SELECTION OF COMPARISON GROUPS.

INDIVIDUAL-LEVEL CAUSAL EFFECTS: effect of {trt / exposure} on the same individual.

POPULATION-LEVEL CAUSAL EFFET: average effect of {trt / exposure} on group.

COUNTERFACTUAL:
Simultaneously examine the {pop / ind} {with / without} {trt / exposure}
↓
impossible, so impossible to be certain
↓
Choose comparison groups to approximate the counterfactual ⟶

ASSUMPTION OF COMPARABILITY OF GROUPS
ASSUMPTION of NO RESIDUAL CONFOUNDING.
↓
Assumption probably violated → bias {point estimates / measures of variability}

→ ↑ Type I error
C.I don't achieve nominal coverage ?
↓
C.I can achieve nominal coverage w randomization

Choice of comparison groups

- INDIVIDUALS : select very similar ind to the trt / exposed ind (clone mice, twins)

- POPULATIONS :

- CLINICAL TRIALS :

■ COMPARABILITY OF EFFECT:

- trt effect : placebo
- expose effect : reference category

■ COMPARABILITY OF POPULATION : randomization.

Avg. incidence of the outcome under non-exposure in all groups is =

Does not eliminate confounding → ↑ n ↓ chance of imbalance = confounding.

Imbalance is caused by differential selection into and out of compared groups, w/ the selection being related to risk factors of the outcome.

Ex: healthy worker effect : ↓ diseased ind. at workplace compared to general pop.

when pops are comparable → no confounding

■ COMPARABILITY OF INFORMATION : blinding.

STUDY BASE PRINCIPLE:

- PRIMARY DEFINITION: 1st define cohort → 2nd identify cases

sum of the individual-level experience ind. / time "
captured through census / sample
as basis for learning about the
reference population it represents

sample → study results → extrapolate → pop that is represented by the sample = "study base"

≠ statistical sample of target pop
= direct referent of the empirical results

STUDY BASE

CLINICAL TRIAL — follow-up experience of enrolled patients.

- COHORT — person-time experience = incidence rates of the entire cohort (including censored ind)
- CASE-CONTROL — ?
  → need to conceptualize case-control as a cohort study to understand study base.

- SECONDARY DEFINITION : 1st identify cases → 2nd define study base

≠ pop. experience from which cases could have come to be included in the study (registry, etc...)

= totality of experience from which each potential case HAD IT OCCURRED would have been included in study.

- Each ind is in a study base at time t_i if were it to become a case, it would be included in study

Ex.: diabetic dogs sampled for Davis vet hospital
study base = all dogs who, were they to become diabetic, would show up at Davis vet hospital

study base is limited { geographically (hospital catchment area)
temporally (study time)

⇓

used in

CASE-CONTROL STUDIES = TROHOC

Alt to cohort studies when dis. is rare.

Recruit all cases → draw sample of study base
cases → controls

constrained by time / place

can yield complete info on the study base.

N° of controls : depends on prevalence of exposure in study base and magnitude of OR.

r = nº of cases

r
r+1

r = 1 → 1/2 → 50% of info from using whole pop.
r = 2 → 2/3 → 60%
r = 3 → 3/4 → 75%
r = 4 → 4/5

→ Max 4 controls

- The case-control approach is particularly valuable when:
  - It is expensive, time-consuming, and inefficient to put an entire cohort under surveillance for disease incidence.
  - The study base is large enough so that adequate information about exposure distribution can be learned by studying a small fraction of it.

How to sample controls ? (no "sampling frame")
From study base → pop. from which potential cases would have been captured. =
ind. who had they developed the illness → would have come to my attention as cases.
controls should be distinguished from cases only by a last-minute differential dx.
ex: dx of G.I tumor in FE → study base is FE w/ other tumors dx in the same way (echography).

controls should represent what the case pop. would be if the exposure had not occurred. = Ho →
where cases de independent of exposure (conditionally on confunders)

Ideal control: cases of another disease :

- that derive from same registry
- where disease :
  - is unrelated to the exposure
  - is ≃ to dis of interest regarding selection

factors that conditio it appears in the registry.

Ex: MacMahon

CASES
carcinoma
≤ 79 y.
Pancreatic
cancer
Hospital X
↓
Physicians X
↳ advised NO coffee

CONTROLS
Physician X patients
≈ time case was
enrolled
NO dis → pancreas
→ biliary tract
→ ass. {smoking
alcohol
↓
also coffee

STUDY BASE
carcinoma
≤ 79 y
if they'd had
P. cancer, they'd
have gone to
Hospital X
include dis:
- pancreas
- biliary tract
- ass. {smoking
alcohol

↓↓
Appropriate controls
of exposed ind
→ bias OR away fro

- "However, *there is no virtue in using a wide variety of diagnoses.* Each diagnosis needs to be justified as to the assumptions, and insofar as they are untenable, the errors have no tendency to cancel out on the basis of variety.
- "The need is to have *defensible inclusions*, and there is *no need to defend exclusions.*"

?

controls don't have to come from the study base, they can come from any pop w/ the same distribution of exposure/incidence, conditional on controlled vbles in the analysis (it is necessary to control for known confounders).

Ex:

