

## Index of Lecture 13a

Page	Title
1	Practical information
2	Introduction to matching
3	Matching in cohort studies
4	Example 13.2 — matching in cohort study
5	Matching in case-control studies
6	Example 13.3 — matching in case-control study
7	Pros/Cons of matching
8	Adjusting for unmeasured confounder
9	Causation scenarios — introduction
10	Exposure-independent variable $Z$
11	Distorter variable $Z$
12	Moderator variable $Z$
13	Suppressor variable $Z$
14	Confounding & interaction summary

## PRACTICAL INFORMATION

### Today's session:

- review of interaction types for risk (L12b–7),
- matching in cohort and case-control studies,
- note on dealing with unmeasured confounders,
- review-type material:
  - \* different causation scenarios,
  - \* operational summary of confounding & interaction.

### Friday's session:

- in small computer lab (225N),
- homework for you:
  - \* Problem 13.9,
  - \* Problem on matching (webpage),
  - \* prepare a causation scenario (L13a–9),
- questions/discussion of Chapter 13 material,
- work on remaining problems for Chapter 13 (webpage),
- evaluation: course + all instructors.

# INTRODUCTION TO MATCHING

## Matching

= idea/principle for study design, to account for potential effect of confounder(s)  $Z$  for exposure/treatment variable(s)  $E$ :

confounding can be eliminated (reduced) by making the  $E$ -groups being compared as equal as possible with respect to  $Z$ ,

i.e., matching attempts to eliminate association between  $Z$  and  $E$ ,

- varies in its implementation in different study designs,
  - \* experiments/randomised trials: matching  $\sim$  blocking by strata of  $Z$  (treatments are randomised within blocks),
  - \* observational studies discussed on following slides,
- individual matching: individuals are selected in pairs (or clusters, if not 1-1 matching) with the same value of  $Z$ ,
- frequency matching: treatment/exposure groups have the same distribution of  $Z$ ,
- caliper-matching: for a continuous  $Z$ , define groups/ranges for  $Z$  to perform the matching on.

Matching — “a double-edged sword”,

- can have both strong advantages and disadvantages,
- some disagreement among epidemiologists on how much to use it,
- will be treated in more detail in VHM 812 (Winter semester).

## MATCHING IN COHORT STUDIES

### Matching procedure:

1. select random sample among exposed individuals,
  2. select sample among unexposed individuals subject to the restriction:  
distribution of confounder(s) should be same as for exposed individuals (observed/theoretically),
- illustrated on next slide in our BRD–Mh–BRSV causal complex and with constructed data.

### Notes:

- if succesful, the matching eliminates confounding and bias (crude and stratified risk measures are the same),
- (additional) analytic control for the confounder may still be produce a smaller variance estimate,
- the matching is undertaken at study onset, and is therefore independent of the outcome,
- examples of commonly used matching variables: age, breed, sex, farm, region/neighbourhood; also stage of disease has been used.

EXAMPLE 13.2 — MATCHING IN COHORT STUDY
---

Aim: study design with 500 Mh+ and 500 Mh−, and frequency matching of Mh− group to distribution of BRSV in Mh+ group:

- prev. of BRSV among Mh+: 0.85 (5500/6500, Ex. 13.1),
- expected no. of BRSV+ individuals in Mh+ group:  
 $0.85 \cdot 500 = 425$ ,
- need to have also 425 BRSV+ indiv. in Mh− group<sup>1</sup>,
- expected counts (based on BRD risks, Ex. 13.1):

BRSV (Z)   BRD (Y)		Mh (E)		total
		+	−	
+	1			
	0			
total		425	425	850

BRSV (Z)   BRD (Y)		Mh (E)		total
		+	−	
−	1			
	0			
total		75	75	150

- stratum-specific OR  $\approx 2$ , and crude OR  $\approx 2$   
 $\Rightarrow$  no confounding bias.

---

<sup>1</sup> In general, we need to have the same prevalence, but as the E+ and E− groups are of the same size, the numbers are also the same.

## MATCHING IN CASE-CONTROL STUDIES

### Matching procedure:

1. select random sample among cases (possibly all cases),
2. select sample among controls subject to the restriction: distribution of confounder(s) should be same as for cases (observed/theoretically),

— illustrated on next slide in our BRD–Mh–BRSV causal complex and with constructed data.

