UNIVERSITY OF
CAMBRIDGE

Mendelian Randomisation

This dissertation is submitted
for the degree of Master of Philosophy
in Statistical Science

by

Panayiota Constantinou

June 25, 2009

Girton College, University of Cambridge

Supervisor: Professor A. P. Dawid

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of the work done in collaboration except where specifically indicated in the text.

acknowledgements

I would like to thank my supervisor Prof. A.P.Dawid for his continuous help and guidance on this project throughout the year. Also, I would like to thank Dr. S.Pitts for her vital encouragement and support and Dr. V.Didelez for the useful indications.
Special thanks to my family and friends.

ABSTRACT

Mendelian Randomisation is a method used to test the hypothesis that a phenotype is causal for a disease in the presence of confounding. This theory has been developed for cases when controlled experiments are not available due to ethical, financial or other reasons. Using a gene that is strongly associated with the phenotype of interest, we can reject the explanation that an observed association between the phenotype and the disease is due to reverse causation or confounding. This approach is known in econometrics literature as an instrumental variable approach and has been developed mostly for continuous variables. In epidemiology we deal with a binary disease variable and categorical genotype variable, therefore some complications occur. Establishing causation we can therefore investigate how modifying the phenotype we can reduce the proportion of the disease in the general population. Using formal mathematical language we distinguish notions that are been confused and in this way identifying which causal parameters can actually be estimated from the data. Case-control studies introduce additional difficulties due to the fact that information is collected conditionally on the disease status.

Mendelian Randomisation

June 25, 2009

Contents

1 Introduction

It is often the case that observed associations between variables are misinterpreted as causal. Especially in epidemiology this is an important issue as variables take the form of phenotypes/exposures and diseases. Phenotypes are traits or characteristics of an individual that can be observed/measured (like blood pressure, triglycerides levels) and it is of great importance to assess whether particular phenotypes or exposures to particular environments (like smoking, increased alcohol consumption) actually cause a disease. If we can establish causation, then we want to investigate how controlling for the phenotype/exposure we can prevent the onset of the disease. This project concentrates on the notion of causality and uses some strong arguments resulting from biology structures to justify some of the assumptions.

In a simple observed correlation between a certain phenotype and a disease status, it is possible to argue in totally different directions. It could be that the phenotype causes the disease or the disease causes the phenotype (reverse causation). It could also be that other observed factors or unobserved factors (confounding) cause both variables to change or even that the observed association is just a coincidence. Maybe time has an effect to both variables or both variables are causal to each other under different circumstances. For example, if we assume an observed association between high blood pressure and increased risk of Coronary Heart Disease (CHD), one could argue that it is high blood pressure that causes CHD to occur, CHD at early stages raises blood pressure or other factors like smoking, bad dietary habits, no exercise etc cause the alteration in both variables.

As this topic gains an increasing interest, especially if the goal is to assess the effect of a modifiable phenotype/exposure on a disease status, different approaches have been suggested. One of them is the instrumental variable approach, in other words 'Mendelian Randomisation'. The main idea behind this approach is the use of a third key variable (usually a gene) that is not affected by reverse causation or confounding to help us deduce causation. Assuming that we want to test whether intervening on variable A causes an alteration in variable B, if we could find a third variable C which is known to be strongly associated with A in a causal way and has an effect on B only through A then maybe we can argue that A is causal for B. The above idea has its origins in the mid-1980s in a reasoning made by Katan[8].

Katan was searching for the causal path between low serum cholesterol levels and tumour growth. There was observational evidence that the two variables were associated but it was not clear which one was the causal to the other. Some scientists argued that it was low serum cholesterol levels that caused increased risk of cancer, others that premature tumour in future cancer patients caused the lowering of serum cholesterol levels and others that other factors such as smoking or bad dietary habits affected both serum cholesterol levels and cancer. An additional observation that helped to unravel the confusion was that patients with a rare genetic disease called abetalipoproteinemia didn't get to have premature tumours even though their serum cholesterol levels were practically zero (because of the disease). This group of people was relatively small in the population to help with the investigations but this observation led Katan to find a larger group of individuals which had low serum cholesterol levels from birth. He counted in his research another variable: a gene called ApoE which was known to be connected with serum cholesterol levels. More specifically, the ApoE2 allele of the gene is related to low cholesterol levels and cholesterol levels increase from the ApoE2 to the ApoE3 to the ApoE4 allele of the gene. He reasoned that many individuals will carry one or two copies of the ApoE2 gene and therefore have low serum cholesterol levels from birth. The idea in which the name Mendelian Randomisation refers to is that the specific allelic combination for a certain gene is random according to Mendel's Second Law (”Independent Assortment”). So since genetic variants are inherited independently of any confounding factors, individuals assigned the ApoE2 variant will not differ systematically from individuals that have been assigned other ApoE alleles (ApoE3, ApoE4). Also, since the ApoE genotype is present since birth, it cannot be affected by disease. In this way he could control for confounding and reverse causation. So examining two groups: patients with cancer and controls, he argued that under the causal hypothesis that low serum cholesterol levels is causal for cancer he would expect to see more individuals with ApoE2 alleles than ApoE3 and ApoE4 alleles in the patients group. Otherwise, if the causal hypothesis didn't hold then the number of individuals with the ApoE2 alleles in the patients and controls group should be similar. The actual case-control study never took place but the idea of converting a cholesterol-cancer test to a gene(ApoE)-cancer test is the one used by the Mendelian Randomisation method.

The above reasoning is known in econometrics literature as an instrumental variable approach. This name reflects the idea of using a gene as an instrument to decide if two variables are associated in a causal way. However, some problems occur when adapting techniques from econometrics since they are mostly about continuous variables and in epidemiology the disease variable is only binary. Also the extend to which this theory can be applied depends on the method used to collect data. Not the same conclusions derive when we examine data from prospective and retrospective view.

In the next sections, the theory of Mendelian Randomisation will be presented. In section 2, we define terms and explain notions that will be used throughout the context and give a brief introduction to the theory of this method. In section 3, we discuss what makes a gene suitable to be used as an instrumental variable that will help us deduce a causal effect and in section 4 we discuss how we measure and estimate causality once we establish its presence. As more assumptions will be made in order to justify the mathematical conclusions, it is going to be more difficult to satisfy them in practice. An attempt to apply the theory in two different datasets will be made in section 5 using R software.

2 Formal language for causality

2.1 Correlation is not causation

As mentioned in the introduction even if two variables are related or associated this does not imply that one causes the other. In other words, when examining two variables A and B, if we observe that as variable A increases also variable B increases, we shouldn't straightforward conclude that A causes the alteration in B. It could be the reverse reasoning (reverse causation) that B causes the alteration in A or other factors causing the increase in both variables (confounding). Observed association is a hint that might lead to the conclusion of causation but the immediate cause and effect conclusion is a fallacy and deserves more consideration. There is more than one explanation to this observation as discussed in the introduction and to establish causation we need additional strong evidence. Using the instrumental variable technique (as this will be presented below), we can adjust for reverse causation and confounding. For this, we will need to use the knowledge of underlying biology which will allow us to conclude a causal pathway.

2.2 Intervention[6]

The need of establishing causality derived from the need to explore how intervening on the causal variable affects the effect variable. In this context causal variables take the form of exposures (smoking, alcohol consumption etc) and phenotypes (high blood pressure, fibrinogen levels etc) and effect variables take the form of a disease outcome (CHD, cancer). Supposing that we can somehow intervene on the exposure/phenotype and change its value we want to know if in this way we can change the disease status. Public health needs to know if prescribing an intervention on the phenotype/exposure will lead to a dicrease of the disease proportion in the whole population. For example, if we force homocysteine levels to take a specific value, we want to assess the change (if any) in the disease associated with this phenotype (stroke). Since folate consumption is known to reduce homocysteine levels, it is of public health interest to know if it also reduces the chances of a stroke. But to deduce such a conclusion we firstly need to establish causation.

The issue of intervention though is subtle. By intervention in a variable we mean change its status and force it to take a specific value without affecting the status of other variables that might be related with it. This however, is a very strong assumption for which we have no evidence of being true in practice. What is important to distinguish is that the effect of intervening on a variable is not the same as if this variable had naturally taken the same value. So the result on the risk of stroke if we reduce an individual's homocysteine levels through folate consumption is not necessarily the same as if the individual had naturally low homocysteine levels. In order to be more accurate, the two notions are distinguished using different notation. Let X denote the exposure/phenotype and Y denote the disease status. Then P(Y|X=x) will be used to reflect the conditional distribution of Y given X and P(Y|do(X=x))[9] will be used to reflect the distribution of Y under intervention on X.

2.3 Confounding

When we want to test the effect of an exposure on a disease, we need to control for other possible factors that affect indirectly the disease. It is reasonable though never to be sure that we have identified and adjusted for all possible influencing factors. Especially when these factors are social, behavioural or physiological there is not an ideal way neither to find a measurement scale nor to control for this measurement scale. These unobserved factors that might not have a direct influence on the disease but an indirect influence are called confounding factors. A confounder is associated with both the probable causal variable and the effect variable. By definition the confounding factor is not the cause of the disease variable nor the effect caused by the phenotype variable but it may be responsible for at least some of the association under investigation. Confounding factors could influence other factors that are causal for the disease such as the phenotype variable. Confounding is the reason why studies on the same set of exposures and diseases produce sometimes totally different results.

2.4 Intervention, Confounding and Mendelian Randomisation

Sometimes intervening in variables is just not possible. The reasons might be practical or financial but more often ethical. If we want to assess the effect of alcohol consumption on blood pressure, we just can't expose individuals to different intake of alcohol and measure its effect. In these cases randomised control trials (RCTs) are not feasible. Also confounding may not have been fully understood and therefore not adjusted for.

Mendelian Randomisation suggests an alternative way to explore data when (RCTs) are not possible because of issues of intervention and confounding. To test for a causal effect of a phenotype on a disease we use a gene as an instrument. The alleles of the genes, according to Mendel's second law, are assigned randomly to individuals during a biological procedure called 'meiosis'. This allocation is independent of any confounding factors. If we establish from biology knowledge that a certain gene has an effect on a disease only through the phenotype of interest then we can check for association between the gene and the disease. If a correlation is observed then it cannot be due to reverse causation. The disease has no effect to the allocation of alleles so it has to be the gene through the phenotype that has an effect on the disease. These reasoning will be justified better in subsection 3.1.

2.5 Causal effect and parameters[3]

The term causal effect refers to a function of the distribution of Y under intervention in X, i.e. a function of P(Y|do(X=x)). Since the original target is to estimate how intervening on X has an effect on Y different functions have been suggested.

1. Average Causal Effect (ACE)

   $$ACE(x_1, x_2) = E(Y|do(X = x_2)) - E(Y|do(X = x_1)) \tag{1}$$

   ACE is defined as the difference in expectation of Y conditioning on different settings of X. The meaning of $E(Y|do(X = x))$, reflects the value we expect to see in Y when we force X to take value x. If the ACE is non zero for different x_1 and x_2, then we say that we have a (non-zero) causal effect of X on Y.

