Causal Inference & Paradoxes

Alexandre Aussem

LIRIS UMR 5205 CNRS
Data Mining & Machine Learning (DM2L) Group
Université Claude Bernard Lyon 1

perso.univ-lyon1.fr/alexandre.aussem
alexandre.aussem@liris.cnrs.fr

Outline of the talk

- Bayesian networks: From probability to causality
- Manipulation theorem to estimate the **effect of external interventions**
- **Confounding**: fundamental impediments to the elucidation of causal inferences from observational data
- Elucidation of some well-known controversies :
  - The selection bias or **Berkson's paradox** (1946),
  - The **birth-weight paradox** (1967)
  - The **Simpson's paradox** (1899)
  - The old debate on the relation between **smoking and lung cancer** (1964),
  - **Sex discrimination**: The « reverse regression controversy » between sex and salary which occupied the social science in the 1970s
- Rules of « **do calculus** »
- Case study: **effect of the pesticides on agricultural yields**
- Unbiased estimates despite **selection bias** and **missing data**

Cause-effect relationships

- The central aim of many studies in the physical, behavioral, social, and biological sciences is the **elucidation of cause-effect relationships** among variables or events, e.g., risk factor exposure on disease occurrence, advertising campaign on benefits, treatment on recovery rate, etc.

- However, the appropriate **methodology for extracting such relationships** from data has been fiercely debated.

- **Graphical models** provide clear semantics for causal claims, and non-trivial causal phenomena, **paradoxes and controversies** in causal analysis that long were regarded as **metaphysical** can now be understood, exemplified, analyzed and solved using **elementary mathematics**.

- Most of the material presented here is borrowed from **Judea Pearl'**s books and papers.

Probabilities…

- **Probabilities** play a central role in machine learning.
- Probability theory can be expressed in terms of two simple equations corresponding to the **sum rule** and the **product rule**.

$$p_X(x) = \sum_{y \in \mathcal{Y}} p_{XY}(x, y).$$

$$p(x|y) = \frac{p(x, y)}{p(y)}.$$

- All of the probabilistic **inference** and **learning** manipulations amount to repeated application of these two equations.

Conditional Independence

Two random variables $\mathbf{X}$ *and* $\mathbf{Y}$ *are independent given a third random variable* $\mathbf{Z}$*, denoted* $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$*, when the following holds for all* $(\mathbf{x}, \mathbf{y}, \mathbf{z}) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$*:*

$$p(\mathbf{x}, \mathbf{y}, \mathbf{z})p(\mathbf{z}) = p(\mathbf{x}, \mathbf{z})p(\mathbf{y}, \mathbf{z}) \ .$$

Any probabilistic independence model satisfies the **semi-graphoid axioms**, so they can be combined to form new independence statements.

- Symmetry: $\langle \mathbf{X}, \mathbf{Y} \mid \mathbf{Z} \rangle \iff \langle \mathbf{Y}, \mathbf{X} \mid \mathbf{Z} \rangle$.
- Decomposition: $\langle \mathbf{X}, \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z} \rangle \implies \langle \mathbf{X}, \mathbf{Y} \mid \mathbf{Z} \rangle$.
- Weak Union: $\langle \mathbf{X}, \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z} \rangle \implies \langle \mathbf{X}, \mathbf{Y} \mid \mathbf{Z} \cup \mathbf{W} \rangle$.
- Contraction: $\langle \mathbf{X}, \mathbf{Y} \mid \mathbf{Z} \rangle \wedge \langle \mathbf{X}, \mathbf{W} \mid \mathbf{Z} \cup \mathbf{Y} \rangle \implies \langle \mathbf{X}, \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z} \rangle$.

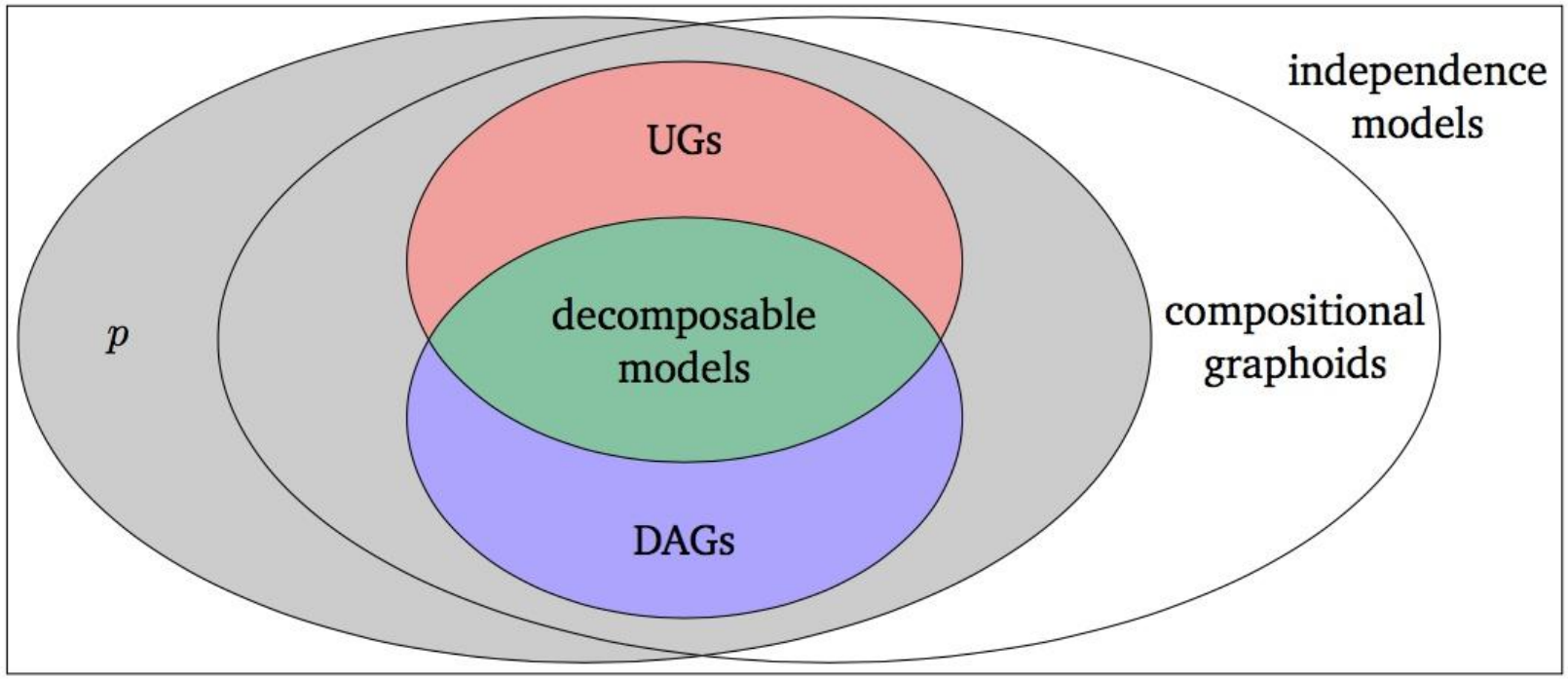
Introduction to Graphical Models

- It is advantageous to augment the analysis using **diagrammatic representations of probability** distributions, called ***probabilistic graphical models***. These offer several useful properties:

  - Simple way to **visualize** the structure of a probabilistic model.
  - Insights into the **conditional independence properties,** can be obtained by inspection of the graph.
  - Complex computations, required to perform **inference** and **learning** in can be expressed in terms of **graphical manipulations**.

- ***Bayesian networks***, also known as ***directed graphical models***, are a major class of graphical models in which the links have directional significance.

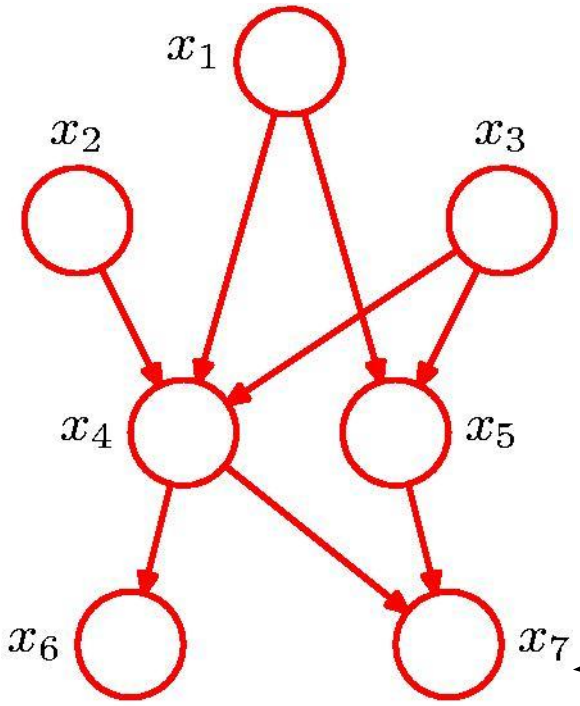
Independence models



Overlapping between probabilistic independence models (p), independence models based on u-separation (UG-faithful), and d-separation (DAG-faithful).

Bayesian Networks

$$p(x_1,\ldots,x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1,x_2,x_3)$$
$$p(x_5|x_1,x_3)p(x_6|x_4)p(x_7|x_4,x_5)$$



General Factorization : $p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k)$

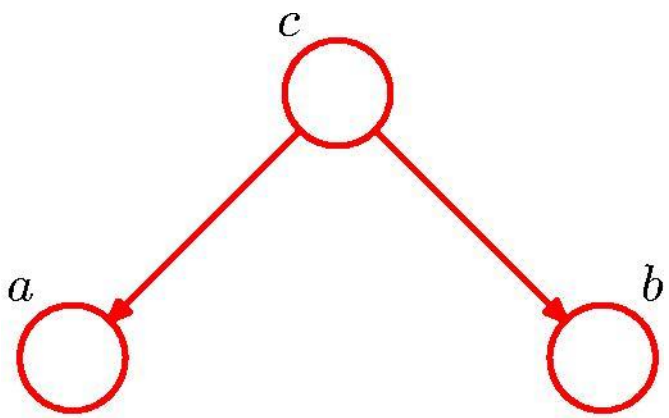
Conditional distribution for (binary) node x_7 :

x_4	x_5	1	0
0	0	0.4	0.6
0	1	0.1	0.9
1	0	0.7	0.3
1	1	0.6	0.4

Corollary (**Markov condition)** : every node given its parents is independent on its non-descendants nodes. Other independencies are entailed (**d-separation** criterion).

