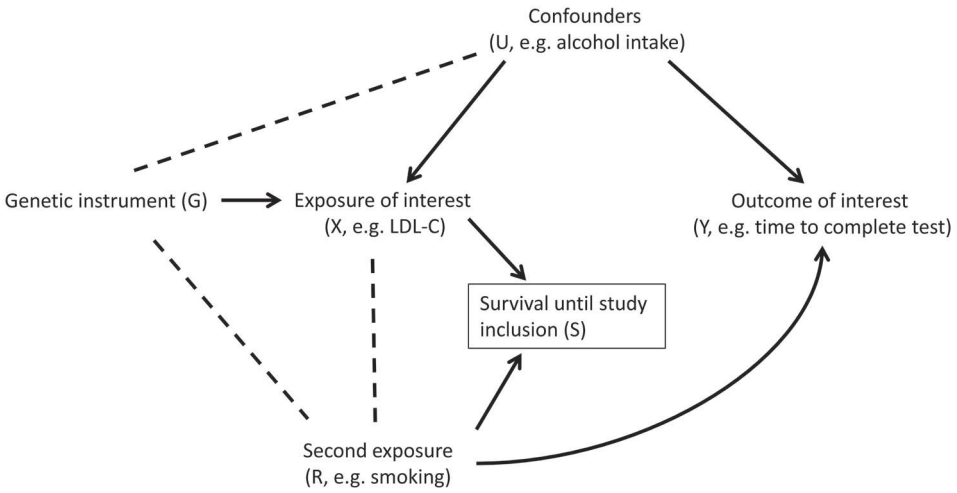
eFigure 4. Causal structure in which selection bias occurs in the presence of confounder U. Conditioning on survival S induces an assocation between genetic instrument G and confounder U. Also presented in Figure 2B.



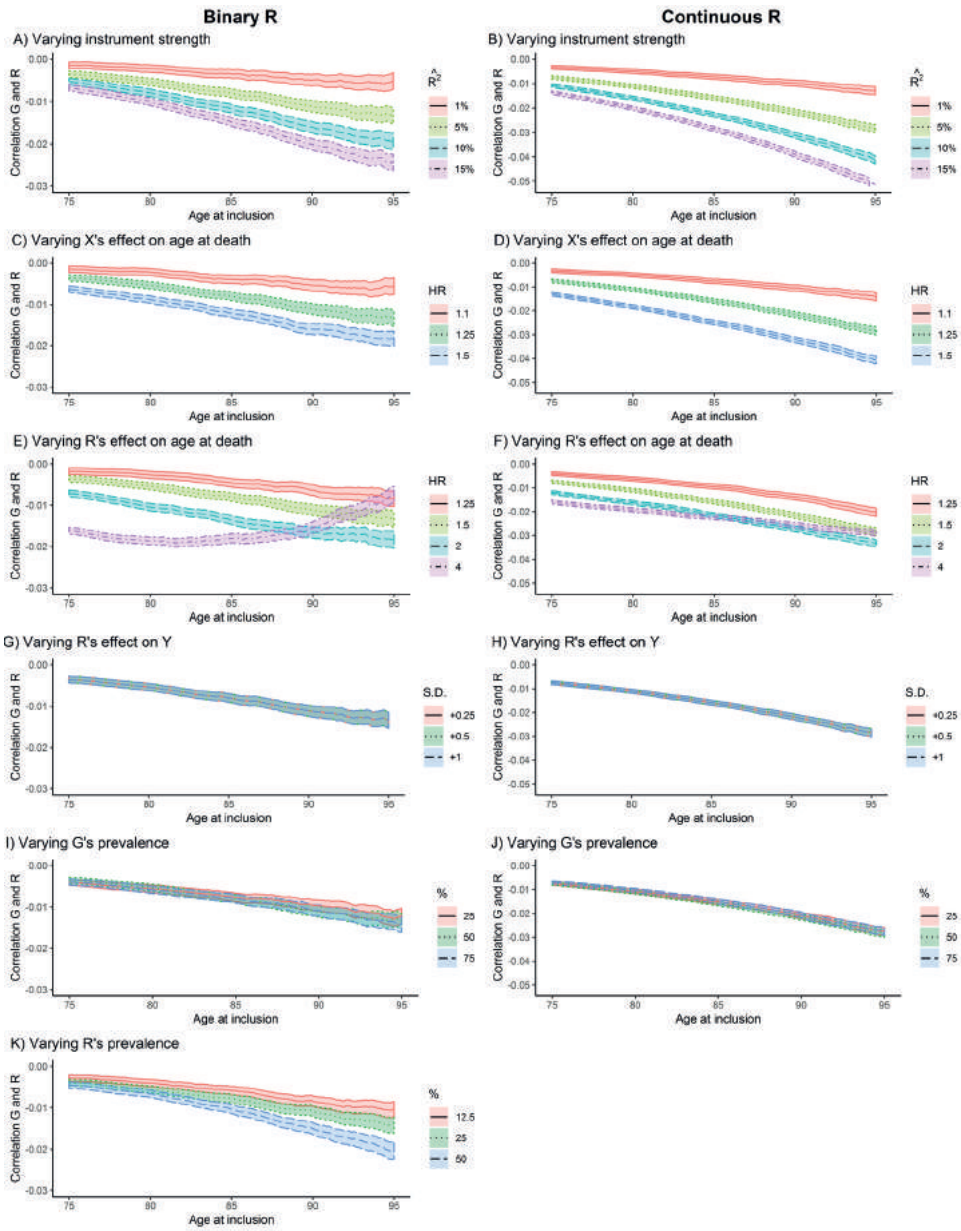
eFigure 5. Effect of varying different parameters on the induced correlation (95% CI) between the genetic instrument *G* and the confounder *U* for an increasingly selected population.



