Causal Models and Learning from Data

PhD Computer Science - University of Milano-Bicocca
Causal Networks 2021 Exam
Georgios Peikos, Gian Carlo Milanese, Paolo Tenti

Assigned Paper

Causal Models and Learning from Data:

Integrating Causal Modeling and Statistical Estimation

Maya L. Petersen and **Mark J. van der Laan**
Divisions of Biostatistics and Epidemiology, University of California, Berkeley, School of Public Health, Berkeley, CA

Published in final edited form as:
Epidemiology. 2014 May ; 25(3): 418–426. doi:10.1097/EDE.0000000000000078.

Introduction

Context: **Causal Modeling in Epidemiology**

The authors argue that:

- The practice of epidemiology requires asking causal questions, to understand:
    - **why patterns of disease and exposure do exist.**
    - **how one can intervene to change them.**

Introduction

A **formal causal framework** can help in:

- **framing** sharper **scientific questions** and **making** transparent the **assumptions** required to answer them.
- **distinguishing** the process of **causal inference** from the process of **statistical estimation**.

Approach

To this aim, the authors **introduce a systematic approach** to answer causal questions, that includes:

- Specification of :
  - a causal model
  - the observed data
  - the target causal quantity

- Assessment of identifiability
- Commitment to a statistical model and estimand
- Statistical estimation
- Interpretation

1. Specification of a Causal Model

Represent knowledge about the system to be studied using a causal model

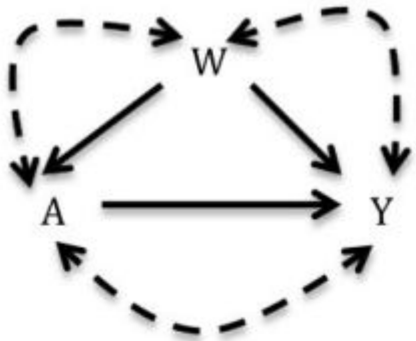
In this paper, the authors focus on the **structural causal model (SCM)**.

Every SCM implies an associated Causal Graph.

- Here, directed **acyclic graphs**.

In which:

- *W*: baseline covariates, for instance age.
- *A*: exposure / treatment
- *Y*: outcome

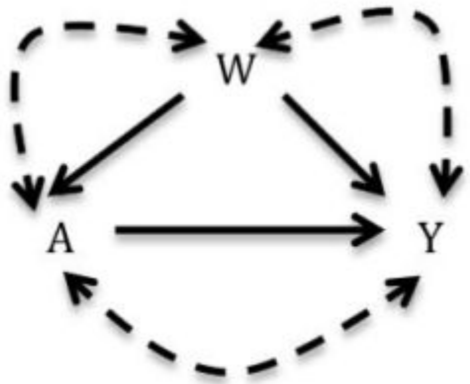


Represent knowledge about the system to be studied using a causal model

To determine the value of a variable, we consider:

- a set of **unmeasured background factors U_W**, together with the **variable's parents**.

Therefore, **graphs encode knowledge** about the possible causal relations among variables.



$$W = f_W(U_W)$$
$$A = f_A(W, U_A)$$
$$Y = f_Y(W, A, U_Y)$$

Set of structural equations.

Represent knowledge about the system to be studied using a causal model

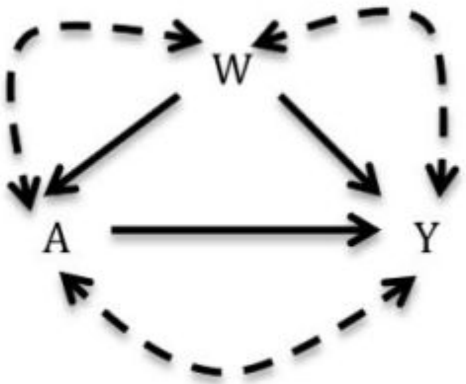
To determine the value of a variable, we consider:

- a set of **unmeasured background factors U_W**, together with the **variable's parents**.

Therefore, **graphs encode knowledge** about the possible causal relations among variables.

Also, **omission** of a double-headed arrow between two variables assumes the variables do not share **an unmeasured cause (background factors)**.