2. Causal Risk Ratio (CRR)

   $$CRR(x_1, x_2) = \frac{P(Y = 1|do(X = x_2))}{P(Y = 1|do(X = x_1))} \tag{2}$$

3. Causal Odds Ratio (COR)

   $$COR(x_1, x_2) = \frac{P(Y = 1|do(X = x_2))P(Y = 0|do(X = x_1))}{P(Y = 0|do(X = x_2))P(Y = 1|do(X = x_1))} \tag{3}$$

The above measures of causality sometimes under additional assumptions can be summarised by one parameter. For example,

$$\text{If } E(Y|do(X = x)) = \beta_1 + \beta_2 x \text{ (linearity)} \Rightarrow ACE(x_1, x_2) = \beta_2(x_2 - x_1)$$

In this case the ACE can be summarised by parameter β_2. So if we can estimate this parameter from the data then we have a measurement of the causal effect. When $E(Y|do(X = x))$ is exponential or log-linear in X it gets more difficult to find quantities of causality that can be summarised by one parameter due to be estimated. This will be discussed further in section 4.2.2

The ACE as a measure of causality is used in cases when $E(Y|do(X = x))$ is linear in X. Generally, when Y is a continuous variable the ACE is the quantity of interest (this is mostly the case in econometrics). When variable Y is binary,

$$
\text{Y} = \begin{cases} 0 & \text{if individual doesn't have the disease} \\ 1 & \text{if individual has the disease} \end{cases}
$$

researches mostly target to estimate the CRR or the COR. In cases of a rare disease when we assume that $P(Y = 0|X = x) = P(Y = 0|do(X = x)) \approx 1$, the COR gives an approximation of the CRR. This approximation turns to be helpful when we have to deal with a case-control study. In a case-control study, we collect information from individuals conditionally on their disease status. So, the only quantities that can be easily estimated are those that involve terms of conditional probabilities on the disease status. Using the assumption that $P(Y = y|X = x) = P(Y = y|do(X = x))$ and Bayes theorem, we get that

$$
\begin{aligned}
COR(x_1, x_2) &= \frac{P(Y = 1|do(X = x_2))P(Y = 0|do(X = x_1))}{P(Y = 0|do(X = x_2))P(Y = 1|do(X = x_1))} \\
&= \frac{P(Y = 1|X = x_2)P(Y = 0|X = x_1)}{P(Y = 0|X = x_2)P(Y = 1|X = x_1)} \\
&= \frac{P(X = x_2|Y = 1)P(X = x_1|Y = 0)}{P(X = x_2|Y = 0)P(X = x_1|Y = 1)}
\end{aligned}
$$

Writing the COR in the above form we can estimate it using observed frequencies from the dataset. When quantities of interest can't be written in an analogous form, it is not easy to estimate them from a case-control dataset. This complication leads to the definition of identifiability below.

2.6 Identifiable causal parameters

Datasets are collected in different ways according to the experimental designs. We could have RCT (random allocation of treatment to patients), case-control study (for every patient there is one-to-one matching for age, gender and geographical origin) etc. Since we assume that intervention is not feasible and we can not allocate different settings of X to individuals to measure the effect in Y, we usually have a case-control study. Modelling the design (for example making assumptions that relate the distributions of the variables like linearity) we aim to estimate the parameters involved using the information from the data. It is possible though, that the data we hold don't contain enough information to allow the estimation of all the parameters. For example, when we want to estimate the intervention effect from case-control study data that don't include information about intervention. In such a case, if we try to estimate $p(y|do(X = x))$ from $p(y|X = x)$ we might end up with wrong results because of confounding. We call identifiable parameters those parameters that can be estimated from the data with some degree of certainty.

3 Using Instrumental Variables to test for a causal effect

The idea of Mendelian Randomisation is to use a genotype as an instrument to test for a causal effect. In other words, replace a test between the phenotype and the disease with a test between the genotype and the disease. This idea may sound reasonable if we have a gene which is strongly related with the phenotype because in this way we control for confounding and reverse causation but has to be justified with more accuracy. We need to establish that the phenotype is causal for the disease and not the genotype alone. In order to do so, we use formal mathematical language to express the relationships between the variables involved. The first step is to set some conditions that will allow us to determine whether a genotype is proper to be used as an instrument. These conditions (core conditions) express conditional dependencies and independencies between the variables of interest (disease, phenotype, genotype, confounding). Once we set these conditions we can derive an expression for the joint distribution of the variables using Directed Acyclic Graphs.

The following notation will be used throughout this context.
X will represent the phenotype or exposure variable.
Y will represent the disease variabe (usually the binary disease outcome).
G will represent the instrumental variable (in our case the gene).
U will represent the confounding variable (observed/unobserved factors related with the phenotype/exposure and the disease).

Also for random variables A, B and C conditional dependencies and independencies will be expressed using $\perp\!\!\!\perp$ and $\not\perp\!\!\!\perp$ [2]where
$A \perp\!\!\!\perp B \mid C$ means A is conditionally independent of B given C and
$A \not\perp\!\!\!\perp B \mid C$ means A is not conditionally independent of B given C.

Using the marginal distributions we will see whether $Y \perp\!\!\!\perp X \mid U \iff \text{Y} \perp\!\!\!\perp G$. The latter when obtained will actually allow us to conclude that a test between the phenotype and the disease is formally equivalent to a test between the genotype and the disease. These conditions are only sufficient to allow us to test for the presence of a causal effect. If we establish a causal effect then we need more assumptions to be able to measure it.

3.1 Core Conditions[1, 4, 5, 7]

With the above notation the core conditions are as follow

1. Core Condition 1 (CC1)
   $G \perp\!\!\!\perp U$
   the genotype (G) must be independent of any confounding (C)

2. Core Condition 2 (CC2)
   $G \not\perp\!\!\!\perp X$
   the genotype (G) must not be independent of the exposure/phenotype (X)

3. Core Condition 3 (CC3)
   $Y \perp\!\!\!\perp G \mid (X, U)$
   the disease (Y) must be conditionally independent of the genotype (G) given the exposure/phenotype (X) and the confounding(U)

Comments

Core Condition 1: Since confounding is the main reason why we get spurious results, we want to establish that it doesn't affect the instrument we use. As we are talking about unobserved and not fully understood confounding factors, this condition cannot be statistically tested. This condition has to be justified from biology knowledge. This is why genetic variants that are proposed to be used as instruments, are variants that are well-studied and their functionality is well-understood. To justify this condition we need to know that the assignment of the alleles of the genes is independent of all confounding factors. For this conditions we use Mendel's second law that states that the allocation of the genes is independent of any confounding factors.

Core Condition 2: Not all genotypes are suitable to be used as instruments. We need genotypes that are strongly related with the exposure/phenotype. For example in Katan's example, the ApoE gene was known to be strongly connected with serum cholesterol and the ApoE2 variant to result in low serum cholesterol levels. This condition as it doesn't involve unobserved confounding is a condition that can be statistically tested. Any statistical test (eg.chi square test) that is suitable to show some form of dependency or independency will do. We will see below that even when the gene is not directly associated with the phenotype but through another gene we still get the same results concerning the test for the causal effect. But the stronger the association is the better the instrument and good instruments are more useful when we want to measure a present causal effect.

Core condition 3: For known exposure/phenotype status and confounding, the genotype must be independent of the disease outcome. Assuming that testing the dependency of the genotype and the disease outcome we conclude that the two variables are dependent, one could argue that it is the genotype that is causal for the disease and not the phenotype. This condition is here to indicate that if we know the phenotype status and the confounder, then the disease outcome is not dependent of the phenotype, i.e. that the genotype affects the disease only through the phenotype. Regarding

Katan's example, if we had information on the cholesterol levels and the confounding factors then the genetic allocation of the alleles shouldn't give any additional information for the outcome of the disease. A subtle point needed to be stressed is that the disease is not independent of the genotype given only the exposure/phenotype status (i.e. $Y \not\perp\!\!\!\perp G \mid X$). We need to condition also on confounding. Again this is an assumption we cannot statistically test as it involves confounding and has to be justified from biology.

Only from the core conditions we can see how easy it is to produce biased results. If this conditions are not adequately justified we might misconclude causation. The underlying knowledge of biomedicine should be sufficient at least for the core conditions. However, if these conditions hold we can straighforward conclude a causal effect or not of the exposure/phenotype X on the disease Y by applying an association test between the genotype G and the disease Y as discussed below.

3.2 Directed Acyclic Graphs (DAGs[1])

In an attempt to explain the conditional independent relationships of the variables involved (X, Y, G, U) we use Directed Acyclic Graphs (DAGs). We call graph $D = (V, E)$ a set V of nodes and a set $E \subseteq V \times V$ of edges. An edge is a link between two nodes. When we say directed we mean that the edges have directions (given by arrows) and when we say acyclic we mean that the directions of the arrows do not create a cycle from one node back to itself. In a DAG when we see arrows connecting nodes we can express the connections with a certain language. When we see $A \longrightarrow B$ we say that A is a parent of B and B is a child of A. We denote the set of parents of node v by $pa(v)$ and the set of children of node v by $ch(v)$. When we see $A \longrightarrow \ldots \longrightarrow B$ we say A is an anscestor of B and B is a descendant of A and when we see that $A \longrightarrow B$ and $C \longrightarrow B$ we say that A and C are co-parents of B. For example, in the DAG presented in Figure 1, Y is a child of X and X is a parent of Y, G is an anscestor of Y and Y is a descentant of G and G and U are co-parents of X. When a node is not a descendant of its own then no cycle occurs.

The nodes in the following graph, see Figure 1, will represent the variables of interest (X, Y, G, U) and the arrows that connect them the conditional dependencies and independencies as presented in core conditions. Generally for the ordered sequence of variables $\boldsymbol{X} = (X_1, \ldots, X_N)$ with joint distribution P, we can construct a DAG in the following way. Let $V = \{v : 1, \ldots, N\}$ denote the set of nodes corresponding to the variables $X_1, \ldots, X_N$. We define $pre(v) = (1, \ldots, v-1)$ (where $pre(1) = \emptyset$) and $X_{pre(v)} = (X_1, \ldots, X_{v-1})$. We introduce nodes in order 1 to N and for each node v we consider the conditional distribution of X_v given $X_{pre(v)}$. Then the set of parents of v can be defined as a subset of $pre(v)$ so as the conditional distribution of X_v depends only on $X_{pa(v)}$ ($X_v \perp\!\!\!\perp X_{pre(v)} | X_{pa(v)}$). Then we draw an arrow from each $w \in pa(v)$ to v. In our case, taking the variables of interest in the order (U, G, X, Y)

we have $pre(U) = \emptyset$, $pre(G) = U$, $pre(X) = (U, G)$ and $pre(Y) = (U, G, X)$. We name the nodes U, G, X, Y and consider them in this order. We have $X_{pre(G)} = U$ and $p(G|U) \overset{CC1}{=} p(G)$. The distribution of G doesn't depend on U, so U is not a parent of G and therefore we have no arrow from U to G. Then we consider $X_{pre(X)} = (U, G)$ and $p(X|U, G)$. The distribution of X depends both on U and G. Therefore, U and G are parents of X and we draw directed arrows from U to X and from G to X. Similarly, $X_{pre(Y)} = (U, G, X)$ and $p(Y|U, G, X) \overset{CC3}{=} p(Y|U, X)$. The distribution of Y depends on U and X so nodes U and X are parents of Y and we draw directed arrows from U to Y and from X to Y. We end up with the DAG presented in Figure 1.

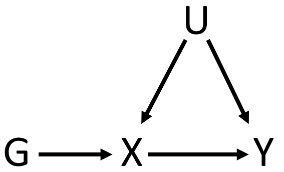


Figure 1: DAG representing Core Conditions

Representing the core conditions using a DAG we can decompose the joint distribution of X, Y, G, U using the fact that every node in the graph is conditionally independent of all other nodes given its parents, children and co-parents of its children. In other words, the joint distribution of X, Y, G and U can be expressed as:

$$\begin{aligned} p(x, y, g, u) &= \prod_{v \in DAG} p(v|pa(v)) \\ &= p(y|u, x)p(x|g, u)p(u)p(g) \end{aligned} \tag{4}$$

Using the above factorisation of the density we can see that in the general case [4, 5]

- $G \not\perp\!\!\!\perp U \mid X$ even if $G \perp\!\!\!\perp U$

$$\begin{aligned} p(g, u|x) &= \sum_y p(y, u, g|x) \\ &= \sum_y \frac{p(y|u, x)p(x|u, g)p(u)p(g)}{p(x)} \\ &= \frac{p(x|u, g)p(u)p(g)}{p(x)} \\ &\neq \frac{p(x|u)p(u)p(g|x)}{p(x)} \\ &= p(g|x)p(u|x) \end{aligned}$$

- $G \not\perp\!\!\!\perp U \mid Y$ even if $G \perp\!\!\!\perp U$

$$\begin{aligned}
p(g,u|y) &= \sum_x p(q,x,u|y) \\
&= \sum_x \frac{p(y|u,x)p(x|u,g)p(u)p(g)}{\sum_{x,u,g} p(y|u,x)p(x|u,g)p(u)p(g)} \\
&= \frac{p(u)p(g)\sum_x p(y|u,x)p(x|u,g)}{\sum_{u,g} p(u)p(g)\sum_x p(y|u,x)p(x|u,g)} \\
&\overset{1*}{=} \frac{p(u)p(g)p(y|u,g)}{\sum_{u,g} p(u)p(g)p(y|u,g)} \\
&= \frac{p(u)p(g)p(y|u,g)}{p(y)} \\
&\neq \frac{p(u)p(g|y)p(y|u)}{p(y)} \\
&= p(g|y)p(u|y)
\end{aligned}$$

$$\begin{aligned}
p(y|u,g) &= \sum_x p(x,y|u,g) \\
&= \sum_x \frac{p(y|u,x)p(x|u,g)p(u)p(g)}{p(u,g)} \\
&\overset{CC1}{=} \sum_x p(y|u,x)p(x|u,g)p(u)p(g) \qquad (1^*)
\end{aligned}$$

The above results indicate that when the data we hold are collected conditionally on a certain variable we need to be careful when concluding independency. Variables U and G that are marginally independent are not independent necessarily once we condition on another variable. Such data appear in a case-control study where all the information is collected conditionally on the disease status. Supposing that we could observe confounding, even though that the alleles allocation in individuals is independent of confounding, we would expect to see some dependency within the groups of patients and controls. However there could be cases where $p(g,u|x) = p(g|x)p(u|x)$ and $p(g,u|y) = p(g|y)p(u|y)$. It could be that for certain numeric figures the confounding and the genotype remain independent even when we condition on the phenotype or the disease.

