Model Selection with Causal Models for Regression Modeling Strategies

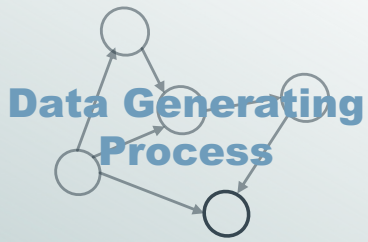
Drew Griffin Levy
Regression Modeling Strategies Short Course
May 2025

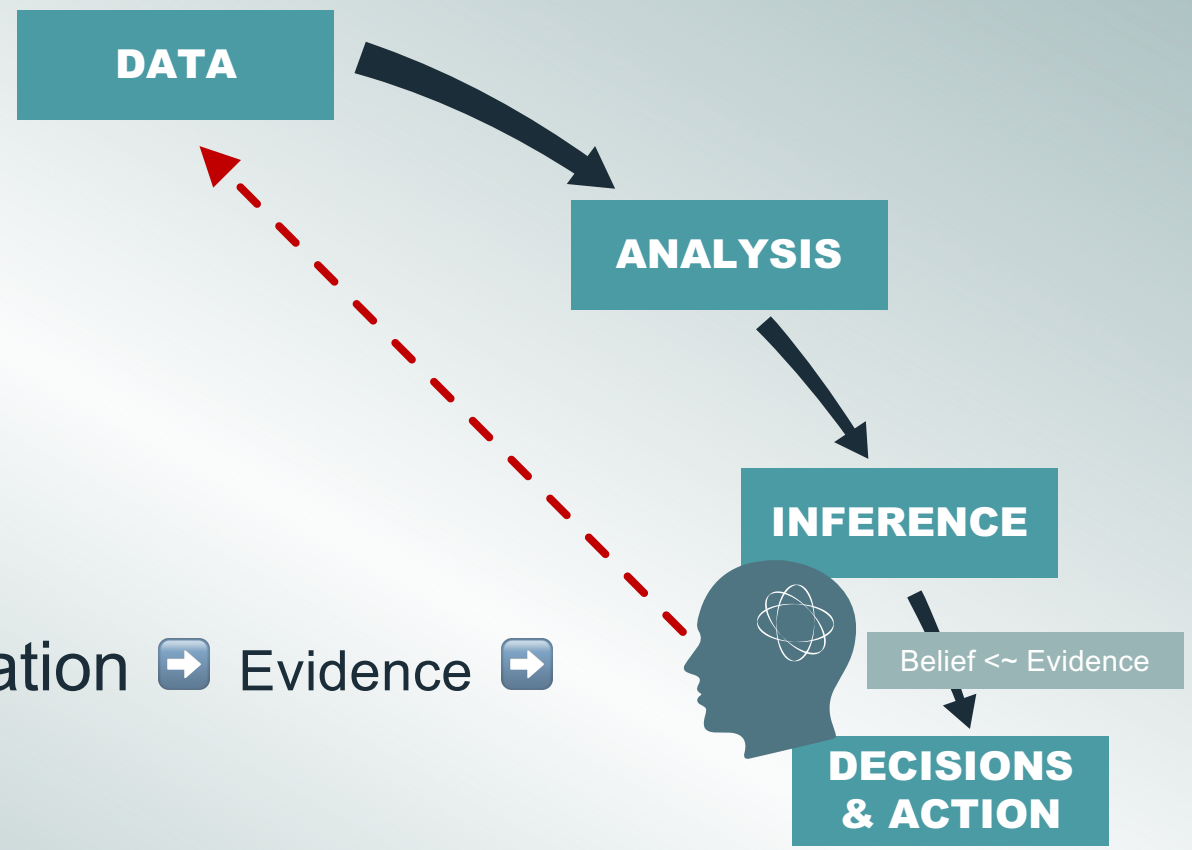
“It’s much easier to get a result than it is to get an answer.”

—Christie Aschwanden, FiveThirtyEight

The 'Epistemologic Arc'

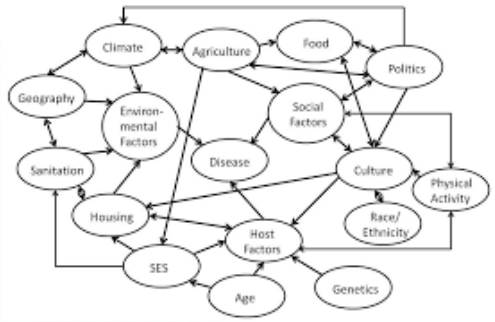
The perceived horizon





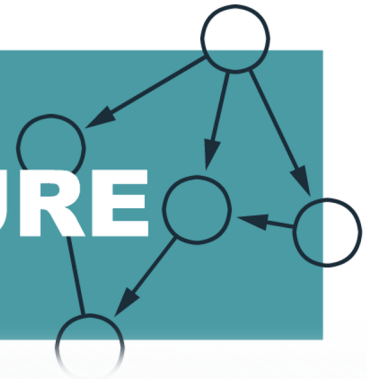
Data ➡️ Information ➡️ Evidence ➡️

POPULATION



Nature is the nexus of causes that produce all phenomena actually or potentially available for empirical study.

NATURE



The ‘Epistemologic Arc’

Nature: The complex and nexus of causes that produce the phenomena of our world that are available for empirical study. The underlying causal structure of nature is often abstruse or inscrutable.

Population: All of the objects (existing, extant and/or possible) in the category of interest for study. The population is the realization of causal process in nature. The Population is the expression of the ‘long run’ probabilistic tendencies in nature’s causes. The population is also the primary object of study and inference.

Sample: The subset of the population available for study and observed.

Data: The actual observations made and recorded on the sample. Not all observations/variables of all possible variables from the sample are collected. Measurements are made imperfectly and recorded with errors. The particular instance of the data (out of many possible instances) are the source of the statistical likelihood on which the analysis is predicated.

Analysis: The mathematical procedures that account for both the structure and randomness of the data. Typically a model is used or is at least implicit. All analyses require assumptions (both strong and weak).

Inference and Belief: The conclusions drawn from the analysis of the data (and in combination with any external information), including whether any associations observed are causal in nature and likely reproducible effects in independent data. Belief depends on the strength of the findings and the research process, coherence with existing knowledge, and numerous cognitive and psychological factors including biases, intentions and motivations.

Decisions and Actions: The consequences, if any, of the research activities. The impact of the research will depend in part on the strength of the belief resulting from the inference, and the relevance for problems faced by others. Consequences include clinical behavior and medical decision making; and scientific behavior including confirmatory reproduction of research, and motivation of additional research.

The process of evidence generation

- Omitted variables
- Missing data
- Measurement issues
- Information bias

DATA

Likelihood: P(data | θ)

SAMPLE

Conventional statistical methods

- Importance of study / experimental design
- Risk of selection bias; confounding by indication

ANALYSIS

Analytic bias
- Model selection
  - $E(\hat{\beta} \mid \hat{\beta}_{"significant"}) \neq \beta_{true}$
- Model misspecification
- Over-fitting
- Residual confounding
- Arbitrary categorization
- Collider bias

Uncertainties
- Model specification
- Model selection
- Assumptions re. distributions

POPULATION

INFERENCE ← **P(θ | data)**

Association vs. Causation

- Cognition/ psychology
- Intentions
- Motivations

Belief ~ Evidence

DECISIONS & ACTION

NATURE

"What we observe is not nature itself, but nature exposed to our method of questioning."

-Werner Heisenberg

The process of evidence generation

- Omitted variables
- Missing data
- Measurement issues
- Information bias

DATA

Likelihood: P(data | θ)

SAMPLE

Conventional statistical methods

- Risk of selection bias; confounding by indication
- Importance of study / experimental design

POPULATION

NATURE

ANALYSIS

Analytic bias

- Model selection
  - $E(\hat{\beta} \mid \hat{\beta}_{"significant"}) \neq \beta_{true}$
- Model misspecification
- Over-fitting
- Residual confounding
- Arbitrary categorization
- Collider bias

Uncertainties

- Model specification
- Model selection
- Assumptions re. distributions

Association vs. Causation

- Cognition/ psychology
- Intentions
- Motivations

INFERENCE ← **P(θ | data)**

Belief ~ Evidence

DECISIONS & ACTION

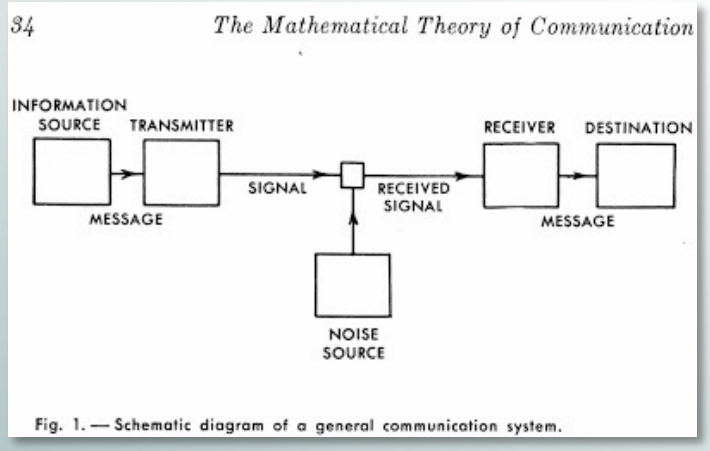
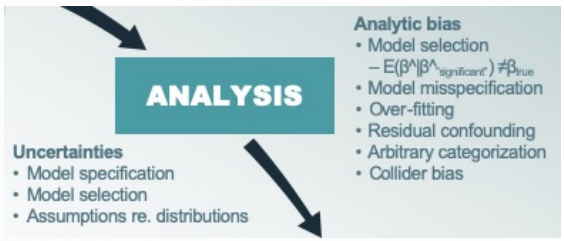


Fig. 1. — Schematic diagram of a general communication system.

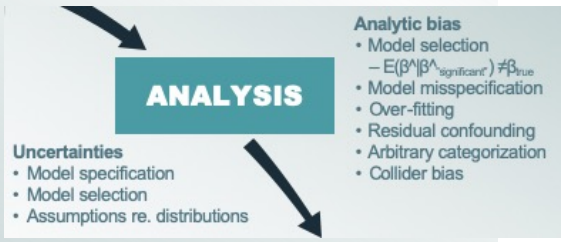
Model selection



4.4 Sample Size, Overfitting, and Limits on Number of Predictors

When a model is fitted that is too complex, that it, has too many free parameters to estimate for the amount of information in the data, the worth of the model (e.g., R^2) will be exaggerated and future observed values will not agree with predicted values. In this situation, *overfitting* is said to be present, and some of the findings of the analysis come from fitting noise and not just signal, or finding spurious associations between X and Y. In this section general guidelines for preventing overfitting are given. Here we concern ourselves with the *reliability* or *calibration* of a model, meaning the ability of the model to predict future observations as well as it appeared to predict the responses at hand. For now we avoid judging whether the model is adequate for the task, but restrict our attention to the likelihood that the model has significantly overfitted the data.

Model selection



4.3 Variable Selection

The material covered to this point dealt with a prespecified list of variables to be included in the regression model. For reasons of developing a concise model or because of a fear of collinearity or of a false belief that it is not legitimate to include "insignificant" regression coefficients when presenting results to the intended audience, stepwise variable selection is very commonly employed. Variable selection is used when the analyst is faced with a series of potential predictors but does not have (or use) the necessary subject matter knowledge to enable her to prespecify the "important" variables to include in the model. But using Y to compute P-values to decide which variables to include is similar to using Y to decide how to pool treatments in a five–treatment randomized trial, and then testing for global treatment differences using fewer than four degrees of freedom.

Stepwise variable selection has been a very popular technique for many years, but if this procedure had just been proposed as a statistical method, it would most likely be rejected because it violates every principle of statistical estimation and hypothesis testing. Here is a summary of the problems with this method.

Structural Causal Models (SCMs) and Causal-Directed Acyclic Graphs (cDAGs)

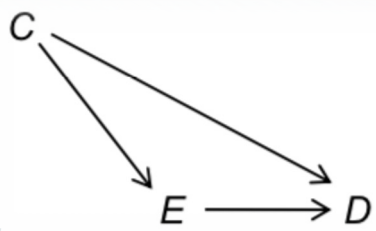
- Modeling decisions can be supported with SCMs and cDAGs (causal diagrams)

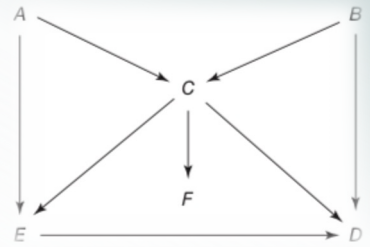
- SCMs can be used to
  - define bias
  - identify confounding
  - *Identify sets of adjustments necessary for unbiased statistical estimation (conditional on assumptions)*

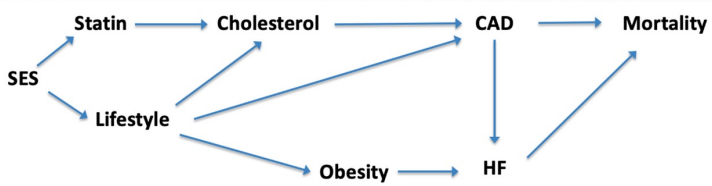
- ! Blind or arbitrary adjustment for confounding may *induce* bias

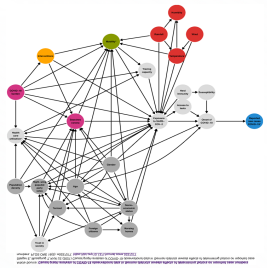
- <u>Minimal sets of required adjustments can help to use data (limited N) efficiently</u>

- Types of systematic bias:
  - <u>*Confounding*</u>
  - Selection bias
  - Measurement bias
  - others





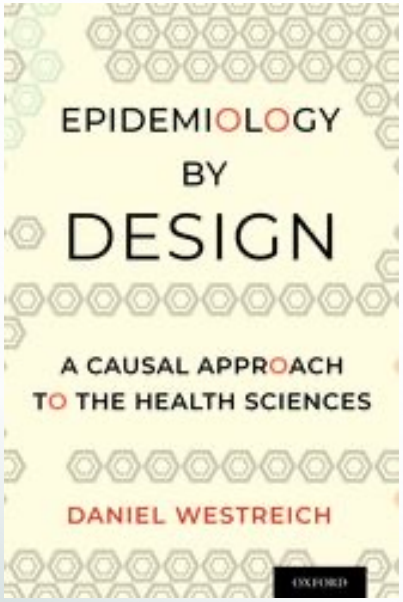


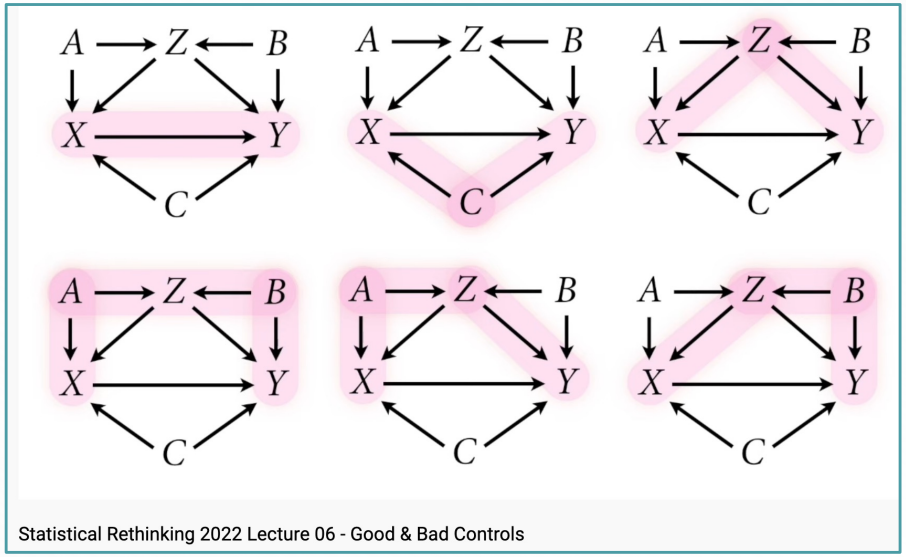


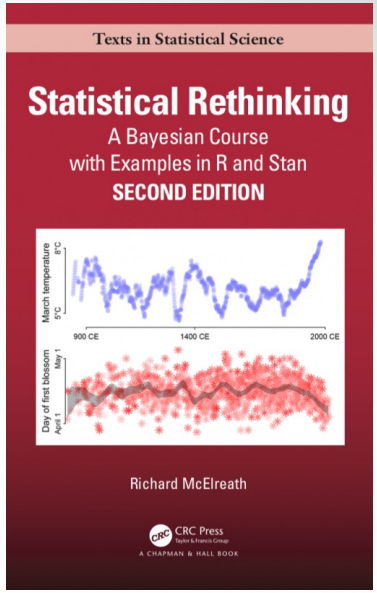
Resources

- Judea Pearl
  1. Causal Inference in Statistics: A Primer, 2016
  2. Causality: Models, Reasoning and Inference, 2009
  3. The Book of Why: The New Science of Cause and Effect, 2018.
- Miguel Hernan
  1. The Causal Inference Book
  2. edX MOOC: Causal Diagrams: Draw Your Assumptions Before Your Conclusions
- Modern Epidemiology, 3rd ed. Rothman, Greenland, Lash: Chapter 12–Causal Diagrams
- Causal Diagrams for Epidemiologic Research. S. Greenland, J. Pearl, J. Robins. Epidemiology 1999;10:37-48.
- Epidemiology by Design: A Causal Approach to the Health Sciences, D. Westreich, 2020
- Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide: Supplement 2, Use of Directed Acyclic Graphs
- DAGitty - drawing and analyzing causal diagrams (DAGs) (www.dagitty.net/)

Re- & Magnifi- cent resources







Epidemiology by Design, 2019

free YouTube lectures! :
Statistical Rethinking 2023

Statistical Rethinking, 2020

The 'Epistemologic Arc' and RMS

- Omitted variables
- Missing data
- Measurement issues
- Information bias

DATA

Likelihood: P(data | θ)

SAMPLE

Conventional statistical methods

- Risk of selection bias; confounding by indication
- Importance of study / experimental design

POPULATION

NATURE

"What we observe is not nature itself, but nature exposed to our method of questioning."
-Werner Heisenberg

ANALYSIS

Analytic bias

- Model selection
  - $E(\hat{\beta} \mid \hat{\beta}_{"significant"}) \neq \beta_{true}$
- Model misspecification
- Over-fitting
- Residual confounding
- Arbitrary categorization
- Collider bias