Illustration from **Christopher Bishop's** book : "*Pattern recognition* and *machine learning*".

Conditional Independence: Example 1

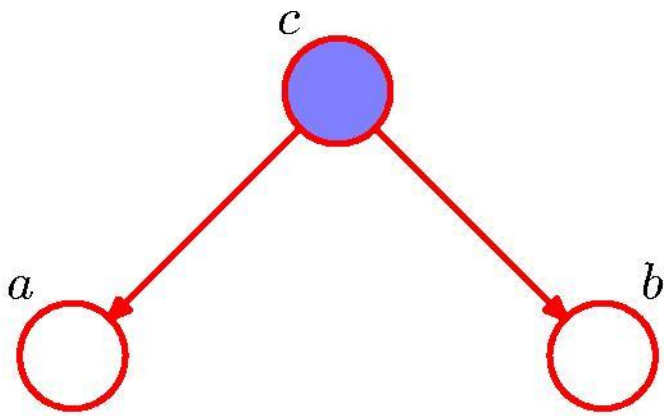


$$p(a, b, c) = p(a|c)p(b|c)p(c)$$

$$p(a, b) = \sum_c p(a|c)p(b|c)p(c)$$

$$a \not\perp\!\!\!\perp b \mid \emptyset$$

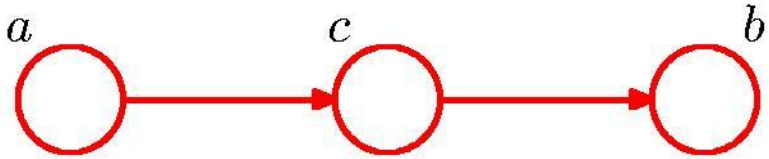
Conditional Independence: Example 1



$$p(a,b|c) = \frac{p(a,b,c)}{p(c)}$$
$$= p(a|c)p(b|c)$$

$$a \perp\!\!\!\perp b \mid c$$

Conditional Independence: Example 2

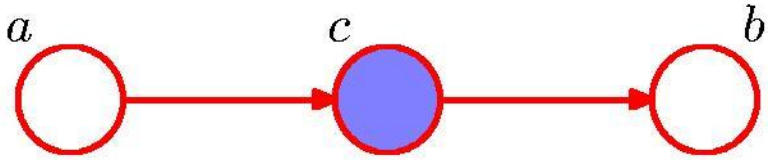


$$p(a, b, c) = p(a)p(c|a)p(b|c)$$

$$p(a, b) = p(a) \sum_c p(c|a)p(b|c) = p(a)p(b|a)$$

$$a \not\perp\!\!\!\perp b \mid \emptyset$$

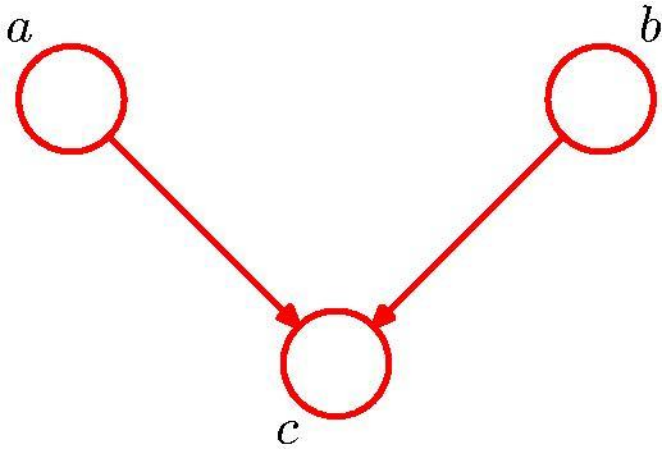
Conditional Independence: Example 2



$$
\begin{aligned}
p(a,b|c) &= \frac{p(a,b,c)}{p(c)} \\
&= \frac{p(a)p(c|a)p(b|c)}{p(c)} \\
&= p(a|c)p(b|c)
\end{aligned}
$$

$$a \perp\!\!\!\perp b \mid c$$

Conditional Independence: Example 3



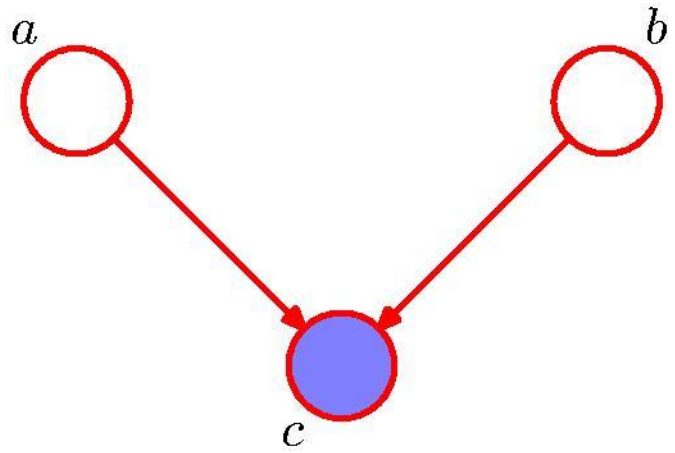
$$p(a, b, c) = p(a)p(b)p(c|a, b)$$

$$p(a, b) = p(a)p(b)$$

$$a \perp\!\!\!\perp b \mid \emptyset$$

Note: this is the opposite of Example 1, with c unobserved.

Conditional Independence: Example 3



$$p(a,b|c) = \frac{p(a,b,c)}{p(c)}$$

$$= \frac{p(a)p(b)p(c|a,b)}{p(c)}$$

$$a \not\perp b \mid c$$

Compared to the previous examples, the opposite is observed:
Two **independent variables become dependent** given a third variable!

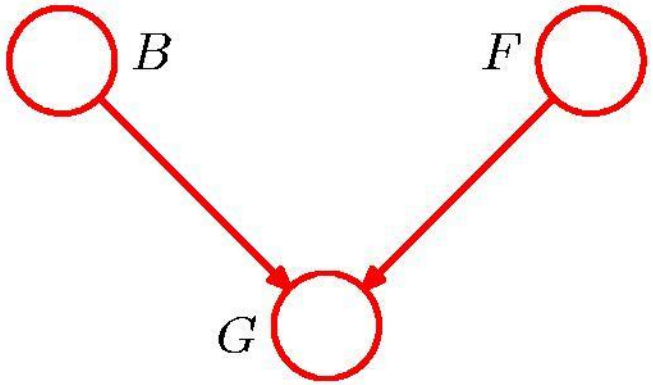
“Am I out of fuel?”

$$p(G=1|B=1,F=1) = 0.8$$
$$p(G=1|B=1,F=0) = 0.2$$
$$p(G=1|B=0,F=1) = 0.2$$
$$p(G=1|B=0,F=0) = 0.1$$

$$p(B=1) = 0.9$$
$$p(F=1) = 0.9$$

and hence

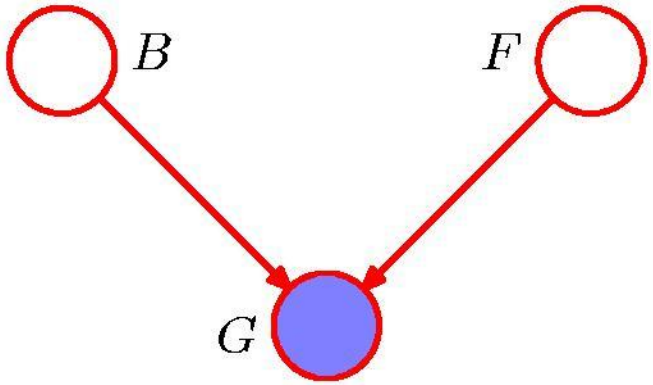
$$p(F=0) = 0.1$$



B = Battery (0=flat, 1=fully charged)
F = Fuel Tank (0=empty, 1=full)
G = Fuel Gauge Reading (0=empty, 1=full)

This illustrative example is borrowed from **Christopher Bishop’s** book : “*Pattern recognition* and *machine learning”*.

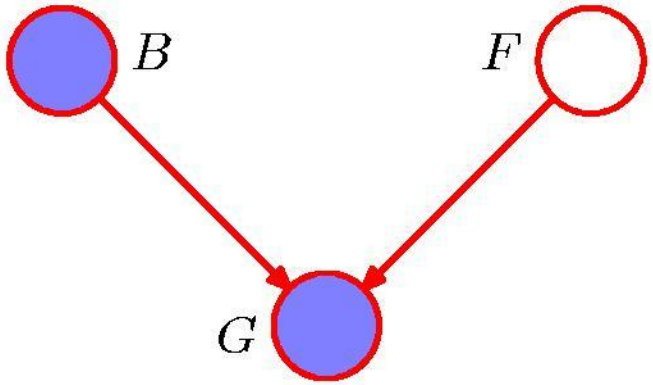
“Am I out of fuel?”



$$
\begin{aligned}
P(F=0 \mid G=0) &= \frac{P(F=0, G=0)}{P(G=0)} \\
&= \frac{\sum_b P(F=0, G=0, B=b)}{P(G=0)} \\
&= \frac{P(F=0)\sum_b P(B=b)P(G=0 \mid F=0, B=b)}{P(G=0)} \\
&= 0.257
\end{aligned}
$$

Probability of an empty tank increased by observing *G=0*, *i.e.* *P(F=0|G=0) > P(F=0).*

“Am I out of fuel?”



$$
\begin{aligned}
P(F=0 \mid G=0, B=0) &= \frac{P(F=0, G=0, B=0}{P(G=0, B=0)} \\
&= \frac{P(F=0)P(B=0)P(G=0 \mid F=0, B=0)}{\sum_f P(F=f, G=0, B=0)} \\
&= 0.111
\end{aligned}
$$

- The probability of an empty tank is reduced by observing *B = 0*,
  i.e. P(F=0|G=0,B=0) < P(F=0|G=0). This referred to as “explaining away”.
- *B* and *F* are ***negatively correlated conditioned on G*** despite being independent.

Limits of Bayesian Networks

- Two given DAGs are **observationally equivalent** if *every* probability distribution that is compatible (or faithful) with one of the DAGs is also compatible with the other (same conditional independences encoded).

