PLOS GENETICS

Strategies to investigate and mitigate collider bias in genetic and Mendelian randomisation studies of disease progression

Ruth E. Mitchell[1,2]* , April E. Hartley[1,2], Venexia M. Walker[1,2,3],
Apostolos Gkatzionis[1,2], James Yarmolinsky[1,2], Joshua A. Bell[1,2],
Amanda H. W. Chong[1,2], Lavinia Paternoster[1,2], Kate Tilling[1,2‡], George Davey Smith[1,2‡]

1 MRC Integrative Epidemiology Unit at the University of Bristol, Bristol, United Kingdom, **2** Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom, **3** Department of Surgery, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, United States of America

☯ These authors contributed equally to this work.
‡ These authors are joint senior authors on this work.
* r.mitchell@bristol.ac.uk

Abstract

Genetic studies of disease progression can be used to identify factors that may influence survival or prognosis, which may differ from factors that influence on disease susceptibility. Studies of disease progression feed directly into therapeutics for disease, whereas studies of incidence inform prevention strategies. However, studies of disease progression are known to be affected by collider (also known as "index event") bias since the disease progression phenotype can only be observed for individuals who have the disease. This applies equally to observational and genetic studies, including genome-wide association studies and Mendelian randomisation (MR) analyses. In this paper, our aim is to review several statistical methods that can be used to detect and adjust for index event bias in studies of disease progression, and how they apply to genetic and MR studies using both individual- and summary-level data. Methods to detect the presence of index event bias include the use of negative controls, a comparison of associations between risk factors for incidence in individuals with and without the disease, and an inspection of Miami plots. Methods to adjust for the bias include inverse probability weighting (with individual-level data), or Slope-Hunter and Dudbridge et al.'s index event bias adjustment (when only summary-level data are available). We also outline two approaches for sensitivity analysis. We then illustrate how three methods to minimise bias can be used in practice with two applied examples. Our first example investigates the effects of blood lipid traits on mortality from coronary heart disease, while our second example investigates genetic associations with breast cancer mortality.

Introduction

There is a growing interest in performing genetic studies of disease progression, with initial studies suggesting that single nucleotide polymorphisms (SNPs) associated with disease
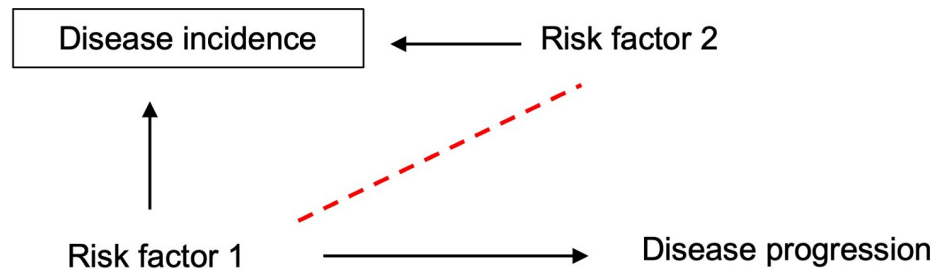
Fig 1. Directed acyclic graph demonstrating the introduction of collider bias in observational case only studies. Conditioning on disease incidence induces the association between previously independent causal risk factor 1 and causal risk factor 2, shown by the dashed line. Because risk factor 1 is also a causal risk factor for disease progression, a case-only setting has led to a biased association between risk factor 2 and disease progression via the path RF1->RF2->DP. The association of the risk factor 2 with disease progression when conditioning on incidence is entirely due to collider bias.

survival often differ from those associated with disease susceptibility [1–7]. "Disease progression," also known as disease prognosis, refers to any event occurring subsequent to disease incidence, such as changes in severity and/or survival. Investigating such events necessitates performing studies restricted to individuals who have the disease of interest, i.e., cases. By design, this involves conditioning on disease incidence, causing it to become a so-called "collider" variable within a causal inference framework [8]. This leads to biased associations between causal risk factors for disease incidence, including inducing associations between risk factors that are truly independent of each other (not correlated) in the source population. This becomes problematic if any of these risk factors for disease incidence, measured or unmeasured, also cause disease progression, because indirect associations may be induced between risk factors for disease incidence and disease progression (red dashed line in **Fig 1**). Therefore, a risk factor that is causal only for incidence (risk factor 2 in **Fig 1**) may falsely appear to cause progression entirely through an induced association with another causal risk factor for incidence (i.e., a noncausal path) (risk factor 1 in **Fig 1**). This can result in biased estimates of the true causal associations between risk factors and disease progression [8,9]; this bias has been termed index event bias (defined in **Box 1**). An example of index event bias is in studies of

Box 1. Terminology commonly used in relation to the bias induced in the case-only setting

Collider bias: Bias induced in the association between two variables when conditioning on their common effect (a "collider").

Selection bias: Bias in the estimated effect of exposure on outcome caused by nonrandom participation in/selection into a study. Collider bias will induce associations between all causes of participation in/selection into a study.

Index event bias: Bias in the estimated effect of exposure on outcome caused by restricting the analysis to cases only. Collider bias will induce associations between all causes of the disease.

Survival bias: Bias in the estimated effect of exposure on outcome caused by conditioning on those who have survived long enough to be in the study. Collider bias will induce associations between all causes of survival.

coronary heart disease (CHD) progression where the restriction of analyses to CHD cases only (i.e., conditioning on disease state) could induce associations between truly independent CHD risk factors. This could explain the so-called "obesity paradox" where higher body mass index (BMI) is associated with longer survival among those with CHD, despite higher BMI being associated with shorter survival in the general population. Indeed, lower levels of other risk factors for CHD measured in individuals with high BMI may be sufficient to induce an association of higher BMI with longer survival [10–12].

The model-dependent nature of the presence and direction of index event bias should be noted [13,14]. Where two independent risk factors are causes of a binary collider variable C, collider (index event) bias will not be induced by conditioning on C if the two risk factors are perfectly multiplicative/log additive in their effect on C on a risk ratio scale [13]. In case-only studies, disease incidence plays the role of the collider C. Different variables may be viewed as colliders in other types of studies; for example, studies affected by survival bias are effectively conditioned on individuals surviving to study onset and collider (index event) bias in such studies is avoided when the two risk factors are multiplicative in their effects on survival [15]. Moreover, collider (index event) bias is expected to induce a positive correlation between the two risk factors if they are supramultiplicative in their effects on disease incidence, and a negative correlation if the two risk factors are submultiplicative in their effects on disease incidence [15]. The extent of the resultant collider (index event) bias will therefore be greater the further away the associations of risk factors with the collider are from the multiplicative/log additive risk model [13]. In chronic disease epidemiology, many causal risk factors may be expected to have a submultiplicative impact on the incidence of disease.

In the case of a genetic epidemiological study of disease progression, index event bias is potentially problematic when a genetic variant causes the onset/incidence of disease, in the presence of a measured/unmeasured common cause (i.e., confounder) for disease incidence and progression. This situation creates spurious and/or biased associations between that genetic variant and the progression phenotype [16,17]. **Fig 2A** illustrates this: In case-only studies, when conditioning on disease incidence, and when there is a shared confounder for incidence and for progression (risk factor 1) in the population, any genetic variant (risk factor 2) that causes incidence will display an induced association with that confounder (risk factor 1). Collider bias in this context has opened up the pathway of genetic variant -> risk factor 1 -> disease progression, and the genetic variant will falsely appear to be associated with progression. Importantly, this confounder could be another genetic variant itself, and, therefore, in a genome-wide association study (GWAS) of case-only samples, more SNPs can appear to be associated with progression than truly are (**Fig 2B**). In another scenario, this spurious association through a noncausal pathway could be in addition to the direct true effect of the SNP on progression, inducing a biased association between the SNP and disease progression, i.e., an overestimate or underestimate of the true causal association (**Fig 2C**). Indeed, a study investigating the association of known common type 2 diabetes variants with BMI (a strong risk factor for type 2 diabetes) found three overestimated and one underestimated associations among 11 type 2 diabetes risk alleles when comparing to a nondiabetic population [16]. Another example uses a polygenic risk score to examine associations between CHD genetic risk variants and cardiovascular outcomes and found that these differ when examined in those with and without prior CHD [18]. These studies highlight the need to address this bias by detecting and accounting for its presence in case-only studies.

Index event bias also has implications for applied genetic epidemiological analyses downstream of GWAS, such as Mendelian randomisation (MR) [19–21]. A consequence of not adjusting for index event bias at the stage of conducting a GWAS would mean that biased association estimates of SNPs with disease progression could be used in MR analyses and result in
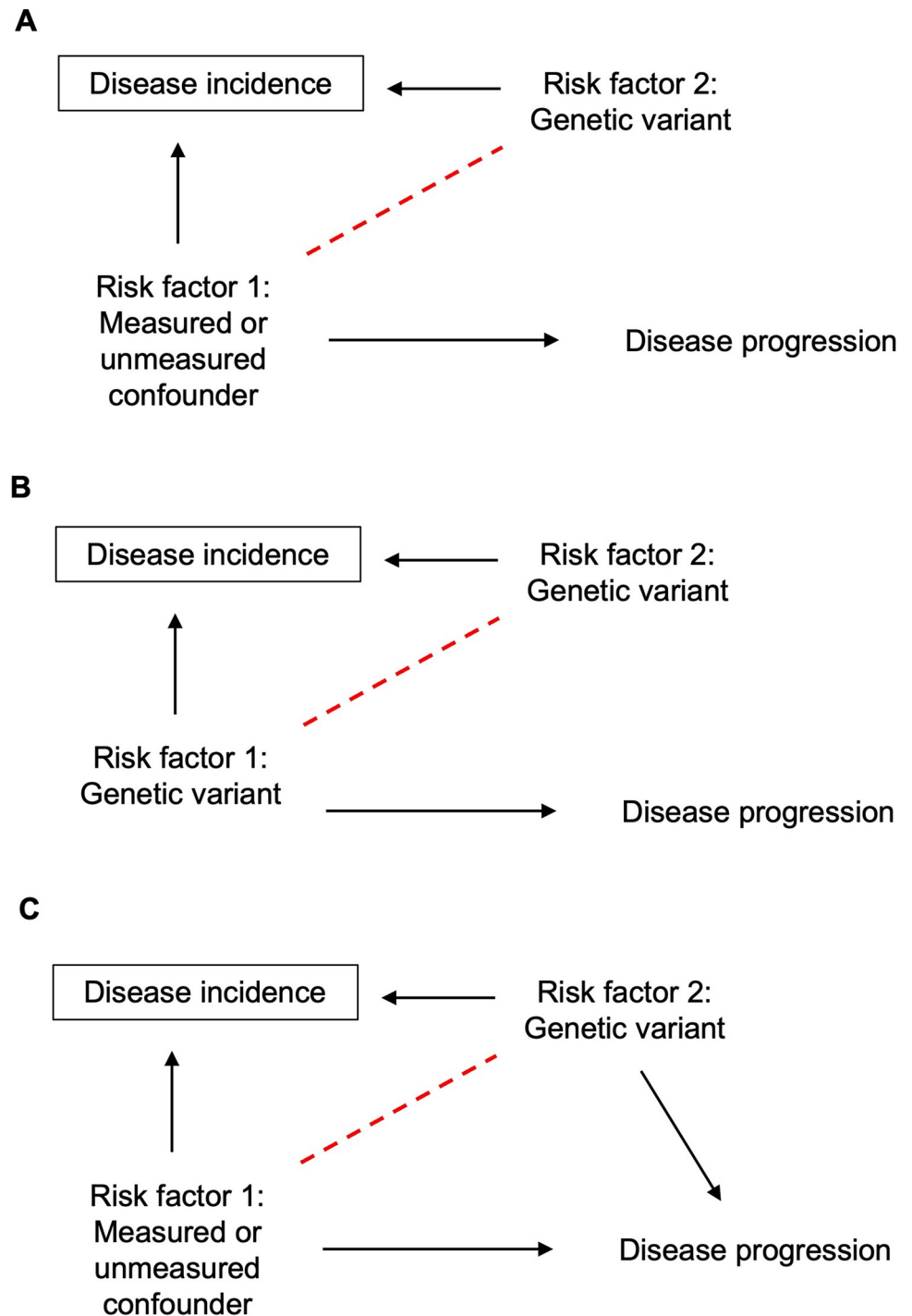
A

