The Case Cotwin Design and Analysis

Using twins in matched studies

Jacob Hjelmborg

University of Southern Denmark

Odense Spring 2018

Overview

University of Southern Denmark

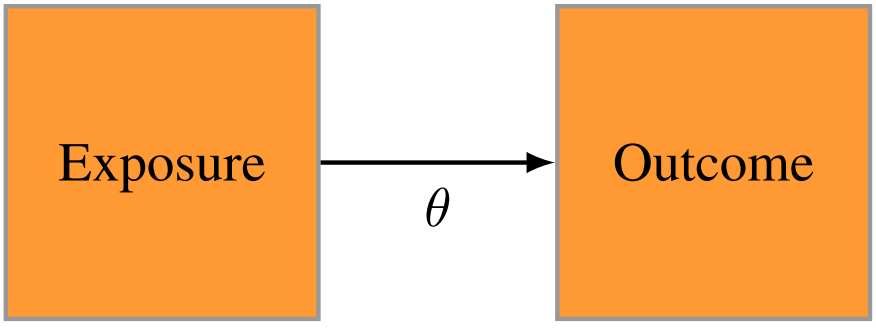
'research to improve human health.'

-based on knowledge from data (and not alternative facts).

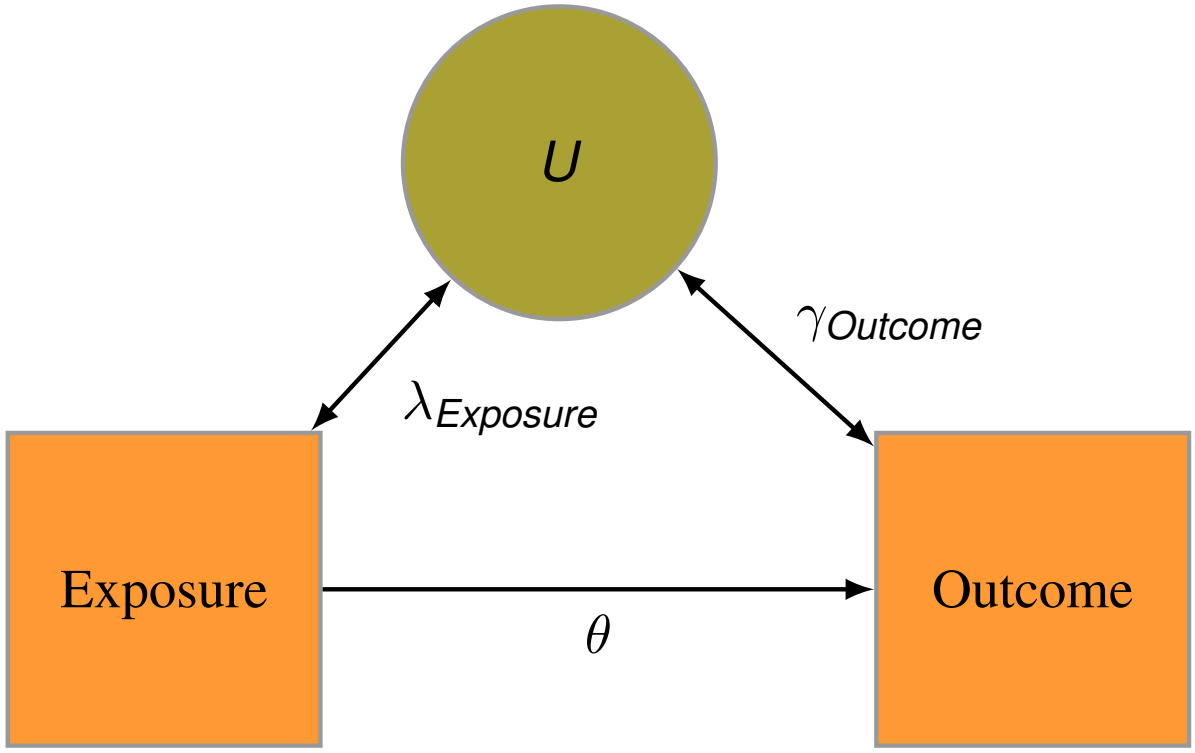
Overview

University of Southern Denmark

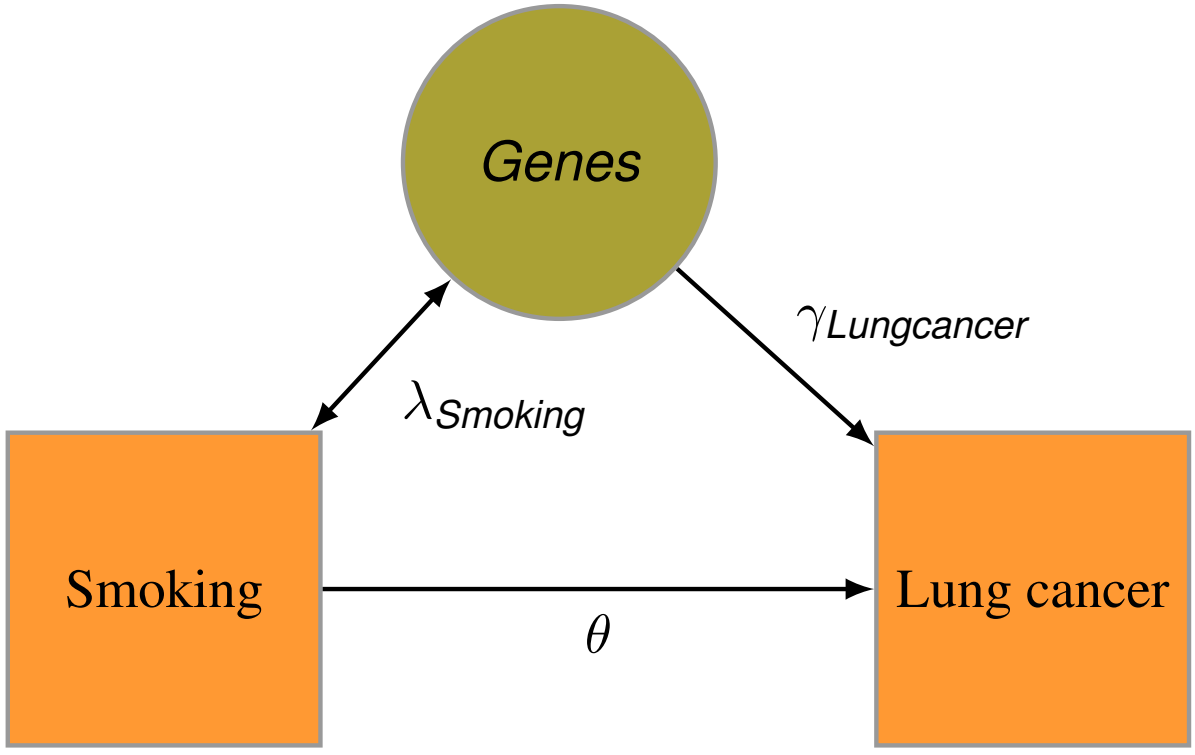
Experiment $\rightarrow$ Design $\rightarrow$ Analysis



