Approaches

to inference about causation in nonexperimental genetically informative studies

David Duffy 20040917

Queensland Institute of Medical Research
Brisbane, Australia

Introduction

There are many philosophical views about **causation**. Hume's (1748) definition of a cause is:

> …an object, followed by another [in time],…where, if the first object had not been, the second had never existed.

An extreme view describes it as a human error: "Beyond such discarded fundamentals as 'matter' and 'force' lies still another fetish amidst the inscrutable arcana of modern science, namely, the category of cause and effect" (Karl Pearson 1911).

In the Popperian view of science, a good scientific model:
a. successfully makes testable predictions about future events (is **falsifiable**),
b. has good explanatory power in terms of mechanism (cause and effect),
c. is consistent with other good scientific models.

This does means that the explanation of cause and effect is contingent scientific knowledge, and may be overthrown by a better model.

Experiments and causation

One ideal type of scientific experiment is a strong test of causality. We **intervene** into an **isolated** system to alter key quantities. If our model is correct, the evolution of the system in time is subsequently altered in the predicted fashion. For the experimenter to be the agent at the head of the causative chain, the intervention must be independent of events within the system. A key method of ensuring independence is **randomisation**.

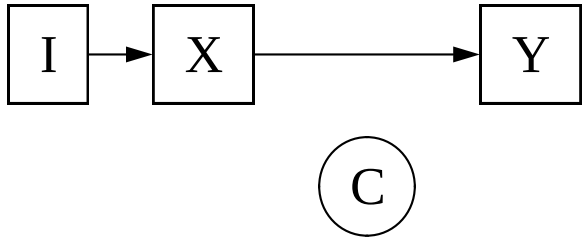
Many experiments in physics are in fact the act of accurate measurement of a quantity eg distribution of the cosmic microwave background. Cause and effect mechanistic explanations are falsified by their failure to **exactly** predict previously unmeasured values. By contrast, the randomised controlled intervention can correctly identify a cause even when the mechanism of cause-and-effect is not known.

Confounding

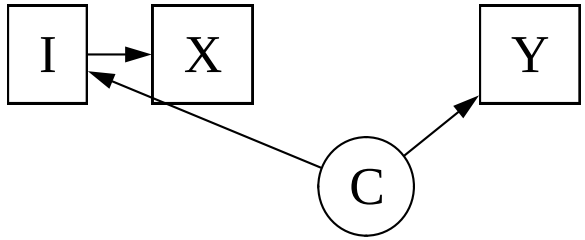
John Stuart Mill (1843), in his book "Of Plurality of Causes, and the intermixture of Effects", talks about the idea of **confounding**, where an observer is confused about the effects of an agent by other "circumstances of the experiment".

In an imperfect experiment (or observational study), we predict a causal relationship between variables *X* and *Y*. We observe association between *X* and *Y*, but this in fact due to changes in another variable *C* that is a cause of both the intervention *I* and *Y*:

What you believe



True State of Nature



Genetics and causation

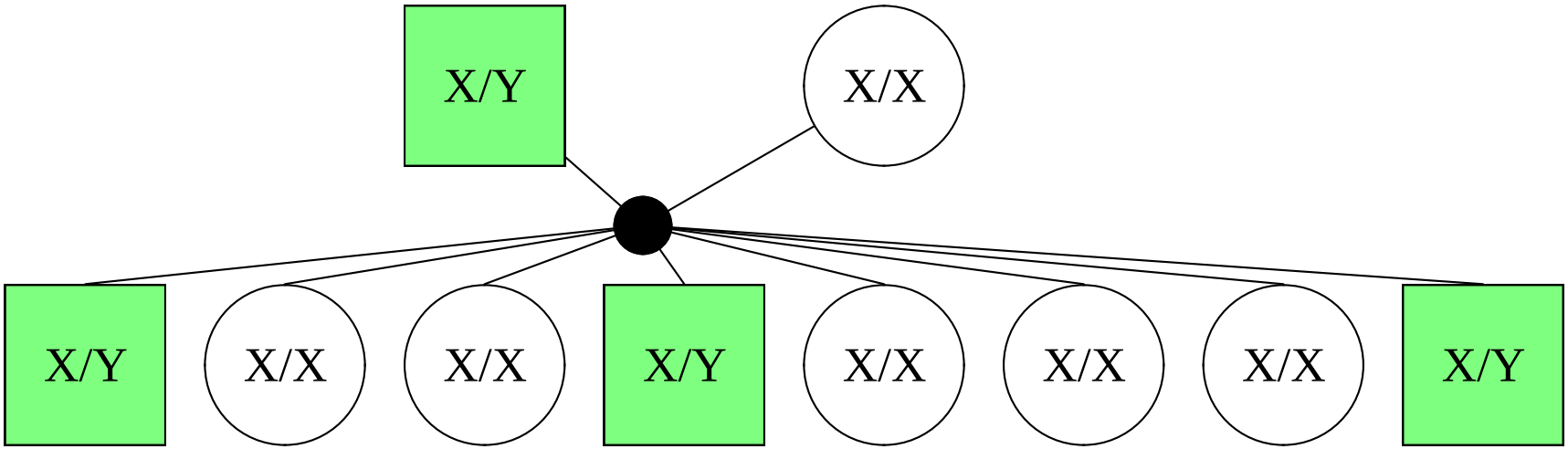
Certain models of the biology of heredity have now withstood 100 years of experimental testing. If we accept these as describing traits we are currently studying,

