Intervention and Causality

Volker Tresp
2016

Avoiding Embarrassments

- Big Data Scientist found: “Taking your breaks outside causes lung cancer!”

Causation and Correlation

- Correlation does not imply causation (well known in statistics)
- No correlation without causation (since we are living in a causal world)

- Causality, randomness, direction of time, destiny, God and all that

Causality

- In this lecture we follow the theory of Judea Pearl (Judea Pearl: Causality, 2nd Edition, 2009)
- Recommended reading: Judea Pearl, Causal inference in statistics: an overview, Technical Report R-350, September 2009
- Causality goes beyond statistical modelling and needs additional constructs

What is Causality? Superfluous?

- See: http://singapore.cs.ucla.edu/LECTURE/
- "All philosophers," says Bertrand Russell, "imagine that causation is one of the fundamental axioms of science, yet oddly enough, in advanced sciences, the word 'cause' never occurs ... The law of causality, I believe, is a relic of bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm ..."
- "Likewise, we say that the ratio f/a helps us DETERMINE the mass, not that it CAUSES the mass. Such distinctions are not supported by the equations of physics, and this leads us to ask whether the whole causal vocabulary is purely metaphysical. "surviving, like the monarchy...etc."

What is Causality? Superfluous? Cont'd

- "I felt like a buccaneer of Drake's days -... I interpreted that sentence of Galton to mean that there was a category broader than causation, namely correlation, of which causation was only the limit, and that this new conception of correlation brought psychology, anthropology, medicine, and sociology in large parts into the field of mathematical treatment." Karl Pearson (1934)

- Laplace's demon was the first published articulation of causal or scientific determinism by Pierre-Simon Laplace in 1814. According to determinism, if someone (the Demon) knows the precise location and momentum of every atom in the universe, their past and future values for any given time are entailed; they can be calculated from the laws of classical mechanics

- Possible arguments in favor of an arrow of time: entropy, quantum mechanics (Copenhagen interpretation), relativity, chaos theory, computational complexity, ...

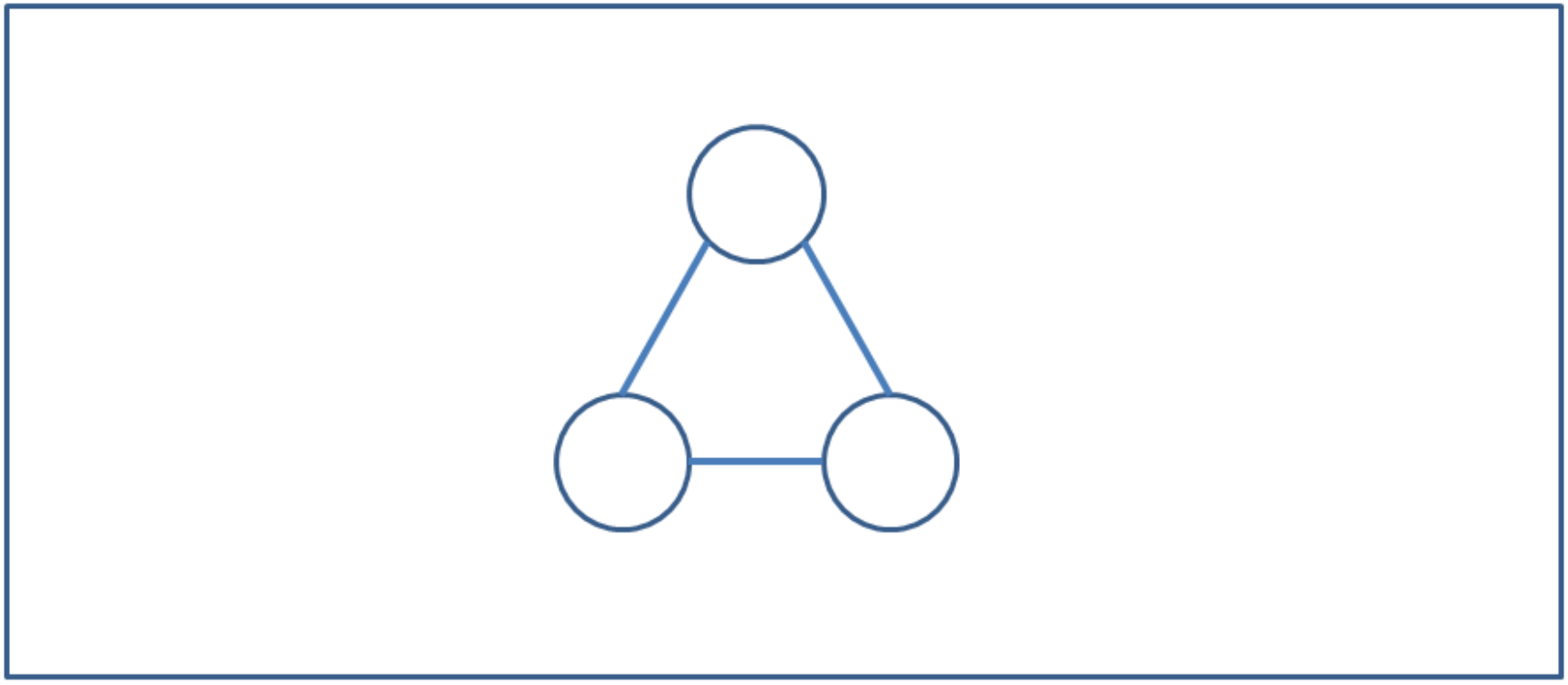
What is Causality? Important Concept Beyond Statistics

- J. Pearl's argument: If you wish to include the entire universe in the model, causality disappears because interventions disappear - the manipulator and the manipulated loose their distinction. However, scientists rarely consider the entirety of the universe as an object of investigation. In most cases the scientist carves a piece from the universe and proclaims that piece: IN namely, the FOCUS of investigation. The rest of the universe is then considered OUT or BACKGROUND, and is summarized by what we call BOUNDARY CONDITIONS. This choice of INs and OUTs creates asymmetry in the way we look at things, and it is this asymmetry that permits us to talk about "outside intervention", hence, causality and cause-effect directionality

Causality Goes Beyond Statistics!

