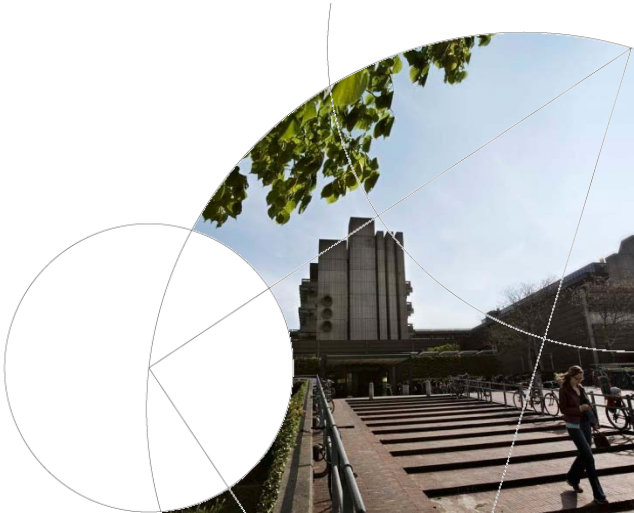
Mediation analysis – moving on

Molslaboratoriet, Oct 2014

DANISH
ramazzini
CENTRE

Theis Lange
thlan@sund.ku.dk
Section of Biostatistics
Faculty of Health and Medical Sciences



What is a causal effect?

- Assume that you are interested in understanding the impact of a given variable (the exposure) on one or more outcomes.

- If you can envies (at least in principle) a way to measure the impact of the exposure through a randomized trial this will be the **causal effect** of the exposure.

- Thus a causal effect is:

  Any effect which can (disregarding costs and ethics) be estimated from a randomized trial.

- The inclusion criterions for the enviesed randomized trial define to which population the causal effect applies.

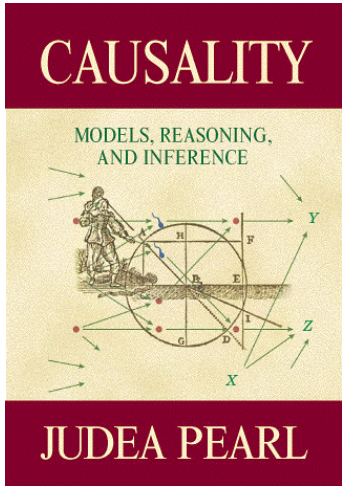
What is causal inference?

- For obvious reasons we cannot always conduct randomized trails.

- Several active areas of research, collectively referred to as causal inference, aim at deducing causal effects from observational studies.

- Participants from many backgrounds including:
  - Epidemiologists
  - Statisticians
  - Economists
  - Philosophers
  - Computer scientists
  - Physicists
  - ...

Causal inference books



The bible – old testament:

J. Pearl (2009)
Causality

The bible – new testament?

M.A. Hernán & J.M. Robins (2011?)
Causal Inference
Available online at:
http://www.hsph.harvard.edu/faculty/miguel-hernan/causal-inference-book/

Some other recent:

- S. Morgan & C. Winship (2007)
  Counterfactuals and Causal Inference
- D. Freedman (2010)
  Statistical Models and Causal Inference

The language of counterfactuals

- **Counterfactuals variables** have been developed to provide a firm (mathematical) footing for arguments about cause and effect.

- The counterfactual variable Y_a denotes the value the outcome (Y) would take if exposure (A) was set to *a* by an intervention.

- For every individual there exists many counterfactual variables corresponding to all possible levels of exposure.

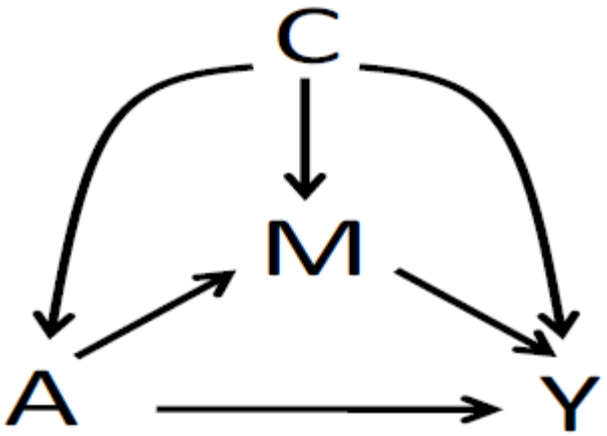
- Thus counterfactual variables describe what would have happened if we had intervened on exposure.