### Notes:

- if succesful, the matching eliminates confounding but will introduce selection bias for the exposure:
  - \* bias for association  $E \rightarrow Y$  in the direction of the null,
  - \* increases with source population strength of  $E - Z$  association,
  - \* intuitive explanation: the matching of controls by  $Z$  will make distribution of  $E$  among controls more similar to that of  $E$  among the cases (because of the  $E - Z$  association)  $\Rightarrow$  reduced effect of  $E$ ,
- analytic control for the confounder is necessary (and eliminates the selection bias).

EX. 13.3 — MATCHING IN CASE-CONTROL STUDY

Aim: study design with 280 BRD cases<sup>2</sup> and 280 controls, and frequency matching of controls to distribution of BRSV among cases:

- prev. of BRSV among cases: 0.82 (230/280, Ex. 13.1),
- expected no. of BRSV+ individuals among cases: 230,<sup>3</sup>
- need to have also 230 BRSV+ indiv. among controls<sup>4</sup>,
- expected counts (based on BRD risks, Ex. 13.1):

BRSV (Z)   BRD (Y)		Mh (E)		total
		+	–	
+	1	220	10	230
	0			230
BRSV (Z)   BRD (Y)		Mh (E)		total
		+	–	
–	1	20	30	50
	0			50

- stratum-spec. and M-H OR  $\approx 2.1$ ; crude OR  $\approx 1.6$   
 $\Rightarrow$  adjustment for Z required,
- selection bias for Mh:

$$\begin{aligned} \text{study : } p(\text{Mh} + | \text{BRD} -) &= 0.79 \quad ((210+12)/280), \\ \text{popul. : } p(\text{Mh} + | \text{BRD} -) &= 0.64 \quad (6260/9720). \end{aligned}$$

<sup>2</sup> All cases in entire population.

<sup>3</sup> Number in total population because all cases included in study; in general, prevalence times no. of cases.

<sup>4</sup> In general, we need to have the same prevalence, but as the case and control groups are of the same size, the numbers are also the same.

## PROS/CONS OF MATCHING

### Advantages:

- may lead to gain in efficiency when random sampling would lead to highly unbalanced and/or sparse data,
- matching on broad variables like farm og neighbourhood may provide efficient adjustment for largely unmeasurable factors,
- matching on convenient variables like closest case may facilitate the logistics of data collection,
- does not preclude analytic control for (additional) confounders (so matching is only for strongest confounders).

### Disadvantages:

- effect of matched factor(s) on the outcome is lost (but interactions are possible),
- low efficiency for factors with strong link to matching variable (possible *overmatching*),
- can be difficult/costly to find appropriate subjects,
- for pair-matching in a case-control design, the matched analysis may have very low efficiency.

### Recommendation against matching on

- factors whose confounding effect is uncertain or weak,
- intervening variables (!), and uncertain/weak risk factors.

## ADJUSTING FOR UNMEASURED CONFOUNDER

yes, it *is* possible, but requires strong external information:

- (i) assume confounding effect (but no interaction) of  $Z$ ,
- (ii) assume known distribution of  $Z$  in exposed and non-exposed groups,
- (iii) assume known association between  $Z$  and  $Y$  after adjustment for exposure,
  - \* information for (ii) and (iii) can perhaps be found in literature and assumed valid for actual study,
  - \* approach invites sensitivity analysis: to try a range of different values for (ii) and (iii), and note the impact on results.

How does it work?

- example given in VER (Ex. 13.10) for case-control study and dichotomous confounder<sup>5</sup>, (constructed data)
- step 1: assumed prevalences of confounder allow calculation of expected no. of non-cases (within  $E$  groups),
- step 2: assumed OR for association ( $Z, Y$ ) within strata of  $E$  allows calculation of expected no. cases (within  $E$ ),
- compute quantities of interest from constructed tables.

---

<sup>5</sup> Rothman & Greenland, Chapter 19, gives more detail, including extensions to other study designs and multiple-category confounders.

## CAUSATION SCENARIOS — INTRODUCTION

Purpose: to provide an overview of different relationships between variables, in order to

- aid in interpretation of findings in the data analysis<sup>6</sup>,
- clarify terminology<sup>7</sup> and concepts.

Methods:

- causal diagrams: single- and double-headed arrows, line (non-headed arrow (!)), absence of arrow,
- Venn diagrams: circles representing factors, amount of overlap indicating strength of association (zero to total), both with a temporal dimension (from left to right).

List of scenarios: ( $Y=BRD$ ,  $E=Mh$ ,  $Z=BRSV$ )

1. Exposure-independent variable — next slide,
2. Simple antecedent variable,
3. Explan. antecedent variable – complete confounding,
4. Explan. antecedent variable – incomplete confounding,
5. Intervening variable,
- 6.–8. Distorter, Suppressor and Moderator variables — following slides.

