UCDAVIS HEALTH

Directed Acyclic Graphs (DAGs) and Regression for Causal Inference

Susan Alber

Feb. 09, 2022

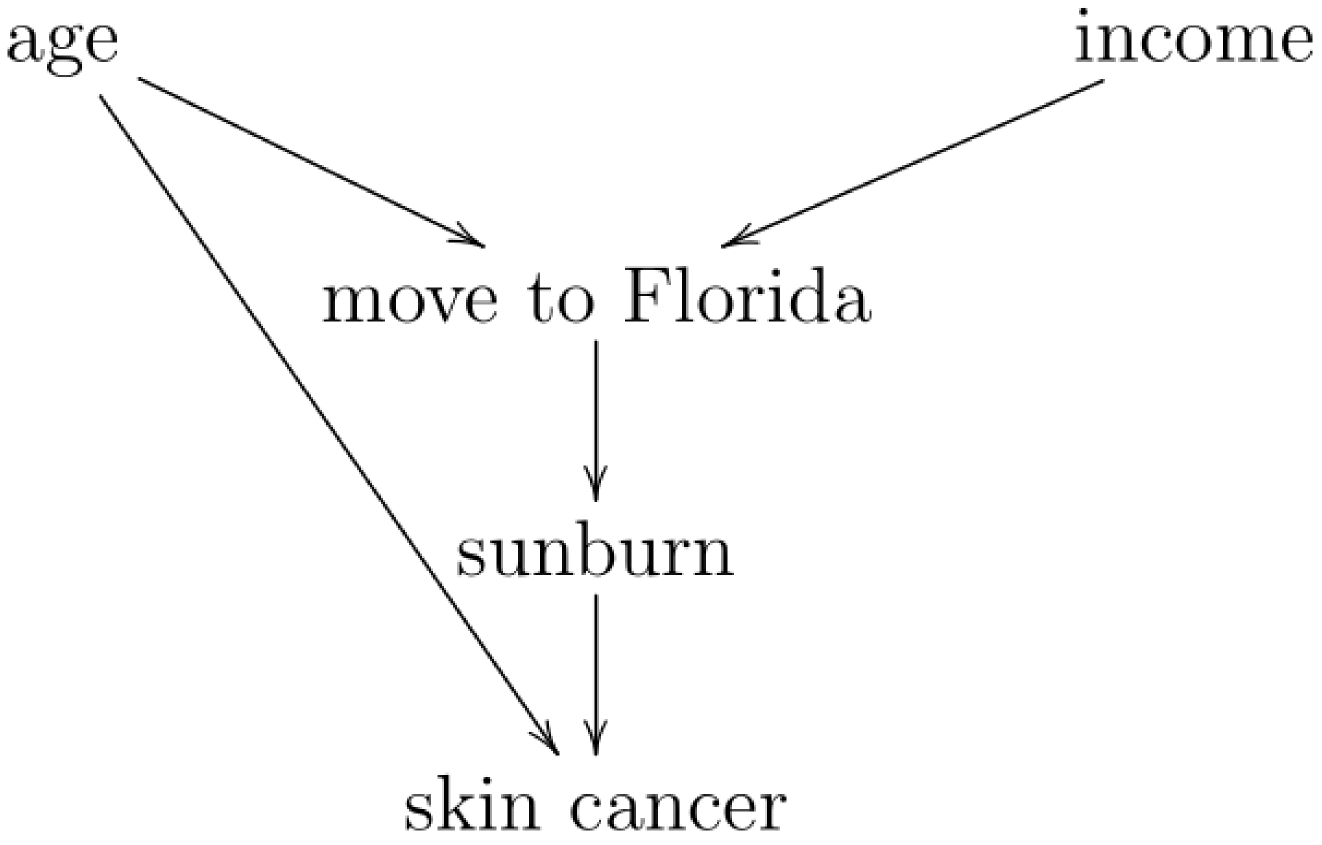
Learning goals

- Understand what a DAG is and what it is useful for.
- Understand how to construct a DAG to represent a set of assumptions about causal relationships.
- Understand how to use a DAG to determine the consequences of including variables in a regression model.
- Understand how to use a DAG to select appropriate statistical analysis methods (with focus on regression models) for estimating causal relationships in non-randomized studies

  how to identify variables that should be included in a regression model

  how to identify variables that should NOT be included in a regression model

DAG example 1



What is a DAG?

A DAG is a graph that provides a visual representation of causal relationships among a set of variables.

These causal relationships are either known to be true or more commonly are only assumed to be true.

Each arrow in a DAG represents a causal relationship. For example

Age → skin cancer

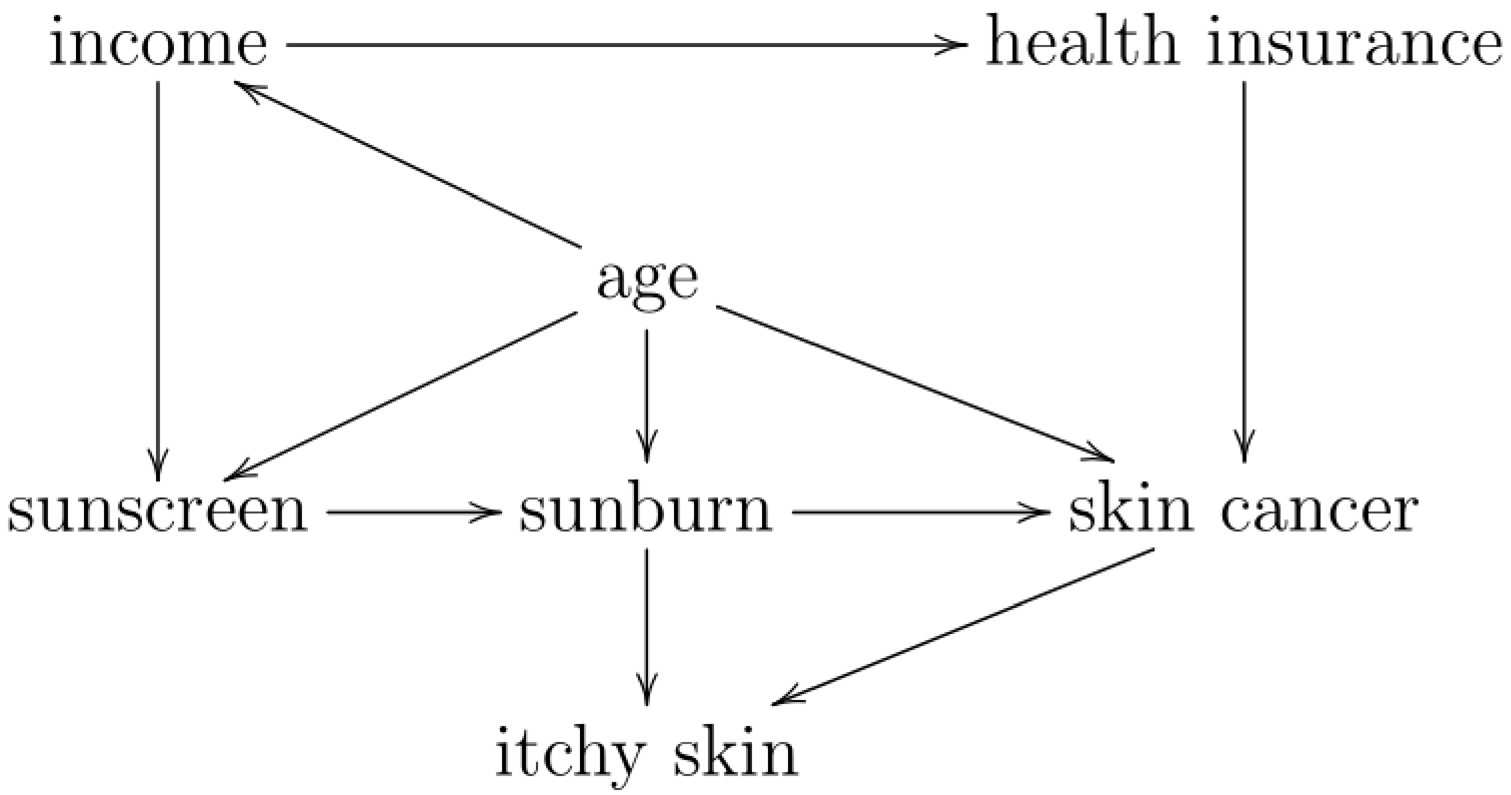
Represents that age has a causal effect of the risk of skin cancer.

Disclaimer

All the DAGs in this talk are for teaching purposes only. (They definitely will not always be the exact correct set of causal relationships.)

In fact, although we will present rules as if the DAGs are correct, it is actually pretty rare that we would really know the exact correct DAG. I will address this issue at the end of the talk.

DAG example 2



What make the graph a DAG?

D = directed (all arrows point in only a single direction)

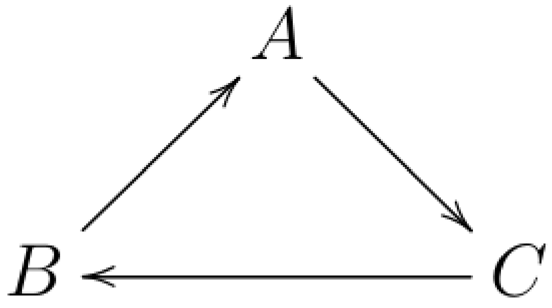
Graph has ⟶ and ⟵

But not ⟷

The direction of the arrow is the direction of causation: A → B means A causes B

A = acyclic (no sequence of arrows forms a closed loop, which would be backwards causation)

graph does not have this



What are DAGs good for?

They provide a simple visual representation of causal relationships among a set of variables. If these causal relationships are know to be true, then great. But if not, then they make it explicit what assumptions are being made.

They can be used to determining if a given pair of variables are independent.

They can be used to select variables to include in regression models, specifically with the aim of estimating causal relationships from non-randomized studies, which are subject to confounding.

Types of studies for estimating causal relationships

Randomized studies

The treatment is randomized → there is no confounding

→ causal relationship between the treatment and the outcome can be estimated using simple statistical methods that do not require adjusting for confounding

Non randomized (also called observational studies)

The treatment is not randomized → there usually will be confounding

→ Can only estimate causal relationships using special statistical methods that adjust for confounding.

This requires that we select an appropriate set of confounding variables for the adjustment, but how do we choose this set? DAGs to the rescue!

What is confounding?

Confounding (dictionary definition)

mix up (something) with something else so that the individual elements become difficult to distinguish.

From an Epidemiology textbook

“Confounding refers to a mixing or muddling of effects that can occur when the relationship we are interested in is confused by the effect of something else.” (Webb, Bain & Pirozzo)

From a Statistics textbook

“Two variables (whether explanatory variables or lurking variables) are confounded when their effects on a response variable are mixed together.” (Moore & McCabe)

Despite these ‘clear as mud’ definitions, confounding is actually a simple concept.

Consider the following situation

You wake up in the morning and find your lamp on the floor.

You know it was on the table when you went to bed.

You also find your neighbor’s cat came in through an open window during the night and is now happily sleeping on the couch.

$$\text{cat in house} \longrightarrow \text{lamp on floor}$$

Can you conclude that the cat knocked over the lamp?

But this only happened one time, so maybe it is just a coincidence.

You decide to collect some data to prove the cat is guilty.

For the next 100 days you collect the following data

day	Cat in house	Lamp on floor
1	yes	yes
2	no	no
3	no	no
.	.	.
.	.	.
100	no	no

Results

90 days where the cat did not come in

all 90 days the lamp was not on the floor

10 days where the cat came in

all 10 days the lamp was on the floor

Can we conclude that these two things (cat and lamp on floor) are not independent?

Results

90 days where the cat did not come in

  all 90 days the lamp was not on the floor

10 days where the cat come in

  all 10 days the lamp was on the floor

Can we conclude that these two things are not independent? Yes.

Can we conclude the cat knocked over the lamp?

Results

90 days where the cat did not come in

all 90 days the lamp was not on the floor

10 days where the cat come in

all 10 days the lamp was on the floor

Can we conclude that these two things are not independent? Yes.

Can we conclude the cat knocked over the lamp? Maybe not.

‘knocked over’ = causation

Guilty cat?

If we conclude the cat caused the lamp to get knocked over, what implicit assumptions are we making?

Guilty cat?

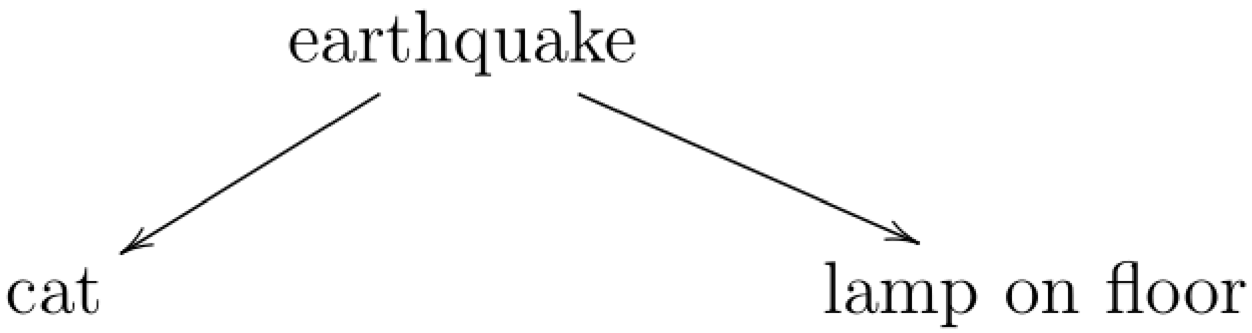
If we conclude the cat caused the lamp to get knocked over, what implicit assumptions are we making?