- For this reason they are also referred to as potential outcomes (or more popular: “What-if-mathematics”)

An illustrative (and famous) example

- The landmark paper Bertrand and Mullainathan (2004) aimed at quantifying the size of (unlawful) discrimination experienced by African Americans.
- The authors collected (at random) CVs from both whites and African Americans.
- Subsequently the names written on the CVs were randomly selected to be either typical white names or typical African American names.
- These CVs were send to many job openings and the rate of call-back was registered.

A formal framework for analyzing call-back mechanisme



- *A* is ethnicity.
- *M* is quality/content of the CV
- *Y* is call-back (0/1)

A formal framework

- For each job-opening one could send a particular CV with either a white or African American name.

- The counterfactual variable *Y(a,m)* denotes the call-back (yes/no) if we, perhaps contrary to fact, had set ethnicity to *a* and the CV to *m*.

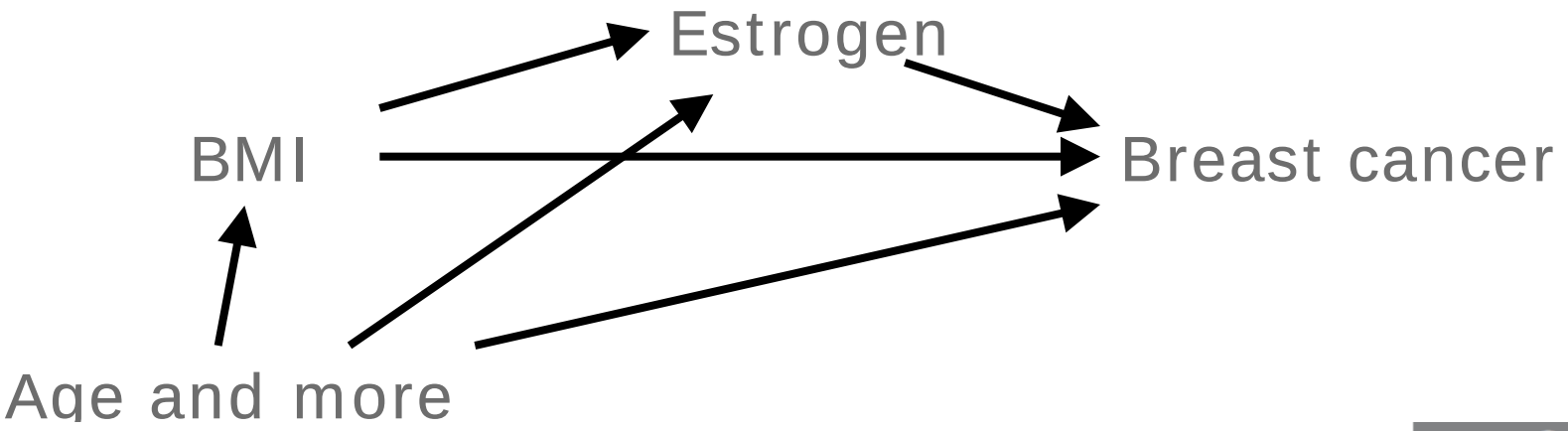
- The counterfactual variable *M(a)* describes the CV obtained if the person, perhaps contrary to fact, had had the ethnicity *a*.

- Unlawful racial discrimination would be

$$E[Y(\text{"afro"}, M(\text{"white"}))] - E[Y(\text{"white"}, M(\text{"white"}))]$$

- This is the so-called **natural direct effect (NDE)**.

One more example (I/III)



Research Article

Cancer Epidemiology, Biomarkers & Prevention

Quantifying Mediating Effects of Endogenous Estrogen and Insulin in the Relation between Obesity, Alcohol Consumption, and Breast Cancer

Ulla A. Hvidtfeldt[1], Marc J. Gunter[4], Theis Lange[2], Rowan T. Chlebowski[5], Dorothy Lane[6], Ghada N. Farhat[8], Matthew S. Freiberg[9], Niels Keiding[2], Jennifer S. Lee[10], Ross Prentice[11], Anne Tjønneland[3], Mara Z. Vitolins[12], Silvia Wassertheil-Smoller[7], Howard D. Strickler[7], and Naja H. Rod[1]

Background: Increased exposure to endogenous estrogen and/or insulin may partly explain the relationship of obesity, physical inactivity, and alcohol consumption and postmenopausal breast cancer. However, these potential mediating effects have not been formally quantified in a survival analysis setting.

Methods: We combined data from two case–cohort studies based in the Women's Health Initiative-Observational Study with serum estradiol levels, one of which also had insulin levels. A total of 1,601 women

One more example (II/III)

Following Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects *Epidemiology*. 1992;3(2):143-155.

