Causal inference in drug discovery

NOVAMATH Thematic Weeks 2024

Short Course Lecture 1

Tom Michoel

18 June 2024

Outline of the course

Today

- A (very) brief history of causal inference
- A brief review of causal inference and applications in drug discovery and development
- Mendelian randomization
- Causal model selection
- A blessing of dimensionality

Tomorrow

- A crash course in genetics and molecular biology
- Causal inference in systems genetics
- Causal gene regulatory network reconstruction and validation

Part I

Correlation and Causation at 100

CORRELATION AND CAUSATION

By SEWALL WRIGHT

Senior Animal Husbandman in Animal Genetics, Bureau of Animal Industry, United States Department of Agriculture

PART I. METHOD OF PATH COEFFICIENTS

INTRODUCTION

The ideal method of science is the study of the direct influence of one condition on another in experiments in which all other possible causes of variation are eliminated. Unfortunately, causes of variation often seem to be beyond control. In the biological sciences, especially, one often has to deal with a group of characteristics or conditions which are correlated because of a complex of interacting, uncontrollable, and often obscure causes. The degree of correlation between two variables can be calculated by well-known methods, but when it is found it gives merely the resultant of all connecting paths of influence.

The present paper is an attempt to present a method of measuring the direct influence along each separate path in such a system and thus of finding the degree to which variation of a given effect is determined by each particular cause. The method depends on the combination of knowledge of the 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 in regard to them.

CORRELATION

Relations between variables which can be measured quantitatively are usually expressed in terms of Galton's (*4*)[1] coefficient of correlation, $r_{XY}=\frac{\Sigma X'Y'}{n\sigma_X\sigma_Y}$ (the ratio of the average product of deviations of X and Y to the product of their standard deviations), or of Pearson's (*7*) correlation ratio, $\eta_{X\cdot Y}=\frac{\sigma_{\left(M_{Y_X}\right)}}{\sigma_X}$ (the ratio of the standard deviation of the mean values of X for each value of Y to the total standard deviation of X), the standard deviation being the square root of the mean square deviation.

Use of the coefficient of correlation (r) assumes that there is a linear relation between the two variables—that is, that a given change in one variable always involves a certain constant change in the corresponding average value of the other. The value of the coefficient can never exceed

[1] Reference is made by number (italic) to "Literature cited," p. 585.

Journal of Agricultural Research,
Washington, D. C.
wh

Vol. XX, No. 7
Jan. 3, 1921
Key No. A-55

17777°—21——4

CORRELATION AND CAUSATION

By SEWALL WRIGHT

Senior Animal Husbandman in Animal Genetics, Bureau of Animal Industry, United States Department of Agriculture

PART I. METHOD OF PATH COEFFICIENTS

INTRODUCTION

The ideal method of science is the study of the direct influence of one condition on another in experiments in which all other possible causes of variation are eliminated. Unfortunately, causes of variation often seem to be beyond control. In the biological sciences, especially, one often has to deal with a group of characteristics or conditions which are correlated because of a complex of interacting, uncontrollable, and often obscure causes. The degree of correlation between two variables can be calculated by well-known methods, but when it is found it gives merely the resultant of all connecting paths of influence.

The present paper is an attempt to present a method of measuring the direct influence along each separate path in such a system and thus of finding the degree to which variation of a given effect is determined by each particular cause. The method depends on the combination of knowledge of the degrees of correlation among the variables in a system with such knowledge 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 in regard to them.

CORRELATION

Relations between variables which can be measured quantitatively are usually expressed in terms of Galton's (4)[1] coefficient of correlation, $r_{XY} = \frac{\Sigma X'Y'}{n\sigma_X\sigma_Y}$ (the ratio of the average product of deviations of X and Y to the product of their standard deviations), or of Pearson's (7) correlation ratio, $\eta_{X \cdot Y} = \frac{\sigma_{\left(\bar{Y}_X\right)}}{\sigma_X}$ (the ratio of the standard deviation of the mean values of X for each value of Y to the total standard deviation of X), the standard deviation being the square root of the mean square deviation.

Use of the coefficient of correlation (r) assumes that there is a linear relation between the two variables—that is, that a given change in one variable always involves a certain constant change in the corresponding average value of the other. The value of the coefficient can never exceed

[1] Reference is made by number (italic) to "Literature cited," p. 585.

Journal of Agricultural Research,
Washington, D. C.
wh
17777°—21——4

Vol. XX, No. 7
Jan. 3, 1921
Key No. A-55

Sewall Wright (1889–1988)

- American geneticist.
- One of the founders of the field of population genetics.
- "Darwin of the 20th century".
- Invented path analysis method 1918–1921.
- First ever use of graphical models.

Yes, this was really written 100 years ago!

The ideal method of science is the study of the direct influence of one condition on another in experiments in which all other possible causes of variation are eliminated. Unfortunately, causes of variation often seem to be beyond control. In the biological sciences, especially, one often has to deal with a group of characteristics or conditions which are correlated because of a complex of interacting, uncontrollable, and often obscure causes. The degree of correlation between two variables can be calculated with well-known methods, but when it is found it gives merely the resultant of all connecting paths of influence.