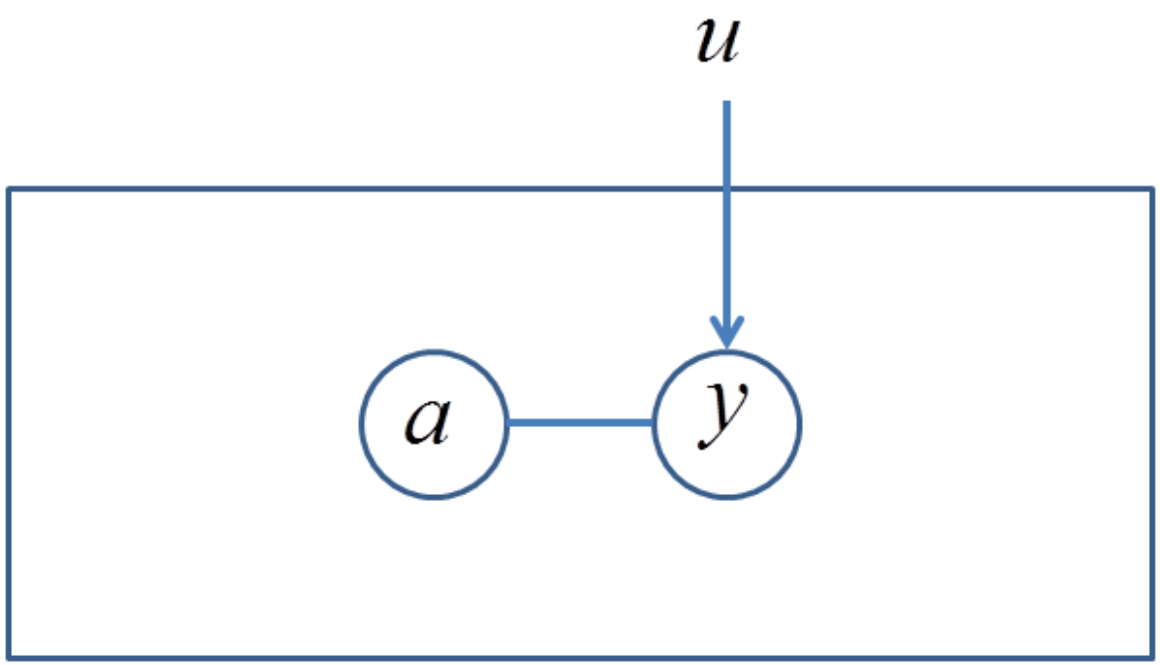
- These are causal questions because they require some knowledge of the data-generating process; they cannot be computed from the data alone, nor from the distributions that govern the data
- The laws of probability theory do not dictate how one property of a distribution ought to change when another property is modified. This information must be provided by **causal assumptions** which identify relationships that remain invariant when external conditions change. Behind every causal conclusion there must lie some causal assumption that is not testable in observational studies
- Causal analysis is about change, and change demands a new vocabulary that distinguishes "seeing" from "doing"
- Pearl's theory also permits to deal with counterfactuals: if there would not have been ice on the road I would not have had that accident and I would have been in time for the job interview, I would have gotten the job and I would be rich by now

The whole universe might be deterministic, non-random, without an identifiable arrow of time and the concept of cause and effect might be meaningless

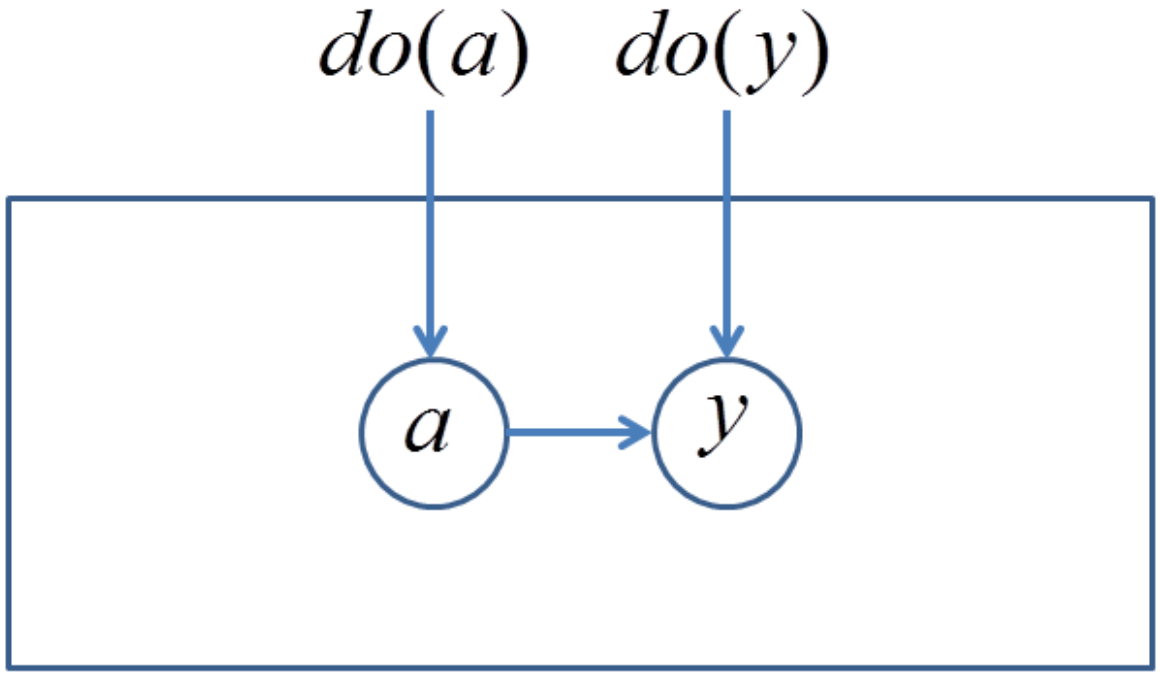


- *From the view point of a part of the world, the external world exerts influence "u" that might appear random*
- *Still I cannot decide if*

$$y <= a+u \quad or \quad a <= y-u$$



- Intervention: an intervention from outside, do(a) or do(y), makes the causal direction clear:
- *do(a)* will affect a and *y*
- *do(y)* will only affect *y* but not *a*

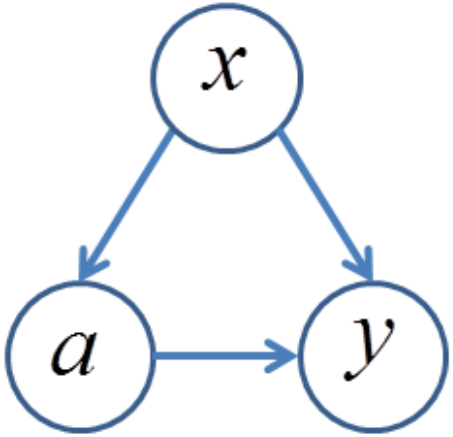


Experimental Control (Randomized Experiments)

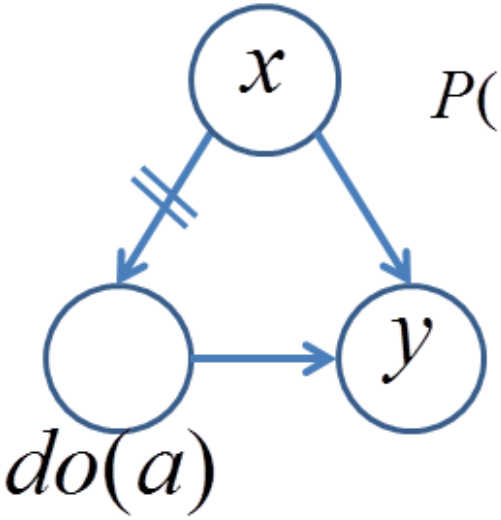
- Causal assumptions, in contrast, cannot be verified, even in principle, unless one resorts to experimental control
- This is exactly what is going on in randomized clinical studies
- But can we determine the effect of intervention from a casual model? This is the topic in the following slides

Do-Calculus

- We start with a Causal Bayesian Network (Causal Probabilistic Network (CPN)) that reflects our **causal assumptions** about a domain
- It must be complete, such that the uncertainties (noise) in the different nodes are independent
- Pearl argues, that such a causal model cannot come from a statistical analysis alone: it must be formulated by the expert
- (Comment: We have been working on the problem of identifying causality from observed data and other groups are still doing that. Pearl would argue that causal directions can only be detected if additional assumptions are made)
- Then, to calculate the causal effect of a on y, one needs to remove all incoming arrows into a, and in this modified CPN, calculate $P(y|do(a))$. In general this is not identical to the $P(y|a)$ in the original CPN



$$P(y \mid a) = \frac{1}{P(a)} \sum_x P(x) P(a \mid x) P(y \mid a, x)$$



$$P(y \mid do(a)) = \sum_x P(x) P(y \mid a, x)$$

What Variables to Condition On?