General definition:

- The ***natural direct effect*** measures the change in outcome that would be observed if we could change the exposure, but leave the mediator at the value it naturally takes when the exposure is left unchanged.

The breast cancer example:

- In the BC example ***natural direct effect*** is the change in BC that would be observed if BMI was increased 5-units without inducing any change in estrogen levels.

One more example (II/III)

Following Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects *Epidemiology*. 1992;3(2):143-155.

General definition:

- The ***natural indirect effect*** measures the change in outcome that would be observed if we could change the mediator as much as it would naturally change when exposure was changed without actually changing the exposure.

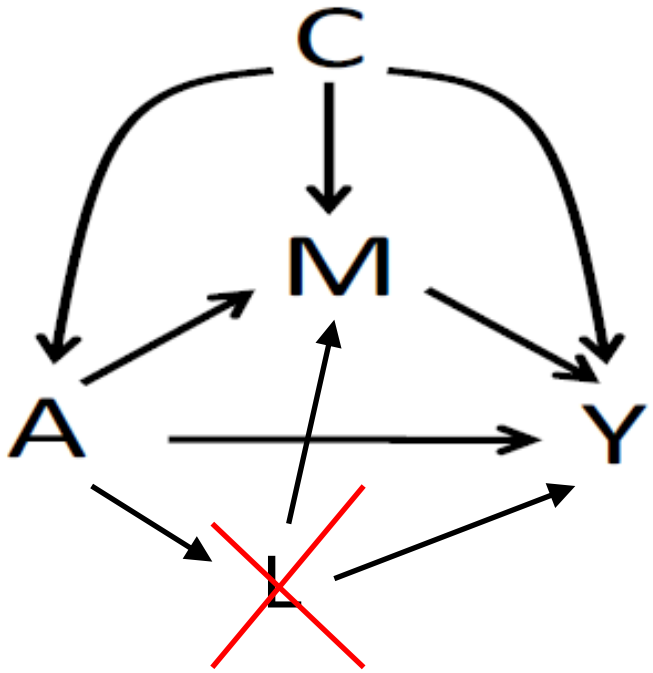
The breast cancer example:

- In the BC example ***natural indirect effect*** is the change in BC that would be observed if estrogen was changed as much as it would naturally change if BMI was increased 5-units without actually changing BMI

Assumptions for natural direct and indirect effects

1: No unmeasured confounders of:

- The exposure-outcome relation
- The exposure-mediator relation
- The mediation-outcome relation



2: No intertwined causal pathways.
(aka. Pearls identifiability condition)

3: Consistency and positivity.
(mostly technical)

REF: Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*. 1992;3(2): 143–155.

OLD vs. NEW

	OLD (Baron & Kenny = Part I of my talk)	NEW (counterfactual based = Part II of my talk)
Coding	Easy	Harder
Underlying idea	An algorithm	A defined parameter of interest.
Bias	Yes (except in purely linear models)	No (for any combination of variables types)

So how to estimate natural direct and indirect effects

Direct and Indirect Effects in a Survival Context

(*Epidemiology* 2011;22: 575–581)

Theis Lange[a] *and Jørgen V. Hansen*[b]

American Journal of Epidemiology

Vol. 176, No. 3
DOI: 10.1093/aje/kwr525
Advance Access publication:
July 10, 2012

Practice of Epidemiology

A Simple Unified Approach for Estimating Natural Direct and Indirect Effects

American Journal of Epidemiology

DOI: 10.1093/aje/kwt270

Practice of Epidemiology

Assessing Natural Direct and Indirect Effects Through Multiple Pathways

- And many more.
- Look for instance at S. Vansteelandt, Tyler VanderWeele, and Imai.

Natural effects models

- We suggest to parameterize the natural direct and indirect effects directly in a model for the (nested counterfactual) outcome:

$$g(E[Y_{a,M_{a^*}}]) = c_0 + c_1 a + c_2 a^* + c_3 a \cdot a^* \qquad (1)$$

g is a link function specifying the requested model for the outcome (e.g. logistic model) and c_3 is an interaction term.

- When $c_3 = 0$ and g is the logit link, then

$$\exp[c_1(a - a^*)] = \frac{\text{odds}[Y_{a,M_{a^*}} = 1]}{\text{odds}[Y_{a^*,M_{a^*}} = 1]}$$

captures the natural direct effect odds ratio.

Natural effects models (cont.)

- Besides all outcomes which can be modeled by generalized linear models the approach can also handle survival outcomes using either Cox or additive hazard models.

- Mediator and exposure can be of any type.

- However, we only see persons with a=a*, so we have to be clever about estimation.