- **Theorem**: Two DAGs are observationally equivalent if and only if they have the same skeletons and the **same sets of v-structures**, that is, two converging arrows whose tails are not connected by an arrow.

- Observational equivalence places a limit on our ability to infer directionality from probabilities alone.

- Networks that are observationally equivalent **cannot be distinguished without** resorting to **manipulative experimentation** or **human knowledge**

Causal Bayesian Networks

Graphs as Models of Interventions

- **Causal models**, unlike probabilistic models, can serve to **predict the effect of interventions**. This added feature requires that the joint distribution *P* be supplemented with a **causal diagram** - that is, a DAG that identifies causal connections.
- The causal diagram may represent the i**nvestigator's understanding** of the major causal influences among measurable quantities in the domain.

- Each child-parent family in a DAG *G* represents a deterministic function:

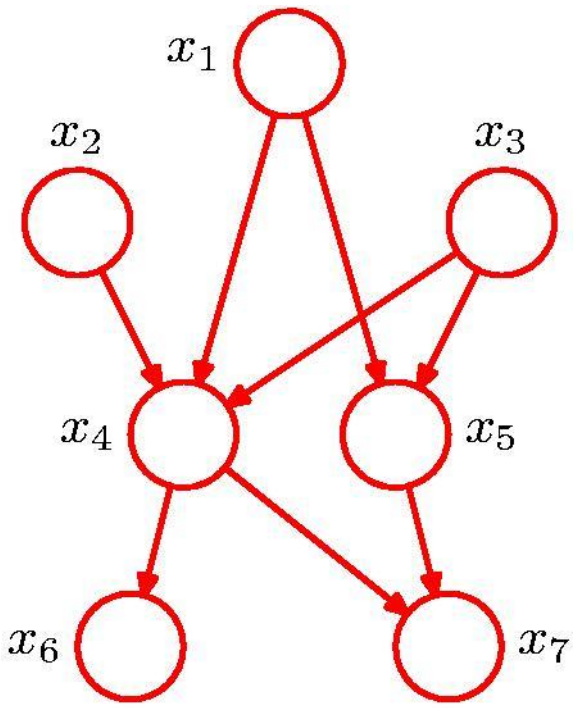
$$x_i = f_i(pa_i, \epsilon_i), \quad i = 1, \ldots, n$$

where $\boldsymbol{pa}_i$ are the parents of variable $\mathbf{x}_i$ in *G*; the ϵ_i (*i=1,…,n*) are mutually **independent**, arbitrarily distributed random disturbances.

- The equality signs in structural equations convey the **asymmetrical relation** of "is determined by“.

Causal Bayesian Networks

$$p(x_1, \ldots, x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)$$
$$p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$$



General Factorization

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k)$$

Now **supplemented** with causal assumptions

$$x_i = f_i(pa_i, \epsilon_i), \quad i = 1, \ldots, n$$

Finding causal relationships

- For finding causal relationships, the gold standard are **randomized controlled trials** initially developed in the context of agricultural research (Fisher, 1926).
- Problem: Not always feasible for ethical, financial or other reasons.

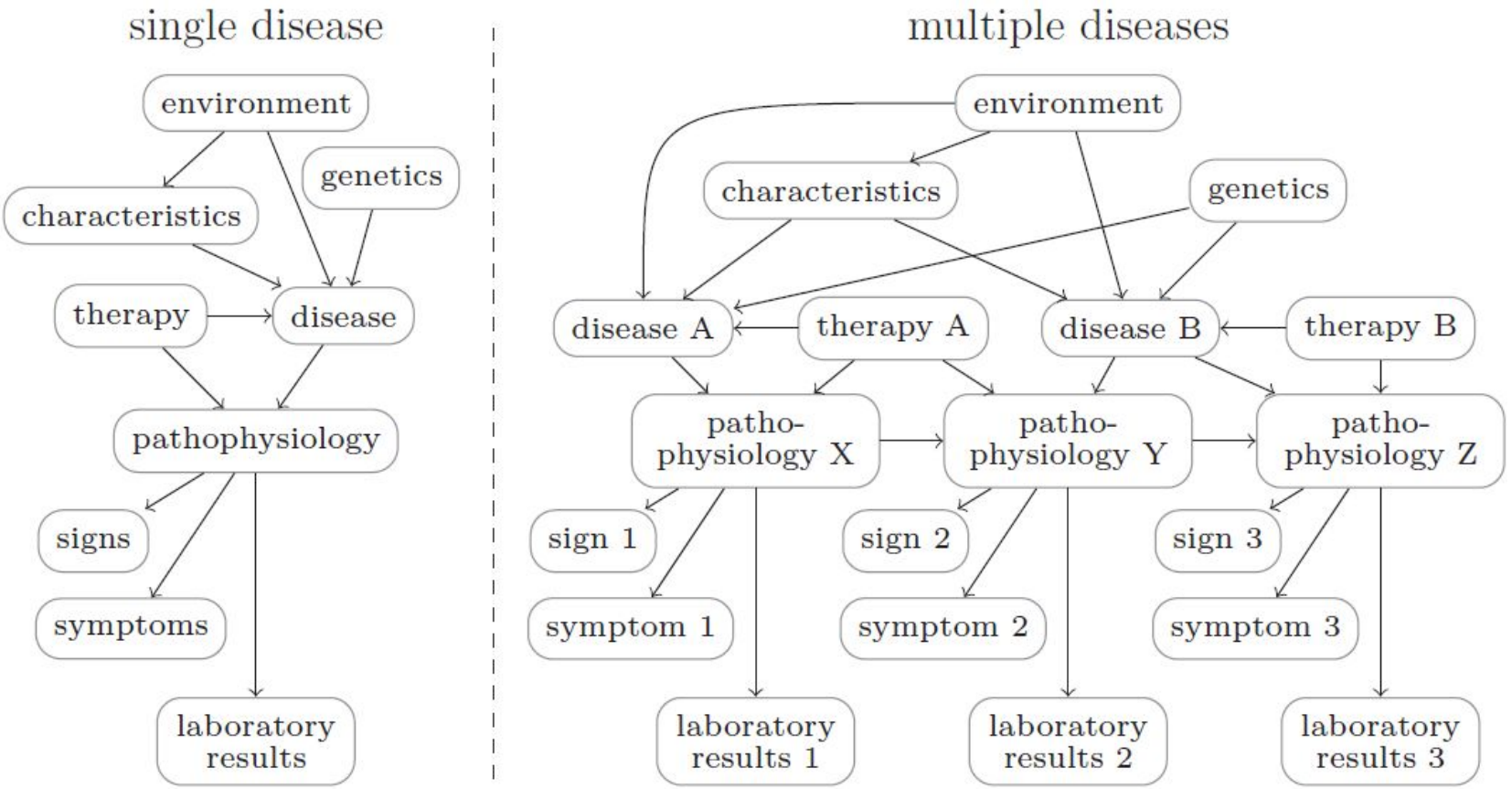
We are left with two problems:

- **Problem 1 (Causal Structure)**: Given observational data, find the DAG representing the causal structure, or, if this is not possible, give a class of DAGs to which the true DAG belongs.
- **Problem 2 (Interventional Distribution)**: Given observational data, find the interventional distribution of a random variable *Y* after some other random variable *X* was set to a certain value by external intervention to make quantitative predictions on the effect of interventions.

Finding causal relationships

- A broad range of methods (search-and-score or constraint based) has been developed for estimating causal structures from observational data assuming *no hidden confounders*
- One can **reduce the size of the equivalence classes** and ideally obtain a unique DAG by,
  - including **background knowledge**, such as time, for further orienting some edges
  - making further assumptions on the structural equation model (e.g. assuming **additive noise models** or non-Gaussian noise
  - performing **targeted experiments** ("active learning")
- With *hidden variables*, a broader class of *ancestral* graphs (MAG, ADMG) are used because DAGs are not closed under marginalization. *Ancestral* graphs are classed of infinitely many DAGs that share the same d-separations on the observed variables.

Abstract model of diseases



M. Lappenschaar et al. “Multilevel Bayesian networks for the analysis of hierarchical health care data”. Artif. Intell. in Medicine, 2013

Manipulation theorem

- The **manipulation theorem** (Spirtes et al. 1993) states that given an external intervention on a variable *X* in a causal graph, we can derive the posterior probability distribution over the entire graph by simply modifying the conditional probability distribution of *X*.

- Intervention amounts to **removing all edges that are coming into** *X*. Nothing else in the graph needs to be modified, as the causal structure of the system remains unchanged.

- Thus, intervention can be expressed in a **simple truncated factorization** formula.

The do(.) operator

- Interventions are defined through a new mathematical operator called **do(*X*=*x*)**, which simulates physical interventions by deleting the probability factor corresponding to variable *X* in the joint factorization, while keeping the rest unchanged elsewhere with *X* fixed to *x*.
-
- The causal effect of *X* on *Y* is denoted *P*(*y*|**do(*X*=*x*)**). It is termed an ***interventional*** **distribution** and should not be confused from the ***observational*** **distribution** *P*(*y*|*x*).

- Interventions can be expressed as a simple ***truncated*** **factorization** formula:

$$P(x_1, \ldots, x_n \mid \mathbf{do}(x_i = x_i')) = \begin{cases} \prod_{j \neq i} P(x_j \mid \mathbf{pa_j}) & \text{if } x_i = x_i' \\ 0 & \text{if } x_i \neq x_i' \end{cases}$$

The do(.) operator

$$P(x_1, \ldots, x_n \mid \mathbf{do}(x_i = x_i')) = \begin{cases} \prod_{j \neq i} P(x_j \mid \mathbf{pa_j}) & \text{if } x_i = x_i' \\ 0 & \text{if } x_i \neq x_i' \end{cases}$$

Can be rewritten as:

$$P(x_1, \ldots, x_n \mid \mathbf{do}(x_i = x_i')) = \begin{cases} P(x_1, \ldots, x_n \mid x_i, \mathbf{pa}_j) P(\mathbf{pa}_j) & \text{if } x_i = x_i' \\ 0 & \text{if } x_i \neq x_i' \end{cases}$$

Summing over all variables except x_i and y leads to the result called **adjustment for direct causes:**

$$P(y \mid \mathbf{do}(x_i = x_i')) = \sum_{\mathbf{pa}_i} P(y \mid x_i', \mathbf{pa}_i) P(\mathbf{pa}_i)$$

In compact form:

$$P(y \mid \mathbf{do}(x)) = \sum_{\mathbf{pa}_x} P(y \mid x, \mathbf{pa}_x) P(\mathbf{pa}_x)$$

Controlling confounding biais

$$P(y \mid \mathbf{do}(x)) = \sum_{\mathbf{pa}_x} P(y \mid x, \mathbf{pa}_x) P(\mathbf{pa}_x)$$

- We ***adjust*** our measurements for possible variations of the **parents of** *X* in the causal DAG *G*, they are acting as "covariates" or « **confounders** ».
- **Adjustment for the direct parents** amounts to **partitioning the population** into groups that are homogeneous relative to $\boldsymbol{pa}_x$ assessing the effect of *X* on *Y* in each homogeneous group, and then averaging the results.
- This expression requires all the parents to be *observed.* Are other **variables appropriate for adjustment**?
- What criterion should one use to decide **which variables are appropriate for adjustment**?

