Causal graphs for the analysis of genetic cohort data

Oliver Hines[1,2], Karla Diaz-Ordaz[1], Stijn Vansteelandt[1,3], and Yalda Jamshidi[2,*]

[1]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, UK
[2]Molecular and Clinical Sciences Institute, St George's, University of London, UK
[3]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium

***Corresponding Author:**
Dr. Yalda Jamshidi
Molecular and Clinical Sciences Institute
St George's, University of London
Cranmer Terrace, London
SW17 0RE
UK
Email: yjamshid@sgul.ac.uk

Keywords: Causal Graphs, GWAS, Mendelian Randomisation

1

Abstract

The increasing availability of genetic cohort data has led to many Genome Wide Association Studies (GWASs) successfully identifying genetic associations with an ever-expanding list of phenotypic traits. Association, however, does not imply causation and therefore methods have been developed to study the issue of causality. Under additional assumptions, Mendelian Randomisation (MR) studies have proved popular in identifying causal effects between two phenotypes, often using GWAS summary statistics. Given the widespread use of these methods, it is more important than ever to understand, and communicate, the causal assumptions upon which they are based, so that methods are transparent, and findings are clinically relevant.

Causal graphs can be used to represent causal assumptions graphically and provide insights into the limitations associated with different analysis methods. Here we review GWAS and MR from a causal perspective, to build up intuition for causal diagrams in genetic problems. We also examine issues of confounding by ancestry, and comment on approaches for dealing with such confounding, as well as discussing approaches for dealing with selection biases arising from study design.

1 Introduction

Genetic cohort data is increasingly used to look for associations between candidate genes or genome regions and specific outcome measures, or else between modifiable risk factors and disease outcomes. Genome Wide Association Studies (GWAS), for example, are a popular and effective approach to analysing Single Nucleotide Polymorphism (SNP) data, which identifies reproducible regions of the genome associated with common traits. Observed GWAS associations, however, are not necessarily indicative of causal relationship, unless one is willing to make additional assumptions on the causal structure of the cohort data.

Mendelian Randomisation (MR) is another popular method, which uses genetic cohort data (or GWAS summary statistics) to establish causal effects between two phenotypes. MR seeks to exploit random genotype allocation, which occurs naturally due to Mendelian inheritance. The requisite MR assumptions are strong, and the causal structure underlying the data must be carefully considered so that biases are not unwittingly introduced. Since both GWAS and MR rely on genetic cohort data, it is more important than ever to understand, and communicate the causal structures found in these datasets, so that findings remain clinically relevant.

Universal frameworks to study causal structures have emerged in the past few decades, based on potential outcomes modelling[31] or causal graphs[27], contributing towards a modern causal understanding of several existing techniques, such as, randomised controlled trials, instrumental variable, and observational data techniques (propensity score methods and sample matching). Causal graphs may inform both the design and analysis of observational studies, and have successfully been applied to problems in epidemiology[13, 14], social science[4] and economics[21] to represent causal assumptions, and derive causal quantities from observed data.

Eliciting and defending causal assumptions requires an expert understanding of the problem at hand. Here we review methods from genomics and genetic epidemiology, highlighting common causal structures which can bias observed associations. We advocate the use of causal graphs, firstly as a formal tool for representing and communicating the causal assumptions regarding data collection and study design, which underly analytical methods, and secondly, for deriving testable implications based on those assumptions. Causal graphs have several attractive properties in this regard. As a communication tool they are inherently diagrammatic and equation-free, aiding interpretability, whilst as a derivation tool one may apply powerful and rigorous mathematical rules, which link causal relations to statistical associations. These rules are summarised in Section 2.1.

We will initially introduce causal concepts which form the basis of our discussion. These are then applied to an example of pleiotropy in Section 1.2. Section 2 discusses causal methods for analysing selection biases, using, as an example, the analysis of case-control data for secondary trait association. Here we see the utility of causal graphs in deriving associations between variables which occur under selection. Section 3 then reviews GWAS assumptions, addressing issues related to population structure, while Section 4 reviews MR causal assumptions, highlighting several ways in which they may be violated.

1.1 Introduction to Statistical Causal Inference

There exists rich philosophical debate on what it means for one thing to *cause* another[40], however, in the study of causal inference an interventionalist definition is used[27, 13, 18]. In this way, questions of causality are reduced to questions of the type: *what would happen if...?*

For example, for two variables A and B, we say that A **causes** B if the value that B takes would be different (or different in probability) if we had intervened by setting A to some other value. In this context we might also say that A **causally influences** B or that B is **causally dependent** on A. Two variables are said to be **statistically dependent** (or associated) if knowing the value of A in some way provides some information about the value of B (or vice-versa). Statistical dependence may arise due to a causal dependence between A and B, but also as a result of a causal dependence of both A and B on a third variable C, as we will see in the example in Section 1.2. Conversely, two variables are **statistically independent** if knowing the value of A does not provide any information about the value of B (and vice-versa).

This notion of causality may also be graphically represented using an arrow[13, 18, 28, 29], for example, $A \rightarrow B$ reads as "A causes B, but B could not possibly cause A". This arrow says nothing about the magnitude or direction of the effect that A has on B, just that if we were to intervene on A, then something would happen to B. Using these arrows one can form **paths**, which are any sequence of variables linked by arrows. For example, if A and B shared a common cause, C, then one may write the path, $A \leftarrow C \rightarrow B$. All possible paths containing three variables are given in Table 1. A path is *causal* if all the arrows point in the same direction. The path $A \rightarrow C \rightarrow B$, for example, is causal since A causes C which causes B, therefore if we were to intervene on A, the value of B could be different. Depending on the directions of the arrows, we also have additional terminology for the intermediate variable, also given in the table.

Path	Description	Terminology for the variable C
$A \rightarrow C \rightarrow B$	A causes B (through C)	Mediator
$A \leftarrow C \leftarrow B$	B causes A (through C)	Mediator
$A \leftarrow C \rightarrow B$	A and B share a common cause C	Confounder
$A \rightarrow C \leftarrow B$	A and B both cause C	Collider

Table 1: All possible paths between three variables (*A,B,C*), with a brief description and additional terminology for the intermediate variable C

On its own, a single path is of limited use, motivating a network structure to represent several paths at once. The causal Directed Acyclic Graph (DAG) is such a structure, which for a set of variables, contains *all possible* paths between them. Causal graphs are said to be **acyclic** if there are no causal paths from one variable back to itself. It may seem obvious to say that any two variables, A and B, on a causal graph could either be linked by the arrow $A \rightarrow B$, the arrow $B \rightarrow A$, or no arrow at all. Each configuration makes different assertions about the impossible causal relationship between A and B. Respectively these are that B is not a direct cause of A, A is not a direct cause of B, or that A and B could not possibly be direct causes of each other. In this sense the arrows which are absent, and those which are present are equally important. Similarly, one must be careful to include common causes of A and B, even if they are unmeasured, since to not do so is to assert that it is impossible for such variables to exist.

At this stage it is also useful to introduce some terminology[13, 18, 27, 28, 29], which will become important later on. Firstly, a **collider** is any variable on a path which is causally dependent on the two variables adjacent to it, as in the final example in Table 1. Secondly, the **ancestors** of a variable are those which causally influence it (i.e. there is a causal path from each ancestor to the variable), and finally the **descendants** of a variable are those which are caused by it (i.e. there is a causal path from the variable to its descendants).