- What can I do if I can make causal assumptions but I cannot quantify the conditional probabilities in the model, but I can access data
- Solution: Although in general $P(y|a) \neq P(y|do(a))$, it might be that

$$P(y|a, S) = P(y|do(a), S)$$

where S is a set of variables that are not descendants of a

- **Then collect data on a, y and the variables in S and model $\widehat{P}(y|a, S) \approx P(y|a, S)$**
- In the previous example, $S = x$ and

$$P(y|a, x) = P(y|do(a), x)$$

from which

$$P(y|do(a)) = \sum_x P(x)P(y|a, x)$$

- S is called the "sufficient set" or "admissible set" for adjustment

Formal Recipe (J. Pearl)

- Definition: (Admissible sets - the back-door criterion for the causal effect of a on y). A set S is admissible (or "sufficient") for adjustment if two conditions hold:
  - 1. No element of S is a descendant of a
  - 2. The elements of S "block" all "back-door" paths from a to y, namely all paths that end with an arrow pointing to a.
- Then: $P(y|a, S) = P(y|do(a), S)$

A: Models with Two Variables

- If we have prior knowledge that a causes y, then $P(y|a) = P(y|do(a))$
- a could be an action (medication, treatment) or a potential disease cause (smoking)
- Thus, observed correlation corresponds to causal influence

Model with 2 Variables

Identifiability of Parameters

- Still there can be a problem if always interventions are applied, for example when a has three states: medication M1, medication M2, no medication. We have a problem if either medication M1 is applied or medication M2 is applied but never none or both. Then the models

$$\hat{P}(y|M_1, M_2) = sig(10M_1 + 9M_2)$$

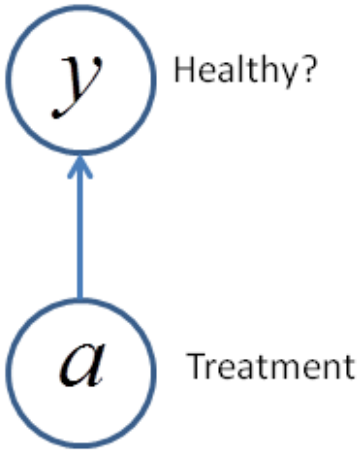
and

$$\hat{P}(y|M_1, M_2) = sig(-M_2 + 10)$$

give the same results if either medication is used. In the second model, medication M2 even appears harmful. The problem is that the case that no medication is given is never observed and the system *cannot be identified* (problem of parameter identification)

- This problem is related to the problem of the identification of controlled systems
- Thus we need so called controls (patients which received neither medication)

Model with 2 Variables



- *a has 4 states: M1, M2, M1 and M2, No Medication*
- *If only M1 or M2 are observed, then parameters cannot be identified*

B: Models with Three Variables

- We want to decide about the causal influence of a on y and there is another variable x, which is correlated with y and potentially with x
- What should we do? Should we include x in the model (correct for x)? The answer is: it depends
- We call $\widehat{P}(y|a)$ the **simple model** and $\widehat{P}(y|a,x)$ the **controlled model**
- We now need to consider several cases

B.I: x and a are independent

- We observe correlations between x and y and a and y but not between x and a
- The joint model is $P(a)P(x)P(y|a,x)$
- In this case the observed and the causal model are identical (there is no link into a)

$$P(y|do(a)) = P(y|a)$$

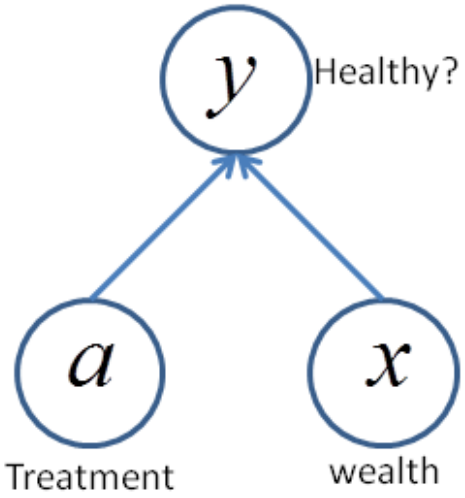
  Thus the **simple model** gives the right answer!

- We can also include x

$$P(y|do(a),x) = P(y|a,x)$$

  and then $\widehat{P}(y|do(a)) = \sum_x \widehat{P}(x)\widehat{P}(y|a,x)$

- Since a and x are independent, there should be no problem quantifying the causal influence (parameter identification), if all combinations of a, x are observed
- In the example, the prospect of becoming healthy might depend on if the patient is wealthy (has a healthier lifestyle in general) but the treatment decision does not consider that


y
Healthy?
a
x
Treatment
wealth


Healthy?
y
x
Treatment

Standard Data Set

- The true model is that a and x are independent and the model is

$$\widehat{y} = -0.18a + 0.82x$$

- The simple model would model (if $x \in \{0, 1\}$, $mean(x) = 0.5$)

$$\widehat{y} = -0.18a + 0.41$$

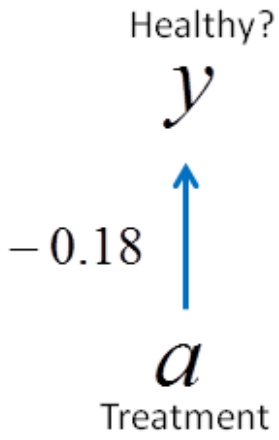
- The controlled model would get

$$\widehat{y} = -0.18a + 0.82x$$

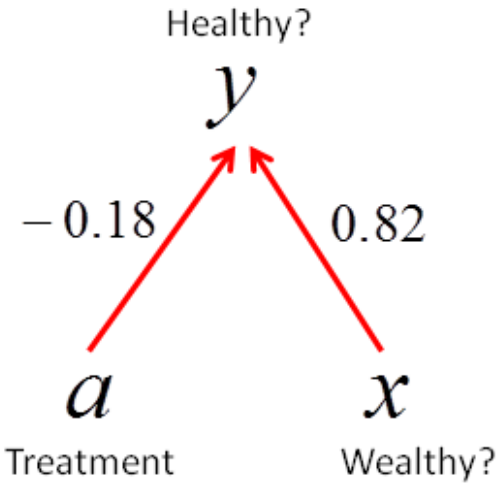
- If a is a medication, both models tell me that the medication is harmful and the recommendation would be not to use it