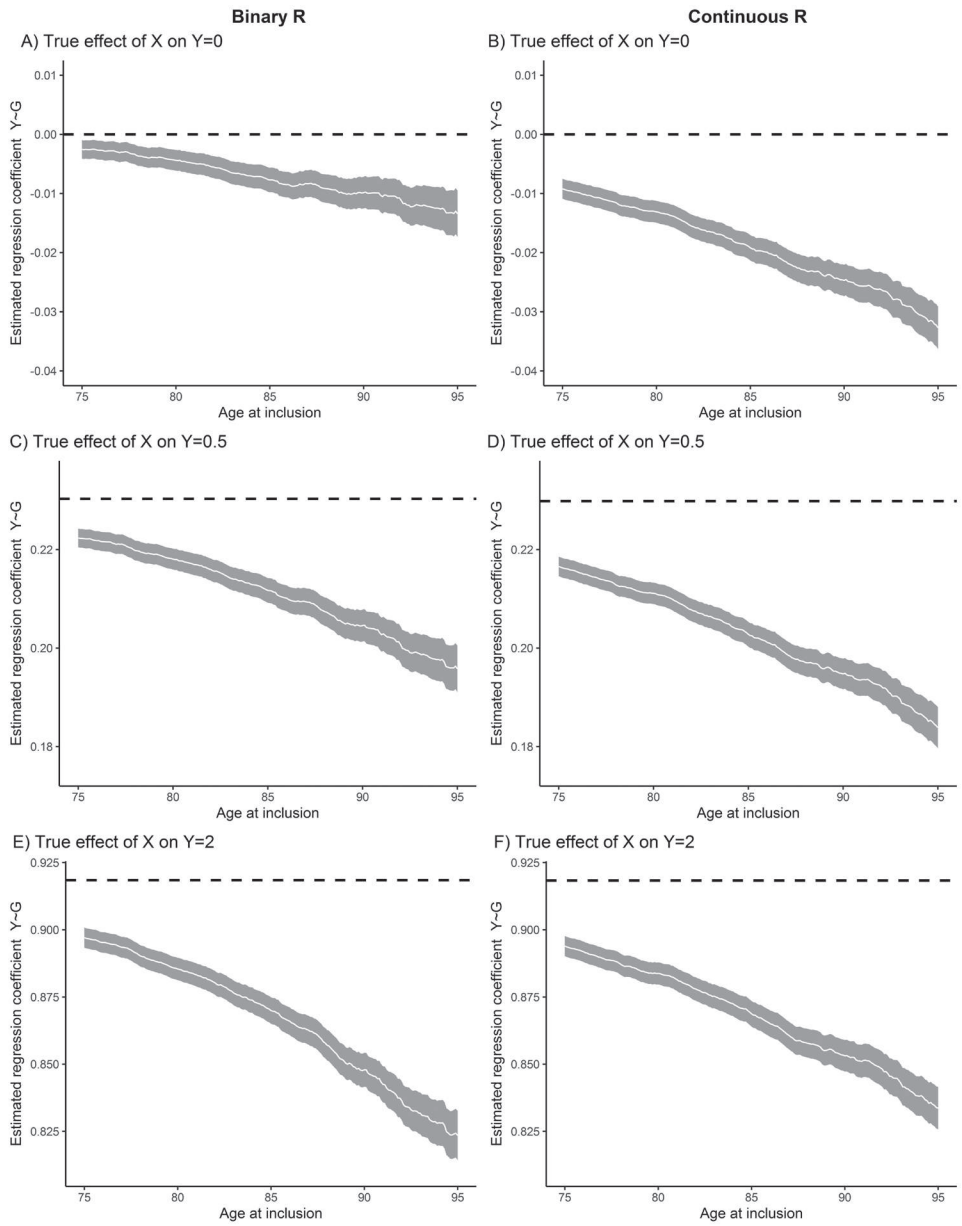
eFigure 6. Effect of survival bias on the association between the genetic instrument G and the outcome of interest *Y* (left panels), and on the Wald ratio IV-estimator (right panels), for different true effects of exposure *X* on *Y*. The true (i.e. unselected) regression coefficient for *G* on *Y*, and of true (i.e. unselected) Wald ratio, are shown as a dashed line in each plot.