Back-Door adjustment

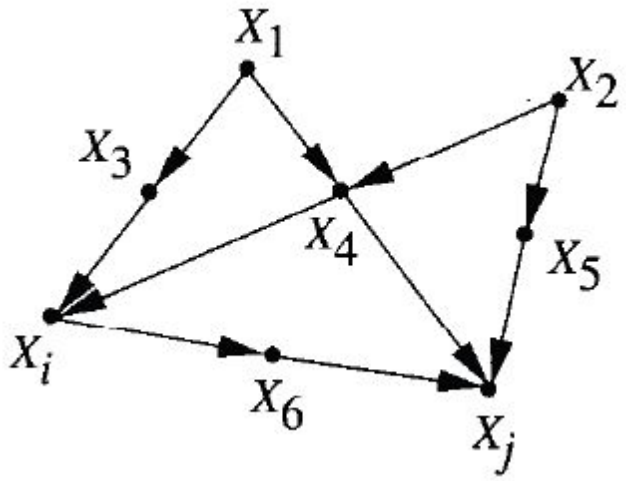
More generally, a set of variables ***Z*** satisfies the **back-door criterion** relative to (*X*,*Y*) in a DAG G iff,

- No node in **Z** is a descendant of X, and
- **Z** blocks every path between X and Y that contains an arrow into X.

Theorem – *If a set of variables **Z** satisfies the back-door criterion relative to (X,Y), then the causal effect of X on Y is **identifiable** and is given by the formula*,

$$P(y \mid do(x)) = \sum_{\mathbf{z}} P(y \mid x, \mathbf{z}) P(\mathbf{z})$$

Example:



- The sets **Z** =$\{X_3, X_4\}$ and **Z** =$\{X_4, X_5\}$ meet the back-door criterion relative to **(**X_i,X_j**)**
- But **Z** =$\{X_4\}$ does not !

Paradoxes & Controversies

Berkson’s paradox

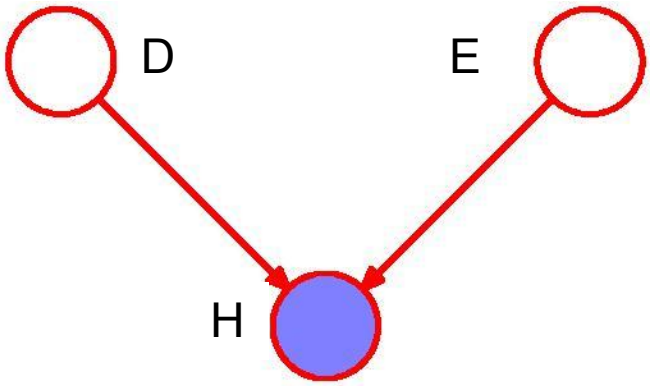
- Berkson's paradox is a result in **conditional probability** (not related de causality) which is counterintuitive for some people: given **two independent events**, if you only consider **outcomes where at least one event occurs, then they become negatively dependent.**

- **Example**: Berkson's original illustration involves a retrospective study examining a risk factor for a disease in a statistical sample. Because samples are taken from a hospital in-patient population, rather than from the general public, this can result in a spurious negative association between the disease and the risk factor

Berkson, Joseph. "Limitations of the Application of Fourfold Table Analysis to Hospital Data". Biometrics Bulletin. 1946

Berkson's paradox

	E^+		E^-	
	D^+	D^-	D^+	D^-
H^+	800	600	400	200
H^-	200	400	600	800



- The prevalence of the disease (***D***) is 50% among exposed (***E***) and unexposed.
- 70% are hospitalized (***H***) among exposed patients (30% among non exposed)
- 60% are hospitalized among diseased patients (40% among non diseased).

- **Within those hospitalized**, the prevalence of the disease is 57% among exposed and 66% among unexposed patients.

Birth weight paradox

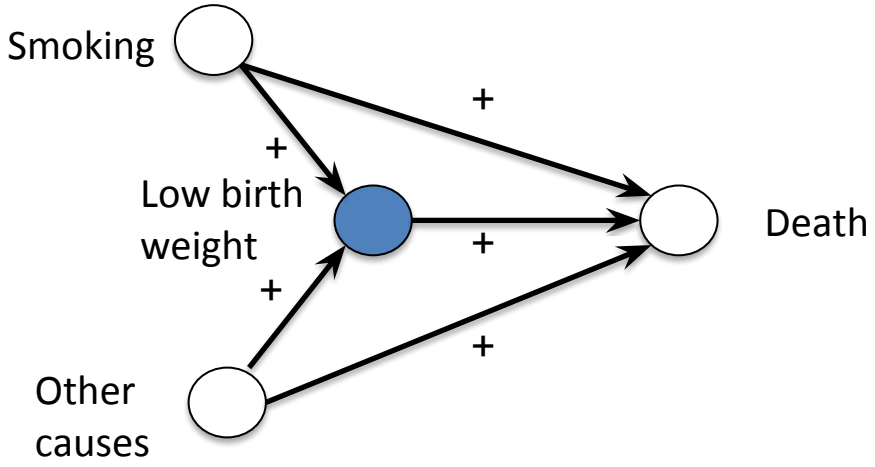
- The birth-weight paradox concerns the relationship between the birth weight and mortality. **Children of smoking mothers are more likely to be of low birth weight** and **low birth weight children have a significantly higher mortality rate than others** (it is in fact 100-fold higher)

- **Contrary to expectations**, low birth weight babies of smoking mothers have a lower child mortality than low birth weight babies of nonsmokers. **Having a smoking mother might be beneficial to one's health!**

- Like the Berkson's paradox, it is counterintuitive as it involves two independent events that become negatively dependent, having observed a third event.

Hernández-Díaz et al. "The birth weight paradox uncovered?" Am J Epidemiol 2006
Wilcox A. "On the importance—and the unimportance—of birth weight". Int J Epidemiol 2001.

Birth weight paradox



- Smoking may be harmful in that it contributes to low birth weight, but other causes (not measured) of low birth weight are generally more harmful.
- Consider a low weight baby, **finding that the mother smokes reduces the likelihood that those other causes are present**.

Simpson's paradox

C : taking a certain drug or treatment
E : recovery
F : gender

Under a causal interpretation the drug seems to be **harmful** to both males and females yet **beneficial** to the population as a whole !

Males	E	$\bar{E}$	Tot.	Recovery rate
Drug (C)	18	12	30	60%
No Drug ($\bar{C}$)	7	3	10	70%
	25	15	40	

Females	E	$\bar{E}$	Tot.	Recovery rate
Drug (C)	2	8	10	20%
No Drug ($\bar{C}$)	9	21	30	30%
	11	29	40	

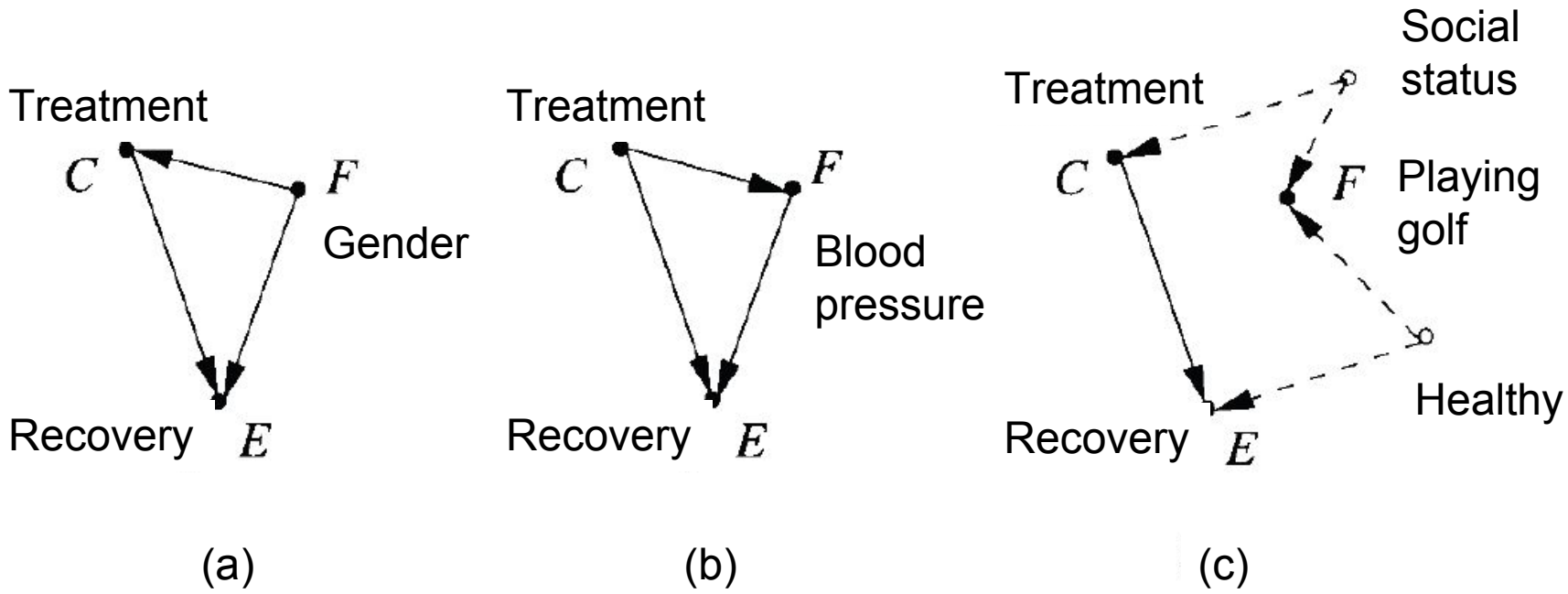
Combined	E	$\bar{E}$	Tot.	Recovery rate
Drug (C)	20	20	40	50%
No Drug ($\bar{C}$)	16	24	40	40%
	36	44	80	

Simpson's paradox

Three causal models capable of generating the data Model

(a) dictates use of the **gender-specific tables**,

whereas (b) and (c) dictates use of the **combined table.**



Simpson's paradox

As **F** connotes gender, the correct answer is the gender specific table, i.e.