B

C

Fig 2. Directed acyclic graph demonstrating the introduction of collider bias in genetic case only studies. (**A**) Conditioning on disease incidence induces the association between a previously independent casual risk factor and causal genetic variant for disease incidence, shown by the dashed line. Because risk factor 1 is also a casual risk factor for disease progression (a confounder of disease incidence and progression), a case-only setting has led to a biased association between the genetic variant and disease progression via the path Genetic variant->Measured/unmeasured confounder->Disease progression. The association of the genetic variant with disease progression when conditioning on incidence is entirely due to collider bias. (**B**) Collider bias will induce an association between genetic variants that both cause disease incidence. This will make a noncausal genetic variant (risk factor 2) to appear associated with disease progression. (**C**) A third scenario is where this induced path is in addition to the direct effect of the genetic variant on disease progression.
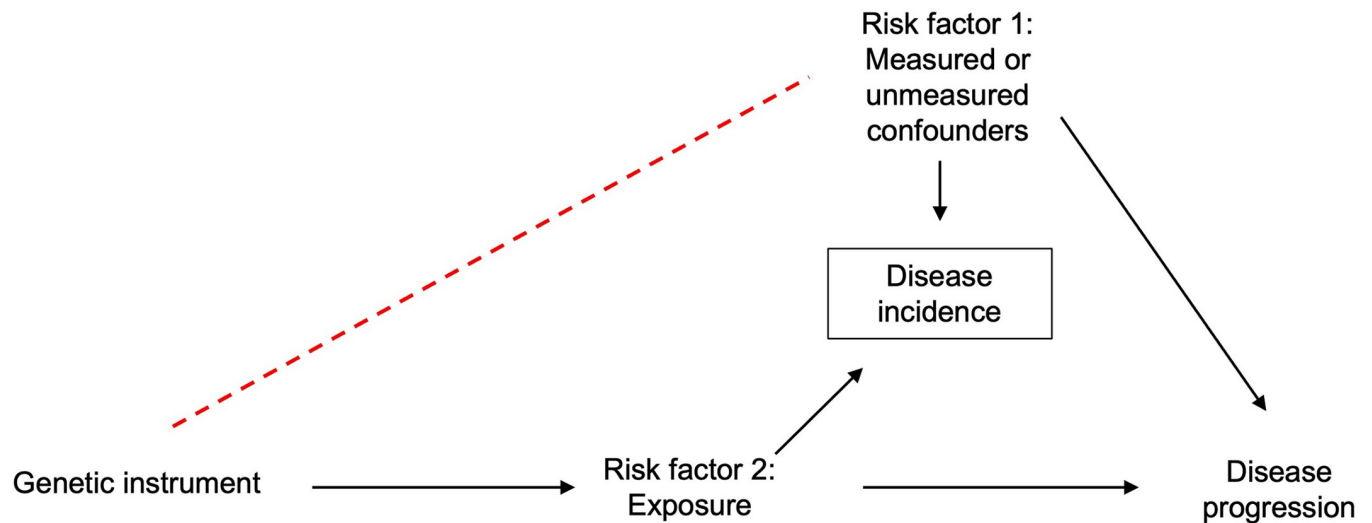
Fig 3. Directed acyclic graph demonstrating the introduction of collider bias in Mendelian randomisation case only studies. In Mendelian randomisation analyses, the exposure is proxied by a causal genetic instrument. Conditioning on disease incidence induces the association between the previously independent genetic instrument and a common cause for disease incidence and disease progression, shown by the dashed line. This would violate the independence MR assumption invalidating the analysis.

https://doi.org/10.1371/journal.pgen.1010596.g003

potentially misleading causal estimates of exposures with disease progression outcomes. In a two-sample MR setting, only the SNP-outcome (disease progression) estimates will be affected by index event bias as the SNP-exposure estimates will be taken from a GWAS that is not restricted to cases only. However, in a one-sample MR setting of a study of case-only samples, if the exposure causes disease incidence, then both the SNP-exposure and the SNP-outcome (disease progression) estimates may be biased. In addition, the MR assumption that the genetic instrument is independent of factors that confound the association of the exposure with the outcome would be violated given that conditioning on disease incidence has opened up the pathway of genetic instrument -> risk factor 1 -> disease progression (**Fig 3**). This would be true for a single genetic variant as well as a combination of variants within a polygenic risk score (PRS) instrument; the use of such scores may increase the potential for this bias. This would invalidate the MR study and lead to an over- or underestimate of the causal effect of exposures on the disease progression outcome of interest [21].

Here, we aim to review several strategies that are currently available to investigate and mitigate index event bias in GWAS and MR studies of disease progression and advise on the interpretation of results from such studies. Although other sources of selection bias can be an issue when studying disease progression, including loss-to-follow-up and missing data [22,23], this review focuses on index event bias. We start with the need to investigate if there is bias in the case-only population. Where index event bias is detected, we discuss three methods that aim to minimise index event bias, according to the data that are available (individual-level or summary-level). We next outline two sensitivity analyses that have been developed to determine the magnitude of bias that would have to be present to explain any observed associations with progression. We conclude with two applied examples of disease progression studies—one concerning blood lipid traits and survival in CHD, and the other concerning breast cancer prognosis.

Detecting index event bias

Index event bias can be investigated using negative controls that are causal for disease onset. For example, age is not genetically determined, and, therefore, the presence of strong
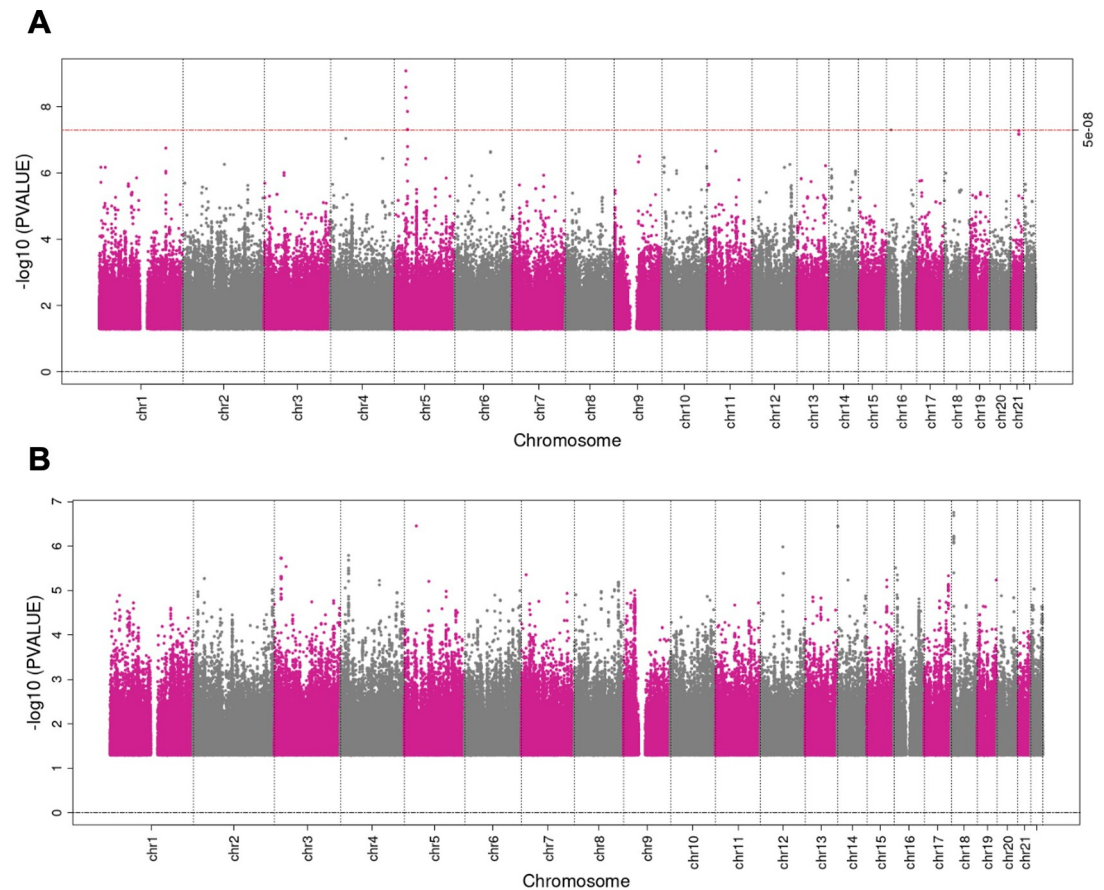
A

B

Fig 4. (A) Manhattan plot of GWAS for age at recruitment in myocardial infarction (MI) incidence cases only in UK Biobank. Cases were defined as individuals who had had an acute MI event using the International Classification of Diseases 10th Revision codes (ICD-10: I21.0-I21.9). GWAS was performed using Plink. This plot illustrates one genetic signal on chromosome 5 that is shown to be strongly associated with age at recruitment ($P < 5 \times 10^{-8}$). This signal could potentially be induced due to collider bias as, in a general random population a GWAS for age should not show any signal. However, this signal could also be due to biases other than collider bias. **(B) Manhattan plot of GWAS for sex in MI incidence cases only in UK Biobank**. Cases were defined as individuals who had had an acute MI event using the ICD10 codes I21.0-I21.9. This plot does not show any strong signal associated with sex, suggesting that no evidence of collider bias is detected.