The problem

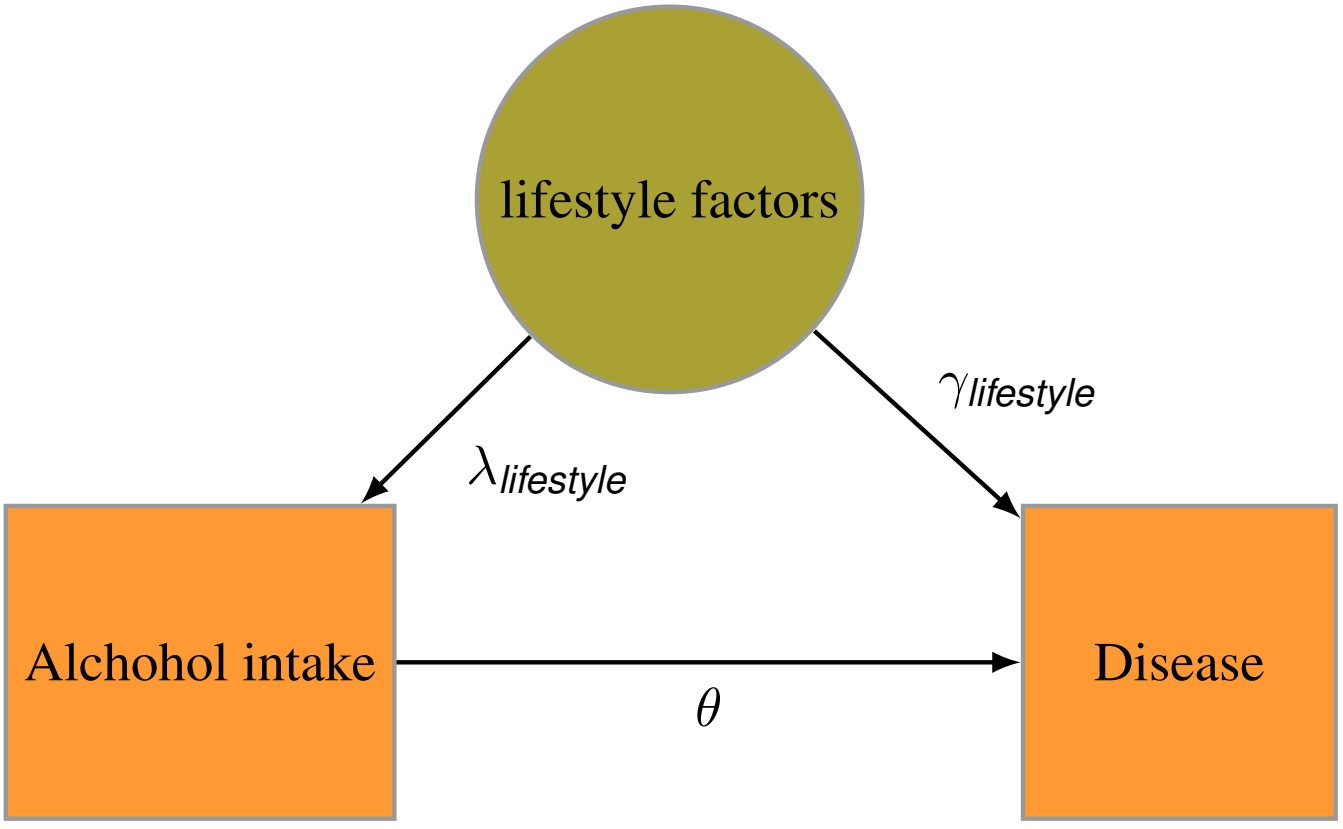


The problem - examples



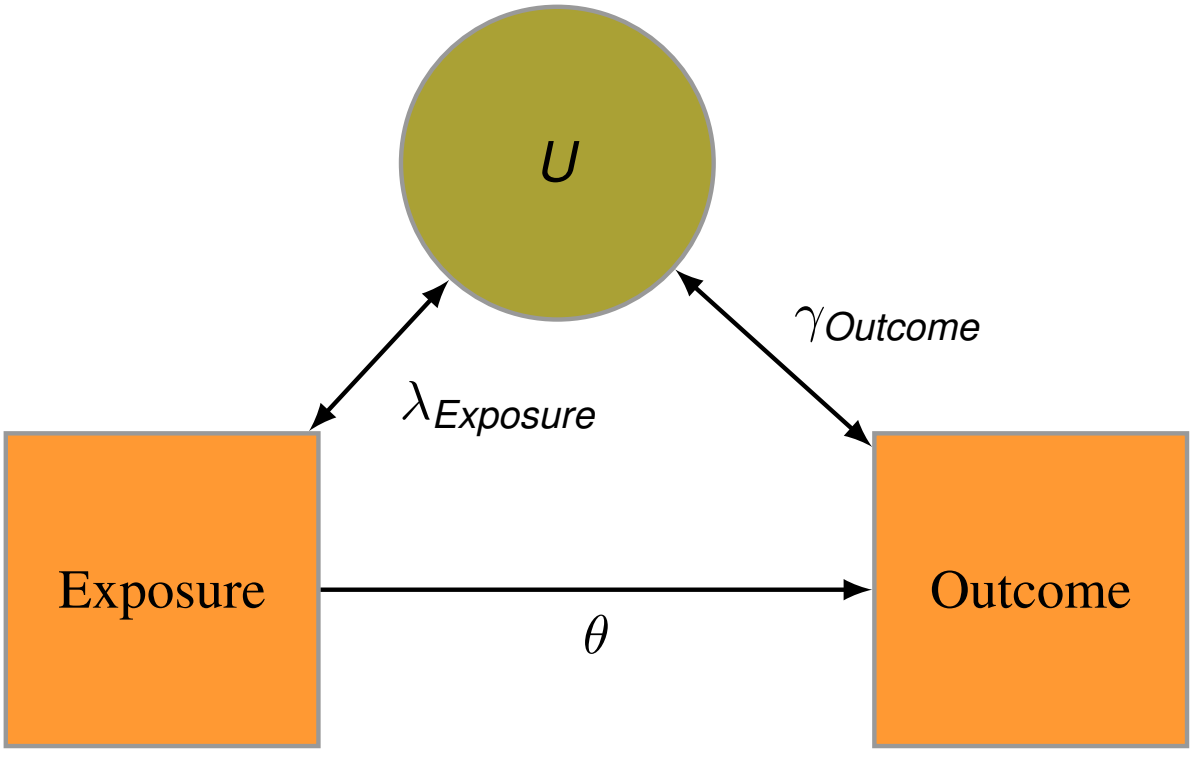
- (Smoking): How harmful is the exposure?
- (Genes): How about unobserved confounders - eg. genetic effects?.

The problem - examples



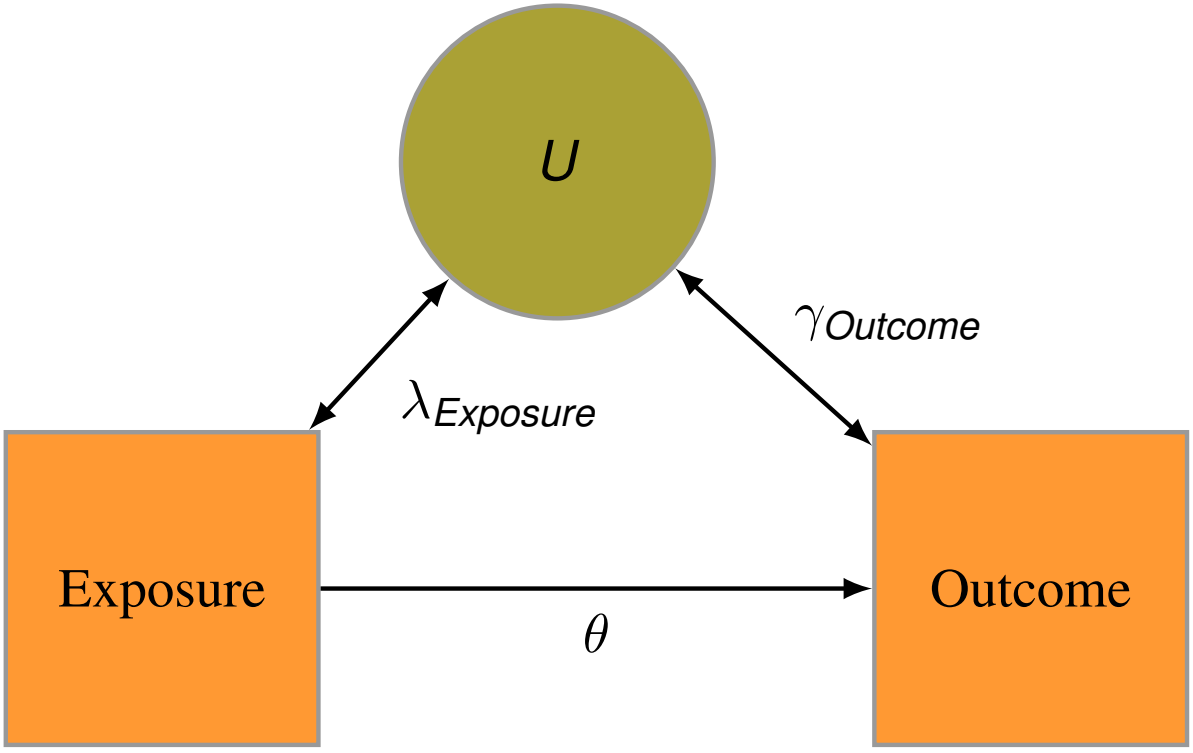
- How harmful is drinking alchohol? (coronary heart disease (J-shape), hypertension, cancers,..)
- How about unobserved lifestyle factors?.
- Causal meaning is doubtful

The problem



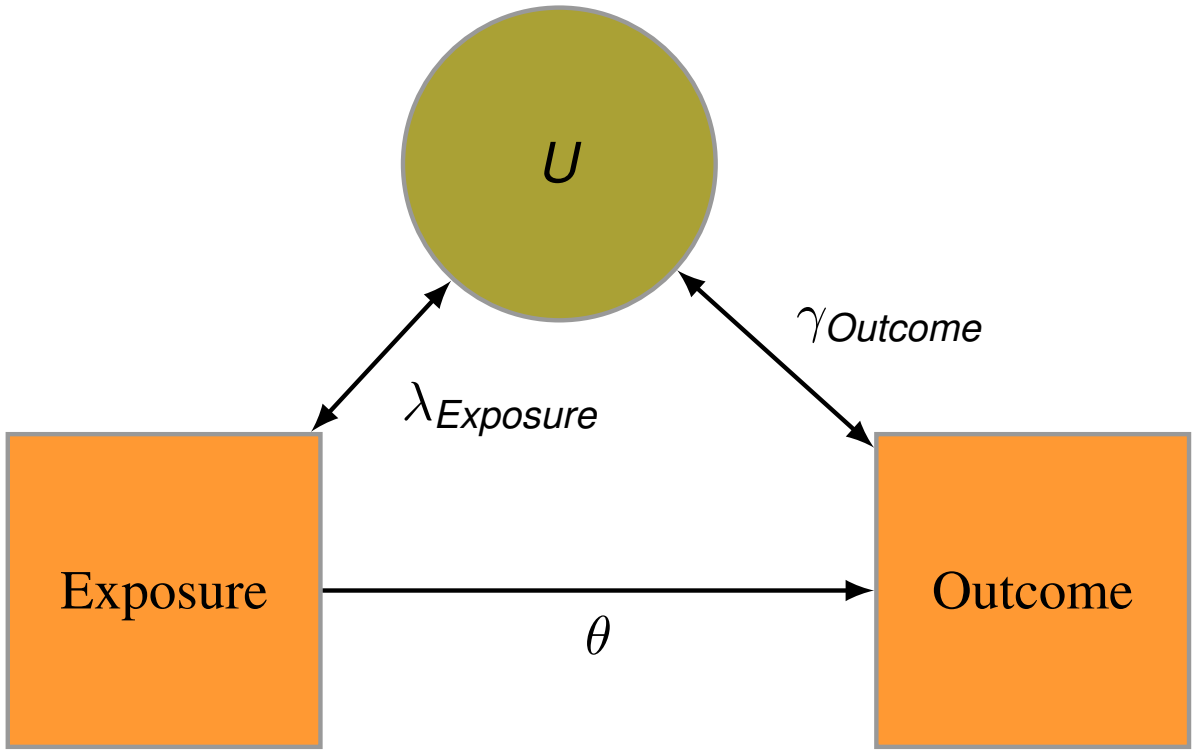
- What is the effect of the exposure?
- confounding? Effects of interest are confused with other effects.
- -unobserved confounders?