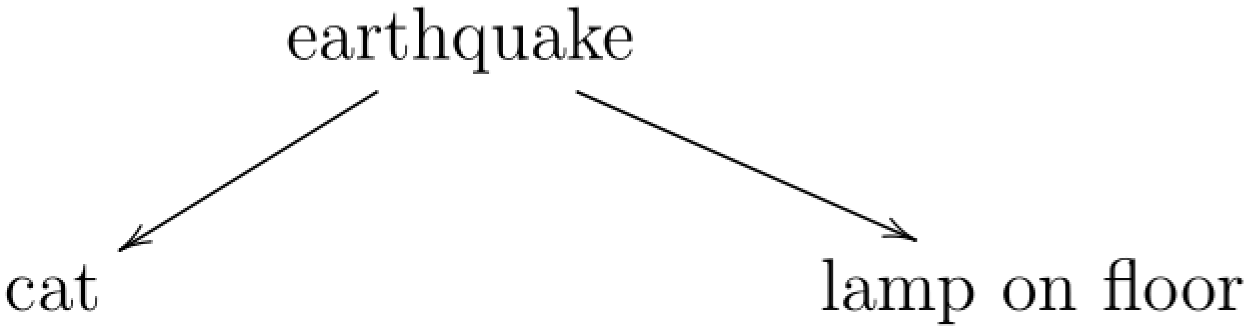
That the ONLY difference between the 10 days the cat came in and the 90 days it did not is the cat. But the days might have other differences.

But suppose now on all the days the cat came in there was an earthquake

day	Cat in house	earthquake	Lamp on floor
1	yes	yes	yes
2	no	no	no
3	no	no	no
.	.		.
.	.		.
100	no	no	no

Perhaps the cat gets scared when there is an earthquake so comes into the house, but does not knock over the lamp. The earthquake did that.





Earthquake is a common cause for cat in house AND lamp on floor.

The causal effect of cat in house on lamp on floor is confounded.

Earthquake is called a confounder.

Example

Hypothesis: Breast feeding decreases the number of infections in the first year of life.

Treatment = breast fed (yes/no)

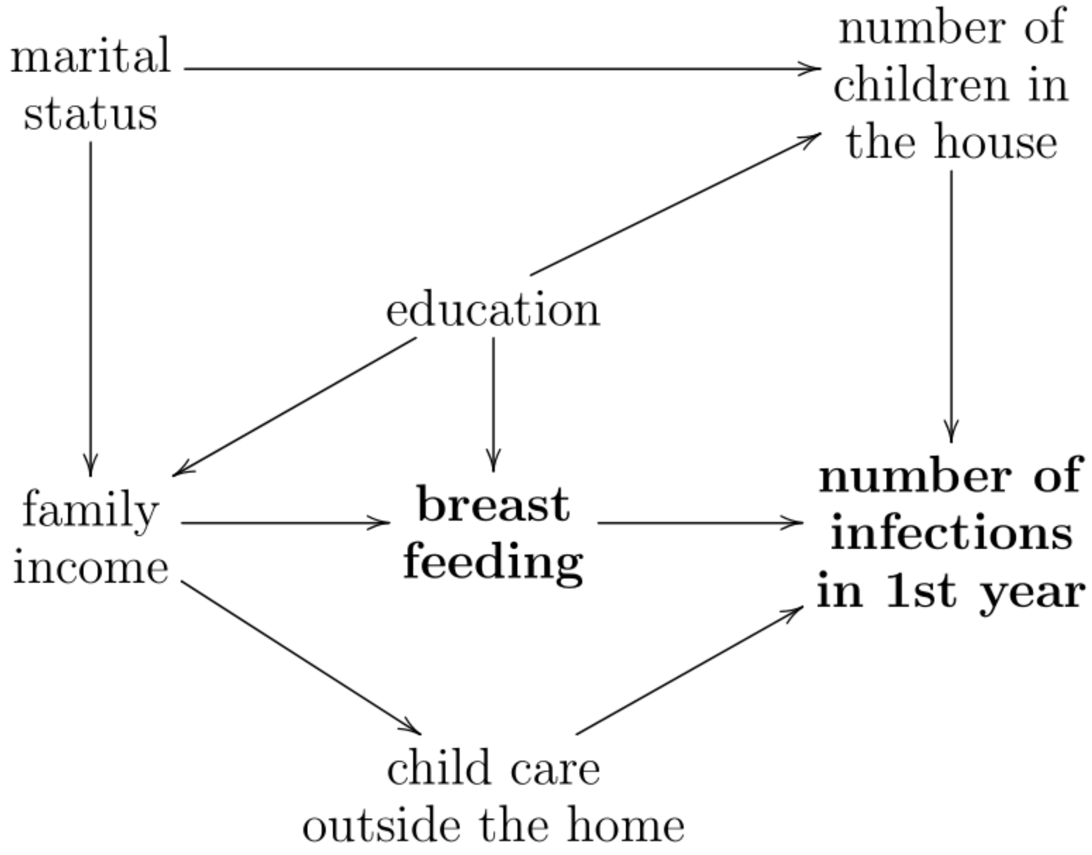
Outcome = number of infections in first year of life

Have the following variables from an observations study (i.e., not randomized)

1. Marital status
2. Family income
3. Education
4. Number of children in the house
5. Child care outside of the home

How do we choose which variables to include in a regression model?

The goal is to estimate the causal path breast feeding → number of infections.


marital status
number of children in the house
education
family income
breast feeding
number of infections in 1st year
child care outside the home

Some DAG terms

Adjacent variables are simply two variables that are next to eachother, for example

C ← B or C→ B

A path is simply a sequence of adjacent variables.

At each variable in the sequence a path is either blocked or not blocked (also called unblocked)

An unblocked path from one variable X to another variable Y is a path that goes through a set of adjacent variables where no variable in the path is blocked

Two variables are correlated (also called ‘dependent’ or ‘not independent’) if there is one or more unblocked path between them.

Blocking and Independence rules

A path is blocked by a variable C if it looks like this

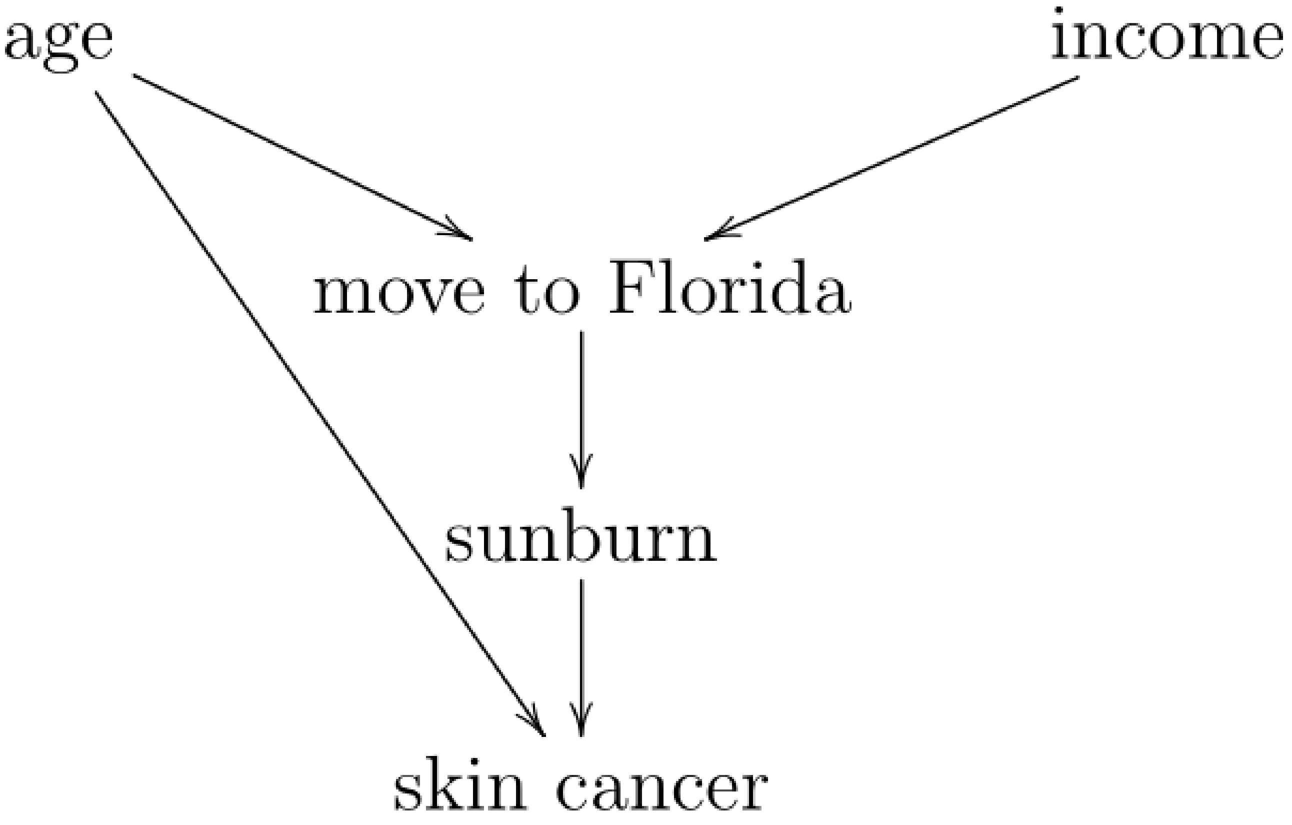
→ C ←

(C is called a ‘collider’, but the name is not important)

A path is not blocked by a variable C if it looks like an of these

→ C →
← C →
← C ←

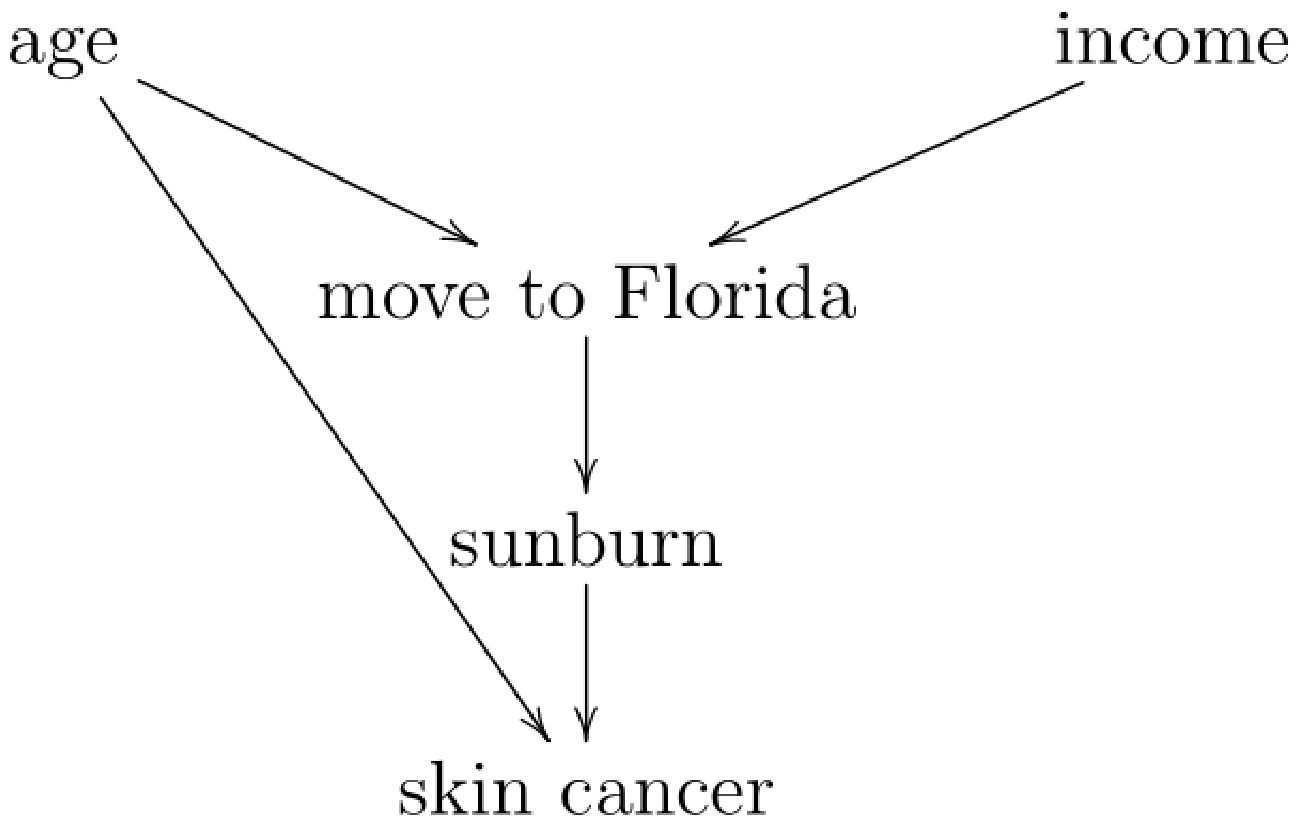
Two variables are correlated (also called ‘dependent’ or ‘not independent’) if there is one or more unblocked path between them.



unblocked paths:

