UNIVERSITY OF YORK

INTEGRATED MASTERS IN MATHEMATICS

Causal inference in epidemiology:

The effect of socio-economic intervention.

Author
Christopher A Oldnall

Supervisor
Dr Ben Powell

May 6, 2021

Abstract

This paper looks to equip the reader firstly with a toolbox of theory surrounding causal inference. Namely graphical modelling, independence, confounding, structural equation models and bias. It highlights some of the most recent key developments in the field, such as the back-door criterion and the do-operator. In addition this paper shows how concepts familiar to a statistician, such as covariance, can be exploited to determine the strength of a causal link. These ideas are introduced to the reader using examples which sit within an epidemiological setting. Once a significant amount of time has been spent developing the theory, the paper then turns to looking at some implementation methodology developed by Judea Pearl, with calculational techniques supported by the work conducted by Kay Brodersen around the use of Bayesian structural time series for causal inference. Finally it address some ethical considerations for utilising causal inference in epidemiology and how its use within artificial intelligence (AI) also has ethical consequences.

Contents

List of Figures

Acknowledgements

With thanks to **Dr Ben Powell**, without whose guidance, enthusiasm and mathematical passion this project would not have come to light.

An additional thank you to my fellow mathematical peers who have always supported me in my endeavours, as well as some who have proof-read this document: *Lizzie Mason, Mia Holley, Harry Pratt, Matthew Johnstone, Grace Clark, Edward Little, Jade Brewer, Thea Handy, Jamie Milne, Megan Parry, Tallulah Coyte, Adam Araim, Dylan Muggleton, Laura Stott, Emily Myles, Alex Cormack, Luke Osborn, Katie Chedgozy, Katelyn Ramshaw, Molly Barker.*

Part I
Introduction

"Shallow men believe in luck...Strong men believe in cause and effect."

Ralph Waldo Emerson [8]

Causality is a field that has been considered for centuries. The problem through time that has continued to arise is that the theory and methodology required to solve problems around causality were claimed by individuals such as Russell [32] to be 'not yet capable of being decisively solved.' Many leading statisticians neglected the field out of an inability to model causality, consistently influencing other mathematicians to follow suit. A key example of this is Pearson [22], who once claimed the field was inscrutable. It took strong mathematicians, like Hill [15], to develop criteria and stand out from the crowd—refuting these claims of impossible modelling. From the acknowledgement that causation modelling could be achieved, leading authors, such as Pearl [25], have started to generate extensive guides to causality. We are at the point where we can utilise causal inference within many fields. This of course includes epidemiology. One might come across scenarios whereby we want to quantify if socio-economic intervention, like a lockdown, has a causal impact on the spread of a disease.

My motivation for wanting to research causal inference comes from my interest in being able to measure the impact a virus can have on the society we live in. Initially I planned to look more at the after-effects of a epidemic. Quickly I realised that I could, with the right toolbox, consider looking at and building models which take in huge amounts of data and provide me with a useful and tangible result.

Theory Surrounding Causal Inference

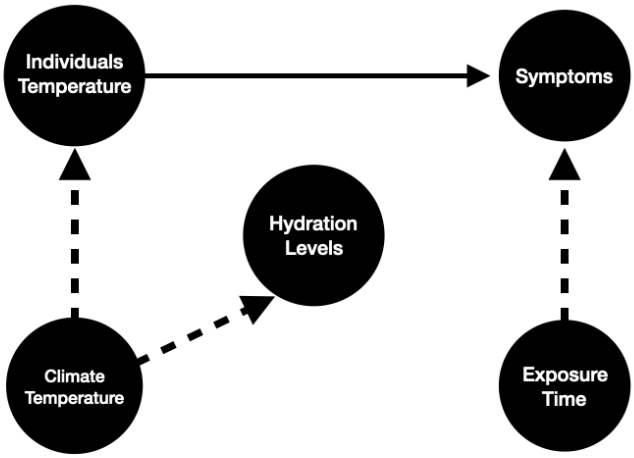
Before delving into the more current methodology of the field, it is important for us to explore a plethora of theorems, results and pieces of statistical inference. Throughout Part II of this paper we will build up our knowledge from basic concepts like **covariance** and **conditional expectation**.

The first complexity of causal inference comes from the selection of variables and the extent to which they are measurable. One could potentially argue that every cause has at least one cause which in itself has a cause and so on. Therefore we decide to split variables into two categories. **Exogenous variables** firstly are causal factors which have influence over other variables in our system, but in themselves we wish to leave them unexplained and independent. **Endogenous variables** are then these other factors which are dependent on both other endogenous and exogenous variables.

In the context of monitoring the spread of a virus we may wish to look at factors such as

climate temperature, an individual's temperature, exposure time (to someone known to have the virus), levels of hydration and symptoms of disease. We could start to construct the following diagram (Figure 1) where we add a directional arrow where we believe one of these factors affects another. A dotted arrow going away from a factor represents an exogenous variable, whilst a factor receiving a dotted arrow or with a bold arrow is an endogenous variable.

Figure 1: Basic 5 Variable Causal Diagram.



We will formalise definitions of these types of variables in addition to delving into the ways that variables relate within section 1.1. A key part of graphical modelling is the relationship between nodes. As such it is sensible for us to touch upon some more sophisticated graph theory where we will learn about terms like **DAG** and **ancestors** in section 1.2.

Once we have explored the connection between variables, it is my intention to introduce the reader to a more in depth discussion around probabilities. In section 2.1 we will take careful note of how the probability of a point in the causal systems taking a value can be modelled by the **Causal Markov Condition** as given below.

Letting a graph, $\mathcal{G}$, be a Directed Acyclic Graph (DAG) with the set of vertices V. The probability of taking the value of the node X_V is given by,

$$p(X = x_V) = \prod_{v \in V} p(x_v | x_{pa_G(v)}).$$

Slowly we will discover how the **'do' operator** allows us to hold a variable constant. Let me provide some motivation for this. Imagine a situation where we want to investigate whether a virus is affecting age groups disproportionately. In this setting it may be useful to hold the gender of the individuals that we examine constant. This would

obviously create a more fair comparison when we look at age disparities. This sort of intervention would be probabilistically represented by denoting gender as the factor W and the following expression:

$$\begin{aligned} p(x_{V\setminus\{w\}}|\text{do}(x_w)) &= \frac{p(x_V)}{p(x_w|x_{pa(w)})}, \\ &= \prod_{v\in V\setminus\{w\}} p(x_v|x_{pa(w)}). \end{aligned}$$

How we achieve this probabilistic expression will be derived and explored in depth throughout section 2.2.

Implementing the Methods

Once we have looked at the probability theory that surrounds causal inference, we then progress to look at methodology around building models. For this we focus primarily on the work presented by Pearl [26], looking to explore how we define the quantities we are interested in as well as what we have to assume about the variables involved. A large part of causal inference methodology is determining whether a variable is **identifiable** hence within section 5.1 we will look at this in detail before turning our attention to how you can calculate quantities either through estimation or approximation.

One starting point for these calculations will be explored in section 6 using modelling suggested by Brodersen et al.[4] in their paper on inferring causal impact using **Bayesian structural time-series models**. The paper provides the background model for the R package 'CausalImpact'. This is a structural equation model, adapted in time

$$\begin{aligned} y_t &= Z_t^T \alpha_t + \epsilon_t, \\ \alpha_{t+1} &= T_t \alpha_t + R_t \eta_t. \end{aligned}$$

Here we assume that the error terms in the observation (ϵ_t) and the systematic error terms (η_t) are normally distributed in the following way

$$\begin{aligned} \epsilon_t &\sim \mathcal{N}(0, \sigma_t^2), \\ \eta_t &\sim \mathcal{N}(0, Q_t). \end{aligned}$$

Whilst section 6.1 will focus on exploring the base model and how it can be adapted to incorporate different elements like a **local linear trend**—a key time series concept and important for the types of data that may be considered in an epidemiological situation—section 6.2 will help us consider the inference needed to evaluate the impact of an intervention, achieving quantifiable results like the **running average effect**.

As we draw the conclusions of this paper, we will take one last detour to consider how the implementation of causal inference comes with some hefty **ethical considerations** both in an epidemiological setting in section 7.1, as well as in section 7.2 when artificial intelligence (AI) is incorporated into the picture.

Part II
Theory

"The new science does not have a fancy name: I call it simply "causal inference," as do many of my colleagues."

Judea Pearl [28]

As we embark on the theory surrounding causal inference, we begin thinking about the graph theory that we need to build our models. Most individuals are able to physically draw out how one factor affects another, whilst few are able to comprehend how we may write a set of equations to represent the same relationships. Therefore we tackle the bridge between the two.

1 Graphical Modelling

It is firstly important to address why simple linear equations don't quite cut it when trying to model causal relationships and why we need to deploy structural equations instead.

Definition 1.1: Linear Equation

For an input variable X and output variable Y we say that there is a linear relationship between the two such that

$$y = \beta x + \alpha,$$

where β quantifies the influence of the input variable on the output, and α represents the additional influences in the system that can be attributed to affect the output.

We must realise that we cannot model a causal relationship using a linear relationship between two variables. For example if we choose y to represent a symptom of a disease whilst x is the disease and write this as a linear relationship it could easily be rearranged and interpreted that the symptoms cause the disease. We know this however is incorrect.

This realisation is one that has been consistent throughout history with causal modelling, with individuals such as Wright [41] still trying to utilise linear equations but assisted by directed graphs. From the discussion which is about to follow we see why the graphs became the easier way to model and why linear equations were thrown out.

1.1 Variable Relations

As identified in our argument as to why linear equations cannot be used when we have a system of variables, some have influence over others but not necessarily vice-versa. We would do well therefore to introduce the following two classes of variables, which can be seen in the work of Pearl [25].

Definition 1.2: Exogenous Variable

We call a causal variable influencing another variable, that we wish to leave unexplained and independent of other variables in the system, exogenous and denote it as following,

$$x_i = U_{x_i}.$$

Definition 1.3: Endogenous Variable

A factor within a causal model is endogenous if it's outcome is dependent on other variables within the system.

$$X = f(x_1, ..., x_n)$$

The following short example demonstrates how a 3 variable system could be written using the definitions we have introduced.

Example 1.1. *Say we have 3 endogenous variables A, B and C. Each has a influencing exogenous variable which we will denote U_a, U_b and U_c. By combining the notations of exogenous and endogenous variables we can write this system as the following set of equations:*

$$\begin{aligned} A &= f_a(U_a), \\ B &= f_b(U_b), \\ C &= f_c(U_c). \end{aligned}$$

The example above of course is the simplest system one could have. The complexity is added when the endogenous variables start to also depend on one another. One may start to wonder—"how can I quantify the strength of the relationship between two variables." This is where we introduce the concept of covariance.

Definition 1.4: Covariance

Let X and Y be two random variables. We define the covariance between the two as the difference between the expected value of the variables multiplied and the expected values of each variable multiplied. This is written in the following manner,

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

It is important to be able to consider what form our covariance might take in a causal relationship. For a linear relationship between two variables this will be a constant value. Therefore we define this constant as the path coefficient which quantifies the causal influence of X on Y. To set a convention we will use the Greek alphabet.