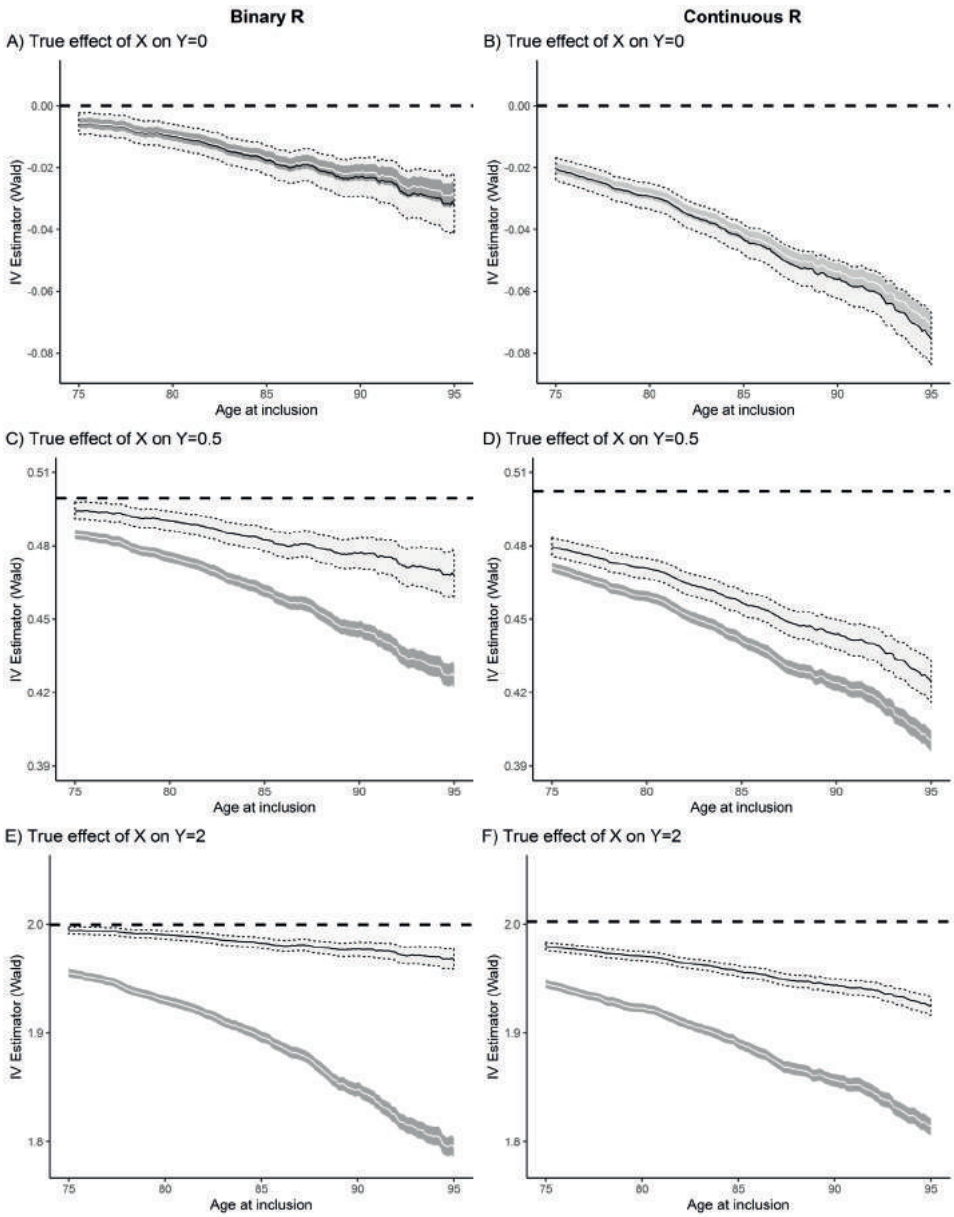
eFigure 7. Causal structure in which selection bias occurs in the context of both a second exposure R and confounder U.



eFigure 8. Effect of varying different parameters on the induced correlation (95% CI) between the genetic instrument *G* and the second exposure *R* for an increasingly selected population. Shown for binary (left column) and continuously (right column) distributed *R*. S.D. denotes standard deviation.



eFigure 9. Effect of survival bias on the association between the genetic instrument G and the outcome of interest *Y*, for different true effects of exposure *X* on *Y*. Data are presented as regression coefficients (95% CI) estimated with linear regression. The true (i.e. unselected) regression coefficient for *G* on *Y* is shown as a dashed line in each plot. Shown for binary (left column) and continuously (right column) distributed *R*.



eFigure 10. Wald ratios (95% CI) based on internally (white ribbon) versus externally (grey ribbon) estimated association between exposure *X* and the outcome *Y*, for different true effects of exposure *X* on *Y*. Shown for binary (left column) and continuously (right column) distributed *R*. Dashed lines denote the true (i.e. unselected) Wald ratio, which equals the true causal effect of *X* on *Y*.

7