1.2 Example using Pleiotropy

Our first example is inspired by a recent discussion of pleiotropy of the fat mass and obesity-related gene *(FTO)*[12]. Consider a Single Nucleotide Polymorphism (SNP) in the *FTO* gene, such as rs1421085, which has been found to be associated with adiposity and brain function[8]. Suppose that a genetic cohort study has been conducted where, for each individual in the study population, an investigator measures body mass index (BMI), B, cerebral blood flow, C, and genotype rs1421085 in the *FTO* gene, denoted by F and coded as 0,1 or 2.

The original authors suggested that reduced cerebral blood flow in the medial prefrontal cortex may effect impulse control and hence BMI [12]. As an illustration, we will attempt to refute the null hypothesis, that there is no causal relationship between cerebral blood flow and BMI by (1) positing the causal relationships that we believe hold amongst the variables involved; (2) representing these causal relationships using a causal graph; and (3) examining the graph, using formal operations, to derive testable assumptions.

Since a person's genome is assigned before their BMI or cerebral blood flow is determined, we argue that it is safe to assume that B and C could not possibly cause F. This assumption, however, says nothing about whether F causes B or C. Since it is possible that F causes B and C we must include the arrows $F \rightarrow B$ and $F \rightarrow C$ in our causal graph. For the purposes of illustration, we will additionally make the strong assumption that no other measured or unmeasured variables causally influence both B and C.

The causal graph in Fig.1 represents the causal assumptions posited between F, B and C under the null hypothesis that there is no causal relationship B and C. These assumptions are unnecessarily strong for the purpose of illustration, since additional variables might be included such as age or physical activity level, which are common causes of both B and C. Other violations of our assumption, which could arise due to population structure, are discussed in Section 3. We remark that while the causal graph in this example is perhaps oversimplified, such assumptions are not uncommon, and by using a causal graph representation we are required to be transparent about them.

[Figure 1 Here]

In the graph in Fig.1, there is no causal path between B and C, but that does not mean that they are statistically independent. In fact one might expect a negative correlation between BMI and cerebral blood flow since those who inherit the *FTO* variant are likely to have a higher BMI and also a lower cerebral blood flow. This statistical dependency can be read off the graph in the form of the possible path: $B \leftarrow F \rightarrow C$. It is a general rule that two variables will be statistically independent if all paths between them that contain colliders. For this reason, we can refer to paths that do not contain a collider as *open paths* and those that do as *closed paths*.

Using our causal graph, we may derive testable assumptions in an attempt to falsify our null hypothesis. Imagine, for example, that we are told the value of B for a particular patient, and are asked to predict their value of C. The value of B may inform our prediction since B and C may be statistically dependent (due to confounding by F). If, however, we are subsequently told the patient's *FTO* variant then, under our causal assumptions, a new prediction based on F and B is no better than a prediction based on F alone, since B only informed our prediction in so much as it may have conferred some information about F.

This important observation is an example of how one may *block* open paths, such as $B \leftarrow F \rightarrow C$, by *conditioning* on an intermediate variable (F). Conditioning on a variable can be done either by stratifying by that variable or by including it as an independent variable in a regression model for B or C. These conditional independences are essential as they allow us to falsify our causal assumptions.

In practice, this means that if one were to stratify our imaginary study population by their *FTO* gene variant, then, under our causal assumptions, no association between B and C should be observed within strata. An association between B and C within strata is, therefore, evidence that our assumptions are invalid. This could be because our null hypothesis does not hold, and B and C are causally related, or else because the relationship between them is confounded by some other variables, which we have not accounted for.

2 Selection Bias

Due to the considerable cost of obtaining original genetic cohort data, it is common for case-control data to be repurposed for analysis of a secondary trait, such as human height[16, 44], obesity[25], or plasma lipid concentration[45]. Methods that fail to account for the case-control study design, are known to result in inflated error rates when testing for null association using GWAS [23]. Indeed it has been argued that epidemiological data analysis depends as much on study design and background information, as on the data itself[30].

Gene-phenotype associations, induced as a consequence of study design, are problematic in GWAS analyses because they are indistinguishable from underlying causal associations in GWAS results. Using causal graphs we may gain some insight into how the non-random selection of individuals to the study cohort propagates to non-randomness in our variables of interest. We will consider an illustrative example, inspired by a real study on the effect of Sex Hormone Binding Globulin (SHBG) on Type 2 diabetes in women[11]. Consider that the study cohort was recruited on a case-control basis and consists of women with a recent Type 2 diabetes diagnosis (D = 1) and controls (D = 0), with genotyping carried out for all women. We shall examine the issues which arise when this cohort is used to conduct a GWAS analysis, with SHBG as the outcome of interest.

SHBG is a glycoprotein, produced in the liver, and the level of SHBG in an individual's blood plasma will be denoted by H. The original authors found that high levels of SHBG were associated with a lower risk of Type 2 diabetes and for this example we shall assume that diabetes status does not causally influence SHBG level. Imagine also a specific SNP, G, which does not causally influence SHBG, but does causally influence diabetes diagnosis by some other mechanism. As with the example in Section 1.2 we shall make the "no unobserved confounding" assumption, i.e. that there are no common causes of H, G, or D that we have not accounted for.

Due to the case-control design, diabetes status D causally influences selection to cohort, S. By definition S = 1 for all women in the cohort and S = 0 for all other women in the population as a whole. Our causal assumptions are represented by the causal graph in Fig.2a.

[Figure 2 Here]

Under these assumptions, G and H are statistically independent as there are no open paths between them. One would expect, therefore, to observe no association between G and H for women sampled from the population. Our cohort, however, is not randomly sampled from the population, but instead we observe only those for whom S = 1. This is equivalent to an unavoidable stratification by S, which allows us to observe only the S = 1 stratum. In this stratum, a "spurious" association between G and H may be induced, which we demonstrate by first examining the D = 1 and D = 0 strata separately.

In the cases group (D = 1) an association between G and H would be observed, since, if an individual's genotype suggests they are unlikely to have diabetes, then their diabetes status is more likely due to a low level of SHBG, and vice-versa. For women in the control group (D = 0) an association between G and H would be observed, since women in this group are less likely to carry the genotype associated with diabetes and are also more likely to have high SHBG.

We see, therefore, that G and H are associated in both the D = 0 and D = 1 strata and that this association must be induced by the stratification process, since G and H are not associated in the population. Worse than this, however, is that stratifying by S also induces associations between G and H because the proportions of each D strata in our cohort are not representative of the population as a whole. For selection problems such as these we have no choice but to consider only the strata S = 1.

In this simple example we were able to reason that selection bias may influence our results, however, in other examples it may not be so clear. Causal graphs may go some way to elucidate selection biases. It is a general rule that conditioning on a collider, or the descendants of a collider, induces statistical dependencies between the ancestors of the collider. In our case-control example D was a collider on the path: $G \rightarrow D \leftarrow H$ and we were forced to condition on S, which is a descendant of D. This conditioning resulted in a statistical dependency between G and H (the ancestors of D). This induced dependency is represented by the dashed line on the causal graph in Fig.2b.

In Section 1.2 we saw how open paths on causal graphs could be blocked by conditioning on intermediate variables. In this example, however, conditioning has the opposite effect. By unintentionally conditioning on colliders, we are effectively unblocking a path that was otherwise closed, thereby inducing associations. Several solutions have been proposed, which allow case-control data to be used for secondary trait analysis in association studies. Example analysis strategies include analysing the cases and controls separately, re-weighting the data using additional models, or including case-control status as a covariate [38, 33].