Principle of Randomization



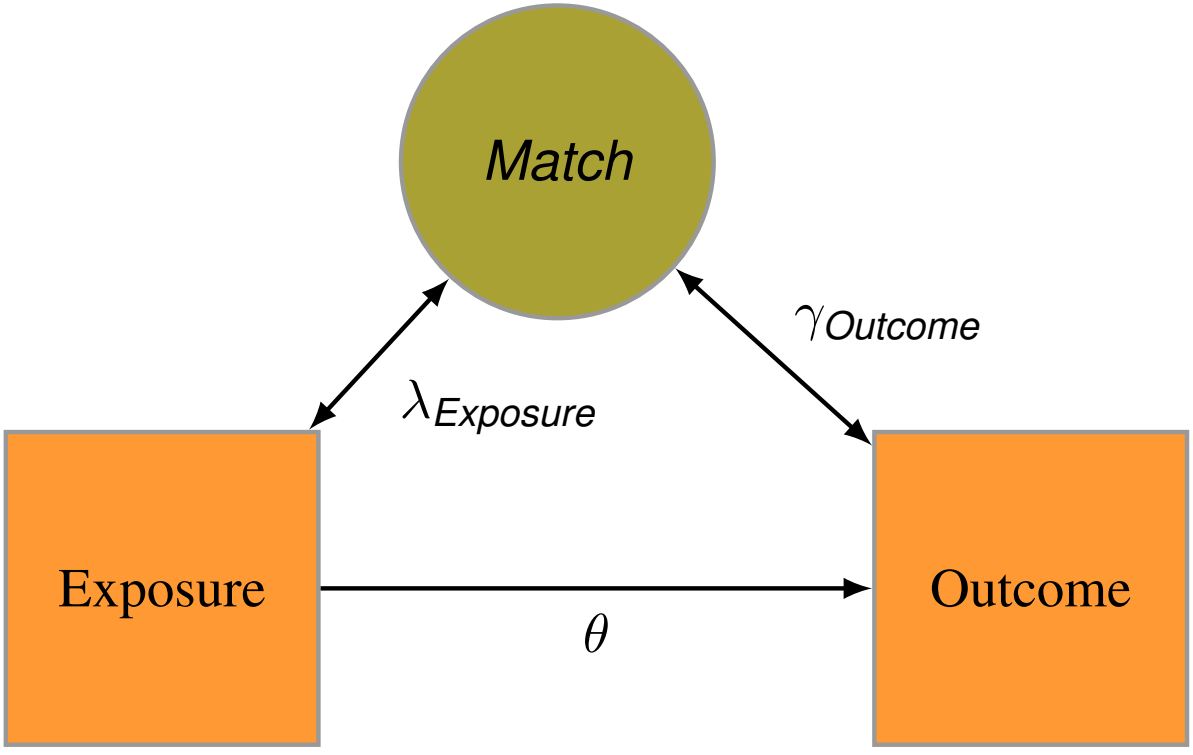
- R.A. Fisher (1935): Randomization negates the effect of confounders.

Principle of Representation

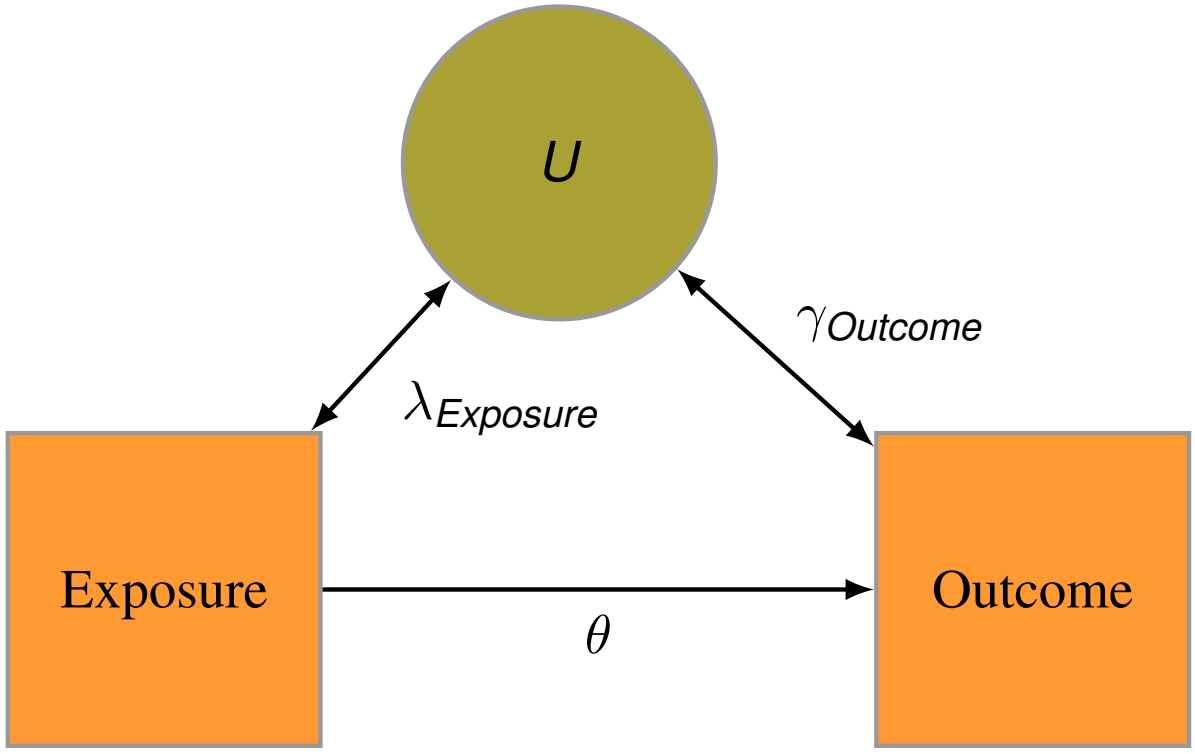


- Representative sampling negates the effect of confounding effect-modifiers.

Principle of Matching



- Matching to ensure that case and control are similar with respect to certain confounding variables.



- Randomization negates the effect of confounders.
- Representative sampling negates confounding effect-modifiers.
- Matching negates the effect of certain confounders.
- -random effects. Give a model for U - the mixed models
- -instrumental variables. Mimic randomization.
- -inverse probability weighting. Mimic representation
- . . .