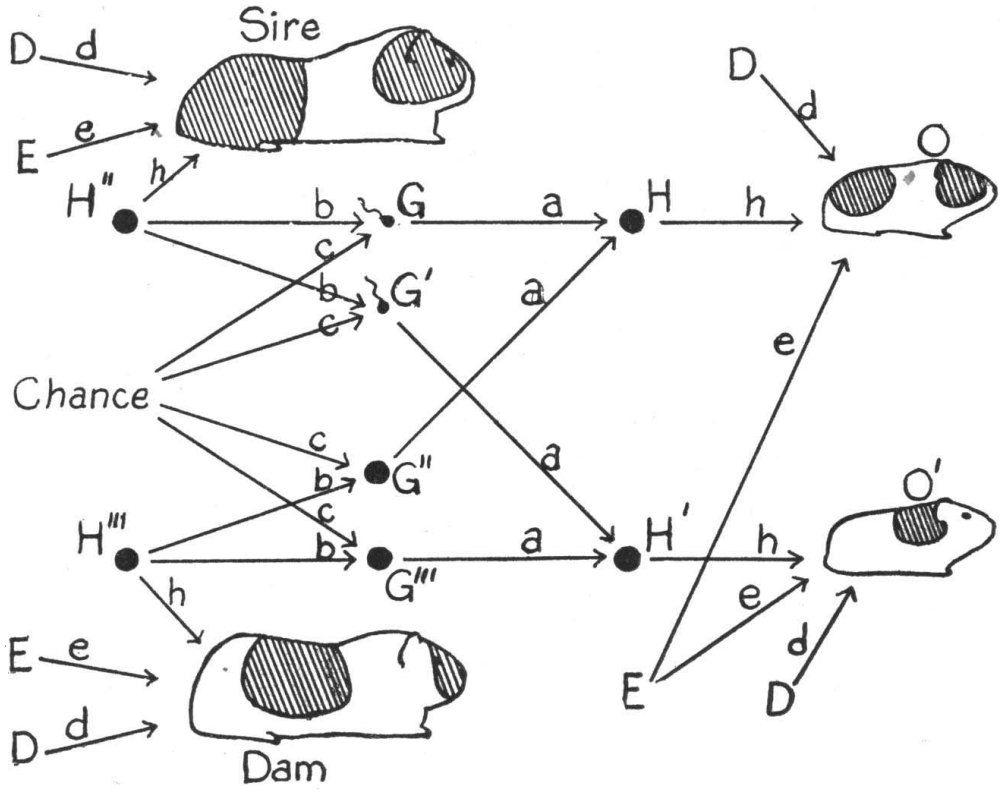
In other words. . .

All the impressive achievements of deep learning amount to just curve fitting. To build truly intelligent machines, teach them cause and effect.

Judea Pearl (2018)

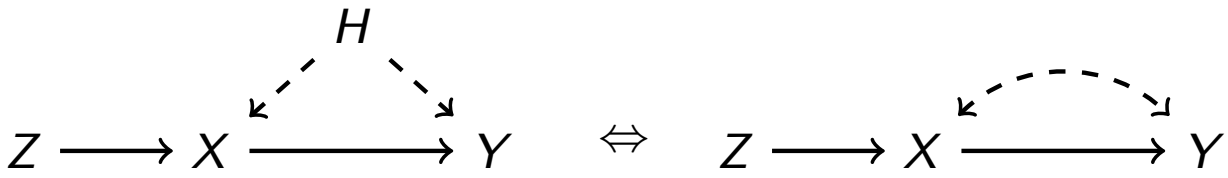
The method of path coefficients

The method depends on the combination of knowledge of the degrees of correlation among the variables in a system with such knowledge as may be possessed of the causal relations.



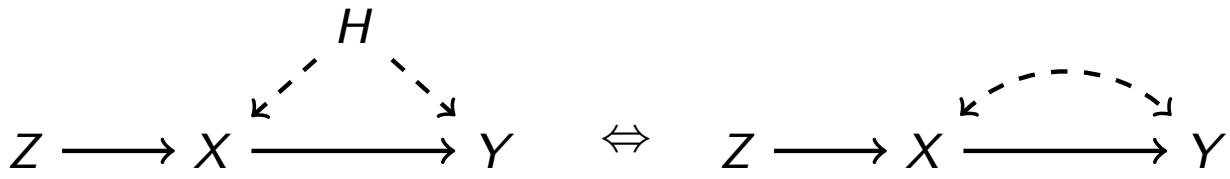
The method of path coefficients

- A **directed acyclic graph (DAG)** summarizes *prior* expert knowledge of the process that generates the data:



The method of path coefficients

- A **directed acyclic graph (DAG)** summarizes *prior* expert knowledge of the process that generates the data:



- **Directed edges** represent (possible) *causal* effects and **bidirected edges** represent the (possible) influence of unknown/unmeasured factors in a **structural equation model (SEM)**:

$$
\begin{aligned}
Y &:= bX + U_Y \\
X &:= aZ + U_X \\
Z &:= U_Z
\end{aligned}
$$

U_Y, U_X, and U_Z are random error terms with

$$
\begin{aligned}
\operatorname{cov}(U_Z, U_X) = \operatorname{cov}(U_Z, U_Y) &= 0 \\
\operatorname{cov}(U_X, U_Y) &= \beta^2
\end{aligned}
$$

The method of path coefficients

- The causal process/SEM determines the covariances/correlations in observational data:

$$\begin{aligned}
\mathrm{cov}(Z, X) &= a + \mathrm{cov}(U_Z, U_X) = a \\
\mathrm{cov}(Z, Y) &= b\,\mathrm{cov}(Z, X) + \mathrm{cov}(U_Z, U_Y) = ab \\
\mathrm{cov}(X, Y) &= b + \mathrm{cov}(X, U_Y) = b + \mathrm{cov}(U_X, U_Y) = b + \beta^2
\end{aligned}$$

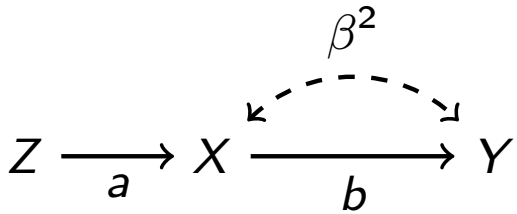
The method of path coefficients

- The causal process/SEM determines the covariances/correlations in observational data:

$$\begin{aligned}\operatorname{cov}(Z, X) &= a + \operatorname{cov}(U_Z, U_X) = a \\ \operatorname{cov}(Z, Y) &= b \operatorname{cov}(Z, X) + \operatorname{cov}(U_Z, U_Y) = ab \\ \operatorname{cov}(X, Y) &= b + \operatorname{cov}(X, U_Y) = b + \operatorname{cov}(U_X, U_Y) = b + \beta^2\end{aligned}$$