Biases introduced by conditioning on colliders are generally referred to as *collider stratification biases*[2]. The inclusion of selection variables in causal graphs, like the variable S in the case-control example, can also be useful for expressing selection and retention assumptions which suffer from similar collider stratification biases[26]. The UK Biobank is an example of a cross-sectional cohort study ($n \approx 500{,}000$) self-selected from a population of 9 million individuals invited to participate. The resultant cohort contains a lower proportion of current smokers (11% in the UK Biobank, vs approximately 19% in the general population), with a similar discrepancy observed in educational qualification attainment. For a highly self-selected cohort, such as the UK Biobank, causal graphs may be useful in exposing subtle biases induced by this self-selection.

2.1 D-separation

The rules discussed in Sections 1.1 and 2 are collectively known as the rules of d-separation (statistical dependence separation). These rules describe statistical dependencies implied by causal graphs before and after conditioning on variables. Table 2 gives a summary of these rules for all possible paths of three variables. To consider longer, more complex paths one must 'chain together' these triplets, and to consider the statistical dependence between variables on the whole causal graph, one must consider all possible paths.

For complex, multivariate causal graphs this could result in a laborious manual analysis. Fortunately, however, the tool www.dagitty.net [20] may be used to examine statistical dependence on causal graphs using an online web tool or R package.

Path	Before conditioning on C	After conditioning on C
$A \rightarrow C \rightarrow B$	open	closed
$A \leftarrow C \leftarrow B$	open	closed
$A \leftarrow C \rightarrow B$	open	closed
$A \rightarrow C \leftarrow B$	closed	open

Table 2: Summary of the rules of d-separation for all possible paths containing three variables. The two additional columns describe the statistical dependence of A and B before and after conditioning on the intermediate variable C.

3 Causal Graphs for Genome Wide Association Studies

GWAS studies are a popular and effective approach to analysing SNP data, which identifies reproducible regions of the genome associated with common traits. As of February 2020, the GWAS Catalogue contains 4439 publications and 175870 associations[6]. Despite their popularity, it is important to remember that the associations discovered by GWAS are not necessarily causal unless one is willing to make additional assumptions. In this section, we use causal graphs to make these assumptions explicit. Genetic relatedness between individuals in the study population poses an additional, well-known challenge that results in individuals with shared ancestry inheriting similar common variants. Heterogeneous study populations, therefore, complicate the task of separating the contributions of individual genetic variants toward phenotypes of interest. We refer to the problem of heterogeneous ancestry as confounding by ancestry, since this more closely aligns with the language of causal inference. It is also referred to as population structure or population stratification, when at the population level, and kinship, at the familial level.

As an illustrative example, we will use Carotid Intima-Media Thickness (CIMT) as a phenotype of interest Y. In its most basic form, one assumes that the study population is in Hardy-Weinberg Equilibrium (HWE), that is, for each individual, the value of their value of a particular SNP of interest, G, is drawn from a binomial distribution with some fixed minor allele frequency for the population.

Common practice is to model a continuous phenotype, Y, using a model which is linear in G, and other relevant variables, such as age and sex, denoted by the 'Environmental' vector, E. When Y is a binary outcome, generalised linear models such as the logistic model, are often used. The linear model for a continuous phenotype, Y, may be written as

$$Y = \alpha G + \sum_{j=1}^{p} \beta_j E_j + \epsilon \qquad (1)$$

where ε is a noise term, with constant mean given G and E, and β is a vector of parameters associated with the p environmental variables contained in the vector E. The unknown model parameters, α and β, may be estimated by Ordinary Least Squares (OLS). Ideally we would like to interpret the α parameter as *a parameter which quantifies the influence that the gene of interest has on the phenotype*, however, to do so is to make a causal assertion, requiring an examination of causal assumptions. We note that for a discussion of causal assumptions, the exact form of the regression model is not important. Instead, from a causal perspective, we are concerned with the variables which are and are not included in the regression model.

One possible causal graph for the basic GWAS analysis, which gives the α parameter the desired causal interpretation is given in Fig.3a. This graph is not unique since it is not strictly required that G and E are independent. Using the running example, the key features of this graph required to interpret α causally are

1. CIMT does not influence the gene of interest, but the reverse may be true.
2. CIMT does not influence age or sex, but the reverse may be true.
3. There are no variables (observed or otherwise), which are common causes of CIMT and the gene of interest, or of CIMT and age or sex.

The first of these assumptions is justified through the biological understanding that G is assigned before phenotypes are determined, hence reverse causation is not possible. Likewise, the second assumption is reasonable from a biological perspective. Assumption 3, however, is where the basic model breaks down. Under modern theories of Mendelian inheritance, the gene of interest depends on an individual's parental genotypes, or more generally on their ancestry. Along with the gene of interest, each individual inherits many other genetic variants, G^*, each of which could also have a causal influence over Y. The ancestry of an individual is therefore a confounder as it may be a common cause of both G and Y.

This effect is, however, negated if one assumes that Y is monogenic, so is causally affected by only one single SNP. Conversely the effect is amplified for polygenic traits, such as CIMT, which are thought to be affected by multiple genetic variants.

[Figure 3 Here]

To adequately adjust for confounding by ancestry, the basic GWAS graph Fig.3a must be updated to reflect Mendelian inheritance assumptions. Fig.3b shows a causal graph, modified to include an unmeasured ancestry variable, C, which affects the phenotype of interest through both the gene of interest, G, and other inherited variants, G^*. In this updated causal graph, we see that there are two open paths by which the gene of interest is associated with CIMT, specifically the $G \rightarrow Y$ causal path and the $G \leftarrow C \rightarrow G^* \rightarrow Y$ non-causal path. If one were able to block the non-causal path, then, the remaining association between G and Y must be due to the causal path.

One strategy for blocking the path is to condition on ancestry by stratification. Since C is unmeasured, one must assume that the population consists of one strata, which is homogeneous in ancestry with a random mating scheme and no natural selection. Under these assumptions, the HWE model is recovered, whereby G is drawn from the same distribution for all individuals, hence G and Y are not confounded by ancestry.

The causal graph in Fig.3b made several additional assumptions regarding the ancestry variable, C. The first is that there is no direct path $C \rightarrow Y$. Modern epigenetic theory, however, does permit such paths through 'imprinting' mechanisms, whereby an individual inherits DNA of the same sequence, whose function is altered by the presence of additional methyl groups.

Furthermore, Fig.3b assumes that C and E are independent. This may not be true, however, for a global study, where individuals from different ethnic groups, may have been brought up in different geographical locations, and hence, different meteorological and socio-economic conditions. It is reasonable, therefore, to posit a $C \rightarrow Y$ path through some unobserved environmental variables. We emphasise again that the arrows absent from a causal graph are important as they represent causal relationships which are assumed not to exist, whilst the arrows represent causal relationships which may exist.

3.1 Using Principal Components to Adjust for Ancestral Confounding

Examining the causal graph in Fig.3b, we discussed how the non-causal path: $G \leftarrow C \rightarrow G^* \rightarrow Y$ may be blocked by conditioning on C when one assumes the study population is homogeneous. For heterogeneous populations, however, stratification by C is not possible because it is unmeasured. Instead, the non-causal path can be blocked by conditioning on the remaining observed SNPs, G^*. This involves using G^* in a regression model for Y, or using G^* for stratification.