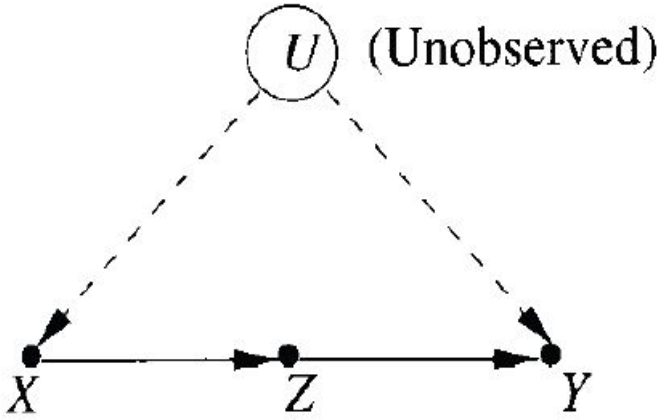
$$P(y|do(x)) = \sum_z P(y|x,z)\, P(z)$$

- **Conclusion**: every question related to the effect of actions must be decided by causal considerations; statistical information alone is insufficient.
- The question of choosing the correct table on which to base our decision is a special case of the **covariate selection problem.**

Front-Door adjustment

A set of variables *Z* is said to satisfy the **front-door criterion** relative to (*X*, *Y*) if

- Z intercepts all directed paths from *X* to *Y*;
- there is no back-door path from *X* to *Z*;
- all back-door paths from *Z* to *Y* are blocked by *X*.



Theorem : If *Z* satisfies the front-door criterion relative to *(X , Y)* and if *P(x,z)* > 0, then the causal effect of X on Y is **identifiable** and is given by the formula:

$$P(y|do(x)) = \sum_{z} P(z|x) \sum_{x'} P(y|z,x')\,P(x')$$

If *Z* were ***not*** **observed**, the causal effect of *X* on *Y* would ***not*** **be identifiable!**

Smoking and Lung Cancer

	Group Type	$P(x, z)$ Group Size (% of Population)	$P(Y = 1 \mid x, z)$ % of Cancer Cases in Group
$X = 0, Z = 0$	Nonsmokers, No Tar	47.5	10
$X = 1, Z = 0$	Smokers, No Tar	2.5	90
$X = 0, Z = 1$	Nonsmokers, Tar	2.5	5
$X = 1, Z = 1$	Smokers, Tar	47.5	85

- Old debate on the relation between **smoking**, X, and **lung cancer**, Y**.**
- If we ban smoking, will the rate of cancer cases be roughly the same as the one we find today among non smokers in the population ?
- **Controlled experiments** could answer the question but they are **illegal** to conduct.

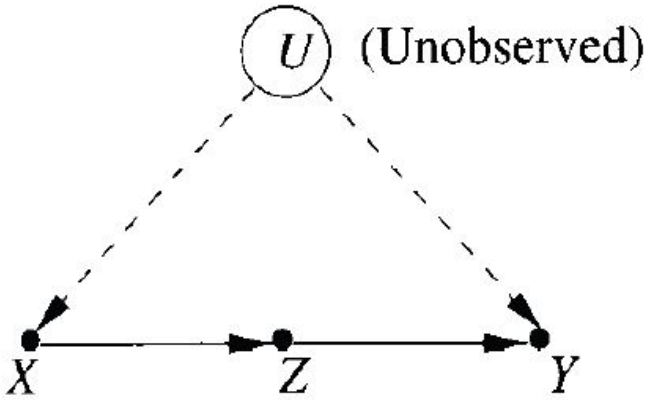
Smoking and Lung Cancer

	Group Type	$P(x, z)$ Group Size (% of Population)	$P(Y = 1 \mid x, z)$ % of Cancer Cases in Group
$X = 0, Z = 0$	Nonsmokers, No Tar	47.5	10
$X = 1, Z = 0$	Smokers, No Tar	2.5	90
$X = 0, Z = 1$	Nonsmokers, Tar	2.5	5
$X = 1, Z = 1$	Smokers, Tar	47.5	85

The tobacco industry has managed to forestall antismoking legislation (1964) by arguing that the observed correlation between smoking and lung cancer could be explained by some sort of **carcinogenic genotype**, *U* (unknown), that involves **inborn craving for nicotine**.

Smoking and Cancer

	Group Type	$P(x,z)$ Group Size (% of Population)	$P(Y=1 \mid x,z)$ % of Cancer Cases in Group
$X=0, Z=0$	Nonsmokers, No Tar	47.5	10
$X=1, Z=0$	Smokers, No Tar	2.5	90
$X=0, Z=1$	Nonsmokers, Tar	2.5	5
$X=1, Z=1$	Smokers, Tar	47.5	85



$$\boldsymbol{P(y|do(x)) = \sum_{z} P(z|x) \sum_{x'} P(y|z,x')\, P(x')}$$

Numerical application

- **Crude analysis**:

$$P(X = 1) = 0.5; P(Z = 1) = 0.5; P(Y = 1) = 0.475$$
$$P(Y = 1 \mid X = 0) = (0.1 \times 0.475 + 0.05 \times 0.025)/0.5 = 0.0975$$
$$P(Y = 1 \mid X = 1) = (0.9 \times 0.025 + 0.85 \times 0.475)/0.5 = 0.8525$$

- These results seem to prove that **smoking is a major contributor to lung cancer.**
- However, the tobacco industry might argue that the table tells a different story - that smoking actually decreases one's risk of lung cancer…

Numerical application

$$P(y|do(x)) = \sum_{z} P(z|x) \sum_{x'} P(y|z,x')\, P(x')$$

$$\begin{aligned}
P(Y=1 \mid do(X=1)) &= 0.05 \times (0.1 \times 0.5 + 0.9 \times 0.5) \\
&\quad +0.95 \times (0.05 \times 0.5 + 0.85 \times 0.5) \\
&= 0.4525 \\
P(Y=1 \mid do(X=0)) &= 0.95 \times (0.1 \times 0.5 + 0.9 \times 0.5) \\
&\quad +0.05 \times (0.05 \times 0.5 + 0.85 \times 0.5) \\
&= 0.4975
\end{aligned}$$

Contrary to expectation, the data prove **smoking** to be somewhat **beneficial to one's health** !

Discrimination controversy

- Another example involves a controversy called « reverse regression », which occupied the social science literature in the 1970s.
- Should we**,** in **salary discrimination cases**, compare **salaries of equally qualified men and women** or instead compare **qualifications of equally paid men and women**?
- Remarkably, the **two choices may lead to opposite conclusions**. It turns out that men earns a higher salary than equally qualified women and, ***simultaneously,*** men are more qualified than equally paid women.
- The moral is that all conclusions are extremely sensitive to which variables we choose to hold constant when we are comparing groups.

Discrimination controversy

- **Men earns a higher salary than equally qualified women** reads:

$$\sum_Q P(S|Male, Q)P(Q) > \sum_Q P(S|Female, Q)P(Q)$$

- **Men are more qualified than equally paid women** reads:

$$\sum_S P(Q|Male, S)P(S) > \sum_S P(Q|Female, S)P(S)$$

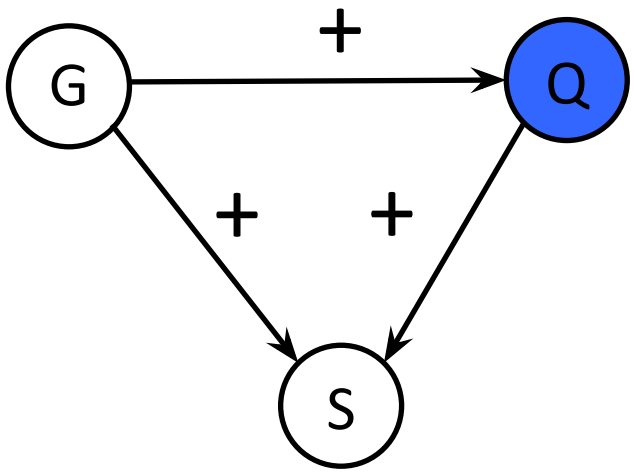
- The question we seek to answer: **does sex *directly* influence salary ?** Which is the court definition of discrimination, and reads:

$$P(S|\mathbf{do}(Male)) > P(S|\mathbf{do}(Female))$$

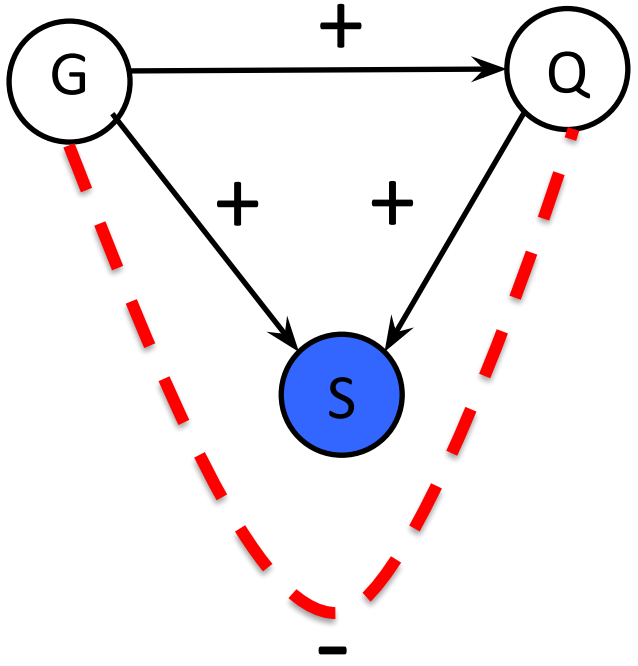
Discrimination controversy

Suppose all direct effects are positive (hence sex discrimination on salary). Conditioned on *S*, *G* and *Q* become negatively correlated via the open path in dotted lines.