- If we can solve this system for the causal parameters, then the causal parameters are **identified**:

$$b = \frac{\operatorname{cov}(Z, Y)}{\operatorname{cov}(Z, X)}$$



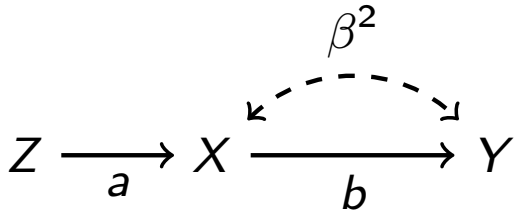
The method of path coefficients

- The causal process/SEM determines the covariances/correlations in observational data:

$$\begin{aligned}\operatorname{cov}(Z, X) &= a + \operatorname{cov}(U_Z, U_X) = a \\ \operatorname{cov}(Z, Y) &= b \operatorname{cov}(Z, X) + \operatorname{cov}(U_Z, U_Y) = ab \\ \operatorname{cov}(X, Y) &= b + \operatorname{cov}(X, U_Y) = b + \operatorname{cov}(U_X, U_Y) = b + \beta^2\end{aligned}$$

- If we can solve this system for the causal parameters, then the causal parameters are **identified**:

$$b = \frac{\operatorname{cov}(Z, Y)}{\operatorname{cov}(Z, X)}$$



- Some correlations do imply causation!

The method of path coefficients

- Wright's profound insight was that the covariance between *any* pair of variables can be determined by inspection of the DAG/path diagram.

The method of path coefficients

- Wright's profound insight was that the covariance between *any* pair of variables can be determined by inspection of the DAG/path diagram.
- The result states that the covariance between X and Y equals the sum of products of path coefficients and error covariances among all paths between X and Y that do not traverse any collider (a pair of head-to-head arrows, as in $X \rightarrow Z \leftarrow Y$).

The method of path coefficients

- Wright's profound insight was that the covariance between *any* pair of variables can be determined by inspection of the DAG/path diagram.
- The result states that the covariance between X and Y equals the sum of products of path coefficients and error covariances among all paths between X and Y that do not traverse any collider (a pair of head-to-head arrows, as in $X \rightarrow Z \leftarrow Y$).
- The result applied to any linear SEM (not only normally distributed variables!).

Let's talk about impact

Let's talk about impact

Statisticians attacked Wright vigorously for daring to say that not all answers are in the data and that prediction (correlation) alone is insufficient.

Let's talk about impact

Statisticians attacked Wright vigorously for daring to say that not all answers are in the data and that prediction (correlation) alone is insufficient.

Economists/sociologists adopted the method, but gradually forgot that *structural* equations are not *algebraic* equations, and have been debating the meaning of SEM parameters ever since.

Let's talk about impact

Statisticians attacked Wright vigorously for daring to say that not all answers are in the data and that prediction (correlation) alone is insufficient.

Economists/sociologists adopted the method, but gradually forgot that *structural* equations are not *algebraic* equations, and have been debating the meaning of SEM parameters ever since.

Biologists completely ignored the method for many decades.

Let's talk about impact

Statisticians attacked Wright vigorously for daring to say that not all answers are in the data and that prediction (correlation) alone is insufficient.

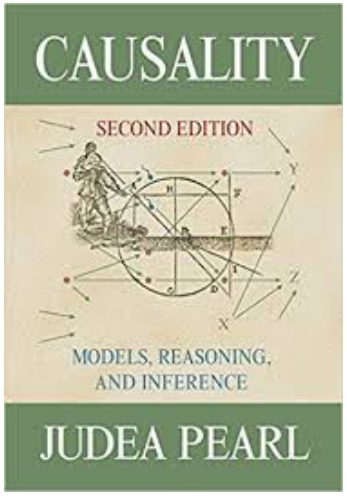
Economists/sociologists adopted the method, but gradually forgot that *structural* equations are not *algebraic* equations, and have been debating the meaning of SEM parameters ever since.

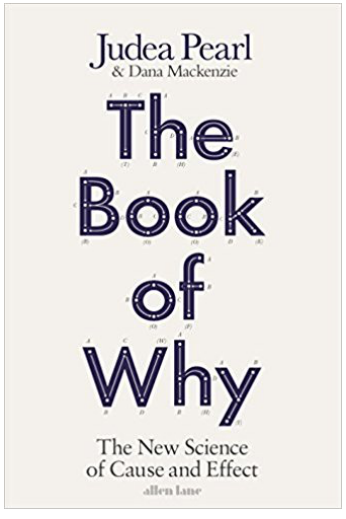
Biologists completely ignored the method for many decades.

Wright considered path analysis one of his most important scientific contributions, and continued to publish on it all his life.

Tellingly, one of his last papers, at the age of 94(!), was to respond to a misrepresentation of the method in a publication in the American Journal of Human Genetics.

Computer science comes to the rescue!

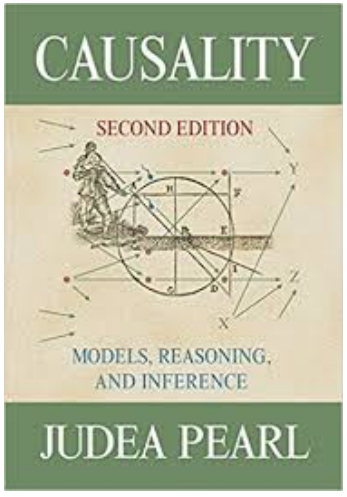


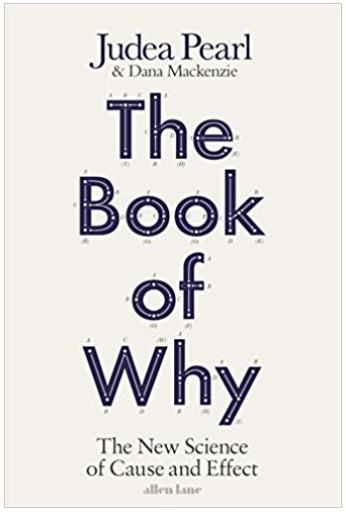


Judea Pearl (1936-)

- American-Israeli computer scientist
- Inventor of Bayesian networks (1985)
- Bayesian networks + prior knowledge graph = causal Bayesian network
- Invented new calculus for computing the effect of interventions from observational data (**do-calculus**)
- Developed general theory for identifying causal effects in non-linear models from the structure of the prior graph.