https://doi.org/10.1371/journal.pgen.1010596.g004

associations between SNPs and age in a case-only sample, where age is a risk factor for being a case, can only be an artefact of index event bias. Similarly, the presence of associations of autosomal SNPs with sex, reflecting differences in allele frequencies between men and women, would be evidence for index event bias if onset of the disease of interest differed by sex. Identification of sex-associated autosomal loci has highlighted potential bias due to sex differences in participation in large cohort studies [24]. It should be noted that these analyses do rely on a large enough sample size so that the analyses have sufficient statistical power. If analyses are underpowered, one cannot be sure that a lack of association, e.g., between SNPs and age, in the case-only population is due to underpowered analyses or the true absence of index event bias. Therefore, power calculations should be performed prior to these analyses. **Fig 4** is an illustration investigating the presence of collider bias in 11,085 myocardial infarction cases in UK Biobank. The signal seen in chromosome 5 associated with age suggests that index event bias may have been induced in this sample (**Fig 4A**). One independent intronic SNP was identified from this signal, rs535799110, located within *C1QTNF3*. This SNP, and SNPs in close linkage disequilibrium, did not show an association with myocardial infarction or age in the

full UK Biobank dataset. However, expression of C1QTNF3 has been shown be associated with traits such as erythrocyte measures, BMI-adjusted waist-up ratio, and kidney function parameters. Equally, known causal risk factors for disease incidence can be used as negative controls, either as an exposure in a GWAS in the case-only population or in a hypothesis-driven manner examining the association of the risk factor with genetic variants associated with disease incidence. When known, genetic variants strongly associated with disease incidence could also be used as negative controls. For example, in a study involving cases of dementia, a GWAS of the ApoE genotype could be performed. In addition, known risk factors for disease initiation can be used as a diagnostic for index event bias. These can be either phenotypic risk factors, such as smoking or BMI, or genetic variants strongly associated with disease incidence, or with causes of disease incidence. Index event bias will alter or induce spurious associations between such risk factors in the case-only population, and these associations can be compared to those from an independent dataset, not restricted to cases. Differences between associations in the cases and the independent dataset will be suggestive of index event bias. Datasets with large sample sizes and deep phenotype data on both cases and non-cases provide external data to explore how divergent the association between risk factors is from the multiplicative/log additive model, although it is worth being aware of the effect of confounding or measurement error in these exploratory analysis. This can indicate the likely quantitative effect of collider (index event) bias. When using this diagnostic, it is worth being aware of the recruitment process in each dataset, as this will change the risk factors chosen to test, e.g., there will be differences in why patients were recruited to a trial versus if they are in a general population-based longitudinal study or sampled through hospital data.

Without access to individual-level data, but with the full set of results from a GWAS of both incidence and progression of the disease of interest, index event bias can still be examined by comparing the magnitude of the effect of a SNP on disease progression with the magnitude of the effect of that same SNP on disease incidence. If there is strong evidence for an association of a SNP with disease incidence, then we cannot rule out the possibility that the association of that SNP with progression is purely an artefact of selection bias (**Fig 2B**) or that the magnitude of association is biased by selection (**Fig 2C**). Associations with progression for SNPs not associated with disease incidence will not suffer collider bias. Miami plots can be generated to visually inspect and compare SNP associations for disease incidence and disease progression on a genome-wide scale. These plots are an extension of a Manhattan plot, where p-values are plotted on the $-\log_{10}$ scale. The Miami plot will present the p-values for incidence on the $-\log_{10}$ scale and the p-values for progression on a $\log_{10}$ scale, for all available SNPs. These can be produced using publicly available code within the software EasyStrata [25]. An example is shown in **Fig 5** plotting the GWAS results of smoking initiation (top) and smoking cessation in a population of smokers (bottom). As well as comparing across the genome for a GWAS, this methodology can also identify potential index event bias in an MR analysis, with comparisons restricted to the instrument(s) for the exposure of interest. As described for GWAS, a lack of evidence for an association between the instrument for the hypothesised exposure and disease incidence is evidence against the presence of index event bias (i.e., the exposure may be specific for disease prognosis), whereas if the instrument is also related to disease incidence we cannot be sure that any relationship with the outcome is not an artefact of index event bias.

Even if these methods do not identify any evidence for index event bias in the case-only population, the next step would be to perform the sensitivity analyses detailed below, to determine the magnitude of bias that would need to be present to explain the observed associations. If there is evidence for index event bias, the subsequent section reviews methods that can be applied to attempt to overcome this index event bias.
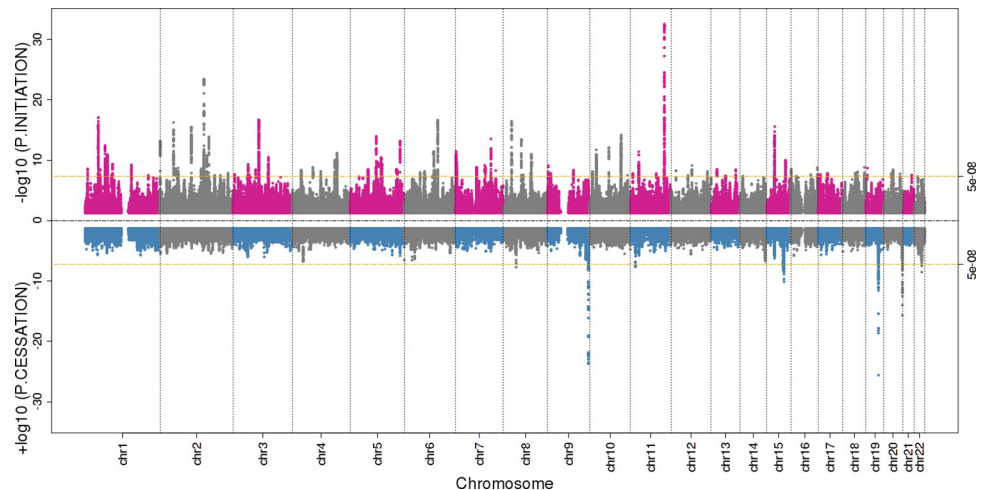
Fig 5. An example of a Miami plot comparing results from a GWAS of smoking initiation (top) and a GWAS of smoking cessation (bottom) in a population of smokers. Plotted using publicly available summary statistics of Liu et al. [60]. There are several loci strongly associated with smoking cessation where there is no strong evidence for an association with smoking initiation (e.g., chr11, 19), suggesting that the association between these loci and smoking cessation is not the product of collider bias. However, further inspection of the magnitude of effect and confidence intervals is required to determine that these loci are not associated with initiation. The locus on chromosome 20 reaching genome-wide significance also appears to be associated with smoking initiation, albeit not at genome-wide significance, suggesting that the association of this locus with smoking cessation may be affected by collider bias.

https://doi.org/10.1371/journal.pgen.1010596.g005

Sensitivity analyses to determine magnitude of bias

Here, we present two sensitivity analyses that can be used to examine the magnitude of bias.

Smith and VanderWeele method

A sensitivity analysis for index event bias was proposed by Smith and VanderWeele [26]. This approach is not specific to genetically informed studies of disease progression but can be used in any epidemiological study where selection bias is suspected, provided that the outcome is binary and the causal parameter of interest is a risk ratio, odds ratio, or risk difference. Smith and VanderWeele derive bounds for the true risk ratio (or odds ratio or risk difference) of the effect of an exposure on an outcome, which can be computed using the observed risk ratio and an additional variable U representing potentially unobserved confounders (or mediators) between the outcome and selection into the study. The variable U must be such that the outcome becomes independent of selection conditional on the exposure and U. In the context of case-only studies, U represents common causes of disease incidence (the selection variable) and progression (the outcome), since conditioning on all such causes and on the exposure will render disease incidence and progression independent. In other words, U represents the "measured or unmeasured confounder" variables in Fig 2. The effects of U on incidence and progression need to be specified to compute the selection bias bounds, but otherwise, the method makes no parametric modelling assumptions. As an additional diagnostic, Smith and Vander-Weele describe how to compute E-values for selection bias. In case-only studies, an E-value quantifies how strong the effects of U on incidence and progression should be for the target risk ratio to take the observed value, if the true effect of the exposure on progression is null. An online calculator to compute risk ratio bounds and E-values is available at http://selection-bias.louisahsmith.com. The use of risk ratio bounds has been advocated in case–control studies with biased selection of controls [27], and the method was recently extended to account for confounding bias and measurement error, in addition to index event bias [28].

Quantitative bias analysis

As with the E-value approach, one form of quantitative bias analysis attempts to quantify the magnitude of bias needed for the observed MR estimates to occur if the association was truly null, using simulations. For the simulations, individual-level data are generated for cases and non-cases, based on user-specified assumptions about the factors associated with disease incidence and prognosis [29]. Note that, as these data are simulated, this approach can be applied to one- or two-sample MR. By repeating this process for multiple simulated samples, investigators can obtain a distribution of estimates that are solely due to the effects of index event bias. This information, presented alongside the main estimates, allows those appraising a study to assess whether the MR estimates represent a plausible association or are more likely a consequence of bias. This has previously been demonstrated in the literature by Noyce et al. in their study of the relationship between BMI and risk of Parkinson's disease, where it is referred to as "frailty modelling" as the bias in their example was thought to be caused by survival effects [29].

Accounting for index event bias

In order to conduct down-stream analysis that correctly accounts for index event bias such as an MR analysis, it is necessary to adjust the beta estimates prior to performing the MR analysis. In this section, we review methods to adjust for index event bias when individual-level data are available, and then explore methods for when either summary-level or individual-level data are available. A summary of methods described in this section is presented in **Table 1**.

Inverse probability weighting (IPW)

IPW can help to address index event bias in case-only studies through the creation of a pseudo-sample where individuals are weighted according to the probability of having the disease of interest [30]. The weighted pseudo-sample aims to mimic a situation where every individual has the same probability of contracting the disease; therefore, the distribution of sociodemographic and behavioural factors in the weighted sample will be similar to that in the overall population. Consequently, IPW will down-weight overrepresented individuals (i.e., those most likely to have the disease) and up-weight underrepresented individuals (i.e., those least likely to have the disease) [31,32]. The probability that an individual is included in the case-only sample is estimated by fitting a statistical model (e.g., logistic regression) for disease incidence. Individuals in the case-only sample are then weighted by the inverse of their estimated probability of disease. To estimate the model used to calculate the probability weightings, at least some information about non-cases must be known. IPW can only truly overcome index event bias when the incidence model is adjusted for enough covariates to render incidence independent of the exposure and progression conditional on these covariates. This means that, ideally, all true causes of incidence and progression should be known and measured within the target population and should be included in the weighting model. Likewise, causes of incidence that are also indirectly related to progression should be included in the weighting model to account for any backdoor paths between incidence and progression and prevent M-bias. Finally, nonlinearities and interactions in the model for incidence should be included. Thus, IPW involves the assumption of no unmeasured confounding of incidence and progression—an assumption that cannot be verified using the observed data.

Even when the weighting model is correctly specified, IPW estimates will have lower precision compared to unweighted estimates because weighting reduces the analysis' effective sample size [33,34]. This can be particularly concerning in applications where a small number of individuals are assigned large weights and hence have a disproportionately large influence in the analysis. In this case, techniques such as weight stabilisation and weight trimming can be used [32,35,36].

Table 1. Summary of methods for detecting and accounting for index event bias and sensitivity analyses to determine the magnitude of bias required to negate observed effect estimates.