Idea: Scenarios 2–5 are for student preparation (Friday).

<sup>6</sup> We caution that inferring causal structure from data has some limitations; see Thompson (1991) paper referred in VER.

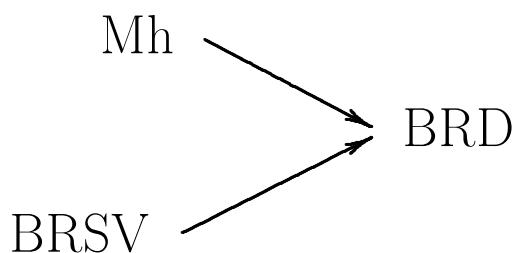
<sup>7</sup> Terminology is from VER, and not all is in general use in epidemiology.

## EXPOSURE-INDEPENDENT VARIABLE $Z$

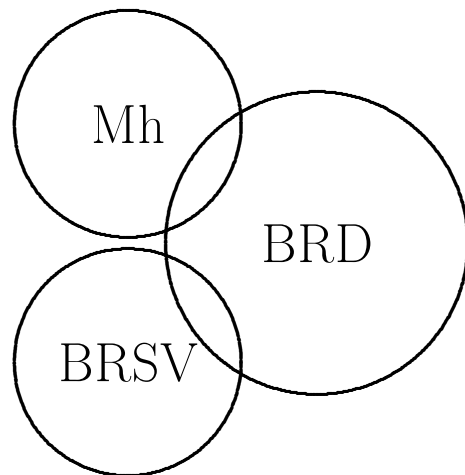
Assumption:

no relation between BRSV and Mh:

Causal model:



Statistical model:



- \* independence of BRSV and Mh: no connecting arrow (left), disjoint circles (right),
- \* associations of Mh and BRSV with BRD: directed arrows (left), overlapping circles (right).

Implications for (statistical) modelling:

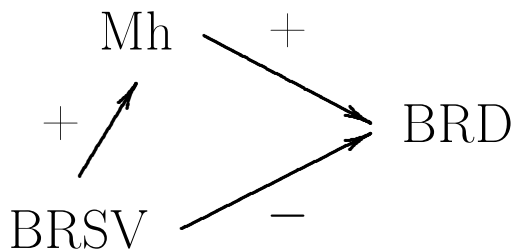
- crude and stratified estimates for Mh should be similar,
- inclusion of BRSV in model may reduce residual variation  $\Rightarrow$  improve estimation precision.

## DISTORTER VARIABLE $Z$

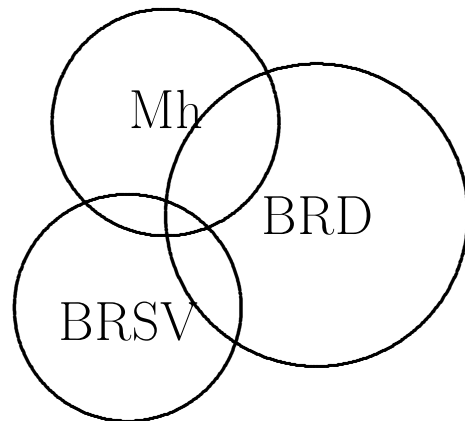
Assumption:

confounding by  $Z$  (BRSV), and different directions (signs) on some of the causal relations:

Causal model:



Statistical model:



- \* positive associations (Mh, BRD) and (BRSV, Mh), but negative association (BRSV, BRD)<sup>8</sup>.

Implications for (statistical) modelling:

- control for confounding by BRSV required,
  - \* will usually increase the strength of association,
  - \* may change the direction of association from negative to positive (or vice versa, for different distribution of + and -),
- for an artificial data example, see Problem 13.2.

---

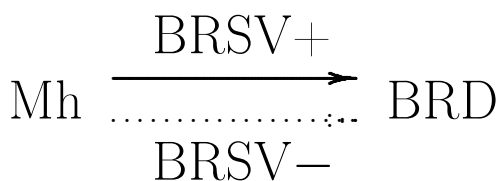
<sup>8</sup> Similar effects are produced by other sign-reversed relations, e.g. a positive association (BRSV, BRD) and a negative association (BRSV, Mh).