Computer science comes to the rescue!





Judea Pearl (1936-)

- American-Israeli computer scientist
- Inventor of Bayesian networks (1985)
- Bayesian networks + prior knowledge graph = causal Bayesian network
- Invented new calculus for computing the effect of interventions from observational data (**do-calculus**)
- Developed general theory for identifying causal effects in non-linear models from the structure of the prior graph.

Pearl's work has led to a renewed interest in rigorous analysis of causality in AI, statistics, economics, sociology, genetics, epidemiology, biology, albeit with some pockets of resistance remaining.

Lessons learned so far

Lessons learned so far

- Data = Correlation = Prediction

Lessons learned so far

- Data = Correlation = Prediction
- Data + Prior knowledge = Causation = Understanding

Lessons learned so far

- Data = Correlation = Prediction
- Data + Prior knowledge = Causation = Understanding
- Data Science was not invented yesterday.

Lessons learned so far

- Data = Correlation = Prediction
- Data + Prior knowledge = Causation = Understanding
- Data Science was not invented yesterday.
- Computational Biology is not only the application of algorithms to biological data, some areas of computer science have their origin in biology itself.

Lessons learned so far

- Data = Correlation = Prediction
- Data + Prior knowledge = Causation = Understanding
- Data Science was not invented yesterday.
- Computational Biology is not only the application of algorithms to biological data, some areas of computer science have their origin in biology itself.
- Big question: How do we apply the ideas and methods of causal inference in a big data context?

References

- Pearl J. *Causality.* 2nd Edition (2009).
- Pearl J. *Linear Models: A Useful "Microscope" for Causal Analysis.* Journal of Causal Inference 1:155-170 (2013).
- Pearl J, *The Book of Why* (2018).
- Shipley B. *Cause and Correlation in Biology*. 2nd Edition (2016).
- Wikipedia
- Wright S. *The relative importance of heredity and environment in determining the piebald pattern of guinea pigs.* Proc. Nat. Acad. Sci. 6: 320-332 (1918).
- Wright S. *Correlation and Causation.* Jour. Ag. Res. 20:557-585 (1921).
- Wright S. *The Method of Path Coefficients*. The Annals of Mathematical Statistics 5:161-215 (1934).

Part II

Causal inference in drug discovery and development[1]

[1] T Michoel and JD Zhang, *Drug Discovery Today*, 28:103737 (2023)

Introduction

- Causal inference is the process of identifying causal effects based on prior knowledge, hypotheses, and correlations observed in data.
- We introduce the statistical causal inference approach and define causality as a probabilistic relationship that satisfies:
  1. Regular probabilistic update: taking the drug modifies the probability of dying from the disease within a defined time window, irrespective of where or when the trial happens.
  2. Manipulation: drug treatment shows additional benefit even when considering all other factors affecting patients' survival.
  3. Counterfactual condition: the death of a patient would not have been postponed had the drug not been taken.
  4. Mechanism of action: we understand why the drug prolongs patients's survival.
- These conditions ensure both statistical correlation and mechanistic understanding.[2]

[2] Williamson J, *Establishing causal claims in medicine.* (2019)

Distinguishing causation from correlation

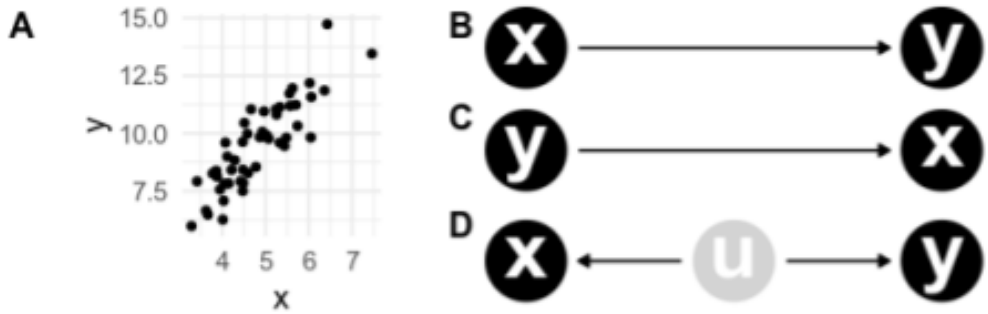


Figure 1: x and y depict correlated expression levels of proteins X and Y in a population of cells. Four scenarios are possible:

1. Causation: expression of X causes expression of Y, or vice versa.
2. Confounding: a third, potentially unobserved, protein U causes expression of both X and Y.
3. Coincidence: the correlation is solely by chance.
4. Conspiracy: the correlation is due to deliberate manipulation of the data or sampling process, for instance removing data from cells where the proteins are not correlated.

Distinguishing causation from correlation

- Correlation is sufficient for predicting one variable from another, but does not inform on causation.

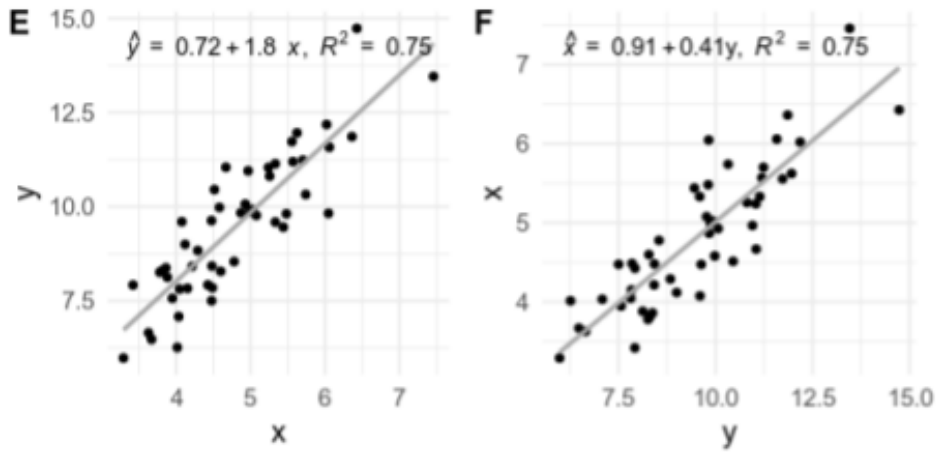


Figure 2: Regression of y on x and x on y give the same correlation coefficients

Distinguishing causation from correlation

- Predicting the outcome of an intervention requires a causal model.