Outline

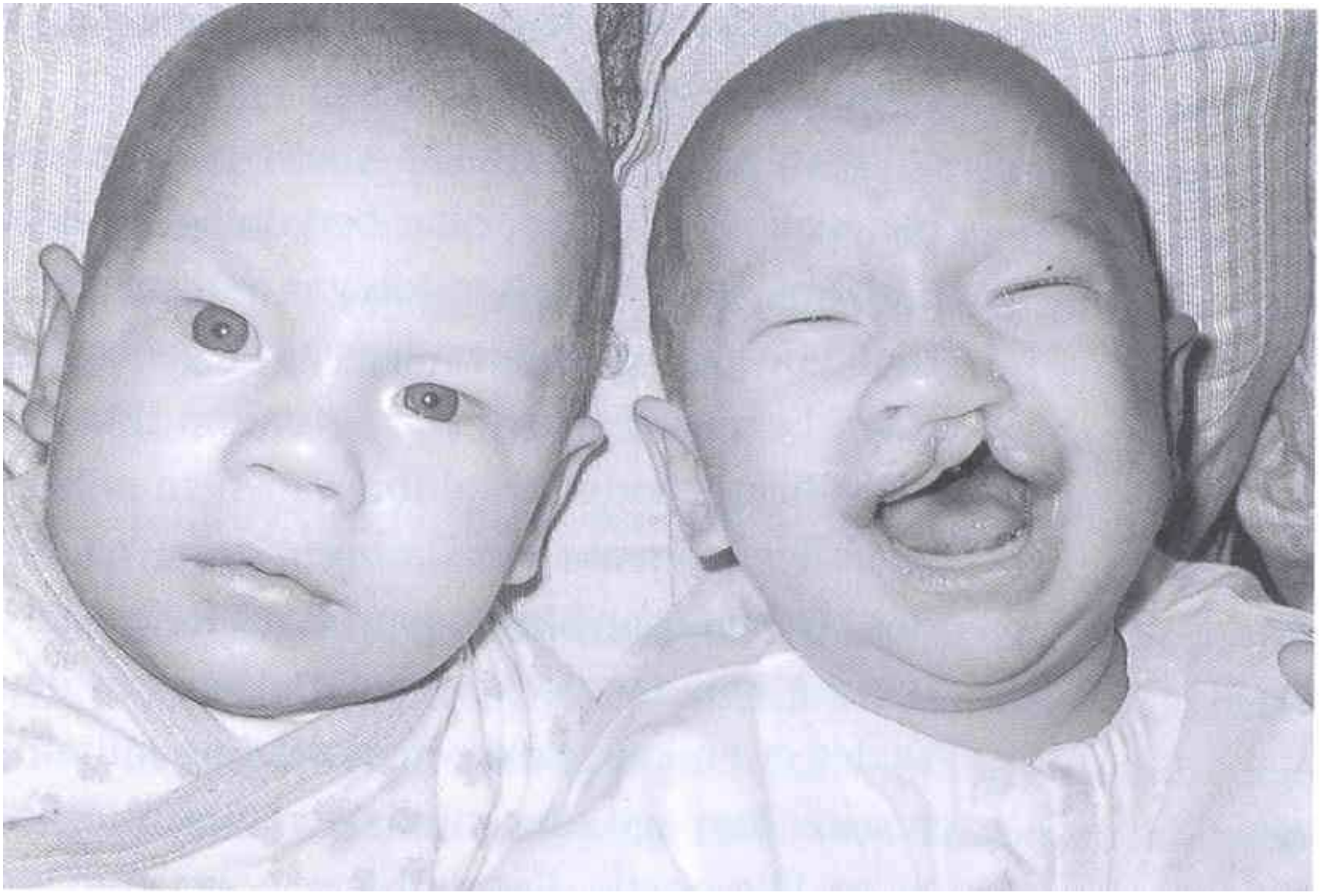
The matched case-so twin design for inferring association of exposure with outcome.

Findings using twin pairs include for instance:

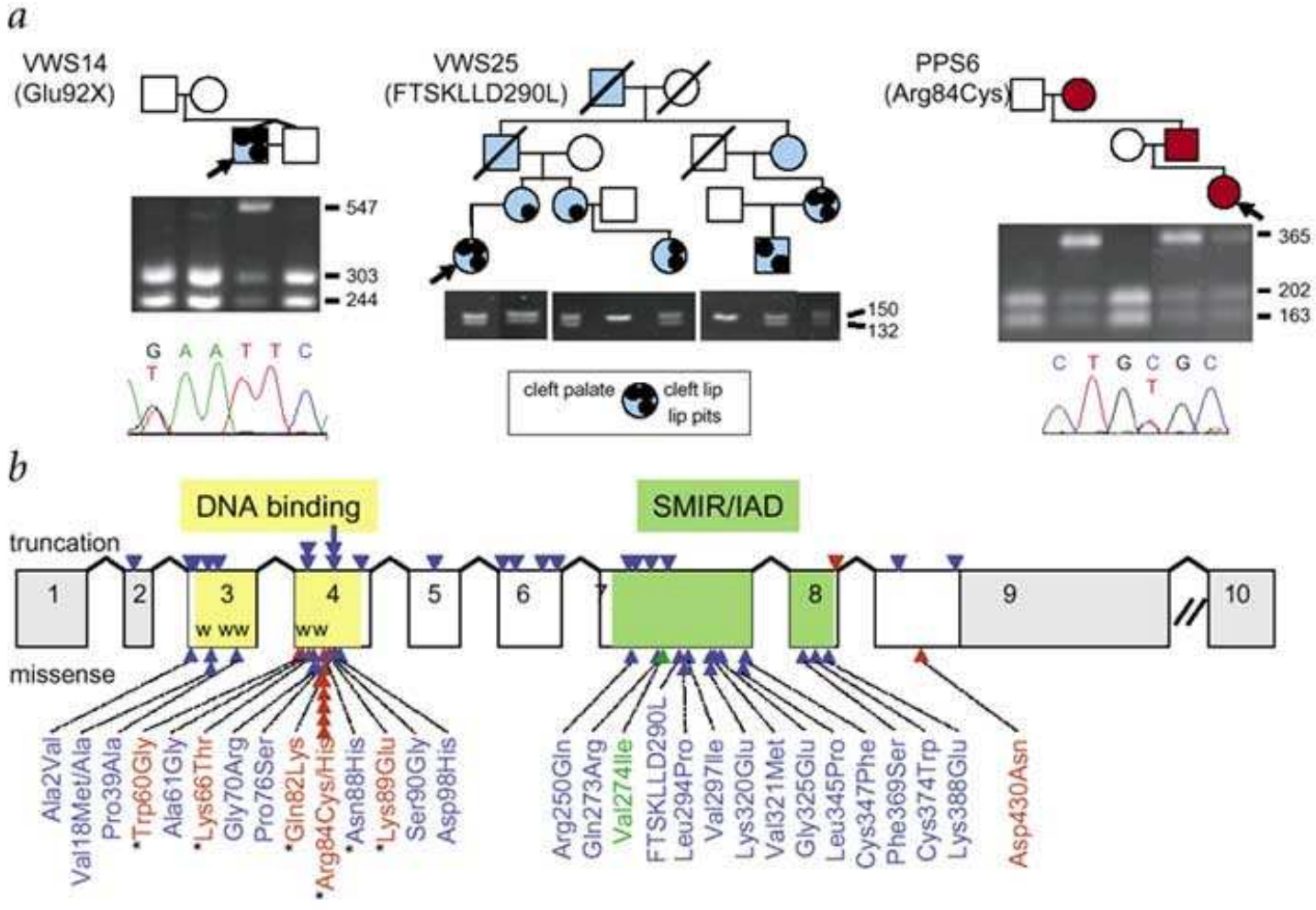
- -human leucocyte telomere length is associated with longevity.
- -the causing mutation for the van der Woude cleft lip palate syndrome.
- -perceived age is associated with longevity.
- -heart rate at rest is associated with longevity.
- -otitis medea as risk factor for dyslexia
- -respiratory symptoms of perfume - Hand eczema
- *-do genetic factors contribute to the association between birth weight and blood pressure?*
- -antibodies for reumatoid arthritis.

And many more results in genetic and epigenetic epidemiology. We consider the methodology, underlying assumptions and pitfalls and work out the analysis for very general cases.

MZ pair discordant for van der Woude syndrome



van der Woude syndrome: Got the position (and insight)!

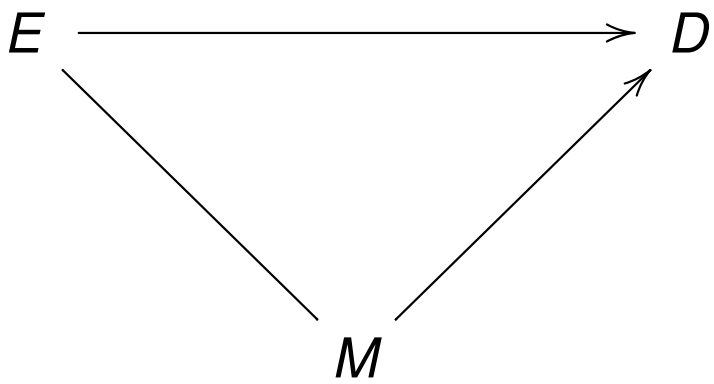


Kondo et al. Nat Genet 32:205, 2002

The individual matched case control design

Principles:

- A matching variable must be regarded as a confounder.



- Efficiency gain: More precise estimate of effect measure.

Overview

University of Southern Denmark

The matched case co-twin design

- Use co-twin as match.
- Discordant pairs for a trait may be highly informative and compared wrt. multiple exposures.
- MZ pairs: controlling for *certain* genetic effects and great many background factors.
- -but confounders should be more shared for the pairs than the exposure - see Pro and Con later.

Overview

University of Southern Denmark

Analysis in matched case cotwin studies

Exposure→Outcome

Exposure	**Outcome**		
	Dichotomous outcome	numeric scale	survivaltime
Binary exposure (Yes/No)	2×2 *pair* table Odds-ratio McNemar χ^2-test	Paired *t*-test mean of differences	
Continuous exposure	Conditional logistic regression, cases vs. controls in same matched set	Intrapair regression, Between - within pair effects model	Cox regression, baseline hazard function for each pair

Examples

- Case cotwin study with dichotomous outcome: Otitis Medea→Dyslexia
- Case cotwin in strata of zygosity: Respiratory symptoms of perfume - Hand eczema
- Case cotwin survival analysis: Human telomere length and lifespan.
- Intrapair analysis for continuos outcome and exposure: *Do genetic factors contribute to the association between birth weight and blood pressure?*
- Sparse numbers analysis: Antibodies for reumatoid arthritis.