Age → move to Florida → sunburn → skin cancer

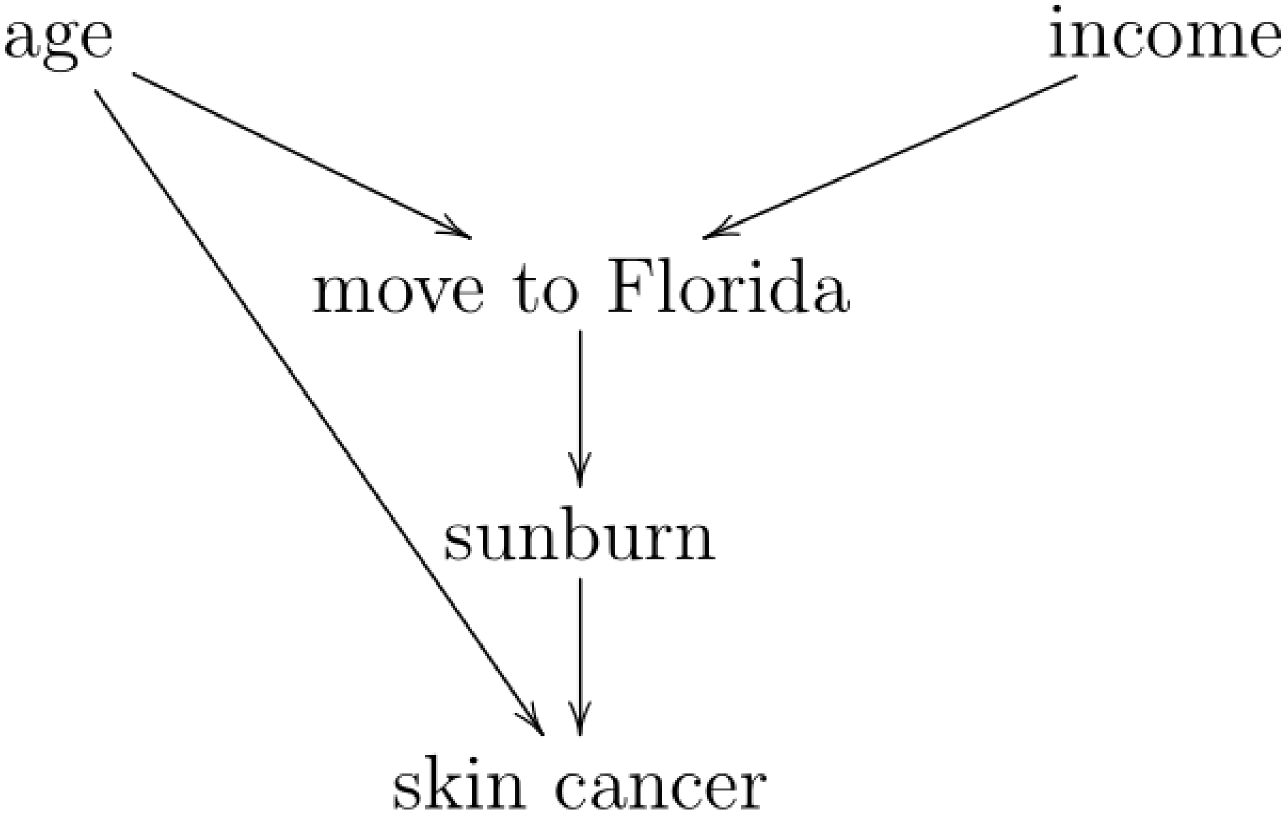
Age → skin cancer



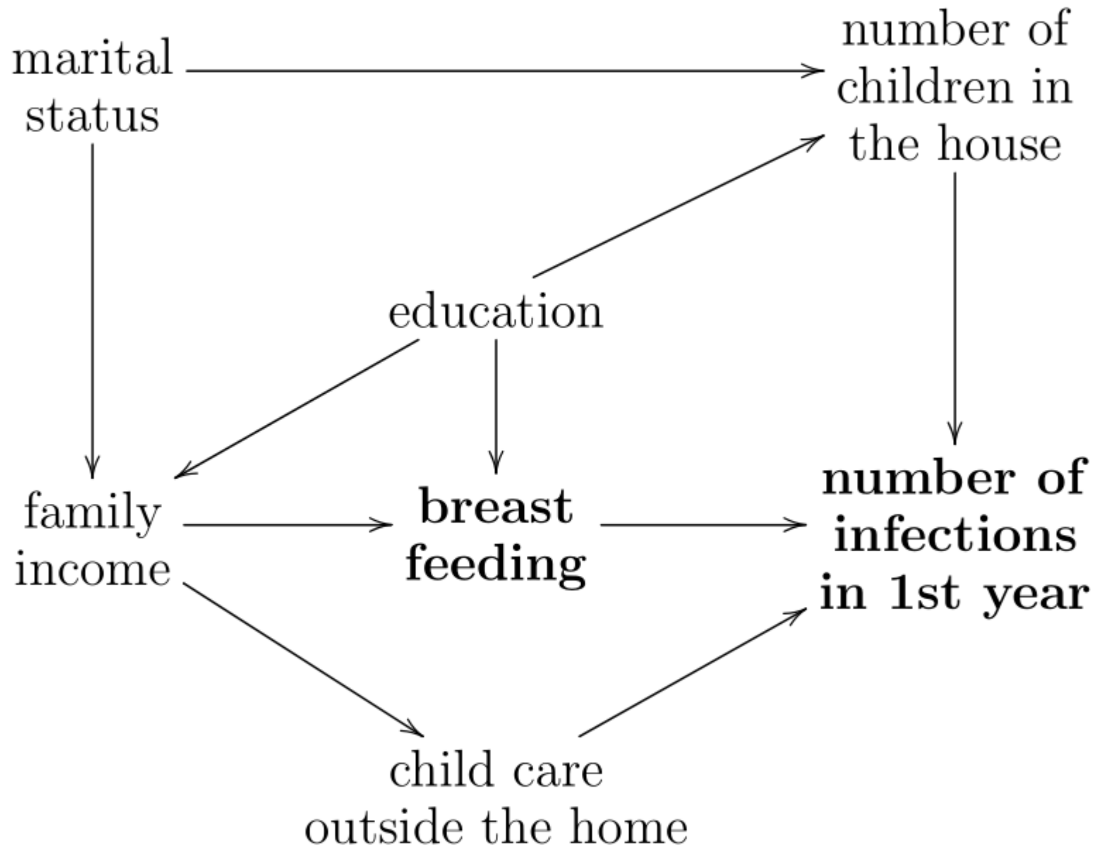
Blocked paths:

Age → move to Florida ← income   (blocked by ‘move to Florida)

Age → skin cancer ← sunburn (blocked by ‘skin cancer’)



Age and income are independent (no unblocked paths between them)


marital status
number of children in the house
education
family income
breast feeding
number of infections in 1st year
child care outside the home

Marital status and education are independent

Types of unblocked paths

We are focused on unblocked paths between a treatment X and an outcome Y

An unblocked front-door path from X to Y starts like this X →

An unblocked back-door path from X to Y starts like this X ←

Unblocked front-door paths from X to Y are causal

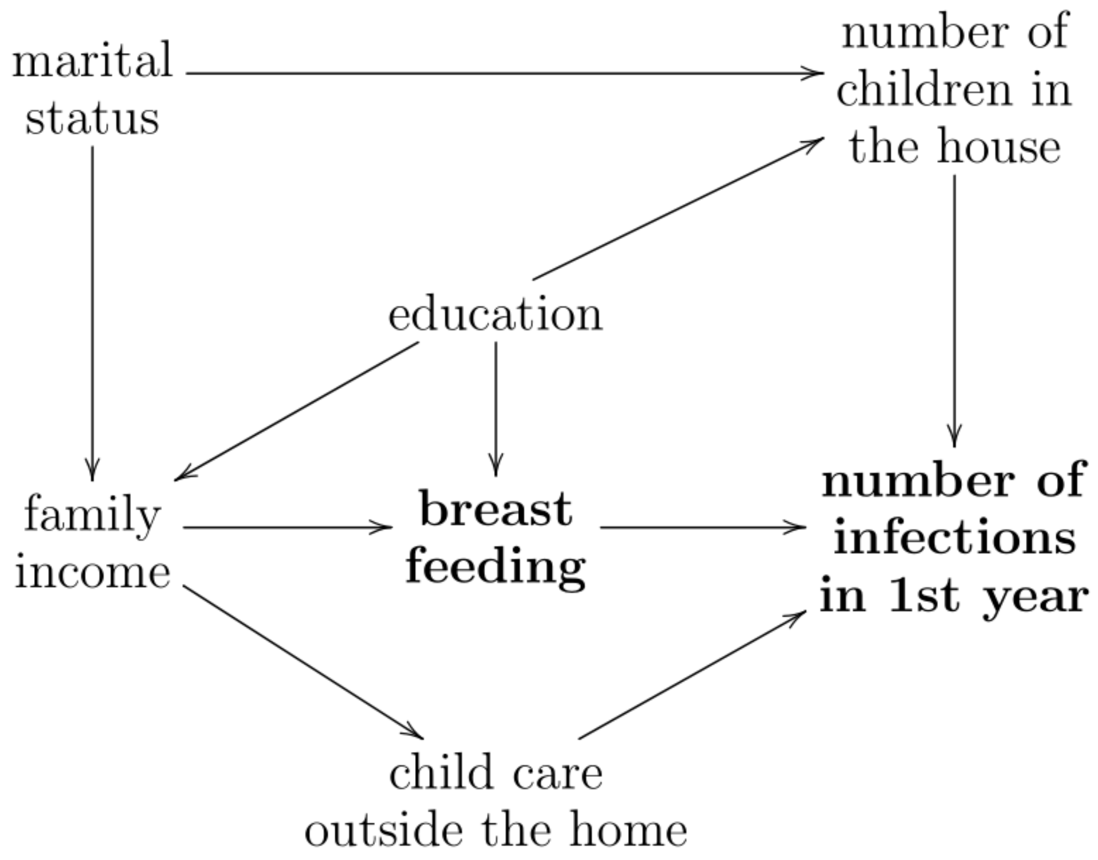
Unblocked back-door paths from X to Y are confounding

X = treatment

Y = outcome

The causal effect of X on Y is not confounded if there are no unblocked back-door paths from X to Y

The causal effect of X on Y is confounded if there is at least on unblocked back-door path from X to Y


marital status
number of children in the house
education
family income
breast feeding
number of infections in 1st year
child care outside the home

The effect of breast feeding on number of infections is confounded because there are several back-door paths from breast feeding to number of infections.

Correlation = causal effect + confounding effect

X = treatment

Y = outcome

$$\text{Correlation between X and Y} = \underbrace{\text{(unblocked \underline{front-door} paths from X to Y)}}_{\text{causal}} + \underbrace{\text{(unblocked \underline{back-door} paths from X to Y)}}_{\text{confounding}}$$

If there are no unblocked back-door paths from X to Y then there is no confounding,

Because the second term will be 0 so the correlation will equal the causal part

We can effectively eliminate confounding by including in a regression model a set of variables (called ‘confounders’) such that conditional on the set there are no unblocked back-door paths from treatment (breast feeding) to outcome (number of infections).

The regression model effectively subtracts out the confounding part leaving us with just the casual relationship between treatment and outcome.

The variables we need is not a single unique set. There typically will be many different sets that will work.

Conditioning on a variable reverses the blocking rules

A path is blocked by a variable C if it looks like this

→ C ←

A path is unblocked by a variable C if it looks like an of these

→ C →

← C →

← C ←

Conditioning on a variable reverses the blocking rules

A path is blocked by a variable C if it looks like this

→ C ←

this path is unblocked conditional on C

A path is unblocked by a variable C if it looks like an of these

→ C →

← C →

← C ←

these paths are blocked conditional on C

Here ‘conditional on C’ effectively means include in a regression model.