Definition 1.5: Path Coefficient

$$\text{Cov}(X, Y) = \alpha, \quad \alpha \in \mathbb{R}$$

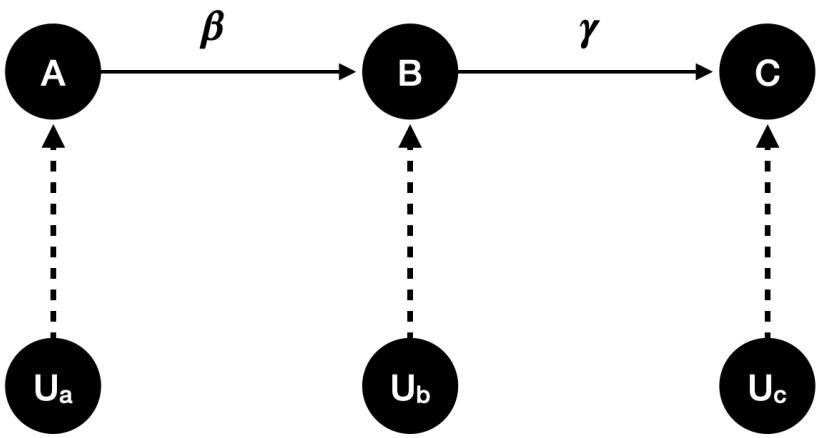
The use of the term path coefficient may be unfamiliar to most readers and we will examine the term further by considering the form they may take later in the paper. Below we return to the previous example, but look to develop a path coefficient between the endogenous variables.

Example 1.1 (Continued)**.** *Say now that the we know there is a link between variables A and B and that path coefficient from A to B is β. There is also now a link between variables B and C with the path coefficient from B to C, γ. We can extend the dependencies in our functions, and write them as follows:*

$$\begin{aligned} A &= f_a(U_a) \\ B &= f_b(A, U_b) \\ C &= f_c(B, U_c) \end{aligned}$$

Alternatively, and perhaps much more usefully we can introduce the concept of a causal graph. A scenario which has a direction as well as variables which cannot cause itself, either directly or through another variable, can be represented by a directed acyclic graph. We give below the causal graph for Example 1.1, before proceeding onto explaining the terminology.

Figure 2: Causal Diagram for Example 1.1



1.2 Directed Acyclic Graphs

It can be seen that figure 2 is in fact a directed acyclic graph. Let us formally introduce the definition for this before looking at how it links into our causal inference material. A compound explanation for the definition and corollary below can be found in the work of Hernán and Roberts [14].

Definition 1.6: Directed Acyclic Graph (DAG)

A graph, $\mathcal{G}$, whose nodes (vertices) are random variables $X = (X_1, ..., X_M)$ with direct edges and no directed cycles is a directed acyclic graph.

Corollary 1.7: Causal DAG

A causal DAG is a DAG in which;
(i) A lack of an arrow from X_j to X_m can be interpreted as the absence of a direct causal effect of X_j on X_m relative to the other variables on the graph.
(ii) All common causes, even if unmeasured, of any pair of variables on the graph are themselves on the graph.
(iii) Any variable is a cause of its descendants.

Most of the definitions above are intuitive in the way that we do not imagine a causal diagram that has a variable which causes itself. When drawing out our graph of 1.1, we used a solid arrow line to represent the relationship between two endogenous variables and a dashed arrow line to signify the link between an exogenous to an endogenous. Within the corollary above we use the term 'descendants'. It is worth us adding in some basic definitions of how we can group sets of nodes based on their connections.

Definition 1.8: Parent and Child

Given (v, w) we write v → w to say v is a parent of w, and conversely that w is a child of v.

The set of parents is denoted $pa_{\mathcal{G}}(w)$.
The set of children is denoted $ch_{\mathcal{G}}(v)$.

Definition 1.9: Ancestor and Descendant

X is an ancestor of v if either $x = v$ or there is a directed path

$$x \to ... \to v.$$

Even more so, Y is a descendant of v if there is a directed path

$$v \to ... \to y.$$

The set of ancestors is denoted $an_{\mathcal{G}}(v)$
The set of descendants is denoted $de_{\mathcal{G}}(v)$

2 Independence and Intervention

We've shown how to express that a variable (either exogenous or endogenous) has influence over another either through encoding or a causal diagram. The next question that naturally arises is: how do we express more formally that a variable has no influence over another? For example in figure 2 it can be observed that C has no influence over A regardless of the choice of B. This is where we introduce 'd-separation' which allows us to conclude that A is independent of C given B. We will do this in a few steps, firstly by the introduction of some notation.

Notation 2.1. *Conditional Independence*
To denote that two random variables X and Z are independent given a third variable Y we write

$$X \perp\!\!\!\perp Z | Y$$

The notion of conditional independence relies heavily on what is going on in this third given variable, Y in our circumstance. Hence, with the aid of the work of Greenland and Pearl [11] we think about what it means to be path blocking in order to build into our d-seperation concept.

Definition 2.1: Path Blocking

A path P is said to be blocked by a set of nodes S if and only if:

(i) P contains a chain $X \to Y \to Z$, or a fork $X \leftarrow Y \to X$ such that the middle node Y is in S

or

(ii) P contains an inverted fork $X \to Y \leftarrow Z$ such that neither the middle node Y nor any of its descendants are in S

This now allows us to introduce the more simplistic definition for d-seperation.

Definition 2.2: d-seperation

If the set of nodes S are path blocking to all paths from X to Z we say that X is d-seperate from Z. Furthermore we conclude that X is independent of Z given S,

$$X \perp\!\!\!\perp Z | S.$$

We will not explore the theory behind d-seperation and how we can be sure that this works in causal inference. For a reader who is intrigued about the soundness of the concept—that is to say whether we can be sure that given the set of nodes S that X

and Z truly are conditionally independent on S—then I would recommend looking at the work presented by Verma and Pearl [40]. Additionally for a logic based argument that shows strong completeness of d-seperation, one can look at the work of Meek [19].

Now that we've addressed how to identify that one variable is d-separated (independent) from another variable given a set of nodes S in the system, we must now think about how we can use this to our advantage. Namely how do we intervene to make two variables d-separated?

Let us firstly consider some motivation for wanting to do this. If we have a variable representing abnormal symptoms A, along with a variable to represent all diseases B and finally a variable representing the severity of the disease, C. We may want to investigate the proportion of individuals who have abnormal symptoms of a set disease and how severe they are. We do this by using the do-operator which holds B to a set disease b_0.

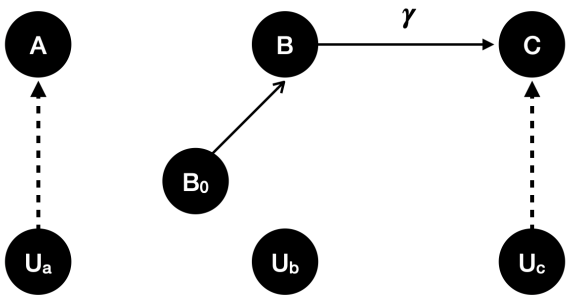
Definition 2.3: do-operator

If we want to physically intervene by holding one variable constant, that is $X = x_0$ we write do(x_0)

We can now have a look again at a basic three model structure in both functional and graphical form.

Example 2.1. *Revisiting the basic six variable structure in terms of the disease scenario set up earlier, we may combine it with the disease scenario that we motivated the definition above with. Our causal diagram would be drawn in the following way:*

Figure 3: 6 Variable Causal Diagram with Intervention for Example 2.1



This would come with the set of equations for our three endogenous variables which demonstrate how the system looks after making the intervention do(b_0).

$$A = f_a(U_a),$$
$$B = b_0,$$
$$C = f_c(B, U_c).$$

It's important to start to now think about how this links to distributions. By firstly defining a probability distribution and then how an intervention alters this. We do this by a small segue into conditional probability.

2.1 Conditional Probability

Let us firstly remind ourselves of some of the more basic probability theory in the form of probability distributions and independence.

Definition 2.4: Probability Distribution

A variable x can be distributed by a probability measure on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and denoted in the following way,

$$p_x(x) = \mathbb{P}(\{X = x\}).$$

Definition 2.5: Independence

Two variables X and Y, with density functions f_x and f_y have a joint density function f_{xy}. If the two variables are independent it holds that

$$f_{xy} = f_x f_y \text{ for all } x \in X, y \in Y$$

This is equivalent to saying...

$$\mathbb{P}(X = x | Y = y) = \mathbb{P}(X = x)$$

Now that we have introduced the notation and terminology for a probability distribution and independence, we can combine these concepts in order to state the form of a conditional probability function. Additionally we also consider what conditional independence looks like in a system of three variables.

Definition 2.6: Conditional Probability Function

Any probability function $p(x|y)$ which may defined in the following way is the conditional probability function,

$$p(x, y) = p(y) \times p(x|y).$$

Definition 2.7: Conditional Independence

For three variables, A, B and C we may say that X and Y are conditionally independent given Z when,

$$p(x|y, z) = p(x|z)$$

We have already seen in the notation point 2.1 that we can denote conditional independence as $X \perp\!\!\!\perp Z|Y$. Below I give a summary theorem of graphoid relationships, as detailed in the earlier work of Pearl and Paz [29]. One should try to consider what each part of the theorem tells us and how they are useful as a system of variables exponentially grows, but not to get too entangled in the detailing for the purposes of this paper.

Theorem 2.1: Graphoid Relationships

1. $X \perp\!\!\!\perp Y|Z \Longrightarrow Y \perp\!\!\!\perp X|Z$
2. $X \perp\!\!\!\perp Y, W|Z \Longrightarrow X \perp\!\!\!\perp Y|Z$
3. $X \perp\!\!\!\perp Y, W|Z \Longrightarrow X \perp\!\!\!\perp W|Y, Z$
4. $X \perp\!\!\!\perp W|Y, Z$ and $X \perp\!\!\!\perp Y|Z \Longrightarrow X \perp\!\!\!\perp Y, W|Z$
5. if $\mathbb{P}(x, y, z, w) > 0$ then $X \perp\!\!\!\perp W|Y, Z$ and $X \perp\!\!\!\perp Y|W, Z \Longrightarrow X \perp\!\!\!\perp Y, W|Z$

Proof. Some of these relationships may seem logical whilst others may require a little more guidance to get to. Therefore here I have outlined a sketch proof to assist the reader.

1. This follows from the symmetry of independence.

2. Here consider probabilistically

$$\mathbb{P}(x, y, w|z) = \mathbb{P}(x|z)\mathbb{P}(y, w|z).$$

Once we 'integrate out' w, that is to say consider all ranges of w, we then have,

$$\mathbb{P}(x, y|z) = \mathbb{P}(x|z)\mathbb{P}(y|z).$$

This gives us the relationship.

3. For this graphoid relationship we recognise that the left hand side is claiming that we can decompose the probability function as follows:

$$p(x, y, w, z) = f(x, z)g(y, w, z).$$

This in itself implies that there is a relationship which exists where W when independent of X is conditioned on Y and Z.

4. The relationship presented here completes the two way connection with property 2 and 3. That is to say on the left hand side we can understand that

$$\mathbb{P}(w|x, y, z) = \mathbb{P}(w|y, z) \quad \text{and} \quad \mathbb{P}(y|x, z) = \mathbb{P}(y|z).$$

Multiplying these together and using the rules we know about conditional independence we have then

$$\mathbb{P}(w, y|x, z) = \mathbb{P}(w, y|z).$$

This leaves us with the expression.

5. This expression combines earlier relationships to say that if we have two things independent of X which are conditioned on both Z and each other, then they are jointly independent of X when conditioned just on Z. □

If we have a system made up of n-endogenous linked variables X_i, each with probability function $p(X_i)$ where $i = \{1, ..., n\}$, we may repetitively apply the conditional independence definition to achieve the following.