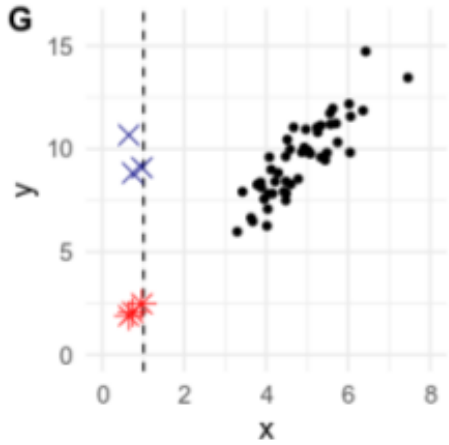


Figure 3: Reducing the expression of X artificially to 1.0 results in different distributions for y depending on whether the causal model is $X \rightarrow Y$ (red stars) or $Y \rightarrow X$ (blue crosses).

Causal modelling with DAGs

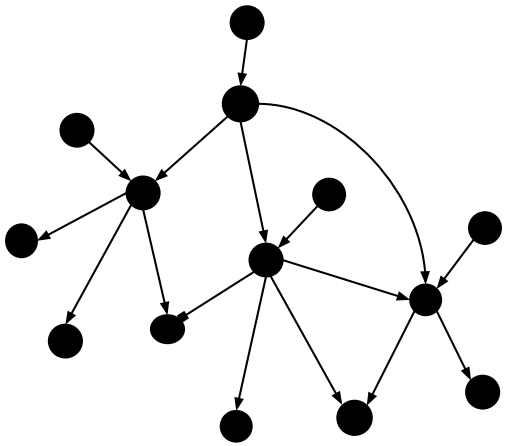


Figure 4: A DAG

- Directed acyclic graphs (DAGs) / Bayesian networks (BNs) are computational models of causal inference.
- DAGs consist of nodes, which represent variables, and edges, which represent knowledge or hypotheses about causal relationships.
- A lack of edge between a pair of nodes means that we exclude the possibility of a direct causal relationship between them.

- The absence of directed cycles (feedback loops) means the joint distribution of the variables (the "correlations") can be expressed as a product of forward conditional probabilities $P(X \mid \mathrm{Pa}_X)$.
- BNs can be viewed as an inference instrument for deducing new independence relationships from those used in constructing the network, and for learning the strength of direct and indirect causal relations (causal identification).

Causal modelling with DAGs

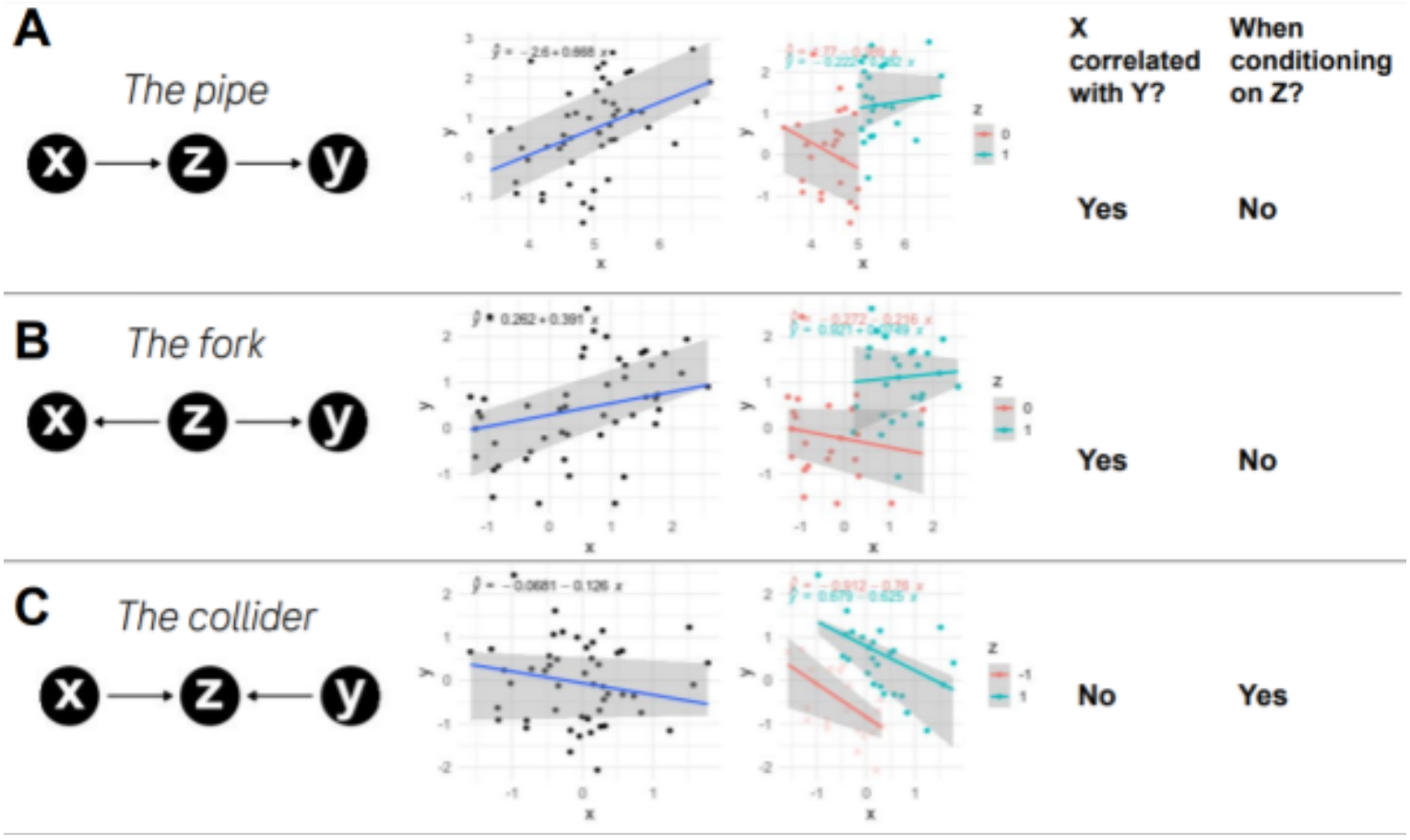


Figure 5: Prevalent 3-node structures allow to interpret more complex causal models.