Intuitively, conditioning on G and G^* removes any dependency between C and Y since, if the full genotype of an individual is used to predict their phenotype, then knowledge of their ancestral genotypes provides no new information to improve our prediction. Using the full genotype in a regression model for Y requires careful consideration, since the number of covariates (SNPs), p, may exceed the number of individuals in the study, $n < p$. Such 'high-dimensional' problems require alternative models and estimation techniques.

Due to the high-dimensionality, modifying the linear model in Eq.1 to include the remaining genes as covariates would result in a model which is impossible to fit by OLS. One very common solution is to drastically reduce the dimensionality of the genetic information, using Principal Components (PCs).

PCs are used in several ways within genomic analysis: (i) PCs can be used to cluster individuals, either by excluding anomalous individuals from the dataset [1], or else clustering the data for use in a Structured Association analysis, (ii) some PC values may be included as fixed effects in a GWAS analysis, thereby accounting for some of the phenotype variation, which can be explained by the remaining SNPs, and (iii) PCs may be included as random effects in the GWAS analysis, an approach which is equivalent to using a Linear Mixed Model (LMM) [19].

Method (i) may be causally interpreted as stratifying the population into one or more sub-populations, for which we believe that HWE holds. Analysis of each sub-population may be conducted using a basic GWAS analysis. Limitations of this method are that confounding by ancestry is not accounted for within strata and it is not clear how to tune the stratification process.

The linear model for methods (ii) and (iii) may be written as

$$Y = \alpha G + \sum_{j=1}^{p} \beta_j E_j + \sum_{j=1}^{q} \gamma_j P_j + \epsilon \qquad (2)$$

where P is the vector of q principal components, summarising the genetic data of a particular individual, each component of which has a coefficient given by the γ parameter vector, and where ε has constant mean given G,E and P. In the fixed effect model (method ii), the q-dimensional parameter vector, γ is treated as a fixed covariate, which may be estimated using conventional methods such as by OLS.

Alternatively, one may treat the parameters γ_j as random effects (method iii), by assuming a normally distributed prior for γ, resulting in a LMM. The use of LMMs in genomic data is not restricted to GWAS analyses. They are frequently applied to phenotype prediction, heritability estimation, and rare-variant analysis [24]. One key feature of LMMs is that the random effect (given by$\sum_{j=1}^{q} \gamma_j P_j$above) may be written in terms of a 'genetic similarity matrix', which is used to model the covariance between any pair of individuals in the cohort. A more detailed discussion of LMMs and methods for measuring genetic similarity can be found in Appendix A.

4 Causal Graphs for Mendelian Randomisation

Mendelian Randomisation (MR) studies also make use of genetic SNP data, or GWAS summary statistics, with the aim of inferring the effect of a genetically modified exposure (e.g. alcohol consumption) on another phenotype (e.g. cardiovascular disease). GWAS results from multiple cohorts may be used to conduct Two- Sample MR analysis. MR base which is a database of GWAS statistics for conducting Two-Sample MR, contained associations from 1673 GWAS, as of May 2018[17]. Another systematic review estimates a 10-fold increase in published MR studies between 2004 and 2015, with the majority (51%) in the fields of cardiovascular disease and diabetes[37]. MR is therefore increasing in popularity, most likely due to the increasing availability of GWAS summary statistics and large cohorts with genetic and phenotypic data.

This section provides an overview of the technique, from the statistical causal inference framework. We refer the interested reader to [40, 7, 32].

4.1 Instrumental Variable Methods

MR exploits the idea that a particular genotype affects the phenotype of interest only indirectly, through the exposure of interest, and that this genotype is assigned randomly (given the parents' genes) at meiosis, independently of the possible confounding factors. This is essentially using the genotype as a so-called *instrumental variable* (IV) for the effect of the exposure on the outcome [9]. This is appealing, as it allows to estimate causal effects event in the presence of exposure-outcome unobserved confounding. Nevertheless, MR makes a number of causal assumptions, known as IV assumptions, which are not always carefully stated and evaluated in applications and are separate from any parametric modelling assumptions, which may also be required.

For illustration, we consider a specific example [22] where the interest is to investigate the causal effect of the level of C-reactive Protein (CRP) on CIMT by exploiting random assignment of a genetic variant, G, associated with CRP. Here CRP is referred to as the exposure, X, CIMT as the outcome, Y , and G as the instrumental variable (or instrumental gene).

[Figure 4 Here]

Note that the IV causal graph permits unmeasured variables that may influence both the exposure CRP and the outcome CIMT, here denoted by U. The IV assumptions encoded by the causal graph in Fig.4a can be written formally as follows

1. CIMT does not influence CRP, but the reverse may be true.
2. Relevance: The instrumental gene is associated with the level of CRP.
3. Exclusion restriction: The instrumental gene may affect CIMT only through its effect on CRP.
4. Unconfoundedness: There is no variable, observed or otherwise, which is a common cause of the instrumental gene and CIMT.

For assumption 1, domain specific knowledge is generally required to defend the $X \rightarrow Y$ causal relationship over the alternative, $Y \rightarrow X$. For this example, it is usually assumed that proteins causally influence disease outcomes, rather than the other way round. Collectively, assumptions 2 to 4 are known as the IV assumptions as they describe the relationship between the IV and the variables U, X and Y. In a randomised control trial (RCT), where the IV is the randomly assigned treatment group, these assumptions are more simple to justify, since the randomisation process is known, and we can engineer the randomised treatment so that it is (a) associated with the exposure, and (b) does not influence the outcome except through the exposure, although in some settings justification of the exclusion restriction remains challenging.

In the MR setting, we justify the relevance condition (assumption 2) by choosing instrumental genes following a GWAS analysis. In practice, several candidate instrumental genes are often used to support or discredit the evidence of a single one. The exclusion restriction (assumption 3) is, however, more problematic as genetic variants may have independent pleiotropic effects on multiple phenotypes. Pleiotropic effects violate the exclusion restriction by introducing alternative paths of the type $G \rightarrow Y$.

Recent developments in MR do allow for some limited pleiotropy, such as MR-Egger[3], which permits a direct path from $G \rightarrow Y$ in Two-Sample studies (under specific assumptions), and the MRGxE method[36], which allows for pleiotropic 'Gene-by-environment' interactions provided they reside on the $G \rightarrow X$ path. Selection of instrumental genes in MR is, however, an open topic of debate, both in terms of statistical and biological considerations[37]. Recent statistical work considers variable selection methods, such as the Lasso, to select IVs[46]. Whilst the exclusion restriction cannot be proven, it may sometimes be possible to show that they are inconsistent with prior evidence. Methods for doing so include leveraging prior causal assumptions, identifying modifying subgroups, or by use of instrument inequality tests[15].

Unconfoundedness (assumption 4) prohibits edges of the type $U \rightarrow G$, which is reasonably well justified on the basis of Mendelian inheritance. As in Section 3, however confounding by ancestry violates this assumption, since unobserved ancestry variables, C, may causally influence the outcome through their effect on other genetic variants as well as causally influencing the instrumental gene itself. Ancestrally heterogeneous populations are therefore known to violate the unconfoundedness in MR, and practitioners are recommended where possible to use homogeneous cohorts, thought to be in HWE.

A modified causal graph, which relaxes the IV assumptions to allow for confounding by ancestry, and limited pleiotropic effects, can be seen in Fig.4b. This graph represents a more general set of causal assumptions, to emphasise the assumptions of the IV graph. The standard IV graph may be recovered by removing arrows from the modified causal graph, or in other words, by assuming certain null causal relationships.