Method	Data required	Theory	Strengths	Limitations	Reference
Detecting index event bias					
Negative control GWAS (e.g., age, sex)	Individual-level phenotype and genotype data for case-only population	Associations of SNPs with age and sex likely to reflect index event bias if age and/or sex cause the disease	• Straightforward GWAS analysis with no knowledge of risk factors for the disease required • Uses variables available in all datasets	• Requires access to individual-level data • Requires sufficient sample size to assume that lack of observed associations is due to absence of bias instead of insufficient power	
Determination of associations between risk factors for incidence and their instruments	Individual-level phenotype and genetic data for case-only and comparator (unselected) population	Presence of associations between risk factors (and/or their instruments) of a different magnitude/direction of effect to an unselected sample likely to reflect index event bias	• Straightforward regression analyses • Does not necessarily require access to genotype data (although PRS are beneficial)	• Requires access to individual-level data • Case-only populations are likely to be smaller than unselected populations, and, therefore, differences in magnitude of relationships could reflect lower statistical power in the case-only population • Requires a comparator population, which is not subject to selection bias	
Comparison of incidence and prognosis GWAS results	Summary-level data from a GWAS of disease incidence and a GWAS of disease prognosis	SNPs associated with incidence may have biased associations with prognosis, but associations of SNPs with prognosis only are not biased by index event bias	• Requires summary-level data, which are often publicly available • Easy to visually inspect using a Miami plot	• Identifies SNPs that may be biased by index event bias but does not identify if index event bias is present	
Accounting for index event bias					
Inverse probability weighting	Individual-level data for all known disease incidence risk factors for all individuals in target population (not just disease cases)	Unbiased estimates (e.g., for a SNP effect on prognosis) can be estimated from a regression model weighted by the inverse probability of an individual being a case	• Can be applied to GWAS and one-sample MR	• Requires data for non-cases • Requires knowledge of all known risk factors for disease incidence • Requires individual-level data • May suffer from a loss of precision if some individuals are assigned large weights	[30]
Dudbridge et al method	Summary-level data for a GWAS of disease incidence and a GWAS of disease prognosis	The bias correction factor can be estimated as the slope of the regression line of SNP-prognosis on SNP-incidence associations using all independent SNPs	• Generates a bias-correction factor that can be applied to all SNPs • Only requires summary-level data • Incidence and prognosis GWAS can be performed on overlapping or independent samples	• Assumes no shared pathways between disease incidence and disease prognosis • Assumes linear effects of SNPs on incidence and prognosis, with no interactions • Inclusion of a large number of SNPs could result in underestimation of the bias correction factor • Assumes constant confounding across SNPs • Affected by allele coding	[12]
Slope-Hunter	Summary-level data for a GWAS of disease incidence and a GWAS of disease prognosis	The bias correction factor can be estimated as the slope of the regression line of SNP-prognosis on SNP-incidence associations using independent SNPs associated with incidence only	• Generates a bias-correction factor that can be applied to all SNPs • Only requires summary-level data • Incidence and prognosis GWAS can be performed on overlapping or independent samples • Doesn't rely on assumption of no shared biological pathways contributing to incidence and prognosis	• Assumes that the variance in incidence explained by SNPs associated with incidence only is at least as great as the variance explained by SNPs associated with incidence and prognosis • Assumes linear effects of SNPs on incidence and prognosis, with no interactions • Assumes constant confounding across SNPs	[41]

(Continued)

Table 1. (Continued)

Method	Data required	Theory	Strengths	Limitations	Reference
Corrected Weighted Least Squares	Summary-level data for a GWAS of disease incidence and a GWAS of disease prognosis	The bias correction factor can be estimated as the slope of a weighted regression of SNP-prognosis on SNP-incidence associations using all independent SNPs	• Generates a bias-correction factor that can be applied to all SNPs • Only requires summary-level data • Incidence and prognosis GWAS can be performed on overlapping or independent samples	• Inclusion of a large number of SNPs could result in underestimation of the bias correction factor • Assumes linear effects of SNPs on incidence and prognosis, with no interactions • Assumes constant confounding across SNPs	[40]
Traditional MR approaches	Summary-level data, for SNPs associated with incidence at genome-wide significance, from a GWAS of disease incidence and a GWAS of disease prognosis	The bias correction factor can be estimated as the effect estimate from traditional MR approaches such as inverse variance weighted or weighted median meta-analyses	• Generates a bias-correction factor that can be applied to all SNPs • Only requires summary-level data	• Slope estimate can be biased by sample overlap • Assumes linear effects of SNPs on incidence and prognosis, with no interactions	[40]
Determining magnitude of bias required to explain observed results					
Smith and Vanderweele method	No data required, but need to specify a factor U such that incidence and progression are rendered independent conditional on U and the exposure. The effects of U on incidence and progression must be elicited.	Computes bounds for the magnitude of selection bias using simple expressions that depend on the effects of U on incidence/progression. The bounds can then be used to assess how strong the selection effects should be for an observed association to be fully explained by selection bias.	• Easy to implement using existing software. • Does not require modelling assumptions about the selection mechanism (other than identifying U). • Works with both individual and summary-level data.	• Only works with a binary outcome. • Requires U to be identified and its effects on incidence and progression elicited, which can be difficult and/or subjective. • Does not provide a point estimate.	[26]
Quantitative bias analysis	No data are required, but it can be useful to inform the simulation.	By simulating individual-level data and an indicator of case status, investigators can obtain a distribution of estimates that are solely due to a given magnitude of index event bias. Comparison of these effects with those observed informs whether the observed effect is likely to be a consequence of this bias.	• No data are required to implement the method, so it can be applied regardless of whether you have performed your main analysis using individual- or summary-level data	• Simulated data may not reflect reality and so be misleading • No formal recommendations on how to compare the observed effect in your main analysis with those obtained from the simulation	[29]

GWAS, genome-wide association study; MR, Mendelian randomisation; SNP, single nucleotide polymorphism; PRS, polygenic risk score.

https://doi.org/10.1371/journal.pgen.1010596.t001

Once a weighted sample has been generated, analysis methods such as MR can then be applied. For example, two-stage least squares estimates can be computed using weighted linear regression instead of ordinary linear regression. IPW can adjust for index event bias in MR provided the probabilities of disease are accurately estimated [37,38]. In practice, IPW can be useful in case-only studies that are nested within a cohort study (e.g., studies utilising individual-level data from the UK Biobank). However, the reliance on individual-level data means that IPW often cannot be used in two-sample MR studies.

Dudbridge et al. method

The Dudbridge et al. method is based on the premise that the association between a SNP and progression is proportional to the true effect of the SNP on progression and a bias that is linear

in the effect of the SNP on incident disease [12]. In equation form this can be summarised as follows:

$$\beta_p^{est} = \beta_p^{true} + b\beta_i^{true}$$

where β_p^{est} is the effect estimate for a SNP from a GWAS of progression, β_i^{true} is the effect of the SNP on incidence, β_p^{true} is the true effect of the SNP on progression (the effect of interest), and b is the slope from a regression of β_p^{est} on β_i^{true} for all independent SNPs and is the bias correction factor. The true effect of the SNP on progression is therefore a combination of the intercept and residual from this regression line [12].

The method developed by Dudbridge et al. uses all SNPs available to determine the correction factor [12]. Linkage disequilibrium (LD) clumping, based on the *p*-value for the SNP effect on incidence, is required prior to analysis to restrict the regression to independent SNPs, although the correction factor can then be applied to all SNPs. As well as the assumption that the SNPs are independent, this method also assumes a linear effect of the SNP on both incidence and progression (with no interactions), that there is no correlation between SNP incidence and SNP-progression effects, and that the effect of common causes of incidence and progression (i.e., genetic and nongenetic confounding) is constant across all SNPs. The assumption of constant confounding across all SNPs may not be true if there is a genetic correlation between incidence and progression, as the genetic component of the unmeasured confounding will be weaker for SNPs that are strongly associated with both incidence and progression. The assumption of no correlation between SNP-incidence and SNP-progression effects is violated for diseases where the same biological pathways, at least in part, contribute to incidence and progression. This assumption may not be justified for the majority of disease traits. For example, these assumptions would not be met in cardiovascular disease; given that lowering LDL cholesterol reduces risk of major vascular events in both primary and secondary prevention trials [39], and thus for SNPs influencing LDL cholesterol, there will be a very strong positive correlation between their associations with vascular disease incidence and risk of secondary events. Although Dudbridge et al. did show that when there is a positive correlation between incidence and progression, type 1 error is lower than an unadjusted analysis [12], we advise caution when using this method.

An updated version of the Dudbridge et al. method was proposed by Cai et al. [40]. The original method is analogous to MR-Egger with inclusion of an intercept, but the intercept model is affected by allelic coding. When including all SNPs, regardless of their effect on incidence, the majority of SNPs will have very small effect estimates, making it difficult to code the alleles in a positive direction. Cai et al. therefore recommended a nonintercept model, which they called the Corrected Weighted Least Squares (CWLS) method. This method still assumes no genetic correlation between incidence and progression. The updated method had a lower type 1 error rate than the original Dudbridge et al. method, but still had a higher type 1 error rate than no adjustment in the presence of a strong negative genetic correlation between incidence and prognosis [40].

We therefore recommend that the CWLS method is used as a sensitivity analysis alongside other correction methods (see Slope-Hunter method), to determine robustness of SNP-progression effect estimates across these methods.

The original Dudbridge et al. method can be performed using an open-source R package (https://github.com/DudbridgeLab/indexevent) and requires full summary statistics from a GWAS of disease incidence and a GWAS of disease progression, which can be generated from

independent or overlapping samples. These summary statistics can be used to perform downstream analyses, such as two-sample MR.

Slope-Hunter method

The Slope-Hunter method developed by Mahmoud et al. [41] is again based on the premise that the association between a SNP and progression measured in a GWAS can be estimated from the true effect of the SNP on progression and bias linear to the effect of the SNP on incidence [12,41]. However, the Slope-Hunter method extends the algorithms generated by Dudbridge et al. and attempts to overcome the limitation of the strong assumption of no genetic correlation between incidence and progression [41]. Slope-Hunter aims to partition all independent SNPs affecting incidence into two categories using cluster-based methods:

1. SNPs only affecting incidence;

2. SNPs affecting both incidence and progression.

   The correction factor is estimated from category 1 SNPs only, assuming unmeasured confounding across these SNPs. The correction factor is estimated as the slope of the regression line of disease progression associations on disease incidence associations for this restricted set of SNPs. This correction factor can be applied to all SNPs. The Slope-Hunter method does not assume that disease incidence and progression are not genetically correlated, but does assume that the SNP effects on both incidence and progression are linear, with no interactions. This is often the case in a logistic model when per-allele effect sizes are small. An additional assumption of the Slope-Hunter method is that the variance in disease incidence explained by category 1 SNPs is at least as large as that explained by category 2 SNPs.

   Slope-Hunter can be performed using an open-source R package (https://github.com/Osmahmoud/SlopeHunter/) and, like the Dudbridge et al. method, requires full summary statistics from a GWAS of disease incidence and a GWAS of disease progression. Both methods are robust to sample overlap and therefore can be used with summary statistics derived from the same population as well as independent populations. Summary statistics from these methods are suitable for use in downstream analyses, such as two-sample MR.