Suppose that we could intervene on a variable and force it to take a specific value. For example, intervene on the phenotype X and $do(X = x_0)$. Then, we shouldn't expect that the joint distribution under intervention should be the same as the joint distribution without intervention. In general, it is not true that $p(y,g,u|x=x_0) = p(y,g,u|do(X=x_0))$. However, if we consider the DAG in Figure 1 and assume that intervening on the phenotype we only affect each node through its parents, i.e. $p(child|parent) = p(child|do(parent))$, then we say that the DAG has a causal interpretation with respect to intervention in X. More specifically, we assume that $p(y|x_0,u) = p(y|do(X=x_0),u)$ and the joint distribution under intervention becomes

$$p(y,g,u|do(X=x_0)) = p(y|u,x_0)p(u)p(g) \tag{5}$$

Graphically, we can represent intervention by removing the arrows leading to X from the parents G and U[9]. This case is shown in Figure 2.

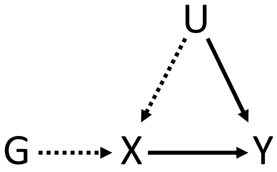


Figure 2: DAG representing Intervention

3.2.1 Intermediate Genotype[4, 5]

In CC2 we establish that G is not independent of X. It is not established though that the gene is causal for the phenotype. In fact, the latter doesn't have to be true. However if it is true, in practice it is easier to detect an association). In this subsection we will see that even if the genotype/phenotype association isn't causal and the association is through another gene that we may not consider in our test we can still derive the same marginal distribution for the gene we observe, the phenotype, the disease and confounding. We will again use another DAG that will help us derive the joint distribution. For this case, let G_1 represent the genotype of interest and let G_2 represent the intermediate genotype. The rest of the notation remains the same. The conditional idependencies expressing this case are

1. $(G_1, G_2) \perp\!\!\!\perp U$
   the genotypes G_1 and G_2 are independent of the confounder C

2. $X \perp\!\!\!\perp G_1 \mid (G_2, U)$
   the exposure/phenotype is independent of the genotype G_1 given the genotype G_2 and the confounder C

3. $Y \perp\!\!\!\perp (G_1, G_2) \mid (X, U)$
   the disease (Y) is conditionally independent of the genotype (G_1, G_2) given the exposure/phenotype (X) and the confounding(U)

We will construct the DAG representing these conditions in the way discribed above. Taking the variables of interest in the order (G_2, G_1, U, X, Y) we have $pre(G_2) = \emptyset$, $pre(G_1) = G_2, pre(U) = (G_1, G_2), pre(X) = (G_1, G_2, U)$ and $pre(Y) = (G_1, G_2, U, X)$. We name the nodes G_1, G_2, U, X, Y and consider them in this order. We have $X_{pre(G_1)} = G_2$ and $p(G_1|G_2)$. The distribution of G_1 depends on G_2, therefore G_2 is a parent of G_1 and we introduce a directed arrow from G_2 to G_1. $X_{pre(U)} = (G_1, G_2)$ and $p(U|(G_1, G_2)) = p(U)$. The distribution of U doesn't depend on (G_1, G_2) because of condition 1 above. Therefore, U has no parents. $X_{pre(X)} = (G_1, G_2, U)$ and $p(X|(G_1, G_2, U)) = p(X|(G_2, U))$. Thus, G_2 and U are parents of X and we draw directed arrows from G_2 to X and from U to X. Similarly, $X_{pre(Y)} = (G_1, G_2, U, X)$ and $p(Y|(G_1, G_2, U, X)) = p(Y|U, X)$ because of condition 2 above. The distribution of Y depends on U and X, so nodes U and X are parents of Y and we draw directed arrows from U to Y and from X to Y. We end up with the DAG presented in Figure 3.

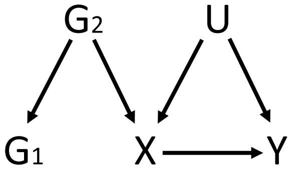


Figure 3: DAG representing Core Conditions

The joint distribution is now given by

$$\begin{aligned} p(x, y, g_1, g_2, u) &= \prod_{v \in DAG} p(v|parents(v)) \\ &= p(y|x, u)p(x|u, g_2)p(g_1|g_2)p(g_2)p(u) \end{aligned}$$

Considering again the matter of intervention in the phenotype X and assuming that $p(y|x_0, u) = p(y|do(X = x_0), u))$ we get the following expression for the joint distribution under intervention.

$$p(y, g_1, g_2, u|do(X = x_0)) = p(y|x_0, u)p(u)p(g_1|g_2)p(g_2) \tag{6}$$

Supposing we don't observe G_2, we want to derive the joint distribution under intervention for the remaining variables. We get that

$$\begin{aligned} p(y, u, g_1|do(X = x_0)) &= \sum_{g_2} p(y|x_0, u)p(u)p(g_1|g_2)p(g_2) \\ &= p(y|x_0, u)p(u)p(g_1) \end{aligned}$$

We see that we get the same marginal distribution as if when we only had one gene involved in the system. Investigating what would the marginal distribution be without intervention in X we also get the same joint distribution as if we only had genotype G_1.

$$\begin{aligned}
p(x, y, g_1, u) &= \sum_{g_2} p(x, y, g_1, g_2, u) \\
&= p(y|x, u)p(u) \sum_{g_2} p(x|u, g_2)p(g_1|g_2)p(g_2) \\
&= p(y|x, u)p(u)p(g_1) \sum_{g_2} p(x|u, g_2)p(g_2|g_1) \\
&\overset{2*}{=} p(y|x, u)p(x|u, g_1)p(g_1)p(u)
\end{aligned}$$

$$\begin{aligned}
p(x|u, g_1) &= \sum_{g_2} p(x|u, g_1, g_2)p(g_2|u, g_1) \\
&\overset{X \perp\!\!\!\perp G_1|(G_2,U)}{=} \sum_{g_2} p(x|u, g_2)p(g_2|u, g_1) \\
&\overset{(G_1,G_2) \perp\!\!\!\perp U}{=} \sum_{g_2} p(x|u, g_2)p(g_2|g_1) \qquad (2*)
\end{aligned}$$

This result allows us to use a 'weaker' instrument in our test. The above form of the joint distribution allows us to consider that the DAG in Figure 1, represents the general case. However, it is always possible that we don't detect a present causal effect of the phenotype on the disease if the gene is not strongly associated with the phenotype. If a gene is causal to a phenotype in practice we would expect to detect a strong association carrying out a statistical test. This is why we consider better instruments genes that are causally related to phenotypes of interest.

Now that we have defined the properties of an instrumental variable, we will discuss how we can use the instrument to test for the presence of a causal effect. As pointed many times above we cannot test for a causal effect only by an association test between variables X and Y mainly because of confounding and reverse causation.

3.3 Testing for the Presence of Causal Effect[4, 5]

The initial goal is to assess the changes in Y under different settings of X. To explore this case firstly we need to establish a causal effect. If there is just an association between X and Y and X is not causal for Y (for a known confounding background), then whatever the status of the phenotype is will not affect the outcome of the disease.

However, if we establish a causal effect of X on Y, so we know that the phenotype is causal for the disease, we have to be careful concluding that if *we* change X then we will cause a change in Y because of the subtle notion of intervention. We don't know how the disease outcome will change (if it will change) if we force the phenotype to take a specific value. In the following context, we will need to assume that the conditional distribution of Y under intervention on X given a certain confounding background is the same as the conditional distribution of Y when we just observe X under the same confounding background. In mathematical language, we assume that $p(y|x, u) = p(y|do(X = x), u)$. This is not necesarily true in practice, but it will help us derive some primary results.

So the first step is to test for a causal link. We can't use a test between X and Y because as mentioned earlier a possible association might appear because of confounding or reverse causation. We can't be sure that the phenotype is causal for the disease. To argue that we have a causal effect, the ACE defined in equation (1) must not be zero. This would imply that under different settings of the phenotype, we would expect a change in the disease outcome (this is what we would expect under the causal link hypothesis). Using the assumption that $p(y|x, u) = p(y|do(X = x), u)$, we will calculate the ACE. The initial goal of this section is to investigate whether $Y \perp\!\!\!\perp X \mid U$ is equivalent to $Y \perp\!\!\!\perp G$. In this case as analysed below, the test of causality between X and Y could be replaced by an association test between G and Y.

We use the equations that express joint distributions obtained above to extend equation (1).
We firstly obtain $p(y|do(X = x))$.

$$\begin{aligned} p(y|do(X = x_0)) &= \sum_{u,g} p(y, u, g|do(X = x_0)) \\ &\overset{(5)}{=} \sum_{u,g} p(y|u, x_0)p(u)p(g) \\ &= \sum_{u} p(y|u, x_0)p(u) \end{aligned} \tag{6}$$

Since U is not observed and we don't have any information on its distribution the above quantity cannot be estimated. When targeting the ACE we obtain

$$\begin{aligned} ACE(x_1, x_2) &= E(Y|do(X = x_2)) - E(Y|do(X = x_1)) \\ &\overset{*3}{=} \sum_{u} (E(Y|U = u, X = x_2) - E(Y|U = u, X = x_1))p(u) \end{aligned} \tag{7}$$

$$\begin{aligned} E(Y|do(X=x)) &= \sum_y yp(y|do(X=x)) \\ &\overset{(6)}{=} \sum_{u,y} yp(y|u,x)p(u) \\ &= \sum_u p(u)E(Y|U=u,X=x) \qquad (*3) \end{aligned}$$

From equation (7) we see that if $E(Y|U=u,X=x) = E(Y|U=u)$, or more strongly if $p(y|u,x) = p(y|u)$, then $ACE(x1,x2) = 0$. So if the disease status only depends on the confounder and not on the phenotype there is no causal effect. But if $ACE(x_1,x_2) = 0$ then we can't conclude that either $E(Y|U=u,X=x) = E(Y|U=u)$ or $p(y|u,x) = p(y|u)$. Only in an imaginary case (since the confounder is unobservable) where

$$\forall\ u \neq u_0,$$
$$E(Y|U=u,X=x_2) - E(Y|U=u,X=x_1) = E(Y|U=u_0,X=x_2) - E(Y|U=u_0,X=x_1)$$

then the zero ACE would actually imply that $E(Y|U,X)$ does not depend on X. However, in practice it is very difficult to obtain a zero causal effect when $E(Y|U,X)$ depends on X since this happens if there exists a very specific formation in the sum obtained in (7), so as to allow cancellations. It might be reasonable to assume that in reality a zero causal effect would actually imply that $E(Y|U,X)$ does not depend on X and therefore we could conclude that the phenotype is not causal for the disease (risky).

What would be ideal if we could prove is $Y \perp\!\!\!\perp X \mid U \iff \mathrm{Y} \perp\!\!\!\perp G$. This is actually Katan's reasoning and if it is formally true it would imply that an association test between G and Y could replace a test between X and Y. If we find that the genotype and the disease present some form of association (using appropriate tests such as chi-square test, t-test, odds-ratio depending on the distributions of the variables), then this would imply that $Y \not\perp\!\!\!\perp G$. From the above assumption, it follows that $Y \not\perp\!\!\!\perp X \mid U$. This means that given the confounder the disease(Y) does depend on the phenotype(X). Showing dependence through this reasoning, implies that the potential observed association between the phenotype and the disease is because of a causal interpretation. Otherwise, we wouldn't expect dependency between the genotype and the disease. So we expect the disease outcome to vary for different 'natural' values of X. With this reasoning, we conclude the presence of a causal effect. Now, if the appropriate test doesn't show any form of association between the genotype and the disease, we can conclude that $Y \perp\!\!\!\perp G$. Assuming that the above relationship is true, it follows that $Y \perp\!\!\!\perp X \mid U$. This means that the disease outcome is independent of the phenotype given a cerain confounder and in this case we can conclude the absence of a causal effect. So even by intervening on the phenotype X we don't expect anything to change in the disease status Y.