Analysis in matched case cotwin studies

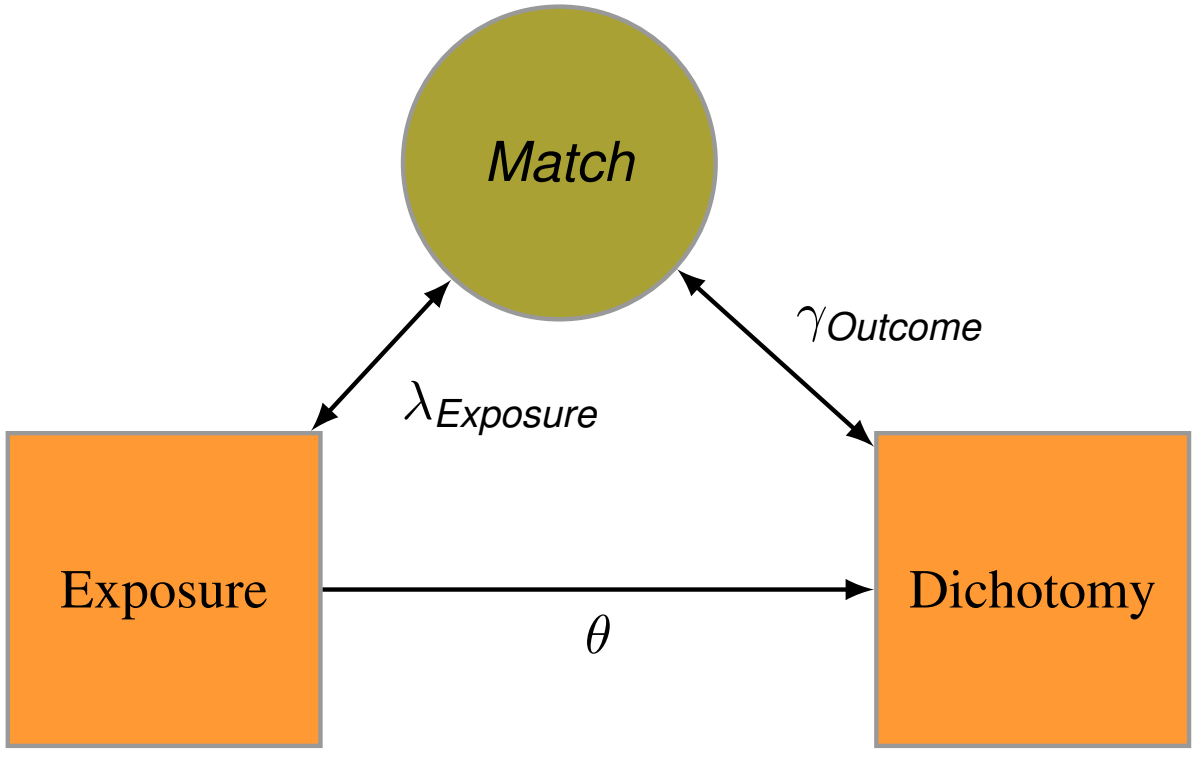
Exposure→Outcome

Exposure	**Outcome**		
	Dichotomous outcome	numeric scale	survivaltime
Binary exposure (Yes/No)	2×2 *pair* table Odds-ratio McNemar χ^2-test	Paired *t*-test mean of differences	
Continuous exposure	Conditional logistic regression, cases vs. controls in same matched set	Intrapair regression, Between - within pair effects model	Cox regression, baseline hazard function for each pair

Examples

- Case cotwin study with dichotomous outcome: Otitis Medea→Dyslexia
- Case cotwin in strata of zygosity: Respiratory symptoms of perfume - Hand eczema
- Case cotwin survival analysis: Human telomere length and lifespan.
- Intrapair analysis for continuos outcome and exposure: *Do genetic factors contribute to the association between birth weight and blood pressure?*
- Sparse numbers analysis: Antibodies for reumatoid arthritis.

Analysis when Matching



Case cotwin study: Riskfactors of Dyslexia

- Danish twin study: *Otitis medea, Dyslexia and Dysphasia.* (conducted by Steen Fibiger)
- Table of *individuals*: Dyslexia (in adulthood) versus Otitis Medea (in childhood).

	otitis medea	**no otitis medea**
dyslexia	768	1657
control	7933	22844

- Odds ratio is $OR = 1.33$ with 95% CI. $(1.22, 1.46)$.
- In R: *glmer(dyslexia* $\sim$ *ottitis+ (1|tvparnr), family=binomial, data=fib)*
- Confounders? We do a matched analysis
- Table of matched pairs discordant for Dyslexia:

	Co-twin	
Twin	**ottitis medea**	**no ottitis medea**
ottitis medea	46	254
no ottitis medea	201	3042

- $\mathrm{OR} = \frac{\#\text{ case is exposed and control is not exposed}}{\#\text{ control is exposed and case is not exposed}} = \frac{254}{201} = 1.26\ (1.05, 1.52)$.
- McNemar's test: $\chi^2 = 6.17$ (1 df.), so $\text{p-value} = 0.013$.
- We may adjust for further confounding by conditional logistic regression.

Case cotwin study: Riskfactors of Dyslexia

- We may adjust for further confounding by conditional logistic regression.
- Conditional logistic model: $\log(\text{odds}(Y = 1)|X, \text{matching}) = \beta X$
- In R: *clogit(dyslexia ∼ ottitis+dysphasia+sex + strata(tvparnr), data=fib)*
- In Stata: *clogit dyslexia ottitis sex, group(tvparnr) or*

```
------------------------------------------------------------------------------
    dyslexia | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      otitis |   1.274064   .1214243     2.54   0.011     1.056983     1.53573
         sex |     1.5474   .1465459     4.61   0.000     1.285257     1.86301
------------------------------------------------------------------------------
```

- We may consider dysphasia as well
- In Stata: *xi: clogit dyslexia i.dysphasia*ottitis sex, group(tvparnr) or*

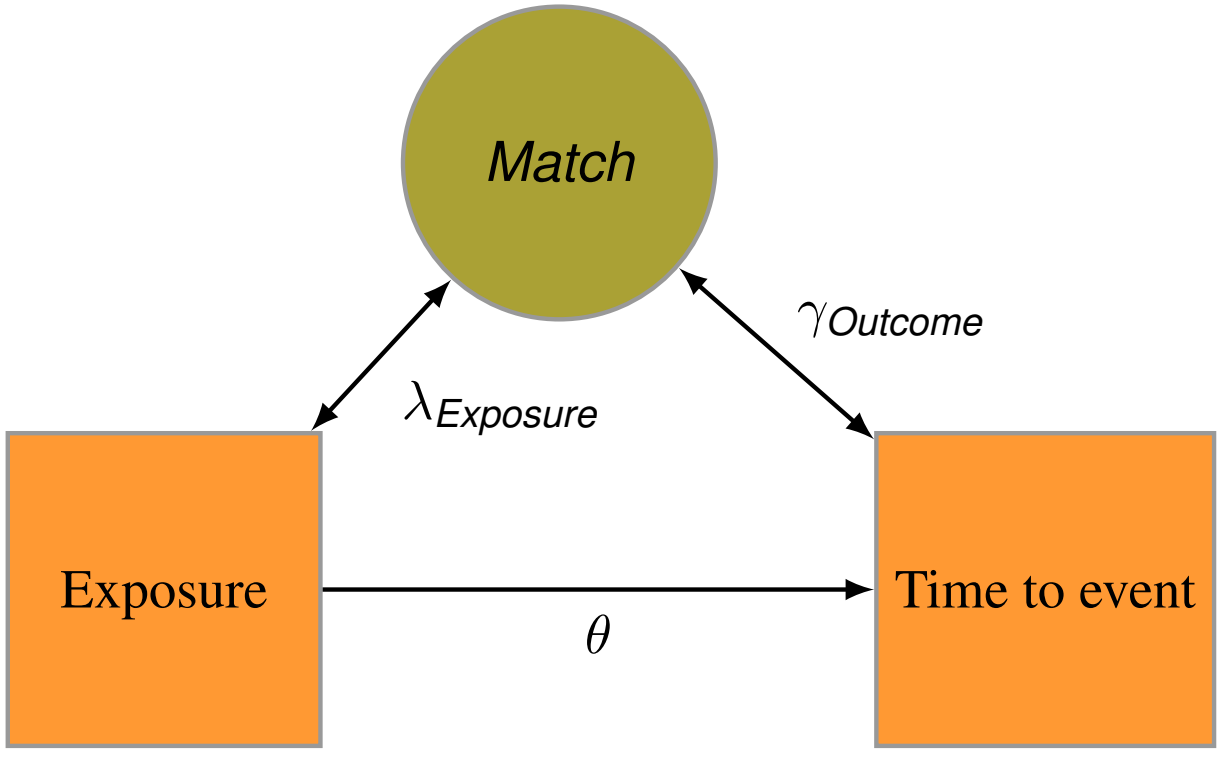
```
------------------------------------------------------------------------------
    dyslexia | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   dysphasia |   2.456097   .4483305     4.92   0.000      1.71739    3.512545
      otitis |   1.188503   .1237568     1.66   0.097     .9690942    1.457587
dysphXotitis |   1.213239   .3601591     0.65   0.515     .6780487    2.170861
         sex |   1.451798   .1413007     3.83   0.000     1.199666     1.75692
------------------------------------------------------------------------------
```

- Model selection and diagnostics as in case of logistic regression.