Other summary-level data approaches

The Slope-Hunter, Dudbridge et al., and CWLS approaches have been developed specifically to identify index event bias from GWAS summary statistics. However, in their recent paper, Cai et al. discuss how traditional two-sample MR approaches (but applied to an MR of incidence on prognosis) can be used to estimate the bias correction factor b [40]. The fundamental difference between this approach and Slope-Hunter is in the SNPs used. While Slope-Hunter aims to identify and use only those that are associated (default $p < 0.001$) with incidence and not progression (valid instruments) [41], Cai et al. argue that including all independent SNPs is preferable to maintain precision and thus power in the progression discovery GWAS, but in doing so, many weak instruments are included [40]. We would recommend performing an MR of incidence on prognosis using these more typical genome-wide significant set of SNPs for incidence. This would reduce the impact of weak instrument bias, but we note that it would still include invalid SNPs that are associated with both incidence and progression. However, this approach would allow the performance of a range of different MR methods such as IVW, weighted median, etc., as further sensitivity analyses. Therefore, in addition to the methods described above, this could be used to check for consistency of methods in estimating the correction factor.

Applied examples of mitigating index event bias

Lipid traits and secondary prevention of CHD

We aimed to examine the existence and mitigation of index event bias in an applied MR study using individual-level data from the UK Biobank cohort (UKB; see S1 Text). We chose to estimate the effects of two well-known lipid traits, low-density lipoprotein cholesterol (LDL-C) and high-density lipoprotein cholesterol (HDL-C), on the risk of CHD mortality, and potential bias induced by selecting on individuals with a history of CHD (the index event). These exposures were chosen because strong evidence exists for both in relation to CHD mortality from randomised controlled trials (RCTs) of lipid-modifying drug therapies, which were conducted among people with a history of CHD, thus providing a valuable (likely unbiased) standard for comparison. This RCT evidence indicates that, among people with CHD history, LDL-C raises CHD mortality risk [42], whereas HDL-C does not alter CHD mortality risk [43]. Because LDL-C likely causes CHD onset (based on positive results from past MR studies and several primary prevention RCTs, in agreement with conventional observational studies; [42,44]), we expected that conditioning on CHD history would induce potential for index event bias for LDL-C estimates (Fig 6). In contrast, because HDL-C likely does not cause CHD onset (i.e., no effect on disease incidence based on null results from past MR studies, in disagreement with conventional observational studies; [45]), we expected that conditioning on CHD history would not induce potential for index event bias for HDL-C estimates. It is plausible, however, that HDL-C may influence CHD mortality indirectly via statin use, leading to an adverse effect of higher HDL-C (if higher HDL-C causes lower statin use, and lower statin use causes higher CHD mortality). Our hypothesised causal relationships between HDL-C, LDL-C, triglycerides, statins, and CHD onset and mortality are shown in Fig A in S1 Text. Using one-sample MR in UKB, we verified a positive effect of LDL-C, and a null effect of HDL-C, when adjusting for each other plus triglycerides (given expectations of pleiotropy) [46], on CHD onset (S1 Text).

Among UKB participants with a history of CHD, we examined the effects of lipids on risk of CHD mortality in an MR framework, using two-stage least squares predictor substitution regression models and genetic risk scores (GRSs) comprised of genetic variants for exposures sourced from external GWAS. In a first-stage linear regression model, the exposure, e.g., LDL-C, is regressed on a GRS for LDL-C, plus age, sex, and the first ten genetic principal components (GPCs). The predicted values from that model were then entered into a logistic regression model as an exposure (plus age, sex, GPCs) with CHD mortality as the outcome. Standard errors were bootstrapped using 100 replications. We expected these prognosis models to create potential for index event bias for LDL-C (because it likely causes CHD onset), but not for HDL-C (because it likely does not cause CHD onset; Fig 6). The pattern of results when conditioning on CHD history was not entirely as expected given the potential for index event bias, however (Table 2). Results of the IEB-naive univariable MR analysis suggest that higher LDL-C (per standard deviation (SD)) raises the odds of CHD mortality, by 2.12 (95% CI = 1.20, 3.73) times higher (Table 2). However, results of a multivariable MR model adjusting for HDL-C and triglycerides (S1 Text) suggested that higher LDL-C lowered the odds of CHD mortality, by 0.81 (95% CI = 0.68, 0.96) times lower. In contrast, higher HDL-C (per SD) appeared to reduce the odds of CHD mortality, and this did not attenuate towards the null upon adjustment for LDL-C and triglycerides, e.g., the estimate for HDL-C was 0.88 (95% CI = 0.66, 1.16) before adjustment and 0.61 (95% CI = 0.51, 0.72) after adjustment (S1 Text). This inverse effect for multivariable-adjusted HDL-C is not consistent with null effects on CHD mortality risk seen in RCTs of HDL-C modification by drug therapies [42,43] and may reflect residual pleiotropy, which is not addressed by the adjustments currently made for other
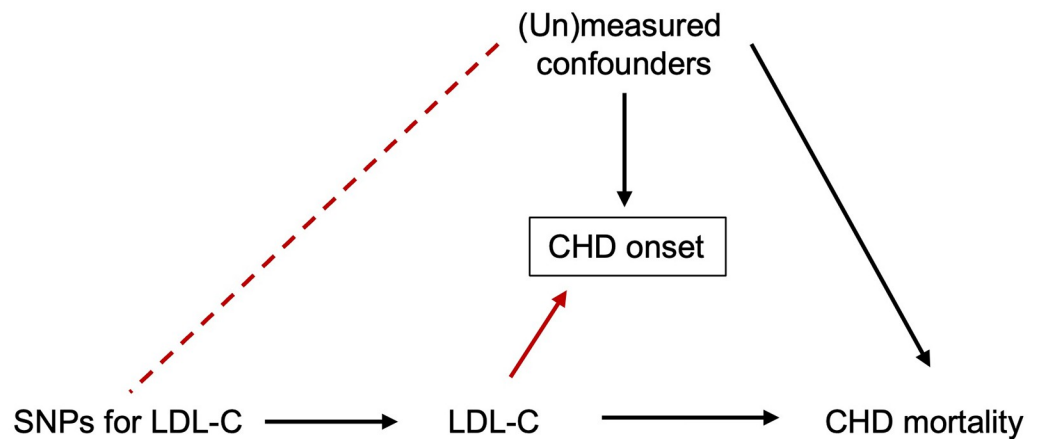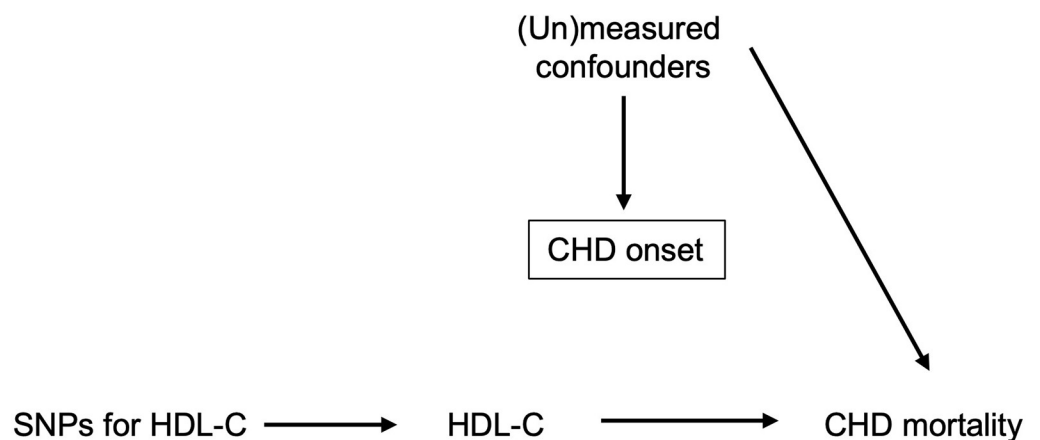
Exposure likely causal for disease onset (index-event bias expected)

Exposure likely non-causal for disease onset (index-event bias not expected)

Fig 6. Contrasting scenarios in which index event bias is or is not expected, based on the likely causality of exposures for disease onset.
Solid black lines indicate assumed causality. Absent lines indicate assumed lack of causality. Dashed black line indicates induced association. Absent dashed line indicates lack of induced association. Boxes indicate a variable that has been conditioned on. CHD, coronary heart disease; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; SNP, single nucleotide polymorphism.

https://doi.org/10.1371/journal.pgen.1010596.g006

lipids (LDL-C and triglycerides); i.e., the genetic variants for HDL-C may be affecting CHD mortality via other pathways beyond HDL-C, which are unaccounted for.

To attempt to mitigate any index event bias induced from conditioning on CHD history, we applied IPW adjustments to MR models of LDL-C and HDL-C with CHD mortality (using predictors and criteria described in **S1 Text**). Results of these IPW adjustments for univariable LDL-C models provided estimates which were directionally consistent with IPW-unadjusted results with modest attenuation towards the null (2.12 (95% CI = 1.20, 3.73) before IPW and

Table 2. One-sample MR estimates of the effect of lipid traits on CHD mortality risk among adults with a history of CHD in UK Biobank (*N* = 20,552 eligible).

	OR (95% CI) for CHD mortality per SD higher exposure
Index event bias expected (exposure likely causes CHD onset)	
LDL-C	
Without IPW	2.12 (1.20, 3.73)
With IPW	1.76 (1.00, 3.11)
LDL-C, adj. for HDL-C and trig.	
Without IPW	0.81 (0.68, 0.96)
With IPW	0.95 (0.74, 1.22)
Index event bias not expected (exposure likely does not cause onset)	
HDL-C	
Without IPW	0.88 (0.66, 1.16)
With IPW	1.08 (0.79, 1.48)
HDL-C, adj. for LDL-C and trig.	
Without IPW	0.61 (0.51, 0.72)
With IPW	0.80 (0.64, 1.00)

Estimates are among adults of European ancestry and are adjusted for age, sex, and the first ten genetic principal components. Models are two-stage prediction substitution regression models with bootstrapped standard errors (100 replications).

CHD, coronary heart disease; HDL-C, high-density lipoprotein cholesterol; IPW, inverse probability weighting; LDL-C, low-density lipoprotein cholesterol; MR, Mendelian randomisation; OR, odds ratio; SD, standard deviation; trig, triglycerides; CI, confidence interval.

https://doi.org/10.1371/journal.pgen.1010596.t002

1.76 (1.00, 3.11) after IPW). A greater attenuation towards the null was seen in multivariable LDL-C models (adjusted for HDL-C and triglycerides) upon IPW adjustment (0.81 (95% CI = 0.68, 0.96) before IPW and 0.95 (95% CI = 0.74, 1.22) after IPW). For univariable HDL-C, the point estimate changed from negative (0.88) to positive (1.08) upon IPW adjustment, although confidence intervals were imprecise (**Table 2**). The estimate for multivariable-adjusted HDL-C also attenuated partly towards the null, this being 0.61 (95% CI = 0.51, 0.72) before IPW and 0.80 (95% CI = 0.64, 1.00) after IPW (**Table 2**).