a. coding variation in genes leads to variation in expressed phenotypes
b. genes are transmitted from parents to offspring with high fidelity
c. Mendelian segregation and recombination are *random*

These facts (assumptions) give us two levers: a detailed model for precise predictions, and a mechanism for randomization in natural experiments.

Genetic linkage analysis and causation

Genetic linkage analysis is a pure test of whether genes cause an observed phenotype, as a test cross is a randomized experiment:



Permutation test of hypothesis that sex and genotype (karyotype) are unassociated: P=0.018

Possible confounders: differential errors in genotyping, sex determination.
Fix: blinded scoring

QIMR

Genetic association analysis and causation

Epidemiologists are interested in **mendelian randomization** so they can assess causation between observed phenotypes, using genotype as an *instrumental variable.* However, they use the term differently to the last pedigree based example, referring to genotypes in the population as a whole.

Minelli and coworkers (2004) note that metaanalysis of traditional epidemiological studies finds that a 3 μM (25%) decrease in serum homocysteine level is associated with an 11% decrease in coronary heart disease (CHD) risk. Fortification of food with folic acid will lower homocysteine level and might be a useful preventative health measure. The estimated strength of association is higher in retrospective studies (eg case control studies), so Minelli worried about **reverse causation**.

Small intervention studies do suggest lowering homocysteine improves brachial artery flow-mediated vasodilatation.

MTHFR genotype, homocysteine and CHD risk

The 677C>T transversion in the MTHFR gene is associated with higher homocysteine levels, therefore an increase in CHD risk would be expected if the **direction of causation** runs from homocysteine level to CHD, and *MTHFR genotype has no other associations with CHD risk*.

	Homocysteine level	CHD Odds Ratio
MTHFR C/C v T/T	2.7 μM (2.02-3.41)	1.21 (1.06-1.40)

Predicted Odds Ratio for CHD per 3 μM increase homocysteine level 1.24 (1.06-1.49).

Possible confounders: Ethnic differences in MTHFR genotype frequency,
multiple effects of MTHFR
Fix: genomic control, Family based association, DOC model