If only the $G \rightarrow Y$ arrow is removed from the causal graph in Fig.4b (i.e. G has no pleiotropic effect on Y) then G may be used as a *conditional instrumental variable*, assuming one collects adequate data on the other genetic variants G^*. In a *conditional instrumental variable* analysis, the gene G acts as an instrumental variable after conditioning on G^* in the models for X and for Y. This conditioning has the effect of blocking the open paths: $G \leftarrow C \rightarrow G^* \rightarrow X$ and $G \leftarrow C \rightarrow G^* \rightarrow Y$. Once blocked, unconfoundedness is no longer violated so G again acts as an instrument, allowing for valid MR analysis with ancestrally heterogeneous cohorts. Conditioning on G^* may be achieved using the methods in Section 3.1.

Violation of any of the IV assumptions would result in invalid causal estimates. We refer the interested reader to [41] for a comprehensive discussion of the challenges faced by MR studies when justifying the IV assumptions and on how to conduct sensitivity analyses.

4.2 Survivor Bias in Mendelian Randomisation

One setting where causal graphs are especially useful for evaluating MR assumptions is in the use of genetic instruments to asses survival biases. Here we consider the example given in [42], namely where an MR analysis of the effect of vitamin D levels on mortality is performed using a cohort of ancestrally homogenous, genotyped individuals between the ages of 40 and 71 years old. Using causal graphs, we show how survivor bias may be introduced because recruitment to the cohort depends on an individual having survived long enough to be eligible for recruitment.

Selection to the cohort depends on T, the lifetime of an individual, being larger than some index time, T_0. By definition, an index time is actually assigned only to individuals in the cohort (who are indexed at some point between the ages of 40 and 71), however, we could imagine that individuals outside the cohort could also be given an index time, for example by sampling from the birth register. As before, we will denote selection to the cohort by the variable S, with $S = 1$ for all individuals in the cohort.

Let D be the level of vitamin D at index and assume that it captures the effect on lifetime of an individual's entire exposure to vitamin D since birth. This assumption is implicit in all MR studies, since to not assume it would generally violate the exclusion assumption, in the sense that we could imagine an additional variable (e.g. adolescent vitamin D level) which causally influences the vitamin D level recorded at index, as well as the lifetime of the individual directly.

Finally, we shall assume that an appropriate genetic instrument (e.g. filaggrin genotype) has been recorded, which we shall denote, G, and assume is randomised by Mendelian inheritance, since the cohort is homogenous. As with the standard MR causal graph, we shall permit unmeasured confounding variables which might causally influence both vitamin D level and lifetime. Our causal assumptions for this example are represented by the causal graph in Fig.5a. In this example, S, is a variable which we have no choice but to condition on, hence we must be very careful to consider collider stratification biases, as discussed in Section2.

[Figure 5 Here]

We see that S is a descendent of D, due to the $D \rightarrow T \rightarrow S$ path, and that D is also a collider on the path $G \rightarrow D \leftarrow U$. Hence, by selecting only individuals who have survived, the ancestors of D (namely G and U) become associated. This violates the exclusion assumption, since association between G and T may arise from either the causal path $G \rightarrow D \rightarrow T$ or from the path $U \rightarrow T$, where U is associated with G.

The association induced by conditioning on selection is illustrated by the dashed line in Fig.5b. Recent work proposes various strategies for MR estimation under survivor bias, using a semi-parametric additive hazard model[42], similar to the canonical Cox proportional hazards model. This relates to similar work on MR for censored survival outcomes[39].

Interestingly, however, this problem of survivor bias disappears when testing the null hypothesis that D has no causal influence on T. Under this null hypothesis, there is, by definition, no $D \rightarrow T$ arrow, hence G is not an ancestor of T and no association between G and U is induced.

5 Conclusion

We have demonstrated, through examples of the most common analytical techniques employed in genetic studies, that a causal inference framework, and in particular the use of causal graphs, allows the analyst to (i) to represent their knowledge of the causal relationships involved in the question at hand, and (ii) use the rules of d-separation, to query the assumptions under which popular genetic analysis methods lead to causal interpretations.

Causal graphs may also inform intuition regarding the advantages and limitations of different analytical techniques from the outset and are useful in deciding which variables should (and should not) be conditioned on to avoid subtle confounding and selection biases, arising from study design or data collection methods. Recognising these biases is necessary so that unbiased estimates of causal effects may be obtained.

Despite their utility, causal inference methods, and in particular causal graphs, do have limitations. Unavoidably, expert knowledge is still required to elicit and defend causal assumptions, and it is recommended that sensitivity analyses be conducted to explore the consequences that departures from causal assumptions have on estimates of interest. Moreover, even in situation where causal assumptions may be well justified, correct specification of regression models remains an issue. These regression models may be required to adequately block open paths. In Section 3.1, we saw that specification of regression models is especially difficult in genomic applications, where dimensionality reduction strategies are required to condition on high-dimensional genetic information. These strategies come with their own model validity assumptions, separate from the causal ones we have discussed.

We reiterate that causal graphs are not the only framework for representing causal assumptions and deriving statistical dependencies, and that this can be done within other causal frameworks, for example[30]. We hope this review may, however, contribute to the discourse of GWAS and MR analyses by allowing causal assumptions to be explicitly acknowledged and communicated in a transparent and intuitive manner. Finally, since causal graphs are common in the communication and development of novel analytical methods, we hope to have contributed to a better understanding of them, thus helping the adoption of new analytical methods in the future.

References

[1] Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., and Zondervan, K. T. Data quality control in genetic case-control association studies. *Nature Protocols 5*, 9 (2010), 1564–1573.

[2] Bareinboim, E., Tian, J., and Pearl, J. Recovering from Selection Bias in Causal and Statistical Inference Elias. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (2014), AAAI Press, pp. 2410 – 2416.

[3] Bowden, J., Smith, G. D., and Burgess, S. Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. *International Journal of Epidemiology 44*, 2 (2015), 512–525.

[4] Brady, H. E. *Oxford Handbooks Online Causation and Explanation in Social Science 1 Causality*. No. April 2017. 2013.

[5] Browning, B. L., and Browning, S. R. A fast, powerful method for detecting identity by descent. *American Journal of Human Genetics 88*, 2 (2011), 173–182.

[6] Buniello, A., Macarthur, J. A., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., Flicek, P., Burdett, T., Hindorff, L. A., Cunningham, F., and Parkinson, H. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research 47*, D1 (2019), D1005–D1012.

[7] Burgess, S., and Thompson, S. G. *Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation*. Chapman & Hall/CRC, 2015.

[8] Chuang, Y. F., Tanaka, T., Beason-Held, L. L., An, Y., Terracciano, A., Sutin, A. R., Kraut, M., Singleton, A. B., Resnick, S. M., and Thambisetty, M. FTO genotype and aging: Pleiotropic longitudinal effects on adiposity, brain function, impulsivity and diet. *Molecular Psychiatry 20*, 1 (2015), 133–139.

[9] Didelez, V., Meng, S., and Sheehan, N. A. Assumptions of IV methods for observational epidemiology. *Statistical Science 25*, 1 (2010), 22–40.

[10] Didelez, V., and Sheehan, N. A. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research 16*, 4 (2007), 309–330.

[11] Ding, E. L., Song, Y., Manson, J. E., Hunter, D. J., Lee, C. C., Rifai, N., Buring, J. E., Gaziano, J. M., and Liu, S. Sex HormoneBinding Globulin and Risk of Type 2 Diabetes in Women and Men. *New England Journal of Medicine 361*, 12 (sep 2009), 1152–1163.