The use of statin medications could potentially complicate the interpretation of effect estimates. Statins are commonly prescribed in adulthood to reduce LDL-C (often on the basis of total cholesterol and HDL-C values) and are known to influence the risk of CHD onset and mortality [42]. The overall prevalence of statin use in UKB was 14%, and use was far more common among participants with versus without a history of CHD (at 57.3% versus 11.3%, respectively). The prevalence of statin use was also high among participants with a CHD history who then died of CHD (68.7%), whereas statin use was lower among those without a CHD history who did not die of CHD (11.3%). LDL-C and HDL-C could each plausibly influence the likelihood of being prescribed statins among adults with CHD history; these assumptions are illustrated in **Fig A in S1 Text** and examined and confirmed using one-sample MR in the same UKB data (**S1 Text** and **Table B in S1 Text**), whereby higher LDL-C strongly raised the odds of using statins (OR = 5.71, 95% CI = 3.62, 9.00) and higher HDL-C lowered these odds (OR = 0.82, 95% CI = 0.71, 0.95). Conditioning on statin use via exclusions/stratifications or adjustment would be problematic, however, as this could heighten the potential for index event bias given the likely role of statin use as a mediator (**Fig A in S1 Text**).

Altogether, the results of this applied example of MR using individual-level data suggest that the impact of index event bias from conditioning on disease status can be modest and is

sensitive to multivariable adjustment, indicative of pleiotropy. Indeed, the extent of induced bias will depend on how divergent the joint effect of lipid traits with other causal risk factors are from the multiplicative/log-additive model, with substantial index event bias expected the more divergent they are. In a univariable MR model in UKB, higher LDL-C appeared to raise CHD mortality risk among adults with a history of CHD, in line with results of RCTs. The direction of this estimated effect was reversed in a multivariable MR model adjusting for HDL-C and triglycerides, however: higher LDL-C appeared to directly reduce CHD mortality. This inverse direct effect, e.g., with LDL-C appearing paradoxically protective against CHD mortality risk among adults with CHD history, may reflect index event bias. Index event bias may also alter the magnitude of a true effect, and so we applied IPW adjustments to LDL-C estimates. This resulted in modest attenuation of effect size for the univariable model with 95% CIs that overlapped those of initial estimates, but more substantial attenuation to the null for the multivariable model. Full attenuations towards the null may not be expected following such IPW adjustments given the necessarily incomplete set of predictive factors on which they are based. In contrast to LDL-C, conditioning on CHD history is not expected to induce index event bias for HDL-C with CHD mortality, because HDL-C is likely noncausal for CHD onset. Our MR effect estimates for HDL-C (univariable and with multivariable adjustment for LDL-C and triglycerides) with CHD mortality were inverse among adults with CHD history, however, which is unexpected given null results from RCTs. This estimate was substantially attenuated upon IPW adjustment in the univariable model, but only modestly attenuated in the multivariable model. This suggests that residual pleiotropy or other forms of selection bias given the unrepresentativeness of UKB may be greater concerns than index event bias in this applied example.

Breast cancer susceptibility PRS and breast cancer–specific mortality

Increasingly, MR and other genetic studies are exploring the prognostic role of risk factors in individuals with cancer. Such studies include those examining the effect of environmental and molecular risk factors on overall and cancer-specific survival, cancer recurrence, and tumour density [47–50]. Prognostic factors examined that are also risk factors for cancer onset are vulnerable to index event bias and, thus, approaches should be taken to examine and mitigate any bias present in these settings. In this applied example, we evaluated the association of a breast cancer susceptibility PRS with subsequent disease progression using summary genetic association data on breast cancer susceptibility in 122,977 cases and 105,974 controls and breast cancer–specific mortality in 96,661 cancer cases (7,976 events) [5,51]. A PRS for breast cancer susceptibility was constructed using 339 SNPs associated with breast cancer risk at genome-wide significance ($p < 5.0 \times 10^{-8}$, $r^2 < 0.10$). Summary statistics for SNPs comprising this PRS were then extracted from the breast cancer progression GWAS and harmonised across datasets by orienting effect estimates in relation to the same allele, resulting in 318 SNPs. In this case-only analysis, the PRS was associated with a lower risk of breast cancer–specific mortality (per unit increase in natural log odds breast cancer liability: HR 0.90, 95% CI 0.86 to 0.96).

To explore whether this finding was influenced by index event bias, we used two methods to detect and account for this bias: the Dudbridge et al. method and Slope-Hunter method. Summary statistics for breast cancer risk and breast cancer progression were harmonised across datasets then pruned for LD ($r^2 < 0.10$), resulting in 94,744 SNPs. SNP-progression effects were then regressed on SNP-risk effects using a SIMEX adjustment for regression dilution to generate a correction factor for SNP-progression effects (under the assumption of no genetic correlation between breast cancer risk and breast cancer–specific mortality). SNP effects on progression were then adjusted using this correction factor (−0.013, 95% CI −0.014
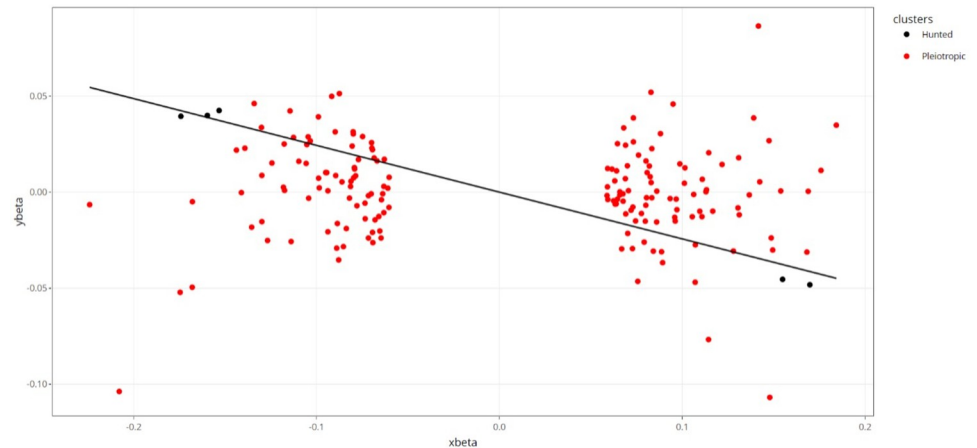
Fig 7. Slope-Hunter fitted model showing assignment of each SNP to "hunted" or "pleiotropic" clusters for an analysis examining the effect of a breast cancer susceptibility PRS on breast cancer–specific mortality. "Hunted" refers to SNPs that only affect incidence (here, breast cancer risk), and "pleiotropic" refers to SNPs that affect both incidence and prognosis (here, breast cancer risk and breast cancer–specific mortality). In this example, there were 5 hunted SNPs (i.e., those used to generate a "correction factor" for index event bias) and 171 pleiotropic SNPs.

to −0.013) and PRS analyses were reperformed, which generated a revised estimate of HR 0.92 (95% CI 0.87 to 0.97) for the effect of the breast cancer susceptibility PRS on breast cancer–specific mortality.

Using the Slope-Hunter method, a correction factor was also generated using a subset of the 94,744 SNPs that only influenced breast cancer risk (i.e., that have no effect on breast cancer–specific mortality, termed "hunted" SNPs), thus being robust to the presence of genetic correlations between disease incidence and progression (Slope-Hunter fitted model showing cluster assignment of each SNP provided in **Fig 7**). In contrast to the Dudbridge et al. method, use of Slope-Hunter generated a larger adjustment factor of −0.243 (95% CI −0.361 to −0.126). When PRS analyses were reperformed using SNP-progression effects adjusted for this correction factor, this generated a revised estimate of HR 1.15 (95% CI 1.09 to 1.22) for the effect of the breast cancer susceptibility PRS on breast cancer–specific mortality. Sensitivity analyses performed using different *p*-value thresholds to generate correction factors for both Dudbridge et al. and Slope-Hunter methods (along with corresponding changes to distributions of cluster sizes and "entropy" values for the Slope-Hunter method) are presented in **Table 3**.

This example demonstrates the potential for large differences in findings when using Dudbridge et al. and Slope-Hunter methods to correct for index event bias, and we recommend further examination of the assumptions behind each method, and sensitivity analyses. For example, the Dudbridge et al. method is more sensitive to the presence of genetic correlation between disease incidence and progression, whereas the Slope-Hunter method assumption is that there are no common causes of incidence and prognosis that explain more of the variance in incidence than the SNPs that only affect incidence. Neither of these assumptions can be verified with the observed data, so sensitivity analyses could explore their implications. Sensitivity analyses using the methods described earlier in this review could be used to either suggest plausible selection effects (i.e., provide estimates of the effects of U on incidence and prognosis) and reestimate the effects of interest, or to investigate what magnitude of selection bias would be needed to have observed the negative effect of PRS for breast cancer on mortality, if in fact there was no effect.

Table 3. Sensitivity analyses employing different *p*-value thresholds to generate correction factors across Dudbridge et al. [12] and Slope-Hunter [41] methods examining the effect of a breast cancer susceptibility PRS on breast cancer–specific mortality.

Method	*p*-value threshold*	Cluster distribution (SH)	Entropy (SH)	Adjustment factor (95% CI)
Dudbridge et al.	1.0×10^{-3}	-	-	−0.013 (−100.011, 99.909)
Slope-Hunter	1.0×10^{-3}	5 (H), 171 (P)	0.577	−0.243 (−0.361, −0.126)
Dudbridge et al.	1.0×10^{-4}	-	-	N/A
Slope-Hunter**	1.0×10^{-4}	0 (H), 42 (P)	NaN	−0.207 (−0.318, −0.097)
Dudbridge et al.	5.0×10^{-8}	-	-	N/A
Slope-Hunter	5.0×10^{-8}	5 (H), 4 (P)	0.971	−0.154 (−0.228, −0.079)

**p-value threshold used to generate correction factor. SH = Only applicable to Slope-Hunter method. H = Hunted, P = Pleiotropic.

**Note: It is advised that users of the Slope-Hunter package interpret findings with caution when no SNPs are assigned to the "hunted" cluster (e.g., findings obtained using a *p*-value threshold $< 1.0 \times 10^{-4}$ in this example) and that users explore multiple *p*-value thresholds as a sensitivity analysis.

https://doi.org/10.1371/journal.pgen.1010596.t003

Conclusions

We have highlighted the potential for bias in genetic studies of disease progression due to their case-only design. We reviewed methods available to detect this index event bias ("Detecting index event bias"), using either individual-level or summary-level data, and showed how access to individual-level data allows a more thorough investigation of biased associations between risk factors for disease incidence. Next, we outlined two sensitivity analyses ("Sensitivity analyses") to assess the magnitude of bias that would have to be present to explain observed associations in an MR analysis; we recommend that results of these analyses are presented alongside causal effect estimates for any case-only MR analysis. We then discussed methods to account for this bias ("Accounting for index event bias") in both GWAS and MR analyses and highlighted the assumptions associated with each method, such as knowledge and availability of all risk factors for IPW analysis in a single-sample setting and assumptions of linearity for the summary-level methods. Finally, we applied methods that account for and minimise bias (IPW, Dudbridge et al., Slope-Hunter) to two real-data examples.