"Independence assumption"

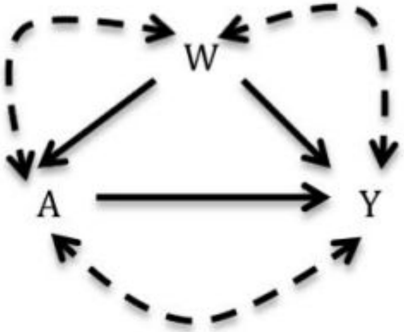


$$W = f_W(U_W)$$
$$A = f_A(W, U_A)$$
$$Y = f_Y(W, A, U_Y)$$

Set of structural equations.

Represent knowledge about the system to be studied using a causal model

The **set of structural equations**, together with any **restrictions** placed on the **joint distribution** of the **error terms U**, constitute a structural causal model.



$$W = f_W(U_W)$$
$$A = f_A(W, U_A)$$
$$Y = f_Y(W, A, U_Y)$$

No restrictions on the distribution of $U = (U_W, U_A, U_Y)$

Set of structural equations.

Represent knowledge about the system to be studied using a causal model

The authors argue that:

- the flexibility of a structural causal model allows us to **avoid assumptions** that are not **supported**.

2. Specification of the observed data and their link to the causal model

Specify how data are linked to the causal model

This involves specifying:

- **what variables** have been (or will be) **measured**.
- **how these variables are generated** by the system described by the causal model.

This provides a **bridge between causal modeling and statistical estimation.**

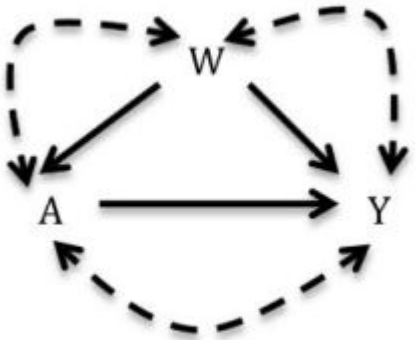
Specify how data are linked to the causal model

Selection and sampling can also be incorporated directly into the causal model.

- For example, a study may have measured (W, A,Y) on an independent random sample of n individuals from some target population.

- Or, the study participants may have been sampled on the basis of **exposure** or **outcome** status.

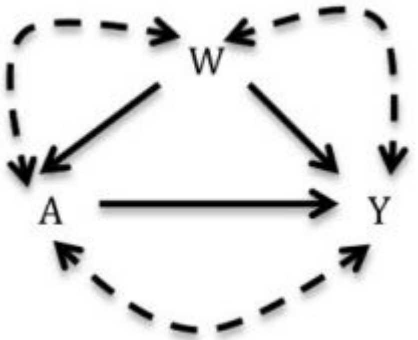
Specify how data are linked to the causal model

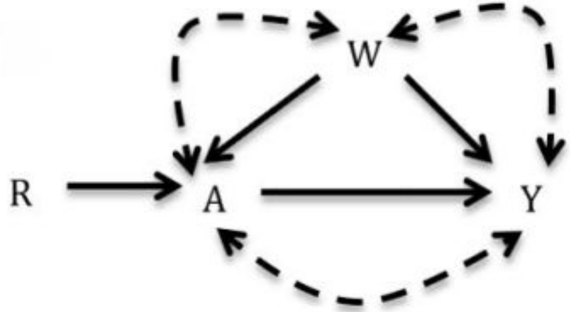


This SCM can generate any possible distribution O = (W, A, Y).

So, it places **no restrictions** on the joint distribution of the observed data, implying a **nonparametric statistical model**.

Specify how data are linked to the causal model





This SCM can generate any possible distribution O = (W, A, Y).

So, it places **no restrictions** on the joint distribution of the observed data, implying a **nonparametric statistical model**.

This SCM, that assumes that **R is independent from W**, can generate only distributions O = (W, R, A, Y).

This causal model **implies a semiparametric statistical model**.

3. Specification of the Target Causal Quantity

Specify the target causal quantity

Translation of the scientific question into a formal causal quantity, defined as some parameter of the counterfactual distribution of data under some ideal intervention.

The following decisions are involved in this step:

1. **Which variables to intervene on** (single variable, multiple variables, …)

2. **How to set the values** of intervention variables: deterministically (on all population), dynamically (based on individual characteristics), stochastically

3. What **summary** of the counterfactual outcome distributions is of interest

4. What **population** is of interest: whole population, a subset of the population, a different population

Specify the target causal quantity

For example, a common counterfactual quantity of interest is the **average treatment effect**, defined as

> *the difference in mean outcome that would have been observed had all members of a population received versus not received some treatment*

With Y_a denoting the counterfactual outcome under an intervention to set *A* = *a*, this quantity is expressed as

$$E(Y_1 - Y_0)$$

4. Assessment of Identifiability

Assessment of identifiability

Previous step: translation of the scientific question into a **parameter of the unobserved counterfactual distribution** of the data under some **ideal intervention**.