Incidence of thromboembolism = OUTCOME

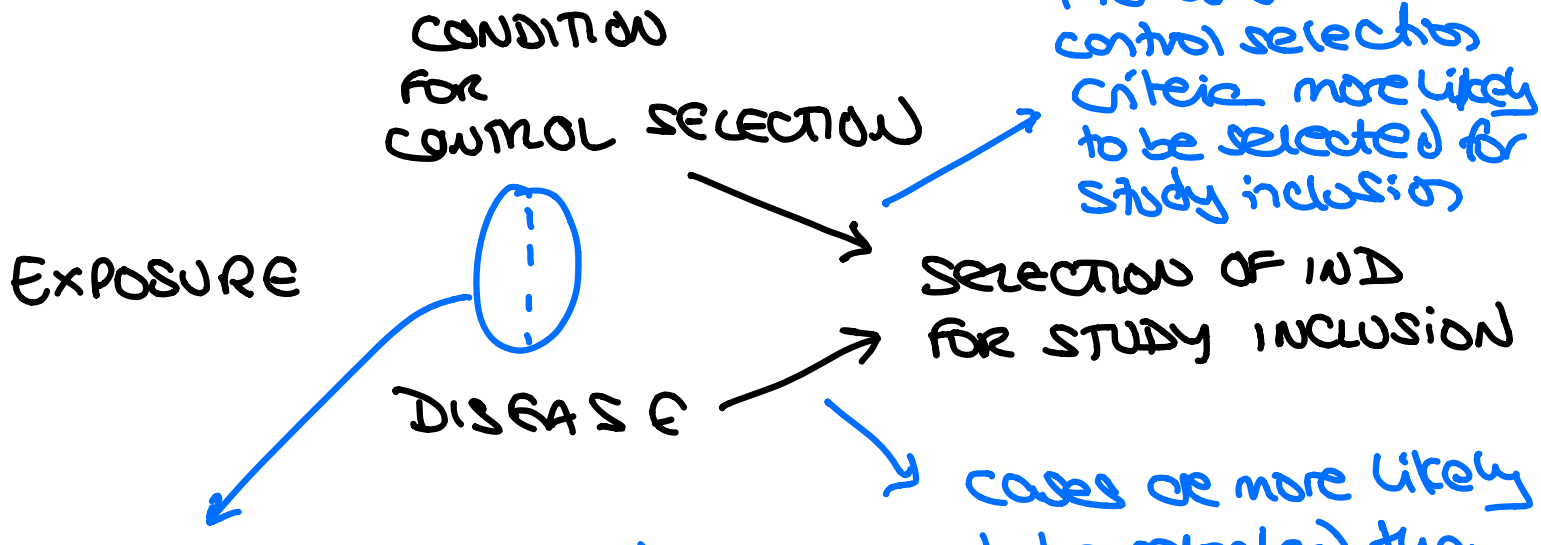
w/ ABO blood type = EXPOSURE

in ♀ using oral contraceptives. = BASE POP.

Control: you could use ♂ or ♀ w/o oral cont.
bc ABO dist = independent of gender, age, oral cont.

≥ 1 control group:

> "Because there tend to be strong feelings for and against various types of control groups by different investigators, an association will have a much greater likelihood of being accepted as valid if it is found when different types of control groups are used."



Marginally unassociated
conditionally associated:
be any exposure-dis association is conditional on being selected for inclusion in the study.

	A=1		A=0	
	B = 1	B = 0	B = 1	B = 0
C = 1	800	600	400	200
C = 0	200	400	600	800
Total	1000	1000	1000	1000

TOTAL 2000 2000 4000

A and B MARGINALLY INDEPENDENT :

Ex: P(A=1|B=1) = 1000/2000 = 0.50 = P(A=1) = 2000/4000 = 0.50
OR (A and B) = (1000/1000) / (1000/1000) = 1

A=1	A=0	
B = 1	B = 1	
800	400	
200	600	
1000	1000	2000

A and B NOT CONDITIONALLY INDEPENDENT.

①

Ex: P(A=1|B=1,C=1) = 800/1200 = 0.67 ≠ P(A=1|C=1) = 1400/2000 = 0.70
Ex: P(A=1|B=1,C=0) = 200/800 = 0.25 ≠ P(A=1|C=0) = 600/2000 = 0.30
Ex: P(B=1|A=1,C=1) = 800/1400 = 0.57 ≠ P(B=1|C=1) = 1200/2000 = 0.60
Ex: P(B=1|A=1,C=0) = 200/600 = 0.33 ≠ P(B=1|C=0) = 800/2000 = 0.40
OR (A and B|C=1) = 0.67 and OR (A and B|C=0) = 0.67