Uncertainties

- Model specification
- Model selection
- Assumptions re. distributions

INFERENCE

P(θ | data)

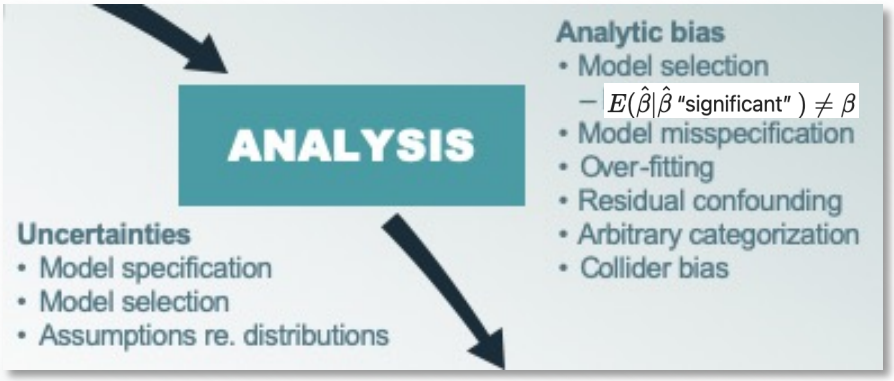
Association vs. Causation

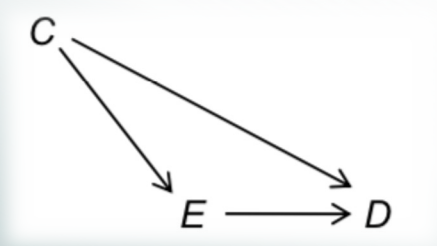
- Cognition/ psychology
- Intentions
- Motivations

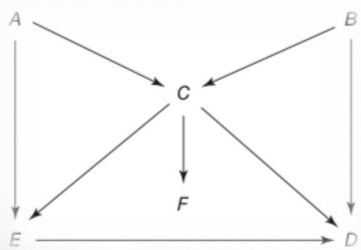
Belief ~ Evidence

DECISIONS & ACTION

<u>Takeaways</u>: Reasons to consider SCMs in model selection for observational studies







SCMs …

1. support identification of biases
2. *recommend a* [minimum] *set of adjustments necessary for unbiased effect estimation*
3. *may* rationalize model selection
4. can help you spend df's effectively
5. help in de-bugging our thinking
6. reduce ambiguity in communication
7. support achieving consensus
8. mitigate 'analysis multiplicity'

Complimentary PoV: Variable selection for model selection

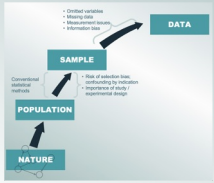
RMS

- Eschew automated variable selection
- Principled data reduction techniques
  - using data reduction methods (masked to Y) to reduce the dimensionality of the predictors and then fitting the number of parameters the data's information content can support
- Shrinkage to mitigate over-fitting
  - use shrinkage (penalized estimation) to fit a large model without worrying about the sample size.

SCMs & causal DAGs

- Subject-matter-knowledge-driven approaches
- Can aid in selecting covariates in regression models by identifying the set(s) of adjustments necessary for estimation of specific effects without bias
- Avoid adjustments that *induce* bias!

We can & *will* be fooled by data!



*"*Using the data to guide the analysis is *almost* as dangerous as not using it!*"*

---Frank Harrell, RMS

"The data are profoundly dumb!"

---Judea Pearl, Book of Why

- Data helps to describe reality—albeit *imperfectly*
- Nature is indifferent to furnishing noise vs. signal; *the computer cannot divine causes*
- It is a prevalent mistake to believe that "all the answers [information] are in the data"
- Relying on statistical approaches to identifying variables for adjustment and control of confounding can be problematic

Confounding is a *causal* phenomenon

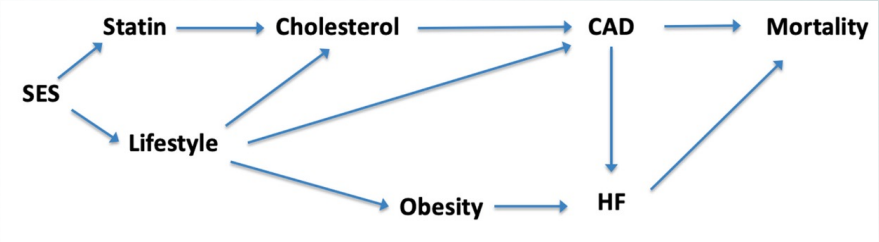
"Data do not understand causes and effects; humans do."
— **Judea Pearl, The Book of Why: The New Science of Cause and Effect**

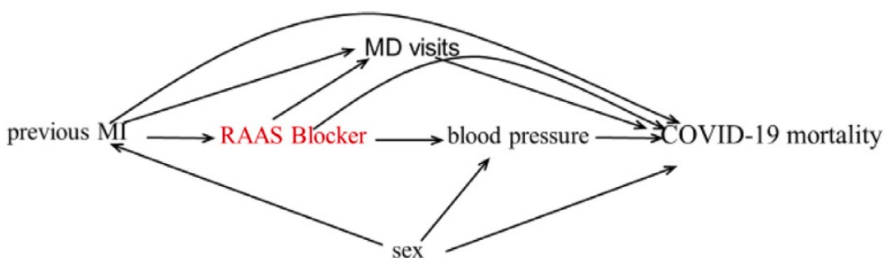
- Statistical data, however large, is insufficient for determining what is "causal," and must be supplemented with extra-statistical knowledge to make sense
- Subject-matter knowledge *must be employed* to effectively prevent bias
- *SCMs/DAGs are concise and explicit expressions of subject-matter knowledge*

“Draw your assumptions before your conclusions.”

—M. Hernan

- Causal diagrams describe the data generating process (DGP)
- Causal diagrams help us summarize what we know about a problem and communicate our assumptions about its causal structure.
- Causal diagrams help us diagnose biases in causal inference
- Causal diagrams help you organize your expert knowledge visually; and therefore, they help make our assumptions assumptions more explicit.





Causal directed acyclic graph of the case scenario depicting the effect of RAAS blockers on the risk of COVID-19 mortality.

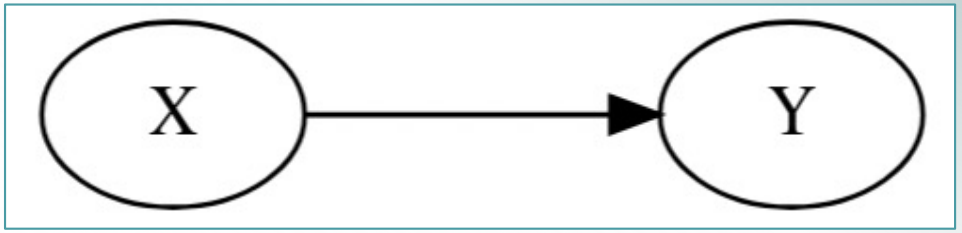
Causal Diagrams: Draw Your Assumptions Before Your Conclusions and The Causal Inference Book

How does a DAG work?

Basic notions in causal models

1. Causal relationship vs. independence
2. Causal paths
3. Biasing structures
   i. Confounder (the “Fork”)
   ii. Mediator (the “Pipe” or “Chain”)
   iii. Collider (the ”Collider”)
4. *Backdoor paths, ‘d-separation’, the ‘do-calculus’*

Cause - effect



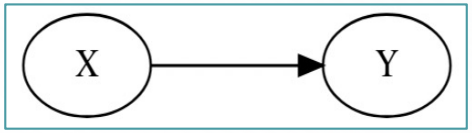
Absence of causal effects imply independencies: e.g., $P(Y|X) = P(Y)$

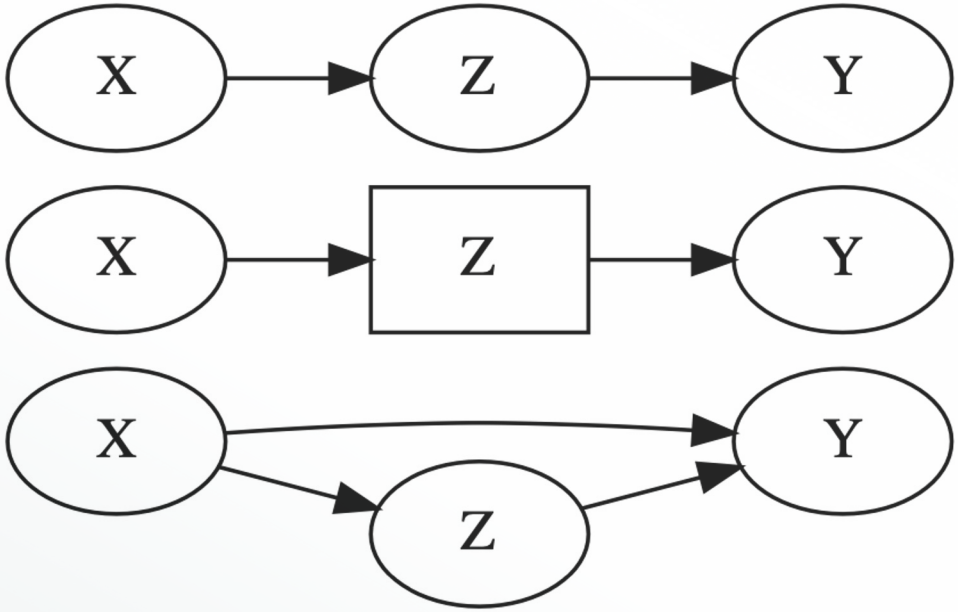
Causal effects imply associations $P(Y=y \mid X=x) \neq P(Y=y)$

- The presence or absence of arrows in DAGs correspond to the presence or absence of individual causal effect *in the population*
- DAGs are both causal models *and* statistical models (i.e., models that represent associations and independencies)

*See Chapter 1, Pearl, Glymour & Jewell, 2016; and M. Hernan's Causal Diagrams: Draw Your Assumptions Before Your Conclusions

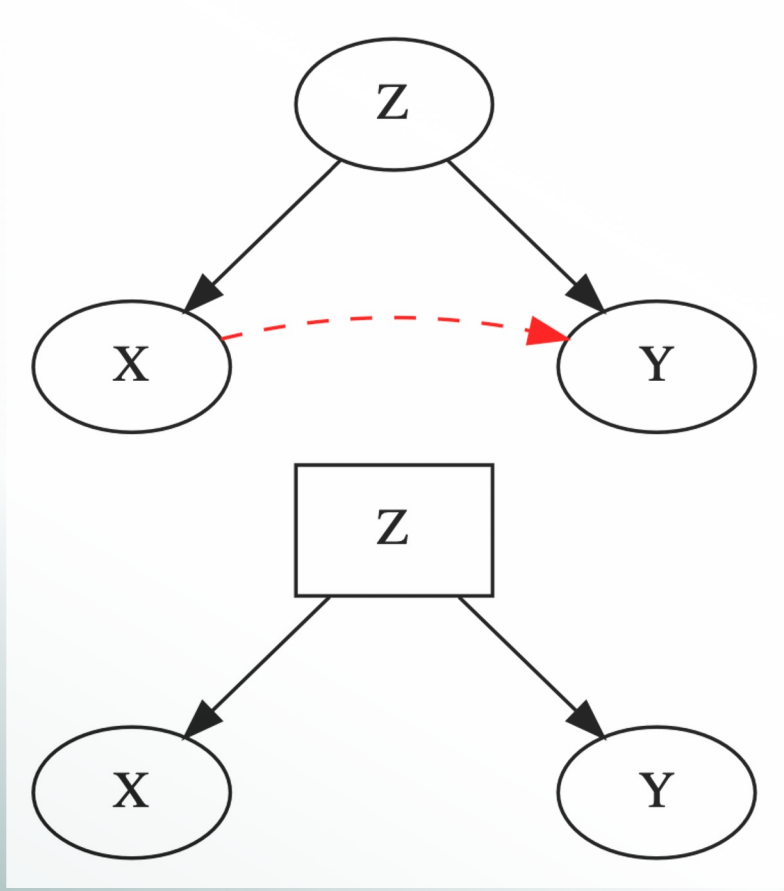
Causal Paths





- Mediation
- Conditional independence, given Z
- Direct vs. indirect effects
- Total effect

Confounder structures



- Causal structure with *common causes*
- Bias: spurious association; X and Y are not expected to be independent
- Conditioning on Z blocks the biasing path

Confounders vs. Mediators (Intermediate variables)

Hospitalization and death among adults with COVID-19

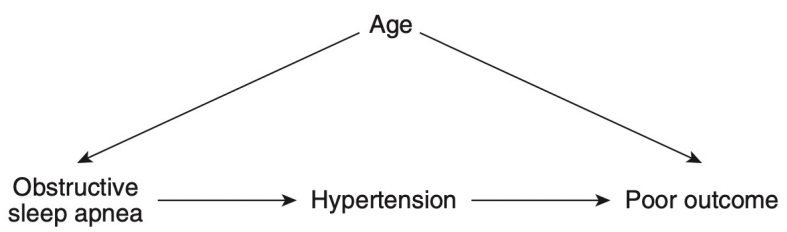
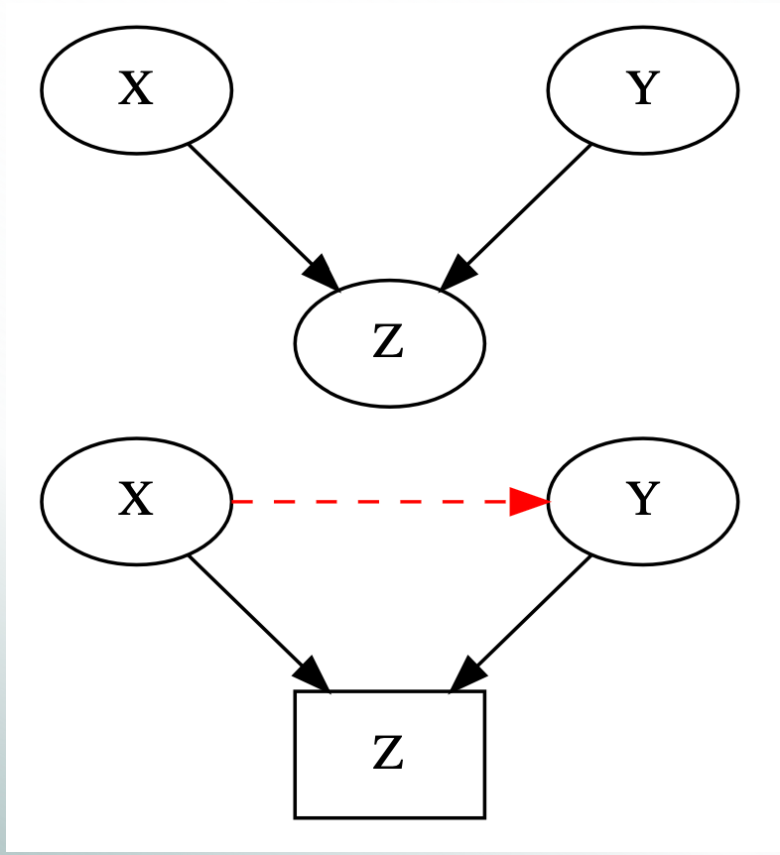


Figure 1. Directed acyclic graph for the effect of obstructive sleep apnea on poor outcome among patients with coronavirus disease (COVID-19). Age is a confounder of the association, whereas hypertension is a causal intermediate.

1. Cade BE, Dashti HS, Hassan SM, Redline S, Karlson EW. Sleep Apnea and COVID-19 Mortality and Hospitalization. Am J Respir Crit Care Med. 2020 Nov 15;202(10):1462-1464. doi: 10.1164/rccm.202006-2252LE. PMID: 32946275; PMCID: PMC7667903.
2. Mulla ZD, Pathak IS. Sleep Apnea and Poor COVID-19 Outcomes: Beware of Causal Intermediates and Colliders. Am J Respir Crit Care Med. 2021 May 15;203(10):1325-1326. doi: 10.1164/rccm.202101-0088LE. PMID: 33684329.

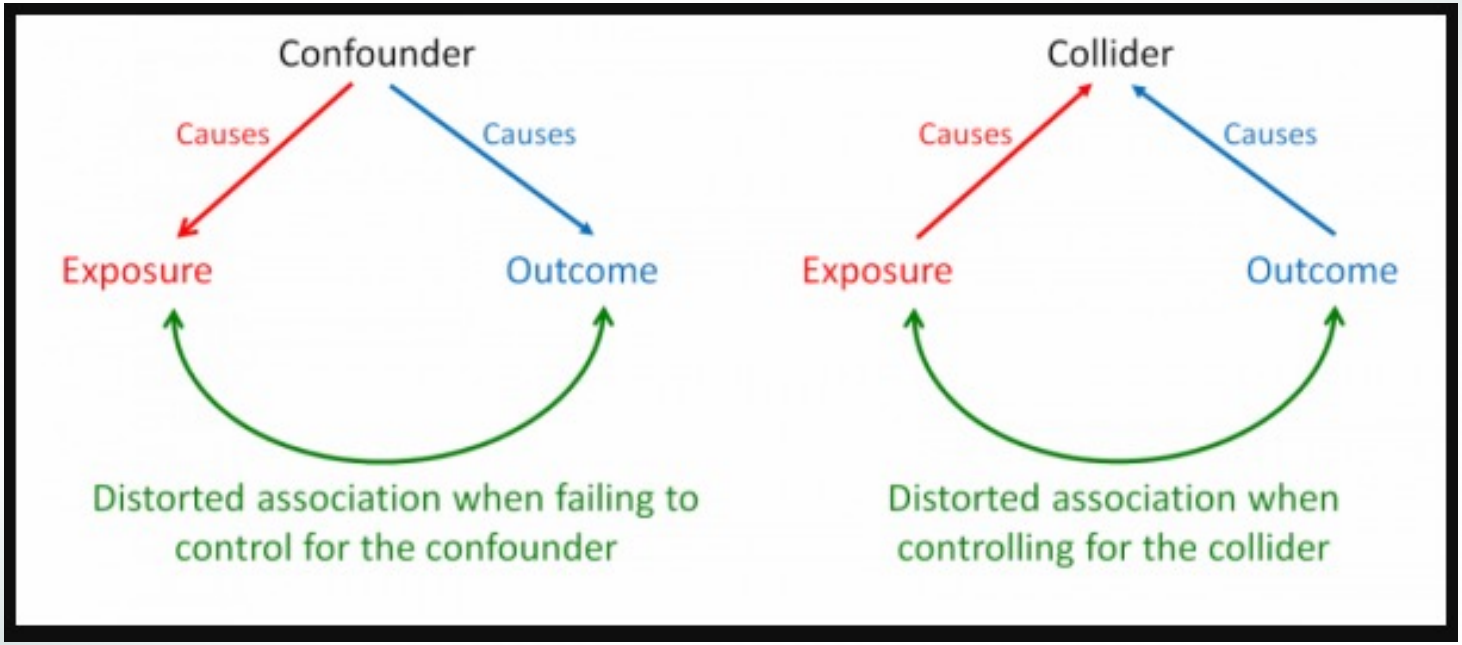
- **Mediator:** variables that are affected by the exposure and also affect the outcome
  - referred to as a mediator because it mediates, at least in part, the effect of hypertension on outcome
- **Confounder:** Variables that are on the common cause path of the exposure and outcome
  - conditioning on this variable through regression modelling, stratification in the analytical stage or restriction and exposure matching in the design stage, can prevent confounding
- Adjusting for a confounder removes bias, while adjusting for a mediator may lead to overadjustment bias.