This step: understand whether the target quantity can be expressed as a **parameter of the distribution of the observed data alone** (an estimand), given the causal model and its link to the observed data. I.e., **identifiability**

Assessment of identifiability

Example: choice of an adjustment set when estimating the average treatment effect (ATE).

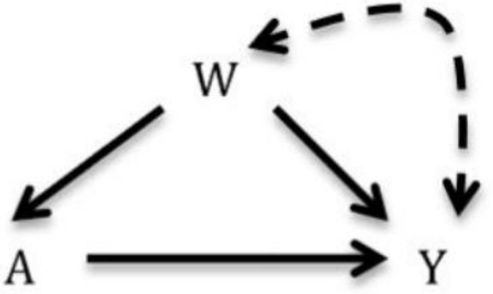
If the pre-intervention covariates *W* block all unblocked backdoor paths from *A* to *Y* (**backdoor criterion**), then the counterfactual quantity

$$P(Y_a{=}y)$$

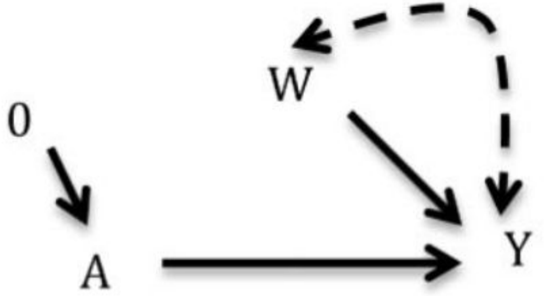
can be identified with the estimand

$$\sum_w P(Y{=}y|A{=}a,\, W{=}w)P(W{=}w)$$

which can be computed from the data alone.



Causal model



Counterfactual distribution under ideal intervention

5. Commitment to a Statistical Model and Estimand

State the statistical estimation problem

Specify the estimand and the statistical model. If knowledge is sufficient to identify the causal effect of interest: commit to the estimand.

In many cases, available knowledge and data are **insufficient to claim identifiability.** Possible steps:

- Understand if further research and data collection can help
- Make further assumptions in order to obtain identifiability, if "current best" answers are needed

State the statistical estimation problem

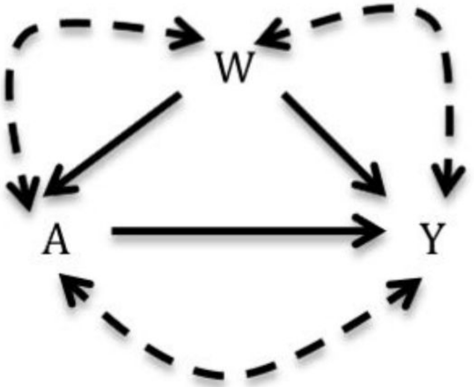
In the latter case, the authors distinguish between two kinds of assumptions:

- **Knowledge-based assumptions**, which represent real knowledge
- **Convenience-based assumptions**, which *do not represent real knowledge*, but which, *if true, would result in identifiability*.

An estimation problem for "current best answers" consists in:

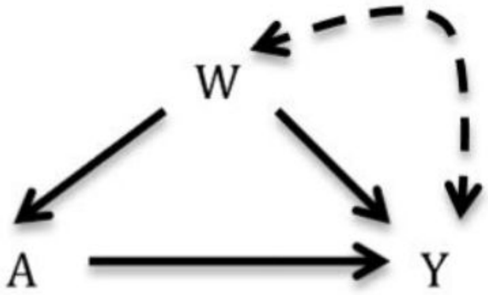
1. A statistical model implied by knowledge-based assumptions
2. An estimand that is equivalent to the target causal quantity under a minimum of convenience-based assumptions
3. A clear differentiation between convenience-based assumptions and real knowledge.