a and *x* are uncorrelated

simple model

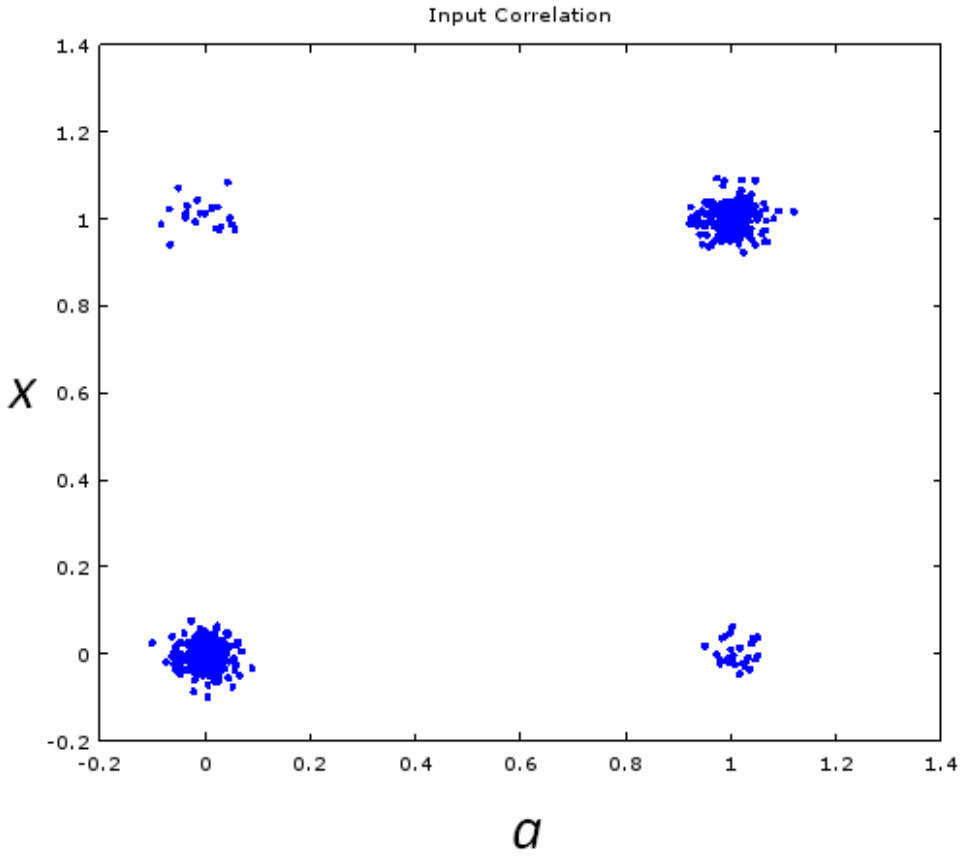


controlled model



B.II: Correlations between x and a

- We observe correlations between x and a
- The following three cases deal with this issue
- Dependent on the underlying causal model, we get different answers



We see the strong correlation between a and x

B.II Case 1: a is the cause of x

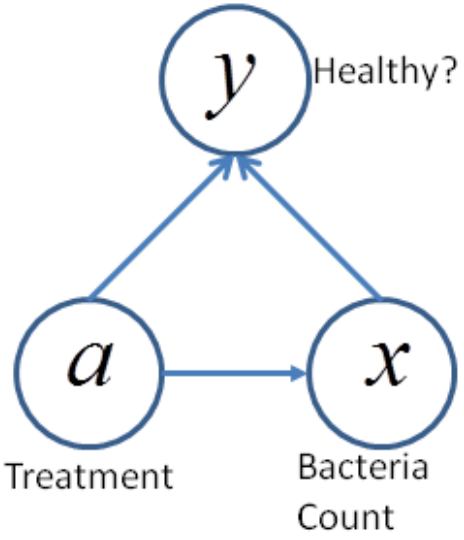
- Here, a is the cause of x. The joint model is $P(a)P(x|a)P(y|a,x)$
- Again,

$$P(y|a) = P(y|do(a))$$

  Thus the **simple model** gives the right answer!

- We cannot include x in the model since x is not known at the time of decision!
- If one would use the controlled model, one would miss the fact that the bacterial count was lowered as a consequence of the medication

a is a cause of x

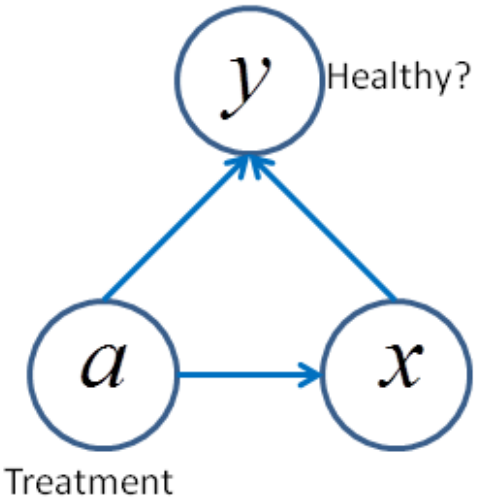


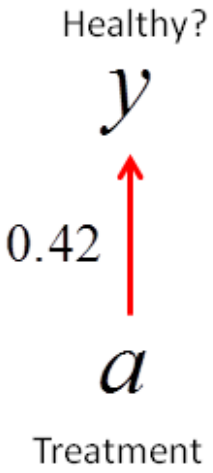
Standard Data Set

- Despite the fact that $P(y|a,x)$ was as before I get

$$\widehat{y} = 0.42a$$

- Now the result is dramatically different: Now it seems that a has a positive influence on health!


y
Healthy?
a
x
Treatment


Healthy?
y
0.42
a
Treatment

B.II Case 2: x is the cause of a

- x is the cause of a. The joint model is $P(x)P(a|x)P(y|a,x)$
- Now,

$$P(y|a) = \frac{1}{P(a)} \sum_x P(x)P(a|x)p(y|a,x)$$

- This is not the same as

$$P(y|do(a)) = \sum_x P(x)p(y|a,x)$$

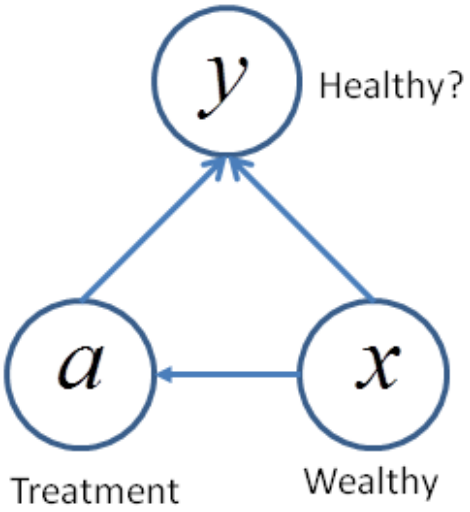
- But

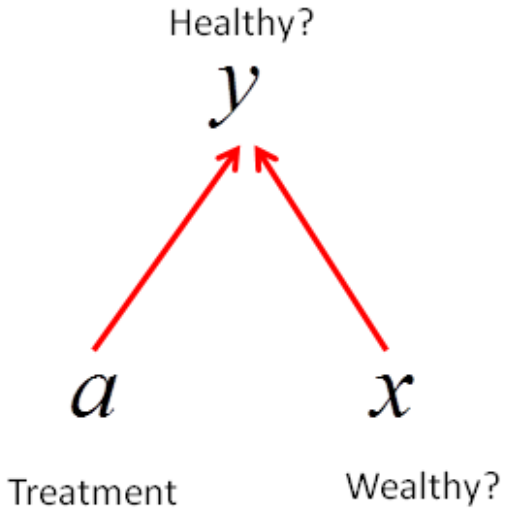
$$P(y|a,x) = P(y|do(a),x)$$

- We must include x in the model! Thus the **controlled model** gives the right answer!
- Collinearity might be a problem in the quantification of the causal influence: all combinations of x, y need to be observed!

- x might be the wealth of the patient. For example it might be that only wealthy people can afford the expensive treatment and that wealthy people have a healthier lifestyle and thus get healthier anyways and not because of the treatment

We need to control for x





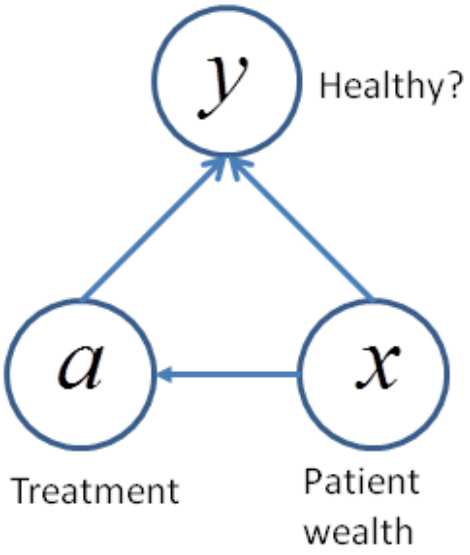
Standard Data Set

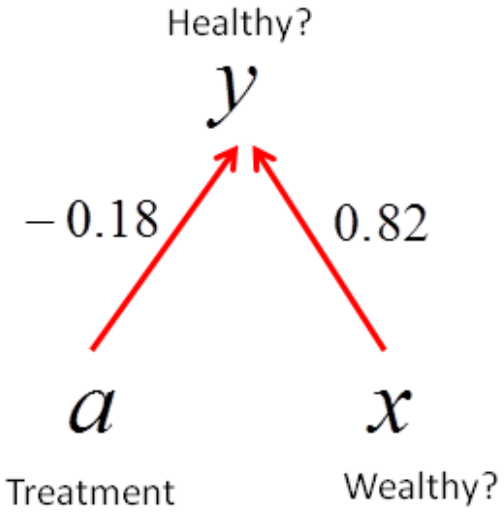
- I get again

$$\widehat{y} = -0.18a + 0.82x$$

- a has a negative effect on health again!