Even if the above assumption holds and we have established a causal effect of the phenotype on the disease, the effect of intervening would need more exploration. This will not be discussed in this project but it's a point that needs to be stressed. We need to be aware that intervening on the phenotype might produce the exact opposite results of the one's we are expecting or no results at all. There is no reason to be certain that the disease variable will change in the same way as if we didn't force the phenotype to change. The issue is that we don't know what is going to happen. Even in cases that we do know, other problems arise. In practice it is difficult to set a phenotype or an exposure to a specific value. For istance, extending the previous example with folate consumption, we don't know the exact amount of folate that an individual needs to consume so as their homocysteine levels to take a certain value. That varies in individuals and it is difficult to be estimated for the general case.

3.3.1 Prospective data

When data are not collected conditionally on a certain variable (such as in a case-control study where data are collected conditionally on the disease status) we can discuss whether the assumption $Y \perp\!\!\!\perp X \mid U \iff \mathrm{Y} \perp\!\!\!\perp G$ holds, using properties of the joint distribution $p(x, y, u, g)$.

It's easy to see that $Y \perp\!\!\!\perp X \mid U$ (i.e.$p(y|x, u) = p(y|u))$) $\Rightarrow$Y $\perp\!\!\!\perp G$.

$$\begin{aligned}
p(y, g) &= \sum_{u,x} p(x, y, u, g) \\
&= p(g) \sum_{u} p(u) \sum_{x} p(y|x, u)p(x|u, g) \\
&= p(g) \sum_{u} p(u)p(y|u) \\
&= p(g)p(u)
\end{aligned}$$

So if given the confounder the disease status and the phenotype are independent, then this implies that also the disease and the genotype are independent. This in practice would mean that if there is no causal link between the phenotype and the disease, then we expect to detect independency between the disease and the phenotype. The reverse argument however is not true, since there is a possibility that in the absence of independence numerical cancellations could induce a factorisation of $p(y, g)$. This means that if we detect independency between the genotype and the disease, applying an appropriate test, it is not mathematically true that also the phenotype and the disease (given the confounding) are independent. In practice though, this would be rare. In this reasoning we regard the above arguments as equivalent (could be risky).

What however is certainly true, is that if we detect dependency, meaning that we could argue that $Y \not\perp\!\!\!\perp G$, then this implies that also $Y \not\perp\!\!\!\perp X \mid U$. So if a test shows dependency between the genotype and the disease, then we can argue that there is also dependency between the phenotype and the disease (given the confounder). Assuming that $p(y|x,u) = p(y|do(X=x),u)$ we give this dependency a causal interpretation.

3.3.2 Retrospective data

The same reasoning holds in the case when we have retrospective data. In this case data are collected conditionally on the disease status so we have to use properties of $p(x,u,g|y)$.

$$\begin{aligned}
p(x,u,g|y) &= \frac{p(y|u,x)p(x|u,g)p(u)p(g)}{\sum_{x,u,g} p(y|u,x)p(x|u,g)p(u)p(g)} \\
&= \frac{p(y|u,x)p(x|u,g)p(u)p(g)}{\sum_{x,u} p(y|u,x)p(u)\sum_g p(x|u,g)p(g)} \\
&\overset{CC1}{=} \frac{p(y|u,x)p(x|u,g)p(u)p(g)}{\sum_{x,u} p(y|u,x)p(u)p(x|u)}
\end{aligned}$$

It can still be shown that $Y \perp\!\!\!\perp X \mid U \Rightarrow$Y $\perp\!\!\!\perp G$ from a retrospective view (i.e. $p(y|x,u) = p(y|u) \Rightarrow p(g|y) = p(g)$). Substituting $p(y|x,u) = p(y|u)$ in the equation above we get

$$\begin{aligned}
p(x,u,g|y) &= \frac{p(y|u)p(x|u,g)p(u)p(g)}{\sum_{x,u} p(y|u)p(u)p(x|u)} \\
&= \frac{p(y|u)p(x|u,g)p(u)p(g)}{\sum_{u} p(y|u)p(u)} \\
&= \frac{p(y|u)p(x|u,g)p(u)p(g)}{p(y)}
\end{aligned}$$

We now see that

$$\begin{aligned}
p(g|y) &= \sum_{x,u} p(x,g,u|y) \\
&= \frac{p(g)\sum_u p(y|u)p(u)\sum_x p(x|u,g)}{p(y)} \\
&= \frac{p(g)\sum_u p(y|u)p(u)}{p(y)} \\
&= p(g)
\end{aligned}$$

We can see again that the reverse does not hold but in practice it is considered safe to assume it holds (considering the same reasoning as in prospective view). This result allows us to conclude the presence or absence of a causal effect also when we have to analyse data from a case-control study. This is the type of data that will be analysed in the application section.

The variables in medical data usually have a certain form. The phenotype (X) is usually continuous and it takes values among a certain range of values. The disease (Y) is binary. It takes the value 0 if the individual doesn't have the disease and the value 1 if the individual has the disease. The genotype (G) is categorical. A simplistic explanation of the seperation of the genotype in 3 categories is as follows. Each individual inherits two genes from their parents. Assuming that a certain gene has two possible outcomes G_1 and G_2, then the child could have 4 possible genotypes G_1G_1, G_1G_2, G_2G_1 and G_2G_2. The cases G_1G_2 and G_2G_1 are considered the same so it is in fact 3 possible genotypes. So if the individual has genotype status G_1G_1, the genotype variable takes the value 0 ($G = 0$), if the individual has genotype G_1G_2 or G_2G_1 the genotype takes the value 1 ($G = 1$) and if the individual has genotype G_2G_2 the genotype variable takes the value 2 ($G = 2$). Since we establish that the genotype and the phenotype are associated (CC2), we expect at least one of the genes to have an effect on the phenotype. In the above simplistic seperation, we could think of G_2 as the one that affects the phenotype. With this reasoning, we expect individuals with genotype status $G = 0$, to present different phenotype status than individuals with genotype status $G = 1$ and individuals with genotype status $G = 1$ to present different phenotype status than individuals with genotype status $G = 2$ and the difference to increase as we move from genotype status $G = 0$ to $G = 1$ to $G = 2$. Using the above conclusion, that in practice we can consider $Y \perp\!\!\!\perp X \mid U \iff \mathrm{Y} \perp\!\!\!\perp G$, we can test for an association between the genotype and the disease. Because of the form of the genotype and disease variables, a suitable test to detect dependency or independency of the two variables is a χ^2 test.

Synopsising, assuming that $p(y|x, u) = p(y|do(X = x), u)$, what we have seen either from a prospective or a retrospective view is that if there is no causal effect ($Y \perp\!\!\!\perp X \mid U$) then a test between Y and G should show no association ($Y \perp\!\!\!\perp G$). Equivalently, if a test between Y and G shows association ($Y \not\perp\!\!\!\perp G$) then we can conclude that there is a causal effect ($Y \not\perp\!\!\!\perp X \mid U$). However, from a retrospective view as we will see in the next section problems occur when we try to estimate the causal effect.

4 Estimating the causal effect

4.1 Core Conditions and Bounds on the ACE [1, 10]

In the previous section, we have seen that a zero ACE will indicate no causal effect and a non zero ACE will indicate the presence of a causal effect. Since the value of the ACE is an indicator of causality, it is of interest to estimate it. In the case where X, Y, G are binary, it can be shown that the conditional probabilities of the disease when we intervene on the phenotype ($P(Y = y|do(X = x))$) and the ACE ($ACE(x_1, x_2) = P(Y = 1|do(X = 1) - P(Y = 1|do(X = 0))$ can be bounded. Let
$\omega_0 = P(Y = 1|do(X = 0)$
$\omega_1 = P(Y = 1|do(X = 1)$
$\phi_{yx.g} = P(Y = y, X = x|G = g)$ and
α_*,α^* represent the bounds of the ACE ($\alpha_* \leq ACE \leq \alpha^*$).
Then the following inequalities hold.

$$\omega_0 \leq \min \begin{cases} 1 - \phi_{00.0} \\ 1 - \phi_{00.1} \\ \phi_{01.0} + \phi_{10.0} + \phi_{10.1} + \phi_{11.1} \\ \phi_{10.0} + \phi_{11.0} + \phi_{01.1} + \phi_{10.1} \end{cases}$$

$$\omega_0 \geq \max \begin{cases} \phi_{10.1} \\ \phi_{10.0} \\ \phi_{10.0} + \phi_{11.0} - \phi_{00.1} - \phi_{11.1} \\ -\phi_{00.0} - \phi_{11.0} + \phi_{10.1} + \phi_{11.1} \end{cases}$$

$$\omega_1 \leq \min \begin{cases} 1 - \phi_{01.0} \\ 1 - \phi_{01.1} \\ \phi_{10.0} + \phi_{11.0} + \phi_{00.1} + \phi_{11.1} \\ \phi_{00.0} + \phi_{11.0} + \phi_{10.1} + \phi_{11.1} \end{cases}$$

$$\omega_1 \geq \max \begin{cases} \phi_{11.1} \\ \phi_{11.0} \\ -\phi_{01.0} - \phi_{10.0} + \phi_{10.1} + \phi_{11.1} \\ \phi_{10.0} + \phi_{11.0} - \phi_{01.1} - \phi_{10.1} \end{cases}$$

$$\alpha_* = \max \begin{cases} \phi_{00.0} + \phi_{11.1} - 1 \\ \phi_{11.0} + \phi_{00.1} - 1 \\ -\phi_{01.0} - \phi_{10.0} + \phi_{11.0} - \phi_{10.1} - \phi_{11.1} \\ -\phi_{10.0} - \phi_{11.0} - \phi_{01.1} - \phi_{10.1} + \phi_{11.1} \\ -\phi_{01.1} - \phi_{10.1} \\ -\phi_{01.0} - \phi_{10.0} \\ -\phi_{00.0} - \phi_{01.0} + \phi_{00.1} - \phi_{01.1} - \phi_{10.1} \\ \phi_{00.0} - \phi_{01.0} - \phi_{10.0} - \phi_{00.1} - \phi_{01.1} \end{cases}$$

$$\alpha^* = \min \begin{cases} 1 - \phi_{10.0} - \phi_{01.1} \\ 1 - \phi_{01.0} - \phi_{10.1} \\ \phi_{00.0} - \phi_{01.0} + \phi_{11.0} + \phi_{00.1} + \phi_{01.1} \\ \phi_{00.0} + \phi_{01.0} + \phi_{00.1} - \phi_{01.1} + \phi_{11.1} \\ \phi_{00.1} + \phi_{11.1} \\ \phi_{00.0} + \phi_{11.0} \\ \phi_{10.0} + \phi_{11.0} + \phi_{00.1} - \phi_{10.1} + \phi_{11.1} \\ \phi_{00.0} - \phi_{10.0} + \phi_{11.0} + \phi_{10.1} + \phi_{11.1} \end{cases}$$

These bounds are derived by only assuming core conditions and it can also be shown that they can't be improved if no additional assumptions are made. The same analysis could be extended generally for discrete variables. However, as the variables split in more possible values then the analysis becomes more complicated and the bounds are not as informative. The conditional probabilities used in the bounds ($\phi_{yx.z} = P(Y = y, X = x|G = g)$) can be estimated from the data using observational frequencies. They can be informative in the case where they collapse to a point estimate or give a narrow interval. Otherwise if they give an interval that includes zero then we can't be sure that the instrument is informative for the causal effect. Also if we compute upper bounds being smaller than lower bounds, this would imply that the core conditions are broken. In the case of Mendelian Randomisation, the disease status(Y) is indeed binary. However the phenotype status(X) is usually continuous and the instrument(G) status usually splits in three categories. What could be done, is dichotomise X with respect to a certain threshold, and also group the genotype outcomes in two categories. Then apply the above bounds and compare them for different tresholds and different groupings. For this to be valid though, we must check that core conditions hold for the dichotimised phenotype. In the case of retrospective data (case-control study) the bounds can't be applied directly, because what we can estimate from frequencies is conditional probabilities given the disease status($p(g, x|y)$). If we additionaly have the distribution of the disease($p(y)$) generally in the population then we can calculate the above bounds using

$$p(y, x|g) = p(g, x|y)\frac{p(y)}{p(g)}$$

where $p(g)$ is obtained by $p(g, x|y)$.