Collider structures



- Paths with *convergent* arrows
- When colliders *are not* conditioned on they block pathways
- Conditioning on a collider opens the path, inducing association between X and Y

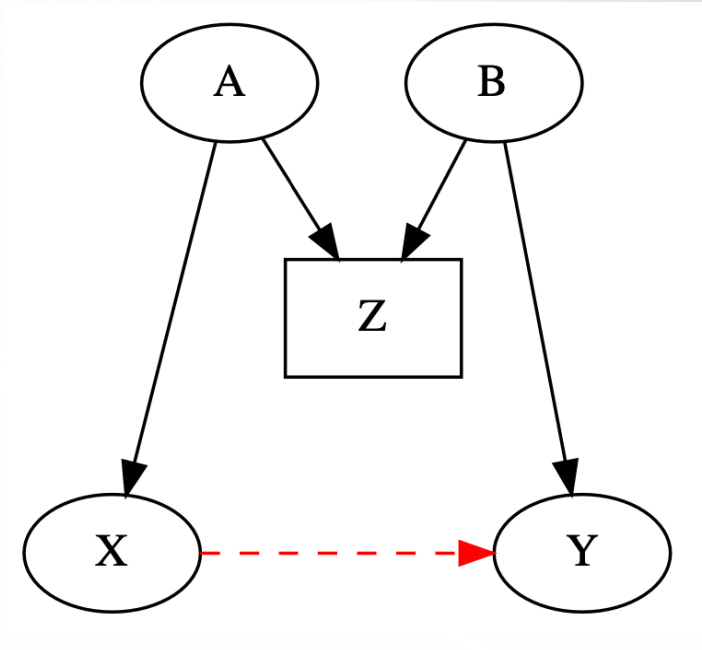
Confounders vs. Colliders



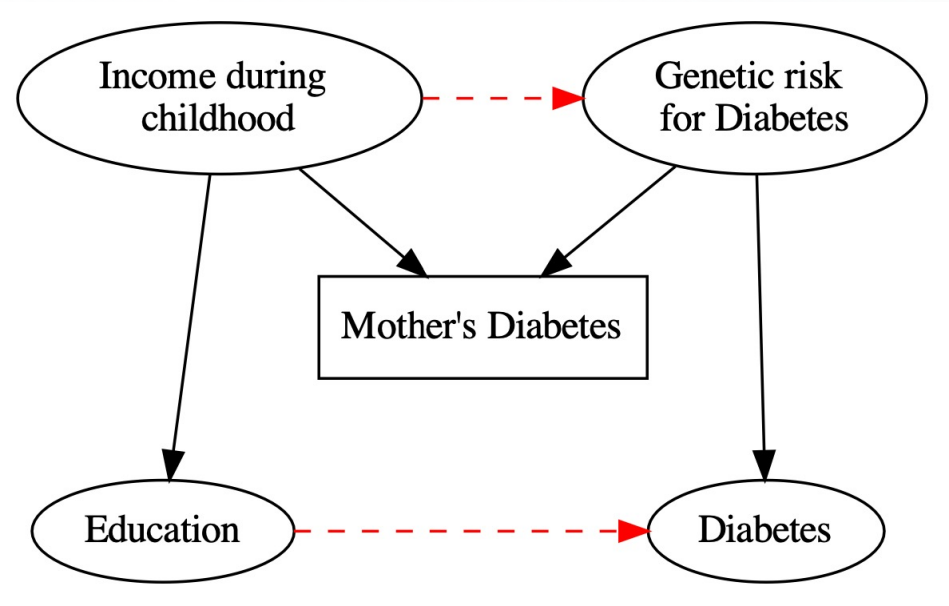
Catalog of Bias

Collider structures

- With colliders X->Z<-Y: $X \perp\!\!\!\perp Y \mid Z$
- The 'back-door' path
  $X \leftarrow \circ \rightarrow Z \leftarrow \circ \rightarrow Y$ *is blocked*
  when Z is *not* conditioned on
- Conditioning on a colliders *opens* a 'back door' path: $X \not\perp\!\!\!\perp Y \mid Z$
- More eloborate collider structures: e.g. "M-bias", etc.



Collider “M-bias”



Conditioning on the common effect (Mother’s Diabetes) imparts an association between two otherwise independent variables (Income and Genetics), leading to confounding via a backdoor path

Common Structures of Bias; Malcolm Barrett; 2021-01-11

Beware of Causal Intermediates *and* Colliders

Hospitalization and death among adults with COVID-19

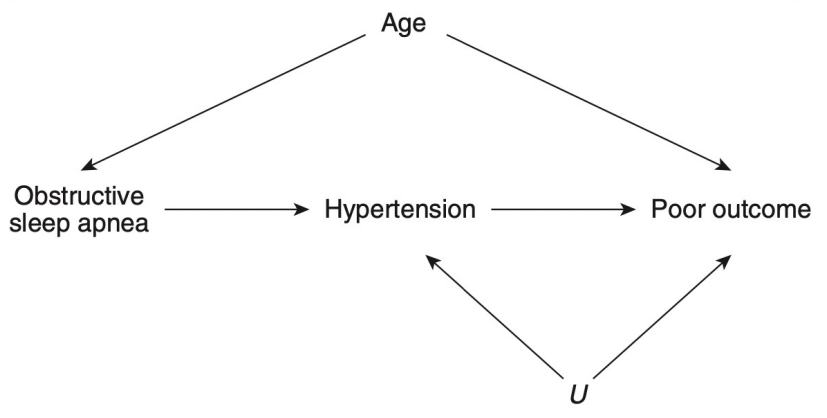


Figure 2. Directed acyclic graph for the effect of obstructive sleep apnea on poor outcome among patients with coronavirus disease (COVID-19). Hypertension is a collider on the path from obstructive sleep apnea to poor outcome. *U* is an unmeasured variable such as a medication or illness.

1. Cade BE, Dashti HS, Hassan SM, Redline S, Karlson EW. Sleep Apnea and COVID-19 Mortality and Hospitalization. Am J Respir Crit Care Med. 2020 Nov 15;202(10):1462-1464. doi: 10.1164/rccm.202006-2252LE. PMID: 32946275; PMCID: PMC7667903.
2. Mulla ZD, Pathak IS. Sleep Apnea and Poor COVID-19 Outcomes: Beware of Causal Intermediates and Colliders. Am J Respir Crit Care Med. 2021 May 15;203(10):1325-1326. doi: 10.1164/rccm.202101-0088LE. PMID: 33684329.

- Variables can be mediators, colliders and confounders (Hypertension is a mediator and *also* a collider)
- A back-door path can be inadvertently opened by conditioning on a collider
- Conditioning on a collider can introduce a spurious association between its causes.
- Controlling for a collider can result in a bias that is strong enough to move the observed association in a direction that is opposite of the true effect.

Beware of Causal Intermediates and Colliders

Hospitalization and death among adults with COVID-19

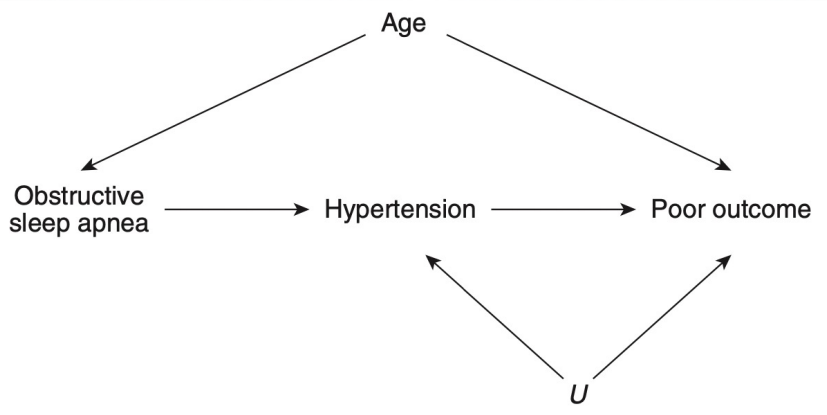
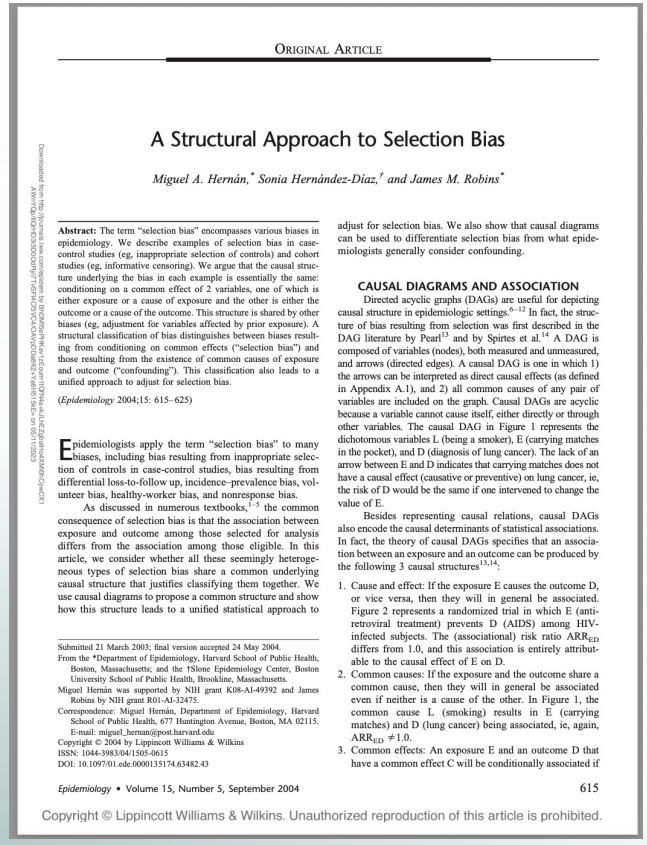


Figure 2. Directed acyclic graph for the effect of obstructive sleep apnea on poor outcome among patients with coronavirus disease (COVID-19). Hypertension is a collider on the path from obstructive sleep apnea to poor outcome. *U* is an unmeasured variable such as a medication or illness.

- Hypertension is a collider on the path from OSA to PO. Variable U is an unmeasured variable, such as a medication or illness, that affects the risk of both hypertension and PO. If the data analyst controls for hypertension but does not control for U in this situation, then collider stratification bias will occur.

1. Cade BE, Dashti HS, Hassan SM, Redline S, Karlson EW. Sleep Apnea and COVID-19 Mortality and Hospitalization. Am J Respir Crit Care Med. 2020 Nov 15;202(10):1462-1464. doi: 10.1164/rccm.202006-2252LE. PMID: 32946275; PMCID: PMC7667903.
2. Mulla ZD, Pathak IS. Sleep Apnea and Poor COVID-19 Outcomes: Beware of Causal Intermediates and Colliders. Am J Respir Crit Care Med. 2021 May 15;203(10):1325-1326. doi: 10.1164/rccm.202101-0088LE. PMID: 33684329.

Collider structures: “Selection bias”



ORIGINAL ARTICLE

A Structural Approach to Selection Bias

Miguel A. Hernán, Sonia Hernández-Díaz,† and James M. Robins**

Abstract: The term “selection bias” encompasses various biases in epidemiology. We describe examples of selection bias in case-control studies (eg, inappropriate selection of controls) and cohort studies (eg, informative censoring). We argue that the causal structure underlying the bias in each example is essentially the same: conditioning on a common effect of 2 variables, one of which is either exposure or a cause of exposure and the other is either the outcome or a cause of the outcome. This structure is shared by other biases (eg, adjustment for variables affected by prior exposure). A structural classification of bias distinguishes between biases resulting from conditioning on common effects (“selection bias”) and those resulting from the existence of common causes of exposure and outcome (“confounding”). This classification also leads to a unified approach to adjust for selection bias.

(*Epidemiology* 2004;15: 615–625)

Epidemiologists apply the term “selection bias” to many biases, including bias resulting from inappropriate selection of controls in case-control studies, bias resulting from differential loss-to-follow up, incidence–prevalence bias, volunteer bias, healthy-worker bias, and nonresponse bias.

As discussed in numerous textbooks,[1–3] the common consequence of selection bias is that the association between exposure and outcome among those selected for analysis differs from the association among those eligible. In this article, we consider whether all these seemingly heterogeneous types of selection bias share a common underlying causal structure that justifies classifying them together. We use causal diagrams to propose a common structure and show how this structure leads to a unified statistical approach to adjust for selection bias. We also show that causal diagrams can be used to differentiate selection bias from what epidemiologists generally consider confounding.

CAUSAL DIAGRAMS AND ASSOCIATION

Directed acyclic graphs (DAGs) are useful for depicting causal structure in epidemiologic settings.[6–12] In fact, the structure of bias resulting from selection was first described in the DAG literature by Pearl[13] and by Spirtes et al.[14] A DAG is composed of variables (nodes), both measured and unmeasured, and arrows (directed edges). A causal DAG is one in which 1) the arrows can be interpreted as direct causal effects (as defined in Appendix A.1), and 2) all common causes of any pair of variables are included on the graph. Causal DAGs are acyclic because a variable cannot cause itself, either directly or through other variables. The causal DAG in Figure 1 represents the dichotomous variables L (being a smoker), E (carrying matches in the pocket), and D (diagnosis of lung cancer). The lack of an arrow between E and D indicates that carrying matches does not have a causal effect (causative or preventive) on lung cancer, ie, the risk of D would be the same if one intervened to change the value of E.

Besides representing causal relations, causal DAGs also encode the causal determinants of statistical associations. In fact, the theory of causal DAGs specifies that an association between an exposure and an outcome can be produced by the following 3 causal structures[13,14]:

1. Cause and effect: If the exposure E causes the outcome D, or vice versa, then they will in general be associated. Figure 2 represents a randomized trial in which E (antiretroviral treatment) prevents D (AIDS) among HIV-infected subjects. The (associational) risk ratio ARR_{ED} differs from 1.0, and this association is entirely attributable to the causal effect of E on D.
2. Common causes: If the exposure and the outcome share a common cause, then they will in general be associated even if neither is a cause of the other. In Figure 1, the common cause L (smoking) results in E (carrying matches) and D (lung cancer) being associated, ie, again, $ARR_{ED} \neq 1.0$.
3. Common effects: An exposure E and an outcome D that have a common effect C will be conditionally associated if

Submitted 21 March 2003; final version accepted 24 May 2004.
From the *Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts; and the †Slone Epidemiology Center, Boston University School of Public Health, Brookline, Massachusetts.
Miguel Hernán was supported by NIH grant K08-AI-49392 and James Robins by NIH grant R01-AI-32475.
Correspondence: Miguel Hernán, Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115. E-mail: miguel_hernan@post.harvard.edu

ISSN: 1044-3983/04/1505-0615
DOI: 10.1097/01.ede.0000135174.63482.43

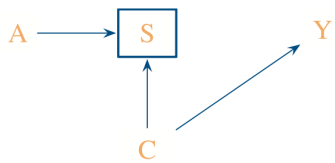


Figure S2.4. DAG illustrating selection bias. Treatment (A) is randomized. Subjects randomized to CCBs (A=1) are more likely to drop out due to adverse drug effects. Subjects with alcohol abuse (C=1) are more likely to drop out of the study and they are also more likely to experience acute liver failure (Y=1). Conditioning on selection (retention in study) (S=1) induces as association between A and C, which results in an open biasing pathway between A and Y.

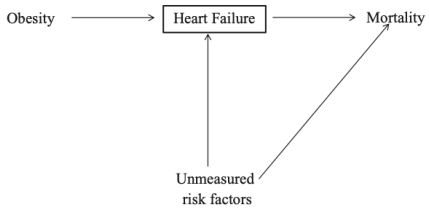


FIGURE. Directed acyclic graph of the hypothesized effects of obesity on mortality among individuals with heart failure. Potential unmeasured risk factors include a genetic factors and lifestyle behaviors.