Our first application was an individual-level data analysis of the effect of LDL-C and HDL-C on CHD mortality in which we applied the IPW method. We showed how the magnitude of the estimated effect of LDL-C on CHD mortality is reduced, albeit modestly, when using IPW to account for index event bias, in the opposite direction to that predicted from the likely direction of collider bias. This analysis is complicated by two factors, namely statin use, and the fact that the incident event (CHD onset) happened before data on the exposures were collected. Thus, the IPW may in fact be modelling variables affected by CHD onset, rather than vice versa. Similarly, earlier LDL-C/HDL-C measures affect statin use at baseline, which is likely to also have been affected by the index event and to cause later mortality. This highlights how it may be difficult to generate accurate causal effect estimates in realistically complex situations. Further methods are needed to address the issue of bias due to pre-index event statin use being both caused by pre-index event measures of exposure and being a cause of the index event, and subsequent statin use to be caused by pre-index event exposure and the index event (and likely an interaction between them). These situations are potentially further complicated by the fact that the data used here (UKB) are known to be highly selected from the general population via nonrepresentative participation and are thus subject to other forms of selection bias.

Although the Dudbridge et al. and Slope-Hunter methods can be applied to both individual- and summary-level data, we chose to illustrate them in our second applied example using summary-level data. We showed how the Slope-Hunter and Dudbridge et al. correction factors produced different estimates of the effect of a breast cancer PRS on breast cancer–specific mortality. This highlights the need to consider whether the assumptions of each method hold, namely the lack of genetic correlation between incident breast cancer and mortality for the Dudbridge et al. method, and that the SNPs only associated with incident breast cancer explain at least as much variance in incidence as the SNPs associated with both incidence and mortality for the Slope-Hunter method. We recommend performing both methods, repeating subsequent PRS/MR analyses with both versions of the corrected data in sensitivity analyses, and presenting results of each method alongside a discussion of the assumptions and why they are/are not likely to hold for the progression example in question.

It should be noted that, in this paper, we have focused on one type of selection bias, index event bias, and there are other forms of selection bias to consider in genetic studies of disease progression, which were beyond the scope of this review. For example, studies of disease progression require a population of individuals who have survived long enough to have developed disease and to allow sufficient time for disease to progress, meaning that such studies are often restricted to an older population and can be susceptible to survival bias. Some of the methods presented in this review could also be applied in the context of survival bias, for example, IPW can be applied so long as data are available for those individuals who did not survive (although this assumes that every factor predicting survival has been measured) [52]. Further methods to deal with survival bias in MR studies have been discussed previously [52–55]. Additionally, analyses of disease progression require longitudinal data; therefore, they can be vulnerable to bias due to missing data and loss-to-follow-up. Equally, bias can arise in MR studies when the genetic variants differ in their association to the exposure in cases of the disease versus the general population [56]. An additional consideration when performing an MR analysis with disease progression as the outcome would be to verify that the SNP-exposure associations are the same in cases as in a healthy population as there may be effect modification by factors relating to having the disease. As a consequence, the SNP-exposure estimates from a healthy population would not be correct.

Finally, we should mention that quantifying and adjusting for index event bias in case-only studies is an active area of research, and we expect new methods to emerge in the coming years. For example, Cinelli et al. [57] recently developed a sensitivity analysis tool for pleiotropic bias and population stratification in MR. Their approach can be used to model collider bias in MR studies where conditioning on a collider opens a pleiotropic path between the instrument and the outcome, meaning that it may be applicable to studies of disease progression, although this was only briefly explored by Cinelli et al. and more research in that direction may be needed. As another example, Heckman-type sample selection models [58] are often used in econometrics to adjust for bias due to missing data, and a recent paper explored whether these models can also be applied to genetic studies [59].

In conclusion, our review summarises established approaches for detecting and adjusting for index event bias in genetic studies of disease progression. We hope that our work will provide a useful resource to applied researchers working on such studies.

Data availability statement

Summary statistics for the age and sex GWAS in myocardial infarction cases in UK Biobank are available through the University of Bristol data repository (https://doi.org/10.5523/bris.2p0ih3u1cpxq12smf079dm1t5a). Fig 5 was generated using publicly available summary

Supporting information

S1 Text. Supplementary methods. Fig A in S1 Text. Directed acyclic graph illustrating potential pleiotropic pathways from LDL-C and HDL-C to CHD mortality. DAG illustrating potential pleiotropic pathways from LDL-C and HDL-C to CHD mortality. Solid black lines indicate assumed causality. Absent solid lines indicate assumed lack of causality. Dashed red lines indicate induced associations. Absent dashed red lines indicate lack of induced associations. Boxes indicate a variable that has been conditioned on. **Table A in S1 Text.** One-sample MR estimates of the effect of lipid traits on CHD onset in UK Biobank (N = 337,288 eligible). Estimates are among adults of European ancestry and are adjusted for age, sex, and the first ten genetic principal components. Models are two-stage prediction substitution regression models with bootstrapped standard errors (100 replications). **Table B in S1 Text.** One-sample MR estimates of the total effect of lipid traits on statin use in UK Biobank. Estimates are among adults of European ancestry and are adjusted for age, sex, and the first ten genetic principal components. Models are two-stage prediction substitution regression models with bootstrapped standard errors (100 replications). (DOCX)

Author Contributions

Conceptualization: Ruth E. Mitchell, April E. Hartley, Venexia M. Walker, Apostolos Gkatzionis, James Yarmolinsky, Joshua A. Bell, Lavinia Paternoster, Kate Tilling, George Davey Smith.

Formal analysis: Ruth E. Mitchell, April E. Hartley, Venexia M. Walker, Apostolos Gkatzionis, James Yarmolinsky, Joshua A. Bell, Amanda H. W. Chong.

Funding acquisition: Kate Tilling, George Davey Smith.

Investigation: Ruth E. Mitchell, April E. Hartley, Venexia M. Walker, Apostolos Gkatzionis, James Yarmolinsky, Joshua A. Bell, Amanda H. W. Chong.

Project administration: Ruth E. Mitchell, Lavinia Paternoster, Kate Tilling, George Davey Smith.

Supervision: Lavinia Paternoster, Kate Tilling, George Davey Smith.

Visualization: Ruth E. Mitchell, April E. Hartley, Venexia M. Walker, Apostolos Gkatzionis, James Yarmolinsky, Joshua A. Bell, Amanda H. W. Chong.

Writing – original draft: Ruth E. Mitchell, April E. Hartley, Venexia M. Walker, Apostolos Gkatzionis, James Yarmolinsky, Joshua A. Bell, Amanda H. W. Chong, Lavinia Paternoster, Kate Tilling, George Davey Smith.

Writing – review & editing: Ruth E. Mitchell, April E. Hartley, Venexia M. Walker, Apostolos Gkatzionis, James Yarmolinsky, Joshua A. Bell, Amanda H. W. Chong, Lavinia Paternoster, Kate Tilling, George Davey Smith.

References

1. Phipps AI, Passarelli MN, Chan AT, Harrison TA, Jeon J, Hutter CM, et al. Common genetic variation and survival after colorectal cancer diagnosis: a genome-wide analysis. Carcinogenesis. 2016; 37 (1):87–95. https://doi.org/10.1093/carcin/bgv161 PMID: 26586795

2. Chang IS, Jiang SS, Yang JC-H, Su W-C, Chien L-H, Hsiao C-F, et al. Genetic Modifiers of Progression-Free Survival in Never-Smoking Lung Adenocarcinoma Patients Treated with First-Line Tyrosine Kinase Inhibitors. Am J Respir Crit Care Med. 2017; 195(5):663–73. https://doi.org/10.1164/rccm.201602-0300OC PMID: 27669169

3. Fogh I, Lin K, Tiloca C, Rooney J, Gellera C, Diekstra FP, et al. Association of a Locus in the CAMTA1 Gene With Survival in Patients With Sporadic Amyotrophic Lateral Sclerosis. JAMA Neurol. 2016; 73 (7):812–20. https://doi.org/10.1001/jamaneurol.2016.1114 PMID: 27244217

4. Lee JC, Biasci D, Roberts R, Gearry RB, Mansfield JC, Ahmad T, et al. Genome-wide association study identifies distinct genetic contributions to prognosis and susceptibility in Crohn's disease. Nat Genet. 2017; 49(2):262–8. https://doi.org/10.1038/ng.3755 PMID: 28067912

5. Guo Q, Schmidt MK, Kraft P, Canisius S, Chen C, Khan S, et al. Identification of novel genetic markers of breast cancer survival. J Natl Cancer Inst. 2015; 107(5):djv081. https://doi.org/10.1093/jnci/djv081 PMID: 25890600

6. Ziv E, Dean E, Hu D, Martino A, Serie D, Curtin K, et al. Genome-wide association study identifies variants at 16p13 associated with survival in multiple myeloma patients. Nat Commun. 2015; 6(1):7539.

7. Liu G, Peng J, Liao Z, Locascio JJ, Corvol JC, Zhu F, et al. Genome-wide survival study identifies a novel synaptic locus and polygenic score for cognitive progression in Parkinson's disease. Nat Genet. 2021; 53(6):787–93. https://doi.org/10.1038/s41588-021-00847-6 PMID: 33958783

8. Munafò MR, Tilling K, Taylor AE, Evans DM, Davey Smith G. Collider scope: when selection bias can substantially influence observed associations. Int J Epidemiol. 2018; 47(1):226–35. https://doi.org/10.1093/ije/dyx206 PMID: 29040562

9. Dahabreh IJ, Kent DM. Index Event Bias as an Explanation for the Paradoxes of Recurrence Risk Research. JAMA. 2011; 305(8):822–3. https://doi.org/10.1001/jama.2011.163 PMID: 21343582

10. Zafrir B, Jaffe R, Rubinshtein R, Karkabi B, Flugelman MY, Halon DA. Influence of Body Mass Index on Long-Term Survival After Cardiac Catheterization. Am J Cardiol. 2018; 121(1):113–9. https://doi.org/10.1016/j.amjcard.2017.09.028 PMID: 29122277

11. Mitchell RE, Paternoster L, Davey Smith G. Mendelian Randomization in Case Only Studies: A Promising Approach to be Applied With Caution. Am J Cardiol. 2018; 122(12):2169–71. https://doi.org/10.1016/j.amjcard.2018.09.035 PMID: 30396509

12. Dudbridge F, Allen RJ, Sheehan NA, Schmidt AF, Lee JC, Jenkins RG, et al. Adjustment for index event bias in genome-wide association studies of subsequent events. Nat Commun. 2019; 10(1):1561. https://doi.org/10.1038/s41467-019-09381-w PMID: 30952951

13. Jiang Z, Ding P. The directions of selection bias. Stat Probab Lett. 2017; 125:104–9.

14. Bartlett JW, Harel O, Carpenter JR. Asymptotically Unbiased Estimation of Exposure Odds Ratios in Complete Records Logistic Regression. Am J Epidemiol. 2015; 182(8):730–6. https://doi.org/10.1093/aje/kwv114 PMID: 26429998

15. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004; 15(5):615–25. https://doi.org/10.1097/01.ede.0000135174.63482.43 PMID: 15308962

16. Yaghootkar H, Bancks MP, Jones SE, McDaid A, Beaumont R, Donnelly L, et al. Quantifying the extent to which index event biases influence large genetic association studies. Hum Mol Genet. 2017; 26 (5):1018–30. https://doi.org/10.1093/hmg/ddw433 PMID: 28040731