4.2 Additional Assumptions

Additional assumptions allow us to express the causal effect in parameters that can actually be estimated from observational data. The most common assumption regards intervening on the phenotype (usually that $E(Y|X = x, U = u) = E(Y|do(X = x), U))$. More assumptions could reflect the relationships between variables. When we target the ACE, assuming linear dependence of the disease on the phenotype and the confounder and linear dependence of the phenotype on the genotype and the confounder, results in an expression of the ACE that can be summarised by only one parameter which can be estimated from the data. However, the ACE is not the causal effect that we target when we deal with binary disease outcome. For binary disease outcome we target on causal parameters such as the CRR and the COR. We will see that even when we make additional assumptions it is still difficult to derive an expression that can be summarised by one identifiable parameter.

4.2.1 Linear model without interactions[4, 5]

The linear model (with respect to X) without interactions is the easiest case. Linearity allows us to use additivity of expectation to derive an expression with only one parameter due to be estimated. More specifically, the additional assumptions in this case are as follow.

1. $E(Y|X = x, U = u) = E(Y|do(X = x), U = u) = \gamma_1 + \gamma_2 x + \gamma_3(u)$

2. $E(X|G = g, U = u) = \delta_1 + \delta_2 g + \delta_3(u)$

The first assumption states that Y depends in a linear way through its mean on X and a function γ_3 of U without interactions. This is a reasonable assumption when we treat Y as a continuous variable. Additionally, we assume that we have the same dependence in the conditional expectation when we intervene on X. The second assumption states that X depends in a linear way through its mean on G and a function δ_3 of U without interactions. Both assumptions include the unobserved confounder and therefore can't be tested or be more specific for the form of γ_3 and δ_3. In this case the causal parameter of interest is the ACE. We have

$$\begin{aligned}
E(Y|do(X=x)) &= E_{U|do(X=x)}E(Y|do(X=x),U) \\
&= E_U E(Y|do(X=x),U) \\
&= E_U(\gamma_1+\gamma_2 x+\gamma_3(U)) \\
&= \gamma_1+\gamma_2 x+E_U(\gamma_3(U)) \\
&= c_1+\gamma_2 x
\end{aligned}$$

The second equality holds because of the intervening action. Since we force X to take a certain value, the change does not depend on confounding. Therefore,

$$\begin{aligned}
ACE(x1,x2) &= E(Y|do(X=x_2)) - E(Y|do(X=x_1)) \\
&= \gamma_2(x_2-x_1)
\end{aligned}$$

We now have an expression of the ACE that is summarised by one causal parameter (γ_2). Even if the functions γ_3 and δ_3 were the indicator function of U, still we wouldn't be able to estimate γ_2 from a linear regression of Y on X and U ($E(Y|X=x,U=u)=\gamma_1+\gamma_2 x+\gamma_3 u$) because U is unobserved. Also, we can't estimate it from a linear regession of Y on $do(X)$ ($E(Y|do(X=x))=c_1+\gamma_2 x$) because the datasets we have don't reflect the intervening effect. When considering estimating it from a linear regression of Y on X alone, we see that

$$\begin{aligned}
E(Y|X=x) &= E_{U|X=x}E(Y|X=x,U) \\
&= E_{U|X=x}(\gamma_1+\gamma_2 x+\gamma_3(U)) \\
&= \gamma_1+\gamma_2 x+E_{U|X=x}(\gamma_3(U))
\end{aligned}$$

The last term $E_{U|X=x}(\gamma_3(U))$ is generally not constant in x. So linear regression of Y on X alone does not allow identification of the causal parameter γ_2. However we could estimate it from a linear regression of Y on G and X on G. We see that

$$\begin{aligned}E(Y|G=g) &= E_{(X,U)|G=g}E(Y|X,U,G=g)\\ &= E_{U|G=g}E_{X|U,G=g}E(Y|X,U,G=g)\\ &\overset{CC3}{=} E_{U|G=g}E_{X|U,G=g}E(Y|X,U)\\ &\overset{CC1}{=} E_U E_{X|U,G=g}E(Y|X,U)\\ &= E_U E_{X|U,G=g}(\gamma_1+\gamma_2 X+\gamma_3(U))\\ &= E_U(\gamma_1+\gamma_2 E_{X|U,G=g}(X)+\gamma_3(U))\\ &= E_U(\gamma_1+\gamma_2(\delta_1+\delta_2 g+\delta_3(U))+\gamma_3(U))\\ &= c_2+\gamma_2\delta_2 g\end{aligned}$$

$$\begin{aligned}E(X|G=g) &= E_{U|G=g}E(X|G=g,U)\\ &\overset{CC1}{=} E_U E(X|G=g,U)\\ &= E_U(\delta_1+\delta_2 g+\delta_3(U))\\ &= c_3+\delta_2 g\end{aligned}$$

So, we can estimate γ_2 from the ratio of the estimators of $\gamma_2\delta_2$ and δ_2. Let $r_{Y|G}$ denote the estimator yield from a linear regression of Y on G alone and $r_{X|G}$ denote the estimator from a linear regression of X on G alone. Then

$$\hat{\gamma_2} = \frac{r_{Y|G}}{r_{X|G}}$$

Here we note that we expect $r_{X|G}$ to be non zero from CC2, since the genotype is supposed to be associated with the phenotype.

4.2.2 Non linear models

[4, 5] When we deal with binary Y, the linearity assumption is not reliable. The most common assumption in this case in epidimiological literature about $E(Y|X,U)$, are logistic regression and log-linear regression. Neither of these assumptions though, will easily summarise the ACE in one parameter. This is why we aim at different measures of causality as the COR and CRR.

Let's consider the case of a logistic regression. More specifically, we assume

$E(Y|X = x, U = u) = E(Y|do(X = x), U = u) = \frac{\exp{(\alpha+\beta_1 x+\beta_2 u)}}{1+\exp{(\alpha+\beta_1 x+\beta_2 u)}}$

When we aim at the $ACE(x1, x2) = E(Y|do(X = x_2)) - E(Y|do(X = x_1))$, we firstly need to work out $E(Y|do(X = x))$. We see that

$$\begin{aligned} E(Y|do(X = x)) &= E_{U|do(X=x)}E(Y|do(X = x), U) \\ &= E_U E(Y|do(X = x), U) \\ &= \int \frac{\exp{(\alpha + \beta_1 x + \beta_2 u)}}{1 + \exp{(\alpha + \beta_1 x + \beta_2 u)}} p(u)du \end{aligned}$$

The last equation can't easily result in an expression of the form

$$E(Y|do(X = x)) = \frac{\exp{(\alpha^* + \beta_1 x + \beta_2 u)}}{1 + \exp{(\alpha^* + \beta_1 x + \beta_2 u)}}$$

because we need to work out the distribution of p(u) which is not identiafiable. Neither it's trivial to find distributions of U that will give a form of $E(Y|do(X = x))$ which will summarise the ACE in one parameter. The same problems occur when we want to estimate COR and CRR.

Considering the case of a log-linear regression we can summarise the CRR in one parameter. But, still the causal parameter is not easily identifiable. In this case, we assume

1. $E(Y|X = x, U = u) = E(Y|do(X = x), U = u) = \exp{(\gamma_1 + \gamma_2 x + \gamma_3 u)}$

2. $E(X|G = g, U = u) = \delta_1 + \delta_2 g + \delta_3 u$

It is important here to stress that the first assumption might not be valid. We defined the disease variable to take values 0 and 1 according to the disease status but the above expression might give an expected value of the disease variable greater than unit. Again this is an assumption we can't statistically test since it involves unobserved confounding. Assuming though it holds, we continue by working out $E(Y|do(X = x))$.

$$\begin{aligned} E(Y|do(X = x)) &= E_{U|do(X=x)}E(Y|do(X = x), U) \\ &= E_U E(Y|do(X = x), U) \\ &= E_U \exp{(\gamma_1 + \gamma_2 x + \gamma_3 U)} \\ &= \exp{(\gamma_1 + \gamma_2 x)} \int \exp{(\gamma_3 u)} p(u)du \\ &= \exp{(\gamma_1^* + \gamma_2 x)} \end{aligned}$$

In the case of binary Y, we have $E(Y|do(X = x)) = P(Y = 1|do(X = x))$. So the CRR becomes

$$\begin{aligned} CRR(x_1, x_2) &= \frac{P(Y = 1|do(X = x_2))}{P(Y = 1|do(X = x_1))} \\ &= \frac{\exp{(\gamma_1^* + \gamma_2 x_2)}}{\exp{(\gamma_1^* + \gamma_2 x_1)}} \\ &= \exp{(\gamma_2(x_2 - x_1))} \end{aligned}$$

We ended up with an expression of the CRR which has only one unknown parameter(γ_2). In an attempt to estimate γ_2, following the linear no interaction case, we try to work out $E(Y|G)$ and $E(X|G)$.

$$\begin{aligned} E(Y|G = g) &= E_U E_{X|U,G=g} E(Y|X = x, U) \\ &= E_U E_{X|U,G=g} (\exp{(\gamma_1 + \gamma_2 X + \gamma_3 U)}) \\ &= \exp{(\gamma_1)} E_U E_{X|U,G=g} (\exp{(\gamma_2 X + \gamma_3 U)}) \end{aligned}$$

The above equation is difficult to summarise more, because we don't know the distribution of the unobserved confounding. Some approximations to the integral involved assume $Y \perp\!\!\!\perp G \mid X$. However, this case addresses a no confounding environment. But if we didn't have confounding, then we could have estimated the causal parameter directly from the data from the first place and wouldn't need the approach of Mendelian Randomisation. For the needs of this project, some very rough approximations follow, to justify why from prospective view the causal parameter γ_2 can be estimated as

$$\hat{\gamma_2} = \frac{\log E(Y|G = 1) - \log E(Y|G = 0)}{E(X|G = 1) - E(X|G = 0)} \tag{8}$$

Continuing the equation above for $E(Y|G = g)$, we approximate the exponential part using $\exp(x) \approx 1 + x$ (Taylor expansion). Then

$$\begin{aligned} E(Y|G = g) &= \exp{(\gamma_1)} E_U E_{X|U,G=g} (1 + \gamma_2 X + \gamma_3 U) \\ &= \exp{(\gamma_1)} E_U [1 + \gamma_3 U + \gamma_2 E_{X|U,G=g} X] \\ &= \exp{(\gamma_1)} E_U [1 + \gamma_3 U + \gamma_2 (\delta_1 + \delta_2 g + \delta_3 U)] \\ &= \exp{(\gamma_1)} [1 + \gamma_2 \delta_1 + \gamma_2 \delta_2 g + (\gamma_3 + \gamma_2 \delta_3) E(U)] \end{aligned}$$

So

$$\log E(Y|G = g) = \gamma_1 + \log{(1 + \gamma_2 \delta_1 + \gamma_2 \delta_2 g + (\gamma_3 + \gamma_2 \delta_3) E(U))}$$

Again we approximate the above logarithm using $\log(1+x) \approx x$. Then

$$\begin{aligned}\log E(Y|G=g) &= \gamma_1 + \gamma_2\delta_1 + \gamma_2\delta_2 g + (\gamma_3 + \gamma_2\delta_3)E(U) \\ &= c_1^* + \gamma_2\delta_2 g\end{aligned}$$

For a binary genotype variable, we get

$$\log E(Y|G=1) - \log E(Y|G=0) = \gamma_2\delta_2$$

Now

$$\begin{aligned}E(X|G=g) &= E_{U|G=g}E(X|G=g,U) \\ &\overset{CC1}{=} E_U E(X|G=g,U) \\ &= E_U(\delta_1 + \delta_2 g + \delta_3 U) \\ &= c_2^* + \delta_2 g\end{aligned}$$