The “Obesity Paradox” Explained

To the Editor:

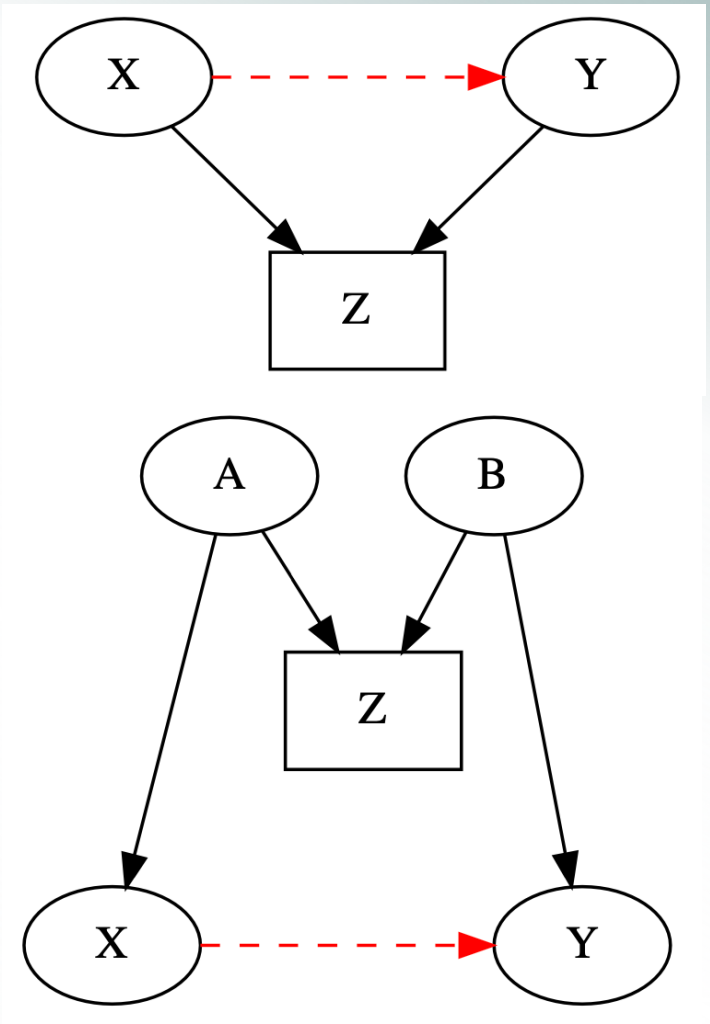
Several prospective studies have reported a J-shaped relationship between obesity and mortality, suggesting increased risk of death in the lowest and highest body mass index (BMI) groups in men and women of all ages, races, and ethnicities.[1] Although obesity is associated with a higher overall mortality risk in the general population, some authors have interpreted these patterns to suggest that obesity confers a survival advantage in surviving clinical subpopulations.[2] This “obesity paradox” has been reported for various disease groups including stroke, myocardial infarction, heart failure, renal disease, and diabetes.[2–5] We propose that this apparent paradox is simply the result of collider stratification, a source of selection bias that is common in epidemiologic research.[6]

The classic manifestation of this selection bias is a result of conditioning on a variable affected by exposure and sharing common causes with the outcome (known as a collider). Conditioning on a collider distorts the association between exposure and outcome among those selected for analysis and can therefore produce a spurious protective association between obesity and mortality in disease groups.

Banack, Hailey R.; Kaufman, Jay S.. The “Obesity Paradox” Explained. Epidemiology 24(3):p 461-462, May 2013.

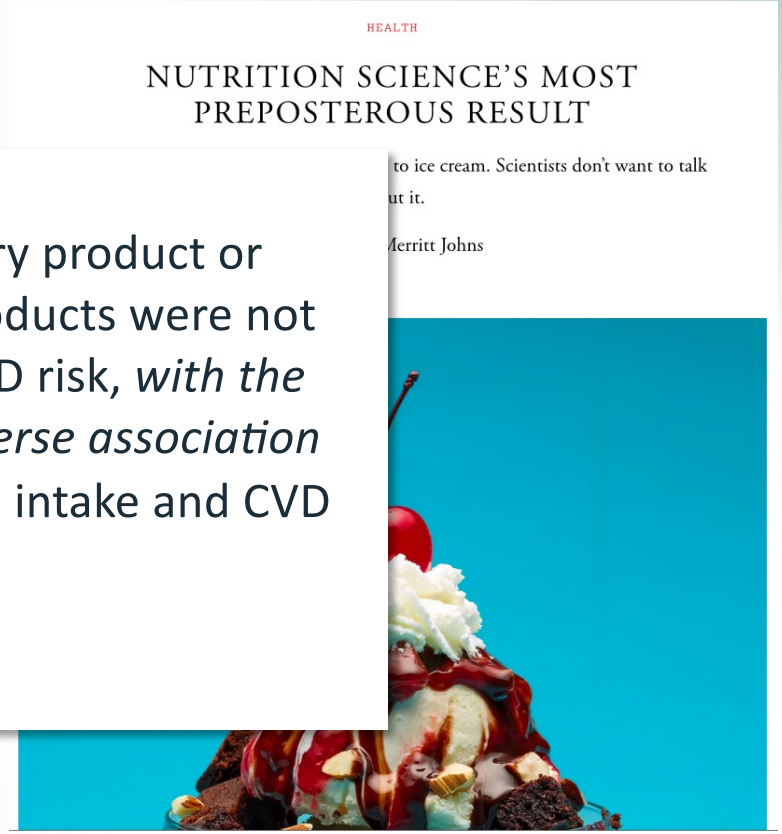
Collider structures

- Collider stratification bias
- Selection bias
  - Type-1
  - Type-2
- “Selection distortion effect”
- Differential follow-up bias
- Berkson's paradox
- Simpson’s paradox
- … paradox’s



Collider "M-bias" as "selection bias" and paradoxes

"Intake of total dairy product or individual dairy products were not associated with CVD risk, *with the exception of an inverse association* between ice-cream intake and CVD health outcomes."



Dairy Products and Cardiometabolic Health Outcomes, Andres Victor Ardisson Korate, 2018

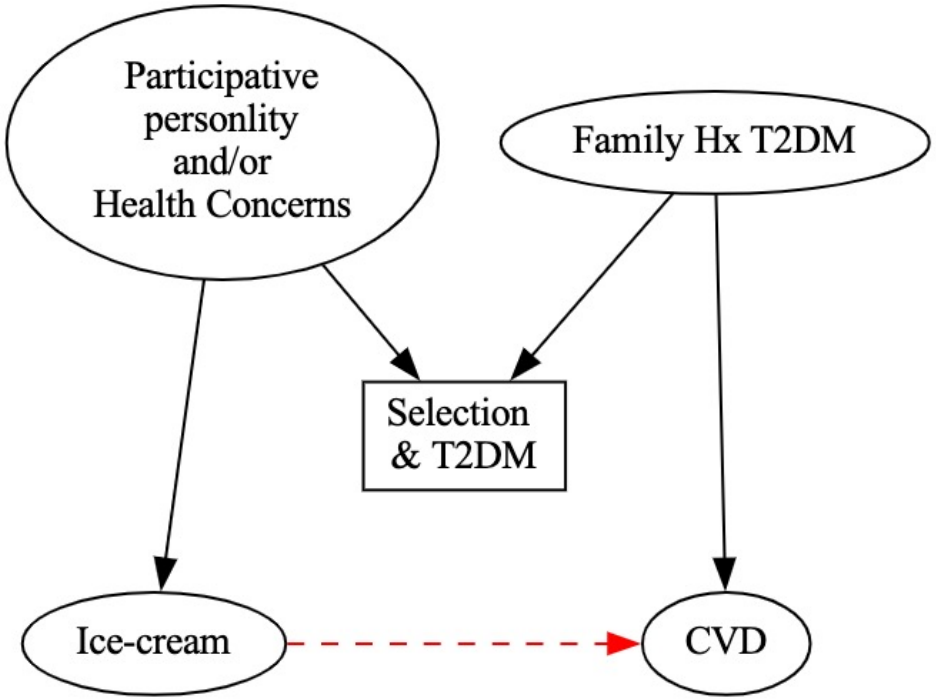
Collider "M-bias" as "Selection bias"

Table S2.3 HRs (95% CI) of cardiovascular disease (CVD) risk according to intakes of various dairy foods in participants with different diet update approaches in participants from both NHS and HPFS cohorts*

	HR (95% CI) for one serving / day		
	Main model[1]	Cancer only[2]	HBP/HC[3]
Total dairy	1.00 (0.97, 1.02)	1.01 (0.98, 1.04)	0.99 (0.97, 1.02)
High-fat dairy	0.96 (0.92, 1.00)	0.97 (0.93, 1.01)	0.96 (0.93, 1.01)
Low-fat dairy	1.02 (0.98, 1.05)	1.02 (0.99, 1.05)	1.02 (0.98, 1.05)
Cheese	1.00 (0.94. 1.07)	1.00 (0.94. 1.06)	1.00 (0.92, 1.08)
Skim/low-fat milk	1.00 (0.96, 1.04)	1.01 (0.97, 1.04)	1.01 (0.96, 1.07)
Whole milk	1.04 (0.94, 1.16)	1.02 (0.91, 1.13)	1.04 (0.94, 1.14)
Yogurt	0.98 (0.85, 1.13)	0.99 (0.86, 1.15)	1.03 (0.83, 1.26)
Fermented dairy products	1.00 (0.94, 1.06)	1.00 (0.94, 1.06)	1.00 (0.90. 1.09)
Cream	0.98 (0.91, 1.05)	0.98 (0.90, 1.05)	1.03 (0.94, 1.12)
Ice cream	0.82 (0.67, 0.99)	0.79 (0.64, 0.96)	0.79 (0.64, 0.96)
Sherbet	0.92 (0.78, 1.11)	0.92 (0.76, 1.09)	1.29 (0.73, 1.26)
Butter[¥]	1.00 (0.95, 1.06)	0.99 (0.94, 1.05)	1.03 (0.98, 1.07)
Dairy fat (1% calories) [§]	0.99 (0.98, 1.00)	0.99 (0.98, 1.00)	1.00 (0.99, 1.01)

* HPFS, Health Professionals Follow-Up Study; NHS, Nurses' Health Study.

[1] Model was adjusted for age (continuous), sex, BMI (4 categories), and total energy intake (quintiles), race, menopausal status [pre or postmenopausal (never, past or current menopausal hormone use)], family history of diabetes (yes/no), family history of myocardial infarction (yes/no), alcohol intake (0, 1-4.9, 5-14.9, >15 g/day), smoking status (never, past, current 1-15 cigarettes/day, >15 cigarettes/day), physical activity (0, 0.1-0.9, 1-3.5, >3.5 hrs./week) current aspirin use (yes/no), current multivitamin use (yes/no), diabetes duration (<5, 5-10, >10 years), baseline hypertension, baseline hypercholesterolemia, lag-time between T2D diagnosis and return of first FFQ, AHEI, and mutually adjusted for other dairy products. Diet update was stopped after diagnosis of cancer, CABG, or angina



Dairy Products and Cardiometabolic Health Outcomes, Andres Victor Ardisson Korate, 2018

Growing awareness of mischief of colliders

MacElreath on Twitter

ARTICLE

Check for updates

https://doi.org/10.1038/s41467-020-19478-2 OPEN

Collider bias undermines our understanding of COVID-19 disease risk and severity

Gareth J. Griffith[1,2,4], Tim T. Morris[1,2,4], Matthew J. Tudball[1,2,4], Annie Herbert[1,2,4], Giulia Mancano[1,2,4], Lindsey Pike[1,2], Gemma C. Sharp[1,2], Jonathan Sterne[2], Tom M. Palmer[1,2], George Davey Smith[1,2], Kate Tilling[1,2], Luisa Zuccolo[1,2], Neil M. Davies[1,2,3] & Gibran Hemani[1,2,4]✉

Numerous observational studies have attempted to identify risk factors for infection with SARS-CoV-2 and COVID-19 disease outcomes. Studies have used datasets sampled from patients admitted to hospital, people tested for active infection, or people who volunteered to participate. Here, we highlight the challenge of interpreting observational evidence from such non-representative samples. Collider bias can induce associations between two or more variables which affect the likelihood of an individual being sampled, distorting associations between these variables in the sample. Analysing UK Biobank data, compared to the wider cohort the participants tested for COVID-19 were highly selected for a range of genetic, behavioural, cardiovascular, demographic, and anthropometric traits. We discuss the mechanisms inducing these problems, and approaches that could help mitigate them. While collider bias should be explored in existing studies, the optimal way to mitigate the problem is to use appropriate sampling strategies at the study design stage.

Adjustment: Information propagation, and interruption

$\boldsymbol{X} \rightarrow Z \rightarrow \boldsymbol{Y}$

- X and Y are associated; *unless conditioning on Z*

$\boldsymbol{X} \leftarrow Z \rightarrow \boldsymbol{Y}$

- X and Y are associated; *unless conditioning on Z*

$\boldsymbol{X} \rightarrow Z \leftarrow \boldsymbol{Y}$

- X and Y are *not* associated; *unless conditioning on Z*

“What *causes* say about *data*”

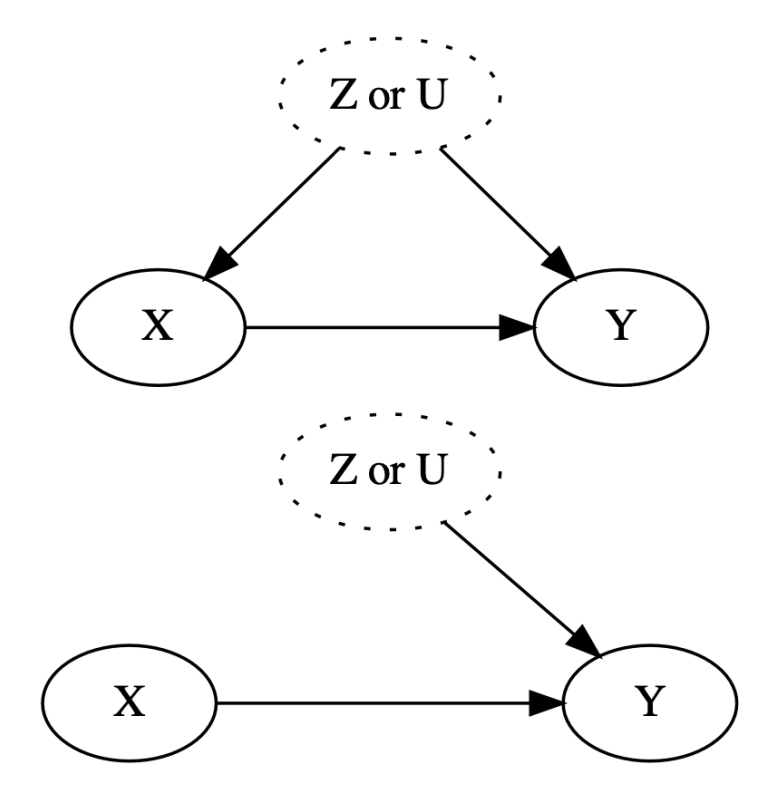
- Causal diagrams show how causal relations are expected to translate into *associations & independencies*
    1. Initially, *associations & independencies* derived from subject matter knowledge are posited in a DAG
    2. Then given the posited model, *associations & independencies* observed in data are are computed
- A credible causal model will reconcile *associations & independencies* observed with the constraints provided by the posited causal model
- Subject to further criticism; revision qualification, elaboration, updating, refinement

Intervention ~ de-confounding

$$P(Y \mid X) = P(Y \mid do(X))$$

Why we really care about pipes, forks and colliders?

P(Y|*do*(X)) ~ Deconfounding



- Heuristic: an RCT helps us define a causal effect in SCMs
- Causal effects of X: arrows leaving X
- Confounding requires an arrow into X
- "*do*(X)": an intervention; no exogenous determinants
  - no arrow into X, no biasing (backdoor) pathways
- An un-confounded estimate emulates instrumental control:

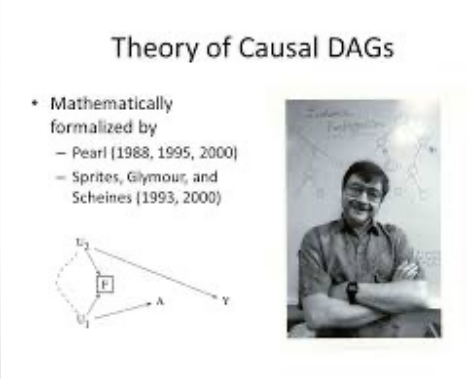
$$P(Y \mid X) = P(Y \mid do(X))$$

De-confounding by emulating P(Y|*do*(X))

- Understanding confounding as $P(Y|X) \neq P(Y|do(X))$, we seek $P(Y \mid X) \cong P(Y \mid do(X))$
- We analyze a DAG for "d-separation": i.e., for any given pattern of paths in the causal model, what pattern of dependencies and independencies we should expect in the data
- We then seek *adjustment strategies* for unbiased estimation of effects [where $P(Y \mid X) \cong P(Y \mid do(X))$]

- Variables are *d-separated* if:
  1. not connected with each other (no pathway)
  2. or pathway is blocked
     - adjusted non-colliders
     - connected only through path on which at least one unadjusted collider
- otherwise there are open pathways and dependencies communicated

The “do-calculus”



A causal path from exposure to outcome
1. Is open (by definition it does not contain any collider variables)
2. Should be left open (do not adjust for any variables on these causal paths)

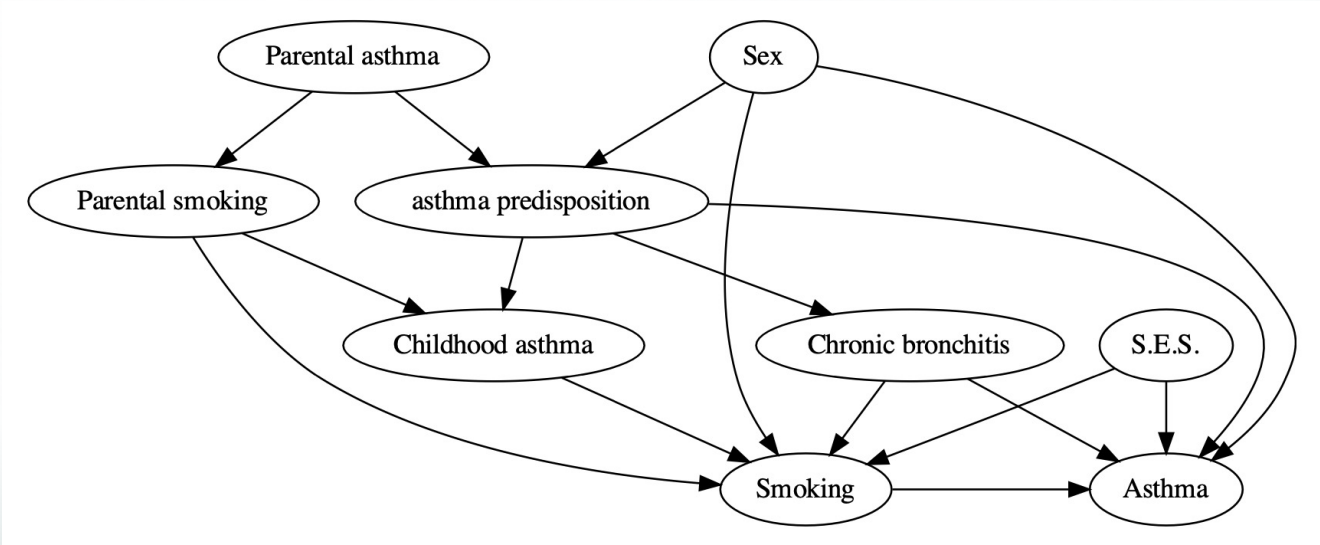
A non-causal path from exposure to outcome containing no collider variables
3. Is open if no variables on the path are adjusted for
4. Is closed if one or more variables on the path are adjusted for

A non-causal path from exposure to outcome containing one collider variable
5. Is closed if no variables on the path are adjusted for
6. Is closed if only non-collider variables are adjusted for
7. Is open if the collider variable,* is the only variable on the path adjusted for
8. Is closed if the collider variable,* and one or more other (non-collider) variables are adjusted for

A non-causal path from exposure to outcome containing more than one collider variable
9. Is closed if no variables (or only non-collider variables) on the path are adjusted for
10. Is closed if at least one collider variable,* is not adjusted for
11. Is open if all the collider variables,* but no non-collider variables, are adjusted for
12. Is closed if all collider variables,* and one or more other (non-collider) variables are adjusted for

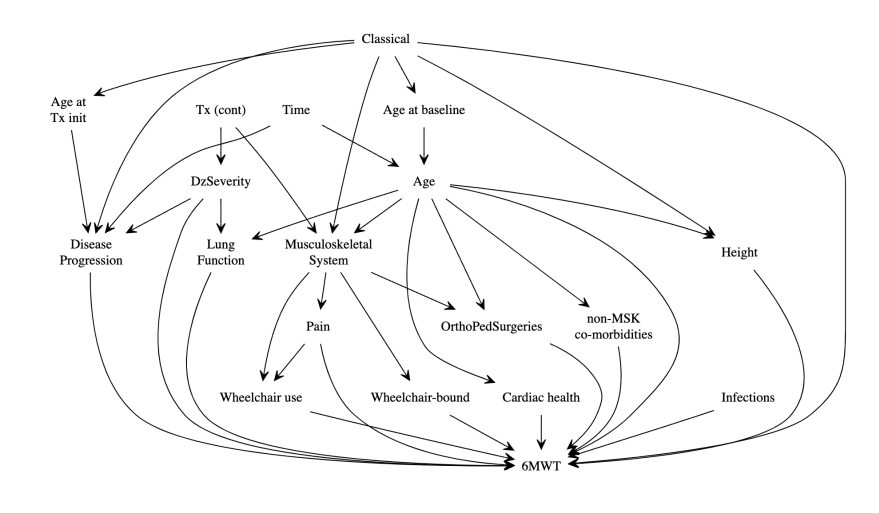
Figure 2 Rules to decide whether a particular path is open or closed in a causal diagram. *The same rules apply if, instead of adjusting for a collider, we adjust for a variable that is caused by that collider.