We need to control for x

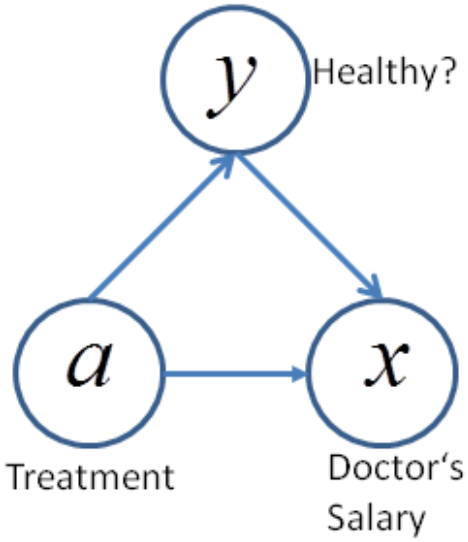




B.II Case 3: a and y are both causes for x

- The joint model is $P(a)P(y|a)P(x|a,y)$
- Again $P(y|a) = P(y|do(a))$. Thus the **simple model** gives the right answer!
- We cannot include x in the model since x is not known at the time of decision! In fact including x would give misleading results!
- In the example, x might be the doctor's salary which depends on the outcome and the expensive treatment
- We need to use the simple model since x is an effect of a (see Pearl's recipe)
- Somewhat against intuition, considering less information is helpful in a causal analysis! The fact that the controlled model gives the wrong answer in this situation is called **Berkson's paradox**

Berkson's paradox



C: Four Variables

- Often neither x is the cause of a nor a is a cause of x, but both have a hidden common cause (which does not influence y directly)

Common Cause

- Thus the joint model is:

$$P(h)P(x|h)P(a|h)P(y|a,x)$$

- Then

$$P(y|a) = \frac{1}{P(a)} \sum_{h,x} P(h)P(x|h)P(a|h)P(y|a,x)$$

and this is not the same as

$$P(y|do(a)) = \sum_{h,x} P(h)P(x|h)P(y|a,x)$$

- But when we condition on x,

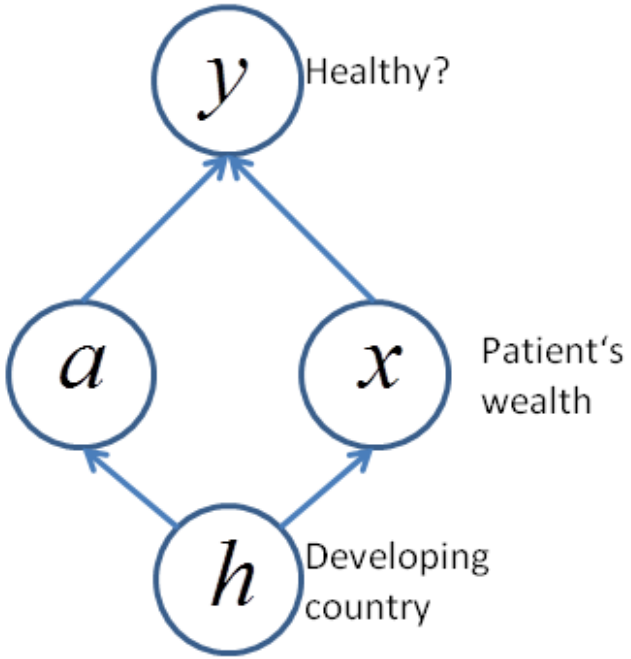
$$P(y|a,x) = P(y|do(a),x) = P(y|a,x)$$

Thus the **controlled model** gives the right answer!

- We can also control for h, since

$$\widehat{P}(y|a,h) = \sum_x \widehat{P}(x|h)\widehat{P}(y|a,x) = \widehat{P}(y|do(a),h)$$

h is a confounder if *x* is unknown (back door criterion)



Standard Data Set

- I get again

$$\widehat{y} = -0.18a + 0.82x$$

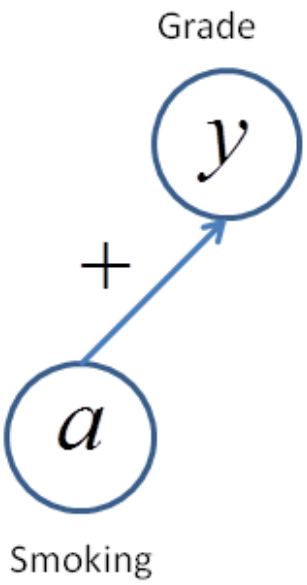
- a has a negative effect on health again!

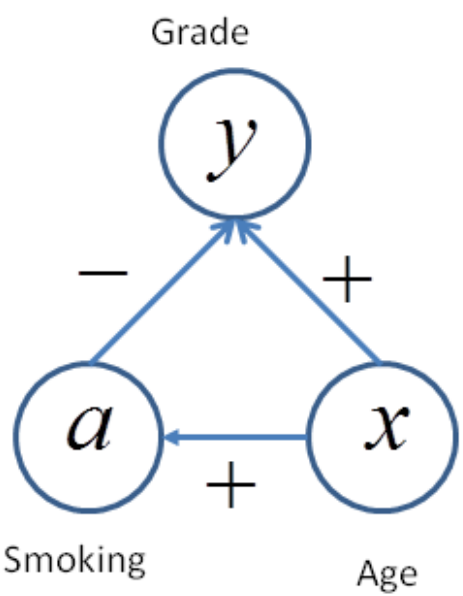
Dealing with Confounders

- When x is a confounder, then there are two options
- First option: As discussed, we include the confounder as input in the model and use the controlled model
- Second option: We train two models without x as input. The first model only sees the data where $x = 1$ and the second model only sees the data with $x = 0$.
- This strategy is called **stratification** and the two groups are called **strats**
- Dealing with confounders is called: "adjusting or controlling for confounders". The set of confounders to be included is called an admissible (or sufficient) set of variables for adjustment.

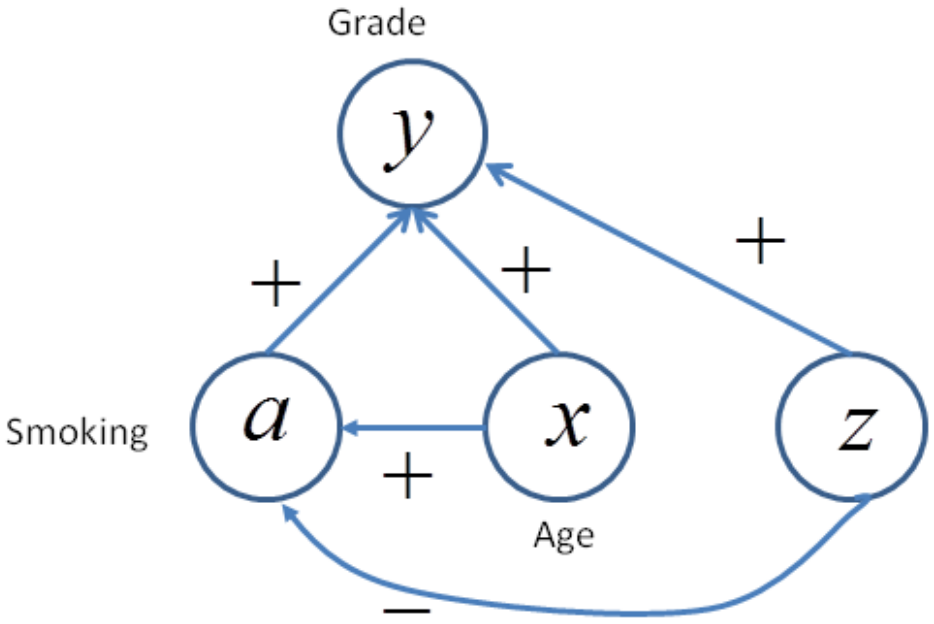
Simpson's Paradox

- When confounders are included sequentially, the intermediate results might look quite confusing