[12] Ganeff, I. M. M., Bos, M. M., van Heemst, D., and Noordam, R. BMI-associated gene variants in FTO and cardiometabolic and brain disease: obesity or pleiotropy? . *Physiological Genomics 51*, 8 (2019), 311–322.

[13] Glymour, M. M. Using causal diagrams to understand common problems in social epidemiology. In *In Methods in Social Epidemiology* (2006), John Wiley and Sons, pp. 393–428.

[14] Glymour, M. M., and Spiegelman, D. Evaluating public health interventions: 5. Causal inference in public health research-do sex, race, and biological factors cause health outcomes? *American Journal of Public Health 107*, 1 (2017), 81–85.

[15] Glymour, M. M., Tchetgen, E. J., and Robins, J. M. Credible mendelian randomization studies: Approaches for evaluating the instrumental variable assumptions. *American Journal of Epidemiology 175*, 4 (2012), 332–339.

[16] Gudbjartsson, D. F., Walters, G. B., Thorleifsson, G., Stefansson, H., Halldorsson, B. V., Zusmanovich, P., Sulem, P., Thorlacius, S., Gylfason, A., Steinberg, S., Helgadottir, A., Ingason, A., Steinthorsdottir, V., Olafsdottir, E. J., Olafsdottir, G. H., Jonsson, T., Borch-Johnsen, K., Hansen, T., Andersen, G., Jorgensen, T., Pedersen, O., Aben, K. K., Witjes, J. A., Swinkels, D. W., Heijer, M. D., Franke, B., Verbeek, A. L., Becker, D. M., Yanek, L. R., Becker, L. C., Tryggvadottir, L., Rafnar, T., Gulcher, J., Kiemeney, L. A., Kong, A., Thorsteinsdottir, U., and Stefansson, K. Many sequence variants affecting diversity of adult human height. *Nature Genetics 40*, 5 (2008), 609–615.

[17] Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., Laurin, C., Burgess, S., Bowden, J., Langdon, R., Tan, V. Y., Yarmolinsky, J., Shihab, H. A., Timpson, N. J., Evans, D. M., Relton, C., Martin, R. M., Davey Smith, G., Gaunt, T. R., and Haycock, P. C. The mr-base platform supports systematic causal inference across the human phenome. *eLife 7* (may 2018), e34408.

[18] Hernan, M., and Robins, J. *Causal Inference: What If.* Boca Raton: Chapman and Hall/CRC, 2020.

[19] Hoffman, G. E. Correcting for Population Structure and Kinship Using the Linear Mixed Model: Theory and Extensions. *PLoS ONE 8*, 10 (oct 2013), e75707.

[20] Holland, P. W. Statistics and Causal Inference. *Journal of the American Statistical Association 81*, 396 (dec 1986), 945–960.

[21] Imbens, G. W. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *NBER Working Paper No.w26104* (2019).

[22] Kivimaki, M., Lawlor, D. A., Smith, G. D., Kumari, M., Donald, A., Britton, A., Casas,´ J. P., Shah, T., Brunner, E., Timpson, N. J., Halcox, J. P., Miller, M. A., Humphries, S. E., Deanfield, J., Marmot, M. G., and Hingorani, A. D. Does high C-reactive protein concentration increase atherosclerosis? The Whitehall II study. *PLoS ONE 3*, 8 (2008), 1–8.

[23] Lin, D. Y., and Zeng, D. Proper analysis of secondary phenotype data in case-control association studies. *Genetic Epidemiology 33*, 3 (apr 2009), 256–265.

[24] Lippert, C., Quon, G., Kang, E. Y., Kadie, C. M., Listgarten, J., and Heckerman, D. The benefits of selecting phenotype-specific variants for applications of mixed models in genomics. *Scientific Reports 3* (may 2013), 1815.

[25] Loos, R. J., Lindgren, C. M., Li, S., Wheeler, E., Hua Zhao, J., Prokopenko, I., Inouye, M., Freathy, R. M., Attwood, A. P., Beckmann, J. S., Berndt, S. I., Bergmann, S., Bennett, A. J., Bingham, S. A., Bochud, M., Brown, M., Cauchi, S., Connell, J. M., Cooper, C., Davey Smith, G., Day, I., Dina, C., De, S., Dermitzakis, E. T., Doney, A. S., Elliott, K. S., Elliott, P., Evans, D. M., Sadaf Farooqi, I., Froguel, P., Ghori, J., Groves, C. J., Gwilliam, R., Hadley, D., Hall, A. S., Hattersley, A. T., Hebebrand, J., Heid, I. M., Herrera, B., Hinney, A., Hunt, S. E., Jarvelin, M. R., Johnson, T., Jolley, J. D., Karpe, F., Keniry, A., Khaw, K. T., Luben, R. N., Mangino, M., Marchini, J., McArdle, W. L., McGinnis, R., Meyre, D., Munroe, P. B., Morris, A. D., Ness, A. R., Neville, M. J., Nica, A. C.,

Ong, K. K., O'Rahilly, S., Owen, K. R., Palmer, C. N., Papadakis, K., Potter, S., Pouta, A., Qi, L., Randall, J. C., Rayner, N. W., Ring, S. M., Sandhu, M. S., Scherag, A., Sims, M. A., Song, K., Soranzo, N., Speliotes, E. K., Syddall, H. E., Teichmann, S. A., Timpson, N. J., Tobias, J. H., Uda, M., Ganz Vogel, C. I., Wallace, C., Waterworth, D. M., Weedon, M. N., Willer, C. J., Wraight, V. L., Yuan, X., Zeggini, E., Hirschhorn, J. N., Strachan, D. P., Ouwehand, W. H., Caulfield, M. J., Samani, N. J., Frayling, T. M., Vollenweider, P., Waeber, G., Mooser, V., Deloukas, P., McCarthy, M. I., Wareham, N. J., Barroso, I., Jacobs, K. B., Chanock, S. J., Hayes, R. B., Lamina, C., Gieger, C., Illig, T., Meitinger, T., Wichmann, H. E., Kraft, P., Hankinson, S. E., Hunter, D. J., Hu, F. B., Lyon, H. N., Voight, B. F., Ridderstrale, M., Groop, L., Scheet, P., Sanna, S., Abecasis, G. R., Albai, G., Nagaraja, R., Schlessinger, D., Jackson, A. U., Tuomilehto, J., Collins, F. S., Boehnke, M., and Mohlke, K. L. Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nature Genetics 40*, 6 (2008), 768–775.

[26] Munafo, M. R., Tilling, K., Taylor, A. E., Evans, D. M., and Davey Smith, G. Collider scope: when selection bias can substantially influence observed associations. *International Journal of Epidemiology 47*, 1 (feb 2018), 226–235.

[27] Pearl, J. Fusion, propagation, and structuring in belief networks. *Artificial Intelligence 29*, 3 (sep 1986), 241–288.

[28] Pearl, J. Causal diagrams for empirical research. *Biometrika 82*, 4 (1995), 669–688.

[29] Pearl, J. *Causality: Models, Reasoning and Inference*. 2000.

[30] Robins, J. M. Data, design, and background knowledge in etiologic inference. *Epidemiology 12*, 3 (2001), 313–320.

[31] Rubin, D. B. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association 100*, 469 (2005), 322–331.