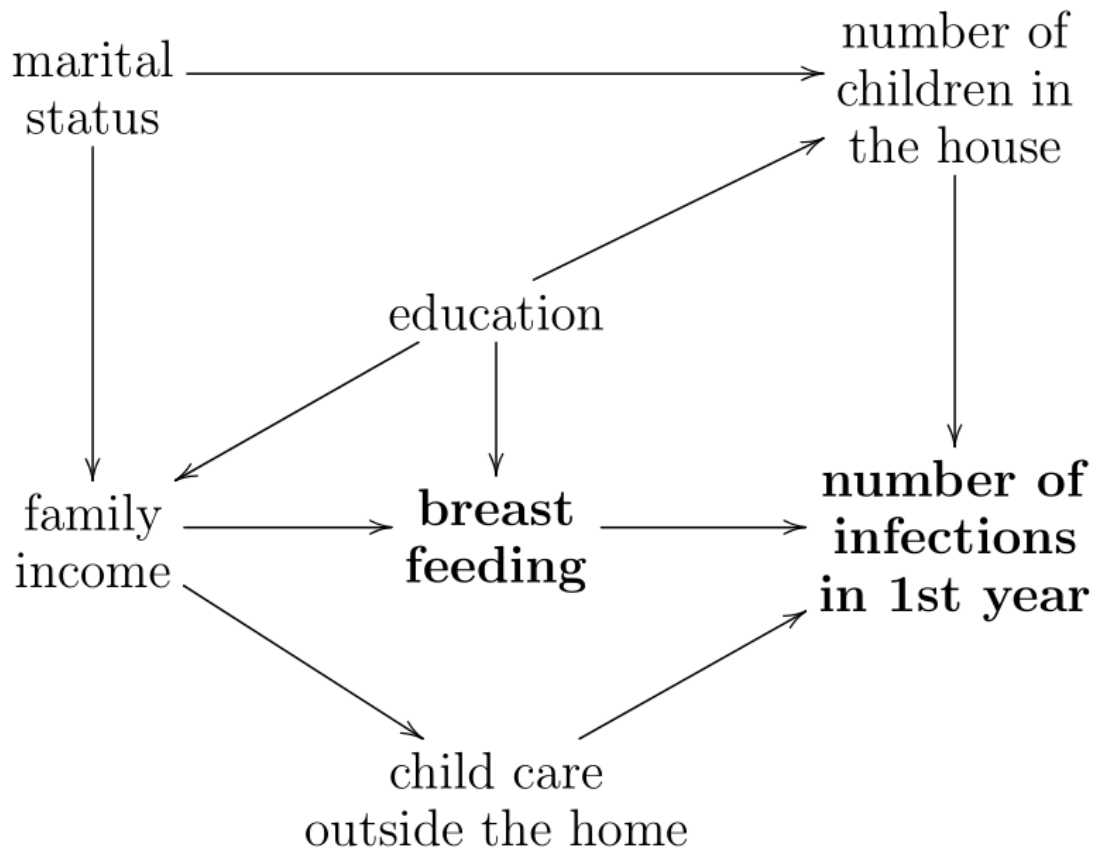
Same No confounding rule conditional on a set of variables

Unconditional on any other variables

The effect of treatment on the outcome is not confounded if there are no unblocked back-door paths from treatment to outcome.

conditional on a set of variables

Conditional on a set of variables, the effect of treatment on the outcome is not confounded if there are no unblocked back-door paths from treatment to outcome.


marital status
number of children in the house
education
family income
breast feeding
number of infections in 1st year
child care outside the home

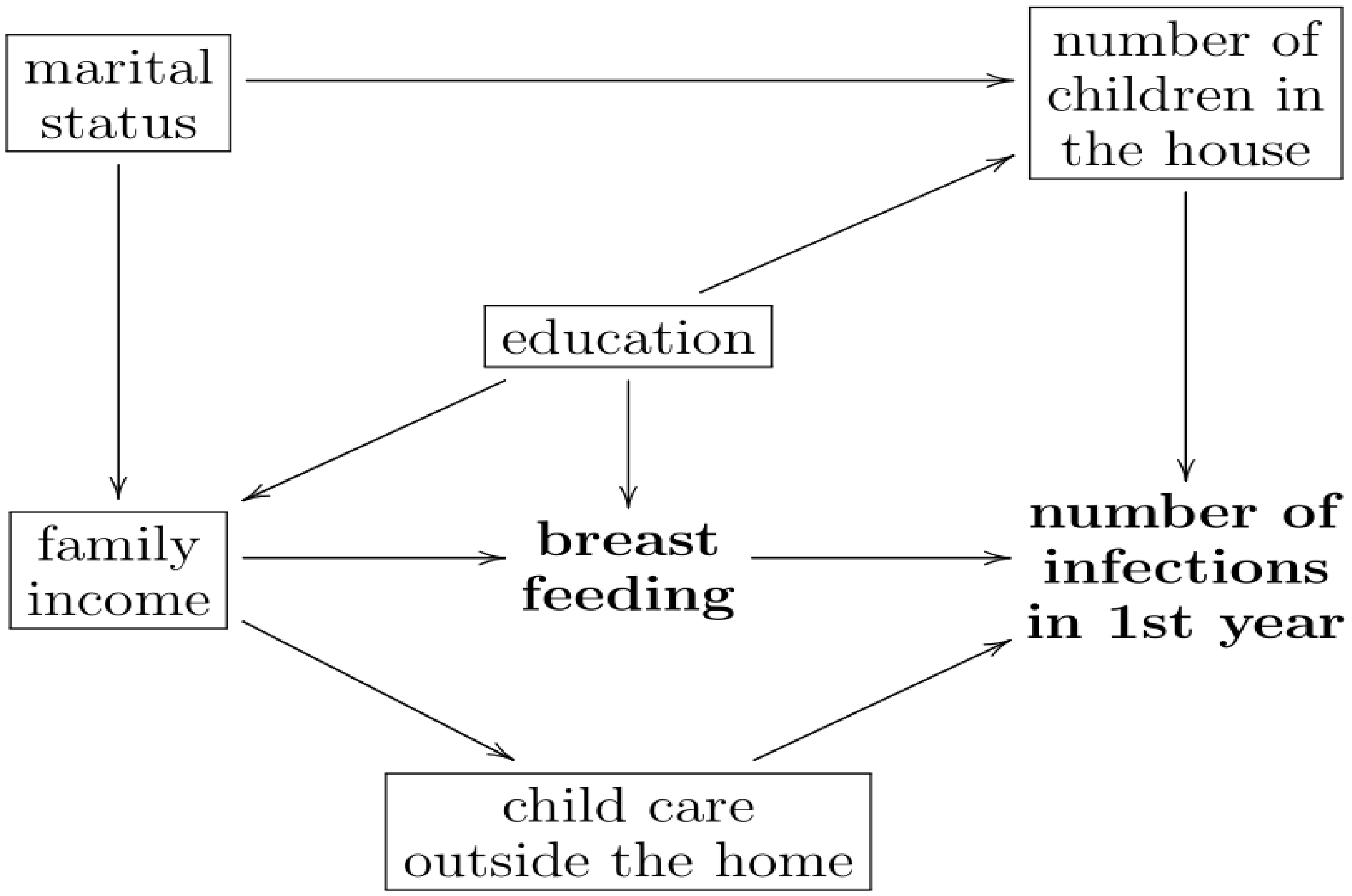
To remove all confounding, we need to condition on variables that will block all back-door paths from breast feeding to number of infections.

A regression model that includes all 5 of these will effectively remove all confounding.

1. Marital status
2. Family income
3. Education
4. Number of children in the house
5. Child care outside of the home

We will represent conditioning on a variable by putting a box around it.

(For the purpose of this talk, conditioning always means including in a regression model.)


marital status
number of children in the house
education
family income
breast feeding
number of infections in 1st year
child care outside the home

All confounding has been removed because there are no unblocked back-door paths from breast feeding to number of infections

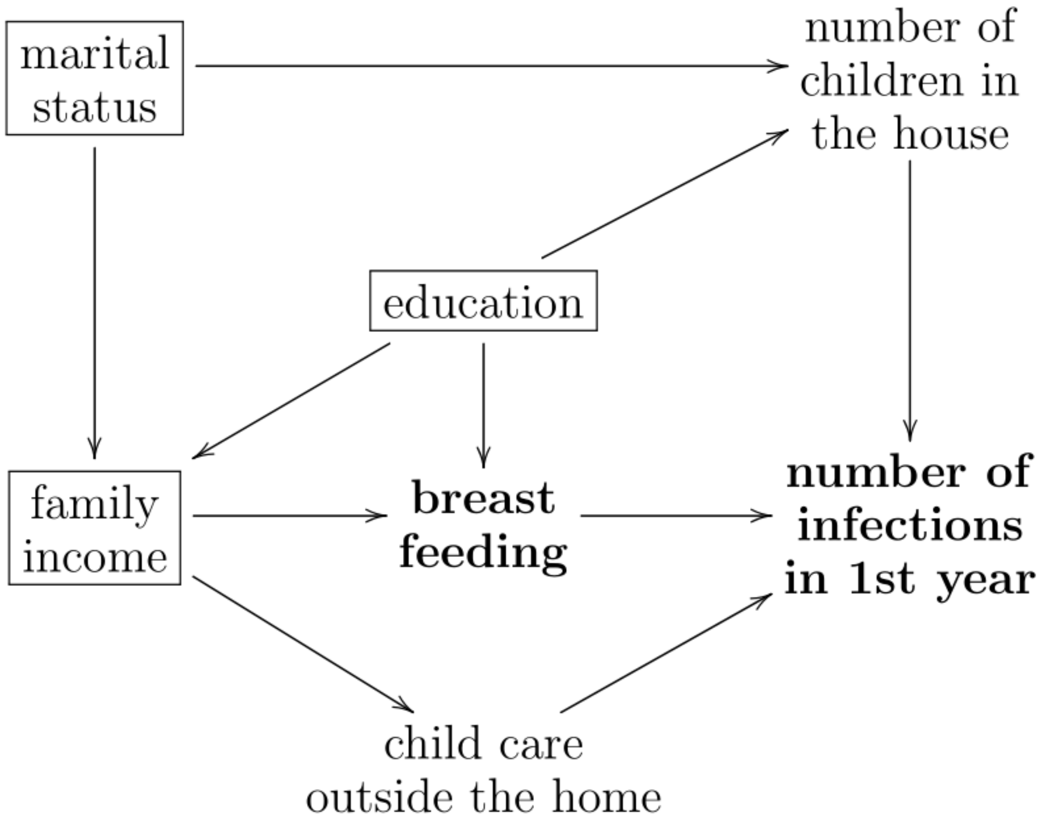
Did we really need to include all 5 variables?

No, several other sets would also work, for example

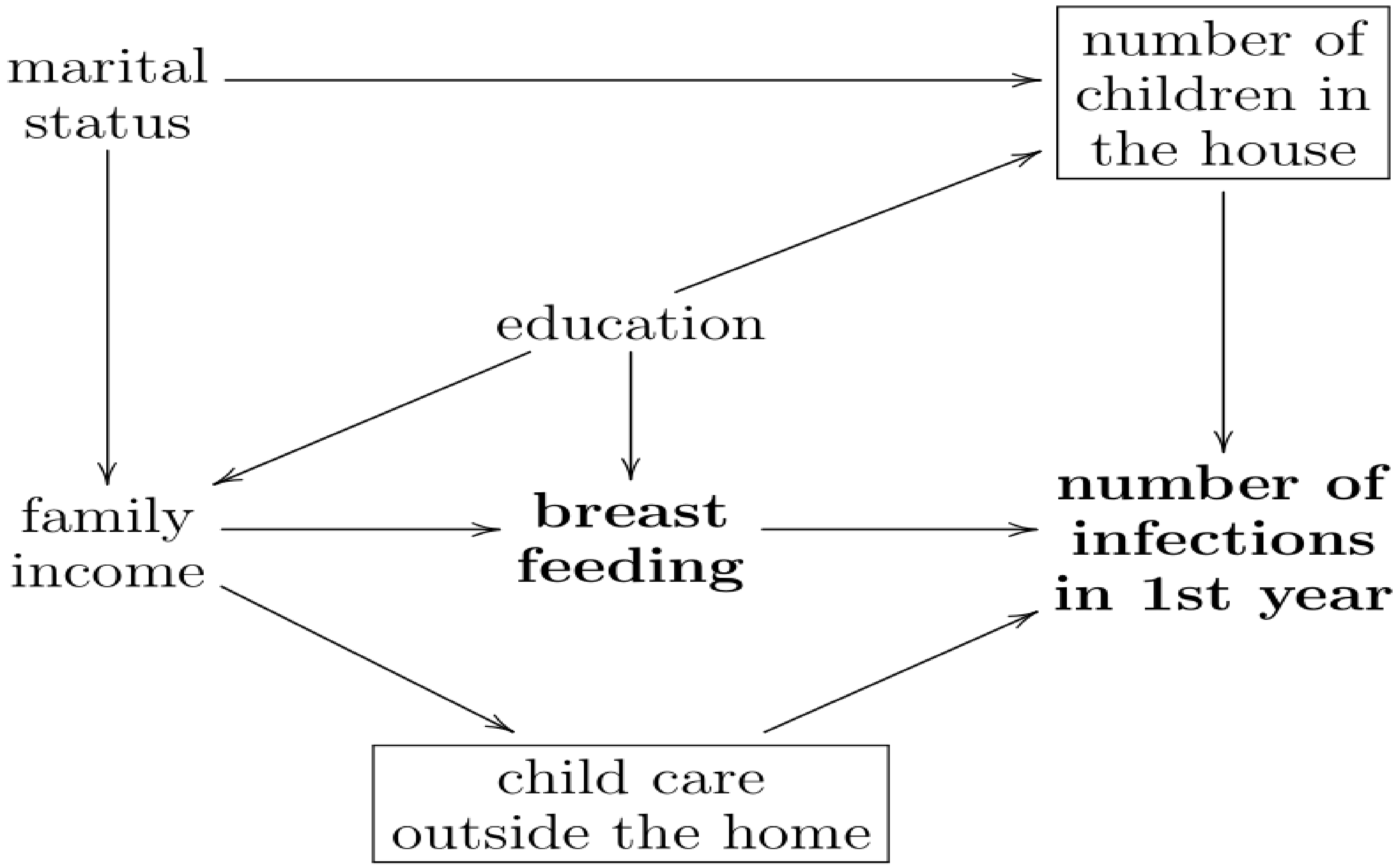
1. Marital status, education, family income
2. Number of children in house, child care outside the home

Why?