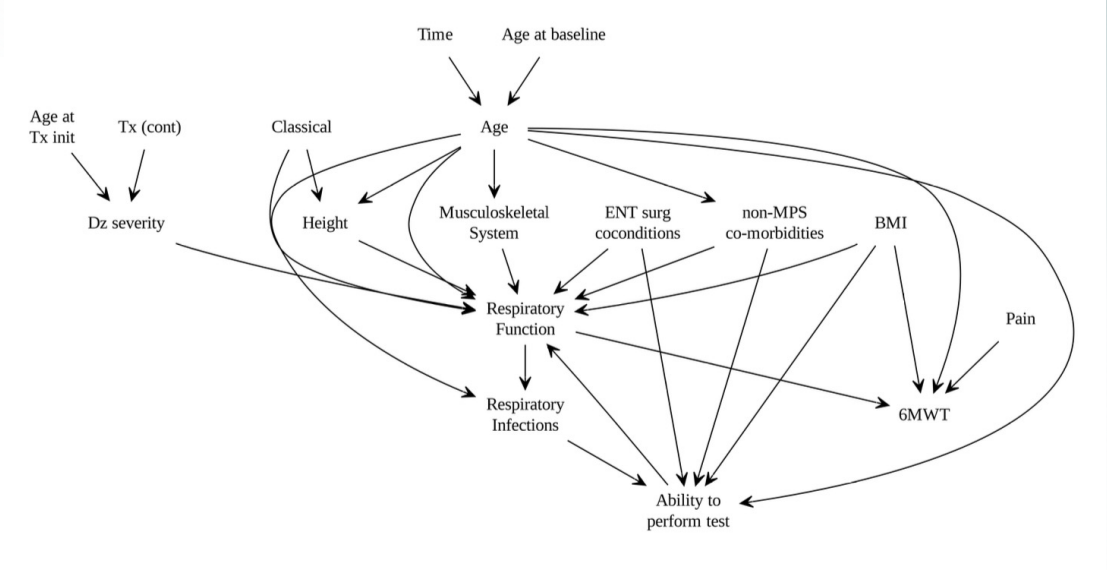
It can get complicated …



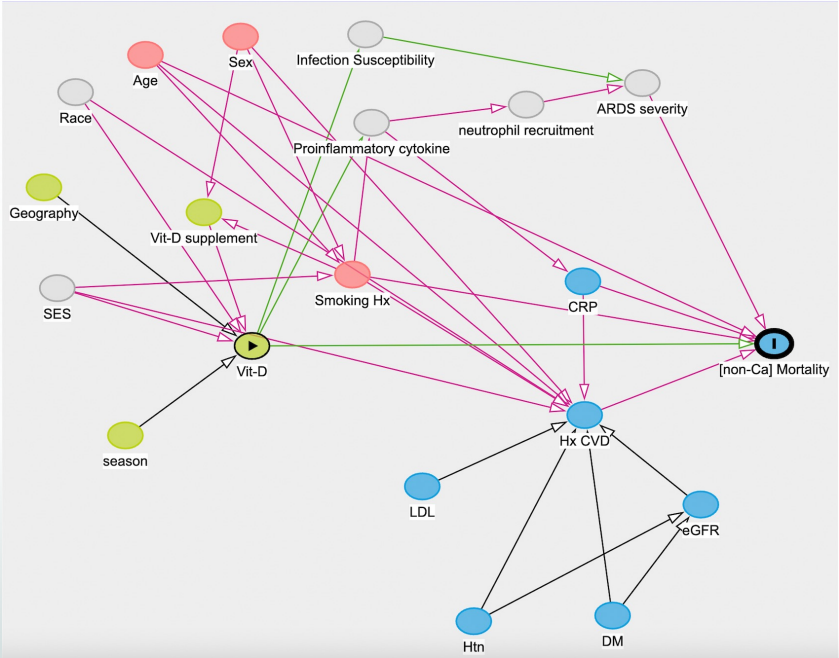
Williamson, et al, 2014

It can get complicated …





Elucidate complexity



"The whole art and practice of scientific [work] is comprised of the skillful interrogation of Nature."

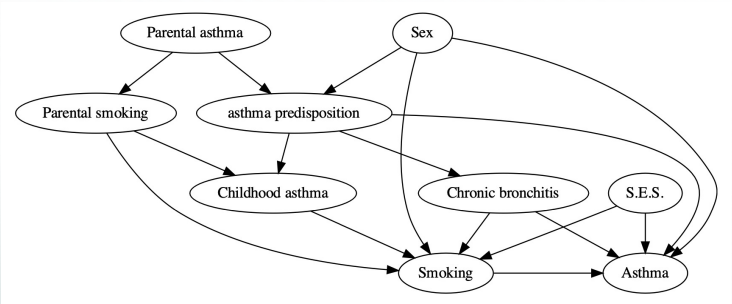
— Joan Fisher Box

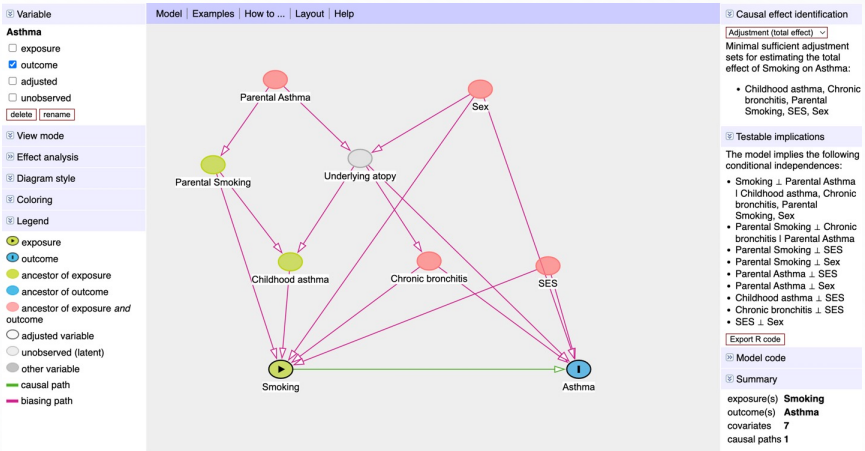
SCMs allow us to

- make our assumptions explicit
- communicate complexity to stakeholders
- qualify our findings
- address sources of uncertainty
- license "transportability" of effects

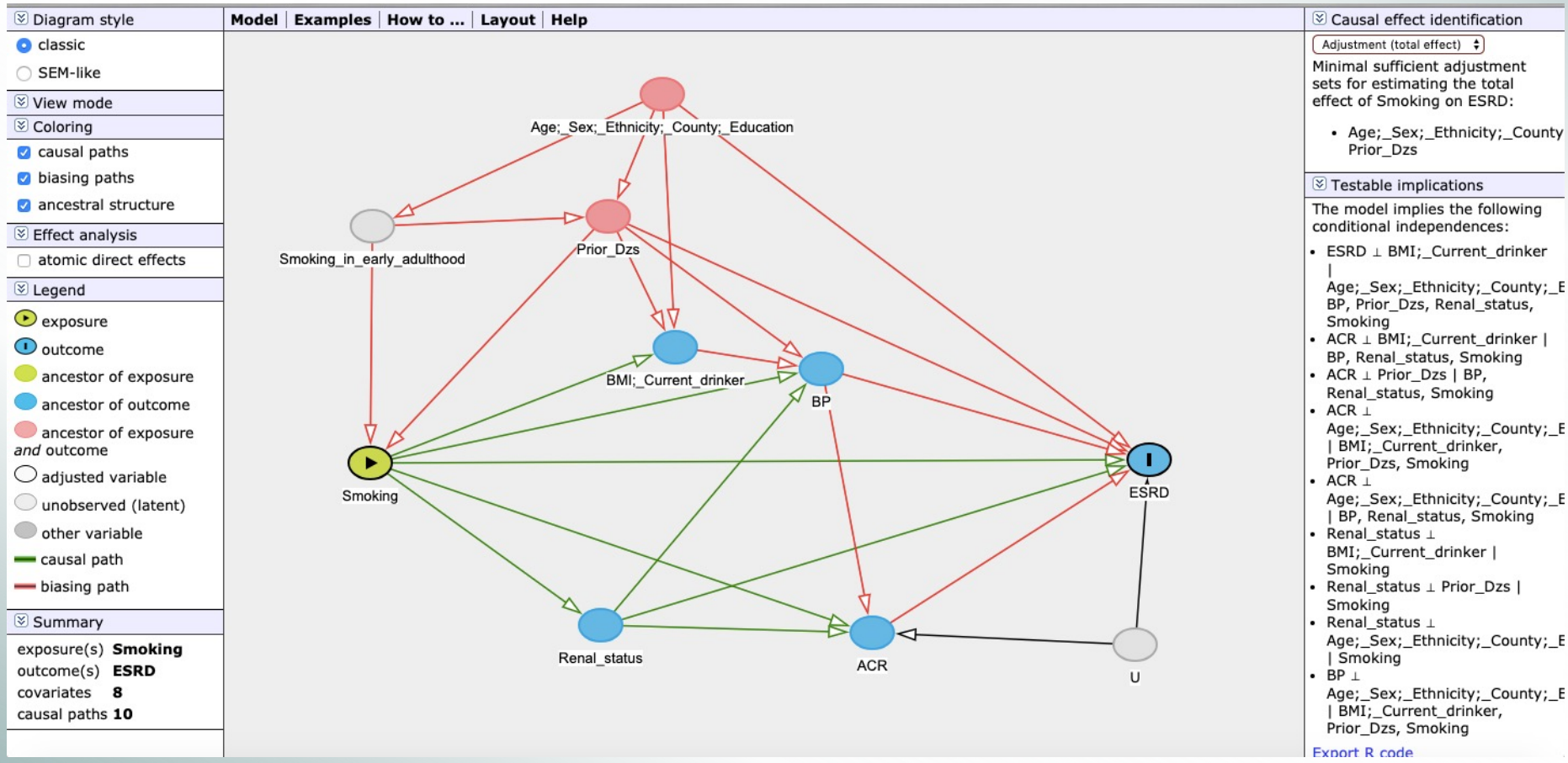
… to analyze the DAG

We have to do the work of positing and articulating a SCM;
but we have tools to do the causal ‘calculus with a DAG





Daggity: - drawing and analyzing causal diagrams (DAGs) (www.dagitty.net/)



Variable

[non-Ca] Mortality

- [ ] exposure
- [x] outcome
- [ ] adjusted
- [ ] unobserved

delete | rename

View mode

Effect analysis

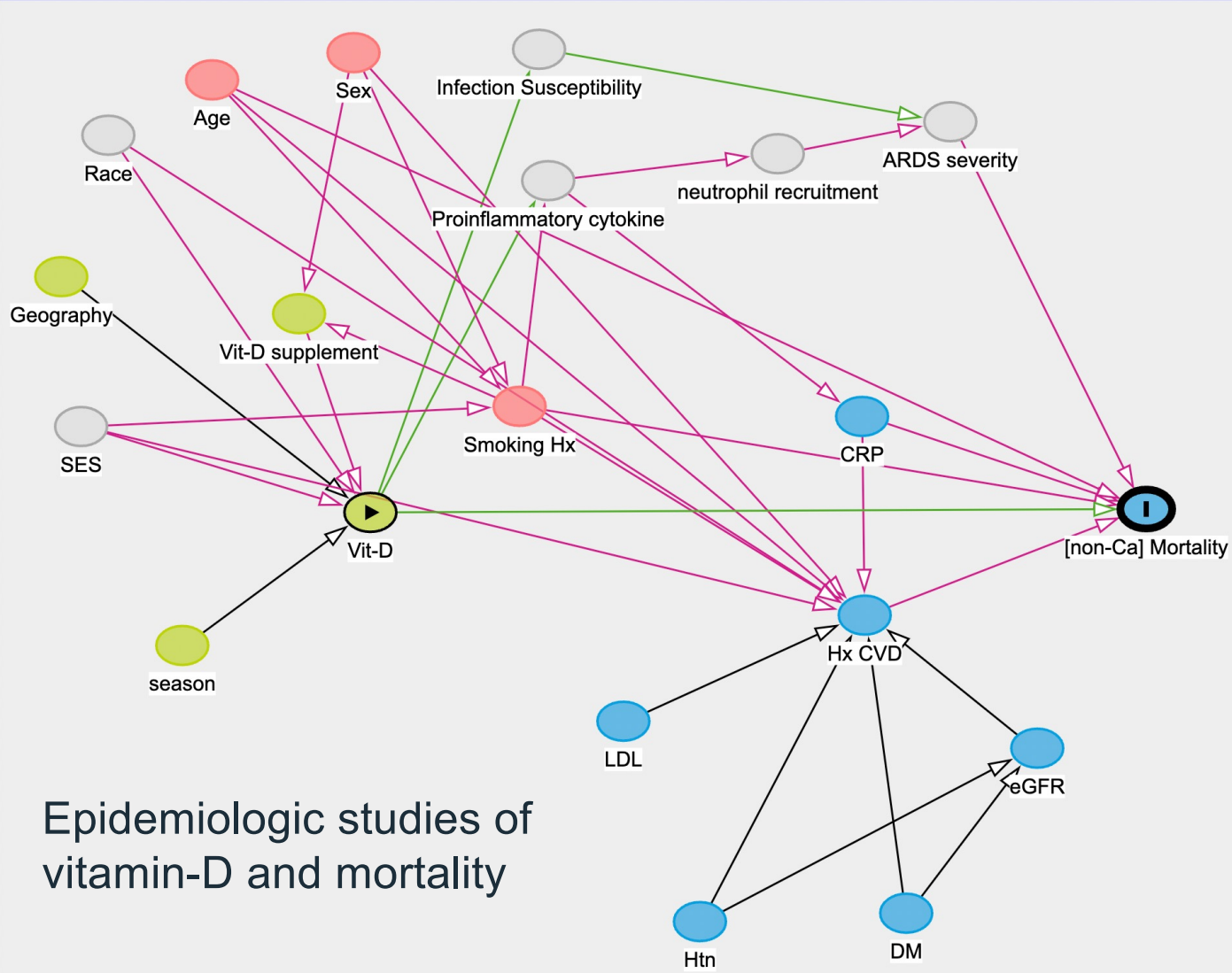
Diagram style

Coloring

Legend

- exposure
- outcome
- ancestor of exposure
- ancestor of outcome
- ancestor of exposure *and* outcome
- adjusted variable
- unobserved (latent)
- other variable
- causal path
- biasing path

Model | Examples | How to ... | Layout | Help



Epidemiologic studies of vitamin-D and mortality

Causal effect identificatio

Adjustment (total effect)

The total effect cannot be estimated by covariate adjustment.

Testable implications

The model implies the follov conditional independences:

- Vit-D supplement ⊥ Age I Sex, Smoking Hx
- Vit-D supplement ⊥ Geography
- Vit-D supplement ⊥ DM
- Vit-D supplement ⊥ Htn
- Vit-D supplement ⊥ eGFI
- Vit-D supplement ⊥ CRP Smoking Hx, Vit-D
- Vit-D supplement ⊥ LDL
- Vit-D supplement ⊥ seas
- Vit-D supplement ⊥ [non- Mortality I Age, CRP, Hx CVD, Smoking Hx, Vit-D

Show all ...