Causal inference for controlled experimental studies

Controlled experiment:

- Test objects (e.g. animals in preclinical studies) assigned to treated and untreated groups.
- If assignment is randomized w.r.t. any relevant attributes of test objects: randomized controlled trial (RCT) – gold standard for establising causality.

Causal inference may be required to identify treatment effect if randomization is broken:

- Non-compliance: some patients do not take treatment as prescribed.
- Missing data: patients drop out of trial.
- Intercurrent events: events occuring after randomization, e.g., development of anti-drug antibodies.

Causal inference for observational studies

Observational studies measure or survey members of a sample without trying to affect them:

- Epidemiological studies
- Electronic health records
- Insurance data
- Omics and behavioural data of healthy individuals and patients

Use of causal models is imperative:

- Integrate knowledge and hypotheses about biases in the data generating process.
- Investigate causal effect of variables of interest, while considering covariates that also affect the outcome.
- Potentially resolve the effect of confounding variables that affect both the independent variable and the outcome.

Six steps of causal inference

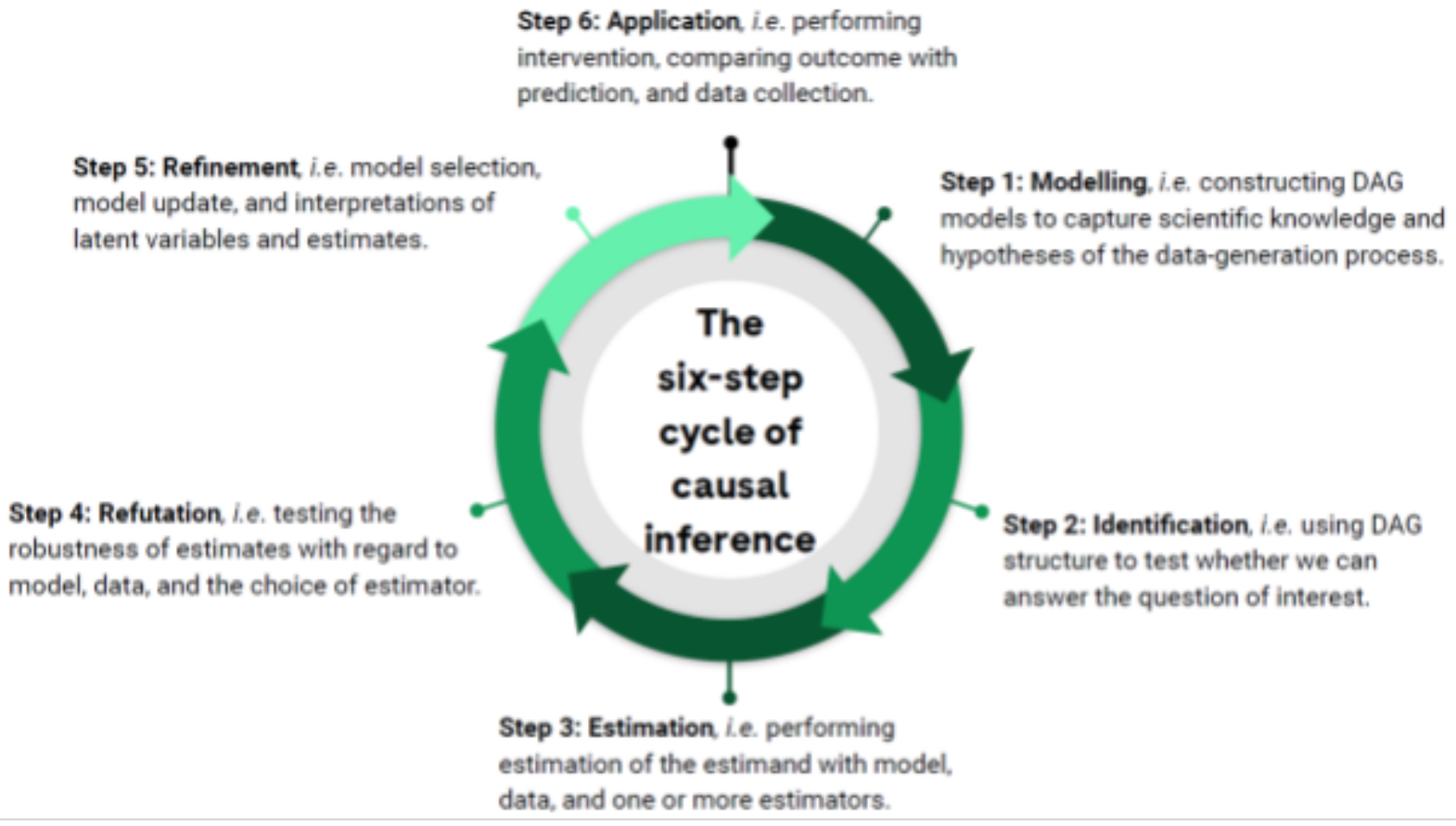
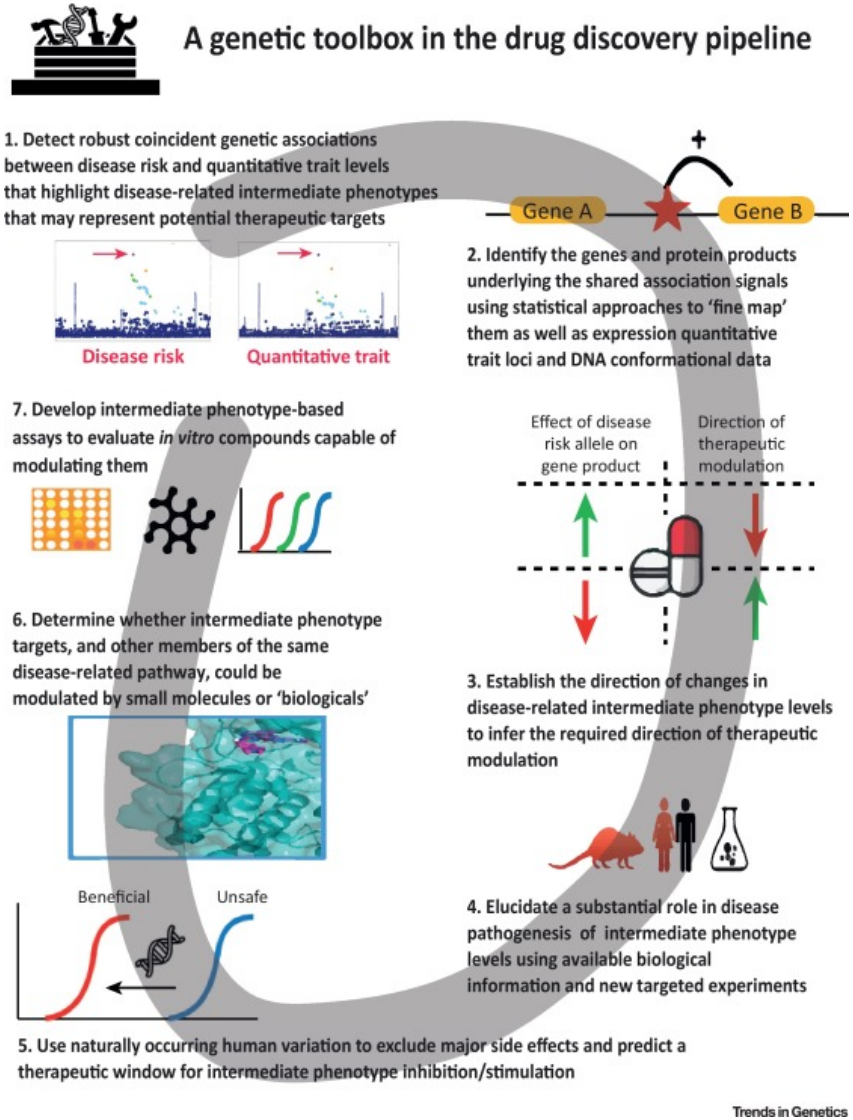