This set also removes all confounding


marital status
number of children in the house
education
family income
breast feeding
number of infections in 1st year
child care outside the home

This also works



marital status

number of children in the house

education

family income

breast feeding

number of infections in 1st year

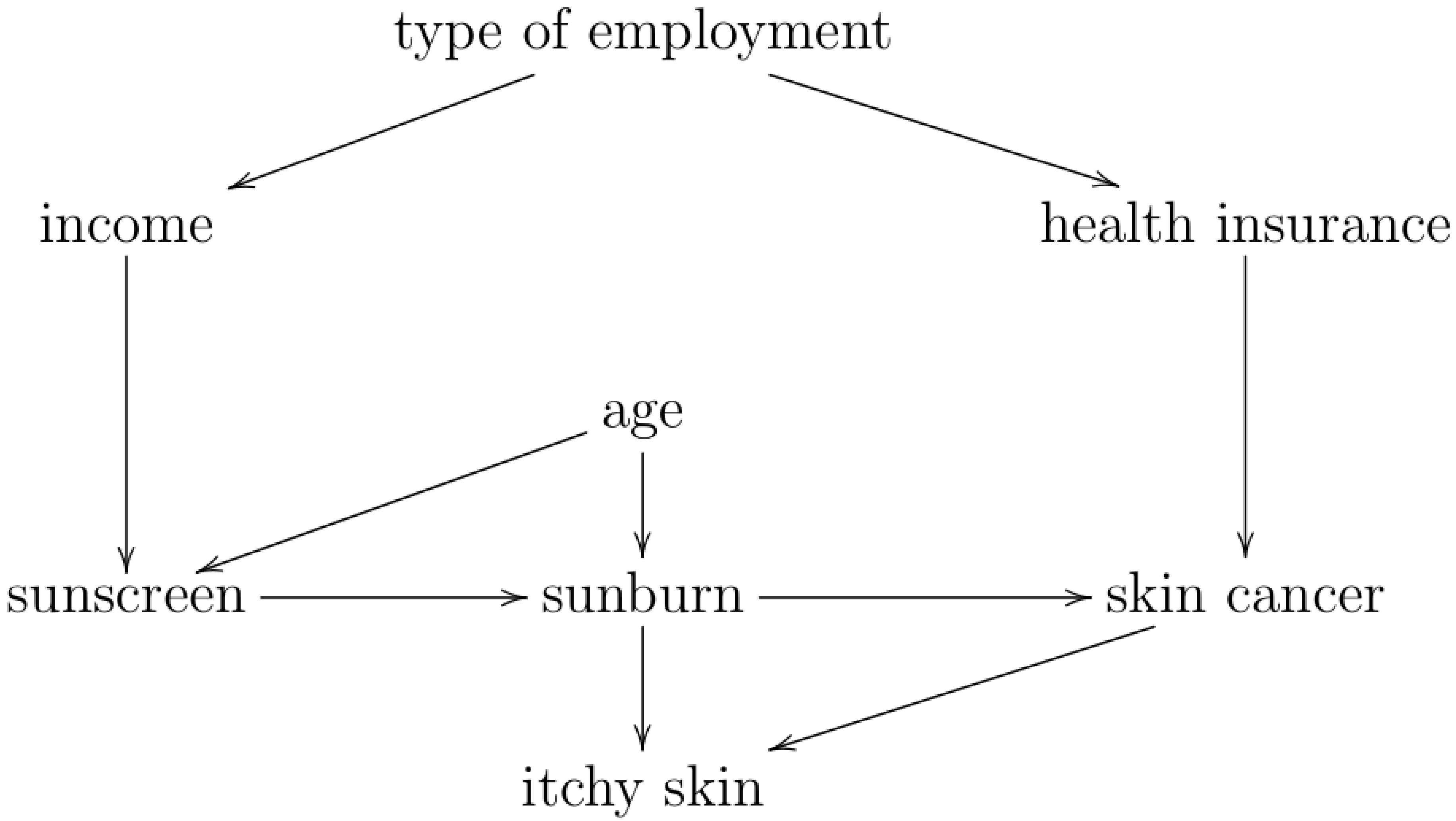
child care outside the home

Another example

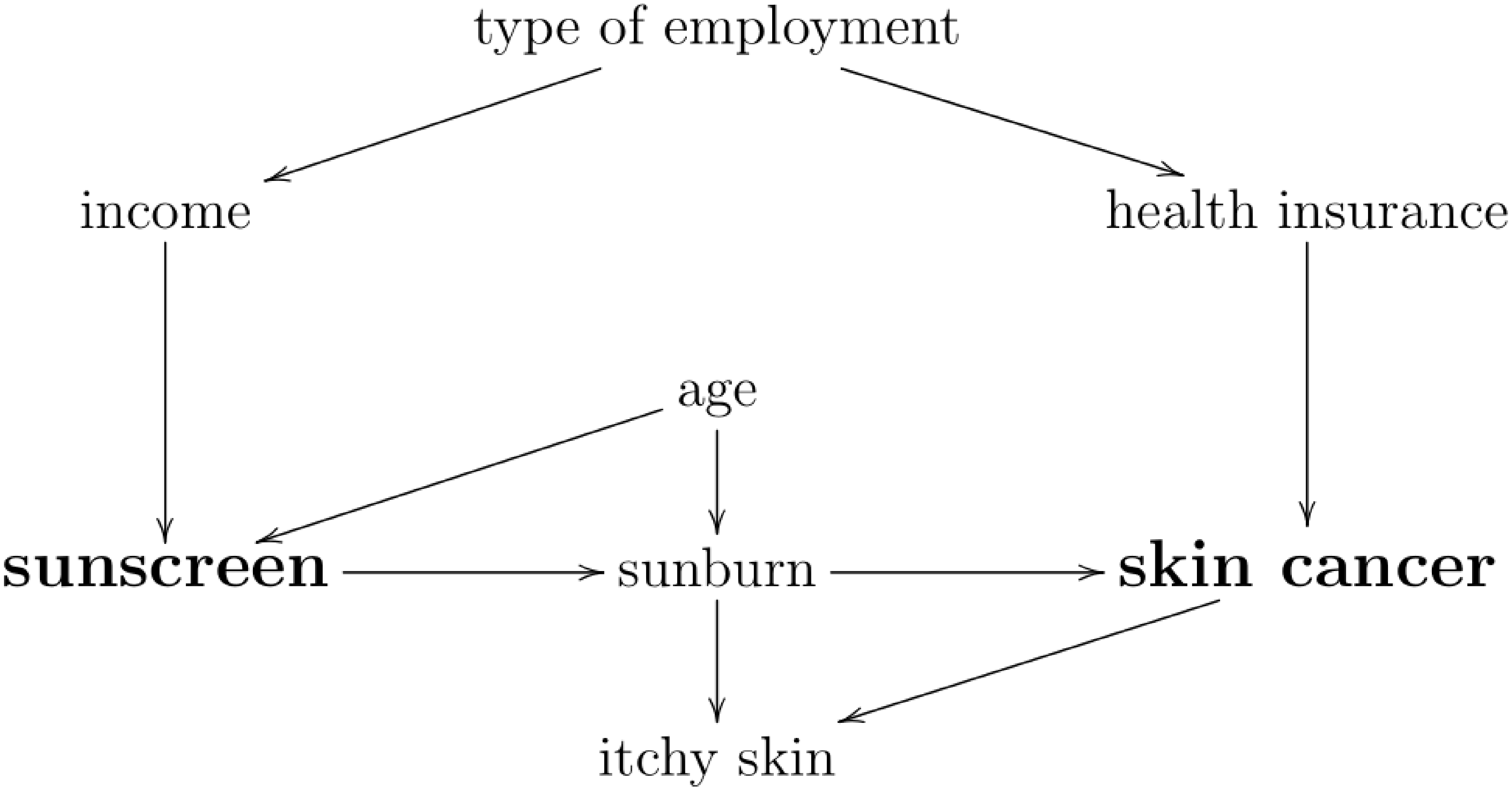
Hypothesis: sunscreen reduces the risk of skin cancer

Variables

1. Type of employment
2. Income
3. Age
4. Sunburn
5. Itchy skin

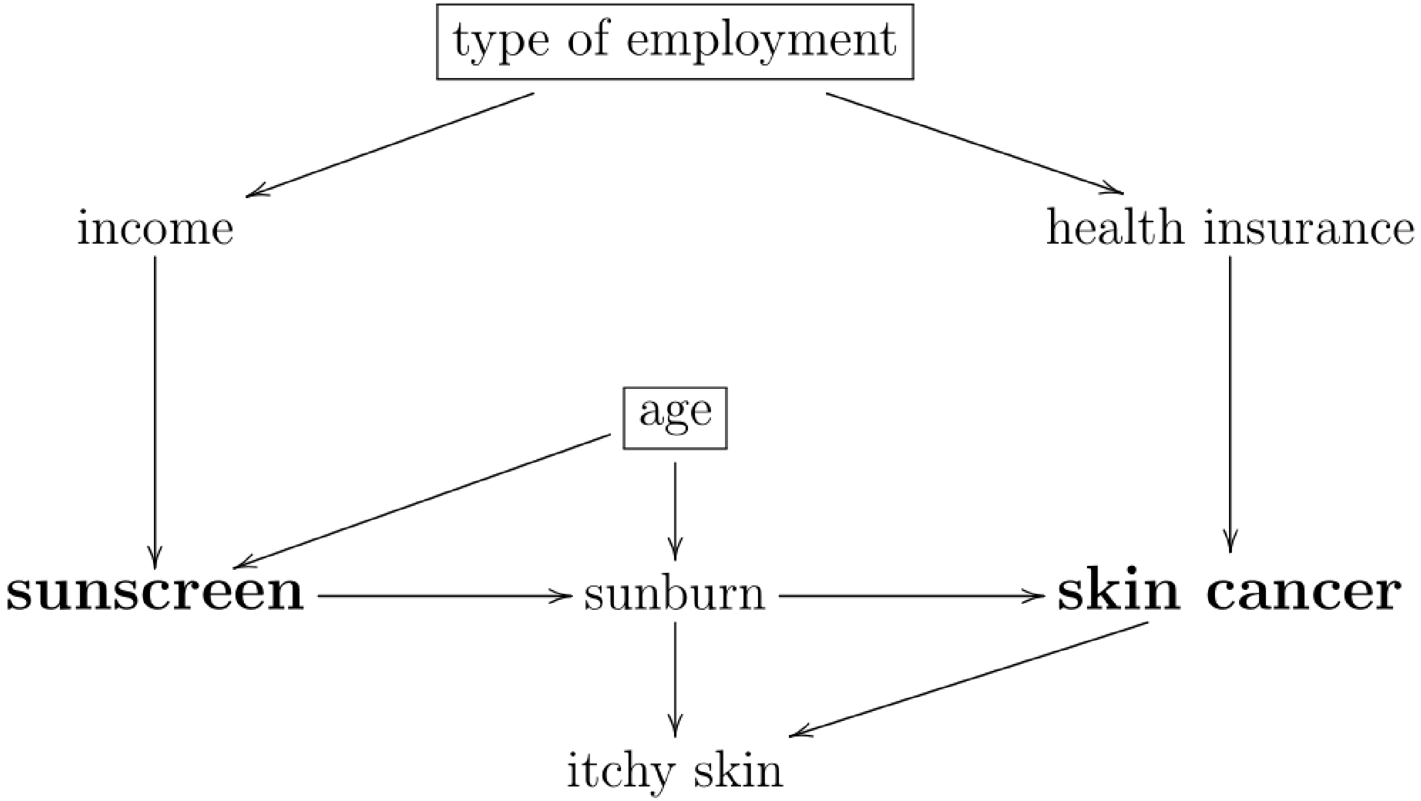


The causal effect of sunscreen on (reduction) of skin cancer is confounded



Type of employment and age are confounders

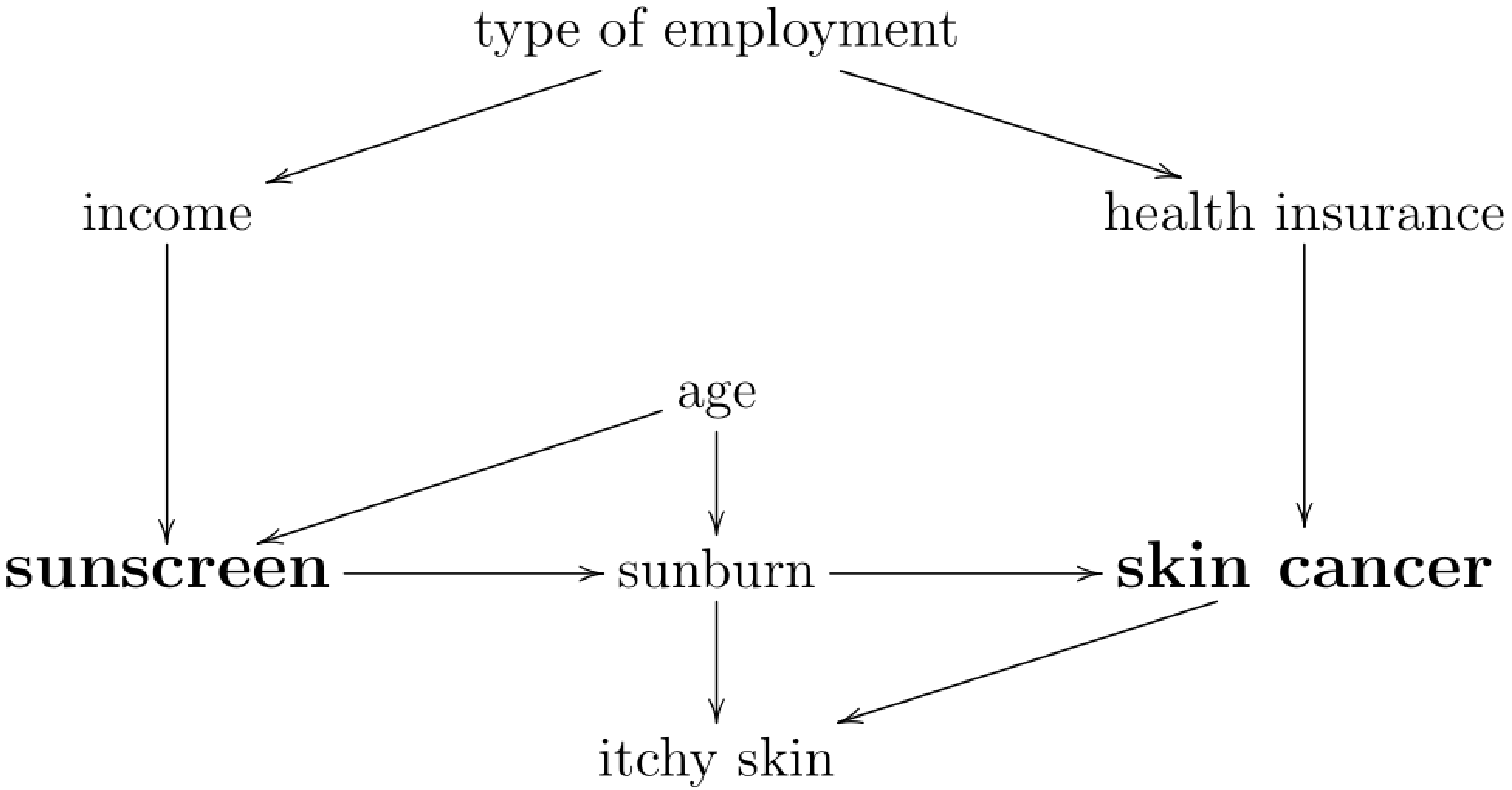
Conditioning on age and type of employment removes confounding



Notice the confounding path that goes through type of employment looks like this

Sunscreen ← income ← type of employment → health insurance → skin cancer

If we did not measure type of employment, but measured health insurance or income, either of those would work (together with age) to remove confounding.