Definition 2.8: Causal Markov Condition

Let $\mathcal{G}$ be a DAG with the set of vertices V.

$$p(x_V) = \prod_{v \in V} p(x_v | x_{pa_G(v)})$$

2.2 Controlling Confounders

Proceeding within a probabalistic context, we consider next what happens when we intervene within a causal system. We have seen that when intervening we can utilise the do-operator. More formally this is referred to as controlling confounders—that is in laymans terms, controlling for other factors that may affect the outcome of our variable of interest. We first take to defining an intervention.

Definition 2.9: Intervention

Let $\mathcal{G}$ be a DAG which represents a causal system, along with p a probability distribution acting on set of variables X_V. Calling the operation do($\{x_w \in X\}$) intervenes as follows...
(i) Graphically this is removing all edges of the DAG pointing into x.
(ii) Probabilistically we can say that

$$\begin{aligned} p(x_{V\setminus\{w\}} | \text{do}(x_w)) &= \frac{p(x_V)}{p(x_w | x_{pa(w)})} \\ &= \prod_{v \in V\setminus\{w\}} p(x_v | x_{pa(w)}) \end{aligned}$$

Whilst it can be effective to talk through the notation and probabilistic interpretation of an intervention, I have developed an extended example below to demonstrate the important nature of the theory.

Example 2.2. *We start by imagining a scenario where someone has to decide whether to catch the train to work or find another method of transport, we denote these situations by an indicator t which either can be t=0 (the person doesn't get the train) and t=1 (they catch the train). There is a 50/50 chance they get the train.*

$$p_T(t) = \begin{cases} 0.5 & t = 0 \\ 0.5 & t = 1 \end{cases}$$

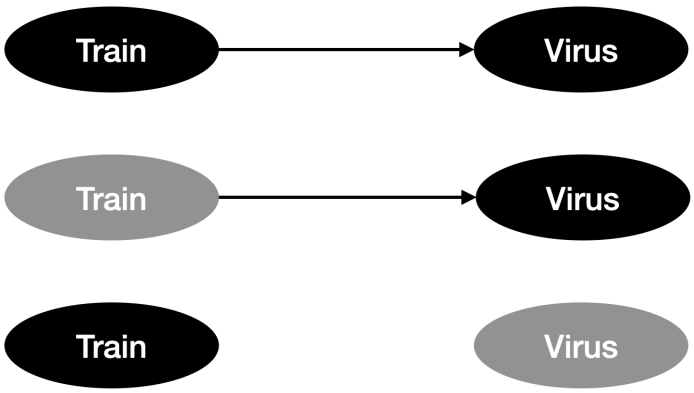
In the world that the individual lives in there is a virus circulating. We will also give the virus an indicator v which can either be v=0 (not contracting the virus) and v=1 (contracting the virus). If they take the train they have a 80% chance of contracting the virus, however if they take other means they have 2% chance of contracting the virus.

$$p_V(v|t = 0) = \begin{cases} 0.98 & v = 0 \\ 0.02 & v = 1 \end{cases}$$

$$p_V(v|t = 1) = \begin{cases} 0.2 & v = 0 \\ 0.8 & v = 1 \end{cases}$$

Here we could ask a question like, is getting the train the cause of contracting the virus? However how can we actually conclude an answer to that question thoroughly. The answer would be to intervene and perform an experiment in which we stop everyone getting the train and observe how many still test positive for the virus. If the train is the primary cause of contracting the virus then we would expect to see rates drop, if catching the train is a result of contracting the virus rates wouldn't change. However intuitively this should seem wrong...and should hopefully draw an important point that in a causal model there is not a symmetric relationship between variables.

Figure 4: Three Potential Setups for the Example 2.2



Above I have included three diagrams, the first displaying the causal link between the train and virus. The next where we have held constant people not getting the train (indicated by the variable being grey) - here the link between the two variables is maintained. The final one demonstrates that if we attempted to hold not getting the virus constant (in practice an impossible task) this would shatter the causal link between the two variables. This is because contracting the virus would not cause you to get the train, hopefully this is intuitive to the reader and at no point will humankind encounter such a rogue virus.

Now let us introduce a third factor, say two different age groups. 18-25 which will be denoted by A=0 and then 26-33 which will be denoted by A=1. Below we have a table which coordinates the three factors for a sample group of 92 people.

	18-25		*26-33*	
	train	*not train*	*train*	*not train*
negative	*30*	*11*	*4*	*2*
positive	*3*	*2*	*27*	*13*

It's important to highlight once again the difference between intervening and simply selecting a group. Say we wished to calculate the probability of testing positive given that the individual didn't get the train simply from the data.

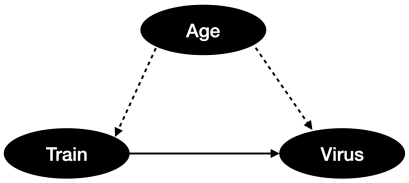
$$\mathbb{P}(v = 1 \,|\, t = 0) = \frac{15}{28} = 0.536$$

Now if we intervene and utilise what we have seen from conditional probability...

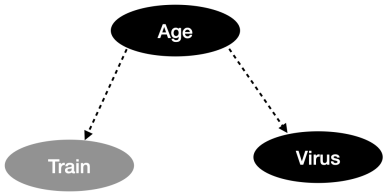
$$\begin{aligned}
\mathbb{P}(v = 1 | do(t = 0)) &= \sum_a \mathbb{P}(v = 1 \,|\, t = 0, A = a)\mathbb{P}(A = a) \\
&= \mathbb{P}(v = 1 \,|\, t = 0, a = 0)\mathbb{P}(a = 0) + \mathbb{P}(v = 1 \,|\, t = 0, a = 1)\mathbb{P}(a = 1) \\
&= \frac{2}{13} \times \frac{1}{2} + \frac{13}{15}\frac{1}{2} \\
&= \frac{199}{390} \\
&= 0.510
\end{aligned}$$

It can be observed that this method produces a lower probability. The difference can be accounted for, in that age changes the chances of an individual getting a virus. Observing individuals who don't get the train can have affects on the chances of them being in a specific age group. However holding not getting the train constant for all in the test doesn't make them more likely to be one age or another.

To complete the example, below we one can observe the two causal diagrams for our set-up. In both it can be seen that age is an exogenous variable which has influence over both the individual getting the train and the individual catching the virus. As by our setup getting the train is the endogenous variable which also has influence over catching the virus. In the second model, the train variable is shown in grey to indicate it has been held constant.



(a) The non-intervention model



(b) The post-intervention model

Figure 5: Intervention Modelling for Example 2.2

Now that we have seen an extended example, we look to other methods for more complicated models which depend on an immeasurable factor. For example say we have a disease which does not affect those who have parents who have at some point been exposed to a certain level of radiation. This sort of generational factor would in most cases be hard to measure.

A way around this is the *back-door adjustment criterion*. This is the process where we hold this immeasurable factor constant and only compare subjects from across different groups who have the same values of the selected factors. Any such factors which we do this with can be put in to an 'admissible set'. Mathematically our motivation for this is so that we can determine, with greater accuracy, the relationship between two nodes regardless of complicated or missing data.

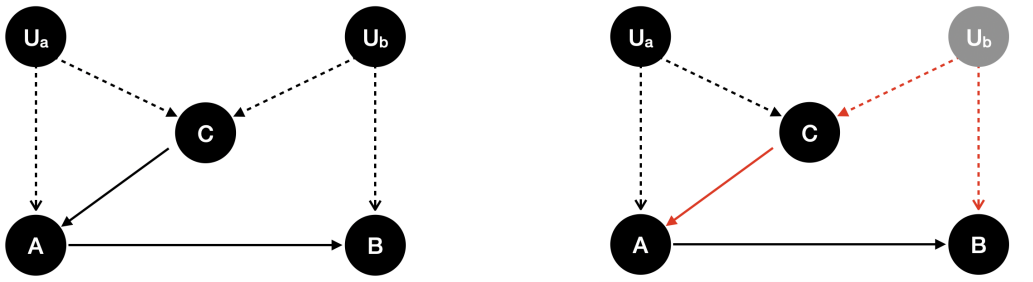
Definition 2.10: Back-Door Adjustment

We say that S is a back-door adjustment set for the ordered pair (v, w) if
(i) no vertex in S is a descendant of v;
(ii) every path from v to w with an arrow into v is blocked by S.

Let us now consider a couple of examples of where employing the back-door criterion is useful and where it is not. The following examples are based on the work presented by Weinburg, later developed by Howards et. al [16].

Example 2.3. *A new strain of a virus is discovered. Regardless of the strain, older age groups are more likely to contract this virus. We wish to investigate the connection between the new viral strain (A) and whether this causes an increase in viral contraction (B). In doing so we also consider location (U_a), age (U_b) and crowding (C). We may represent this as in figure 6a.*

Whilst it may seem obvious, we should look to utilise the back-door criterion and control for age to remove confounding paths. Figure 6b shows the controlled age (U_b) variable in grey with the 'backdoor path' in red that would be deleted due to the backdoor criterion.



(a) The causal model

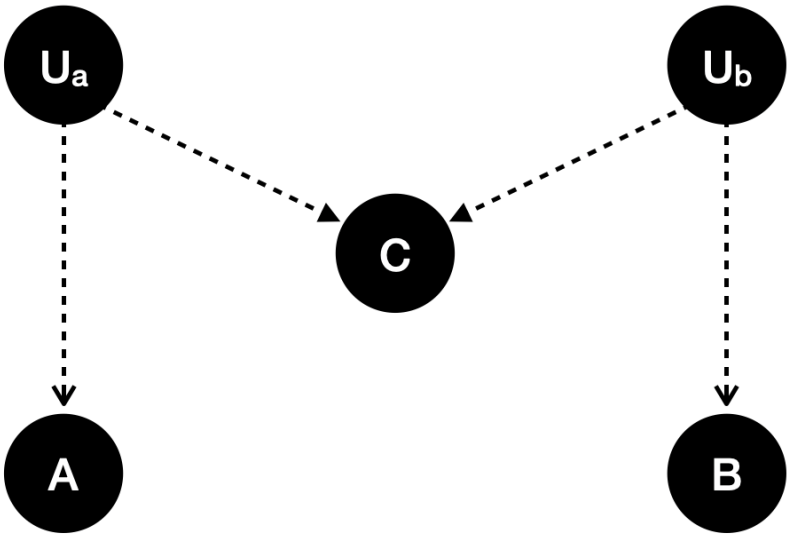
(b) The model back-door adjusted

Figure 6: Causal Diagrams for Example 2.3

In the following example we highlight how one should not necessarily consider a variable to be a confounder simply because it has an association with both the observational variable and the outcome variable that we wish to investigate.

Example 2.4. *Consider the situation where we wish to investigate the causal system which may indirectly link tooth decay (A) and nosebleeds (B). Of course there are many different causal diagrams that could be brought up here but we consider one which involves consuming sugar (U_a), high blood pressure (U_b) and Diabetes (C). Diagrammatically we would draw this in the following way—often refered to as an M-diagram.*

Figure 7: "M" Causal Diagram for Example 2.4



Here the back-door path between tooth decay (A) and nosebleeds (B) is already blocked by the Diabetes variable (C). We hence do not need to apply the back-door criterion. We would be considered in the wrong to count it as a confounder simply because it is associated with both tooth decay and nosebleeds.

Now let us address what this does for our probability function.

Theorem 2.2: Back-Door Probablistic Representation

Let S be the back-door adjustment set for (v, w).

$$p(x_w|do(x_v)) = \sum_{x_S} p(x_S)p(x_w|x_v, x_S)$$

The following proof builds upon one that can be found in Pearl's work [23]. We have utilised the notation and concepts provided in his work.