Practical's

- We consider if ottitis medea might be associated with dyslexia using the case co-twin design matching for numeruos factors.
- The scripts "dyslex.R" (also in appendix of slides) and "dyslex.do" contains extracts of analysis using R and Stata, respectively.
- Apply the logistic regression model. Why cluster on twin pairs?
- Analyze the scenario to recover above results by applying the conditional logistic regression model.
- Digression: Do you think there is evidence for genetic influence on Dyslexia? Examine this using the pairwise odds regression approach.

Analysis when Matching



Case cotwin study: Time to event

- Digression: How about time to event using case-cotwin design?
- Suppose it takes one time unit until event, i.e., a variable 'time' is 1 for individuals until dyslexia (for those who gets it).
- -then the Cox regressionmodel recovers results above:
- In Stata: *stcox ottitis sex, strata(tvparnr) hr nolog exactp*

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
ottitis	1.278802	.1214791	2.59	0.010	1.061558	1.540504
sex	1.540007	.1456308	4.57	0.000	1.279466	1.853602

- Hence we may do case cotwin survival analysis.

Case cotwin study stratifying by zygosity

Exposure→Outcome

- *in strata of zygosity.*

Respiratory symptoms of perfume - Hand eczema

Twin	Co-twin	
	eczema	**no eczema**
eczema	4	36
no eczema	10	126

Respiratory symptoms of perfume - Hand eczema

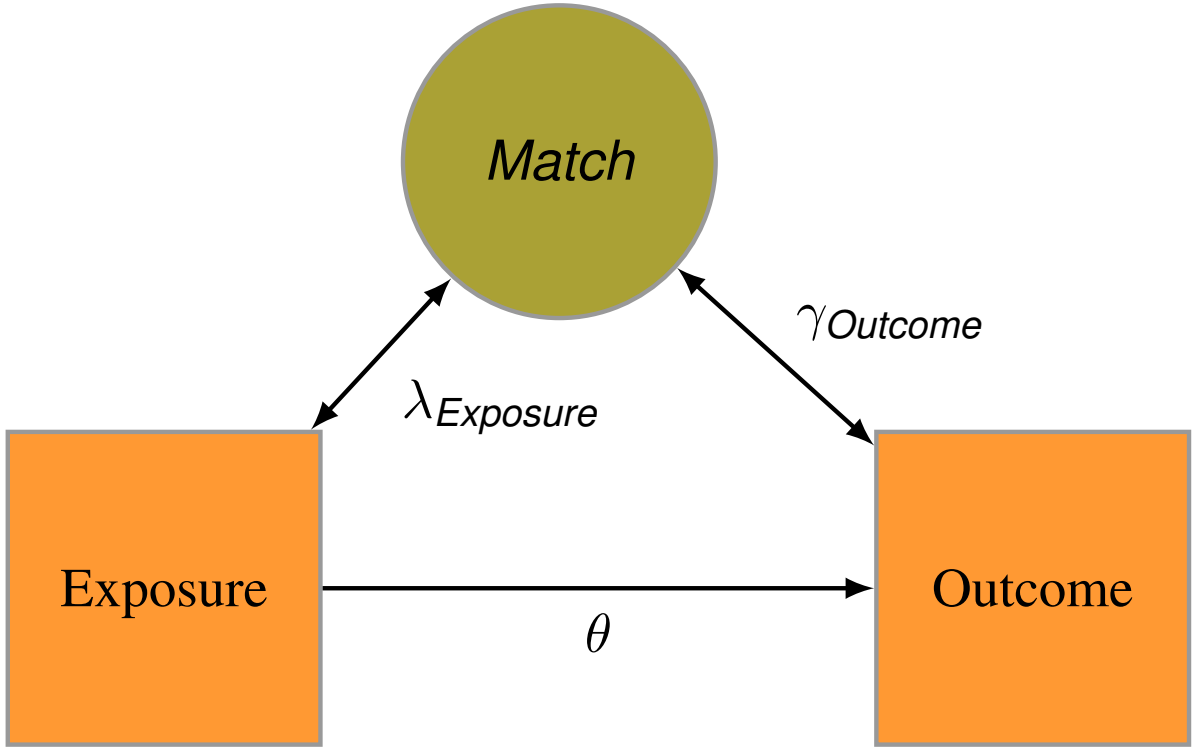
	Co-twin	
Twin	**eczema**	**no eczema**
eczema	4	36
no eczema	10	126

- The discordant pairs are informative for the Exposure-Disease association.
- Difference in OR for MZ and DZ pairs may indicate common genetic effects.

	Co-twin	
Twin	**Informative pairs**	**Odds-ratio (95% CI.)**
All	46	3.60 (1.75, 8.13)
MZ	16	4.33 (1.19, 23.71)
DZ	30	3.29 (1.36, 9.07)

Elberling et al. 2008

Analysis when Matching



- Matching when sparse data, but informative.

Sparse data - association?

Autoantibodies in twins discordant for rheumatoid arthrititis

Table 1 Number of RA discordant pairs by antibodies and zygosity

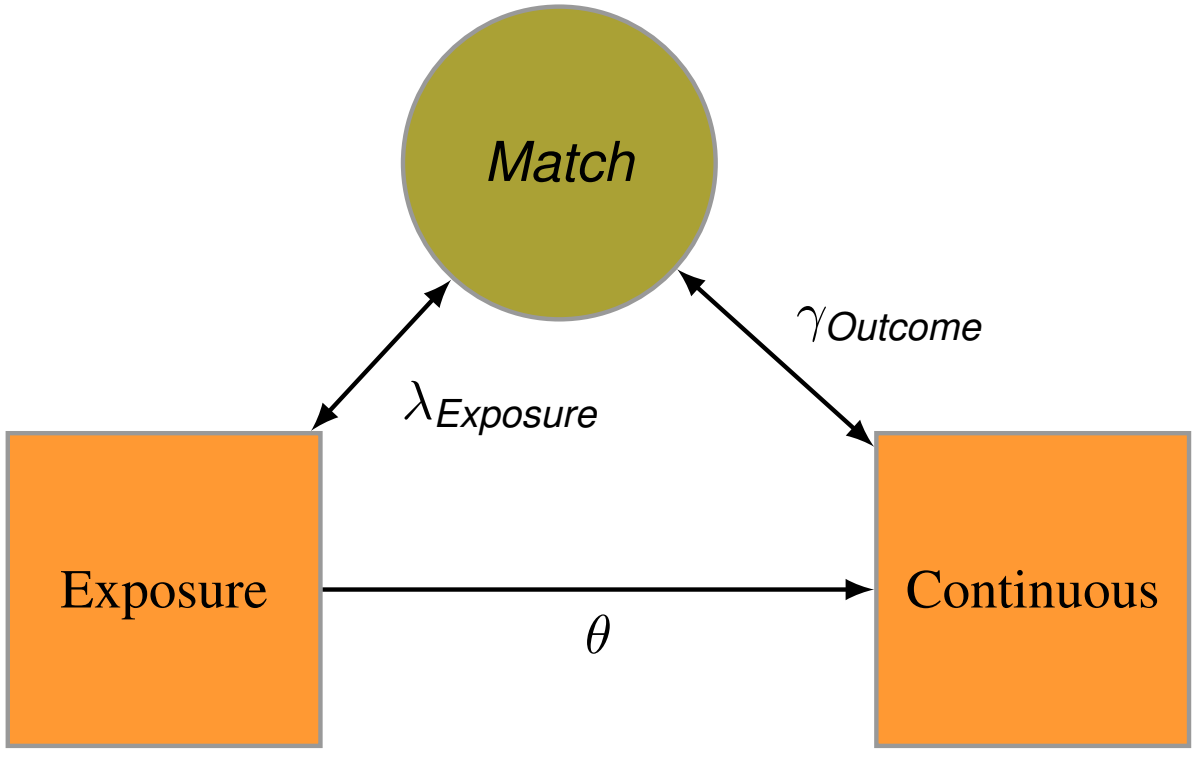
Antibody	Zygosity group	Antibody-positive RA twins with a non-RA antibody-negative co-twin	Antibody-positive non-RA twin with an RA antibody-negative co-twin	OR for RA (95% CI) according to presence or absence of antibody		Ratio DZ/MZ ORs
AKA	DZ	14	0	179.4	(3.6 to 591.9)	46
	MZ	3	1	3.9	(0.5 to 17.8)	
CCP	DZ	15	0	290.0	(3.9 to 622.0)	2
	MZ	6	0	146.1	(1.4 to 281.6)	
IgA-RF	DZ	16	0	204.5	(4.1 to 660.8)	29
	MZ	6	1	7.0	(0.9 to 30.4)	
IgM-RF	DZ	18	2	9.5	(2.3 to 32.1)	3
	MZ	7	2	3.0	(0.8 to 13.0)	

Owing to cells with zero observations, OR and 95% plausibility intervals are estimated from standard Bayesian techniques (assuming a Dirichlet-multinomial distribution of probabilities of counts).[10]
AKA, antikeratin antibody; CCP, cyclic citrullinated peptide; DZ, dizygotic; MZ, monozygotic; RA, rheumatoid arthritis; RF, rheumatoid factor.