Conditioning on age and (either type of employment or income or health insurance) removes confounding

Notice the shape of our two confounding paths

Sunscreen ← income ← type of employment → health insurance → skin cancer

And

Sunscreen ← age → sunburn → skin cancer

Both have similar pattern in the middle

← Type of employment →

← Age →

I define these as primary confounders

A variable C is a primary confounder if

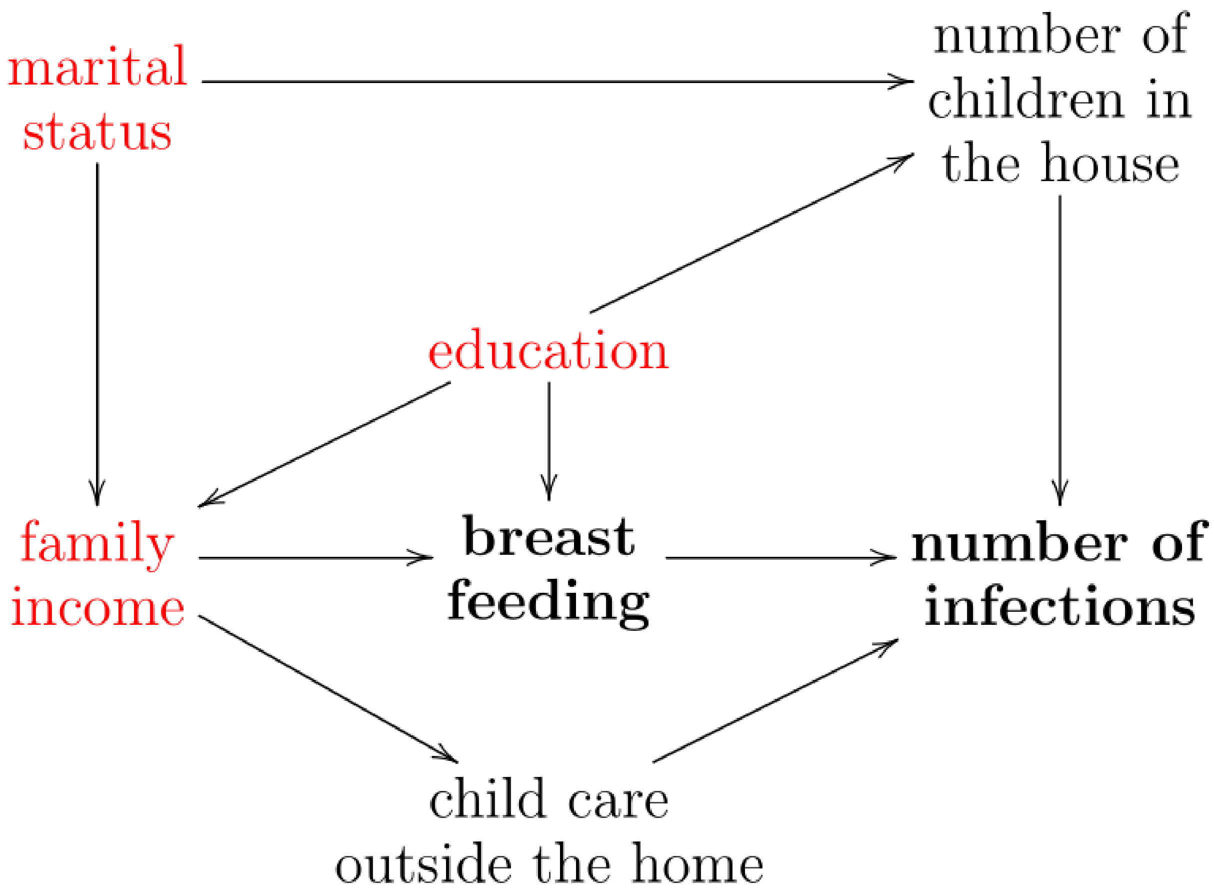
There is an unblocked front-door path from C to the treatment

And an unblocked front-door path from C to the outcome

Every unblocked back-door path from treatment to outcome has exactly one primary confounder and conditioning on all primary confounders will remove confounding.

In order to identify all confounding we only need to include the primary confounder.

The red are the primary confounders



So far we have looked at ways where conditioning on variables (equivalently, including in regression models) removes confounding.

But can conditioning on variables possibly cause confounding? Or other types of bias?

Answer is yes (can cause confounding) and yes (can cause other types of bias).

Conditioning on a variable induces (rather than removes) confounding if it unblocks a back-door path from treatment to outcome that was previously blocked.

This can only happen if the variable you condition on looks like this

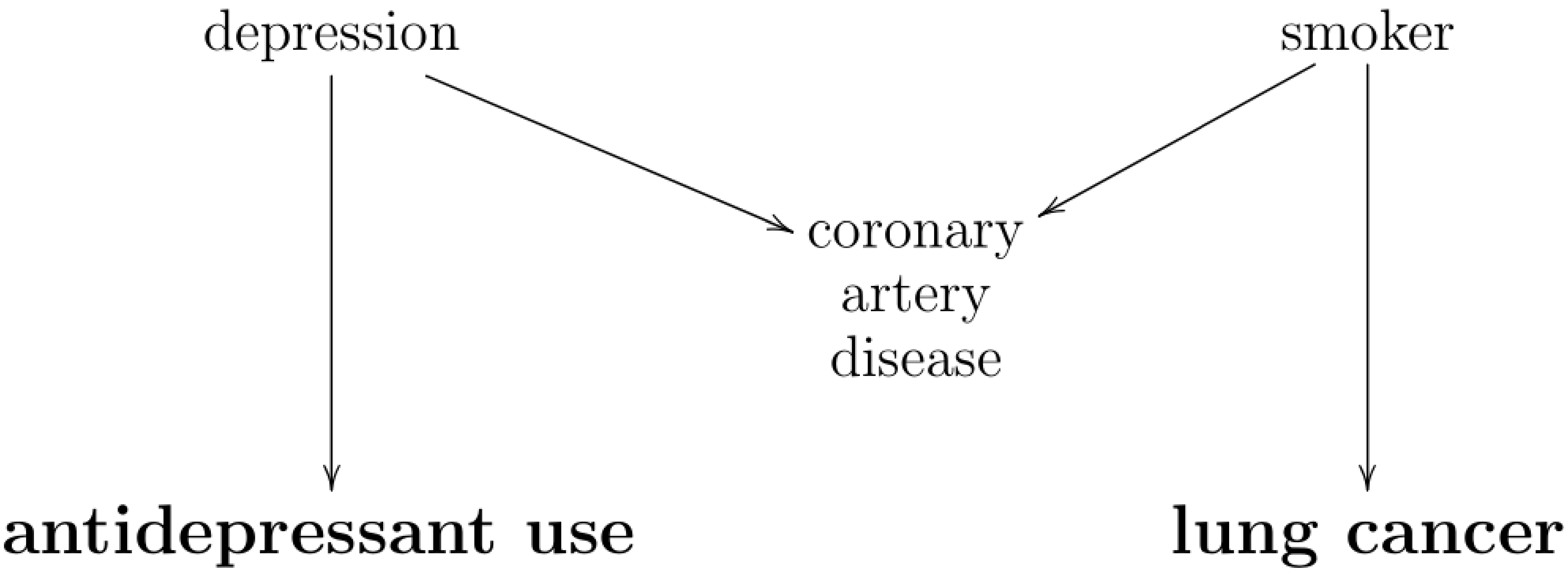
→ C ←

Recall C is called a collider. But conditioning on a collider does not always induce confounding. It only induces confounding if it unblocks a back-door path from treatment to outcome.

Recall the Blocking rules

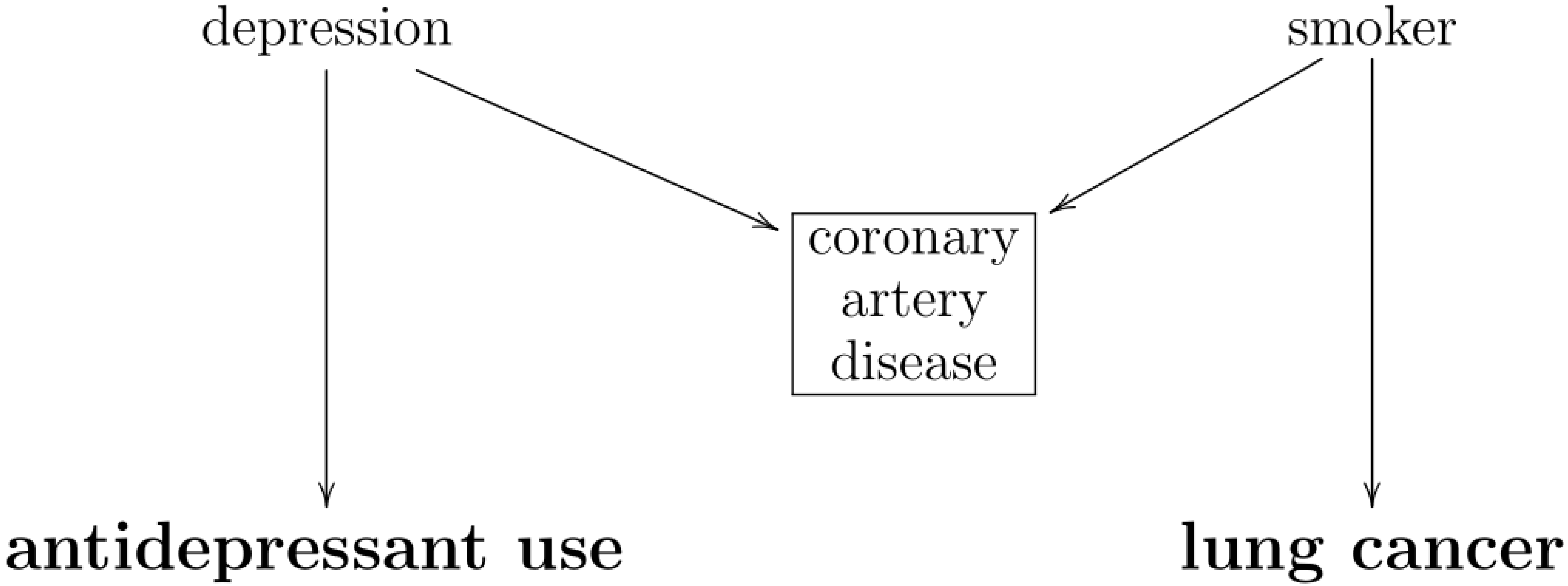
→ C ← Is blocked unconditional on C and unblocked conditional on C

Does antidepressant use increase risk of lung cancer?



No confounding, but what we if condition on coronary artery disease?

Conditioning on COD induces confounding

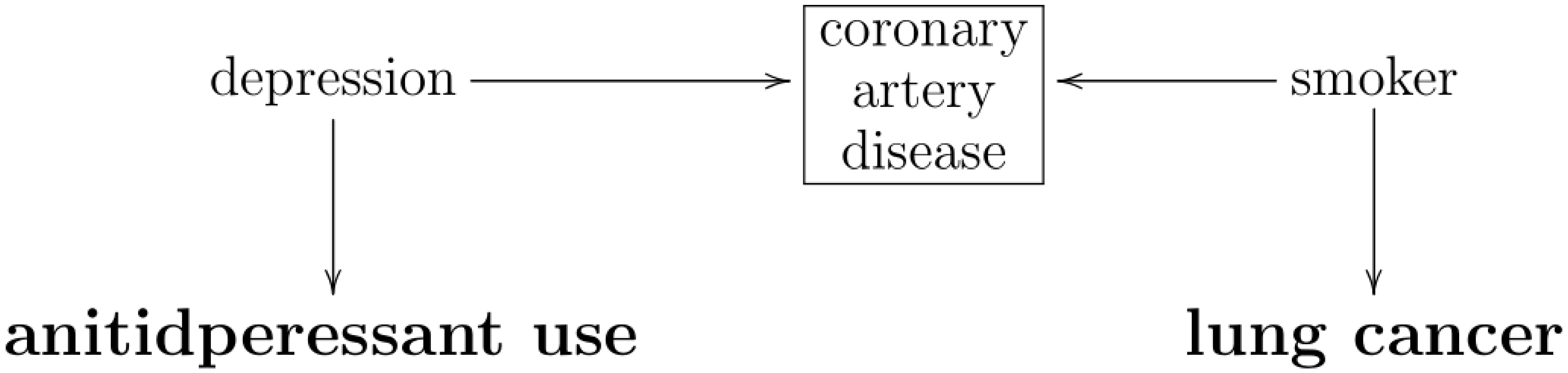


Now we have confounding because there is an unblocked back-door path

This is often called M-bias because the graph has the shape of an M.