It follows that

$$E(X|G=1) - E(X|G=0) = \delta_2$$

This is how we obtained equation (8). In the case of retrospective study, we can't estimate γ_2 directly from equation (8), because data is collected conditionally on the disease status. We can only estimate $E(X|Y)$ and $E(G|Y)$. Therefore we need a different approach.
Let

$RR_{Y|G} = \frac{P(Y=1|G=1)}{P(Y=1|G=0)}$

$OR_{Y|G} = \frac{P(Y=1|G=1)P(Y=0|G=0)}{P(Y=1|G=0)P(Y=0|G=1)}$

$OR_{G|Y} = \frac{P(G=1|Y=1)P(G=0|Y=0)}{P(G=0|Y=1)P(G=1|Y=0)}$

Then

$$\begin{aligned}\hat{\gamma_2} &= \frac{\log\frac{E(Y|G=1)}{E(Y|G=0)}}{E(X|G=1) - E(X|G=0)} \\ &= \frac{\log\frac{P(Y=1|G=1)}{P(Y=1|G=0)}}{E(X|G=1) - E(X|G=0)} \\ &= \frac{\log RR_{Y|G}}{E(X|G=1) - E(X|G=0)}\end{aligned}$$

For rare diseases, it is reasonable to assume $P(Y = 0|G) \approx 1$ (even if the individual has the genotype that affects the trait). So we can approximate $RR_{Y|G}$ with $OR_{Y|G}$. Then

$$\begin{aligned}
RR_{Y|G} &\approx OR_{Y|G} \\
&= \frac{P(Y = 1|G = 1)P(Y = 0|G = 0)}{P(Y = 1|G = 0)P(Y = 0|G = 1)} \\
&= \frac{P(G = 1|Y = 1)P(G = 0|Y = 0)}{P(G = 0|Y = 1)P(G = 1|Y = 0)} \\
&= OR_{G|Y}
\end{aligned}$$

In this way we found a suitable form for the numerator to be estimated from a case-control study. For estimating the denominator $E(X|G = 1) - E(X|G = 0)$ we need a different approach.

$$\begin{aligned}
E(X|G = g) &= E(X|G = g, Y = 0)P(Y = 0|G = g) + E(X|G = g, Y = 1)P(Y = 1|G = g) \\
&= E(X|G = g, Y = 0)P(G = g|Y = 0)\frac{P(Y = 0)}{P(G = g)} \\
&+ E(X|G = g, Y = 1)P(G = g|Y = 1)\frac{P(Y = 1)}{P(G = g)}
\end{aligned}$$

In the last equation all the above quantities can be estimated from the data except $P(Y = y)$ and $P(G = g)$. For the distribution of the disease and the genotype we could use information from other studies if present that estimate the patients proportion and the general allocation of the genotype in the whole population. If additional information is not available we could still approximate the denominator in the case of a rare disease. We consider estimating the genotype distribution only from the controls group (i.e. $P(G = g) = P(G = g|Y = 0)$) and approximate $p(Y = 0)$ with a value close to 1 and $p(Y = 1)$ with a value close to zero. We could fit an approximation value of $P(Y = 0)$ and $P(Y = 1)$ to the following expression.

$$\begin{aligned}
E(X|G = g) &= E(X|G = g, Y = 0)P(G = g|Y = 0)\frac{P(Y = 0)}{P(G = g|Y = 0)} \\
&+ E(X|G = g, Y = 1)P(G = g|Y = 1)\frac{P(Y = 1)}{P(G = g|Y = 0)} \\
&= E(X|G = g, Y = 0)P(Y = 0) \\
&+ E(X|G = g, Y = 1)P(G = g|Y = 1)\frac{P(Y = 1)}{P(G = g|Y = 0)}
\end{aligned}$$

So, when we have a case-control study, under the assumption of a rare disease, we could approximate γ_2 by

$$\hat{\gamma_2} = \frac{\log OR_{G|Y}}{A} \tag{9}$$

where

$$A = [E(X|G=1,Y=0)P(Y=0) + E(X|G=1,Y=1)P(G=1|Y=1)\frac{P(Y=1)}{P(G=1|Y=0)}]$$
$$- [E(X|G=0,Y=0)P(Y=0) + E(X|G=0,Y=1)P(G=0|Y=1)\frac{P(Y=1)}{P(G=0|Y=0)}]$$

In the application part of this project, a point estimate of γ_2 will be given using R and fitting figures in this formula. Unfortunately, there is no measure of the uncertainty around this estimation.

5 Application

In this section two datasets will be analysed using the theory of Mendelian Randomisation. Both datasets originate from an Italian genetic study of early-onset myocardial infarction and are collected from 125 Coronary Care units in Italy. The study involves individuals that have been hospitalised for a myocardial infarction before the age of 45. For every patient, a control match has been selected, with one-to-one matching for age, gender and geographical origin. The first dataset will explore whether high levels of fibrinogen (a glycoprotein in the blood) is causal for myocardial infarction and the second dataset will explore whether high levels of triglyceride (a glyceride in the blood) is causal for the same disease. The first dataset will prove to be a negative example of Mendelian Randomisation, because we will see that a test of association between the relative genotype and the disease will show no association. Therefore there will be no causal effect to be estimated. However, the second dataset will give a positive association test between the phenotype and the disease, which will allow us to use the genotype as an instrument and estimate the causal effect. Due to the fact that data is collected retrospectively, application will be possible up to a certain point. Both datasets are analysed using R.

5.1 Negative example of Mendelian Randomisation

This dataset will be used to explore whether the phenotype fibrinogen has a causal effect on the disease myocardial infarction(MI). Fibrinogen is a soluble plasma glycoprotein synthesised by the liver. Fibrinogen levels can be measured in venous blood with normal levels ranging at about 1.5-3.0 g/L. It is suspected that higher levels of fibrinogen are associated with myocardial infarction. In an effort to test whether this is true we will include in the analysis the G455 gene. Summarising, we have the disease variable Y, a binary variable that takes values 0 for controls and 1 for cases, the phenotype variable X, a continouous variable that records the fibrinogen level in the blood and the genotype variable G, which takes values 0, 1, 2 according to the number of copies of the minor allele. In this dataset we have a case-control study of 2646 individuals. From the 2646 individuals, 1385 consist the cases group and 1261 the controls group. We note that there are 865 missing fibrinogen level values of which 452 belong to the controls and 413 to the cases. For the following analysis (test of association between fibrinogen(phenotype) and G455(genotype)), the missing observation vectors are removed from the dataset.

As a first step to the exploration of the data, Figure 4 shows a plot of the disease status(controls=0, cases=1) against fibrinogen levels(100xg/L). What we see, is that individuals that actually had MI have a wider range of fibrinogen levels with very high levels included in opposition to the controls group that only show a range of low levels. So, it looks like high fibrinogen level might be a cause for MI. But because of reverse causation and unobserved confounding, we can't be certain for causality only from an observed association between the phenotype and the disease. Therefore, we will use the G455 gene in the analysis, which from biology knowledge is assumed to be independent of confounding (satisfying CC1).

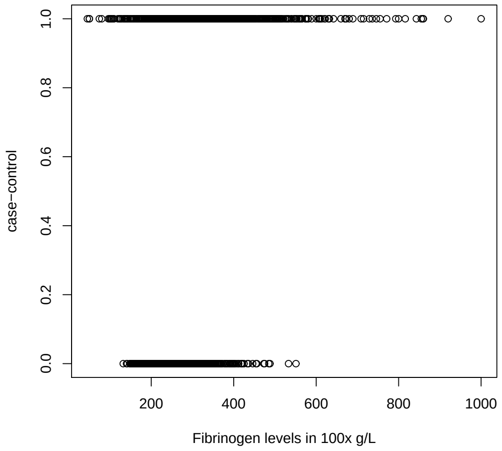


Figure 4: Possible association between high fibrinogen levels and disease.

To deduce possible association of the G455 gene with fibrinogen levels, we could use the controls group as it can be reasonably assumed it represents a sample from the general population (since MI is considered to be a rare disease). In Figure 5, we see how the level of fibrinogen varies according to the status of the G455 gene in the controls group. The horizontal line in the middle of the box is the median and the box extends between the lower and the upper quartiles. Even in the scaled figure, it is not clear whether individuals with different G455 code have different fibrinogen levels on average. Using R we get that the medians for controls group with G455 code $0, 1, 2$ are $256.5, 269.0, 287.5$ respectively and the means for the same group with G455 code 0, 1, 2 are 265.3648, 271.6642, 287.6250 respectively. The means for different G455 codes, do not have a remarkable difference between them.

To deduce whether there is an association between fibrinogen and G455, we use R to analyse the data and fit the model

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 0, 1, 2 \quad j = 1, \ldots, 809$$

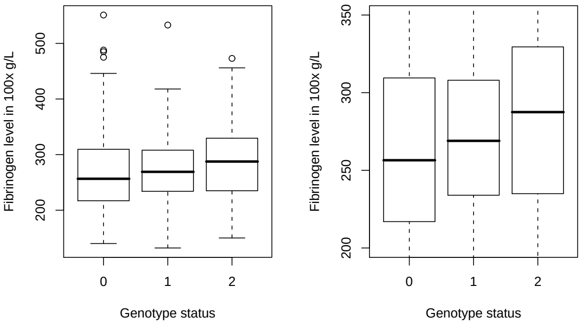


Figure 5: Association between gene G455 and fibrinogen levels in controls group.

where
$i = 0, 1, 2$ corresponds to genotype code $0, 1, 2$ respectively and $j = 1, \ldots, 809$ corresponds to the j^{th} individual in the controls group.
X_{ij} is the fibrinogen level for the j^{th} individual in the controls group with G455 code i, μ is the overall mean response, α_i is the effect due to the i^{th} level of the G455 gene and ε_{ij}'s are iid $N(0, \sigma^2)$ random variables. We use the constraint $\alpha_0 = 0$ so as to compare α_0 with the estimators of α_1 and α_2.

Using R we get the following ANOVA table and summary of the model.

Anova Table

```
Analysis of Variance Table

Response: x1
           Df  Sum Sq Mean Sq F value Pr(>F)
g1          2   27755   13877  3.5879 0.0281 *
Residuals 806 3117515    3868
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Summary

```
Call:
aov(formula = x1 ~ g1, subset = (y1 == "0"))