Proof. Let us firstly divide the variables that sit within our system in to the following; Y, Z, $X_{pa(Z)}$ and X_W - where X_W is any other variable that is not Y, Z or a parent of Z. Wishing to hold Z constant, we may utilise definition 2.9. This leads us to,

$$p(y, x_{pa(z)}, x_W|do(z)) = \frac{p(y, z, x_{pa(z)}, x_W)}{p(z|x_{pa(z)})} = p(y, x_W|z, x_{pa(z)})p(x_{pa(z)})$$

Summing over all the X variables in the equation above we can obtain the following expression,

$$\begin{aligned} p(y|do(z)) &= \sum_{x_W, x_{pa(z)}} p(y, x_W|z, x_{pa(z)})p(x_{pa(z)}) \\ &= \sum_{x_{pa(z)}} p(x_{pa(z)}) \sum_{x_W} p(y, x_W|z, x_{pa(z)}) \\ &= \sum_{x_{pa(z)}} p(x_{pa(z)})p(y|z, x_{pa(z)}) \end{aligned}$$

Give S to be the set of nodes such that w is d-separated from $pa_{\mathcal{G}}(v)$ by $S \cup \{v\}$. Combining this with the above result we can conclude,

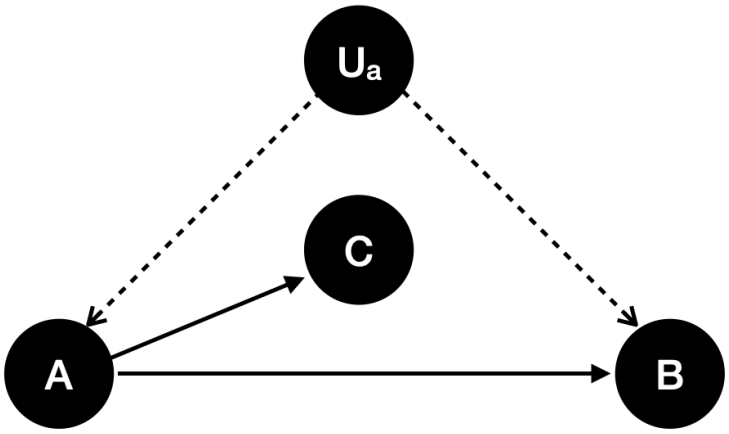
$$\begin{aligned} p(x_w|do(x_v)) &= \sum_{x_{pa(v)}} p(x_{pa(v)})p(x_w|x_v, x_{pa(v)}) \\ &= \sum_{x_{pa(v)}} p(x_{pa(v)}) \sum_{x_S} p(x_w, x_S|x_v, x_{pa(v)}) \\ &= \sum_{x_{pa(v)}} p(x_{pa(v)}) \sum_{x_S} p(x_w|x_S, x_v, x_{pa(v)})p(x_S|x_v, x_{pa(v)}) \\ &= \sum_{x_{pa(v)}} p(x_{pa(v)}) \sum_{x_S} p(x_w|x_S, x_v)p(x_S|x_{pa(v)}) \\ &= \sum_{x_S} p(x_w|x_S, x_v) \sum_{x_{pa(v)}} p(x_{pa(v)})p(x_S|x_{pa(v)}) \\ &= \sum_{x_S} p(x_S)p(x_w|x_v, x_S) \end{aligned}$$

□

Of course removing hard to measure factors is not the only implementation of the back-door criterion. We also want to try and remove bias from our model. You can see the later section 4 to discuss bias in more detail—in essence the back-door criterion removes confounding bias. However when using the criterion for this purpose we must be very careful as we do not wish to introduce an alternative bias as we see in the following example.

Example 2.5. *Consider the epidemiological situation where we wish to instruct individuals with the symptom of an increased temperature to not go to work. We could consider that the weather (U_a) is an exogenous variable of our system which impacts both an individuals temperature (A) as well as whether an individual goes to work (B). We then have in this context that an individuals temperature impacts whether they go to work as well as if they have the virus of concern (C). This forms the following causal diagram.*

Figure 8: Causal Diagram for Example 2.5



Here we would be wise to de-confound (control) the weather variable U_a and look to the back-door criterion to remove it. In the case where we have no way of measuring the weather then one might be inclined to attempt to simplify the system and choose to control the virus variable C in some way. This would not be a good idea however as we will look to explore in section 4.

3 Structural Equation Models

The most common solution to the complex problem of modelling a causal situation is to look to a Structural Equation Model (SEM). We will firstly give the definition of what it means to be an SEM before also discussing some Markovian properties that we can exploit.

Definition 3.1: Structural Equation Model (SEM)

We said that if we have a graph $\mathcal{G}$ and a probability distribution p, which follows the multivariate normal distribution, that we call the the pair ($\mathcal{G}$, p) a structural equation model.

We have that a node in a SEM is Markov if and only if the model can be written recursively with error ϵ. That is to say,

$$X_i = \sum_{j \in \text{pa}_{\mathcal{G}}(i)} b_{ij} X_j + \epsilon_i, \quad \forall i \in V,$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ are independent errors. We denote the covariance matrix of the terms in the system by Σ, and here b_{ij} is the $[i, j]$ th entry to the matrix B, which is lower triangular, given by $B = \Sigma_{ij}(\Sigma_{jj})^{-1}$.

Let us motivate why we would look to work with an SEM which is Markovian, with a 3 variable formulation of a SEM. This will also allow us to see the form that the functions introduced in section 1.1 take.

Example 3.1. *Take a system that has three variables whereby A has influence over B and C, and then B has influence over C. Each variable is measured and comes with an associated error ϵ_i. We may write the system of equations as follows,*

$$\begin{aligned} A &= \epsilon_a, \\ B &= \alpha A + \epsilon_b, \\ C &= \beta A + \gamma B + \epsilon_c. \end{aligned}$$

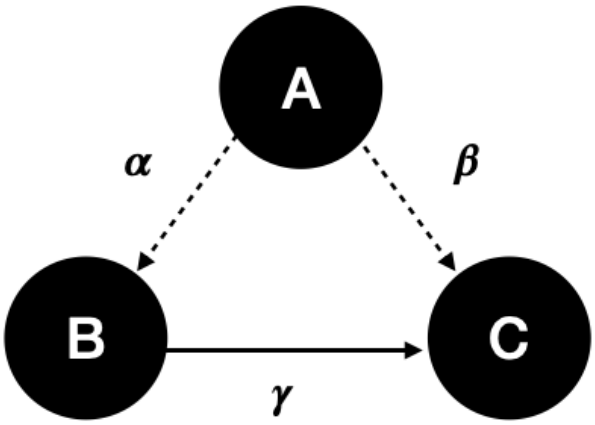
In matrix notation we can write this is as,

$$\begin{pmatrix} A \\ B \\ C \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ \alpha & 0 & 0 \\ \beta & \gamma & 0 \end{pmatrix} \begin{pmatrix} A \\ B \\ C \end{pmatrix} + \begin{pmatrix} \epsilon_x \\ \epsilon_y \\ \epsilon_z \end{pmatrix}.$$

This then becomes...

$$\begin{pmatrix} 1 & 0 & 0 \\ -\alpha & 1 & 0 \\ -\beta & -\gamma & 1 \end{pmatrix} \begin{pmatrix} A \\ B \\ C \end{pmatrix} = \begin{pmatrix} \epsilon_x \\ \epsilon_y \\ \epsilon_z \end{pmatrix}$$

Figure 9: Causal Diagram for Example 3.1



Hence we now achieve

$$\begin{pmatrix} A \\ B \\ C \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \alpha & 1 & 0 \\ \beta + \alpha\gamma & \gamma & 1 \end{pmatrix} \begin{pmatrix} \epsilon_x \\ \epsilon_y \\ \epsilon_z \end{pmatrix}$$

The question arises how do we now interpret these values. Well naturally one should be able to spot that the $\beta + \alpha\gamma$ term in the [1, 3] place in the matrix is a combination of the paths from $A \to C$ and $A \to B \to C$. Therefore each [i, j] entry within the first matrix are the sum of the paths from node i to node j.

One might now wonder, why this is a useful model in terms of causal inference. Well now that we have a way to identify paths between nodes, which we will formalise, in matrix notation, the notion of covariance can be used to identify the strength of a relationship between two variables. This is an incredibly useful measure when looking to identify causal links.

Theorem 3.1: SEM Covariance

We may write the covariance of any graphical model, $\mathcal{G}$, that is Markov in the following matrix notation,

$$\Sigma = \text{Cov}(X) = (I - B)^{-1} D (I - B)^{-T},$$

where I is the identity matrix, B is a lower triangular matrix given by given by $B = \Sigma_{ij}(\Sigma_{jj})^{-1}$ with non-zero entries for b_{ij} when $j \to i$ and D is a diagonal matrix. D is set to be the matrix which is equal to the covariances of the errors, $D = \text{Cov}(\epsilon)$.

As this theorem may not appear completely intuitively to the reader I have included below a sketch proof to guide an individual as to why this result is as such, based off the earlier introduced recursive model.

Proof. It is given that an SEM is Markov if and only if

$$X_i = \sum_{j \in \text{pa}_{\mathscr{G}}(i)} b_{ij} X_j + \epsilon_i, \quad \forall i \in V.$$

In matrix notation we may write this,

$$\begin{aligned} X_V &= BX_V + \epsilon \\ &= (I - B)^{-1}\epsilon \end{aligned}$$

Therefore the next step is to now take the covariance which leaves us with the following,

$$\text{Cov}(X_V) = (I - B)^{-1}(\text{Cov}(\epsilon))(I - B)^{-T}.$$

The final step is to use the denotion of Σ and D, and we achieve our result. □

3.1 Treks

With an increasing amount of variables, the SEM Covariance matrix gets increasingly complicated to calculate. Instead we look to be able to work out one entry of this matrix which links two variables. The work of Sullivant, Talaska, and Draisma [35] allows us to introduce the concept of a trek.

Definition 3.2: Trek

A trek in a graph, $\mathscr{G}$, from x to y is an ordered pair of directed paths (P_1, P_2), where both the nodes x and y are sinks (terminal nodes) for the system and both paths come from a common node z.

We can therefore interpret a trek as being the path without colliders in them. This should ring a bell as this concept has been visited before - in the form of the back-door criterion. Let us revisit the 3 variable SEM introduced in Example 3.1, but now with the notions above.

Example 3.1 (Continued)**.** *Firstly, for the simplicity of the example, we will pressume that the error terms of the variables within the system are uncorrelated. That is to say in this case we set $D = I$. This leaves us with the following,*

$$\begin{aligned} \Sigma &= (I - B)^{-1}(I - B)^{-T} \\ &= \begin{pmatrix} 1 & 0 & 0 \\ \alpha & 1 & 0 \\ \beta + \alpha\gamma & \gamma & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ \alpha & 1 & 0 \\ \beta + \alpha\gamma & \gamma & 1 \end{pmatrix}^T \\ &= \begin{pmatrix} 1 & 0 & 0 \\ \alpha & 1 & 0 \\ \beta + \alpha\gamma & \gamma & 1 \end{pmatrix} \begin{pmatrix} 1 & \alpha & \beta + \alpha\gamma \\ 0 & 1 & \gamma \\ 0 & 0 & 1 \end{pmatrix}. \\ &= \begin{pmatrix} 1 & \alpha & \beta + \alpha\gamma \\ \alpha & 1 + \alpha^2 & \alpha\beta + \gamma + \alpha^2\gamma \\ \beta + \alpha\gamma & \alpha\beta + \gamma + \alpha^2\gamma & 1 + \gamma^2 + \beta^2 + 2\alpha\beta\gamma + \alpha^2\gamma^2 \end{pmatrix} \end{aligned}$$

We can take from above that the covariance of the node A with itself is 1. This is to be expected since the node represents an exogenous variable.