Estimation procedure for natural effects models (I/II)

Under standard assumptions of no-unmeasured confounders and no exposure dependent confounding the MSMs in (1)-(3) can be estimated by the following approach:

1. Estimate a suitable model for the mediator conditional on exposure and baseline variables using the original data set.
2. Construct a new data set by repeating each observations in the original data set twice and including an additional variable A^*, which is equal to the original exposure for the first replication and equal to the opposite of the actual exposure for the second replication.

   See the paper for details regarding categorical or continuous exposures.

Estimation procedure for natural effects models (II/II)

3. Compute weights given by

$$W_i = \frac{P(M = M_i | A = A_i^*, C = C_i)}{P(M = M_i | A = A_i, C = C_i)}.$$

through applying the fitted model from step 1 to the new data set. In most software packages this can be done using predict-functionality.

4. Fit a suitable MSM model to the outcome including A, A^*, (perhaps their interaction) and baseline variables as covariates and weighted by the weights from the previous step.

5. Conservative confidence intervals can be obtained using robust standard errors or better yet using bootstrap.

An example: the rage question

(collaboration with S.B. Hansen, Retspsykiatrisk Kompetencecenter)

- **Exposure:** Number of personality disorders (PD)
  - Integers between 0 and 7
  - Measured by semi-structured interview (SCID-II)
- **Mediator:** Attachment style
  - Four types: secure, fearful, preoccupied, and dismissing
  - Measured by questionnaire
- **Outcome:** Rage type
  - Two types: Impulsive and premeditated (last is most dangerous)
  - Measured by IPAS-30 questionnaire
- In total 108 patients.

An example: the rage question

- Attachment type's dependence on PD is modeled by a Multinomial logit model.
  R code:

```
myData <- suneData
myData$id <- 1:nrow(myData)
myData$myATemp <- suneData$SCID_PD

fitM <- vglm(RQ_A_Style ~ myATemp + factor(Gender),
             data = myData, family=multinomial())
```

An example: the rage question

- Next create a new data set

```
exposureLevels <- unique(suneData$SCID_PD)
newMyData <- NULL
for(i in 1:length(exposureLevels))
{
    myDataTemp <- myData
    myDataTemp$Astar <- exposureLevels[i]
    newMyData <- rbind(newMyData, myDataTemp)
}
```

An example: the rage question

- Then compute weights

```
newMyData$myATemp <- newMyData$SCID_PD
temp <- predict(fitM,type = "response", newdata=newMyData)
tempDir <- temp[cbind(1:nrow(temp),newMyData$RQ_A_Style)]
            # requires that mediator takes on values like 1,2,3,4

newMyData$myATemp <- newMyData$Astar
temp <- predict(fitM,type = "response", newdata=newMyData)
tempIndir <- temp[cbind(1:nrow(temp),newMyData$RQ_A_Style)]
            # requires that mediator takes on values like 1,2,3,4

newMyData$weightM <- tempDir/tempIndir
```

An example: the rage question

- Finally, fit the MSM model

```
require(geepack)
newMyData <- newMyData[order(newMyData$id), ]
fitYbinary <- geeglm(I(IPAS_IA_PM_C==2) ~ SCID_PD + Astar + factor(Gender)
family='binomial', data=newMyData, weights=1/weightM, id=newMyData$id,
scale.fix = T)
```

- Result:

```
                  Estimate  Std.err   Wald Pr(>|W|)
(Intercept)      -0.91437  0.36974  6.116   0.0134 *
SCID_PD           0.41173  0.19445  4.483   0.0342 *
Astar             0.21730  0.03863 31.640 1.86e-08 ***
factor(Gender)2 -0.58616  0.53210  1.213   0.2706
```

so natural direct effect OR: exp 0.41 = 1.51, natural indirect effect OR: 1.25, and total effect OR: $1.51 * 1.25 = 1.89$.

How to obtain confidence intervals for TE and IE/TE

Step 1: Collect information

Source	Estimate	Std. error	Covariance*
A	0.4173	0.19	$2.82*10^{-5}$
A*	0.2173	0.03863	
Dummy (do not change)	1	0	No need.

* In R you obtain this using the **vcov**-function.

Step 2: Use Excel sheet

Discussion

The good news:

- In linear models simple multiplication along paths will suffice.
- We have suggested a unified approach to estimate direct and indirect effects.
- The paper includes SAS and R code. Moreover, some Stata code is available from me upon request.

However, do not forget:

- 3 times no-confounding (even in RCTs!)
- Measurement error on mediator bias IE towards zero
- What if the mediator is in fact a process evolving continuously?
- Mediation analysis require a lot of data.

Thanks for the attention – questions?