Export R code

Model code

Summary

exposure(s)	**Vit-D**
outcome(s)	**[non-Ca] Mortality**
covariates	**18**
causal paths	**5**

Variable

Vit-D ICU baseline

- exposure
- outcome
- adjusted
- unobserved

delete | rename

View mode

Effect analysis

- atomic direct effects

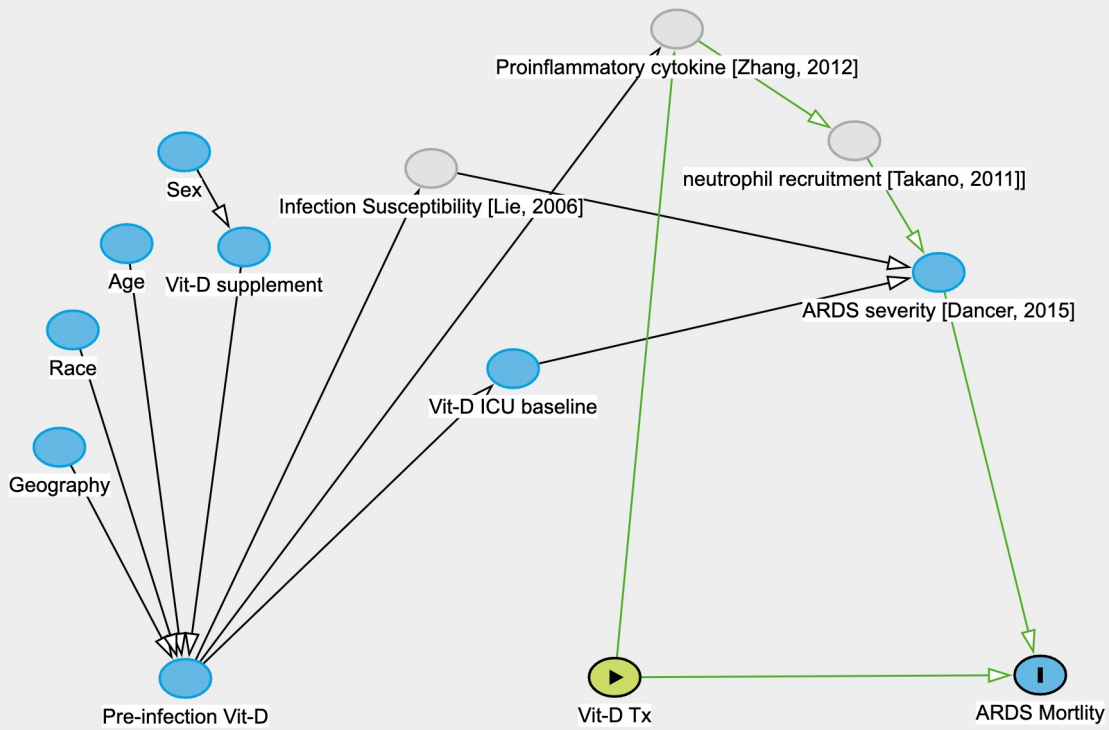
Diagram style

Coloring

Legend

- exposure
- outcome
- ancestor of exposure
- ancestor of outcome
- ancestor of exposure *and* outcome
- adjusted variable
- unobserved (latent)
- other variable
- causal path
- biasing path

Model | Examples | How to ... | Layout | Help



The VIOLET Randomized Controlled Trial of vitamin-D and mortality

Causal effect identification

Adjustment (total effect)

No adjustment is necessary to estimate the total effect of Vit-D Tx on ARDS Mortlity.

Testable implications

The model implies the following conditional independences:

- ARDS Mortlity ⊥ Pre-infection Vit-D | ARDS severity [Dancer, 2015], Vit-D Tx
- ARDS Mortlity ⊥ Vit-D ICU baseline | ARDS severity [Dancer, 2015], Vit-D Tx
- ARDS Mortlity ⊥ Vit-D supplement | Pre-infection Vit-D
- ARDS Mortlity ⊥ Vit-D supplement | ARDS severity [Dancer, 2015], Vit-D Tx
- ARDS Mortlity ⊥ Age | Pre-infection Vit-D
- ARDS Mortlity ⊥ Age | ARDS severity [Dancer, 2015], Vit-D Tx
- ARDS Mortlity ⊥ Geography | Pre-infection Vit-D
- ARDS Mortlity ⊥ Geography | ARDS severity [Dancer, 2015], Vit-D Tx
- ARDS Mortlity ⊥ Race | Pre-infection Vit-D

Show all ...

Export R code

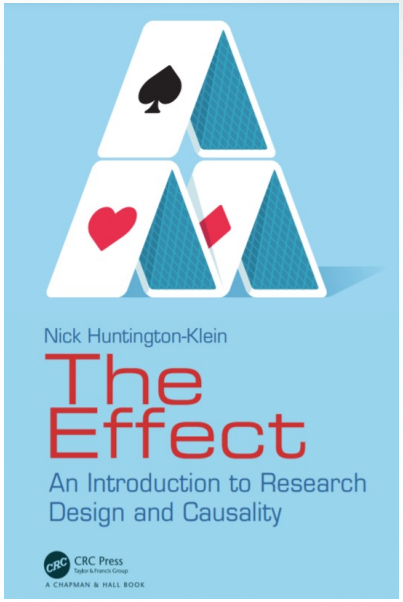
Model code

Summary

exposure(s)	**Vit-D Tx**
outcome(s)	**ARDS Mortlity**
covariates	**11**
causal paths	**2**

Proposed process for using SCMs and DAGs

1. Model the *data generating process*
2. List out all paths
3. Find a set of variables that close all back doors
4. Measure and control for all those variables



The Effect: An Introduction to Research Design and Causality
Nick Huntington-Klein, 2022

Proposed process for using SCMs and DAGs

1. Think hard about the research question and problem of effect identification ("skillful interrogation of Nature")
2. Develop DAGs based on subject matter knowledge *without* looking at data: *do not contort the DAG based on data availability*
3. Do the '*causal calculus*' in Daggity to identify the set of minimum necessary adjustment for unbiased effect estimation
4. Do analysis and reconcile observations with causal model (this *is* science)
5. Publish the DAG with the research report

The Limitations!

The limitations

Eur J Epidemiol
DOI 10.1007/s10654-015-9995-7

METHODS

Limitations of individual causal models, causal graphs, and ignorability assumptions, as illustrated by random confounding and design unfaithfulness

Sander Greenland · Mohammad Ali Mansournia

Received: 26 March 2014 / Accepted: 22 January 2015

Abstract We describe how ordinary interpretations of causal models and causal graphs fail to capture important distinctions among ignorable allocation mechanisms for subject selection or allocation. We illustrate these limitations in the case of random confounding and designs that prevent such confounding. In many experimental designs individual treatment allocations are dependent, and explicit population models are needed to show this dependency. In particular, certain designs impose unfaithful covariate-treatment distributions to prevent random confounding, yet ordinary causal graphs cannot discriminate between these

Keywords Causal graphs · Confounding · Directed acyclic graphs · Ignorability · Inverse probability weighting · Unfaithfulness

Introduction

Potential-outcome (counterfactual) and graphical causal models are now standard tools for analysis of study designs and data. Expositions can be found in modern textbooks [1–3]; in most applications we see, however, the

Journal of Epidemiology

Special Article J Epidemiol 2020;30(4):153-162

Causal Diagrams: Pitfalls and Tips

Etsuji Suzuki[1], Tomohiro Shinozaki[2], and Eiji Yamamoto[3]

[1]Department of Epidemiology, Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama University, Okayama, Japan
[2]Department of Information and Computer Technology, Faculty of Engineering, Tokyo University of Science, Tokyo, Japan
[3]Okayama University of Science, Okayama, Japan

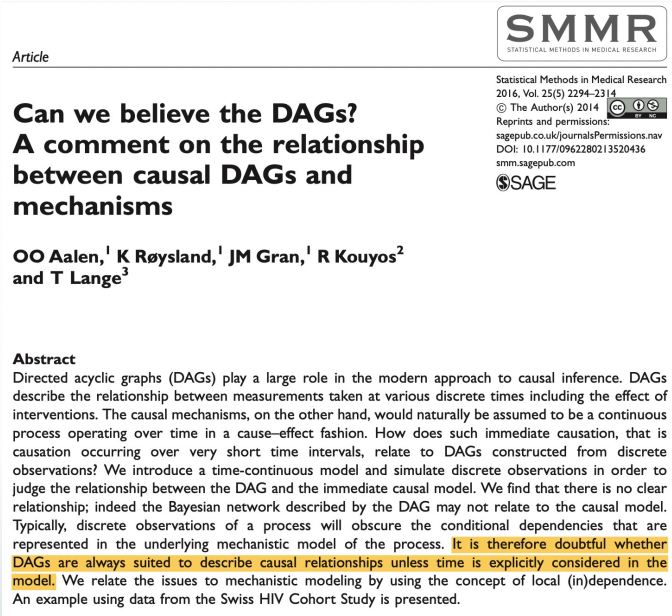
Received August 6, 2019; accepted October 11, 2019; released online February 1, 2020

ABSTRACT

Graphical models are useful tools in causal inference, and causal directed acyclic graphs (DAGs) are used extensively to determine the variables for which it is sufficient to control for confounding to estimate causal effects. We discuss the following ten pitfalls and tips that are easily overlooked when using DAGs: 1) Each node on DAGs corresponds to a random variable and not its realized values; 2) The presence or absence of arrows in DAGs corresponds to the presence or absence of individual causal effect in the population; 3) "Non-manipulable" variables and their arrows should be drawn with care; 4) It is preferable to draw DAGs for the total population, rather than for the exposed or unexposed groups; 5) DAGs are primarily useful to examine the presence of confounding in distribution in the notion of confounding in expectation; 6) Although DAGs provide qualitative differences of causal structures, they cannot describe details of how to adjust for confounding; 7) DAGs can be used to illustrate the consequences of matching and the appropriate handling of matched variables in cohort and case-control studies; 8) When explicitly accounting for temporal order in DAGs, it is necessary to use separate nodes for each timing; 9) In certain cases, DAGs with signed edges can be used in drawing conclusions about the direction of bias; and 10) DAGs can be (and should be) used to describe not only confounding bias but also other forms of bias. We also discuss recent developments of graphical models and their future directions.

Key words: bias; causal inference; causality; confounding; directed acyclic graphs

The limitations



SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Statistical Methods in Medical Research
2016, Vol. 25(5) 2294–2314

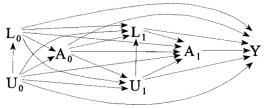
DOI: 10.1177/0962280213520436
smm.sagepub.com
SAGE

Can we believe the DAGs? A comment on the relationship between causal DAGs and mechanisms

OO Aalen,[1] K Røysland,[1] JM Gran,[1] R Kouyos[2] and T Lange[3]

Abstract
Directed acyclic graphs (DAGs) play a large role in the modern approach to causal inference. DAGs describe the relationship between measurements taken at various discrete times including the effect of interventions. The causal mechanisms, on the other hand, would naturally be assumed to be a continuous process operating over time in a cause–effect fashion. How does such immediate causation, that is causation occurring over very short time intervals, relate to DAGs constructed from discrete observations? We introduce a time-continuous model and simulate discrete observations in order to judge the relationship between the DAG and the immediate causal model. We find that there is no clear relationship; indeed the Bayesian network described by the DAG may not relate to the causal model. Typically, discrete observations of a process will obscure the conditional dependencies that are represented in the underlying mechanistic model of the process. It is therefore doubtful whether DAGs are always suited to describe causal relationships unless time is explicitly considered in the model. We relate the issues to mechanistic modeling by using the concept of local (in)dependence. An example using data from the Swiss HIV Cohort Study is presented.

- It can be difficult: "Causal Inference" ("the skillful integration of Nature") is a complex scientific task
- Specifying SCMs/DAGs is not easy
  - achieving consensus on SCM even harder
  - a 'complete' SCM (no omitted variables) harder still
- Static causal problems are easier; time-dependent confounding requires special methods



"What is simple is always wrong. What is not is unusable." —Valéry, Paul (1942)

The limitations

- It's not 'automatic': Specifying SCMs/DAGs is not easy
- Regression assumptions, C(Y|X)=$\boldsymbol{X\beta}$, include *no omitted predictors*
- DAGs should include all relevant variables, *including those where direct measurements are unavailable*
  - Explicitly depicting unobserved variables helps to highlight potential sources of unobserved confounding.
- Not clear that a "complete" SCMs ever achieved.

“identifiability” does not imply “estimability”

What can be estimated? Identifiability, estimability, causal inference and ill-posed inverse problems

Oliver J. Maclaren OLIVER.MACLAREN@AUCKLAND.AC.NZ
Department of Engineering Science
University of Auckland
Auckland 1142, New Zealand.

Ruanui Nicholson RUANUI.NICHOLSON@AUCKLAND.AC.NZ
Department of Engineering Science
University of Auckland
Auckland 1142, New Zealand.

Editor: TBD

Abstract

We consider basic conceptual questions concerning the relationship between statistical estimation and causal inference. Firstly, we show how to translate causal inference problems into an abstract statistical formalism without requiring any structure beyond an arbitrarily-indexed family of probability models. The formalism is simple but can incorporate a variety of causal modelling frameworks, including 'structural causal models', but also models expressed in terms of, e.g., differential equations. We focus primarily on the structural/graphical causal modelling literature, however. Secondly, we consider the extent to which causal and statistical concerns can be cleanly separated, examining the fundamental question: 'What can be estimated from data?'. We call this the problem of estimability. We approach this by analysing a standard formal definition of 'can be estimated' commonly adopted in the causal inference literature – identifiability – in our abstract statistical formalism. We use elementary category theory to show that identifiability implies the existence of a Fisher-consistent estimator, but also show that this estimator may be discontinuous, and thus unstable, in general. This difficulty arises because the causal inference problem is, in general, an ill-posed inverse problem. Inverse problems have three conditions which must be satisfied to be considered well-posed: existence, uniqueness, and stability of solutions. Here identifiability corresponds to the question of uniqueness; in contrast, we take estimability to mean satisfaction of all three conditions, i.e. well-posedness. Lack of stability implies that naive translation of a causally identifiable quantity into an achievable statistical estimation target may prove impossible. Our article is primarily expository and aimed at unifying ideas from multiple fields, though we provide new constructions and proofs.

Keywords: identifiability, estimability, causal inference, structural causal models, inverse problems, stability, robust statistics, statistical learning theory, sensitive parameters, applied category theory

1. Introduction

A common idea in much of the causal inference literature (see e.g. Pearl, 2009, and related work) is that there is a natural separation of concerns between causal inference and statistical estimation of the form:

Models, identifiability, and estimability in causal inference

Oliver J. Maclaren [1] **Ruanui Nicholson** [1]

Abstract

Here we discuss two common but, in our view, misguided assumptions in causal inference. The first assumption is that one requires potential outcomes, directed acyclic graphs (DAGs), or structural causal models (SCMs) for thinking about causal inference in statistics. The second is that identifiability of a quantity implies estimability of that quantity. These views are not universal, but we believe they are sufficiently common to warrant comment.

1. Overview

The focus of this extended abstract is two common but, in our view, misguided assumptions in causal inference. While these assumptions are not universal, and causal inference is diverse and multidisciplinary, we believe explicit discussion of them is worthwhile. The first assumption concerns the role and meaning of **models** in causal inference. It is common to assume that causal inference in statistics necessarily requires special causal modelling formalisms such as potential outcomes, directed acyclic graphs (DAGs), or structural causal models (SCMs). The second assumption concerns the relationship between **identifiability** and **estimability**. Formal logics of causal inference often take identifiability of a quantity to imply its statistical estimability, then giving identification primary importance. Here estimability means, intuitively, that statistical estimation with finite error guarantees is possible. Maclaren & Nicholson (2019) give a detailed background and analysis of the above assumptions and explain why they are misguided. The present work gives a condensed overview of their article.

1.1. Causal models and statistical frameworks

The first assumption above is closely related to how the term 'model' should be understood in causal inference and statistics. For example, is a model a single probability distribution, a family of distributions, a 'generative mechanism', or a set of structural equations? Or something else? A more general, informal definition of 'model' is simply: 'theoretical construct that implies distributions over observables'. Starting from this perspective, Maclaren & Nicholson (2019) translate a standard DAG/SCM causal inference framework into an abstract statistical framework. In (1), we give a high-level view of this translation, with the left-hand side based on Pearl & Bareinboim (2014), and the right-hand side a further abstracted version of the statistical framework for inverse problems given by Evans & Stark (2002):

$$\begin{aligned} \mathcal{M} &\leftrightarrow \Theta \\ M_1, M_2 \in \mathcal{M} &\leftrightarrow \theta_1, \theta_2 \in \Theta \\ Q(M) &\leftrightarrow q(\theta) \\ P : \mathcal{M} \to \mathcal{P},\ M \mapsto P(M) &\leftrightarrow P : \Theta \to \mathcal{P},\ \theta \mapsto P(\theta). \end{aligned} \tag{1}$$

In the above, structural causal models in the sense of Pearl & Bareinboim (2014), symbolised by M_1, M_2, correspond to abstract models or 'theories' θ_1, θ_2; the causal class of Pearl & Bareinboim (2014), $\mathcal{M}$, corresponds to the abstract model space Θ to which θ_1, θ_2 belong, and causal queries $Q(M)$ correspond to (interest) parameters or 'queries' $q(\theta)$. The function P on the left, which maps any fully-specified structural causal model M to its probability distribution $P(M)$, is translated as the so-called 'forward mapping' P in the abstract framework.

Both **interventional** and **counterfactual** concepts can be expressed as **interest parameters** in the above abstract statistical framework. Importantly, these are defined as functions or functionals on a basic 'model space', *rather than the space of distributions*. This translation is fully compatible with specific causal modelling frameworks like SCMs or DAGs but also expands the scope of causal inference to include model types often neglected in the causal inference literature, for example differential equations, agent-based models, or continuous-time stochastic process models.