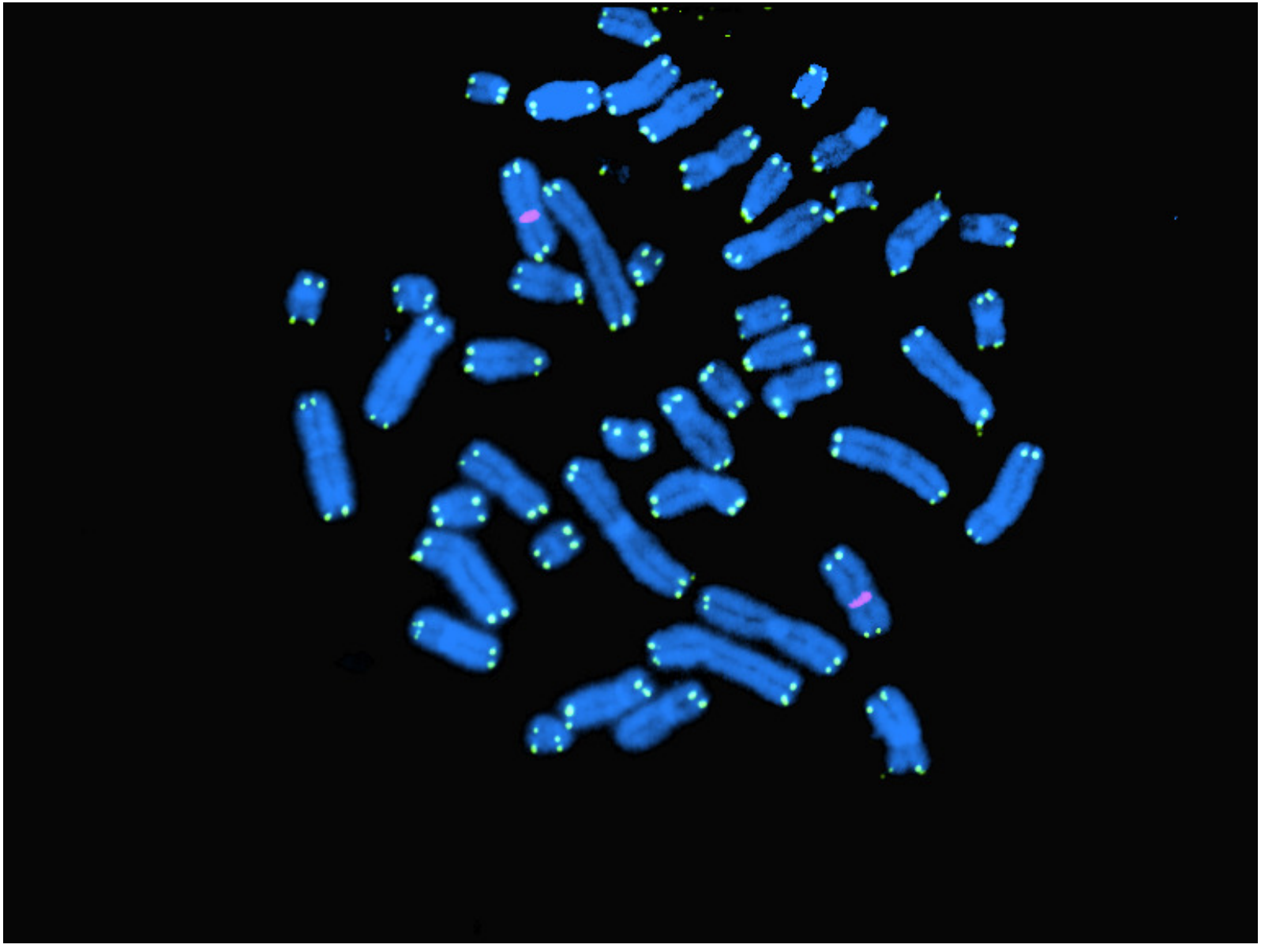
Svendsen et al. Ann Rheum Dis 2011: 70:708-709

-use function ’rdirichlet’ from the R package ’LearnBayes’.

Analysis when Matching



Human telomeres $= (\mathrm{TTAGGG})_n$ (approx. $5 - 12$kb)



Telomere length

- Inversely related to chronological age
- Attrition due to replication of the cell and oxidative stress
- Highly heritable and linked with the X-chromosome
- Females have longer telomere length (Aviv)
- Is telomere end the end? (the cell will no longer replicate)
- Biological marker of aging?

Human telomeres $= (\mathrm{TTAGGG})_n$

- Telomere length:
  - 65% of variation is explained by genetic effects.
  - 20% common environmental effects.
  - 15% individual environmental effects.
- Attrition rate in 10 year follow-up:
  - 30% of variation in change by genetic effects.
  - 70% individual environmental effects.

Hjelmborg, Christensen, Aviv, et al. Journal of Medical Genetics (2014).

Telomeres and survival

- Telomere length and survival: (Cox regression)

TRF length	male ($N = 180$)	female ($N = 368$)
hazard rate	0.85 (0.56, 1.28)	0.62 (0.44, 0.88)

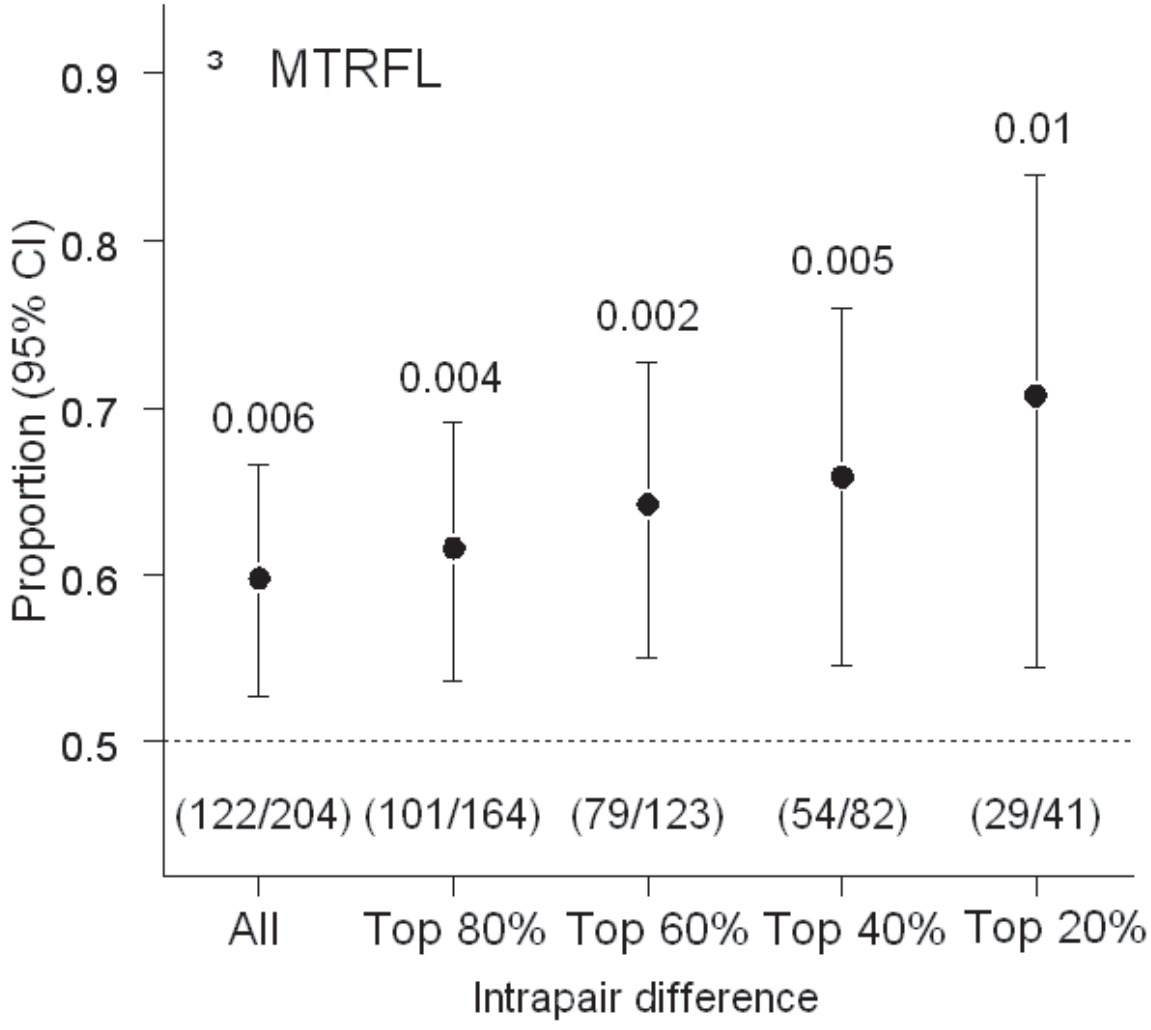
(Masayuki et al. 2008)