Figure 6: A six-step model of causal inference

Literature review of causal inference in drug discovery and development

Figure 7: Review of >800 publications on causal inference in drug discovery and development

Learning causal associations from natural experiments (genetics) and observational studies



[Floris et al. Trends Genet (2018)]

Selecting genetically supported targets could double the success rate in clinical development.

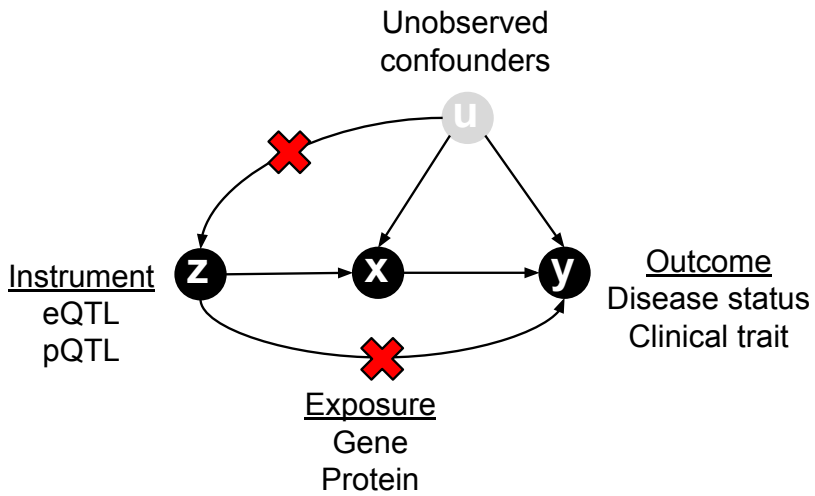
Nelson *et al.* Nat Genet (2015)

When causal genes are clear [. . .] the use of human genetic evidence increases approval by greater than two-fold.

King *et al.* PLOS Genet (2019)

Mendelian randomization identifies causal associations between heritable traits

- GWAS identify genetic loci associated with disease risk, drug response, susceptibility to adverse drug reactions, etc.
- When GWAS loci overlap with loci with an effect on transcriptome/proteome (eQTL/pQTL), MR can estimate causal effects and suggest drug repurposing opportunities or new candidate drug targets.



- **But:** MR tests effects of genes/proteins on disease one-by-one, whereas molecules operate through complex pathways and networks.

Causal model selection

Statistical model selection is used to orient the direction of causality among *all* pairs of correlated genes or proteins, using *cis*-acting eQTLs or pQTLs as instrumental variables.