Men earns a higher salary than equally qualified women



Men are more qualified than equally paid women



The Rules of do-calculus

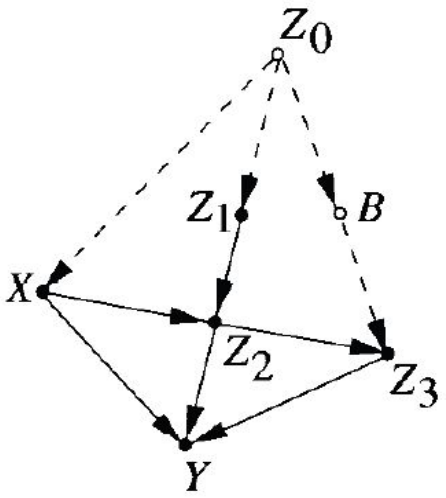
- When a query is given in the form of a do-expression, for example *P(y|**do(x)**,z),* its **identifiability** can be decided systematically using an **algebraic procedure** known as the do-calculus.

- The **do-calculus** was developed by **J. Pearl in 1995** to facilitate the identification of causal effects in non-parametric models.

- It consists of **three inference rules** that permits to map interventional and observational distributions whenever certain conditions hold in the causal diagram G.

- The do-calculus was shown to be *complete* (Tian and Pearl 2002a; Huang and Valtorta 2006; Shpitser and Pearl 2006; Bareinboim and Pearl 2012a).

Causal graphs: illustration

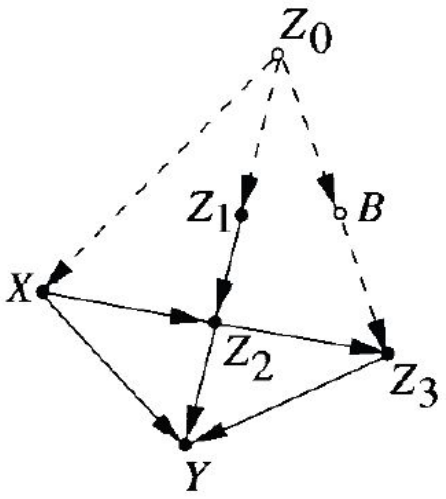


- We wish to assess the **total effect of the fumigants *X* on yields *Y*.**
- The causal diagram represents the investigator's understanding of the major causal influences among measurable quantities in the domain.

- Z_1, Z_2, Z_3 represent the **eelworm population** *before* treatment, *after* treatment, and *at the end* of the season, respectively.
- Z_0 represents *last year's* eelworm population.
- *B* is the population of birds and other predators.

Unmeasured quantities are designated by hollow circles and dashed lines.

The Rules of do-calculus



- **Using the do-calculus,** one can establish that the total effect of *X* on *Y* can be estimated consistently from the observed distribution of *X*, Z_1, Z_2, Z_3, and *Y*.

- These conclusions are obtained by **performing a sequence of symbolic derivations** (the 3 inference rules).

$$P(y \mid \mathbf{do}(x)) = \sum_{z_1,z_2,z_3} P(y \mid z_2, z_3, x) P(z_2 \mid z_1, x)$$

$$\times \sum_{x'} P(z_3 \mid z_1, z_2, x') P(z_1, x')$$

Confounding & Selection bias

Confounding & Selection bias

- The biases arising from confounding and selection are fundamentally different, though both constitute threats to the validity of causal inferences.
  - The **confounding bias** is the result of treatment *X* and outcome *Y* being affected by common ancestral variables,
  - The **selection bias** is due to treatment *X* or outcome *Y* (or ancestors) affecting the inclusion of the subject in the sample.

- In both cases, we have extraneous "flow" of information between treatment and outcome, which falls under the rubric of "spurious correlation," since it is not what we seek to estimate.

- What are the conditions for **recoverability of interventional distributions** for when selection and confounding biases are both present?

Controlling confounding biais

- Whenever we undertake to evaluate the effect of one factor, *X*, on another, *Y*, the question arises as to whether we should ***adjust*** our measurements for possible variations in some other factors *Z*, otherwise known as “covariates” or « **confounders** ».
- **Adjustment** amounts to **partitioning the population** into groups that are homogeneous relative to *Z*, assessing the effect of *X* on *Y* in each homogeneous group, and then averaging the results.
- The practical question that it poses - whether an adjustment for a given covariate is appropriate - has resisted mathematical treatment.
- Epidemiologists often adjust for wrong sets of covariates: is the prevailing practice misguided?
- What criterion should one use to decide **which variables are appropriate for adjustment**?

Confounding with latent variables

- Some relevant confounders are difficult to measure in many real-world applications (e.g., intention, mood, DNA mutation), which leads to the need of modelling explicitly **latent variables** that affect more than one observed variable in the system (Semi- Markovian models).

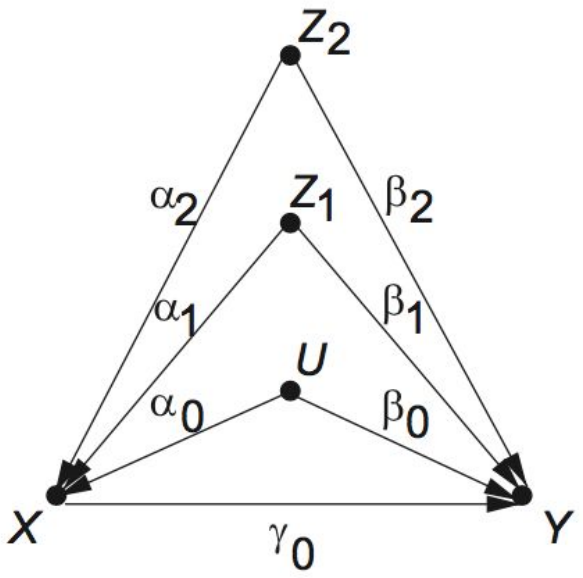
- In such models, identifiability is *not always achievable*.

- **Causal Effects Identifiability**: Let be ***V*** the set of *observable* variables, ***U*** is the set of *unobservable* variables. The causal effect of an action, **do**$(X = x)$ is said to be identifiable from P in G if $P(y|\mathbf{do}(x))$ is uniquely computable from $P(v)$.

- The evaluation of identifiability goes through a **non-trivial algebraic process**, namely the *do-calculus*.

Confounding: Bias amplification

Linear structural model with
two **instrumental variables** Z_1 *and* Z_2 and
one unobserved **confounder** U



- Z_1 and Z_2 are *observed* **confounder** and U is an *unobserved* confounder.
- We are seeking to approximate $P(Y|\mathbf{do}(X))$ but conditioning on U is not possible. The **total bias** is :

$$\alpha_0\beta_0 + \alpha_1\beta_1 + \alpha_2\beta_2$$

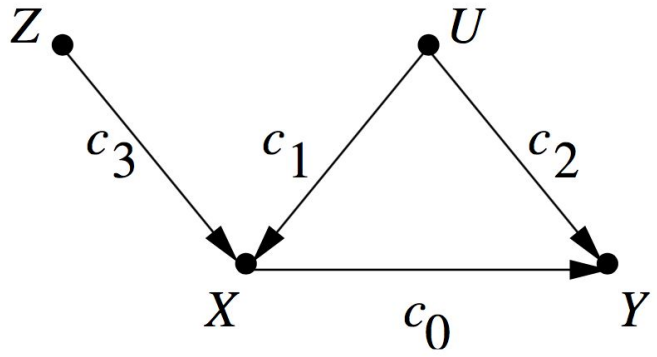
- **Is conditioning Z_1 and Z_2 a good idea?** Not always...

- Using **linear structural model**, one can show that conditioning on Z_1 and Z_2 produces a bias equal to :

$$\frac{\alpha_0\beta_0}{(1-\alpha_1^2-\alpha_2^2)}$$

J. Pearl “Linear Models: A Useful “Microscope” for Causal Analysis”, Tech. report 2013.
S. Wright. “Correlation and causation”. Journal of agricultural research, 1921.

Instrumental variable



A **linear structural equation model** with **instrumental variable** *Z* and **confounder** *U*

- *Z* is a **pre-treatment variable**, i.e. it is not affected by the treatment, nor does it interfere with the causal pathways from treatment X to outcome Y.
- *Z* seems to behave just **like an ordinary confounder** *U*, i.e. *Z* is dependent on *X* and *Y*, still is dependent on *Y* given *X*.
- We are seeking to approximate *P(Y*|**do***(X))* but conditioning on *U* is not possible. **Is conditioning *Z* a good idea? No!**
- Using a **linear structural equation model**, one can show that conditioning on *Z* amplifies the unconditioned bias c_1c_2 by a factor $1/(1-c_3^2)$

J. Pearl "On a Class of Bias-Amplifying Variables that Endanger Effect Estimates", Proceedings of UAI, 2012.

Confounding : risks and pitfalls

- Researchers must weigh the benefit of reducing confounding bias carried by those covariates against the risk of **amplifying residual bias** carried by **unmeasured confounders.**

- According to Judea Pearl, epidemiologists often adjust for wrong sets of covariate (usually *Sex* and *Age* but other covariates are missing*)* .

- Is the prevailing practice in epidemiology misguided?

Controlling selection bias

- Another major challenge that needs to be addressed when evaluating the effect of interventions is the problem of selection bias, caused by **preferential exclusion of samples** from the data.

- Selection bias is a **major obstacle to valid causal and statistical inferences**; it can hardly be detected in either experimental or observational studies.

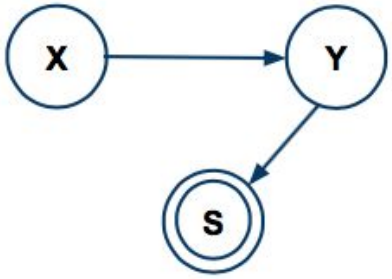
- **Example**: in a typical study of the effect of training program on earnings, subjects achieving higher incomes tend to report their earnings more frequently than those who earn less.

Selection bias

●To illuminate the nature of this bias, consider a variable *S* affected by both *X* (treatment) and *Y* (outcome), indicating entry into the data pool.