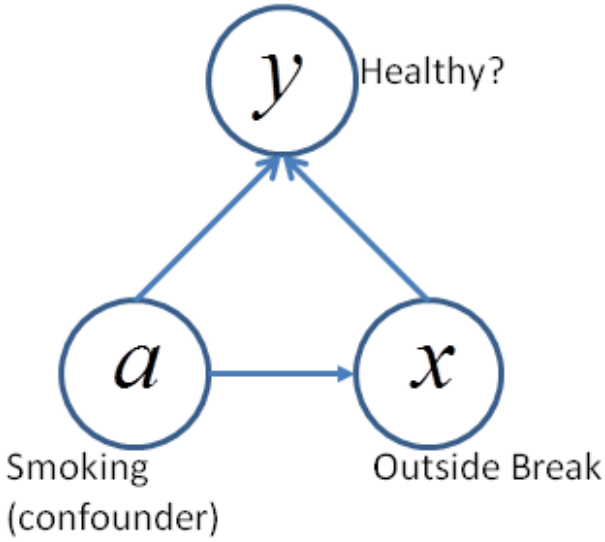
SIMPSON'S PARADOX

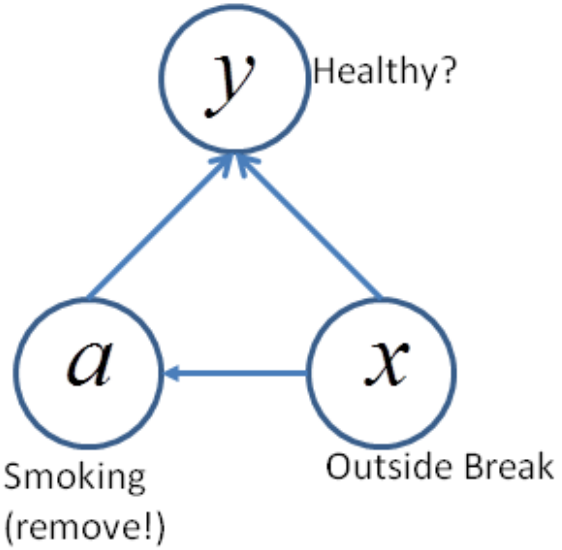


Which factors to include and which not?

... and so on ... until all confounders are included

Smoking and Spending Breaks Outside





- Both models are plausible, although the left one probably is more plausible
- If you believe that both is possible for different groups of people than you would need to analyze both groups in separate strats

Warning on Controlling

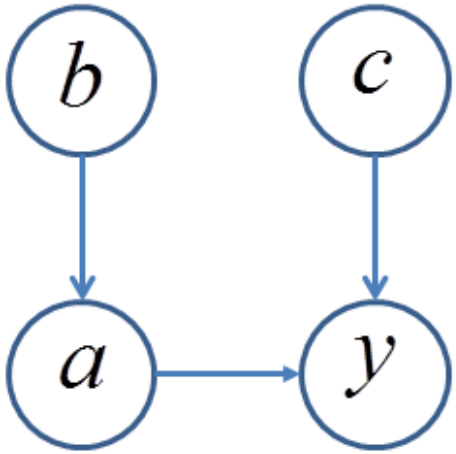
- Warning: The rule is NOT, controlling on anything that happened prior to a is safe!
- Consider the model $P(b)P(c)P(a|b)P(y|a,c)$
- Here

$$P(y|do(a)) = P(y|a) = \sum_c P(c)P(y|a,c)$$

and this is equal to

$$\sum_b P(b)P(y|a,b) = \sum_c P(b)P(y|a,c) = \sum_{b,c} P(b)P(y|a,b,c)$$

- So here it does not really matter if I control or not


b
c
a
y

Counter Example

- Consider the model $P(b)P(c)P(d|b,c)P(a|b)P(y|a,c)$
- Here, again

$$P(y|do(a)) = P(y|a)$$

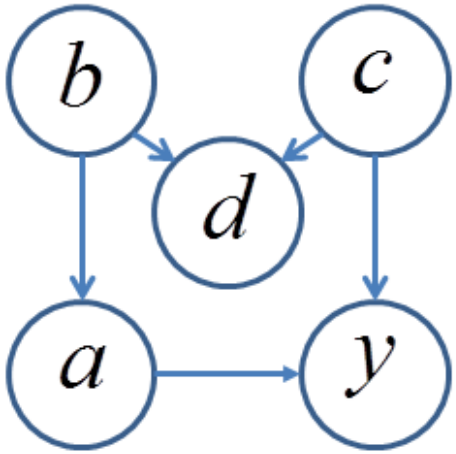
- But,

$$P(y|a,d) = \frac{1}{P(a,d)} \sum_{b,c} P(b)P(c)P(d|b,c)P(a|b)P(y|a,c)$$

$$P(y|do(a),d) = \frac{1}{P(d)} \sum_{b,c} P(b)P(c)P(d|b,c)P(y|a,c)$$

  and these two expressions are not identical

- So I should not control on d. The reason is that knowing d opens the path between b and c