It would seem that in some way the notion of a trek should connect to the entries in the covariance matrix and be easily obtainable. Therefore we look to define the 'trek covariance'.

Definition 3.3: Trek Covariance

Given a trek $T = (P_1, P_2)$ that have a common node z, we call the covariance between the two paths the trek covariance and this is given by the following expression,

$$c(T) = d_{kk} \prod_{(i \to j) \in P_1} b_{ji} \prod_{(x \to y) \in P_2} b_{yx}$$

where d_{kk} is the kth element of the D matrix from theorem 3.1 and b_{ji}, b_{yx} are two different elements of the B matrix.

Now that we have an expression for the covariance of each individual trek, we can sum over these trek covariances to establish the total covariance between two nodes. We present the Trek Rule below, but will refrain from looking into the details too much—the reader should simply be able to take from the result that an element of the covariance matrix, presented in earlier theorem, is the sum of the treks.

Theorem 3.2: Trek Rule

Given the covariance matrix Σ which is Markov with respect to $\mathcal{G}$, then each element of the matrix is given by,

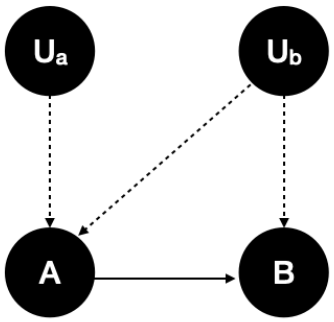
$$\sigma_{ij} = \sum_{T \in \mathcal{T}_{ij}} c(T),$$

where $\mathcal{T}_{ij}$ represents the set of treks between the nodes i and j.

For a proof of the theorem, and to investigate the concept of a trek more, one could look at the work of Robeva and Seby [30]. To summarise the concepts we have come across here we look at an epidemilogical based scenario of an SEM.

Example 3.2. *Consider the scenario where we are trying to see whether a socio-economic intervention on the number of people using non-essential retail affects the spread of a virus. We could assign the variable 'individuals shopping' to A, then the number who 'contract the virus' to B. The endogenous variables U_a could represent 'self-implemented policies' of retailers, and U_b could represent the rules put in place to 'limit the number of individuals in a store'. We say that U_a has a causal link to A, and U_b has a causal link to both A and B. We wish to investigate how much the number of individuals shopping (A) affects the number of individuals who contract the virus (B).*

Figure 10: Causal Diagram for Example 3.2



Above is the causal graph for this set-up.
We could now consider the treks between A and B to quantify the covariance between the two. Below is a table of the treks which exist with their corresponding trek covariance.

Treks from A to B	*Trek Covariances*
$A \to B$	$d_{aa} b_{ba}$
$A \leftarrow U_b \to B$	$d_{U_b U_b} b_{aU_b} b_{bU_b}$
$A \leftarrow U_b \to A \to B$	$d_{U_b U_b} b^2_{aU_b} b_{ba}$
$A \leftarrow U_a \to A \to B$	$d_{U_a U_a} b^2_{aU_a} b_{ba}$

We can now obtain an expression for the covariance between A and B by summing the trek covariances.

$$\begin{aligned} Cov(X_A, X_B) = \sigma_{ab} &= d_{aa} b_{ba} + d_{U_b U_b} b_{aU_b} b_{bU_b} + d_{U_b U_b} b^2_{aU_b} b_{ba} + d_{U_a U_a} b^2_{aU_a} b_{ba} \\ &= b_{ba}(d_{aa} + d_{U_b U_b} b^2_{aU_b} + d_{U_a U_a} b^2_{aU_a}) + d_{U_b U_b} b_{aU_b} b_{bU_b} \end{aligned}$$

Take a moment to consider the treks from A to A. This would be the covariance of A with itself, which is also simplified to the variance of A.

Treks from A to A	*Trek Covariances*
A	d_{aa}
$A \leftarrow U_b \to A$	$d_{U_b U_b} b^2_{aU_b}$
$A \leftarrow U_a \to A$	$d_{U_a U_a} b^2_{aU_a}$

Hence we may simplify our covariance expression between A and B to the following,

$$Cov(X_A, X_B) = \sigma_{ab} = b_{ba} Var(A) + Var(U_b) b_{aU_b} b_{bU_b}.$$

One can observe that the covariance simply depends on the causal path from A to B and the backdoor path that exists in the form of $A \leftarrow U_b \to B$.

4 Bias

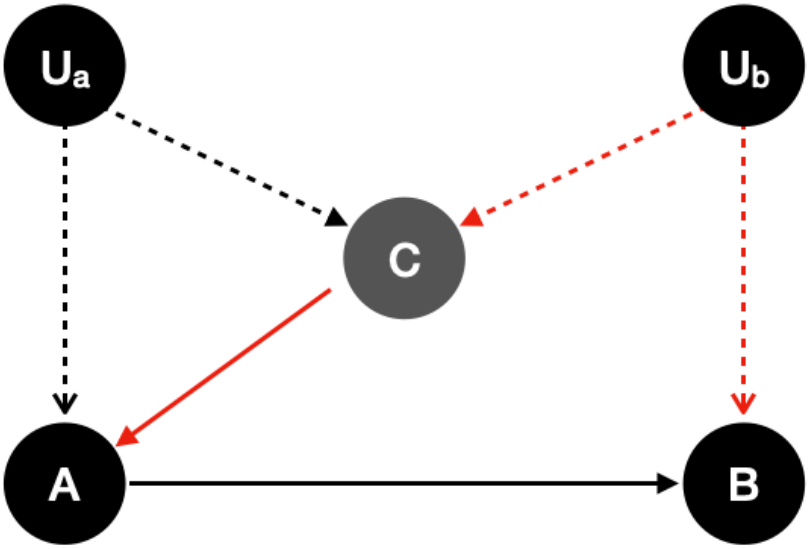
We say that there is bias present in a causal system if the conditional dependence between two variables does not equal the causal effect between the same two variables. It should come as no suprise that both controlling and not controlling variables are ways in which this equallity is thrown off. The first way in which we introduce bias into our system is through conditioning on a descendant of an exogenous variable which we cannot accurately control or which is not necessary to control as it does not block a causal path. We call this overcontrol.

4.1 Overcontrol

Overcontrol bias arises from blocking a causal path from treatment X to outcome Y by conditioning on a descendant of X. If we do this then we end up removing either all or part of the association between X and Y resulting from the total effect of X on Y. To put this into practice we revisit the example that we saw back in Example 2.3.

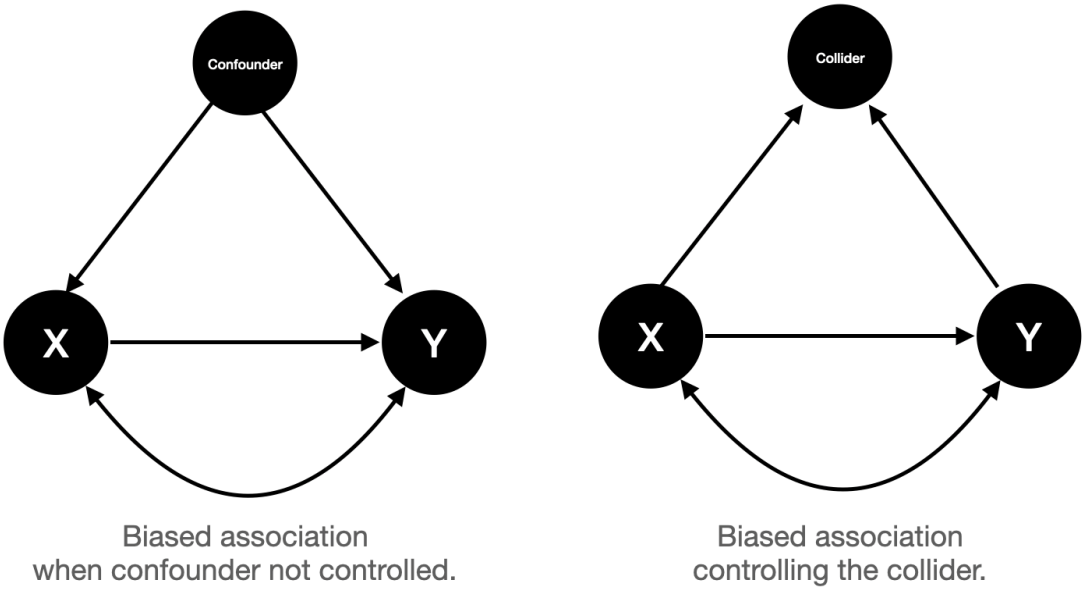
Example 4.1. *Had we chosen originally in our example to control the crowding variable 'C', which we denote in grey for control, we see that the path in red is removed and hence part of the association between the new viral strain and viral contraction is removed. We have introduced overcontrol bias.*

Figure 11: Overcontrolled Causal System



Next we will consider confounding and collider biases. This is where we can introduce biased associations when either controlling the wrong variable or failing to control the right variable. The contrast is shown in the diagram, recreated from the work of Lee, Aronson and Nunan [18], below for ease of comparison.

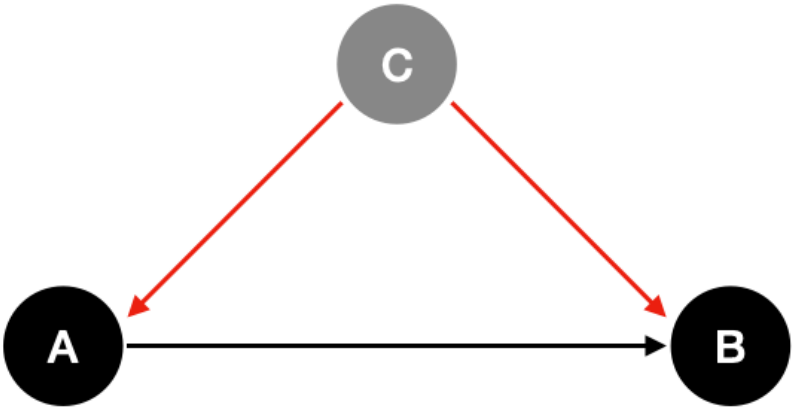
Figure 12: Collider Bias vs Confounder Bias



4.2 Confounding Bias

When we have a common ancestor on both of the variables of interest this introduces confounder bias. We present a basic 3 variable example to see this.

Figure 13: Confounding Bias



Example 4.2. *If it were the case the the confounder C did not exist the only path between treatment A and outcome B would be* $A \to B$*. However we must be careful here because C is in our system an example of a backdoor path. This means if we were to not control C then there would be bias in our system in the form of confounding bias. This can be seen in Figure 13, where the variable in question is in grey and its affects on A and B highlighted in red.*

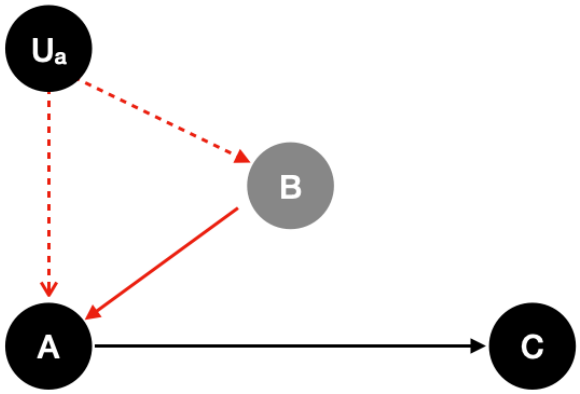
4.3 Collider/Endogenous Selection Bias

When we model, there is an important decision an individual must think about, that of which endogenous variables we include before they start to do more damage than good. That is to say the bias of the model outways the causal impact they have. Collider bias arises from noncausal association due to conditioning on a collider on a path between variable X and outcome Y (or conditioning on the descendant of such a collider). The work of Elwert and Winship [7] addresses how difficult it is in practice to condition on a collider variable and why we need to take extreme caution.