●Such preferential selection to the pool **amounts to conditioning on** *S*, which creates spurious association between *X* and *Y*.

●Our assumption about the selection mechanism are embodied in an **augmented causal graph G_s**.



●**Illustration** : Effect of training program on earnings

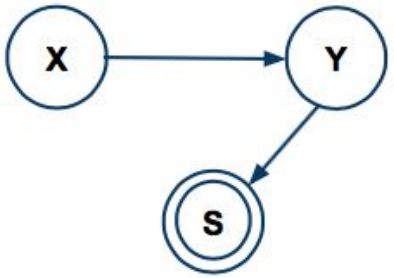
●S represents the selection mechanism. *S=1* indicates presence in the sample, and *S=0* exclusion.

Recoverability

- Under what conditions $P(y|\mathbf{do}(x))$ can be recovered from data drawn from $P(y, x|S = 0)$?

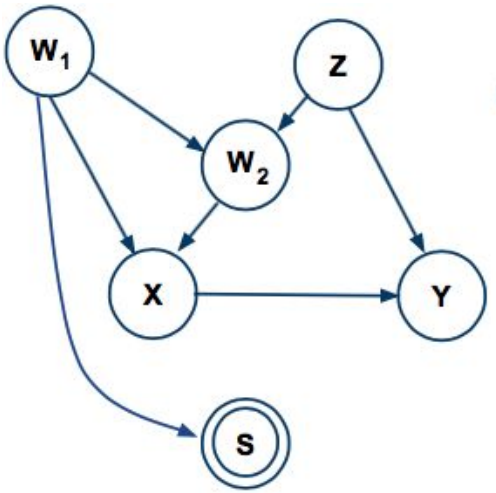
- **Recoverability from Selection Bias:** Given a causal graph G_s augmented with S, $P(y|\mathbf{do}(x))$ is said to be recoverable from selection biased data in G_s if $P(y|\mathbf{do}(x))$ is expressible in terms of the distribution under selection bias $P(v|S = 0)$.

- In this example, $P(y|\mathbf{do}(x))$ is *not* recoverable



Recoverability

$$P(y \mid \mathbf{do}(x)) = \sum_{w_1,w_2} P(y \mid x, w_1, w_2) P(w_1, w_2)$$
$$= \sum_{w_1,w_2} P(y \mid x, w_1, w_2, S = 1) P(w_1, w_2)$$



- It may appear that $P(y|\mathbf{do}(x))$ is not recoverable since the second term $P(w_1, w_2)$ is *not* recoverable, however

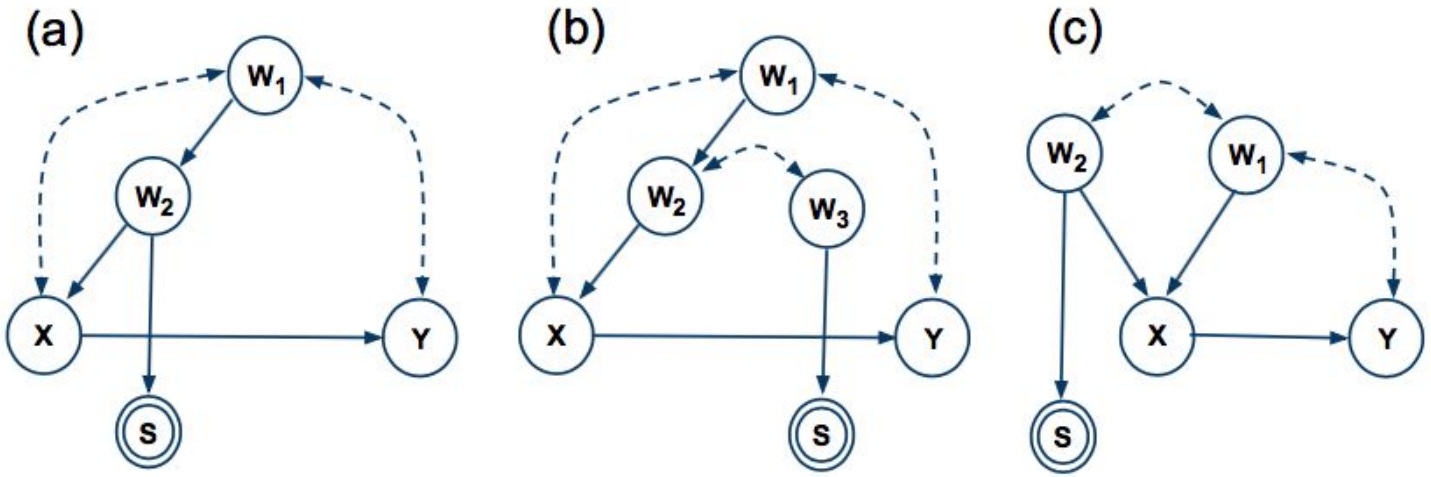
$$P(y \mid \mathbf{do}(x)) = \sum_{z} P(y \mid x, z) P(z)$$
$$= \sum_{z} P(y \mid x, z, S = 1) P(z \mid S = 1)$$

.

- This is another expression witnessing the identifiability of $P(y|\mathbf{do}(x))$, but in this case, it is recoverable.
- If not recoverable, we need an additional *unbiased* data set.

E. Bareinboim and J. Tian. “Recovering Causal Effects from Selection Bias”, Proc. AAAI 2015.

Recoverability: Illustration

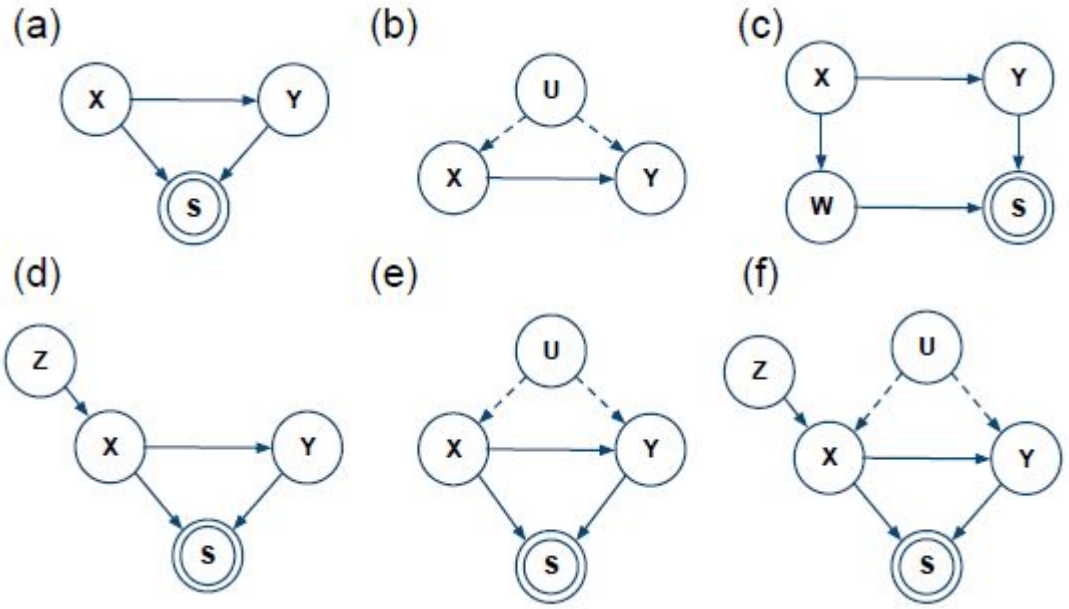


- Non-trivial scenarios involving intricate relationship of the counfounded structure and the S-nodes.

- *P(y*|**do***(x))* is not recoverable in (a) but is in (b) and (c).

E. Bareinboim and J. Tian. “Recovering Causal Effects from Selection Bias”, Proc. AAAI 2015.

Bias scenarios in social and medical sciences



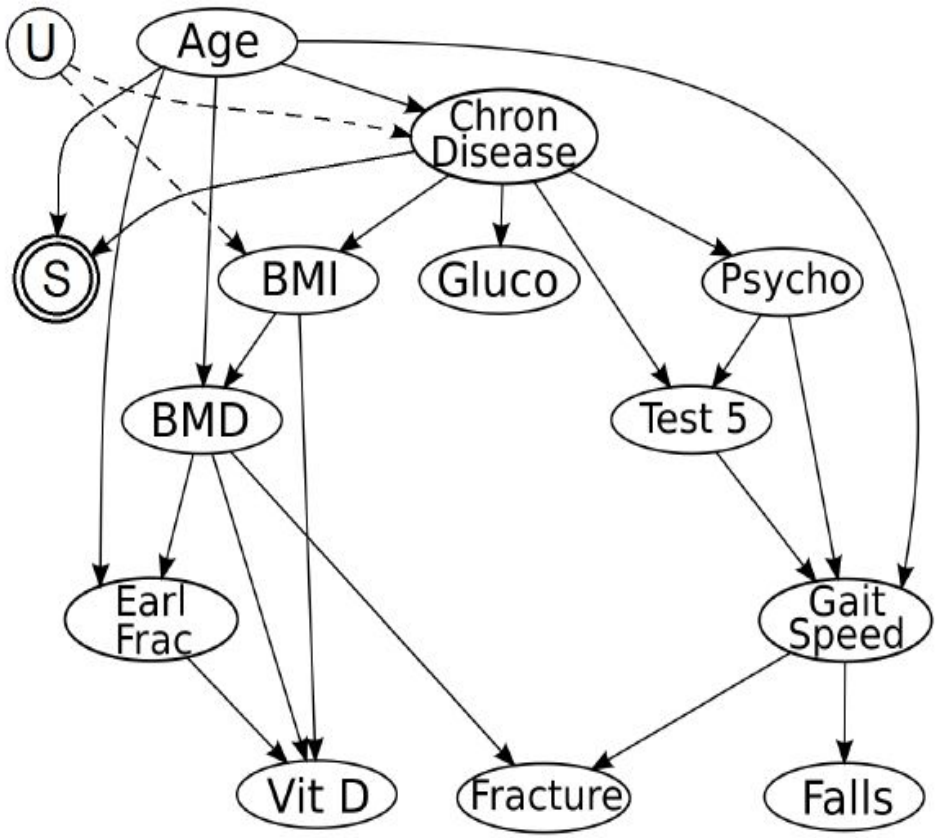
Z is an Instrumental variable.