①

	A=1	A=0	
	B = 1	B = 1	
C = 1	800	400	1200

	A=1		A=0		
	B = 1	B = 0	B = 1	B = 0	
C = 1	800	600	400	200	
	1400		600		2000

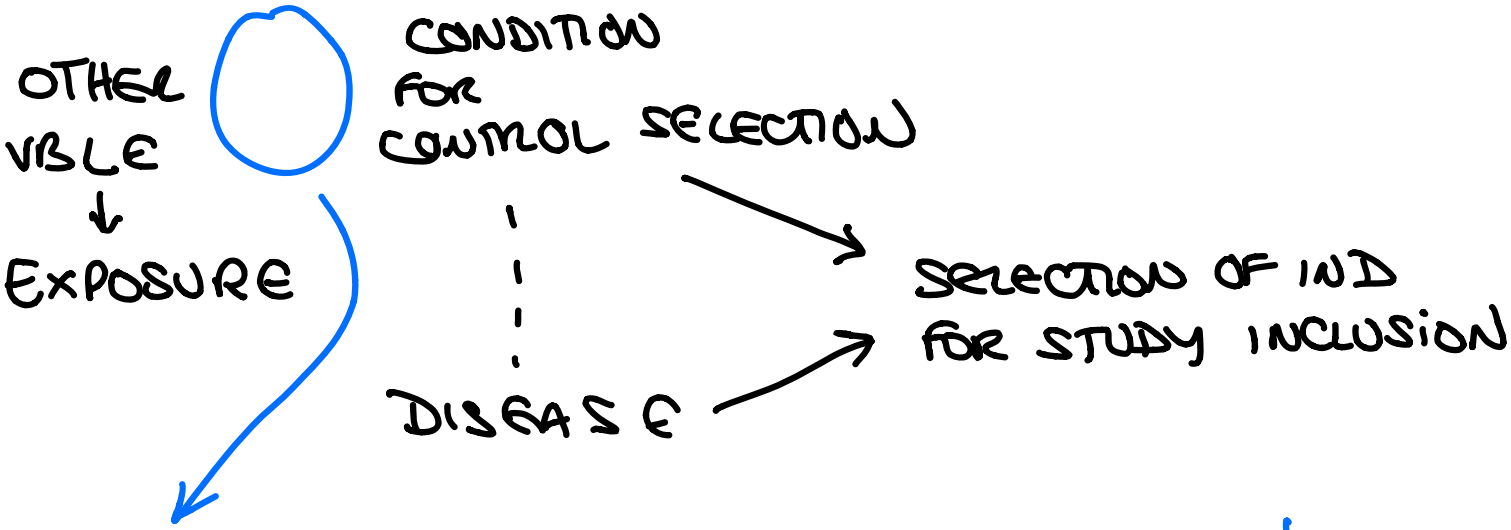
	A=1	
	B = 1	B = 0
C = 1	800	600

A=0

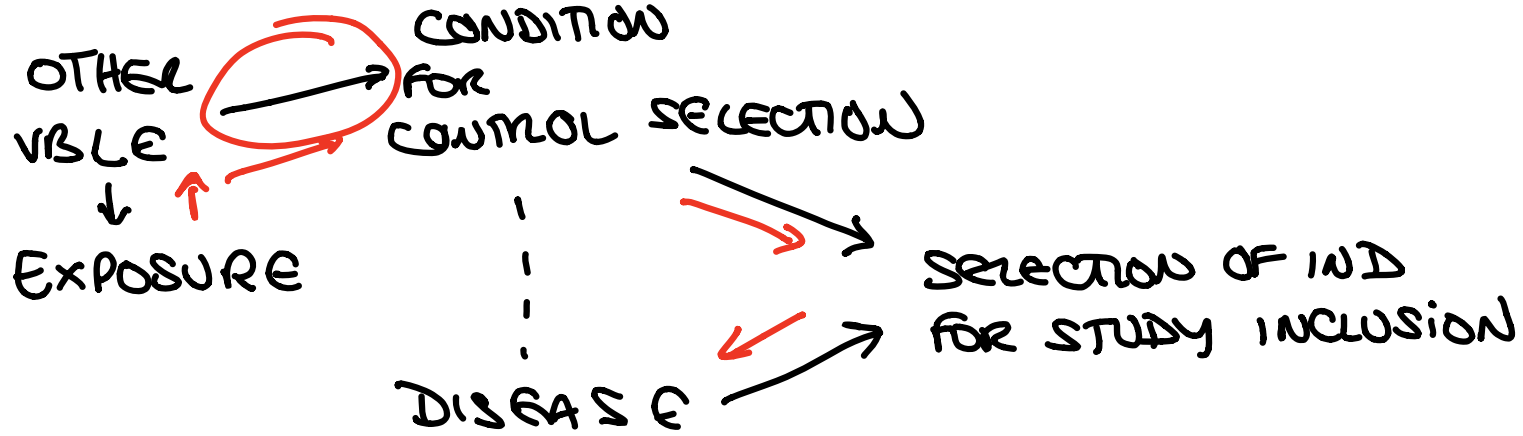
B = 1	B = 0
400	200

$\frac{800 \times 200}{400 \times 600} = 0.67 = OR\ (A\ ^{AND}\ B\ |C=1)$

$\frac{200 \times 800}{400 \times 600} = 0.67 = OR\ (A\ ^{AND}\ B\ |C=0)$

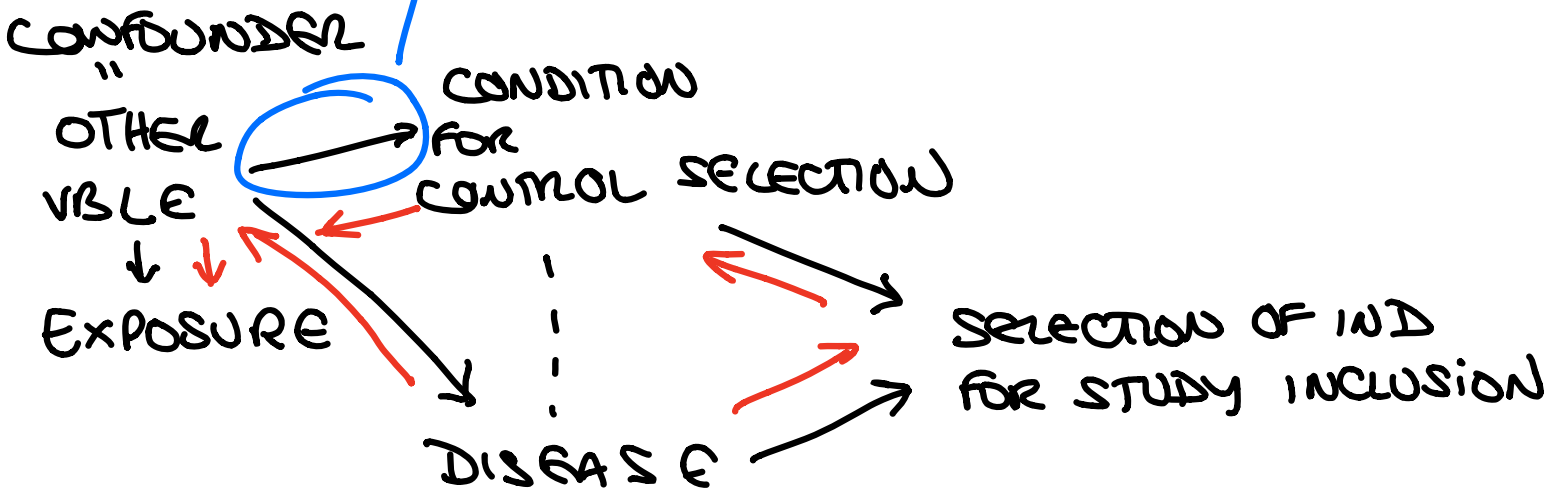


controls poorly selected. No ass. between condition for selection and exposure