QIMR

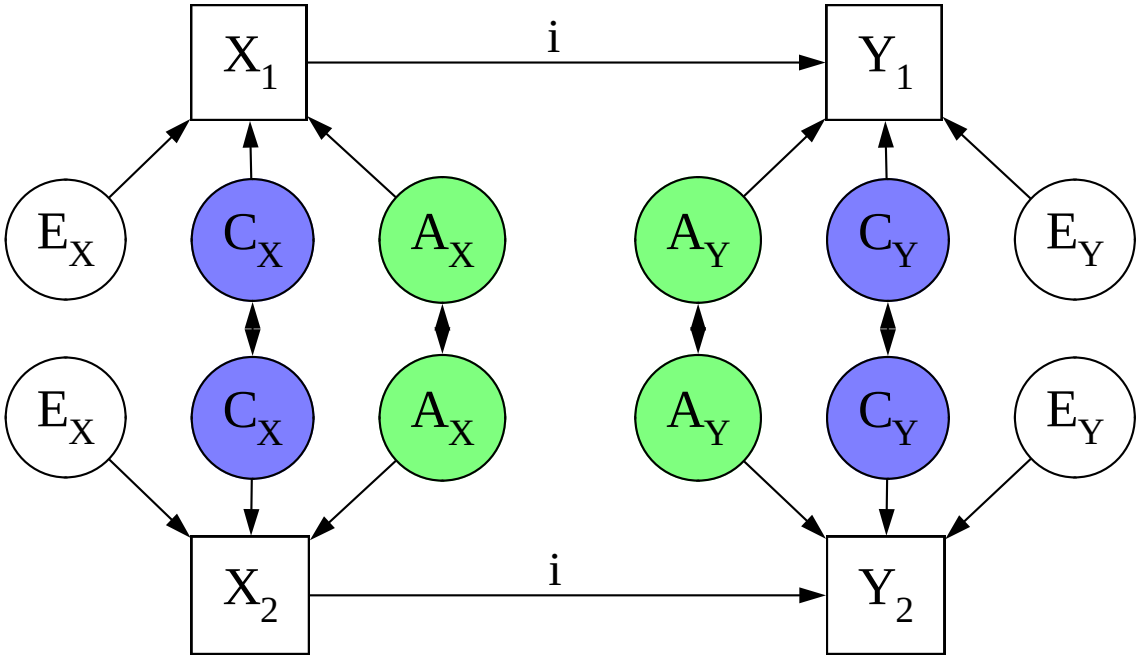
Path analysis and causation

“Systems of correlation coefficients may be dealt with from two radically different points of view: from that of purely statistical description and from that of interpretation in terms of paths of causation. The method of path analysis (Wright 1921) was designed for the purpose of interpretation but becomes identical with the methods of statistical description (multiple regression etc) when the appropriate symmetrical pattern of relations is imposed.” (Wright 1968).

“The method itself depends on the combination of knowledge of degrees of correlation among the variables in a system with such knowledge as may be possessed of the causal relations. In cases in which the causal relations are uncertain, the method can be used to find the logical consequences of any particular hypothesis” (Wright 1921).

"Direction of causation" (DOC) models

Andrew Heath (Heath et al 1989, 1992, 1993) described a cross-sectional observational design to distinguish between alternative theories of causation for the relationship between two phenotypes. It uses pairs of relative of different degrees of relationship. Genotypes are not directly measured.



Relative 1

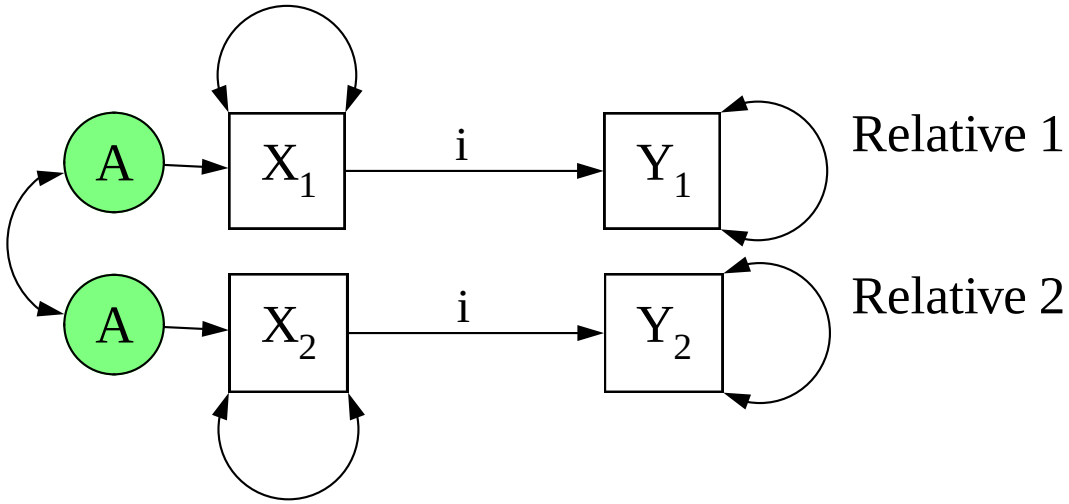
Relative 2

Bivariate ACE model with **phenotypic** causation from X to Y.