Residuals:
     Min       1Q   Median       3Q      Max
-139.664  -46.365   -6.365   39.336  285.635
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  265.365      2.815  94.258   <2e-16 ***
g11            6.299      4.746   1.327   0.1848
g12           22.260      8.775   2.537   0.0114 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 62.19 on 806 degrees of freedom
Multiple R-squared: 0.008824,   Adjusted R-squared: 0.006365
F-statistic: 3.588 on 2 and 806 DF,  p-value: 0.0281
```

The anova table tests the hypothesis H_0 that all the group means(α_i) are zero. To test this hypothesis, we refer the value 3.5879 to an $F_{2,806}$ distribution. The $p-value$ for the relative test is 0.0281. Therefore, we can reject H_0 and conclude that $\alpha_i \neq 0$ for all $i = 1, 2$. So the G455 genotype code is associated with fibrinogen levels.

From the summary of our model we get $\hat{\sigma^2} = 62.19^2$. The parameter estimates with the standard error in brackets are:
$\hat{\mu} = 265.365(2.815)$
$\hat{\alpha_0} = 0$, $\hat{\alpha_1} = 6.299(4.746)$, $\hat{\alpha_2} = 22.260(8.775)$
From the difference in the estimations of α, it seems that one allele can also influence fibrinogen levels($\hat{\alpha_1} = 6.299$). Two alleles seem to have a bigger effect($\hat{\alpha_2} = 22.260$). Also, from the summary of the model, we see that a test for the null hypothesis $H_0 : \alpha_1 = 0$ gives $p - value = 0.1848$. We don't have enough evidence from the data to reject this H_0. So there might not be an effect due to one allele present in the genotype code compared with no alleles at all. For the same hypothesis about α_2 ($H_0 : \alpha_2 = 0$), we get $p - value = 0.0114$. So we can reject this H_0 and conclude that $\alpha_2 \neq 0$. So two alleles in the genotype actually affect fibrinogen status. Since we established association of genotype-phenotype, also CC2 holds. We can a priori exclude a direct effect of the G455 gene on the disease status (using underlying biology knowledge), satisfying CC3.

Using the arguments above, the genotype G455 satisfies all three core conditions. Therefore, it can be used as an instrument in our analysis. The DAG representing this case is presented in Figure 6.

To test whether fibrinogen is causal for MI, it's enough to check whether there is an association between the G455 gene and MI. This reasoning seems to be suitable whether we have prospective or retrospective data. For testing the presence of a causal effect, we will use a chi-square test for independence. A summary of the dataset is presented in the table in Figure 7.

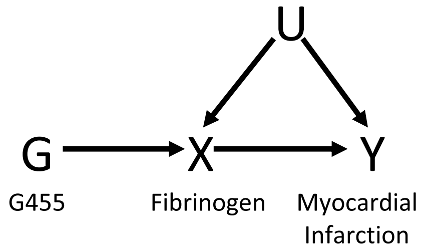
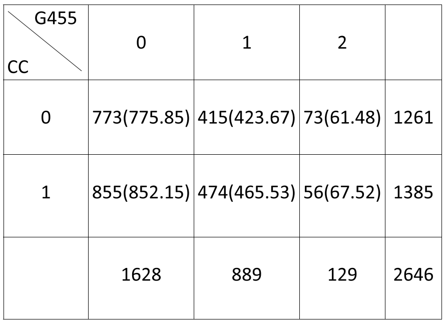


Figure 6: DAG representing first dataset



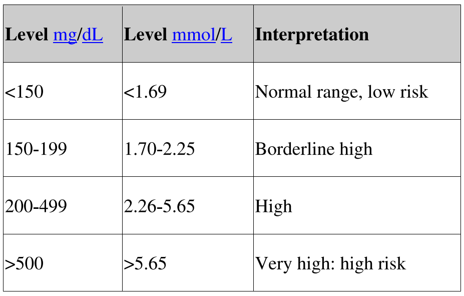
G455 / CC	0	1	2	
0	773(775.85)	415(423.67)	73(61.48)	1261
1	855(852.15)	474(465.53)	56(67.52)	1385
	1628	889	129	2646

Figure 7: Summary of first dataset

In this test, the observations with missing fibrinogen values are included. Figures are obtained using R. The numbers outside brackets indicate the real numbers that belong to the specific disease and genotype status. The numbers in brackets indicate the expected numbers under the null hypothesis that the row into which an observation falls is independent of the column into which it falls. Let p_{ij} be the probability that an individual belongs to cell (i, j), where i represents the disease status ($i = 0, 1$) and j represents the genotype status ($j = 0, 1, 2$). If p_i is the probability that an individual belongs to the i category of disease status and q_j is the probability that an individual belongs to the j category of genotype status, we wish to test the null hypothesis $H_0 : p_{ij} = p_i q_j$. The relevant χ^2 statistic for the test is $\chi^2 = 4.483$. Under H_0, this would be an observation from a χ^2 distribution with 2 degrees of freedom. We may reject H_0 if $\chi^2 > \chi_2^2(\alpha)$, where $\chi_2^2(\alpha)$ is the upper α-point of the χ_2^2 distribution. For $\alpha = 0.05$, $\chi_2^2(0.05) = 5.99$. Since $\chi^2 = 4.483 < 5.99 = \chi_2^2(0.05)$, we don't have enough evidence to reject H_0. Therefore, we conclude that the event that an individual is a patient or not is independent of their G455 gene code. So we didn't find any association between the genotype and the disease. This result is practically equivalent to the fibrinogen level not being causal to MI. So the analysis of this dataset ends here, since we don't have a causal effect to estimate.

5.2 Positive example of Mendelian Randomisation

This dataset includes more or less the same patients and controls as the previous dataset but with different recordings of phenotype and gene code. The aim of this analysis is to conclude whether the phenotype triglycerides has a causal effect on the disease myocardial infarction(MI). Triglyceride is a glyceride (the major form of fat) and comes from the food we eat as well as from being produced by the body. It can be measured in bloodstream. High levels of triglycerides have been accused for various heart related diseases so that the American Heart Association(AHA) has set guidelines (presented in Figure 8). In this application, we will see that using the Mendelian Randomisation approach, we also deduce a causal effect of triglycerides on MI. The instrumental variable in this case will be a gene called APOA5. Again the disease variable Y, is a binary variable that takes values 0 for controls and 1 for cases, the phenotype variable X, is a continuous variable that records the triglycerides levels in the bloodstream and the genotype variable G, takes values 0, 1, 2 according to the number of copies of the minor allele in the APOA5 gene. This study involves 3728 individuals and is a case-control study. From the 3728 individuals, 1864 consist the cases group and 1864 the controls group. In this dataset we have 348 missing values of triglycerides level and 18 missing values of the APOA5 code of which one is in common. In the following analysis, whenever incomplete observations are removed from the dataset it will be stated.



Level mg/dL	Level mmol/L	Interpretation
<150	<1.69	Normal range, low risk
150-199	1.70-2.25	Borderline high
200-499	2.26-5.65	High
>500	>5.65	Very high: high risk

Figure 8: Guidelines for triglyceride levels from AHA

In Figure 9 we see a plot of the disease status(controls=0, cases=1) against triglycerides levels(mg/dL). Individuals in the controls group don't have very high levels of triglycerides while individuals in the patients group show a wider range of triglycerides levels with high levels included. Using the Mendelian Randomisation approach we will see that in this case association is probably causation. The instrumental variable in this case will be the APOA5 gene code. Using the same argument as in the previous application we conclude that CC1 holds.

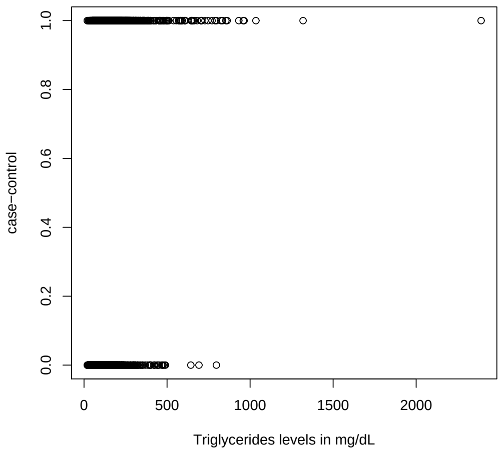


Figure 9: Possible association between high triglycerides levels and disease.

In figures 10 and 11, we see box plots of triglycerides levels with the APOA5 gene in the controls and cases groups respectively. Box plots on the right are scaled figures of the box plots on the left. In figure 10, comparing 50% of the observations included in the box, we see that individuals with 1 or 2 number of minor alleles have higher levels of triglycerides than individuals with 0 number of alleles in the APOA5 code. Also individuals with APOA5 code 2 show higher levels of triglycerides than individuals with APOA5 code 1. The same pattern appears also in the cases group (Figure 11). To infer association of the APOA5 gene with triglycerides level we will use mostly the controls group since possible association in the cases group might appear exactly because we condition on cases. Analysis of the association between APOA5 and triglycerides levels follows for both groups. Because we have 348 missing values of triglycerides levels and 18 missing values of APOA5 code (1 in common), the dataset reduces from 3728 to 3363 observations of which 1697 belong to the controls group and 1666 belong to the cases group.

Firstly we analyse data from the controls group. We use R to fit the model

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1, 2, 3 \quad j = 1, \ldots, 1697$$

where
$i = 0, 1, 2$ corresponds to genotype code $0, 1, 2$ respectively and $j = 1, \ldots, 1697$ corresponds to the j^{th} individual in the controls group.

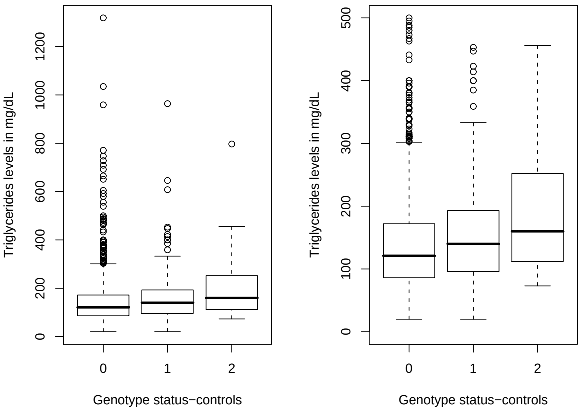


Figure 10: Association between APOA5 and triglycerides levels in controls group.

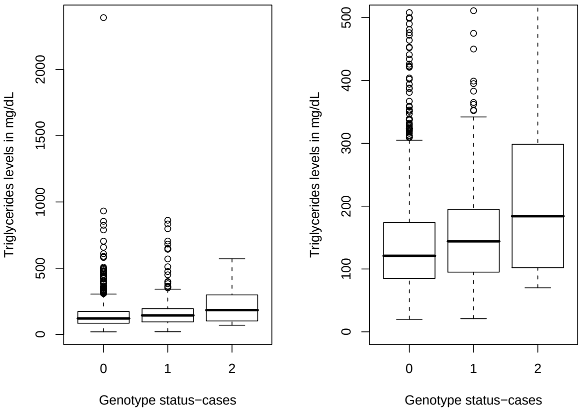


Figure 11: Association between APOA5 and triglycerides levels in cases group.

X_{ij} is the triglycerides level for the j^{th} individual in the controls group with APOA5 code i, μ is the overall mean response, α_i is the effect due to the i^{th} level of the APOA5 gene and ε_{ij}'s are iid $N(0, \sigma^2)$ random variables. The constraint $\alpha_0 = 0$ is used so as to compare α_0 with the estimations of α_1 and α_2.

Using R we get the following ANOVA table and summary of the model.

Anova table for controls

```
Analysis of Variance Table

Response: x22
```

```
            Df  Sum Sq Mean Sq F value    Pr(>F)    
g22          2  173322   86661  17.727 2.402e-08 ***
Residuals 1694 8281282    4889                      
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Summary for controls

```
Call:
aov(formula = x22 ~ g22, subset = (y22 == "0"))

Residuals:
    Min      1Q  Median      3Q     Max 
-167.44  -44.67  -15.67   29.53  574.33 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  117.666      1.836  64.083  < 2e-16 ***
g221          12.801      4.890   2.618  0.00893 ** 
g222         126.779     23.378   5.423 6.71e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 69.92 on 1694 degrees of freedom
Multiple R-squared: 0.0205,     Adjusted R-squared: 0.01934 
F-statistic: 17.73 on 2 and 1694 DF,  p-value: 2.402e-08
```

The anova table tests the hypothesis H_0 that all the group means(α_i) are zero. To test this hypothesis, we refer the value 17.727 to an $F_{2,1694}$ distribution. The p-value for the relative test is $2.402 * 10^{-8}$. This is a very small $p - value$, therefore, we can reject H_0 and conclude that not all $\alpha_i = 0$ for $i = 1, 2$. So the APOA5 gene code is associated with triglycerides levels in the controls group.

From the summary of the model we get $\hat{\sigma^2} = 69.92^2$. The parameter estimates with the standard error in brackets are:
$\hat{\mu} = 117.666(1.836)$
$\hat{\alpha_0} = 0$, $\hat{\alpha_1} = 12.801(4.890)$, $\hat{\alpha_2} = 126.779(23.378)$
From the difference in the estimations of α, it seems that one allele can also influence triglycerides levels. Two alleles though have a very strong effect on the triglycerides level as 126.779 is a very large figure compared to 0 or 12.801. The $p - values$ for the null hypotheses $H_0 : \alpha_1 = 0$ and $H_0 : \alpha_2 = 0$ are 0.00893 and $6.71 * 10^{-8}$ respectively. Therefore, both H_0 are rejected. The genotype-phenotype association in controls group is well established in this dataset, indicating that CC2 holds.

Then we analyse data from the cases group. We use R to fit the model

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1, 2, 3 \quad j = 1, \ldots, 1666$$

where
$i = 0, 1, 2$ corresponds to genotype code $0, 1, 2$ respectively and $j = 1, \ldots, 1666$ corresponds to the j^{th} individual in the cases group.
X_{ij} is the triglycerides level for the j^{th} individual in the cases group with APOA5 code i, μ is the overall mean response, α_i is the effect due to the i^{th} level of the APOA5 gene and ε_{ij}'s are iid $N(0, \sigma^2)$ random variables. The constraint $\alpha_0 = 0$ is used so as to compare α_0 with the estimations of α_1 and α_2.

Using R again we get the following ANOVA table and summary for this model.