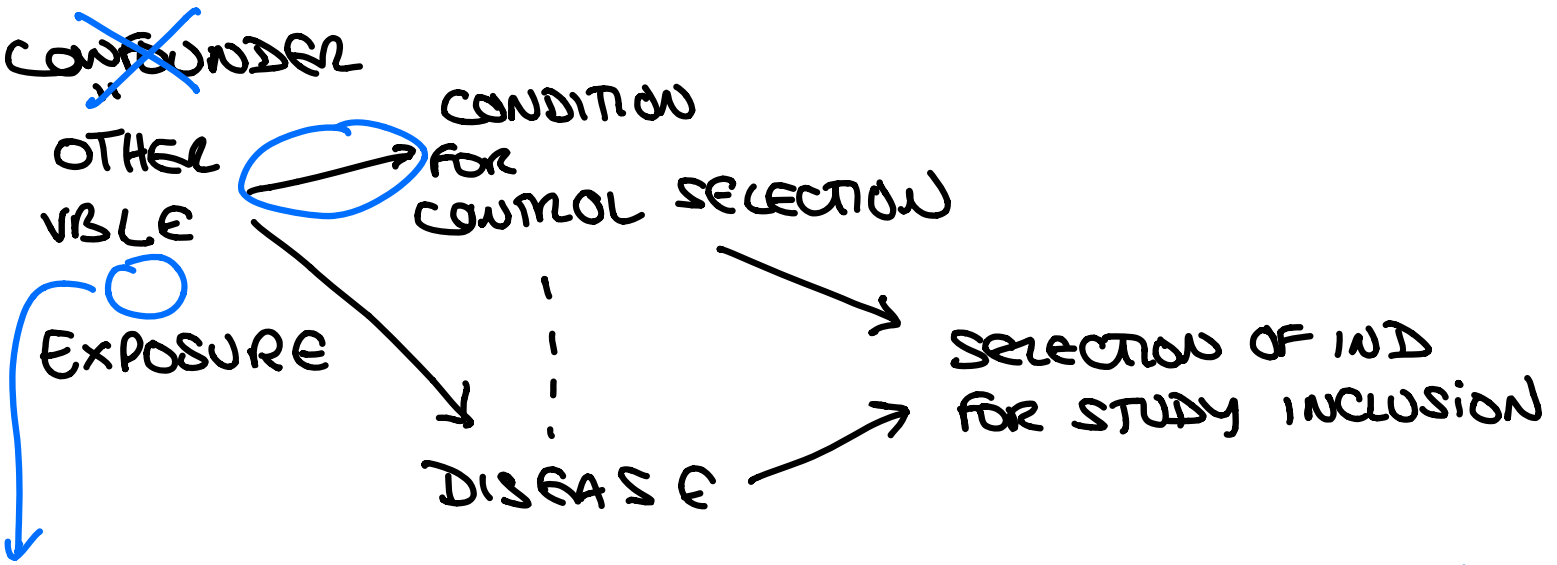


Backdoor path: controls selected for reasons not independent of exposure → exposure → condition → disease are associated → SELECTION BIAS → shows exposure–dis ass even when there is no causal relationship.

MATCHING: cases are matched to controls on a confounder



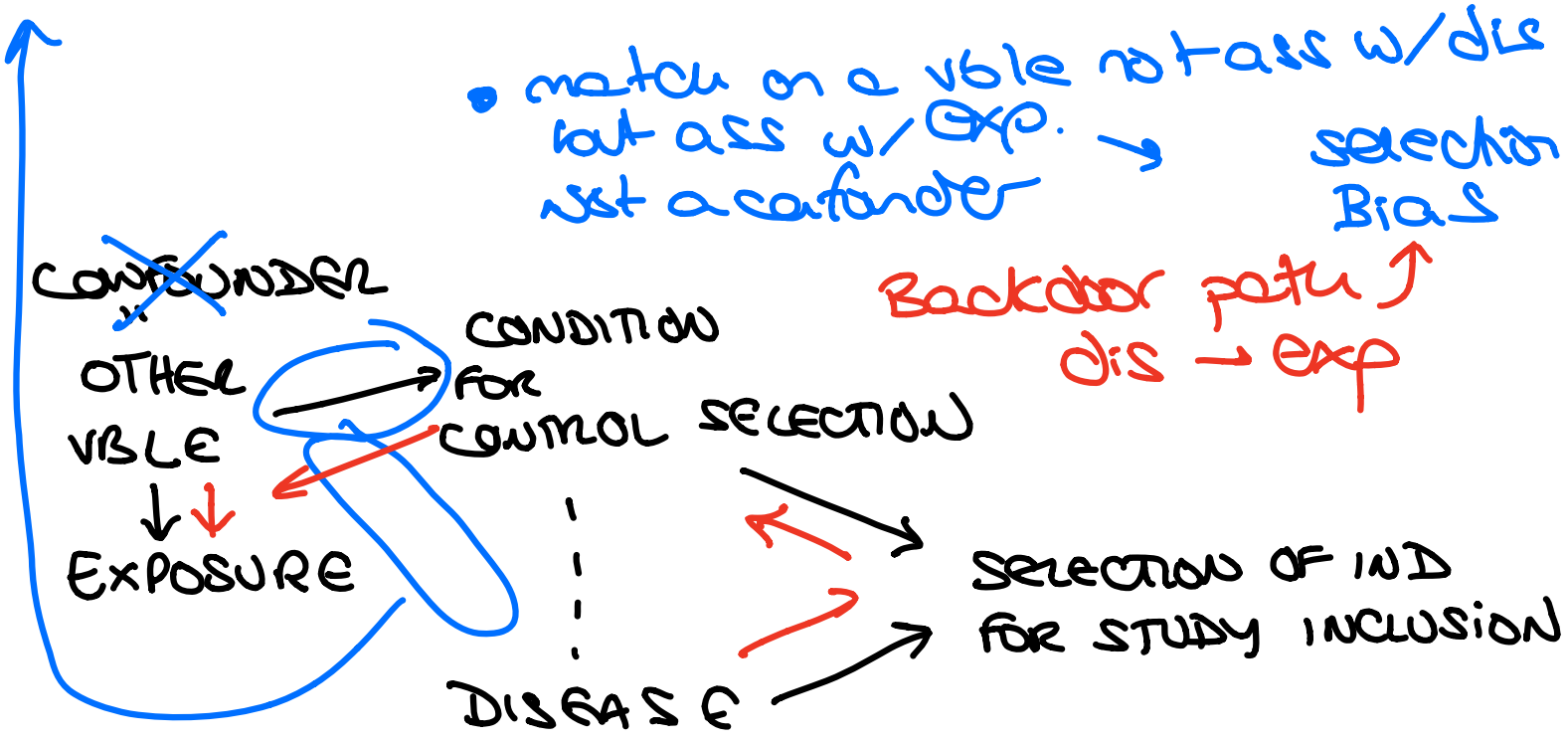
Backdoor path: you still have to control on the confounder, even you match by it, to block the backdoor path DIS → EXP.