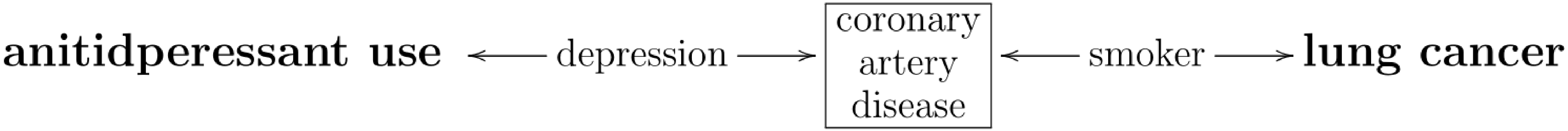
But I think that is pretty stupid since the way we choose to shape the graph is arbitrary.

Same exact graph



What is the bias now? Upside down U-bias?

Same exact graph



anitidperessant use ← depression → coronary artery disease ← smoker → lung cancer

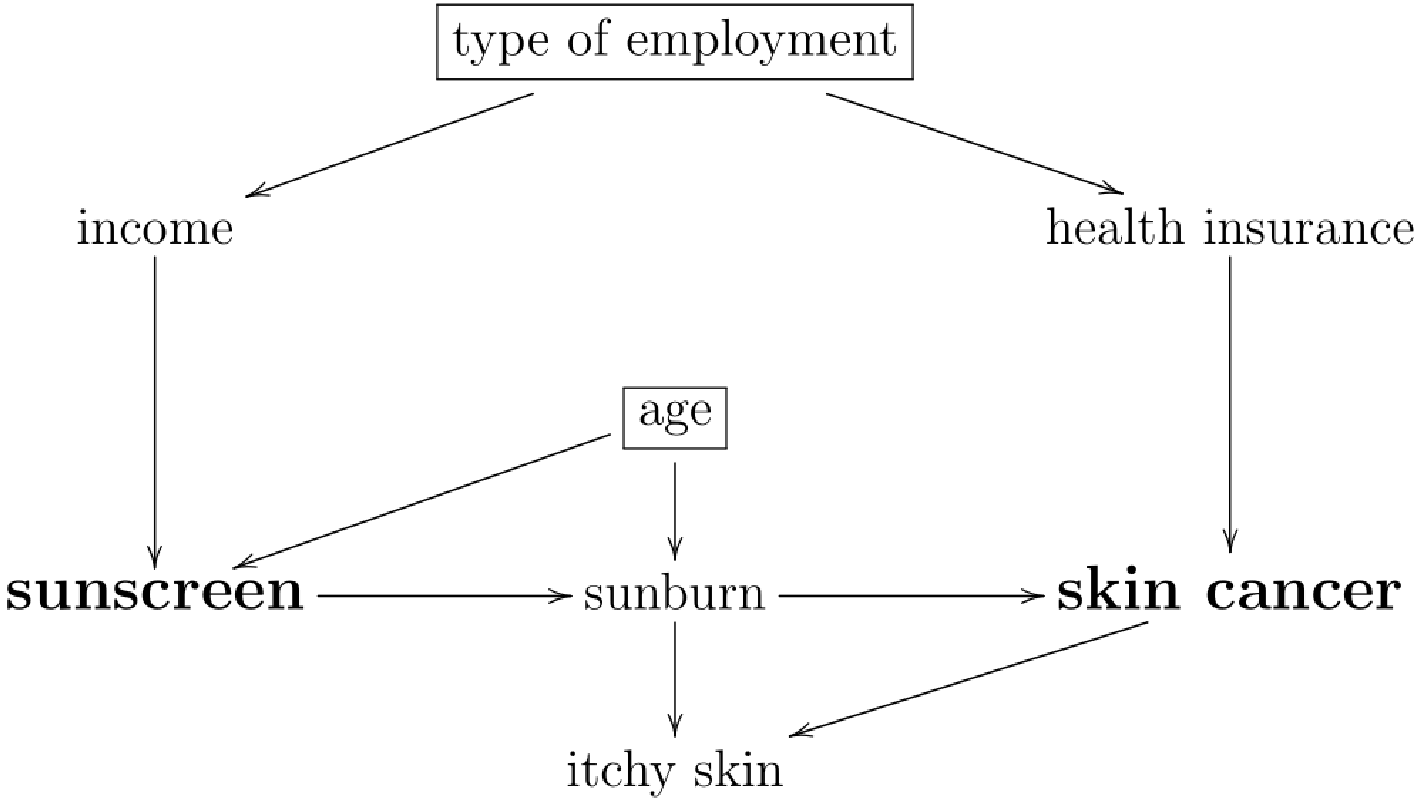
How about now? Still want to call this M-bias?

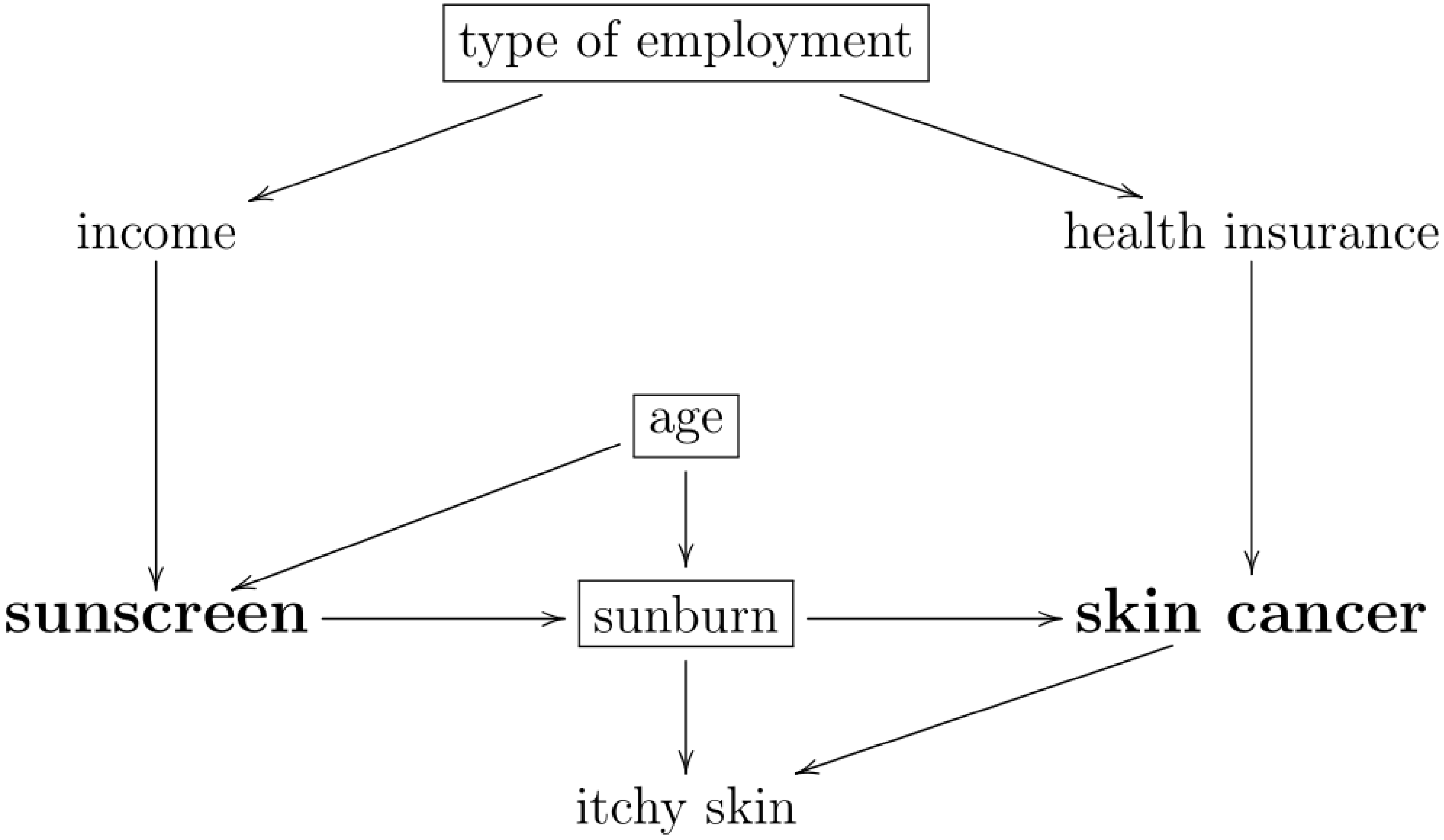
Another way to mess up the estimation of the causal relationship is to condition on something that effectively alters the strength and/or direction of the causal relationship between the treatment and the outcome.

One example of this is conditioning on something in the casual path between treatment and outcome.

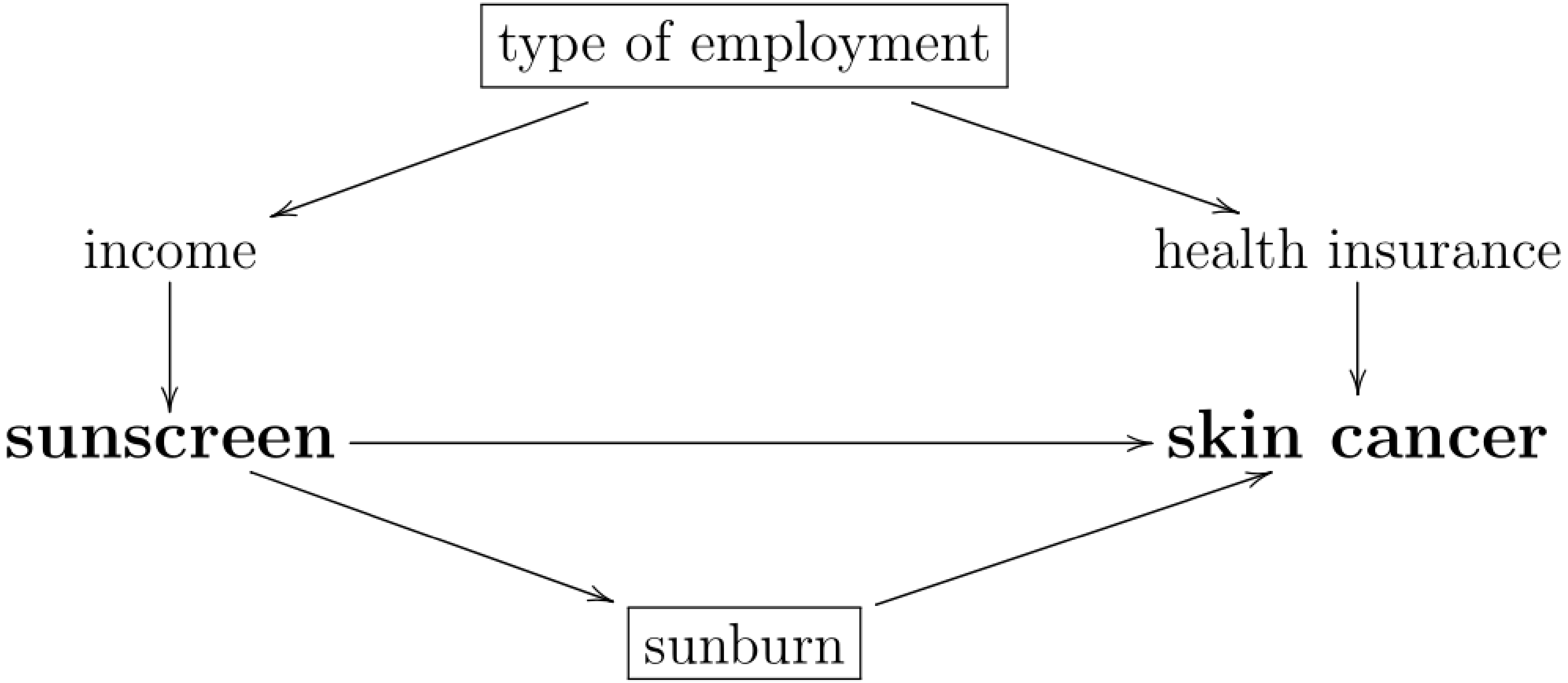
In this situation we do not induce confounding, but rather cause a different type of bias, but one that still results in a biased estimate of the causal effect.