[32] Sheehan, N. A., and Didelez, V. Human Genetics Epidemiology, Genetic Epidemiology and Mendelian Randomisation: more need than ever to attend to detail? Epidemiology, Genetic Epidemiology and Mendelian Randomisation: more need than ever to attend to detail? *Human Genetics*, 0123456789 (2018).

[33] Song, X., Ionita-Laza, I., Liu, M., Reibman, J., and Wei, Y. A general and robust framework for secondary traits analysis. *Genetics 202*, 4 (2016), 1329–1343.

[34] Speed, D., and Balding, D. J. Relatedness in the post-genomic era : is it still useful ? *Nature Publishing Group*, November (2014), 1–12.

[35] Speed, D., Hemani, G., Johnson, M. R., and Balding, D. J. Improved heritability estimation from genome-wide SNPs. *American Journal of Human Genetics 91*, 6 (2012), 1011–1021.

[36] Spiller, W., Slichter, D., Bowden, J., and Davey Smith, G. Detecting and correcting for bias in Mendelian randomization analyses using Gene-by-Environment interactions. *International Journal of Epidemiology* (2018), 1–11.

[37] Swerdlow, D. I., Kuchenbaecker, K. B., Shah, S., Sofat, R., Holmes, M. V., White, J., Mindell, J. S., Kivimaki, M., Brunner, E. J., Whittaker, J. C., Casas, J. P., and Hingorani, A. D. Selecting instruments for Mendelian randomization in the wake of genome-wide association studies. *International Journal of Epidemiology 45*, 5 (2016), 1600–1616.

[38] Tchetgen Tchetgen, E. J. A general regression framework for a secondary outcome in case-control studies. *Biostatistics 15*, 1 (2014), 117–128.

[39] Tchetgen Tchetgen, E. J., Walter, S., Vansteelandt, S., Martinussen, T., and Glymour, M.
Instrumental Variable Estimation in a Survival Context. *Epidemiology 26*, 3 (may 2015), 402–410.

[40] Vandenbroucke, J. P., Broadbent, A., and Pearce, N. Causality and causal inference in epidemiology: The need for a pluralistic approach. *International Journal of Epidemiology 45*, 6 (2016), 1776–1786.

[41] VanderWeele, T. J., Tchetgen, E. J. T., Cornelis, M., and Kraft, P. Methodological challenges in mendelian randomization. *Epidemiology (Cambridge, Mass.) 25*, 3 (2014), 427.

[42] Vansteelandt, S., Dukes, O., and Martinussen, T. Survivor bias in Mendelian randomization analysis. *Biostatistics 19*, 4 (2018), 426–443.

[43] Vilhjalmsson, B. J., and Nordborg, M.´ The nature of confounding in genome-wide association studies. *Nature Reviews Genetics 14*, 1 (2013), 1–2.

[44] Weedon, M. N., Lettre, G., Freathy, R. M., M, C., Voight, B. F., Perry, J. R. B., Elliott, K. S., Guiducci, C., Shields, B., Zeggini, E., Lango, H., Lyssenko, V., Timpson, N. J., Burtt, N. P., Rayner, N. W., Ardlie, K., Tobias, J. H., Ness, A. R., and Ring, S. M. UKPMC Funders Group UKPMC Funders Group Author Manuscript A common variant of HMGA2 is associated with adult and childhood height in the general population. *October 39*, 10 (2011), 1245–1250.

[45] Willer, C. J., Sanna, S., Jackson, A. U., Scuteri, A., Bonnycastle, L. L., Clarke, R., Heath, S. C., Timpson, N. J., Najjar, S. S., Stringham, H. M., Strait, J., Duren, W. L., Maschio, A., Busonero, F., Mulas, A., Albai, G., Swift, A. J., Morken, M. A., Narisu, N., Bennett, D., Parish, S., Shen, H., Galan, P., Meneton, P., Hercberg, S., Zelenika, D., Chen, W. M., Li, Y., Scott, L. J., Scheet, P. A., Sundvall, J., Watanabe, R. M., Nagaraja, R., Ebrahim, S., Lawlor, D. A., Ben-Shlomo, Y., Davey-Smith, G., Shuldiner, A. R., Collins, R., Bergman, R. N., Uda, M., Tuomilehto, J., Cao, A., Collins, F. S., Lakatta, E., Lathrop, G. M., Boehnke, M., Schlessinger, D., Mohlke, K. L., and Abecasis, G. R. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature Genetics 40*, 2 (2008), 161–169.

[46] Windmeijer, F., Farbmacher, H., Davies, N., and Davey Smith, G. On the Use of the Lasso for Instrumental Variables Estimation with Some Invalid Instruments. *Journal of the American Statistical Association 1459* (2018).

[47] Zhou, X., and Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods 11*, 4 (apr 2014), 407–409.

List of Figures

that conditioning on selection to the cohort, S, which depends on an individual surviving to index time T_0, introduces associations between G and U which violate the IV exclusion assumption. [Page 11]

Appendix A Linear Mixed Models

Consider again the linear model in Eq.2. When the model parameters are estimated by OLS, one effectively makes no prior assumptions about the parameter values, other than that they are fixed to some true unknown value. Considering *P* as a random effect, however, we impose, in a Bayesian sense, a normally distributed prior for $\gamma \sim \mathcal{N}_p\left(0, \sigma_g^2 I_p\right)$, where I_p is a *p* by *p* identity matrix, $\sigma_g{}^2$ is a hyper parameter and $N_p(\mu, \Sigma)$ is a *p*-multivariate normal distribution with mean μ and variance Σ.

By making this prior assumption we arrive at a LMM, which may be written as a model for the full *n*-dimensional observed phenotype vector, *Y* . Here bold notation is used to refer to vector (or matrix) quantities with *n* entries (or rows), each representing a single individual in the cohort. Again I_n is the *n* by *n* identity matrix,

$$Y \sim \mathcal{N}_n\left(\alpha \boldsymbol{G} + \boldsymbol{E}\beta, \sigma_g^2 \boldsymbol{K} + \sigma_e^2 \boldsymbol{I_n}\right) \tag{3}$$

where $K = PP^{\mathrm{T}}$ and *P* is an *n* by *q* matrix where each row represents the vector of PCs for a particular individual. The *n* by *n* matrix, *K* is referred to as the genetic similarity matrix, since the entry K_{ij} is a measure of the genetic similarity between the i^{th} and j^{th} individuals in the cohort, obtained by comparing their PCs. In general one is not restricted to using PCs to define the genetic similarity matrix. In fact several different methods can be expressed by the LMM equation above, using different measures of genetic similarity [19].

Measures of Genetic Similarity

Methods for measuring genetic similarity may be broadly separated into two categories: Those related to the Principal Component Analysis (Principal Components like), and those where some biologically motivated measure of genetic similarity is made. We will refer to methods of the latter type as Identity By Descent like, since they often measure similarity by finding genetic regions which are thought to be identical by descent in two individuals. A brief overview of these approaches is provided below.

Principal Component like

In a conventional PC analysis, the variables from which PCs are constructed (in this case the SNP values) are standardised. Variations exist, however, in how the SNPs are selected, how they are weighted in the standardisation step, and how the resultant PCs are selected. These include:

1. Selection of which SNPs to use for PC analysis: It is possible to include all available SNPs, however, it has been suggested that only variants thought to be causally related to the phenotype of interest should be included [43, 24], since these are the ones which lie on the causal pathway between *C* and *Y* . The process of selecting SNPs is known as pruning or thinning.
2. The choice of SNP dependent scaling constant before constructing PCs: The intuition behind scaling the SNP value is that sharing a rare variant is greater evidence of common ancestry than sharing a common variant. Scaling values are often estimates of the SNP standard deviation. This may be estimated by the sample standard deviation or using the standard deviation under the Hardy-Weinberg equilibrium model.