OVERMATCHING: match on a vble not ass w/exp.

- not ass w/ dis → no selection Bias

not a confounder



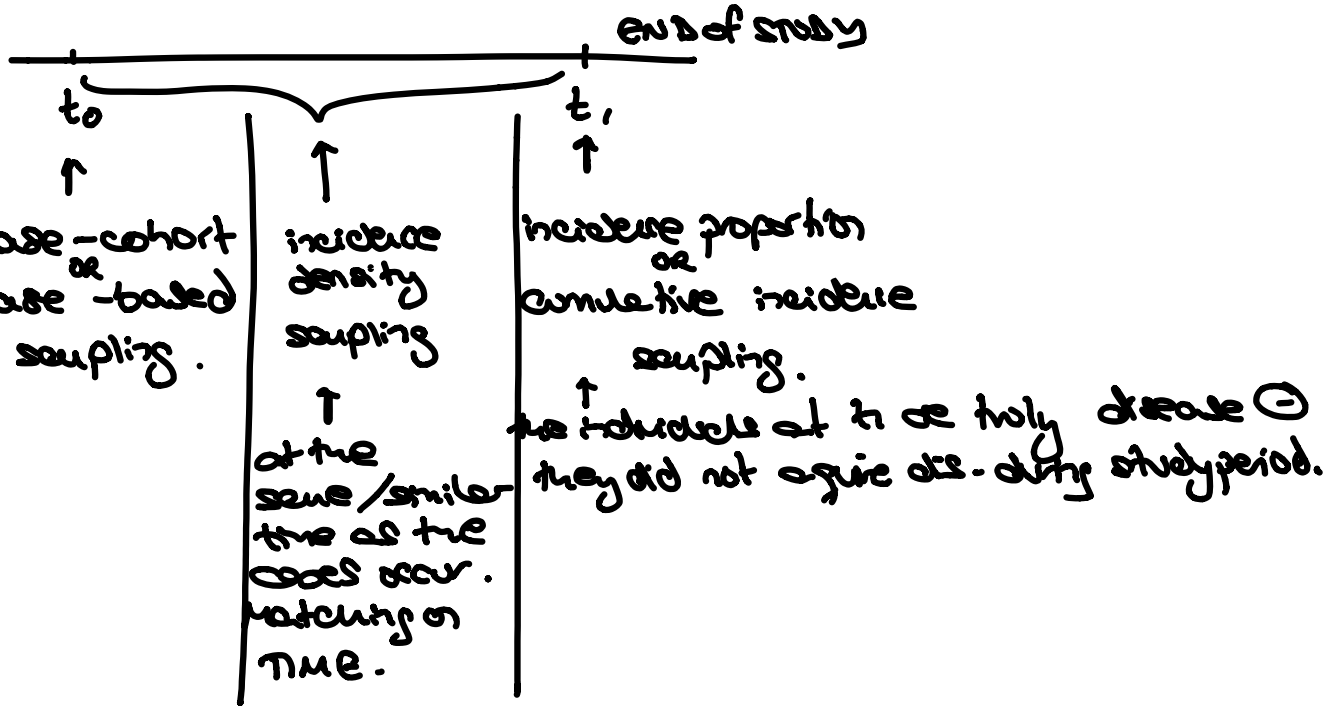
To avoid backdoor path you NEED to control for this vble, but if you hadn't matched you wouldn't have needed to in 1st place.

METHODS OF CONTROL SAMPLING:

- CASE-BASED SAMPLING: to

- INCIDENCE DENSITY SAMPLING: match on time w/cases

- INCIDENCE PROPORTION or CUMULATIVE INCIDENCE SAMPLING: t_1 end of study.



ODDS RATIO

Diseased to nondiseased | Exposure (E^+)

COHORT STUDY

	Exposed E^+	Unexposed E^-
CASE D^+	A_1	A_0
POP. AT RISK D^-	B_1	B_0

COHORT EXPOSURE OR :

$$\frac{\text{odds Diseased | Exposed}}{\text{odds Diseased | unexposed}} = \frac{A_1 / B_1}{A_0 / B_0}$$

Algebraically similar but ≠ means.

CASE CONTROL STUDY

	Exposed E^+	Unexposed E^-
CASE D^+	A_1	A_0
CONTROL D^-	B_1	B_0

CASE-CONTROL EXPOSURE OR :

$$\frac{\text{odds Exposure | case}}{\text{odds Exposure | control}} = \frac{A_1 / A_0}{B_1 / B_0}$$

Exposed to unexposed | $dis^{(+)}_{(-)}$

- CONTROL SELECTION NOT MATCHED ON TIME :

  t_0: CASE-BASED. t_1: INCIDENCE PROP. SAMPLING

ASSUMPTION (1) : a constant IDR ; ie $\frac{ID_1(t)}{ID_0(t)} = \overline{IDR}$ over (t_0, t_1)

$$\overline{IDR} : IDR(t) = \frac{ID_1(t)}{ID_0(t)} \Rightarrow ID_1(t) = \overline{IDR} \cdot ID_0(t)$$

ASSUMPTION (2) $P_i(t) = P_i$ stable pop. assumption with respect to exposure

This means that time is not a CONFOUNDER => we do not have to control for time.

$$\text{OR}_{\text{UNMATCHED}} = \frac{A_1/A_0}{B_1/B_0} = \frac{\int_{t_0}^{t_1} a_1(t)\,dt \int_{t_0}^{t_1} b_0(t)\,dt}{\int_{t_0}^{t_1} a_0(t)\,dt \int_{t_0}^{t_1} b_1(t)\,dt} =$$

$$a_+(t) = a_1(t) + a_0(t)$$

= exposed cases + unexposed cases

CONTROLS

$$\left(\begin{array}{l}\text{total nº of cases}\\ \text{expected per unit of } t\end{array}\right)\left(\begin{array}{l}\text{proportion of } N(t)\\ \text{in exposure group } i\end{array}\right)$$

$$= \frac{\int_{t_0}^{t_1} N(t)\cdot P_1(t)\cdot ID_1(t)\,dt \int_{t_0}^{t_1} a_+(t) P_0(t)\,dt}{\int_{t_0}^{t_1} N(t)\cdot P_0(t)\cdot ID_0(t)\,dt \int_{t_0}^{t_1} a_+(t) P_1(t)\,dt} =$$