QIMR

A simple DOC setup

In the homocysteine-CHD example, if this were the true state of nature,



then

a. X is an intermediate variable in the pathway from the genes to Y,
b. a genetic correlation between Y_1 and Y_2 is induced
c. a correlation between Y_1 and X_2 is induced
d. conditioning on X would abolish the Y_1-Y_2 correlation

DOC expected covariances in simple model

One can then test alternative hypotheses about the direction of causation:

X causes Y

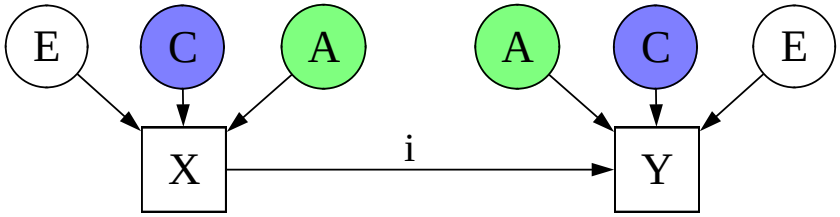
	X_1	Y_1	X_2	Y_2
X_1	1			
Y_1	i	1		
X_2	Rh^2	iRh^2	1	
Y_2	iRh^2	i^2Rh^2	i	1

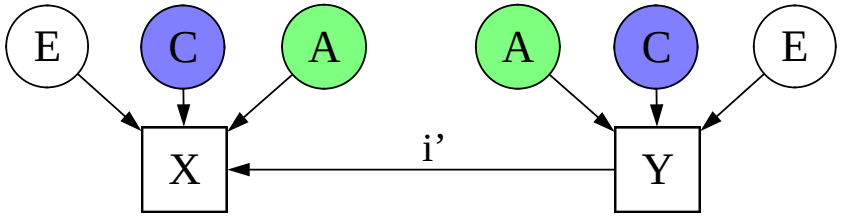
Y causes X

	X_1	Y_1	X_2	Y_2
X_1	1			
Y_1	i	1		
X_2	Rh^2	0	1	
Y_2	0	0	i	1

Reciprocal causation (mutual instantaneous or lagged feedback) is another possible model. This can be the base model for hierarchical testing of phenotypic causation models, and can be easily fitted in fine software such as **Mx**.

DOC expected covariances more general model





	X_1	Y_1	X_2	Y_2
X_1	1			
Y_1	i	1		
X_2	$Rh_x^2 + c_x^2$	$i(Rh_x^2 + c_x^2)$	1	
Y_2	$i(Rh_x^2 + c_x^2)$	$Rh_y^2 + c_y^2 + i^2 (Rh_x^2 + c_x^2)$	i	1

	X_1	Y_1	X_2	Y_2
X_1	1			
Y_1	i	1		
X_2	$Rh_x^2 + c_x^2 + i^2 (Rh_y^2 + c_y^2)$	$i(Rh_y^2 + c_{y)}^2$	1	
Y_2	$i(Rh_y^2 + c_{y)}^2$	$Rh_y^2 + c_y^2$	i	1

QIMR

Identifiability and power

This approach is strongest in situations such as the first example, where the sources of genetic determination are quite different.

If the two traits are similar in heritability, then falsification power is poor (resolution depends on differences in total variances for each trait under the alternatives).