   It has also been suggested that, rather than pruning SNPs, SNPs should be weighted according to their degree of LD, to account for replication of causal information by neighbouring, imputed, SNPs in LD [35].

Their proposal uses weights, chosen such that SNPs with high LD are down-weighted. This is implemented in their LDAK software package.

3. The number of PC dimensions chosen for inclusion in the linear model: This is often determined using heuristic measures. Each successive PC accounts for a smaller amount of genetic variation in the chosen SNPs. Most methods use estimates for the proportion of variance explained by each PC, for example selecting PCs to exceed some threshold of the total proportion of variance explained, or else choosing an arbitrary number of PCs.

   In the LMM, it is possible to include all PCs. This is the choice made in the GEMMA software package [47]. This approach is equivalent to measuring the covariance between two individuals based on all chosen SNPs.

Identity By Descent like

Traditional measures for relatedness pre-date modern genomic study and were originally used to study trait inheritance within pedigrees. Using known pedigree information one can construct the probabilities that genomic regions of two individuals are identical-by-descent (IBD) from a recent common ancestor ('recent' in so far as it is assumed that there is no intermediate mutation or recombination event).

Pedigree based relatedness measures are broadly obsolete in modern genomic analysis for several reasons [34]: (i) When studying natural populations pedigree information is often unavailable or insufficient to account for population structure. (ii) Even when pedigree information is available, it is usually unrealistic to assume that pedigree founders have zero genetic similarity. (iii) The relatedness of any two individuals tends towards one, as the size of the pedigree is increased.

Rather than using pedigree information to estimate IBD probabilities, modern theories instead measure IBD by appealing to SNP data itself. These methods generally examine the length and frequencies of similar genomic regions in two individuals and are based on biochemical theories regarding the process by which gametes divide and recombine from two parents. Examples include: FastIBD [5], which estimates the frequencies of shared haplotype distributions; and shared segment detection in PLINK [1]. Reviewing these methods is beyond the scope of this review.

Table 1: All possible paths between three variables (*A*,*B*,*C*), with a brief description and additional terminology for the intermediate variable *C*

Path	Description	Terminology for the variable *C*
$A \rightarrow C \rightarrow B$	*A* causes *B* (through *C*)	Mediator
$A \leftarrow C \leftarrow B$	*B* causes *A* (through *C*)	Mediator
$A \leftarrow C \rightarrow B$	*A* and *B* share a common cause *C*	Confounder
$A \rightarrow C \leftarrow B$	*A* and *B* both cause *C*	Collider

Table 2: Summary of the rules of d-separation for all possible paths containing three variables. The two additional columns describe the statistical dependence of A and B before and after conditioning on the intermediate variable C.

Path	Before conditioning on C	After conditioning on C
$A \rightarrow C \rightarrow B$	open	closed
$A \leftarrow C \leftarrow B$	open	closed
$A \leftarrow C \rightarrow B$	open	closed
$A \rightarrow C \leftarrow B$	closed	open

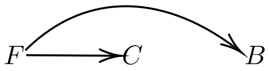

Fig.1 Causal graph representing the causal assumptions between a patients FTO gene variant, F, body mass index, B, and cerebral blood flow, C.

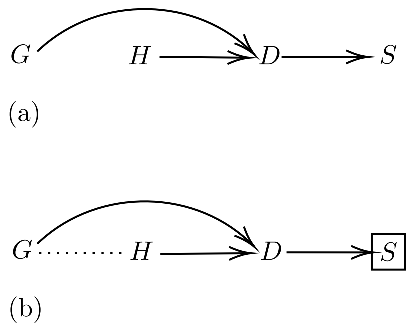

Fig.2 (a) Causal graph representing the causal assumptions between a specific gene of interest, G, Type 2 diabetes status, D, SHBG level, H, and selection to the cohort, S. (b) Causal graph when considering only individuals in the cohort (S = 1). The selection variable has been conditioned on, indicated by the box around it. The induced association between G and H is represented by the dashed line.

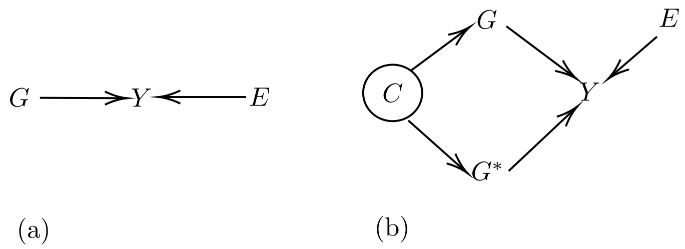

Fig.3 Causal graphs for GWAS analysis. Graph (a) shows the basic causal GWAS model, where the phenotype of interest, Y , is dependent on the gene of interest, G, and some other environmental factors, E. Graph (b) accounts for confounding by the ancestry of the individual, C, which affects the gene of interest, and the remaining genes, G^*. This modified graph assumes that a polygenic trait, Y , depends on both the gene of interest, and the remaining genes. By convention, unobserved (or latent) variables, such as the ancestry variable, C, are circled.

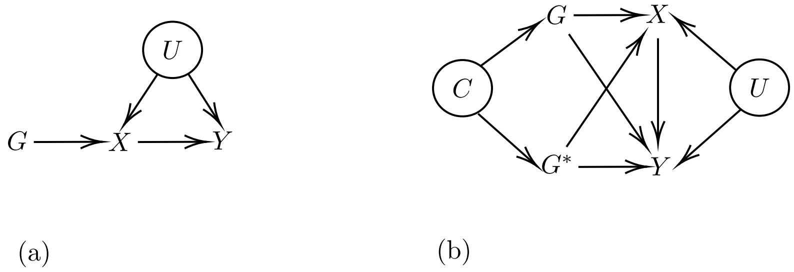

Fig.4 Causal graphs for MR analysis. Graph (a) shows the traditional IV causal graph, where the gene, G, acts as an IV for the $X \rightarrow Y$ relationship of interest, itself confounded by the unmeasured variable, U. Graph (b) shows modifications to graph (a) which relax assumptions by allowing for confounding by ancestry, and some pleiotropic effects.

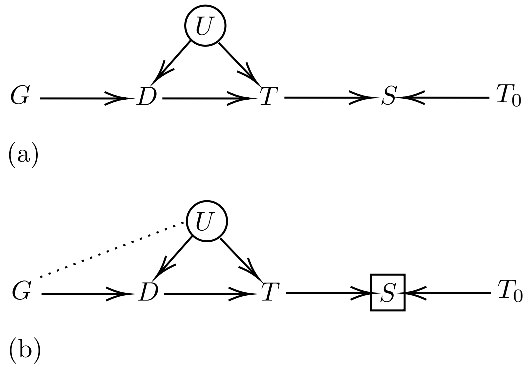

Fig.5 Causal graphs for MR analysis of a survival outcome. Graph (a) shows the instrumental gene, G, acts as an IV for the $\mathrm{D} \rightarrow \mathrm{T}$ relationship of interest where D is vitamin D level and T is lifetime. Graph (b), however, shows that conditioning on selection to the cohort, S, which depends on an individual surviving to index time T_0, introduces associations between G and U which violate the IV exclusion assumption.