U is a latent variable, i.e. unobserved variable acting as a *confounder.*

S represents the selection mechanism. S=1 indicates presence in the sample, and S=0 otherwise.

(a) Simplest example of selection bias
(b) Simplest example of confounding bias
(c) Intermediary variable W between X and selection S
(d) Instrumental variable with selection bias
(e) Selection combined with confounding
(f) Instrumental variable with confounding and selection bias simultaneously present

Osteoporotic fracture risk assessment



- Prospective cohort study with 7500 elderly osteoporotic women followed-up during 4 years.
- A *plausible* causal BN was learned from a combination of **non-experimental data** and qualitative assumptions that are deemed likely by health experts.
- Inclusion of a **selection mechanism** and an **unobserved confounder.**
-
- We seek to estimate the strength of the **causal effect of psychotropic drugs** on the risk of **hip fracture:**

$$P(Fracture \mid \mathbf{do}(Psycho)) = ?$$

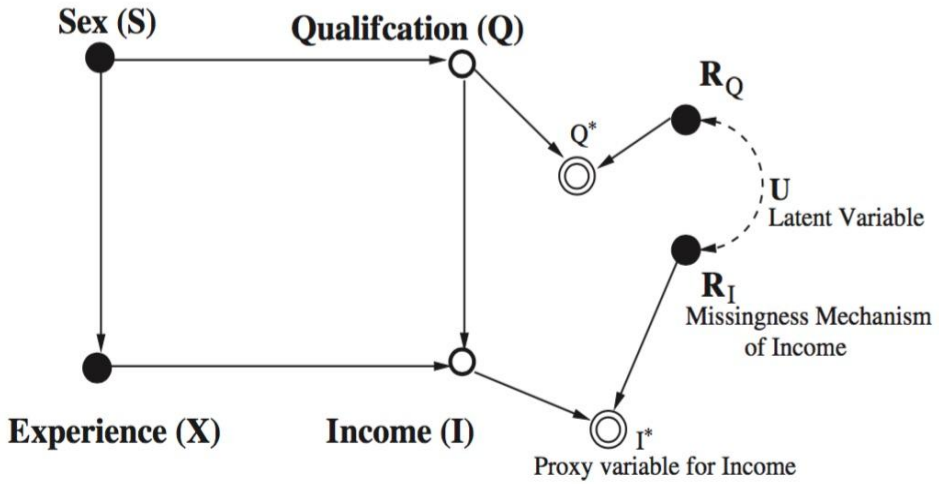
A. Aussem et al. “Analysis of risk factors of hip fracture with causal Bayesian networks”. IWBBIO 2014.
P.Caillet et al. “Hip fracture in the elderly: a re-analysis of the EPIDOS Study with causal Bayesian Networks”, Plos One, 2015

Missing data

Missing data

- All branches of experimental science are plagued by missing data
- The “missing data” problem arises when values for one or more variables are missing from recorded observations
- Occurs often in social science, epidemiology, biology and survival data analysis etc.
- Caused by varied factors such as high cost involved in measuring variables, failure of sensors, reluctance of respondents in answering certain questions
- **Improper handling of missing data** can bias outcomes and potentially **distort the conclusions** drawn from a study.

Misingness mechanism : *m*-graph



- Associated with every partially observed variable $V_j \in V_{miss}$ are **two other variables** R_j and V_j^*
- V_j^* is a proxy variable that is actually observed.
- R_j represents the status of the causal mechanism responsible for the missingness of V_j*

$$v_i^* = f(r_{v_i}, v_i) = \begin{cases} v_i & \text{if } r_{v_i} = 0 \\ m & \text{if } r_{v_i} = 1 \end{cases}$$

Observed and partially missing variables are represented by full and hollow circles respectively.

Missing data

Let R represents the status of the causal mechanism responsible for the missingness variables, and V_{obs} and V_{miss} denote the fully observed and partially missing variables.

- **Missing Completely At Random** (MCAR) if

  $P(R|V_{obs}, V_{miss}) = P(R)$

  Example: when respondents decide to reveal their income levels based on coin-flips.

- **Missing At Random** (MAR) if

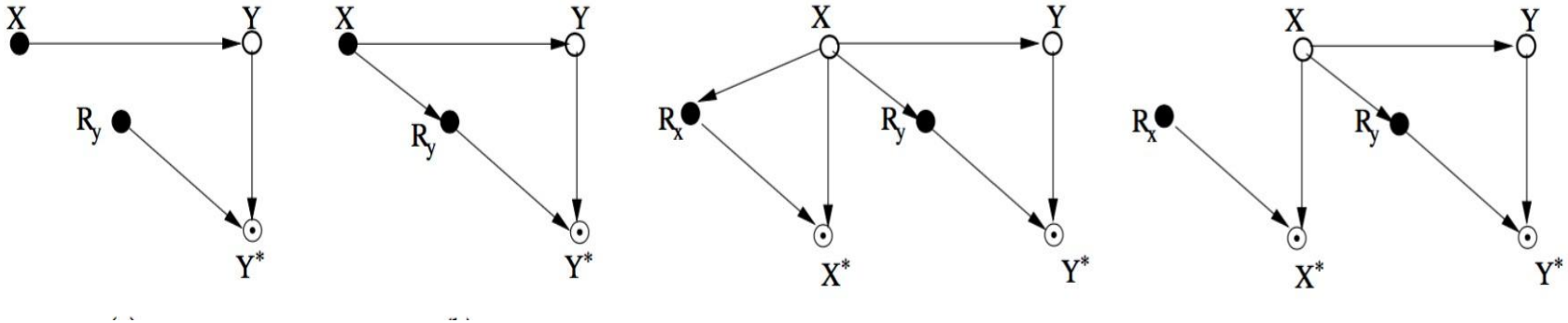
  $P(R|V_{obs}, V_{miss}) = P(R|V_{obs})$

  Example: Women in the population are more likely to not reveal their age.

- **Not Missing At Random** (NMAR) if data are neither MAR nor MCAR.

  Example: The probability that a customer supplies a rating is dependent on he's underlying liking.

Missingness mechanisms



- V_j^* is a proxy variable that is actually observed, and R_i represents the status of the **causal mechanism responsible for the missingness** of V_j.

- Data that are: (a) MCAR, (b) MAR, (c) & (d) MNAR. Hollow and solid circles denote partially and fully observed variables respectively

---

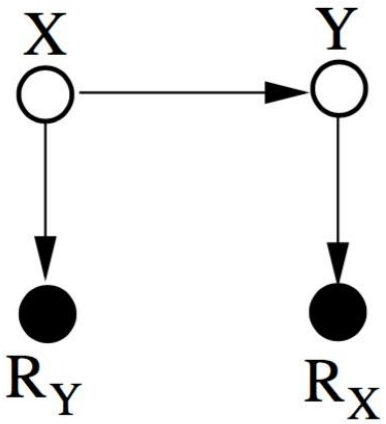
Mohan K. et al. *"Missing Data as a Causal Inference Problem".* NIPS, 2013

Recoverability with missing data

Let V_{obs}, V_{miss} be the set of observed and missing variables

- **Recoverability from Data Missingness Bias:** Given a causal graph G augmented with the missingness variables R, $P(y|\mathbf{do}(x))$ is said to be recoverable in G if $P(y|\mathbf{do}(x))$ is expressible in terms of the distributions $P(V_{obs}, V_{miss}|R = 0)$.

Recoverability even when data is NMAR!



$$P(X,Y) = P(X,Y)\frac{P(R_x,R_y|X,Y)}{P(R_x,R_y|X,Y)}$$
$$= \frac{P(R_x,R_y)P(X,Y|R_x,R_y)}{P(R_x|Y,R_y)P(R_y|X,R_x)}$$

- *P(X,Y)* is decomposed into a product of terms, namely $P(R_x=0,R_y=0)$, $P(X,Y|R_x=0,R_y=0)$, $P(R_x=0|X,R_y=0)$ and $P(R_y=0|Y,R_x=0)$, that are **all recoverable**.

- So P(y|**do**(x)) = P(y|*x*) = P(x,y)*/P(x) is also recoverable despite being NMAR.*

Conclusions

- Testing for cause and effect is difficult, discovering cause effect is even more difficult.
- But, once the **causal diagram** is provided (both from *expert knowledge* and data), identification of causal effects is straightforward using the *do-calculus* rules.
- Many **paradoxes** and **controversies** in social and medical sciences can be illustrated and understood by simple graphical means.
- The **data missingness** and **selection mechanisms** can easily be represented in the diagram for **bias correction** purposes.
- **Inference of causal relationships from massive data sets** is still a challenge but may eventually lead to new discoveries (e.g. cancer)

References

- J. Pearl. Causality: *Models, Reasoning, and Inference*. New York: Cambridge University Press, 2009.
- J. Pearl "Understanding Simpson's Paradox'', UCLA Cognitive Systems Laboratory, Tech. Rep. R-414, 2013.
- J. Pearl "*Do*-Calculus Revisited" UCLA Cognitive Systems Laboratory, Conference on Uncertainty in Artificial Intelligence (UAI) 2012.
- S. Lauritzen. *Graphical Models*. Clarendon Press: Oxford, 1996.
- J. Pearl "Myth, confusion, and science in causal analysis", UCLA Cognitive Systems Laboratory, Tech. Rep. R-348, 2009.
- J. Pearl "On a Class of Bias-Amplifying Variables that Endanger Effect Estimates", Proceedings of UAI, 417-424, 2012.
- E. Bareinboim and J. Tian. "Recovering Causal Effects from Selection Bias", Proc. AAAI 2015.
- J.A. Myers et al. « Effects of adjusting for instrumental variables on bias and precision of effect estimates. Am. J. of Epidemiology 2011.
- A. Goldberger. "Reverse regression and salary discrimination". The Journal of Human Resources 1984.
- J. Berkson. "Limitations of the application of fourfold table analysis to hospital data ». Biometrics Bulletin 1946.
- P. Spirtes, et al. *Causation*, *Prediction* and *Search*, MIT Press, 1993.
- A.P. Dawid: *Fundamentals of Statistical Causality*. Research Report No. 279, University College London, 2007.
- *M. Kalisch et P. Bühlmann. Causality: A selective review, Quality Technology & Quantitative, 2014*
- K. Mohan et al. *"Missing Data as a Causal Inference Problem".* NIPS, 2013

Thank you for your attention, any question ?