17. Hu YJ, Schmidt AF, Dudbridge F, Holmes MV, Brophy JM, Tragante V, et al. Impact of Selection Bias on Estimation of Subsequent Event Risk. Circ Cardiovasc Genet. 2017; 10(5). https://doi.org/10.1161/CIRCGENETICS.116.001616 PMID: 28986451

18. Howe LJ, Dudbridge F, Schmidt AF, Finan C, Denaxas S, Asselbergs FW, et al. Polygenic risk scores for coronary artery disease and subsequent event risk amongst established cases. Hum Mol Genet. 2020; 29(8):1388–95. https://doi.org/10.1093/hmg/ddaa052 PMID: 32219344

19. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?. Int J Epidemiol. 2003; 32(1):1–22. https://doi.org/10.1093/ije/dyg070 PMID: 12689998

20. Richmond RC, Davey Smith G. Mendelian Randomization: Concepts and Scope. Cold Spring Harb Perspect Med. 2021.

21. Paternoster L, Tilling K, Davey Smith G. Genetic epidemiology and Mendelian randomization for informing disease therapeutics: Conceptual and methodological challenges. PLoS Genet. 2017; 13(10): e1006944. https://doi.org/10.1371/journal.pgen.1006944 PMID: 28981501

22. Howe CJ, Cole SR, Lau B, Napravnik S, Eron JJ Jr., Selection Bias Due to Loss to Follow Up in Cohort Studies. Epidemiology. 2016; 27(1):91–7. https://doi.org/10.1097/EDE.0000000000000409 PMID: 26484424

23. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009; 338:b2393. https://doi.org/10.1136/bmj.b2393 PMID: 19564179

24. Pirastu N, Cordioli M, Nandakumar P, Mignogna G, Abdellaoui A, Hollis B, et al. Genetic analyses identify widespread sex-differential participation bias. Nat Genet. 2021; 53(5):663–71. https://doi.org/10.1038/s41588-021-00846-7 PMID: 33888908

25. Winkler TW, Kutalik Z, Gorski M, Lottaz C, Kronenberg F, Heid IM. EasyStrata: evaluation and visualization of stratified genome-wide association meta-analysis data. Bioinformatics (Oxford, England). 2015; 31(2):259–61. https://doi.org/10.1093/bioinformatics/btu621 PMID: 25260699

26. Smith LH, VanderWeele TJ. Bounding Bias Due to Selection. Epidemiology. 2019; 30(4):509–16. https://doi.org/10.1097/EDE.0000000000001032 PMID: 31033690

27. Smith LH, VanderWeele TJ. Simple Sensitivity Analysis for Control Selection Bias. Epidemiology. 2020; 31(5):e44–e5. https://doi.org/10.1097/EDE.0000000000001207 PMID: 32404769

28. Smith L, Mathur M, Vander Weele T. Multiple-bias sensitivity analysis using bounds. 2020.

29. Noyce AJ, Kia DA, Hemani G, Nicolas A, Price TR, De Pablo-Fernandez E, et al. Estimating the causal influence of body mass index on risk of Parkinson disease: A Mendelian randomisation study. PLoS Med. 2017; 14(6):e1002314. https://doi.org/10.1371/journal.pmed.1002314 PMID: 28609445

30. Hernan MA, Robins JM. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC; 2020.

31. Griffith GJ, Morris TT, Tudball MJ, Herbert A, Mancano G, Pike L, et al. Collider bias undermines our understanding of COVID-19 disease risk and severity. Nat Commun. 2020; 11(1):5749. https://doi.org/10.1038/s41467-020-19478-2 PMID: 33184277

32. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res. 2013; 22(3):278–95. https://doi.org/10.1177/0962280210395740 PMID: 21220355

33. Cole SR, Hernán MA. Constructing Inverse Probability Weights for Marginal Structural Models. Am J Epidemiol. 2008; 168(6):656–64. https://doi.org/10.1093/aje/kwn164 PMID: 18682488

34. Kish L. Survey sampling. New York: John Wiley & Sons; 1965.

35. Zhou Y, Matsouaka RA, Thomas L. Propensity score weighting under limited overlap and model misspecification. Stat Methods Med Res. 2020; 29(12):3721–56. https://doi.org/10.1177/0962280220940334 PMID: 32693715

36. Chesnaye NC, Stel VS, Tripepi G, Dekker FW, Fu EL, Zoccali C, et al. An introduction to inverse probability of treatment weighting in observational research. Clin Kidney J. 2022; 15(1):14–20. https://doi.org/10.1093/ckj/sfab158 PMID: 35035932

37. Canan C, Lesko C, Lau B. Instrumental Variable Analyses and Selection Bias. Epidemiology (Cambridge, Mass). 2017; 28(3):396–8. https://doi.org/10.1097/EDE.0000000000000639 PMID: 28169934

38. Gkatzionis A, Burgess S. Contextualizing selection bias in Mendelian randomization: how bad is it likely to be? Int J Epidemiol. 2019; 48(3):691–701. https://doi.org/10.1093/ije/dyy202 PMID: 30325422

39. Silverman MG, Ference BA, Im K, Wiviott SD, Giugliano RP, Grundy SM, et al. Association Between Lowering LDL-C and Cardiovascular Risk Reduction Among Different Therapeutic Interventions: A Systematic Review and Meta-analysis. JAMA. 2016; 316(12):1289–97. https://doi.org/10.1001/jama.2016.13985 PMID: 27673306

40. Cai S, Hartley A, Mahmoud O, Tilling K, Dudbridge F. Adjusting for collider bias in genetic association studies using instrumental variable methods. Genet Epidemiol. 2022; 46:303–316. https://doi.org/10.1002/gepi.22455 PMID: 35583096

41. Mahmoud O, Dudbridge F, Davey Smith G, Munafo M, Tilling K. A robust method for collider bias correction in conditional genome-wide association studies. Nat Commun. 2022; 13(1):619. https://doi.org/10.1038/s41467-022-28119-9 PMID: 35110547

42. Ference BA, Ginsberg HN, Graham I, Ray KK, Packard CJ, Bruckert E, et al. Low-density lipoproteins cause atherosclerotic cardiovascular disease. 1. Evidence from genetic, epidemiologic, and clinical studies. A consensus statement from the European Atherosclerosis Society Consensus Panel. Eur Heart J. 2017; 38(32):2459–72. https://doi.org/10.1093/eurheartj/ehx144 PMID: 28444290

43. Schwartz GG, Olsson AG, Abt M, Ballantyne CM, Barter PJ, Brumm J, et al. Effects of Dalcetrapib in Patients with a Recent Acute Coronary Syndrome. N Engl J Med. 2012; 367(22):2089–99. https://doi.org/10.1056/NEJMoa1206797 PMID: 23126252

44. Voight BF, Peloso GM, Orho-Melander M, Frikke-Schmidt R, Barbalic M, Jensen MK, et al. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. Lancet. 2012; 380 (9841):572–80. https://doi.org/10.1016/S0140-6736(12)60312-2 PMID: 22607825

45. Võsa U, Claringbould A, Westra H-J, Bonder MJ, Deelen P, Zeng B, et al. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. bioRxiv. 2018:447367.

46. Richardson TG, Sanderson E, Palmer TM, Ala-Korpela M, Ference BA, Davey Smith G, et al. Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable Mendelian randomisation analysis. PLoS Med. 2020; 17(3):e1003062. https://doi.org/10.1371/journal.pmed.1003062 PMID: 32203549

47. Escala-Garcia M, Morra A, Canisius S, Chang-Claude J, Kar S, Zheng W, et al. Breast cancer risk factors and their effects on survival: a Mendelian randomisation study. BMC Med. 2020; 18(1):327. https://doi.org/10.1186/s12916-020-01797-2 PMID: 33198768

48. Guo Q, Burgess S, Turman C, Bolla MK, Wang Q, Lush M, et al. Body mass index and breast cancer survival: a Mendelian randomization analysis. Int J Epidemiol. 2017; 46(6):1814–22. https://doi.org/10.1093/ije/dyx131 PMID: 29232439

49. Hua X, Dai JY, Lindström S, Harrison TA, Lin Y, Alberts SR, et al. Genetically Predicted Circulating C-Reactive Protein Concentration and Colorectal Cancer Survival: A Mendelian Randomization Consortium Study. Cancer Epidemiol Biomarkers Prev. 2021; 30(7):1349–58. https://doi.org/10.1158/1055-9965.EPI-20-1848 PMID: 33972368

50. Afzal S, Brøndum-Jacobsen P, Bojesen SE, Nordestgaard BG. Genetically low vitamin D concentrations and increased mortality: Mendelian randomisation analysis in three large cohorts. BMJ. 2014; 349:g6330. https://doi.org/10.1136/bmj.g6330 PMID: 25406188

51. Michailidou K, Lindström S, Dennis J, Beesley J, Hui S, Kar S, et al. Association analysis identifies 65 new breast cancer risk loci. Nature. 2017; 551(7678):92–4. https://doi.org/10.1038/nature24284 PMID: 29059683

52. Smit RAJ, Trompet S, Dekkers OM, Jukema JW, le Cessie S. Survival Bias in Mendelian Randomization Studies: A Threat to Causal Inference. Epidemiology. 2019; 30(6):813–6. https://doi.org/10.1097/EDE.0000000000001072 PMID: 31373921

53. Vansteelandt S, Walter S, Tchetgen ET. Eliminating Survivor Bias in Two-stage Instrumental Variable Estimators. Epidemiology. 2018; 29(4):536–541. https://doi.org/10.1097/EDE.0000000000000835 PMID: 29652757

54. Schooling CM, Lopez P, Au Yeung SL, Huang JV. Survival bias and competing risk can severely bias Mendelian Randomization studies of specific conditions. bioRxiv. 2019:716621.

55. Vansteelandt S, Dukes O, Martinussen T. Survivor bias in Mendelian randomization analysis. Biostatistics. 2018; 19(4):426–43. https://doi.org/10.1093/biostatistics/kxx050 PMID: 29028924

56. Zheng J, Tang H, Lyon M, Davies NM, Walker V, Floyd JS, et al. Genetic effect modification of cis-acting C-reactive protein variants in cardiometabolic disease status. bioRxiv. 2021:2021.09.23.461369.

57. Cinelli C, La Pierre N, Hill BL, Sankararaman S, Eskin E. Robust Mendelian randomization in the presence of residual population stratification, batch effects and horizontal pleiotropy. Nat Commun. 2022; 13(1):1093. https://doi.org/10.1038/s41467-022-28553-9 PMID: 35232963

58. Heckman JJ. Sample Selection Bias as a Specification Error. Econometrica. 1979; 47(1):153–61.

59. Gkatzionis A, J. TTE, Heron J, Northstone K, Tilling K. Using Instruments for Selection to Adjust for Selection Bias in Mendelian Randomization. arXiv. 2022.

60. Liu M, Jiang Y, Wedow R, Li Y, Brazel DM, Chen F, et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nat Genet. 2019; 51 (2):237–44. https://doi.org/10.1038/s41588-018-0307-5 PMID: 30643251