State the statistical estimation problem



Model based on real knowledge

The causal effect of *A* on *Y*
can not be identified

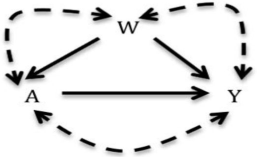


Model based on convenience assumptions

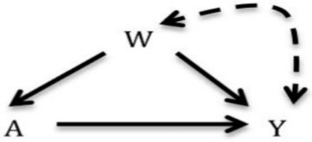
The causal effect of *A* on *Y*
can be identified

6. Statistical Estimation

Statistical Estimation



Want to estimate the causal effect of *A* on *Y*
Knowledge is captured by a SCM
The analyst would like to use ATE



Under some convenience-based assumptions, we have identifiability, and the (1) holds

$$P(Y_a{=}y){=}\sum_w P(Y{=}y|A{=}a,\ W{=}w)P(W{=}w) \quad (1)$$

$$\sum_w (E(Y|A{=}1,\ W{=}w){-}E(Y|A{=}0,\ W{=}w))P(W{=}w) \quad (2)$$

(2) can be used as an estimand to evaluate ATE
We need an estimator to obtain an estimation for ATE!

Statistical Estimation

There is nothing causal about the resulting estimation problem,

Estimation itself is a purely statistical problem

- The analyst is free to choose among several estimators
- e.g. regression of Y (outcome) on A (exposure), followed by averaging with respect to the empirical distribution of W (covariates)

Any estimator itself requires, as "ingredients," estimators of specific components of the observed data distribution

- *The true structural formula that generate the distribution is unknown*

$$\sum_w \left(\boxed{E(Y|A=1, W=w)} - \boxed{E(Y|A=0, W=w)} \right) \boxed{P(W=w)} \quad (2)$$

Estimators have important differences in their statistical properties, which can result in meaningful differences in performance

7. Interpretation

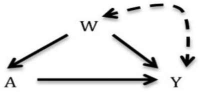
Interpretation

Estimate is given for (2)

$$\sum_{w}(E(Y|A=1, W=w)-E(Y|A=0, W=w))P(W=w) \quad (2)$$

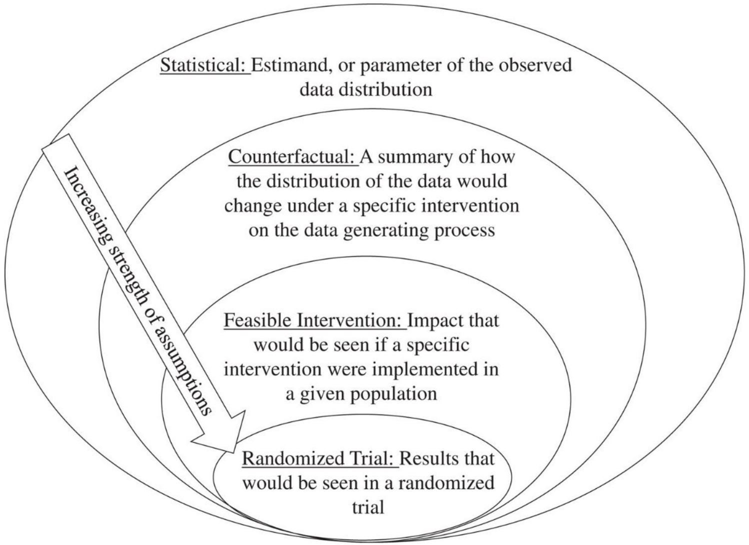
How can we interpret (2)?

As a purely statistical quantity…



As a causal quantity (i.e., ATE) under certain convenience-based assumptions that are explicit in the SCM (identifiability)

Interpretation can be further expanded by considering stronger assumptions by the investigator, that concerns conceivable and well-defined intervention in the real world



Interpretation

The decision of how far to move along this hierarchy can be made by the analyst based on the specific application at hand.

The assumptions required are explicit and, when expressed using a causal graph, readily understandable by subject matter experts.

The debate continues as to whether causal questions and assumptions should be restricted to quantities that can be tested and thereby refuted via theoretical experiment.

Conclusions

Conclusions

Epidemiologists continue to debate whether and how to integrate formal causal thinking into applied research.

Like any tool, the benefits of a causal inference framework depend on how it is used. Good epidemiologic practice requires to:

- *Learn about how data are generated*
- *Be clear about the question to be addressed*
- *Design an analysis that answers this question using the available data*
- *Avoid or minimize assumptions not supported by knowledge*
- *Be transparent and skeptical when interpreting results*

A formal causal framework, when used appropriately, provides an invaluable tool for integrating the following principles into applied epidemiology.