We are able to manipulate our model to reduce collider bias by conditioning on measured non-colliders which have no causal link between the two variables of interest. This is the same adjustment process that we use to reduce confounding bias. Below we present an example where we could introduce collider bias.

Example 4.3. *Consider the following system. Here we see that A is affected by both B and U_a. One might incorrectly try to condition on B, however as demonstrated in the following diagram this would have an indirect effect on the relationship between U_a and A—an introduction of collider bias.*

Figure 14: Introduction of Collider Bias



As we end this section it is worth commenting that it may be difficult to sometimes consider how the different forms of bias differ. One can very much consider the three types to be summarised as followed:

- Overcontrol Bias: Conditioning on a variable on a path between X and Y.
- Confounding Bias: Created by an existing ancestor of X and Y.
- Collider Bias: Conditioning on a variable which is a collider between X and Y.

No large system will ever be entirely without bias, however one can reduce bias by carefully selecting which variables which they control and condition upon.

Part III
Implementation

> "The great advantage of the model-based over the adhoc approach, it seems to me, is that at any given time we know what we are doing."
>
> George Box [3]

Now that we have gathered some appropriate theory surrounding causal inference, this paper takes a turn to looking a little more at the implementation within the field. We will firstly consider the methodology introduced by Pearl [24] to create our causal model then the work of Brodersen [4] in utilising Bayesian Structural Time Series to measure the causal impact of an intervention. To finish this paper we will briefly consider the ethical arguments which surround the implementation of causal inference in both epidemiological settings and also Artificial Intelligence (AI).

5 Methodology

When building a causal model, we follow methodology that is very simialar to many other statistical model building procedures. This will be explained below in a causal inference setting, using the format in which Pearl introduces it.

1.**Define:**
Decide on a relationship which one wishes to focus on, $X \to Y$. From this choose a quantifiable measure (see Section 5.2 for a discussion about potential measures). Express this a function of the model X(M).

2.**Assume:**
Allow the assumption that the system can be modelled using structural equations. Write down the structural equation and graphical form to represent all the casual relationships that may exist between the variables in the system.

3.**Identify:**
Determine whether the target quantity, this is the quantifiable measure from step 1, is identifiable. If it is not, do not proceed and return to step 1 and reconsider the relationship that one wants to investigate.

4.**Calculate:**
If the quantity is identifiable then we proceed by calculating it. There are a couple of ways that we could look to calculate the value of interest: estimation through data or approximation via simulation.

The first two steps should come easily to an individual who has read through part one of this paper. Therefore I will now look to expand on the concept of identifiablity.

5.1 Identifiability

We can see that Pearl [25] defines identifiability in a way which is akin to the following.

Definition 5.1: Identifiability

When two models, M_1 and M_2, equivalently satisfy a set of assumptions encoded in a causal diagram (including the relationships and parameters, P, identified) we say that the quantity of interest, Q is identifiable when

$$P(M_2) = P(M_1) \Longrightarrow Q(M_1) = Q(M_2).$$

That is to say, the details of the two models do not matter, simply the assumptions which limit the variability of the parameterisation. When this is true it follows that when a common parameterisation is applied to either model, the quantity of interest will also be equal regardless of which model is selected.

Whilst this defintion has a very firm mathematical backing, I do believe it may leave some readers not entirely clear on the concept of identifiability in this context. Therefore one could look at a problem as such. Split your model into two divisions: the first, a group of variables which cannot be changed without altering the distribution of the observable variables. The second, the variables which we can alter with no observable consequences. These are the identifiable and unidentifiable parts.

There are a few methods which can be utilised to identify variables. One of which is the back-door criterion (see Definition 2.2). A detailed explanation of why this allows us to conclude identifiability, along with other methods such as the front-door criterion, can be found in the work of Morgan and Winship [20]. We will consider a more simplistic example which will show the commonly used target quantity, the expected value of the resultant variable when holding the influencing variable constant, is indeed identifiable.

Example 5.1. *We consider the system which is set-up by the following equation, and graphically depicted as in Figure 15.*

$$A = f_A(U_a)$$
$$B = f_B(U_b, A)$$

We want to work out what happens when we hold $A = a_0$. This is interventionally done via the do-operator. We would claim in this scenario that

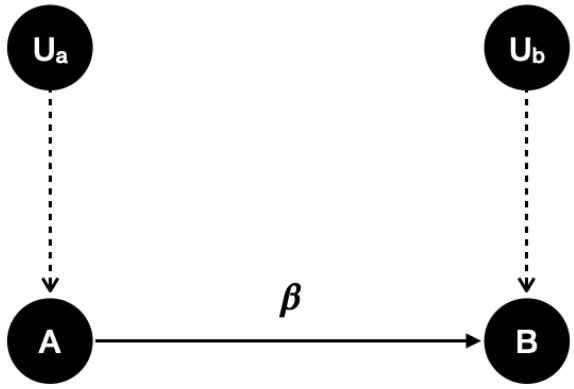
$$\mathbb{E}[B|do(a_0)] = \mathbb{E}[f_B(U_b, A)].$$

Without explicit use of the do-operator this would be achieved by setting $A = a_0$. When looking at the expected value we have,

$$\begin{aligned}\mathbb{E}[B|A = a_0] &= \mathbb{E}[f_B(U_b, A|A = a_0)],\\ &= \mathbb{E}[(f_B(U_b, a_0|A = a_0)],\\ &= \mathbb{E}[f_B(U_b, A)]\\ &= \mathbb{E}[\beta a_0 + U_b]\\ &= \beta a_0\end{aligned}$$

Hence we see here, the value $\mathbb{E}[B|do(a_0)]$ is identifiable and can be estimated by $\mathbb{E}[B|A = a_0]$.

Figure 15: Causal Diagram for Identifiability



The example above showed the expected value of an intervention is identifiable. The most other common measure which we may wish to determine the identifiability of is $\mathbb{P}(Y|\text{do}(x))$. A slightly more involved argument to that shown in Example 5.1 can be used to show this quantity is identifiable. However the theorem presented by Tian and Pearl [38] gives us a simplified condition for this.

Theorem 5.1: Probabalistic Intervention, Identifiability Criterion

A sufficient condition for identifying the causal effect $\mathbb{P}(Y|\text{do}(x))$ is that every path between X and Y (and any of its children) contains at least one arrow which comes away from a measured variable.

We won't look into this theorem, it is rather one included for the completeness of this section. Instead we turn to the more pressing concept of being able to measure an interventional impact once knowing it is identifiable.

5.2 Measuring Interventional Impact

The final part of this methodology section looks to address, perhaps, the biggest question one might ask when using causal inference. How do we measure an interventional impact? The first common method of measuring an interventional impact, in fact one which will appear in Section 6, is the average differencing.

Many causal situations will involve variables which are binary in some way. That is to say they can take one value or another. In the vast manjority of examples in this paper that has been the case—often commenting on whether something makes an individual catch a disease or not. We can thus give the following definition.

Definition 5.2: Average Differencing

For a variable X which can take only two values, either $X = x_0$ or $X = x_1$, we say that the averaged difference impact on the variable Y is given by,

$$\mathbb{E}(Y|\text{do}(X = x_0)) - \mathbb{E}(Y|\text{do}(X = x_1)).$$

A different, yet equally useful way of measuring interventional impact of a binary variable is the averaged Risk Ratio. Once again we will use the conditional expectations of two alternative scenarios. Here the larger the ratio, the more 'impactful' we can deem the intervention at that value compared to the alternative.

Definition 5.3: Average Risk Ratio

For a variable X which can take only two values, either $X = x_0$ or $X = x_1$, we say that the averaged risk ratio on the variable Y is given by,

$$\frac{\mathbb{E}(Y|\text{do}(X = x_0))}{\mathbb{E}(Y|\text{do}(X = x_1))}.$$

6 Bayesian Structural Time Series

Having looked at some of the conventional ways which are employed in measuring interventional impact, what remains is to investigate a way in which may think about calculating these quantities of interest and the causal impact of one variable on another. There are many examples of individuals who have developed extended modelling for this. An interesting example that we will follow in this section is the model developed by Brodersen et. al [4]. We will explore how we can utilise a Bayesian structural time series to measure the causal impact of an intervention. The work below can very much be utilised in epidemiology.

Consider the situation in which we want to measure whether a singular government instruction on social mobility, such as a 'stay at home' order impacts the number of cases of a virus which is in wide circulation. This quantity will be the difference between the observed value of the response, number of viral cases, and the value that would have been obtained had the intervention not been put into place. We will address the inference that is required to obtain these values.

Definition 6.1: Bayesian Structural Time Series

Given an observable variable, Y, we may model the change its change in time via the following,

$$y_t = Z_t^T \alpha_t + \epsilon_t,$$
$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t.$$

$\epsilon_t \sim \mathcal{N}(0, \sigma_t^2)$ and $\eta_t \sim \mathcal{N}(0, Q_t)$ are assumed to be independent of all other unknowns.

The first equation in the definition is known as the 'Observational Equation', which links the observed data y_t (cases of a virus) to a latent d-dimensional state vector α_t. We then have the 'State Equation' which governs the evolution of α_t. The terms in this set of equations can be identified as follows,

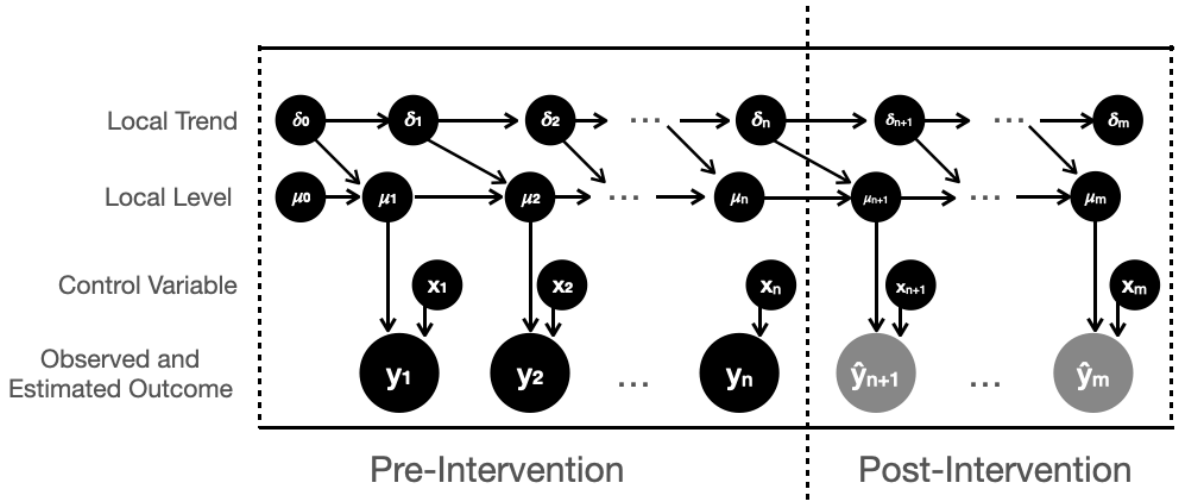
y_t : scalar observation,
Z_t : d-dimensional output vector,
T_t : dxd transition matrix,
R_t : dxq control matrix,
ϵ_t : scalar observational error with noise variance σ_t,
η_t : q-dimensional system error with qxq state diffusion matrix Q_t where $q \leq d$

Within this model we make two key assumptions that are,

- Assume errors of different state component models are independent.
- Covariates are unaffected by the effects of the intervention.