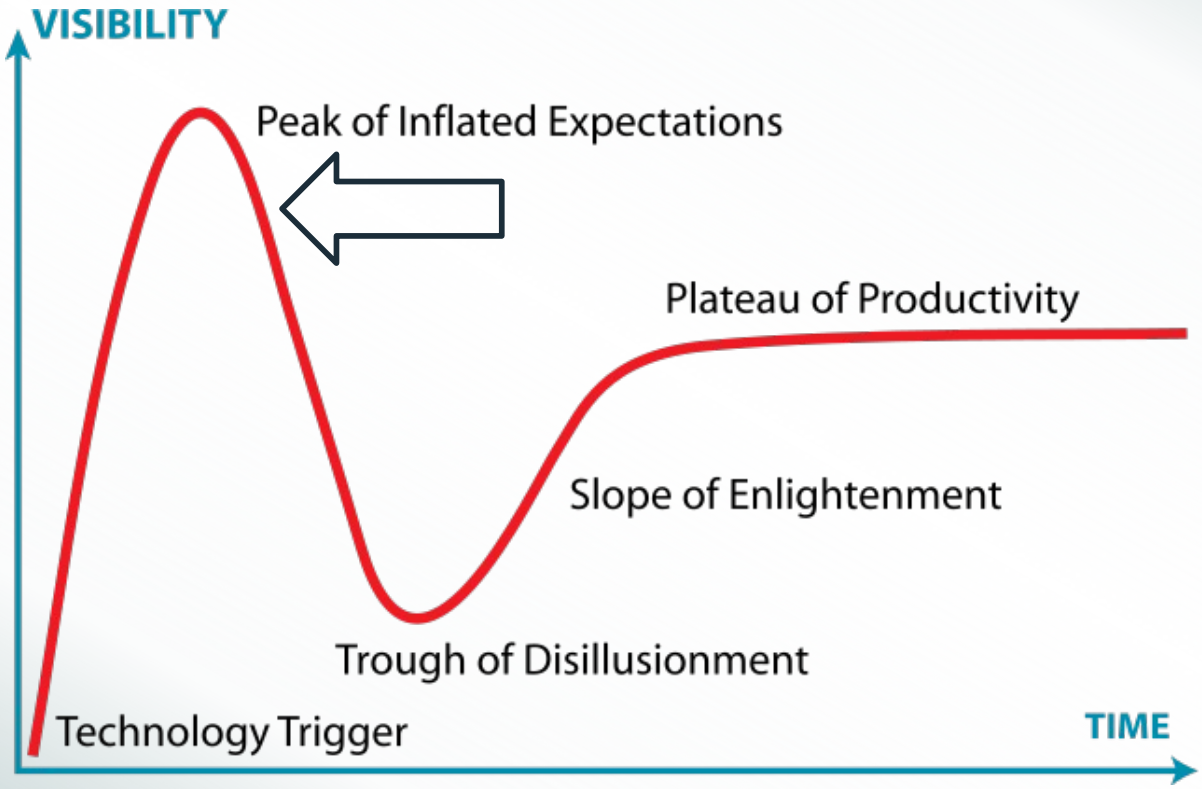
1.2. Identifiability and estimability

The second assumption arises from a common idea in the formal causal inference literature (e.g. Pearl & Bareinboim, 2014, and references therein). This idea is that there is a natural separation of concerns between causal inference and

[1] Department of Engineering Science, The University of Auckland, Auckland, New Zealand. Correspondence to: Oliver J. Maclaren <oliver.maclaren@auckland.ac.nz>.

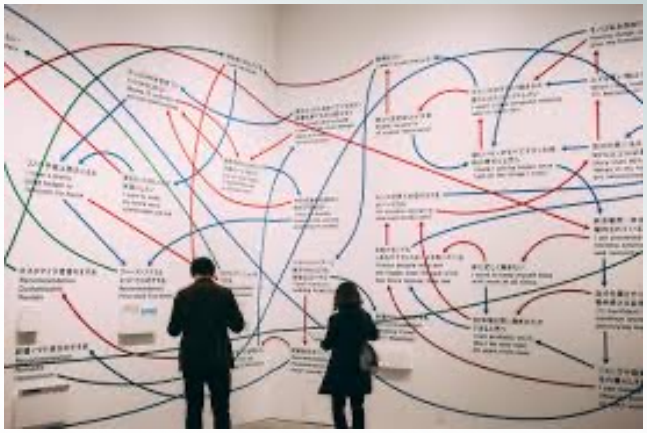
Workshop on the Neglected Assumptions in Causal Inference (NACI) at the 38th International Conference on Machine Learning, 2021

The limitations



Perhaps the hardest part: bringing *ingenuity* to generating the DAG

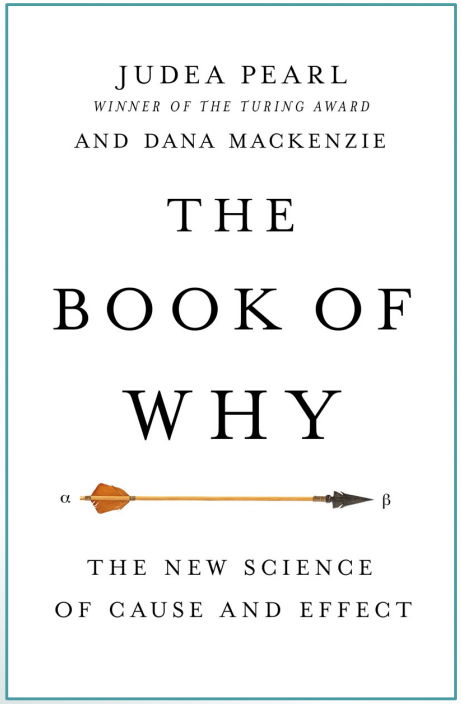
- A DAG is a narrative…
- describing the processes that gave rise to the data
- No infinite regress: for a DAG to be complete, the shared cause of any two variables in the DAG must be included
- requires
  - abstraction
  - lateral and orthogonal thinking
  - collaboration with SME's
  - iteration and revision
  - time, perseverance
  - and ideally, consensus



Writing out DAG means '*sticking your neck out*'.

But positing assumptions so conjectures about implications can be made *is* 'doing science'!

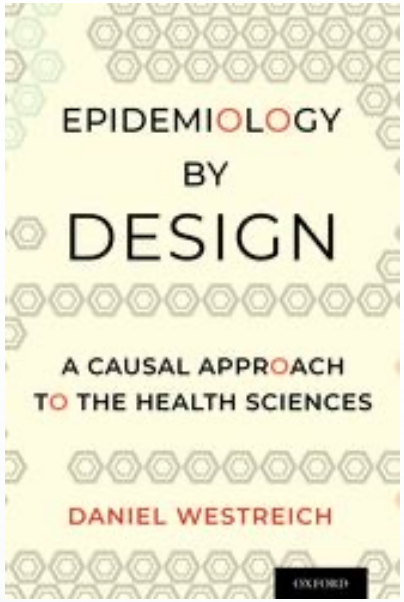
Recommendation



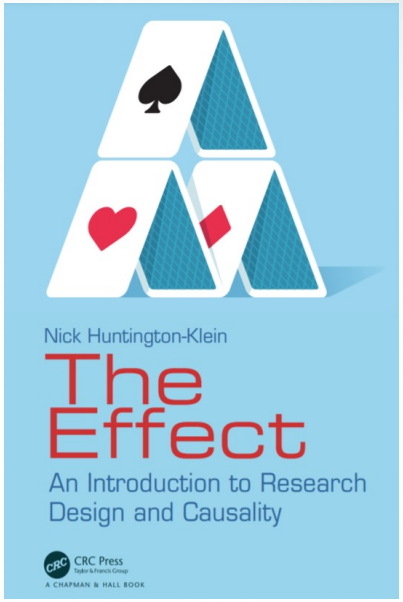
The Book of Why: The New Science of Cause and Effect, by Pearl & Mackenzie, 2018

free YouTube lectures!
Statistical Rethinking 2023

Recommendation

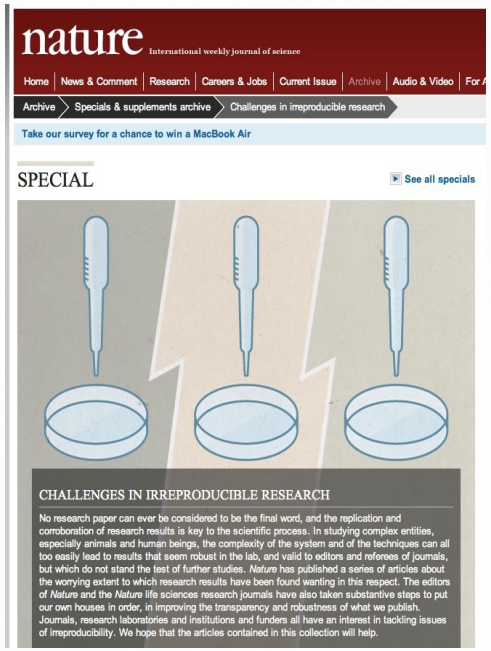


Epidemiology by Design
Daniel Westreich, 2019



The Effect: An Introduction to Research Design and Causality
Nick Huntington-Klein, 2022

The ‘Reprodicibility-’, ‘Replication-’, ‘Statistical-’, ‘… ‘, ‘*Crisis*’



- A ‘crisis’ in Science: research findings often do not replicate on independent data
- How are SCMs and RMS connected to the crisis of scientific “credibility”?

Multiplicity of analysis strategies

- Omitted variables
- Missing data
- Measurement issues
- Information bias

DATA

Likelihood: P(data | θ)

Analytic bias
- Model selection
  - $E(\hat{\beta} \mid \hat{\beta}_{"significant"}) \neq \beta_{true}$
- Model misspecification
- Over-fitting
- Residual confounding
- Arbitrary categorization
- Collider bias

SAMPLE

Conventional statistical methods

- Risk of selection bias
  confounding by indic
- Importance of study /
  experimental design

ANALYSIS

tainties
el specification
el selection
mptions re. distributions

ciation
ausation

gnition/
chology
entions
tivations

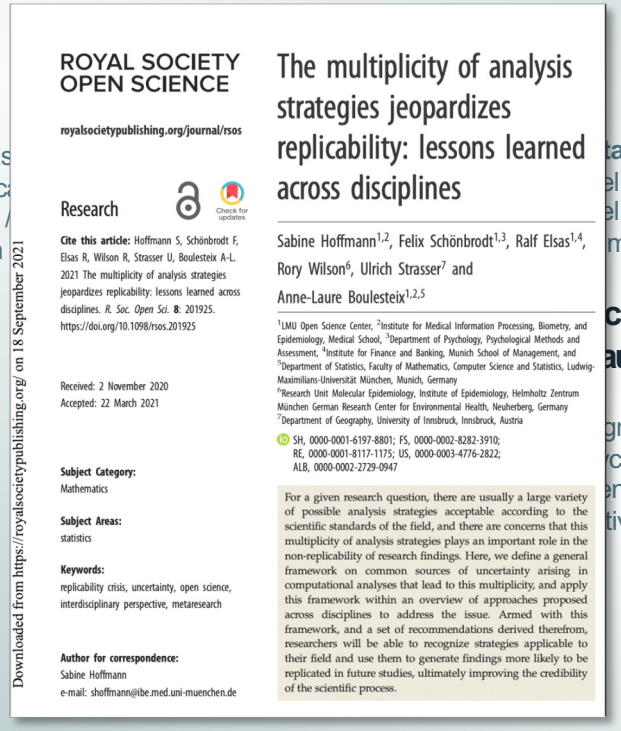
POPULATION

INFERENCE ← **P(θ | data)**

Belief ~ Evidence

DECISIONS & ACTION

NATURE



ROYAL SOCIETY OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

Research

Cite this article: Hoffmann S, Schönbrodt F, Elsas R, Wilson R, Strasser U, Boulesteix A-L. 2021 The multiplicity of analysis strategies jeopardizes replicability: lessons learned across disciplines. *R. Soc. Open Sci.* **8**: 201925. https://doi.org/10.1098/rsos.201925

Received: 2 November 2020
Accepted: 22 March 2021

Subject Category:
Mathematics

Subject Areas:
statistics

Keywords:
replicability crisis, uncertainty, open science, interdisciplinary perspective, metaresearch

Author for correspondence:
Sabine Hoffmann
e-mail: shoffmann@ibe.med.uni-muenchen.de

The multiplicity of analysis strategies jeopardizes replicability: lessons learned across disciplines

Sabine Hoffmann[1,2], Felix Schönbrodt[1,3], Ralf Elsas[1,4], Rory Wilson[6], Ulrich Strasser[7] and Anne-Laure Boulesteix[1,2,5]

[1]LMU Open Science Center, [2]Institute for Medical Information Processing, Biometry, and Epidemiology, Medical School, [3]Department of Psychology, Psychological Methods and Assessment, [4]Institute for Finance and Banking, Munich School of Management, and [5]Department of Statistics, Faculty of Mathematics, Computer Science and Statistics, Ludwig-Maximilians-Universität München, Munich, Germany
[6]Research Unit Molecular Epidemiology, Institute of Epidemiology, Helmholtz Zentrum München German Research Center for Environmental Health, Neuherberg, Germany
[7]Department of Geography, University of Innsbruck, Innsbruck, Austria

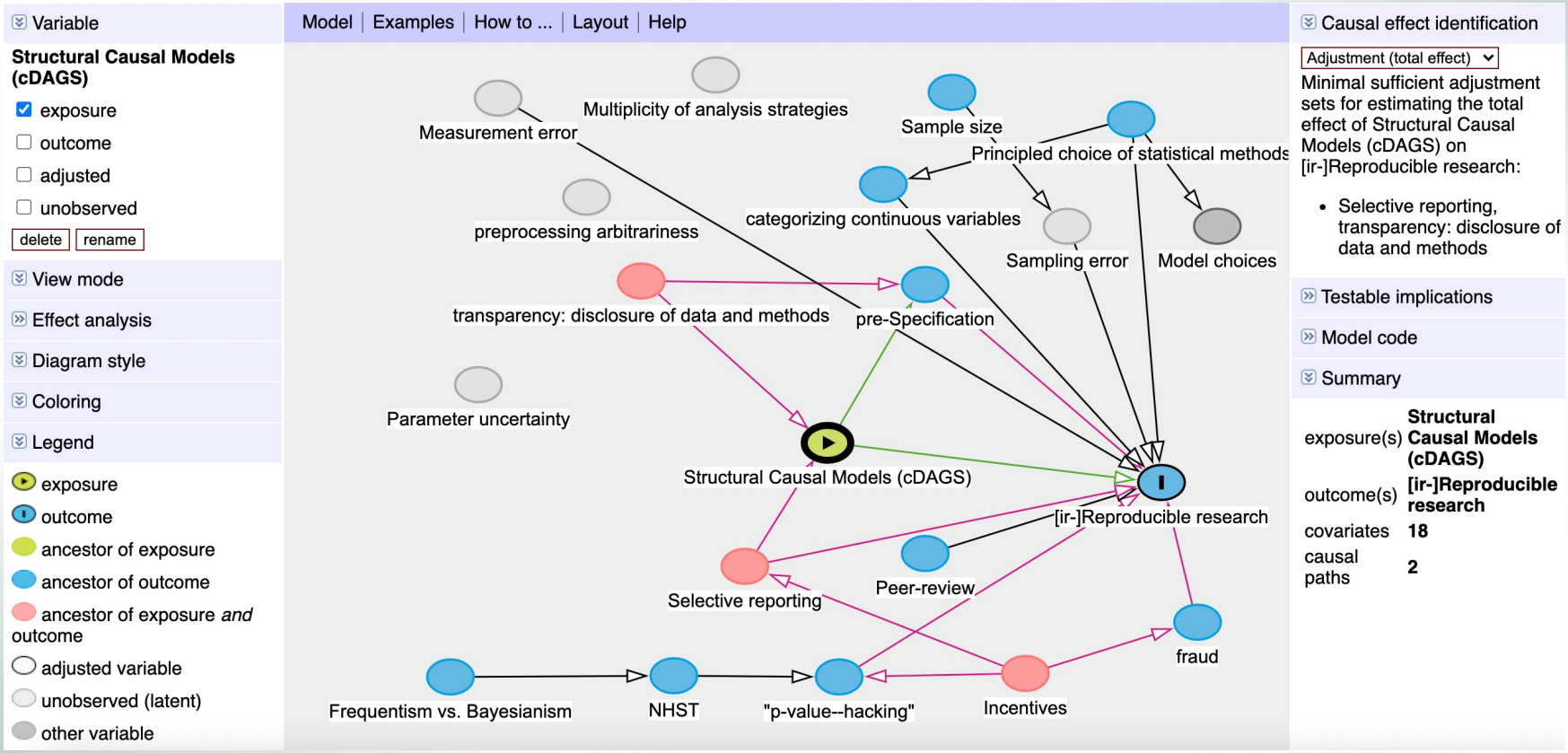
SH, 0000-0001-6197-8801; FS, 0000-0002-8282-3910; RE, 0000-0001-8117-1175; US, 0000-0003-4776-2822; ALB, 0000-0002-2729-0947

For a given research question, there are usually a large variety of possible analysis strategies acceptable according to the scientific standards of the field, and there are concerns that this multiplicity of analysis strategies plays an important role in the non-replicability of research findings. Here, we define a general framework on common sources of uncertainty arising in computational analyses that lead to this multiplicity, and apply this framework within an overview of approaches proposed across disciplines to address the issue. Armed with this framework, and a set of recommendations derived therefrom, researchers will be able to recognize strategies applicable to their field and use them to generate findings more likely to be replicated in future studies, ultimately improving the credibility of the scientific process.