b
c
d
a
y

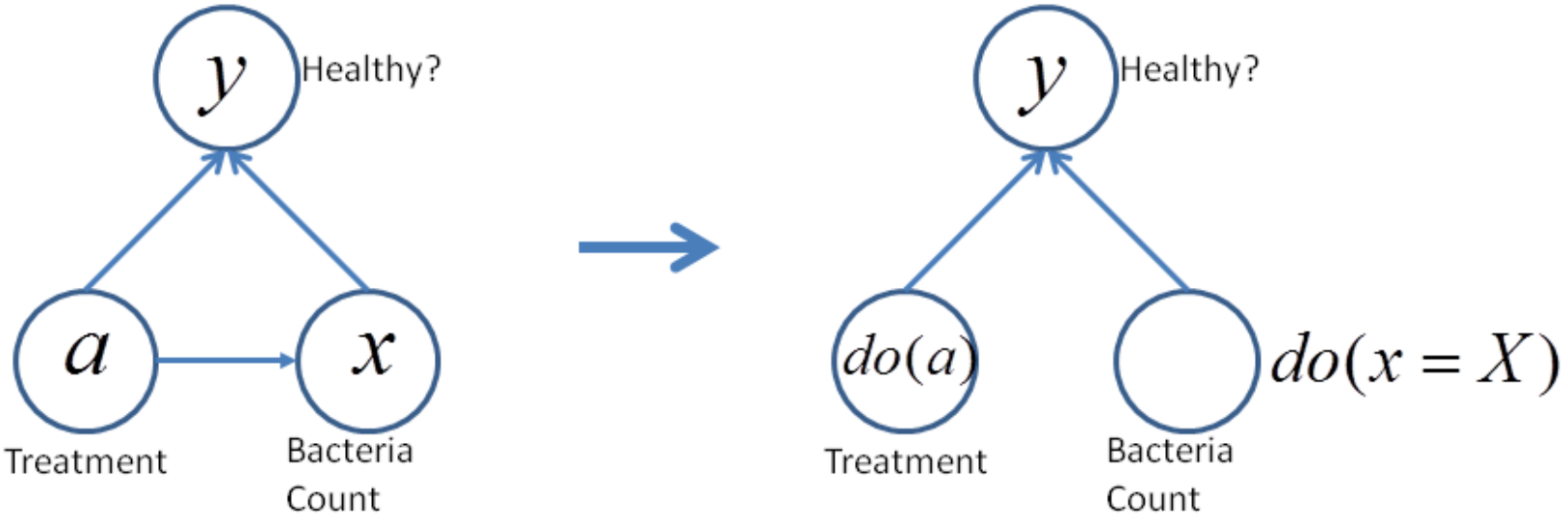
Direct Causal Effect

- So far we focussed on the *total effect* of a on y
- The term "direct effect" is meant to quantify an effect that is not mediated by other variables in the model or, more accurately, the sensitivity of y to changes in a while all other factors in the analysis are held fixed
- To analyse the direct effect, we hold all other variables fixed and analyse

$$P(y|do(a), do(x = X)$$

Here we mean that we analyse the effect of a on y by removing all arrows into x (and a) and by keeping x at the value X

Direct Cause



By analyzing the direct cause, one might capture the side effects of a

A causal chain: What is the real cause of the liver problems?

Liver problems

Red wine drinking

Attending Dagstuhl seminars

Being a world-class computer scientist researcher

„Being a world-class computer scientist researcher“ has a total causal effect on „liver problems“ but no direct causal effect

Effect Modification

- In nonlinear models, a might have a positive effect for some states of x and a negative effect for others. This issue is called: *effect modification, interaction, or heterogeneity between strata*
- Example: A medication is only effective in adults

Counterfactuals

- Pearl's theory also permits to deal with counterfactuals: if there would not have been ice I on the road I would not have had that accident A and I would have been in time for the job interview J, I would have gotten the job and I would be rich R by now
- Note that counterfactual reasoning can be applied to a an individual
- The conditional probabilities need to be formulated as structural equation models (SEMs) of the form

$$P(y|a) = f_y(a, u_y)$$

  where u_y is an exogenous variable
- The model in the above example is $P(I)P(A|I)P(J|A)P(R|J)$

Counterfactuals (cont'd)

- Let the exogenous variables u_I, u_A, u_J, u_R all be independent Gaussian zero mean variables with variance one
- The SEM model is (*Ind* is the indicator function)

$$I = f_I(u_I) = \mathit{Ind}(u_I > 0)$$

$$A = f_A(I, u_A) = \mathit{Ind}(I - 1 + u_A > 0)$$

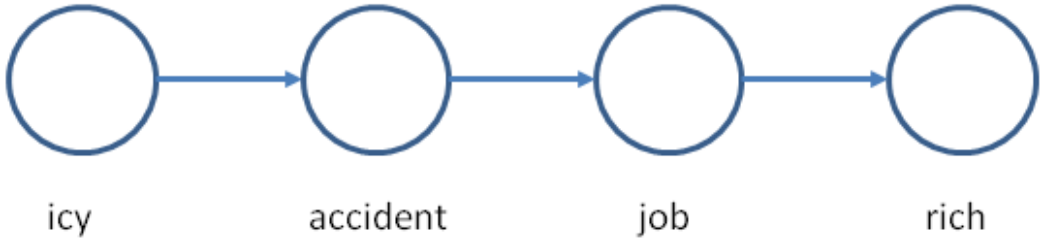
$$J = f_J(A, u_J) = \mathit{Ind}(-A + u_J > 0)$$

$$R = f_R(J, u_R) = \mathit{Ind}(J - 1 + u_R > 0)$$

- Note that all functions $f()$ map to $\{0, 1\}$
- After the fact, all states of the endogenous variables are observed $I = 1, A = 1, J = 0, R = 0$ and the exogenous variables $u_I = 0.3, u_A = 0.2, u_J = 0.4, u_R = 0.1$ are known and observed

- To analyse counterfactual reasoning, we simply change $I = 1$ into $I = 0$ and leave the values of the exogenous variables as is
- Then, using the same values for the exogenous variables, $A = 0, J = 1, R = 1$.
- Thus A becomes false, J becomes true, and R becomes true as well
- Thus is fact, if the road would not have been icy, I would be rich by now!

Direct Cause

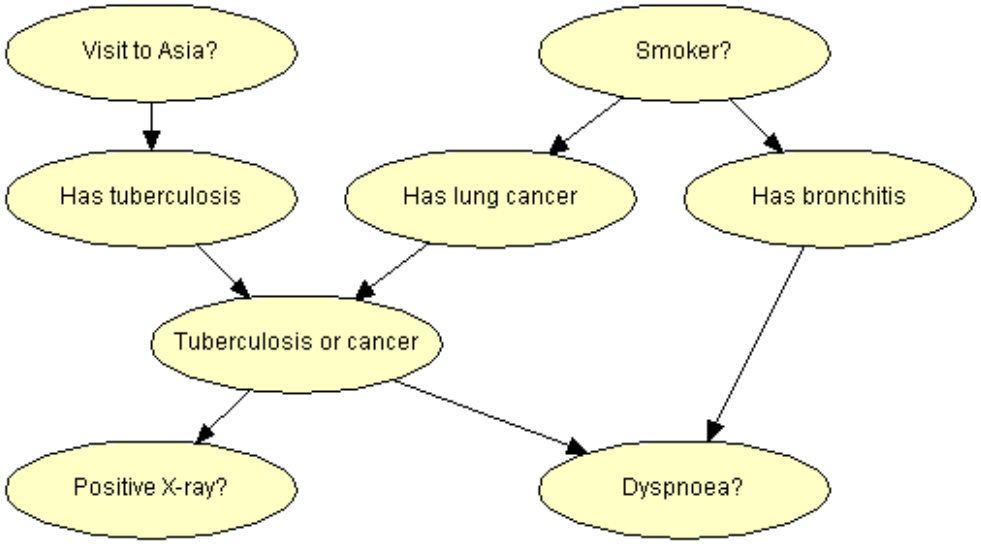


$$P(I)P(A \mid I)P(J \mid I)P(R \mid J)$$

Example: Chest Clinic

What is the effect of Lung Cancer (LC) on Dyspnea (D)?