Below I have an included a graphical representation of this model and how it develops both pre and post and intervention. In essence the model builds up to the time of intervention and then makes a predictive outcome of what would have come next had the intervention not occured. This generates a series of unobserved counterfactual responses, $\hat{y}_{n+1}, ..., \hat{y}_m$. One can subtract these values away from the observed data for these points, that is the true values $y_{n+1}, ..., y_m$ in order to determine a probability density, over post-intervential period of time, for the causal impact of the intervention. The model below includes a combination of some of the potential forms of components of the state, which can be explored in the succeeding section.

Figure 16: Bayesian Structural Time Series (simplified)



6.1 Components of State

There are a few different forms that our components of state can take. We will look at two different components of state which I feel are more interesting for the purposes of epidemiological data. Firstly the local linear trend, then the more involved linear regression.

Consider the following pair of equations for a local linear trend.

Definition 6.2: Local Linear Trend

Given that a Bayesian structural time series can be used to model a set of data, we say that the local linear trend is given by,

$$\mu_{t+1} = \mu_t + \delta_t + \eta_{\mu,t},$$
$$\delta_{t+1} = \delta_t + \eta_{\delta,t},$$

where $\eta_{\mu,t} \sim \mathcal{N}(0, \sigma_\mu^2)$ and $\eta_{\delta,t} \sim \mathcal{N}(0, \sigma_\delta^2)$.

The μ_t is the 'trend equation', from which the value of the trend can be read off at any time, t, either pre or post-intervention. We can interpret the equation given by δ_t are the 'slope-equation'. That is to say it gives us the expected increase in the trend between times t and t+1. For a local linear trend we look to set

$$Z_t = [1,0]^T, \quad \alpha_t = (\mu_t, \delta_t)^T, \quad T_t = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad R_t = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \eta_t = (\eta_{\mu,t}, \eta_{\delta,t})^T.$$

It must be noted carefully that we use the local linear trend to model situations which make short-term predictions. This is because a local linear trend allows for stochastic behaviour from the trend and slope. For a model which can cope with a longer term uncertainty we now look at the linear regression model, specifically one with dynamic coefficients.

Definition 6.3: Linear Regression with Dynamic Coefficients

Assuming the covariates in a Bayesian structural time series co-exist, we may write the dynamic regression component as follows,

$$\underline{x}_t^T \beta_t = \sum_{j=1}^{J} x_{j,t} \beta_{j,t},$$
$$\beta_{j,t+1} = \beta_{t,j} + \eta_{\beta,j,t},$$

where $\eta_{\beta,j,t} \sim \mathcal{N}(0, \sigma^2_{\beta_j})$.

For this case we look to set $Z_t = \underline{x_t}$ and $\alpha_t = \beta_t$, with $T_t = R_t = I_{JxJ}$ as well as $Q_t = \text{diag}(\sigma^2_{\beta_j})$. When there is a relationship, as there has been throughout this paper, between the control covariates and the intervention, then we would definitely consider having a linear regression component of state to be the more attractive option.

For the purposes of a reader who is interested in developing their knowledge behind building these models, in particular building models around time series, then the work of Chatfield [5] would prove accessible and develops on these ideas. For our purposes, the above provides an introduction to the models in which the R packages 'bsts' [34] and 'CausalImpact' [4] run on. We now look to delving into the inference which is utilised in our causal model to measure a causal impact of an intervention.

6.2 Evaluating Impact

The first key element to the inference that we carry out when we simulate data to continue the pre-intervention trend in order to compare with the true interventional data is the formation of a prior distribution. The form of the prior in terms of a Bayesian structural time series is a complex process that would perhaps provide too much of an aside for the purposes of this paper. Hence I have tried to provide the reader with a brief summary of the formation that takes place.

Firstly let θ denote the set of all the model parameters.

$$\theta = \{Z, T, R, \eta\} \tag{1}$$

Then let $\alpha = (\alpha_1, ..., \alpha_m)$ denote the full state sequence through time, as instructed by the model. We then may form the following,

- $p(\theta)$ the prior distribution on model parameters.
- $p(\alpha_0|\theta)$ the prior distribution based on the initial state values.
- $p(\alpha, \theta|\underline{y})$ the sampler should we want to simulate data through methods like MCMC.

Having laid out this process for being able to simulate the model parameters and state sequences, we can now think about the three step process in measuring causal impact.

1.**Simulate θ and α:**
Simulate draws of the model parameters θ and state vectors α given the set of data prior to the intervention taking place. That is to say using the set of data $y_1, ..., y_n$.

2.**Posterior Simulation:**
Use posterior simulations to simulate the posterior prediction distribution

$$p(\tilde{y}_{n+1:m}|y_{1:n}).$$

3.**Point-wise Impact Prediction:**
Use the posterior prediction samples to compute the posterior distribution of point-wise impacts

$$y_t - \tilde{y}_t \text{ for each } t = 1, ..., m.$$

Once we have our pointwise impacts and performed n-draws we may denote each sample within each draw in the following way.

$$\phi_t^{(n)} = y_t - \tilde{y}_t^{(n)}$$

From this we are best placed to then look to the running average effect following the intervention given by the summation,

$$\frac{1}{t-n} \sum_{i=n+1}^{t} \phi_i^{(n)} \quad t = n+1, ..., m.$$

There are some limitations in using a running average. As we increase the counterfactual forecast period we approach 0. For the purpose in which we will evaluate impact in a relatively short time period (of a few months) the running average effect is simplistic and useful.

Within this section we have considered a way to calculate the quantity that we wish to investigate within our system. The methods which we have explored above are an example of approximation via simulation.

7 Ethical Considerations

In order to bring this paper to a close, I felt it would be interesting to provide the reader with an overview of the ethical considerations and philosophical questions that have been raised as potential issues with the blunt nature that causal inference enforces when looking at relationships between variables.

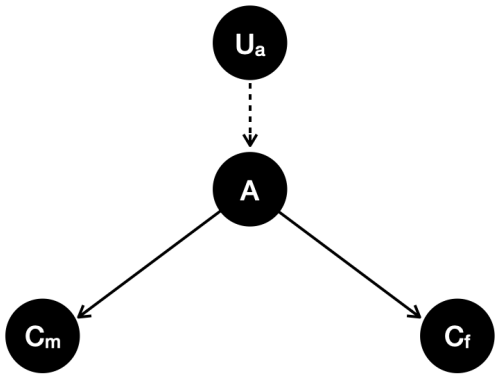
7.1 Epidemiology

A starting point of this is the commentary provided by Glymour [10] around counterfactuals that are intelligible. He provokes here the thought that causal relations are indeed also intelligible.

For a more attacking case, where priori knowledge is not required when considering a causal link and instead that causal relationships arise from experiences, Sanford [33] proves to be an interesting read. Below I have adapted an example provoked by Glymour within epidemiology which relates to this concept that there sits an ethical consideration as to whether conscious and immeasurable choice impacts a causal link.

Example 7.1. *Consider the scenario in which two individuals have to choose the name for a new-born baby. They do not know the sex of the child. With some crudeness we assume the probabilities of the child being a male or a female are equal. An hour before the birth they choose to name the child 'Chris' if it is a boy, and 'Christine' if it is a girl. A potential causal diagram could be the following: Let the chromosomes of the child be denoted U_a and let U_a have influence over the sex of the child, denoted by A. This causes the child to be named one of two names C_m or C_f. The diagram presented*

Figure 17: Intelligent Causal Relations



makes it seem as if this decision is completely inevitable. However we may consider the system to be 'intelligent', where the variables in the system are too complex and cannot be simplified to a computation level. From a philosophical level, the situation above presents the argument around whether we have free will or whether causal determinism exists at all. An argument too vast to discuss here—but also an important concept which an individual should be aware of when utilising causal inference.

Another pressing point to make around an ethical consideration stems from this 'cut and dry' nature. Epidemiologists often will be conducting studies which could influence the socio-economic environment that we live in. The discovery that smoking leads to an increased chance of lung cancer had a huge impact not only the field of medicine but also on implementation of social policies, a journey well told by Ruegg [31].

One could be put under intense pressure, either by funders or political figures, to conclude a result that isn't necessarily thorough or accurately representative of the data which they hold. After all mathematics may be a universal language but it is not easily interpreted by everyone. This has lead to documents such as the Council for International Organisations of Medical Science to create guides [1] which engage those in the argument to always act objectively when conducting studies.

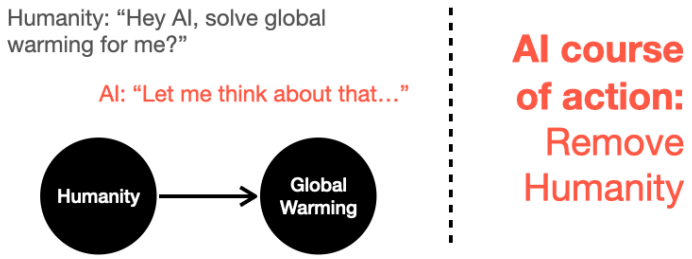
The final ethical consideration surrounding epidemiology that I will mention here is the issue regarding drawing conclusions about a risk factor's causal effect on an individual. There is some heavy literature such as the work of Glass et. al [9], who argue that we should not look to draw direct causal conclusions about risk factors but instead provide estimates of the consequences of an action on a variable.

7.2 Computation and Artificial Intelligence

There is natural logic we follow in saying that the next step towards the development of causal inference will be based around machine learning and mass computation. Indeed the works of leading authors like Pearl have in recent years turned to contemplating how the two interlink [27] and more importantly how should we allow the two to interact?

As computers learn to spot causal patterns, some hefty considerations fall on humans. For example if a computer was set-up to build and learn the causal factors of type 1 diabetes, it may also start to notice that these individuals will always require more medical attention than those who do not. Would an artificial intelligence provide any real guidance on what we could do here or simply provide humanity with the ultimatum of not allowing individual ancestral chains to continue? I consider myself to be part of the beneficial AI movement, in the terminology of Tegmark [36], but still lie wary of the ethical conundrums that we will face as artificial intelligence merges into the causal inference area and how far the decision making ability will lie with AI.

Figure 18: AI Gets Causal Inference Wrong



Part IV
Conclusions

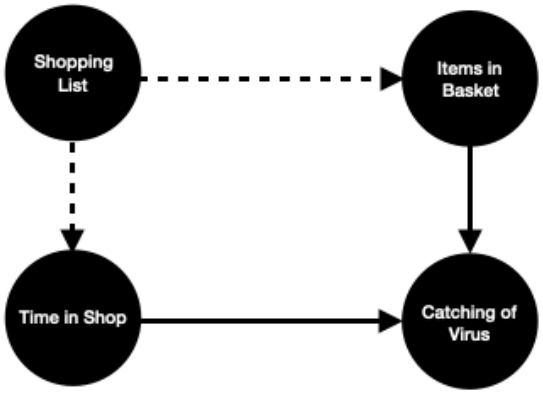
> "A man who sets out to justify his existence and his activities has to distinguish two different questions. The first is whether the work which he does is worth doing; and the second is why he does it, whatever its value may be."
>
> Godfrey Harold Hardy [12]

The quote above taken from Hardy's 'A Mathematicians Apology' resonates greatly with me when approaching the field of causal inference. There are a few key observations that I have made along the way when writing this paper—some have been explicitly mentioned, others simply implied. I give below to the reader three take-away messages surrounding the field of casual inference.