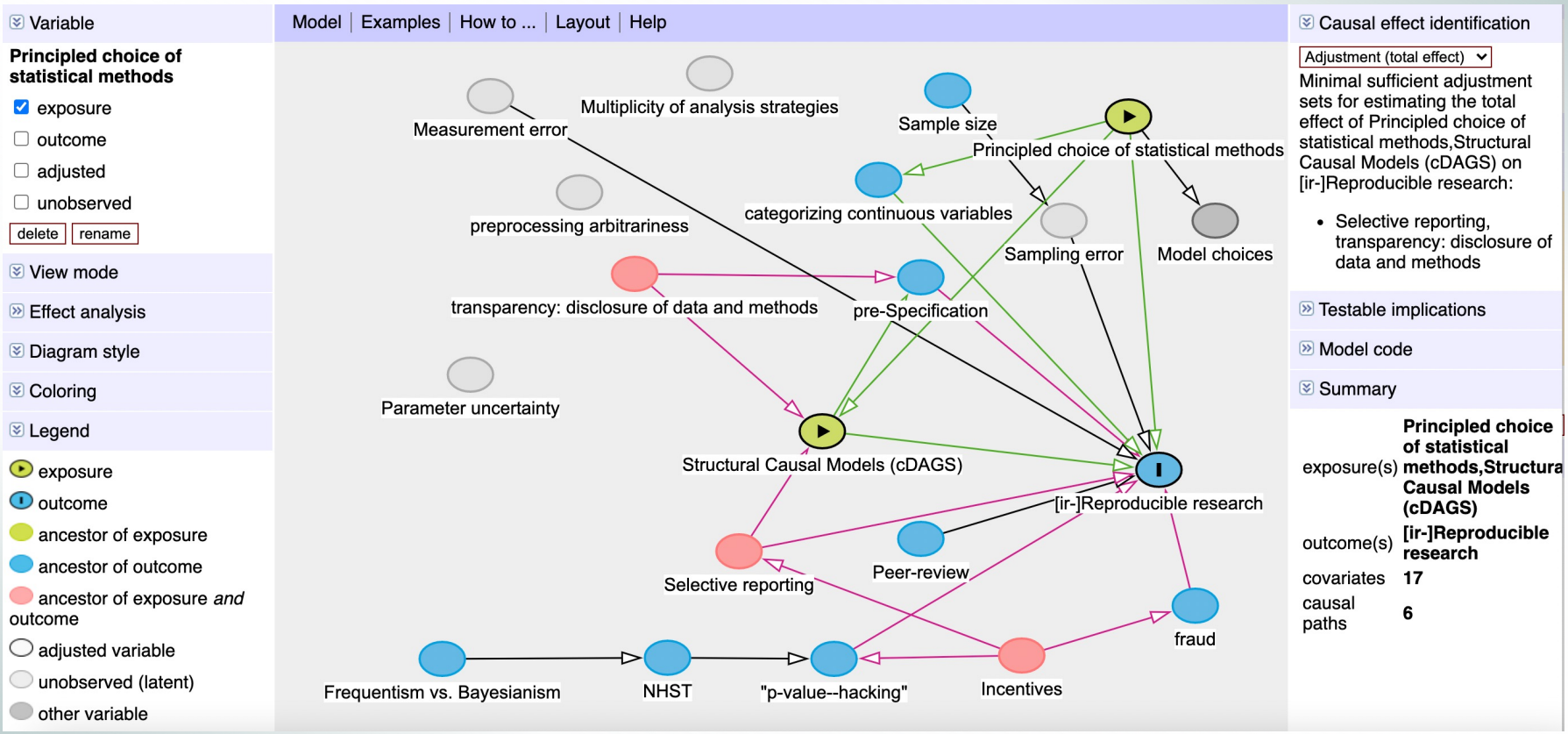
Downloaded from https://royalsocietypublishing.org/ on 18 September 2021

Hoffmann S, *et al*. The multiplicity of analysis strategies jeopardizes replicability: lessons learned across disciplines. R Soc Open Sci. 2021 Apr 21;8(4):201925

The multiplicity of analysis strategies jeopardizes replicability

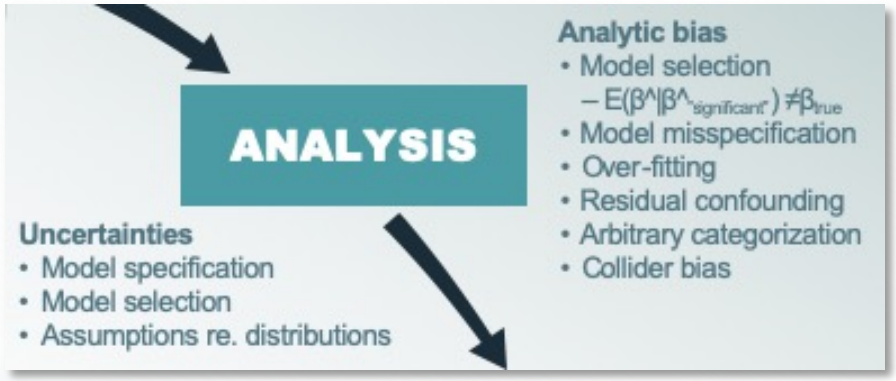


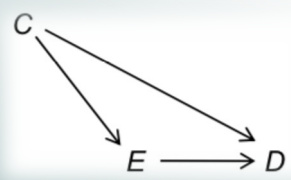
The multiplicity of analysis strategies jeopardizes replicability

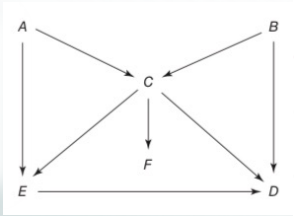


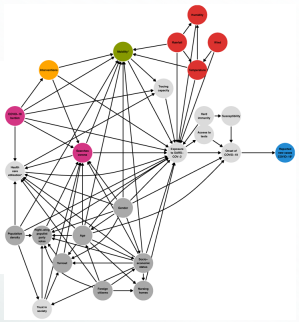
See RMS as "*Principled choice of statistical methods*"

<u>Takeaways</u>: Reasons to consider SCMs in regression modeling strategies for observational studies









SCMs …

1. are a great way of de-bugging your thinking
2. support identification of biases
3. can recommend adjustments necessary for unbiased effect estimation
4. can rationalize model selection
5. can help you spend df's effectively
6. reduce ambiguity in communication
7. support achieving consensus

Explanation vs. Prediction

- Evaluates the validity of using prediction as a proxy for explanation in Bayesian statistical models
  i. a conceptual introduction and overview of the relationship of explanation and prediction as well as their connection to causality;
  ii. large-scale simulations of Bayesian generalized-linear models to study said relationship under various causal and statistical misspecifications;
  iii. initial evidence that causality is indeed the missing link that connects prediction and explanation when comparing statistical models

- *Using prediction as a proxy for explanation is valid and safe <u>only</u> when the considered models are sufficiently consistent with the underlying causal structure of the true data generating process.*

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION
https://doi.org/10.1080/00949655.2024.2449534

Taylor & Francis
Taylor & Francis Group

Check for updates

Prediction can be safely used as a proxy for explanation in causally consistent Bayesian generalized linear models

Maximilian Scholz and Paul-Christian Bürkner

Cluster of Excellence SimTech, University of Stuttgart, Stuttgart, Germany

ABSTRACT
Bayesian modeling provides a principled approach to quantifying uncertainty and has seen a surge of applications in recent years. Within the context of a Bayesian workflow, we are concerned with model selection for the purpose of finding models that best explain the data or underlying data generating process. Since insight into the true process is rare, what remains is incomplete causal knowledge and model predictions of the data. This leads to the important question of when the use of prediction as a proxy for explanation for the purpose of model selection is valid. We approach this question by means of large-scale simulations of Bayesian generalized linear models where we investigate various causal and statistical misspecifications. Our results indicate that the use of prediction as proxy for explanation is valid and safe if the models under consideration are sufficiently consistent with the underlying causal structure of the true data generating process.

ARTICLE HISTORY
Received 19 July 2023
Accepted 17 December 2024

KEYWORDS
Bayesian workflow; causal inference; explanation; prediction; generalized linear models; simulation study

Scholz, M., & Bürkner, P. C. (2025). Prediction can be safely used as a proxy for explanation in causally consistent Bayesian generalized linear models. Journal of Statistical Computation and Simulation, 95(6), 1226–1249. https://doi.org/10.1080/00949655.

Advice


DATA
ANALYSIS
INFERENCE
Belief ~ Evidence
DECISIONS & ACTION
NATURE
POPULATION
SAMPLE

Remember to anchor on the ideal [1]

1. Analysts should be [pro-]actively involved in study design & measurement design!
   - Simulate the design, the analysis, and the expression of results for stakeholders.
     - Simulation is especially useful for
       - sample size estimation,
       - setting realistic expectations about precision/uncertainty in results
         - exposing *futility*
       - exposing sources of uncertainty in the evidence generating process.

Remember to anchor on the ideal [2]

2. Receive data from a well-designed `experiment`, with optimal measurement, either restricting or blocking on important &/or relevant sources of variability
   - *Count your blessings!*
   - Treatment assignment / exposure has no association with any other independent variables
     - ‘The “unreasonable effectiveness” of Randomization in Natural Sciences’
   - Adjust for efficiency / precision in estimation
   - Follow principles and examples in RMS, and use RMS tools

Remember to anchor on the ideal [3]

3. [“Degenerate situation”] Receive observational data (including SDA of RCTs)

- use DAGs to expose and summarize your assumptions about the relevant system for the estimation
- identify the variables that must be measured and controlled to obtain unconfounded effect estimates given those assumptions
- use Daggity, until you get good at parsing paths by eye
- simulate the DGP, and confirm that your analysis methods can recover the posited estimate to everyone’s satisfaction
- simulate the design, the analysis, and the expression of results for stakeholders in advance of analysis

[“Degenerate situation”] “External comparator”

Statistical Thinking Frank Harrell

About Posts Talks Courses Datamethods News Links Bio Publications

Contents

Incorporating Historical Control Data Into an RCT

Code

DRUG-EVALUATION BAYES DESIGN DRUG-DEVELOPMENT INFERENCE OBSERVATIONAL POSTERIOR PRIOR 2023

Historical data (HD) are being used increasingly in Bayesian analyses when it is difficult to randomize enough patients to study effectiveness of a treatment. Such analyses summarize observational studies' posterior effectiveness distribution (for two-arm HD) or standard-of-care outcome distribution (for one-arm HD) then turn that into a prior distribution for an RCT. The prior distribution is then flattened somewhat to discount the HD. Since Bayesian modeling makes it easy to fit multiple models at once, incorporation of the raw HD into the RCT analysis and discounting HD by explicitly modeling bias is perhaps a more direct approach than lowering the effective sample size of HD. Trust the HD sample size but not what the HD is estimating, and realize several benefits from using raw HD in the RCT analysis instead of relying on HD summaries that may hide uncertainties.

AUTHOR
Frank Harrell

AFFILIATION
Department of Biostatistics
Vanderbilt University School of Medicine

PUBLISHED
November 4, 2023

MODIFIED
May 4, 2024

EHRs and RCTs: Outcome Prediction vs. Optimal Treatment Selection

Some ‘exotic’ situations and solutions

Propensity Score Adjustment:
- Covariate
- Matching
- Weighting
- Stratification

In BBR, see
Misunderstandings About Propensity Scores
Reasons for Failure of Propensity Analysis

Some ‘exotic’ situations and solutions

- Front-Door Criterion: use mediators when confounders are unmeasured
- Instrument affects treatment
  - Independent of outcome except through treatment
  - Not associated with confounders
  - Examples: policy changes, random assignment
- Marginal Structural Models (MSMs) with IPTW
- Sensitivity Analysis & Negative Controls

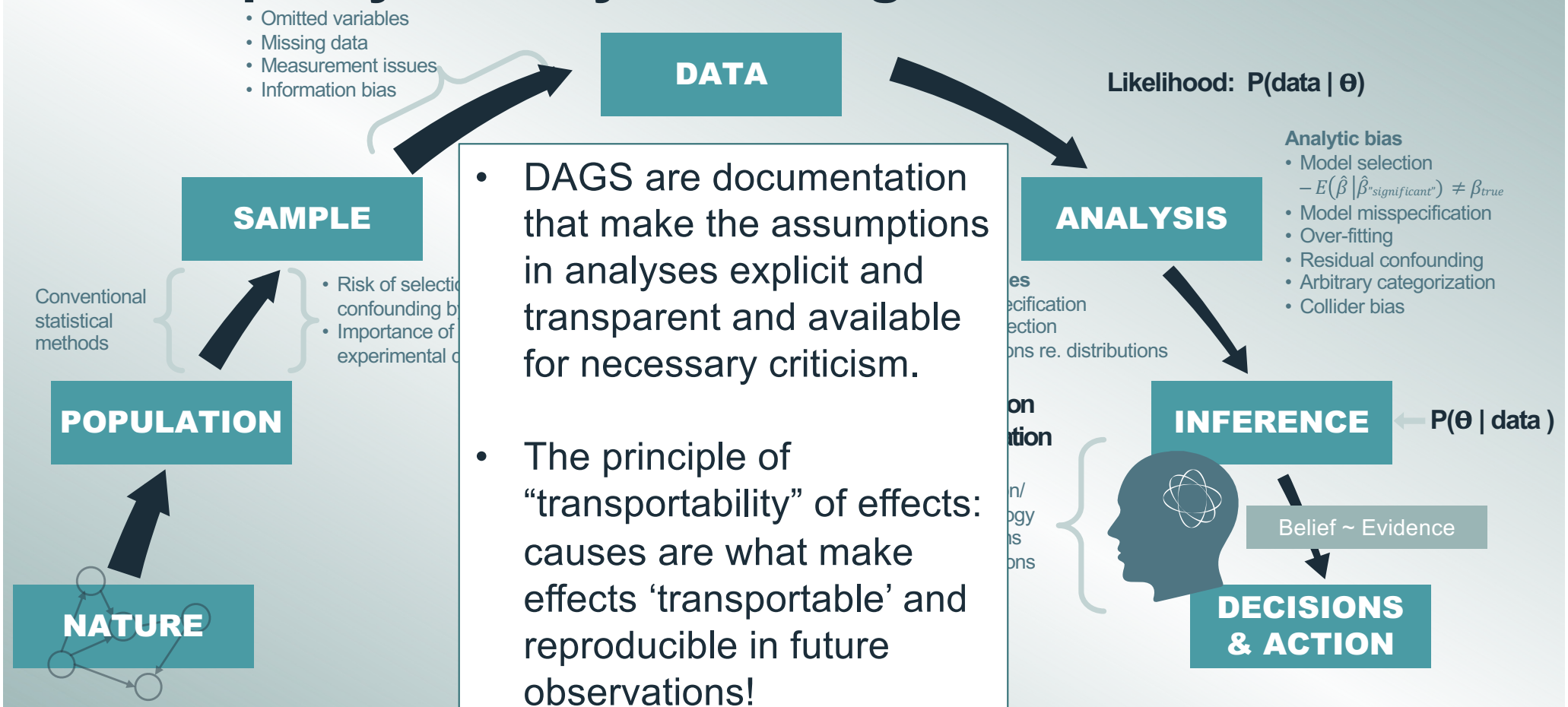
Some ‘exotic’ situations and solutions

- G-Methods
  - G-Formula (Parametric G-Computation)
  - G-Estimation of Structural Nested Models

- Machine Learning + Causal Inference
  - Targeted Maximum Likelihood Estimation (TMLE)
  - Double Machine Learning (DML)
  - Causal Forests and HTE estimation

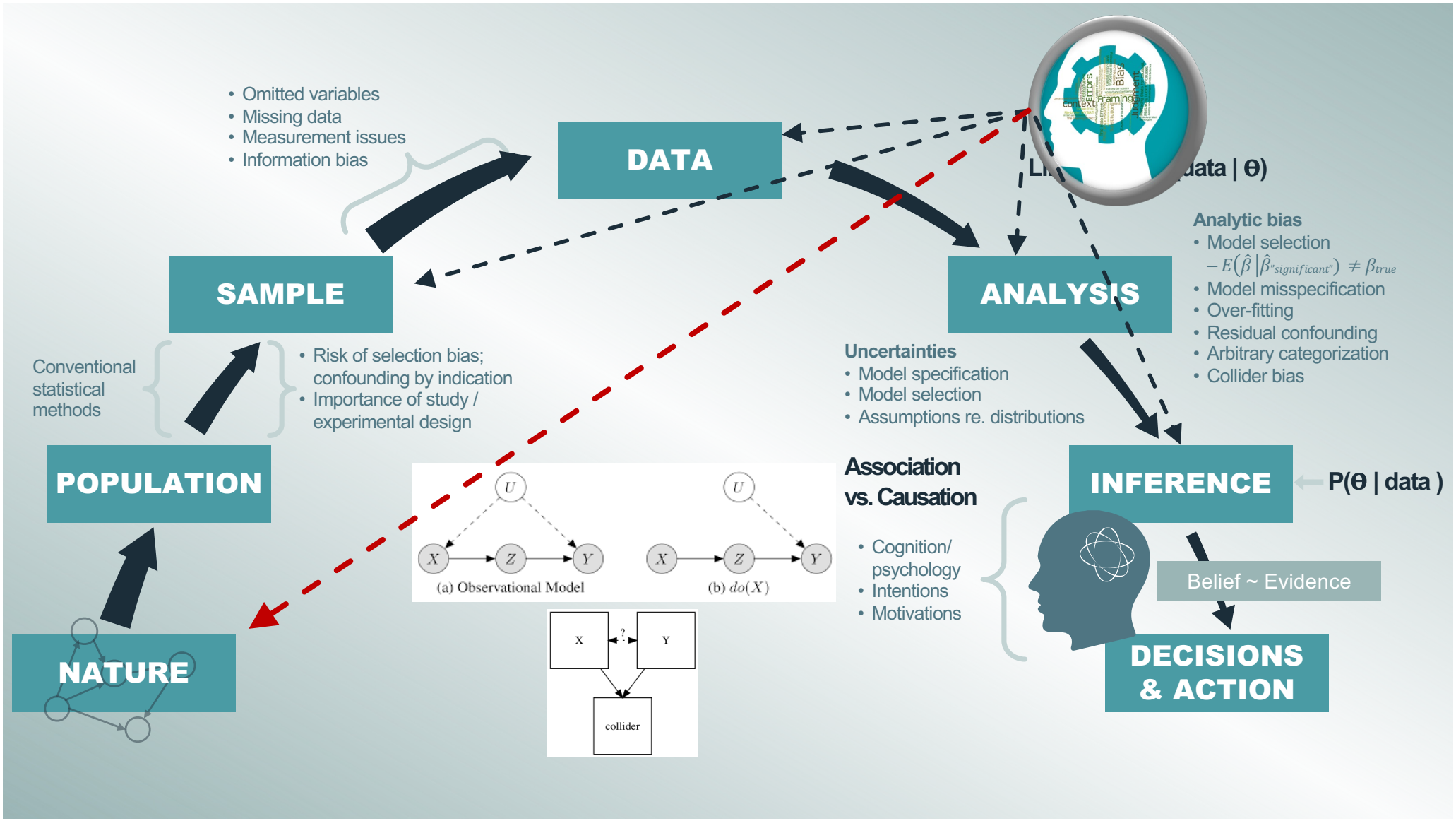
Keep your standards up!!

Multiplicity of analysis strategies



- DAGS are documentation that make the assumptions in analyses explicit and transparent and available for necessary criticism.
- The principle of "transportability" of effects: causes are what make effects 'transportable' and reproducible in future observations!

Hoffmann S, *et al*. The multiplicity of analysis strategies jeopardizes replicability: lessons learned across disciplines. R Soc Open Sci. 2021 Apr 21;8(4):201925


DATA
• Omitted variables
• Missing data
• Measurement issues
• Information bias
SAMPLE
Conventional statistical methods
• Risk of selection bias; confounding by indication
• Importance of study / experimental design
POPULATION
NATURE
ANALYSIS
Uncertainties
• Model specification
• Model selection
• Assumptions re. distributions
Analytic bias
• Model selection
– $E(\hat{\beta} \mid \hat{\beta}_{"significant"}) \neq \beta_{true}$
• Model misspecification
• Over-fitting
• Residual confounding
• Arbitrary categorization
• Collider bias
INFERENCE
P(θ | data)
Belief ~ Evidence
DECISIONS & ACTION
Association vs. Causation
• Cognition/ psychology
• Intentions
• Motivations
(a) Observational Model
(b) do(X)
X
Y
collider

Thank you

Any questions?