Chest Clinic



$$P(A)P(S)$$

$$x\ P(T|A)P(LC|S)P(B|S)$$

$$x\ P(ToC|\ T, LC)P(X|ToC)P(D|ToC, B)$$

Conditional Probability

- The joint probability is

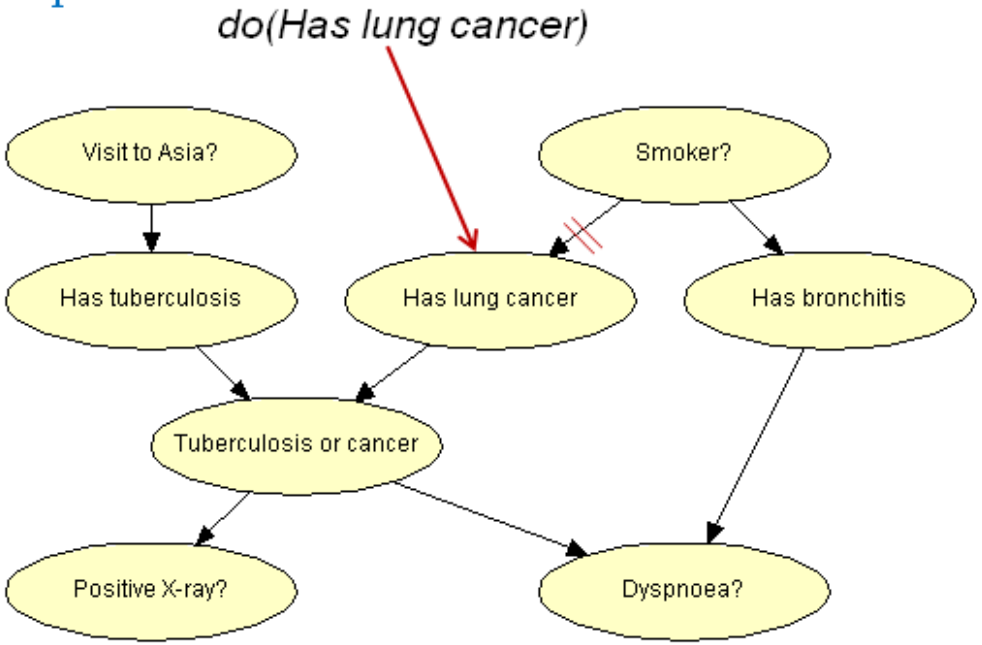
$$P(A)P(S)P(T|A)P(LC|S)P(B|S)P(\mathit{ToC}|T,LC)P(X|\mathit{ToC})P(D|\mathit{ToC},B)$$

- The conditional probability is

$$P(D|LC) = \frac{1}{P(LC)}$$

$$\sum_{A,S,T,B,ToC} P(A)P(S)P(T|A)P(LC|S)P(B|S)P(\mathit{ToC}|T,LC)P(D|\mathit{ToC},B)$$

Chest Clinic: to analyze the casual effect of "Has lung cancer" on "Dyspnoea", one removes the links into the node "Has lung cancer" and performs normal inference with "Has lung cancer = 1" and "Has lung cancer=0" . Of course this is a virtual experiment since the actual experiment is neither ethical nor possible!



$$P(A)P(S)$$

$$x\ P(T|A)\ \ P(LC)\ \ P(B|S)$$

$$x\ P(ToC|\ T,\ LC)P(X|ToC)P(D|ToC,\ B)$$

Causal Effect

- The joint probability is

$$P(LC)P(A)P(S)P(T|A)P(B|S)P(\mathit{ToC}|T,LC)P(X|\mathit{ToC})P(D|\mathit{ToC},B)$$

- The conditional probability is

$$P(D|do(LC)) =$$

$$\sum_{A,S,T,C,B,\mathit{ToC}} P(A)P(S)P(T|A)P(B|S)P(\mathit{ToC}|T,LC)P(D|\mathit{ToC},B)$$

and this is not the same as $P(D|LC)$, since $P(LC|S)$ is missing

Observing Causal Effect

- If I randomly manipulate LC, and then observe D, then I can estimate $P(D|do(LC))$ from this data. This is called a randomized controlled study
- If I simply observe LC and D in patients, then in general $P(D|do(LC)) \neq P(D|LC)$

Conditioning/Controlling

- The solution is to condition on variables (which are not effects of the potential cause), such that both conditional probabilities become identical
- Let's assume that I condition on S. Then

$$P(D|LC, S) =$$

$$\sum_{A,T,B,ToC} P(A)P(B|S)P(ToC|T, LC)P(D|ToC, B)$$

and

$$P(D|do(LC), S) =$$

$$\sum_{A,T,B,ToC} P(A)P(B|S)P(ToC|T, LC)P(D|ToC, B)$$

and both are the same

- This conditioning on a variable is called "controlling for a variable" or stratification

- If you do not measure and control for S, then S is called a hidden confounder
- (Alternatively one could control for B)
- We then have

$$P(D|do(LC)) = \sum_{S,B} P(D|LC,B)P(B|S)P(S)$$

Example: SNP Data with Confounders

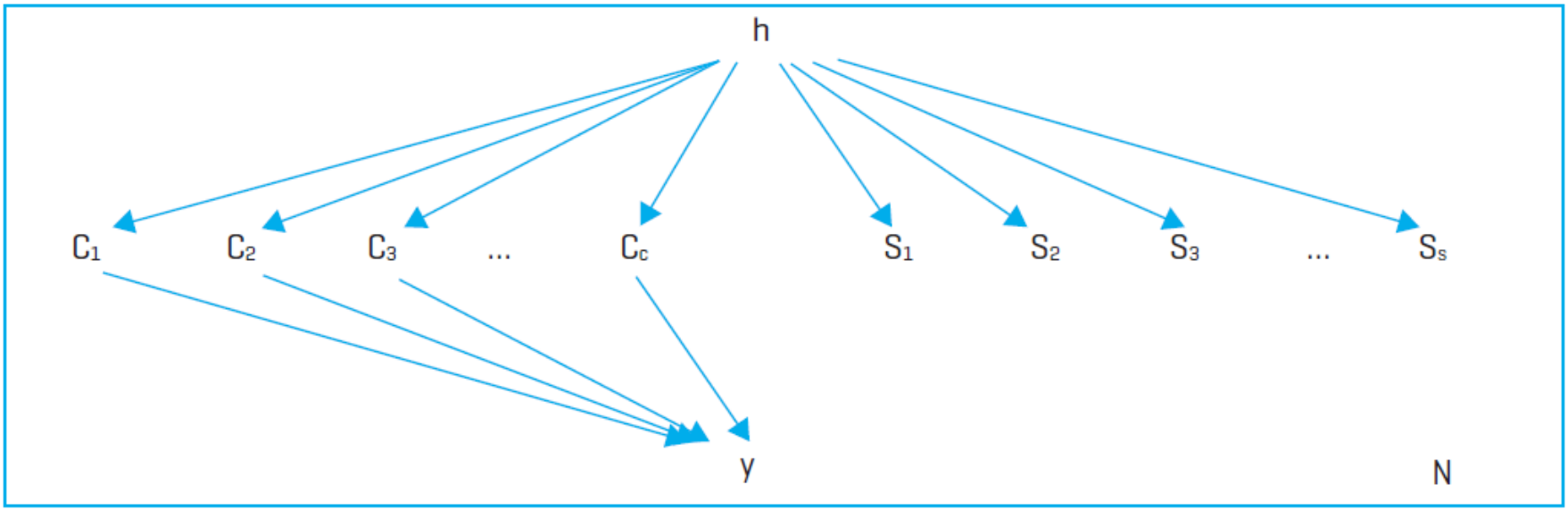
Hidden Confounders

- It might not always be known what the hidden confounders are or they cannot be measured
- With observed data one can include as many variables as possible as inputs and hope that the confounders are somehow represented (but never include possible effects of a)!

Confounding in SNP-Analysis

- h is a confounder (family identity, population identity)
- The SNPs $C_1, ..., C_N$ have a causal influence on the trait (disease) y
- The SNPs $S_1, ..., S_s$ do not have a causal influence on the trait (disease) y
- A univariate SNP analysis would have problems with confounders and the second set of SNPs might appear to have an influence as well. A reason might be that, e.g., population identity leads to correlations amount SNPs
- The first solution is to do include h in the model or to do different models for different populations/families. This can be done but typically not all confounding factors are known or can be measured
- The state-of-the-art solution solution is to do a multivariate model with all SNPS, thus the confounding influence of h might be blocked

Figure 3: A graphical model for GWAS, showing causal SNPs c_i, non-causal SNPs s_j, trait y, and unobserved confounder h for an individual. The box around the model indicates that observations are repeated for N individuals.



Secondary Use of Clinical Data

- If the decision law used by the doctor is known, all variables which were used by the doctor need to be included
- This is typically a problem since not all information that is used by the doctor is documented