CASES

(pop. at risk) (proportion of the pop. in exposure group i) (incidence rate in exposure group i)

$$\overline{IDR} = \frac{ID_1}{ID_0} \Rightarrow ID_1 = \overline{IDR} \times ID_0$$

$$= \frac{[P_1\,\overline{IDR} \int_{t_0}^{t_1} N(t)\,ID_0(t)\,dt]\,[P_0 \int_{t_0}^{t_1} a_+(t)\,dt]}{[P_0 \int_{t_0}^{t_1} N(t)\cdot ID_0(t)\,dt]\,[P_1 \int_{t_0}^{t_1} a_+(t)\,dt]} = \overline{IDR}$$

we have interpreted the $\overline{IDR}$ from the OR without making the rare disease assumption. Importantly, we assumed that time is NOT a confounder.

• CONTROL SELECTION MATCHED ON TIME:

INCIDENCE DENSITY SAMPLING

ASSUMPTION ①: a constant IDR; ie $\frac{ID_1(t)}{ID_0(t)} = \overline{IDR}$ over (t_0, t_1)

$$\overline{IDR}: IDR(t) = \frac{ID_1(t)}{ID_0(t)} \Rightarrow ID_1(t) = \overline{IDR} \cdot ID_0(t)$$

CASE CONTROL STUDY

CASES		CONTROLS: Exposed E^+	CONTROLS: Unexposed E^-
	Exposed E^+	M_{11}	M_{10}
	Unexposed E^-	M_{01}	M_{00}

Matched pairs OR

$\frac{M_{10}}{M_{01}}$ → proves H

M_{01} → disproves H

C	Co
E	ONE
ONE	E

The expected number of pairs · are $(t_0, t+dt)$ in which the case is n exposure group i and the control B is in exposure group j $= a_i(t)\, P_j\, dt$

OR MATCHED $= \frac{\int_{t_0}^{t_1} m_{10}(t)\,dt}{\int_{t_0}^{t_1} m_{01}(t)\,dt} = \frac{\int_{t_0}^{t_1} a_1(t)\, P_0(t)\,dt}{\int_{t_0}^{t_1} a_0(t)\, P_1(t)\,dt} =$

$$= \frac{\int_{t_0}^{t_1} N(t) \cdot P_1(t) \cdot ID_1(t) \cdot P_0(t)\,dt}{\int_{t_0}^{t_1} N(t) \cdot P_0(t) \cdot ID_0(t) \cdot P_1(t)\,dt} =$$

$$= \frac{\int_{t_0}^{t_1} N(t) \cdot P_1(t) \cdot \overline{IDR} \cdot ID_0(t) \cdot P_0(t)\,dt}{\int_{t_0}^{t_1} N(t) \cdot P_0(t) \cdot ID_0(t) \cdot P_1(t)\,dt} = \overline{IDR}$$

Incidence is a f(x) of time and exposure is a f(x) of time ⇒ time is a confounder. We match on time.
we have 1 less assumption than before and we get that OR ≃ $\overline{IDR}$ still without having to invoke the rare disease assumption.

If you have a rare disease you get a better estimate if you do incidence density sampling rather than cumulative incidence sampling. You have the most flexibility in terms of interpreting the OR, if you match cases on time.

You're better off designing a study that allows you to interpret the OR on a rate or risk scale.

For each individual i, there is a certain incidence probability

r_{1i} = probability disease will occur when an individual is exposed

r_{0i} = " "unexposed

$1 - r_{1i}$ = probability that dis. does not occur when ind is $E^+ = s_{1i}$

$1 - r_{0i} = s_{0i}$

Define risk diseased odds : $\omega_{1i} = r_{1i}/s_{1i}$ } on the individual
$\omega_{0i} = r_{0i}/s_{0i}$ } " individual level risk odds

Define risk difference : $r_{1i} - r_{0i}$ on an individual's

Define risk ratio : r_{1i}/r_{0i} " "

Define risk (disease) odds difference : $\omega_{1i} - \omega_{0i}$ on ind. i

Define incidence/risk/disease odds ratio : $\omega_{1i} / \omega_{0i}$ on ind. i.

The expected number of individuals over the risk period :

	Exposed	Not exposed	
Disease occurs	$A = \sum_1 r_{1i}$	$B = \sum_0 r_{0i}$	
Disease not occur	$C = \sum_1 s_{1i}$	$D = \sum_0 s_{0i}$	
	N_1	N_0	

s = survivors
r = prob. disease will occur

Incidence proportion ^{1}exposure : $\frac{A}{N_1} = \frac{\sum_1 r_{1i}}{N_1}$

Incidence proportion ^{No}exposure : $\frac{B}{N_0} = \frac{\sum_0 r_{0i}}{N_0}$

} measures of average risk

Disease odds ^{1} exposure : $\frac{A}{C} = \frac{\sum_1 r_{1i} / N_1}{\sum_1 s_{1i} / N_1}$

Disease odds ^{No} exposure : $\_ = \frac{\sum_0 r_{0i} / N_0}{\sum_0 s_{0i} / N_0}$

} Ratio of average risk to average survival probability.

Individual level risk odds.

$\omega_{1i} = r_{1i} / s_{1i}$
$\omega_{0i} = r_{0i} / s_{0i}$
} on ith individual

$\Rightarrow \frac{\sum r_{1i}}{\sum s_{1i}} \neq \frac{\sum r_{1i}}{\sum s_{1i}} \Leftarrow$ Average disease odds

$\sum \omega_{1i} / N_1$

$\sum \omega_{0i} / N_0$

Ratio of average risk to average survival probability. $\neq$ Average disease odds

It's NOT equal

odds are complicated and hard to explain!!