- However, the case cotwin survival analysis using all pairs shows:

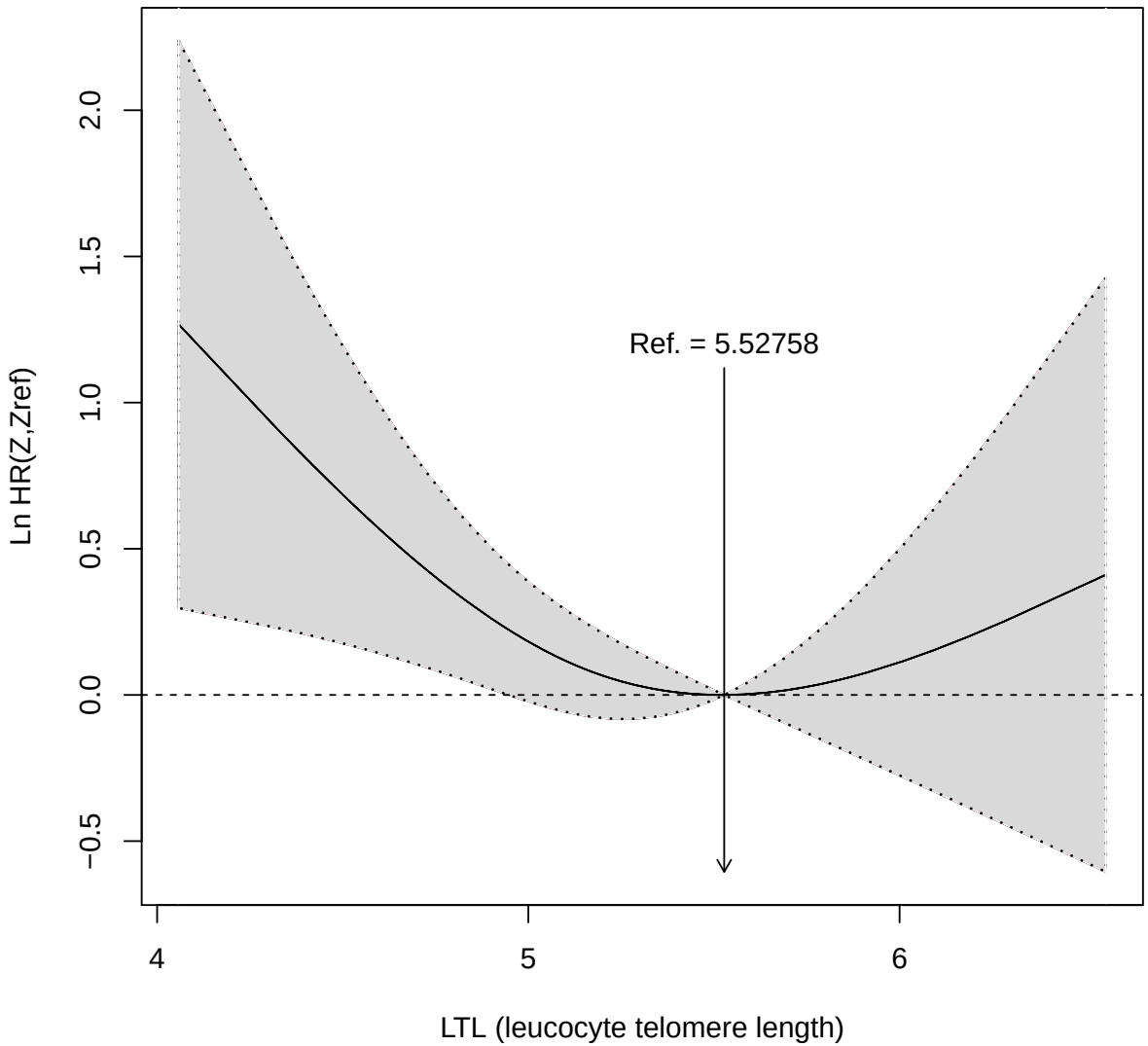
```
. stcox mtrf1, strata(tvparnr) hr nolog  exactp

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  TRF length |   .4707662   .1508366    -2.35   0.019     .2512318    .8821366
------------------------------------------------------------------------------
```

- Now the pairs are matched on gender etc.
- Does the twin with longest leukocyte telomere also live longer?

Telomeres and survival



Leucocyte telomere length survival



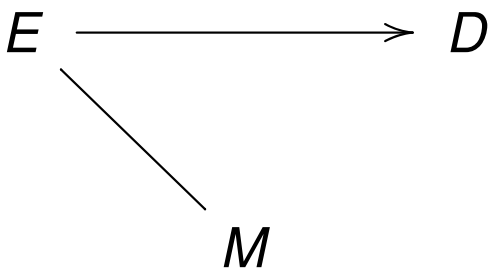
Hjemborg, Christensen, Aviv et al.

The New Yorker June 2005

"You're fifty-seven years old. I'd like to get that down a bit."

Pro and Con

- Efficiency gain: More precise estimate of effect measure.
- Overmatching: Matching variable is associated with exposure and not with disease,



  Narrows the range of exposure, hence loss in precision. Method still correct.
- Non-shared confounding: Pairs might be discordant for confounders as well. In comparison with usual regression:
  - less bias when set of confounders is more strongly shared (correlated) by siblings than the exposure.
  - see Duffy (1994) and Frisell (2012).

Conclusion

- Case cotwin control studies as a powerful tool for studying association.
- A matching variable must be regarded as a confounder.
- Efficiency gain: More precise estimate of effect measures.
- Discordant pairs for at trait may be highly informative and compared wrt. multiple exposures.
- Difference between MZ and DZ effect measure may indicate common genetic effects for the traits.

Appendix: R code page 1

```
# Case co-twin analysis
#
# Dyslexia, ottitis medea and dysphasia (Steen Fibiger, 2011)

library(survival)
library(lme4)
library(mets)
# install packages by eg.: install.packages("lme4")


load("fibigerSim")
str(fib2)
head(fib2)
fib <- fib2

table(fib$sex)

with(fib,table(ottitis,dyslexia))

## Classical estimation of logistic model
# (ignoring dependence within pairs)
glm <- glm(dyslexia ~ ottitis+factor(sex), family=binomial, data=fib)
print(summary(glm))
exp(glm$coef[2])
exp(glm$coef[2]+c(-1,1)*1.96*summary(glm)$coef[4])
exp(cbind(OR = coef(glm), confint(glm)))  # based on profile likelihood.


# We could model the within-pair dependence by
# the following random effects model.
# (or perform robust variance estimation clustering on twin pairs).
glmr <- glmer(dyslexia ~ ottitis+factor(sex)+ (1|tvparnr), family=binomial,
      data=fib)
summary(glmr)
```

Appendix: R code page 2

```
##
## Now, for conditional logistic model:
fib$dys <- (fib$dyslexia=="yes")*1
with(fib,table(dys,dyslexia))
glmc<-clogit(dys~ ottitis+factor(sex) + strata(tvparnr),data=fib)
summary(glmc)

# including dysphasia as well
glmcd<-clogit(dys~ ottitis*dysphasia+sex + strata(tvparnr),data=fib)
summary(glmcd)

anova(glmcd,glmc)

glmcdfin <-clogit(dys~ ottitis+dysphasia+sex + strata(tvparnr),data=fib)
summary(glmcdfin)

anova(glmcdfin,glmcd)

#model-check etc. to follow. eg. Hosmer Lemeshow test
```

Appendix: R code page 3 (extra)

```
# How about pairwise dependence?
# Pairwise odds-ratio model, POR:

theta.des <- model.matrix( ~-1+factor(zyg),data=fib)
margbin <- glm(dys~ ottitis+dysphasia+sex,data=fib,family=binomial())
bin <- binomial.twostage(margbin,data=fib,
         clusters=fib$tvparnr,theta.des=theta.des,detail=0,
                 score.method="fisher.scoring")
summary(bin)
summary(margbin)


### does dysphasia influence the dependence

#twinstut$cage <- scale(twinstut$age)
theta.des <- model.matrix( ~-1+factor(zyg)+dysphasia,data=fib)
bina <- binomial.twostage(margbin,data=fib,
         clusters=fib$tvparnr,theta.des=theta.des,detail=0,
                 score.method="fisher.scoring")
summary(bina)

theta.des <- model.matrix( ~-1+factor(zyg)+factor(zyg)*dysphasia,data=fib)
binai <- binomial.twostage(margbin,data=fib,
         clusters=fib$tvparnr,theta.des=theta.des,detail=0,
                 score.method="fisher.scoring")
summary(binai)

# ottitis
theta.des <- model.matrix( ~-1+factor(zyg)+ottitis,data=fib)
bino <- binomial.twostage(margbin,data=fib,
         clusters=fib$tvparnr,theta.des=theta.des,detail=0,
                 score.method="fisher.scoring")
summary(bino)

theta.des <- model.matrix( ~-1+factor(zyg)+factor(zyg)*ottitis,data=fib)
binoi <- binomial.twostage(margbin,data=fib,
         clusters=fib$tvparnr,theta.des=theta.des,detail=0,
                 score.method="fisher.scoring")
summary(binoi)
```