Things Can Go *Very* Wrong

It should come as no shock to an individual, having read in particular Section 4, that when we start to model causal systems we can be very far from the mark. This could perhaps be as early on as not selecting the correct variables to be exogenous and endogenous. Below I present a model one could use to consider an individual catching a virus whilst in a supermarket. It slightly misses the mark—hopefully and having read this paper you are hopefully able to conclude just as intuitively as I, why this is. Perhaps replacing the shopping list and items in basket with meal planning and existing health conditions respectively would improve our model.

Figure 19: An Imperfect Shopping-Virus Model



There Is Not One Set Method

Throughout we have built our theory and our understanding around SEMs in order to define and create assumptions around our causal modelling. There is still much misunderstanding as to what SEMs aim to do, many of which are addressed in Bollen and Pearl's [2] myth busting work. In particular however there is not a universal agreement that this is the best way to approach causal inference.

Other potential methods include looking to agent-based models (AGMs) which can be seen in the work of Murray [21], as well as in a current research project in the Turing Institute, UK—led by Proffessor Heppenstall and Proffessor Gilthorpe.

Where Do We Go From Here?

It should come as no surprise to the reader that the next step we should take in the development and implementation of causal inference is a technological step. Many recent studies have looked to implement the earlier mentioned 'CausalImpact' R package to model socio-economic changes due to the COVID-19 pandemic. The work of Edgar, Mussgnug and Kroll [6] is particularly recent and interesting to look at; they decide to utilise a combination of the afforementioned 'CausalImpact' R package as well as the R package 'MASS' [39] to utilise the linear regression model we introduced in Section 6.1.

The theory of causal inference is still a vastly developing field. The concept of an 'adjustment set' is still widely debated by many, with Henckel, Perkovic and Maathius [13] providing a persuadable commentary. An adjustment set can be thought of as the set of variables which can be adjusted for without altering the integrity of a causal system, which of course is risky business. R packages, 'daggity' [37] and 'pcalg' [17], have been developed to tackle adjustment sets. This could be of interest to the reader, although I add a note of caution with all the R packages mentioned throughout that I cannot speak to their soundness as I have not investigated them in their entirety myself yet.

The concept of the adjustment set of course links in with my earlier discussion around the ethical consideration of computation and AI. If left to its own devices, a computer may decide that the optimal adjustment set does not include a factor which as humans we would deem integral. This is what I am perhaps most conscious of when it comes to developing all these technological methods to work with causal systems. If a machine has the ability to make completely objective decisions about what should and shouldn't be included this, when left without human emotion and reasoning, could end in some terrible decisions being made. All in the name of Causal Inference.

References

[1] BEAUCHAMP, T. L., COOK, R. R., FAYERWEATHER, W. E., RAABE, G. K., THAR, W. E., COWLES, S. R., AND SPIVEY, G. H. Ethical guidelines for epidemiologists. *Journal of clinical epidemiology*, **44** (1991), 151–169. doi:10.1016/0895-4356(91)90192-c.

[2] BOLLEN, K. A. AND PEARL, J. Eight myths about causality and structural equation models. In *Handbook of Causal Analysis for Social Research*, pp. 301–328. Springer Netherlands (2013). doi:10.1007/978-94-007-6094-3_15.

[3] BOX, G. Some problems of statistics and everyday life, presidential address. *Journal of the American Statistical Association*, **74** (1979). doi:10.1080/01621459.1979.10481600.

[4] BRODERSEN, K. H., GALLUSSER, F., KOEHLER, J., REMY, N., AND SCOTT, S. L. Inferring causal impact using Bayesian structural time-series models. *The Annals of Applied Statistics*, **9** (2015), 247–274. doi:10.1214/14-aoas788.

[5] CHATFIELD, C. *The Analysis of Time Series: An Introduction*. Chapman and Hall/CRC, 6th edn. (2003). doi:10.4324/9780203491683.

[6] EDGAR, S., MUSSGNUG, T., AND KROLL, L. Causal analysis of covid-19 observational data in german districts reveals effects of mobility, awareness, and temperature. *medRxiv*, (2020). doi:10.1101/2020.07.15.20154476.

[7] ELWERT, F. AND WINSHIP, C. Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, **40** (2014), 31. doi:10.1146/annurev-soc-071913-043455.

[8] EMERSON, R. W. The conduct of life: Worship. In *The Works of Ralph Waldo Emerson, in 12 vols.*, vol. 6. Fireside edition (Boston and New York, 1909). edn. (1909). Accessed: 16/04/21. Available from: https://oll.libertyfund.org/title/emerson-the-works-of-ralph-waldo-emerson-vol-6-the-conduct-of-life.

[9] GLASS, T. A., GOODMAN, S. N., HERNÁN, M. A., AND SAMET, J. M. Causal inference in public health. *Annual review of public health*, **34** (2013), 61–75. doi:10.1146/annurev-publhealth-031811-124606.

[10] GLYMOUR, C. Statistics and causal inference: Comment: Statistics and metaphysics. *Journal of the American Statistical Association*, **81** (1986), 964–966. doi:10.2307/2289067.

[11] GREENLAND, S. AND PEARL, J. Causal diagrams. In *International Encyclopedia of Statistical Science* (edited by M. Lovric), pp. 208–216. Springer, Berlin, Heidelberg (2011). doi:10.1007/978-3-642-04898-2_162.

[12] HARDY, G. *A Mathematicians Apology*. Cambridge University Press, 1st edn. (1940). ISBN 0521095778.

[13] HENCKEL, L., PERKOVI, E., AND MAATHUIS, M. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. (2019). arXiv: 1907.02435.

[14] HERNÁN, M. A. AND ROBINS, J. M. *Causal Inference: What If*. Chapman and Hall/CRC (2020). Accessed: 16/04/21. Available from: https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1268/2021/03/ciwhatif_hernanrobins_30mar21.pdf.

[15] HILL, S. A. B. The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, **58** (1965), 295. doi:10.1177/003591576505800503.

[16] HOWARDS, P., SCHISTERMAN, E., POOLE, C., KAUFMAN, J., AND WEINBERG, C. Toward a clearer definition of confounding revisited with directed acyclic graphs. *American journal of epidemiology*, **176** (2012), 506. doi:10.1093/aje/kws127.

[17] KALISCH, M., MÄCHLER, M., COLOMBO, D., MAATHUIS, M., AND BÜHLMANN, P. Causal inference using graphical models with the r package pcalg. *Journal of Statistical Software*, **47** (2012), 1. doi:10.18637/jss.v047.i11.

[18] LEE, H., ARONSON, J., AND NUNAN, D. Collider bias. In *Catalogue of bias collaboration* (2019). Accessed: 16/04/21. Available from: https://catalogofbias.org/biases/collider-bias/.

[19] MEEK, C. Strong completeness and faithfulness in bayesian networks. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, p. 411–418. Morgan Kaufmann Publishers Inc. (1995). arXiv:1302.4973.

[20] MORGAN, S. L. AND WINSHIP, C. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Analytical Methods for Social Research. Cambridge University Press, 2nd edn. (2014). doi:10.1017/CBO9781107587991.

[21] MURRAY, E. J. *Agent-Based Models for Causal Inference.* Ph.D. thesis, Harvard T.H. Chan School of Public Health (2016). Accessed: 16/04/21. Available from: http://nrs.harvard.edu/urn-3:HUL.InstRepos:27201721.

[22] PEARL, J. Causal inference without counterfactuals: Comment. *Journal of the American Statistical Association*, **95** (2000), 428–431. doi:10.2307/2669380.

[23] PEARL, J. *Causal Diagrams and the Identification of Causal Effects*, p. 65–106. Cambridge University Press, 2nd edn. (2009). doi:10.1017/CBO9780511803161.005.

[24] PEARL, J. Causal inference in statistics: An overview. *Statistics Surveys*, **3** (2009), 96–146. doi:10.1214/09-SS057.

[25] PEARL, J. *Causality*. Cambridge University Press, 2nd edn. (2009). doi:10.1017/cbo9780511803161.

[26] PEARL, J. The science and ethics of causal modeling. In *Handbook of Ethics in Quantitative Methodology* (edited by A. Panter and S. Serba), pp. 383–414. Taylor and Francis Group (2010). Accessed: 16/04/21. Available from: https://escholarship.org/uc/item/41c9q6z3.

[27] PEARL, J. The seven tools of causal inference, with reflections on machine learning. *Commun. ACM*, **62** (2019), 54–60. doi:10.1145/3241036.

[28] PEARL, J. AND MACKENZIE, D. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., 1st edn. (2018). ISBN 9780141982410.

[29] PEARL, J. AND PAZ, A. Graphoids: A graph-based logic for reasoning about relevance relations or when would x tell you more about y if you already know . In *Proceedings of the 7th European Conference on Artificial Intelligence*, vol. 2 of *ECAI'86*, p. 357–363. North-Holland (1986). Accessed: 16/04/21. Available from: https://ftp.cs.ucla.edu/pub/stat_ser/r53-L.pdf.

[30] ROBEVA, E. AND SEBY, J.-B. Multi-trek separation in linear structural equation models (2020). arXiv:2001.10426.

[31] RUEGG, T. A. Historical perspectives of the causation of lung cancer: Nursing as a bystander. *Global Qualitative Nursing Research*, **2** (2015). doi:10.1177/2333393615585972.

[32] RUSSELL, B. On the notion of cause. *Proceedings of the Aristotelian Society*, **13** (1913), 1–26. doi:10.1093/aristotelian/13.1.1.

[33] SANFORD, D. H. Causation and intelligibility. *Philosophy*, **69** (1994), 55–67. doi:10.1017/S0031819100046611.

[34] SCOTT, S. L. *bsts: Bayesian Structural Time Series* (2020). R package version 0.9.5. Accessed: 16/04/21. Available from: https://CRAN.R-project.org/package=bsts.

[35] SULLIVANT, S., TALASKA, K., AND DRAISMA, J. Trek separation for Gaussian graphical models. *The Annals of Statistics*, **38** (2010), 1665–1685. doi:10.1214/09-aos760.

[36] TEGMARK, M. *Life 3.0: Being Human in the Age of Artificial Intelligence*. Knopf Publishing Group (2017). ISBN 1101970316.

[37] TEXTOR, J., VAN DER ZANDER, B., GILTHORPE, M., LIKIEWICZ, M., AND ELLISON, G. Robust causal inference using directed acyclic graphs: the r package 'dagitty'. *International Journal of Epidemiology*, **45** (2017), 1887. doi:10.1093/ije/dyw341.

[38] TIAN, J. AND PEARL, J. A general identification condition for causal effects. In *Eighteenth National Conference on Artificial Intelligence*, p. 567–573. American Association for Artificial Intelligence (2002). ISBN 0262511290.

[39] VENABLES, W. AND RIPLEY, B. *Modern Applied Statistics with S*. Springer, New York, 4th edn. (2002). doi:10.1007/978-0-387-21706-2.

[40] VERMA, T. S. AND PEARL, J. Causal networks: Semantics and expressiveness. In *Proceedings of the Fourth Conference on Uncertainty in Artificial Intelligence*, vol. 9 of *Machine Intelligence and Pattern Recognition*, pp. 69–76 (1990). doi:10.1016/B978-0-444-88650-7.50011-1.

[41] WRIGHT, S. Correlation and causation. *Journal of Agricultural Research*, **20** (1921), 557–585.