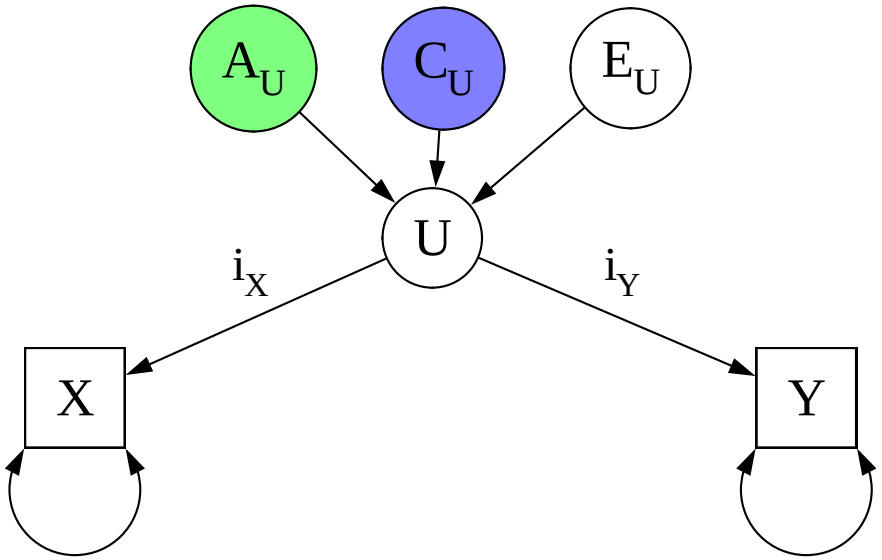
Identifiability and power

Probability of failing to reject a false model (α = 0.05, 150 MZ and 150 DZ twin pairs). Excerpted from Table 1 (Duffy & Martin 1994).

True Model	h_X^2	h_Y^2	Incorrect Model	*Pr*(Fail to Reject Incorrect Model)
X $\rightarrow$ Y, i=0.25	0.1	0.8	Gen Factor model	0.00
			Env Factor model	0.92
			X $\rightarrow$ Y	0.00
	0.5	0.8	Gen Factor model	0.00
			Env Factor model	0.28
			X $\rightarrow$ Y	0.03
	0.8	0.8	Gen Factor model	0.04
			Env Factor model	0.01
			X $\rightarrow$ Y	0.87

Complications in interpretation of DOC models I

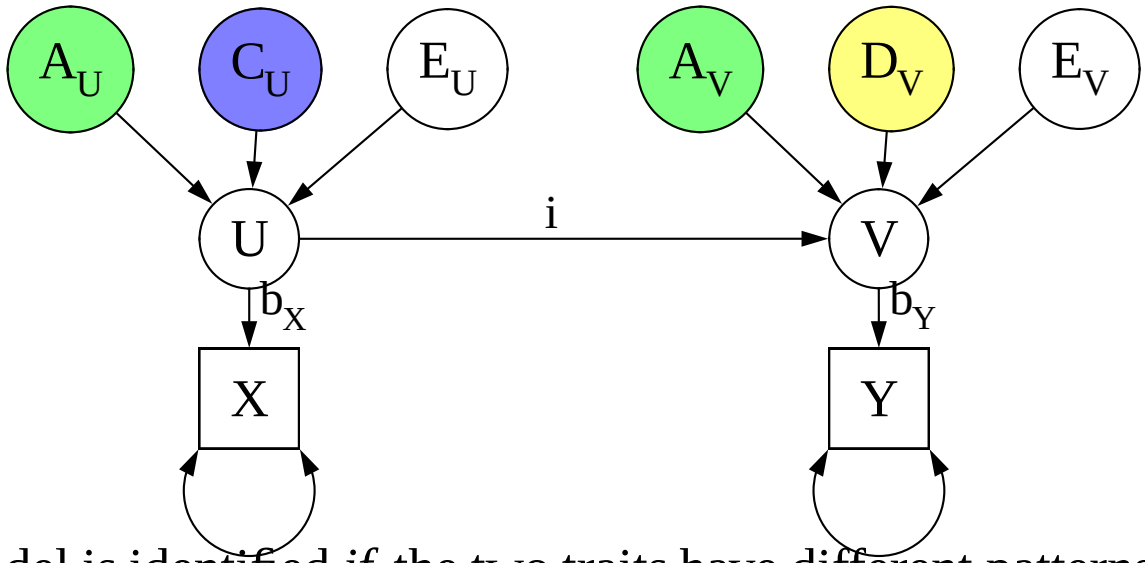
If true state of nature is that a more proximal cause for both phenotypes exist, then the more heritable trait will appear as the causative trait if a phenotypic causation model is fitted.



If $i_X > i_Y$, then X is a better indicator of the true cause than Y is.

Complications in interpretation of DOC models II

If the trait is measured imperfectly, this must be explicitly modelled in the DOC model, else the heritabilities are attenuated, and so too the evidence used to assess causation.



This type of model is identified *if* the two traits have different patterns of genetic determination.