Assume NO CONFOUNDING (i.e. group comparability)

then the average risk in the exposed and unexposed subcohorts would be the same if the exposed group was unexposed.

This is a critical counterfactual assumption.

$$\frac{\sum_0 r_{0i}}{N_0} = \frac{\sum_1 r_{0i}}{N_1}$$

The incidence proportion (risk) difference

$$\frac{A}{N_1} - \frac{B}{N_0} = \frac{\sum_1 r_{1i}}{N_1} - \frac{\sum_0 r_{0i}}{N_0} = \frac{\sum_1 r_{1i}}{N_1} - \frac{\sum_1 r_{0i}}{N_1} \Rightarrow$$

This only holds if NO CONFOUNDING.

Hypothetical unexposed r_{0i} and exposed r_{1i} in the same cohort N_1

Absolute change in the average risk of the exposed subcohort produced by exposure

$$= \frac{\sum_1 (r_{1i} - r_{0i})}{N_1} \Rightarrow$$

Average absolute change in risk produced by exposure in exposed individuals.

This is a statement with CAUSAL MEANING!!

That's why we need to avoid bias, in order to have answers that have causal significance.

Incidence proportion ratio:

$$\frac{\frac{A}{N_1}}{\frac{B}{N_0}} = \frac{\sum_1 r_{1i}/N_1}{\sum_0 r_{0i}/N_0} = \frac{\sum_1 r_{1i}/N_1}{\sum_1 r_{0i}/N_1} =$$

proportionate change in the average risk ($\sum_1 r_{1i}/N_1$, $\sum_0 r_{0i}/N_1$) of the exposed subcohort (N_1) produced by exposure (r_0 vs r_1)

$\neq$ average proportionate change in risk produced by exposure

$$\frac{\sum_1 (r_{1i}/r_{0i})}{N_1}$$

The causal interpretation looks at what happen to the exposed cohort if they had not been exposed. ⇒ this is the point of the unexposed cohort. It's a proxy for the exposed cohort if no exposure had occured.

The disease odds ratio

$$\frac{A/C}{B/D} = \frac{\sum_1 r_{1i} / \sum_1 s_{1i}}{\sum_0 r_{0i} / \sum_0 s_{0i}}$$

NO CONFOUNDING ASSUMPTION ⇓ =

$$\frac{\sum_1 r_{1i} / \sum_1 s_{1i}}{\sum_1 r_{0i} / \sum_1 s_{0i}} = \frac{\sum w_{1i}}{\sum w_{0i}}$$

The proportionate change in the incidence odds (i.e the ratio of the average risk to the average survival probability) in the exposed population produced by exposure (r_0 vs r_1, s_0 vs s_1)

This is the correct interpretation of the OR, but it is hard to comprehend.

$$\frac{\sum_1 w_{1i} / N_1}{\sum_1 w_{0i} / N_1}$$ = proportionate change in the average odds in the exposed population produced by exposure

$$\frac{\sum_1 \left(\frac{w_{1i}}{w_{0i}}\right)}{N_1}$$ = average of the individual odds ratios among the exposed

These are more intuitive but incorrect interpretations of the OR

Thus, the incidence odds ratio lacks any simple interpretation in terms of

average risk or average odds or

an average exposure effect on individual risk or odds.

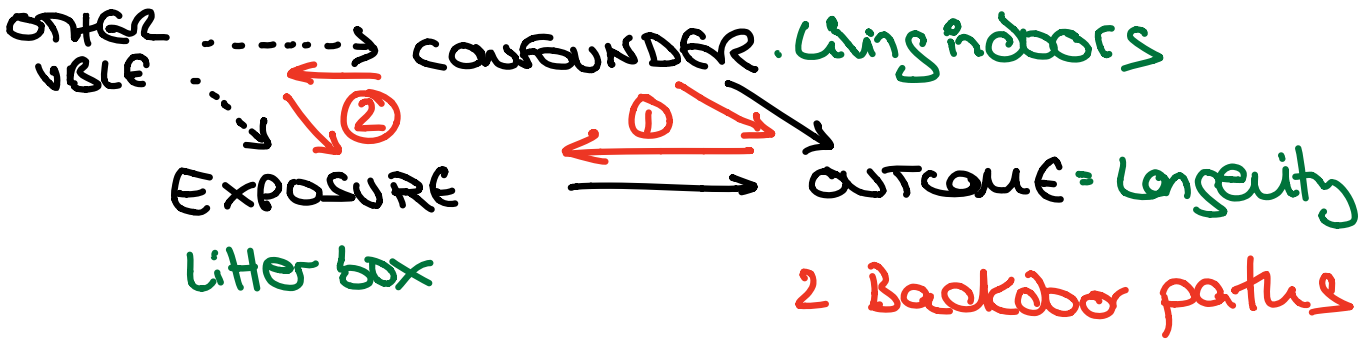
CONFOUNDING. BIAS

Error of "comparison validity" → cohorts not comparable → biased effect measure estimates.

CONFOUNDER: factor responsible for confounding.

CONFOUNDING present if Exp ass w/ Dis.
aka if all Exp effects were blocked/erased.

Ex: cats w/ litterbox live longer. Actually due to living indoors?



- IN COHORT STUDIES:

① C ASS. W/O
OR $_{COL\bar{E}} \neq 1$

Living indoor
Longevity
confounds ass. w/ outcome
regardless of exposure { presence / absence
litterbox use

Confounder predictive of outcome in unexposed ind.

② C ASS. W/ E
OR $_{CE} \neq 1$

Living indoors
Litter box
confounds ass w/ exposure
(d-connected)

This can only be tested w/ data if the assumption of RANDOM SAMPLING = COMPARABILITY of COHORTS = True

- IN CASE-CONTROL STUDIES

① CASE w/O
OR $_{COL\bar{E}} \neq 1$

Living indoor
confounds ass. w/ outcome
regardless of exposure (presence / absence)
Longevity
Litterbox use

Confounder predictive of outcome in unexposed ind.