MODERATOR VARIABLE  $Z$

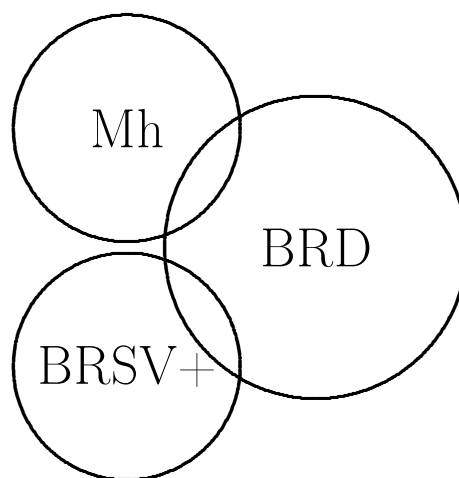
Assumption:

interaction between  $Z$  (BRSV) and  $E$  (Mh):

Causal model:



Statistical model:



- \* positive association (Mh, BRD) only for BRSV+,
- \* statistical model for BRSV- (not shown): no association between Mh and BRD (disjoint circles),
- \* other interaction effects possible, but would not be called moderator effects (in VER terminology).

Implications for (statistical) modelling:

- statistical model needs to include BRSV+ (either as single term, or as part of the interaction  $BRSV * Mh$ ),
- test for homogeneity ( $X^2_{hom}$ ) statistically significant,
- Mh effect reported separately for BRSV+ and BRSV-.

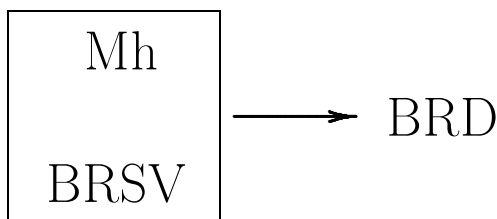
## SUPPRESSOR VARIABLE $Z$

### Assumptions:

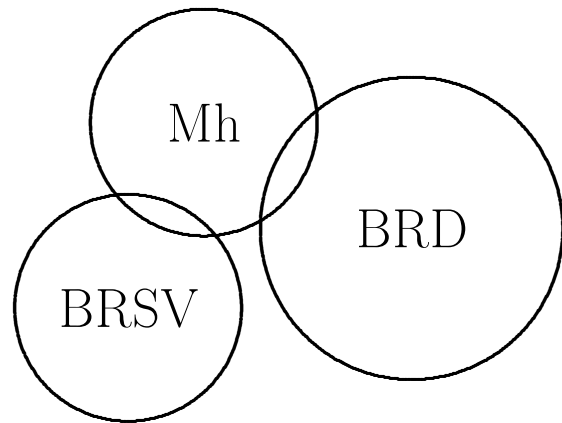
no relation between BRSV and BRD, but Mh and BRSV measured only via “proxy variable”:

### Causal model:

cattle contact



### Statistical model:



- \* indirect measurement of BRSV and Mh by proxy variable (in the example, cattle contact).

### Implications for (statistical) modelling:

- need to refine measurement of predictors whereby the (irrelevant) BRSV component is eliminated,
- statistical association of proxy variable with BRD will increase by removal of BRSV,
- note: suppression of outcome possible as well (e.g., by including multiple reasons of mortality of which only one is of interest).

## CONFOUNDING & INTERACTION SUMMARY

Setting: risk outcome  $Y$ , binary/categorical exposure  $E$ , binary/categorical potential confounder  $Z$ .

Approach of Chapter 13 (without multivariable modelling):

1. Is interaction  $Z * E$  present?
  - \* subjective assessment: compare stratum-specific estimates for  $E \rightarrow Y$  (RD, RR, OR, IRR),
  - \* statistical assessment: compute chi-square test ( $X_{\text{hom}}^2$ ) for homogeneity of across strata,
    - 1a. interaction present: report stratum-specific point estimates instead of a single estimate,
    - 1b. interaction absent: compute stratified estimate (Mantel-Haenszel estimate), and proceed to Question 2,
2. Is  $Z$  a confounder for the relation  $E \rightarrow Y$ ?
  - \* check necessary conditions for confounding (using graphical approach), and use the data to assess the relations between  $Z$  and both  $E$  and  $Y$  (recall slightly different approaches for  $Z - E$  relation in cohort and case-control studies),
  - \* assess the relative difference between the crude estimate and M-H estimate for the association  $E \rightarrow Y$  (“20-30% rule”),
    - 2a. confounding present: report the M-H estimate and the evidence obtained for confounding,
    - 2b. confounding absent: report the crude estimate, and note that  $Z$  was not found to be a confounder (perhaps only in cases where a confounding effect might have been expected).