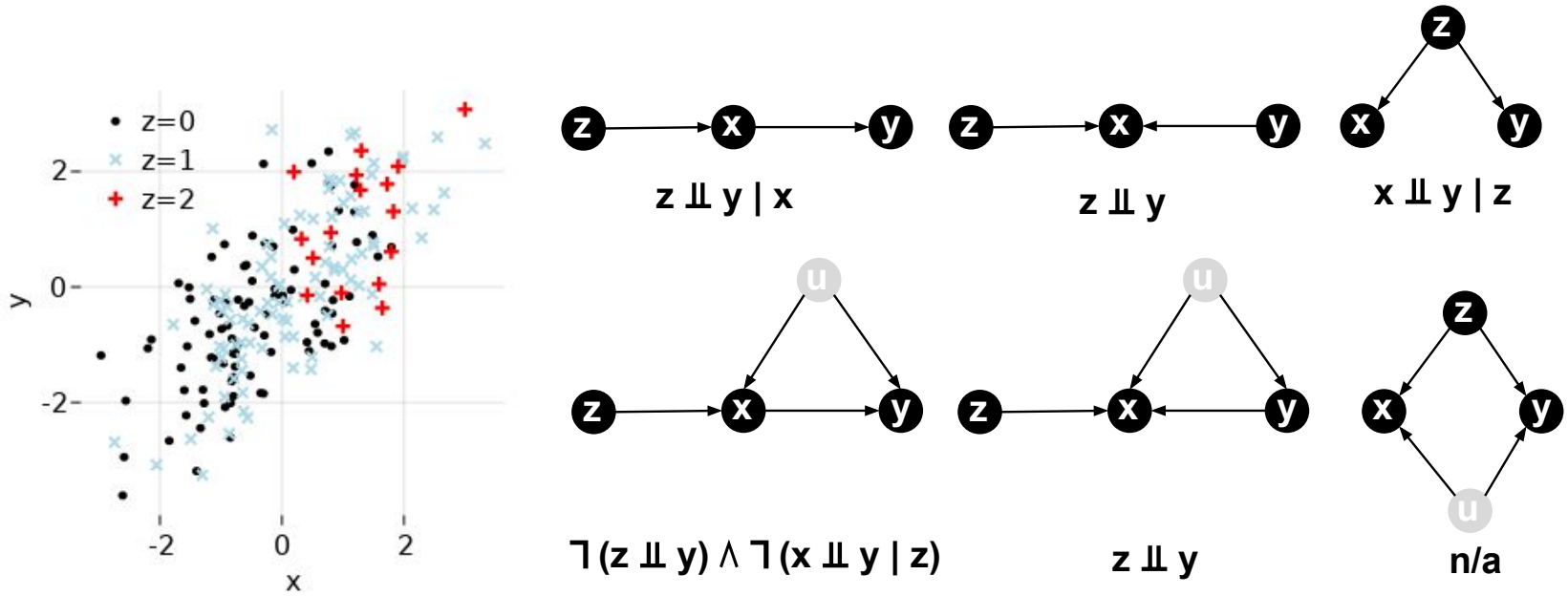


Figure 8: (Simulated) scatter plot of coexpressed genes X and Y with samples colored by genotype of a genetic marker Z for X. Model selection tests conditional independencies implied by each possible DAG.

Causal model selection

Statistical model selection is used to orient the direction of causality among *all* pairs of correlated genes or proteins, using *cis*-acting eQTLs or pQTLs as instrumental variables.

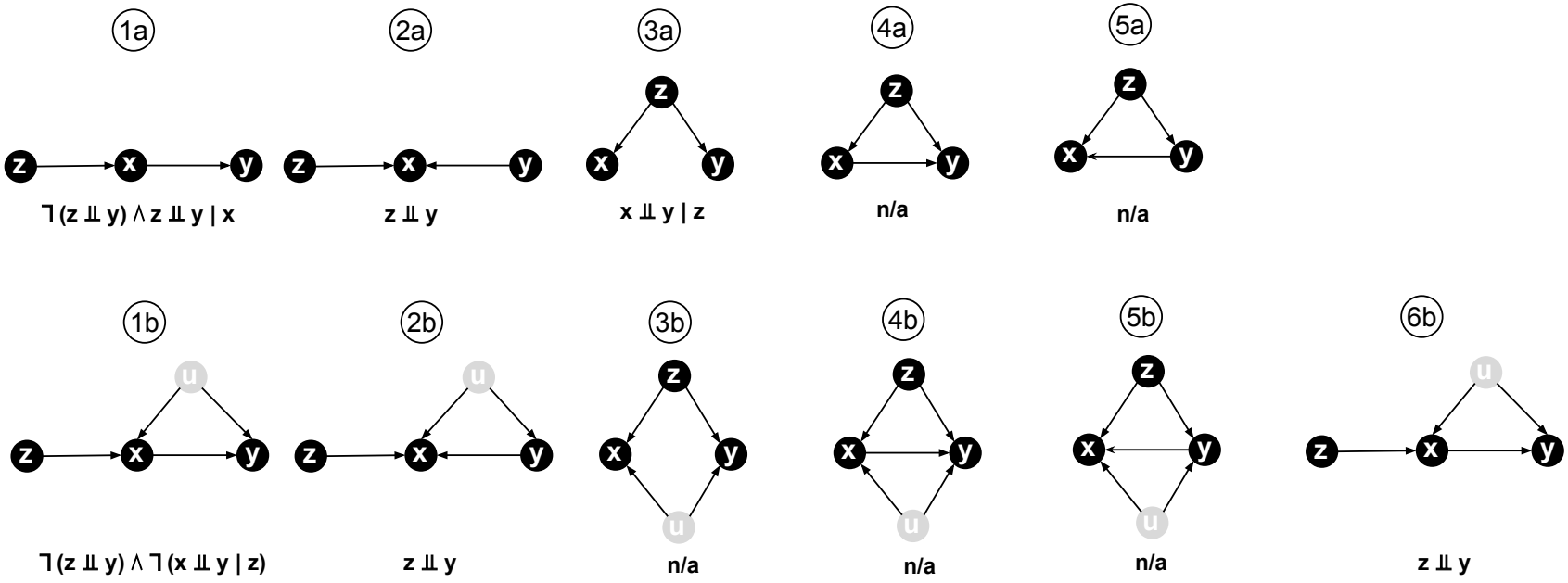


Figure 9: All models that satisfy (1) Z $\rightarrow$ X, (2) X and Y correlated, (3) Z independent of U, (4) Z has no incoming arrows.

Causal model selection tests

A sufficient condition for $X \to Y$ (mediation)

- **If** Y is not independent of Z, and Y is independent of Z given X, **then** X $\to$ Y.
- High specificity, low sensitivity.

A necessary condition for $X \to Y$

- **If** X $\to$ Y, **then** Y is not independent of Z, and Y is not independent of X given Z.
- High sensitivity, reduced specificity.

There is no condition that is both necessary and sufficient for X $\to$ Y

Reference

T Michoel and JD Zhang, *Causal inference in drug discovery and development*, Drug Discovery Today, 28:103737 (2023)

https://doi.org/10.1016/j.drudis.2023.103737