Conditioning on age and type of employment removes confounding

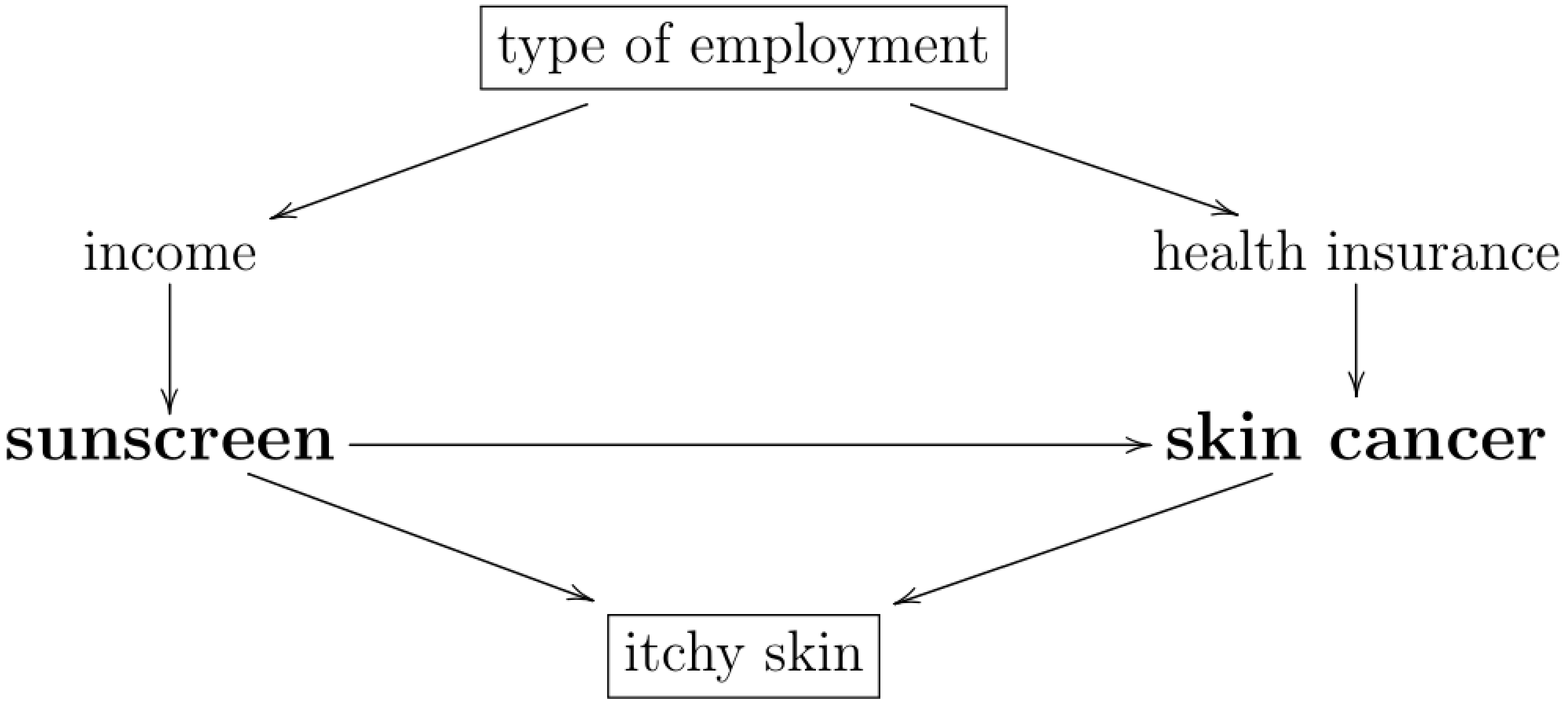




No confounding, but conditional on sunburn sunscreen and skin cancer are independent. So estimate of causal relationship is clearly biased.



Conditioning on sunburn is still causes bias even if it does not completely block the path



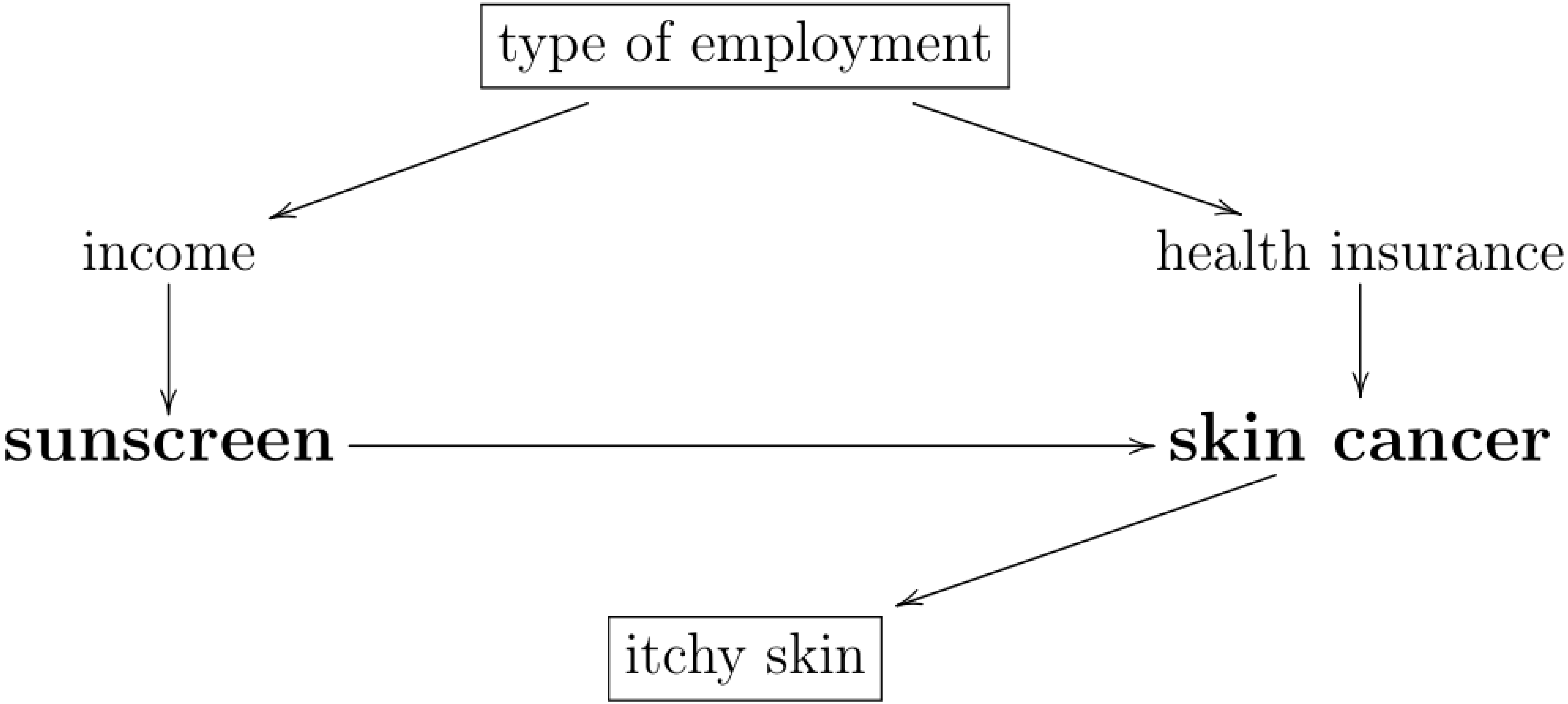
Conditioning on itchy skin causes bias. In this case it is not because it blocks the casual path. So is there a general rule for these types of bias? Yes

General rule for when conditioning on a variable C causes non-confounding bias is any time you have

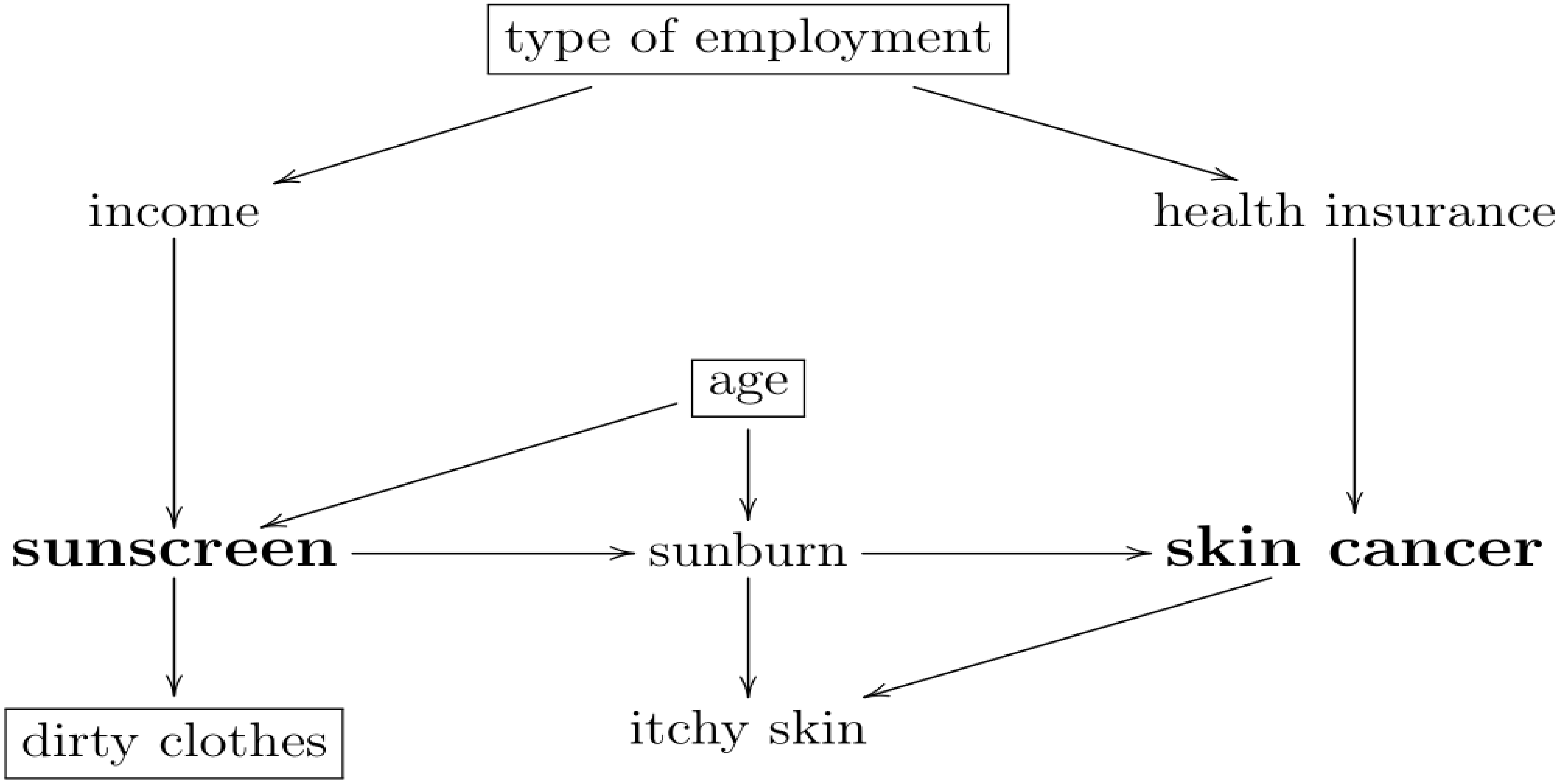
At least one unblocked front-door path from treatment to C (and this path may go through the outcome: so it could be X→Y→C)

And

Conditional on the treatment, there is at least one unblocked path (front-door or back-door) from C to the outcome.



Conditioning on itchy skin causes bias



Conditioning on dirty clothes does not cause bias because conditional on sunscreen there is no unblocked path from dirty clothes to skin cancer

There are two problems with all of this

Problem 1

We rarely know the true DAG. Whatever we draw is probably not correct.

But what other choice do we really have? Head in the sand?

Better to at least make all your assumptions explicit.

If you only get the primary confounders correct (and you are able to include them all in the regression model) then you do not need to include any other variables in the DAG.

Still can be helpful for identifying variables that should not be included in the regression model. (Will usually be able to identify these even if rest of the DAG is not correct.)

There are two problems with all of this

Problem 2

Even if we have the correct DAG we may not have all the variables we need to remove all confounding.

What then?

Will including a partial set of confounders in a regression model still reduce confounding?

Unfortunately, the answer is not always. In fact sometimes it will actually increase confounding (even if the variables are true confounders).

Why including on a subset of confounders might increase bias.

Confounding bias always has a direction. When there is confounding, (unconditional on any confounders) the correlation between X and Y is either larger or smaller than the true causal effect of X on Y.

Suppose

(Correlation between X and Y) < (causal effect of X on Y)

Which means the confounding is a negative bias.

Suppose also that you have only one confounder C, but that confounder causes positive bias. If you include C in your regression model your treatment effect estimate will <u>decrease</u>, which moves it <u>further away from the true causal effect</u>.

Why make the DAG when you know it will almost certainly be wrong?

1. What is the alternative? Bury your head in the sand?
2. Without it you are making a set of implicit assumptions. Better to make your assumptions explicit (even if they are wrong).
3. Can still be useful for identifying variables that should not be included in regression (those that induce confounding or non-confounding bias).
4. Can help you choose a set of variables with best chance to reduce confounding (even if you cannot completely remove it).

Practical guidance for using DAGs in nonrandomized studies

Accept that you will not be able to draw the correct DAG, but do your best to construct one anyway AND do this BEFORE you begin collecting data, so you can know what variables to collect.

Make sure to also include all the primary confounders on the graph. If either after the study is complete you have not measured the primary confounders or for some reason cannot collect them then look for the secondary confounders to block the primary confounders. A secondary confounder is a variable that blocks either the path from the primary confounder to X or the path from the primary confounder to Y.

Try to choose variables close to Y (especially if you think you do not have all the confounders), but make sure they do not have any unblocked front-door path from X.

Help is available

- CTSC and Cancer Center Biostatistics Office Hours
  - Every Tuesday from 12 – 2:00 currently via WebEx
  - 1st & 3rd Monday from 1:00 – 2:00 currently via WebEx
  - Sign-up through the CTSC Biostatistics Website
- EHS Biostatistics Office Hours
  - Upon request
- Request Biostatistics Consultations
  - CTSC
  - MIND IDDRC
  - Cancer Center Shared Resource
  - EHS Center