Anova table for cases

```
Analysis of Variance Table

Response: x22
            Df   Sum Sq  Mean Sq F value  Pr(>F)
g22          2   111951    55976   3.389 0.03398 *
Residuals 1663 27467544    16517
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Summary for cases

```
Call:
aov(formula = x22 ~ g22, subset = (y22 == "1"))

Residuals:
    Min      1Q  Median      3Q     Max
-170.97  -69.85  -23.24   29.24 2218.49

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  172.508      3.529  48.879   <2e-16 ***
g221          18.457      8.045   2.294   0.0219 *
g222          36.075     26.470   1.363   0.1731
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
Residual standard error: 128.5 on 1663 degrees of freedom
Multiple R-squared: 0.004059,   Adjusted R-squared: 0.002861
F-statistic: 3.389 on 2 and 1663 DF,  p-value: 0.03398
```

The anova table tests the hypothesis H_0 that all the group means(α_i) are zero. To test this hypothesis, we refer the value 3.389 to an $F_{2,1663}$ distribution. The p-value for the relative test is 0.03398. This $p-value$ is on the borderline, but we decide to reject H_0. So even in the cases group, the APOA5 genotype code seems to be associated with triglycerides levels.

From the summary of the model we get $\hat{\sigma^2} = 128.5^2$. The parameter estimates with the standard error in brackets are:
$\hat{\mu} = 172.508(3.529)$
$\hat{\alpha_0} = 0$, $\hat{\alpha_1} = 18.457(8.045)$, $\hat{\alpha_2} = 36.075(26.470)$
So, even in the cases group there is a difference in the estimators of α. However in this group, for the null hypothesis $H_0 : \alpha_2 = 0$ we get $p - value = 0.1731$ and therefore can't reject H_0. This means that two alleles in the APOA5 gene don't affect triglycerides level differently than no alleles in the the APOA5 gene. But for $H_0 : \alpha_1 = 0$ we get $p - value = 0.0219$ which gives enough evidence to reject this null hypothesis. So one allele in the APOA5 gene gives a different effect on triglycerides level than no alleles. The fact that one allele seems to have an impact on the phenotype and two alleles not is odd and it's probably because we check for this association in the cases group. Even with one allele affecting the phenotype we establish the genotype-phenotype association satisfying CC2 holds.

There is a big difference in the p-value for the specific test in controls and cases, but as MI is concidered a rare disease in the whole population we will use the controls group as the appropriate group to derive conclusions on APOA5 and triglycerides association. Again from underlying biology knowledge we exclude a direct effect of the APOA5 gene on the disease status satisfying CC3. Using the arguments above, the genotype APOA5 satisfies all three core conditions. Therefore, it can be used as an instrument in the estimation of the causal effect. The DAG representing this case is presented in Figure 12.

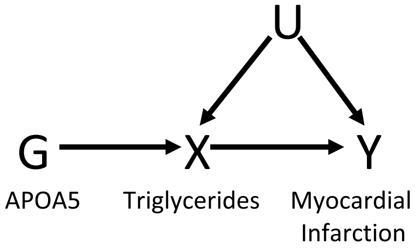


Figure 12: DAG representing second dataset

For this dataset, we will assume a log-linear model and target to estimate the Causal Risk Ratio. More specifically, we assume:

1. $E(Y|X=x,U=u)=E(Y|do(X=x),U=u)=\exp(\gamma_1+\gamma_2 x+\gamma_3 u)$

2. $E(X|G=g,U=u)=\delta_1+\delta_2 g+\delta_3 u$

The Causal Risk Ratio, as shown in subsection 4.2.2 is

$$\begin{aligned} CRR(x_1,x_2) &= \frac{P(Y=1|do(X=x_2))}{P(Y=1|do(X=x_1))} \\ &= \exp(\gamma_2(x_2-x_1)) \end{aligned}$$

Assuming that MI is a rare disease, we use equations 9 and **??** to estimate the causal parameter γ_2. So

$$\hat{\gamma_2}=\frac{\log OR_{G|Y}}{A} \tag{9}$$

where $OR_{G|Y}=\frac{P(G=1|Y=1)P(G=0|Y=0)}{P(G=0|Y=1)P(G=1|Y=0)}$ and

$$\begin{aligned} A=&[E(X|G=1,Y=0)P(Y=0)+E(X|G=1,Y=1)P(G=1|Y=1)\frac{P(Y=1)}{P(G=1|Y=0)}] \\ &-[E(X|G=0,Y=0)P(Y=0)+E(X|G=0,Y=1)P(G=0|Y=1)\frac{P(Y=1)}{P(G=0|Y=0)}] \end{aligned}$$

Using R, we estimate the probabilities in the $OR_{G|Y}$ formula from the relative frequencies in the dataset. Because this formula is derived for a binary genotype status, we consider $G=0$ to represent APOA5 status 0 and 1 and $G=1$ to represent APOA5 status 2. APOA5 status 1 was grouped with APOA5 status 0 because in the above analysis for the APOA5-triglycerides association in the controls group, the estimation of the effect of APOA5 genotype on triglycerides was closer for APOA5 status $0(\alpha_0=0)$ and $1(\alpha_1=12.801)$ than for APOA5 status $1(\alpha_1=12.801)$ and $2(\alpha_2=126.779)$. So in some sense the effect of APOA5 status 0 and 1 is considered relatively similar. For these estimations the 18 vectors with missing APOA5 information were removed from the dataset. We get

$P(G=\hat{1}|Y=1)=0.0188$, $P(G=\hat{0}|Y=0)=0.9896$, $P(G=\hat{0}|Y=1)=0.9812$ and $P(G=\hat{1}|Y=0)=0.0104$

So $OR_{G|Y}=1.8231$

For estimating A we need additional information from other studies regarding $p(Y = y)$. Unfortunately we don't have this information but we know that MI is a rare disease($P(Y = 0) \approx 1$). So we will assume for this application that $P(Y = 0) = 0.9999$ and $P(Y = 1) = 0.0001$. Using R we get

$E(X|G \hat{=} 0, Y = 0) = 120.1332$, $E(X|G \hat{=} 1, Y = 0) = 244.4444$

$E(X|G \hat{=} 0, Y = 1) = 176.5288$, $E(X|G \hat{=} 1, Y = 1) = 208.5833$

$P(G = \hat{0}|Y = 0) = 0.9896$, $P(G = \hat{1}|Y = 0) = 0.0104$ and $P(G = \hat{1}|Y = 1) = 0.0188$

Combining the above estimations we get $\hat{\gamma_2} = 4.83 * 10^{-3}$. Finally, we get the following expression for the CRR

$$\begin{aligned} CRR(x_1, x_2) &= \frac{P(Y = 1|do(X = x_2))}{P(Y = 1|do(X = x_1))} \\ &= \exp\left(4.83 * 10^{-3}(x_2 - x_1)\right) \end{aligned}$$

For a unit increase in the phenotype status, the CRR becomes $\exp\left(4.83 * 10^{-3}\right) \approx 1.0048$, meaning that $P(Y = 1|do(X = x_1 + 1)) \approx 1.0048 P(Y = 1|do(X = x_1))$. This indicates that as we intervene on an individual's phenotype and change it status to a greater value, we also increase the probability that the individual becomes a patient. In this application, with some approximations and assumptions about the disease and the APOA5 alleles proportion in the population we were able to identify the causal effect of interest. However, the modelling assumptions about the conditional expectations are assumptions tha can't be tested as they include the unobserved confounding so we can't be certain for the validity of the results.

6 Conclusions

Observed associations between modifiable phenotypes and diseases are sometimes misleading. It is not easy to argue which variable is the cause and which the effect, especially when other factors are suspected to interfere. Mendelian Randomisation, is a method that allows us to conclude causation when confounding is present but not fully understood. Using the genetic background, we can reject the explanation of reverse causation and also deal with the problem of confounding. Using Mendel's second law, which states that the allocation of the genes is random and independent of confounding factors, we no longer need to assume no unobserved confounding when analysing medical data. Establishing that a risk factor (phenotype/exposure) actually causes the disease, we can then address the problem of how modifying the risk factor we can reduce the proportion of the disease in the general population aiming to improve population health.

An important step to the development of this method is the use of formal mathematical language to describe accurately the parameters of interest. Distinguishing the intervening action (do operator) on the phenotype allows us to explain more accurately what can and what can't be estimated from the data. After formally expressing the quantities of interest, we set the core conditions that will allow us to use a genotype associated with the phenotype as an instrument to test for the presence of a causal effect. When we use the instrumental variable approach, we assume that the Core Conditions hold. Since these conditions involve confounding and can't be statistically tested, we need to be careful when justifying that they hold. Usually we need the underlying biology knowledge to justify their validity. This is why only well studied genes should be used as instruments. If the gene is not suitable, any inference about the causal effect is proned to be biased. So full understanding of the functionality of the instrument is crucial. These genotypes though are difficult to find.

Directed Acyclic Graphs can help us picture the conditional dependencies and independencies indicated by Core Conditions. In this project, we assumed the simple DAG in Figure 1 to describe the joint distribution. However, there are more complicated cases to consider. Cases where more than one gene affects the phenotype (genetic heterogeneity), cases were by confounding we mean behavioural patterns which actually are not independent of the genotype and generally cases where several genes affect several phenotypes which together affect the disease and confounding can have a multible effect on them.

The complications are many but as a first step we can consider the simple case and show that if conditional dependencies and independencies implied by Core Conditions hold then a test of association between the genotype and the disease can replace a test of association between the phenotype and the disease and unravel the confusion of association or causation. This replacement of the test, seems to be suitable whether analysing data from prospective or retrospective view. This is practically very helpful since medical data very often take the form of a case-control study.

Testing for the causal effect is the easiest step. When trying to estimate it though, additional stronger parametric assumptions need to be made. These assumptions as they include confounding effect just can't be tested and we don't know if they actually hold. Also, in an effort to get an identifiable causal parameter some of the assumptions we make might not be reasonable. For example linearity when the disease outcome is binary. Econometrics literature is a source of information on instrumental variables but mostly provides techniques on manipulating continuous variables. In epidimiology since the disease outcome is binary additional reasearch needs to be made.

In the application part of this project it was straightforward to decide for the presence of a causal effect. When trying to estimate it though, exactly because we had to deal with retrospective data, we needed additional approximations to identify the causal parameter of interest. We also made further approximations to estimate the causal parameter from the data and therefore only able to give a point estimate and no confidence intervals. The simplicity of the statistical analysis in this dataset indicates that further research needs to be made. Because of the nature of this method, additional research should involve collaboration of both biologists and statisticians.

References

[1] A.P.Dawid, *Fundamentals of Statistical Causality*, RSS/EPSRC Graduate Training Programme, Version of April 10, 2008.

[2] A.P.Dawid, *Influence Diagrams for causal modelling and inference*, International Statistical Review, 70:161-189, 2002.

[3] V.Didelez, S.Meng and N.Sheehan, *On the Bias of IV estimators for Mendelian Randomisation* , Research Report 08-20, Statistics Group, University of Bristol, 2008.

[4] V.Didelez, N.Sheehan, *Mendelian Randomisation as an Instrumental Variable Approach to Causal Inference* , Stat. Methods Med. Res. 16 (2007), 309-330.

[5] V.Didelez, N.Sheehan, *Mendelian Randomisation and Instrumental Variables: What Can and What Can't be Done* , Technical Report 05-02, Department of Health Sciences, University of Leicester, 2005.

[6] V.Didelez, N.Sheehan, *Mendelian Randomisation: Why epidemiology needs a Formal Language for Causality* , In Causality and Probability in the Sciences, F.Russo and J. Williamson, Eds., vol. 5 of Texts in Philosophy. London College Publications, 2007, pp.263-292.

[7] N.Sheehan, V.Didelez, P.R.Burton, M.D.Tobin *Mendelian Randomisation and Causal Inference in Observational Epidemiology* , PLoS Medicine Vol 5, No 8, e177 doi:10.1371/journal.pmed.0050177, 2008.

[8] Katan *Apolipoprotein E Isoforms, Serum Cholesterol and Cancer* , Lancet i, 507-508

[9] J.Pearl *Causality*, Cambridge University Press, 2000.

[10] A. A. Balke and J.Pearl *Computational Methods, bounds and applications*, In R.L. Mantaras and D.Poole, editors, Proceedings of the 10th Conference on Uncertainty in Artificial Inteligence, pages 46-54, 1994.