② CASE w/E
OR $_{CE} \neq 1$

Living indoors
Litter box
confounds ass w/ exposure
(d-connected)

⇓

Only applies in c-c studies if we know about this association in the source population; we cannot look at this ass. in the data bc we don't have access of pop. cases and controls but selectively sample non-cases

EXPOSURE CONFOUNDER

CASES FE die young →
CONTROLS FE die old →

Litter box ↔ Living indoors
we expect association

↑ conditional on being selected in study

To examine the data for CONFOUNDER-EXP association
⇓

② ALT C ASS W/ E

OR $CE|\bar{D} \neq 1$

Litterbox

Living indoor

confounds ass. w/ exposure regardless of outcome [presence / absence]

longevity

The C-E association (statistical dependence between the putative confounder and exposure) is not always identifiable from the data, because it requires comparing the distribution of the "confounder" in unexposed individuals with the distribution it would have had in the exposed individuals if exposure had been absent or blocked.

Confounder predictive of exposure in unexposed ind.

③ CONFOUNDER NOT ON CAUSAL PATH

EXPOSURE ⟶ OUTCOME.

can't verify from the data

STATISTICAL TESTING OF CONFOUNDING ↘ should be avoided

It reflects:

- magnitude of association
- n° of observations ⟶ sufficient n° = everything is significant ↘

The magnitude of confounding should have nothing to do with the size of the study.

Instead, look at changes in the estimates [with / without] confounder var → if the estimate changes > cut-off (5-10%) → significant association

COMPETING DEFINITIONS OF CONFOUNDING

- **COLLAPSIBILITY:**

A measure of effect is confounded if the strata-spp measures ≠ crude collapsed measures.

Can give ≠ results depending on effect measure of choice in the same pop!

(RR might show collapsibility but OR might not)

- **COMPARABILITY:**

confounding = invalid comparison of groups of ind.

No confounding: groups comparable in absence of Exp. => same incidence

Equivalent results between these two definitions may be obtained when using measures that compare averages of individual responses, such as incidence proportion ratios and differences, and rate ratios.

Odds ratios do not compare average responses in populations!

- This implies that collapsibility-based definitions of confounding will fail when using odds ratios unless the study outcome is rare; i.e., the odds ratio is interpretable as an incidence proportion or rate ratio.
- This criterion must hold in all strata of controlled variables.

crude SMR = O/E = 656/258 = 2.54 = Adjusted SMR, = O/E = 656/258 = 2.54

O: observed cases: 656

E: expected cases: $\frac{258}{4000} \times 4000 \rightarrow$ exposed pop $\rightarrow$ unexposed cases $= 258$

O: 124 + 532 = 656

E: $\frac{43}{2000} \times 2000 + \frac{215}{2000} \times 2000 = 258$

⇓ No confounding by gender

Table 1—Example showing the association of corticosteroid (Co) treatment and canine urinary tract infection (UTI) by gender

	Male		Female		Crude	
Disease	Co	No Co	Co	No Co	Co	No Co
UTI	124	43	532	215	656	258
No UTI	1,876	1,957	1,468	1,785	3,344	3,742
Total	2,000	2,000	2,000	2,000	4,000	4,000
Incidence	0.062	0.022	0.266	0.108	0.164	0.065
Risk difference	. . .	0.041	— ≠ —	0.159	— ≠ —	0.100
Risk ratio	. . .	2.86	— ≠ —	2.47	— ≠ —	2.54
Odds ratio	. . .	3.01	— = —	3.01	— ≠ —	2.85

Heterogeneity RD and RR } collapsibility = confounding
Non-collapsibility OR } comparability ≠ "

CRUDE SMR: O/E = 1.775 ≠ ADJUSTED SMR = 5120 / 3167.162 = 1.62 ⇒ confounding by gender

O: 5120

E: $\frac{2075}{6000} \times 8850 = 3887.7$

O: 5120 = 4588 + 532

E: $\frac{1859}{4000} \times 6350 + \frac{216}{2000} \times 2000 = 3167.162$

Table 4—Numerical example of odds ratio collapsibility, but risk difference and risk ratio noncollapsibility

	Male		Female		Crude	
Disease	Co	No Co	Co	No Co	Co	No Co
UTI	4,588	1,859	532	216	5,120	2,075
No UTI	1,762	2,141	1,468	1,784	3,230	3,925
Total	6,350	4,000	2,000	2,000	8,350	6,000
Incidence	0.723	0.465	0.266	0.108	0.613	0.346
Risk difference		0.258		0.158		0.267
Risk ratio		1.55		2.46		1.77
Odds ratio		3.0		3.0		3.0

Co = corticosteroid; UTI = urinary tract infection.

Heterogeneity RD and RR } collapsibility = confounding
Collapsibility OR } comparability = "

Table 2—Numerical example demonstrating that control of a factor meeting the three criteria for confounding can produce bias

Type of individual	Male EXP	Male UEXP	Female EXP	Female UEXP	Crude EXP	Crude UEXP
Disease* inevitable	100	67	18	72	118	139
Exposure causal	100	0	18	0	118	0
Immune to disease	100	33	164	328	264	361
Total	300	100	200	400	500	500
Incidence	0.67	0.67	0.18	0.18	0.47	0.28
Incidence ratio	1.00		1.00		1.70	

* = urinary tract infection. EXP = exposed to corticosteroids; UEXP = unexposed to corticosteroids.

NON-COLLAPSIBILITY IR

Incidence proportion if there had been no exposure in the exposed group? $\frac{(0.995 \times 5000 + 0.0002 \times 1000)}{6000} = 0.829$

Confounding?

non-comparability

Incidence prop. in the unexposed group: 0.1660

effect modification depends on the measurement scale (multiplicative vs additive)

TABLE 1

Hypothetical cohort showing inadequacy of odds-ratio collapsibility as a criterion for ignoring a covariate

	Stratum 1		Stratum 2		Combined (crude)	
	Exposed	Unexposed	Exposed	Unexposed	Exposed	Unexposed
Diseased	4,999	995	5	1	5,004	996
Nondiseased	1	5	995	4,999	996	5,004
Total	5,000	1,000	1,000	5,000	6,000	6,000
Incidence proportion	0.9998	0.9950	0.0050	0.0002	0.8340	0.1660
Incidence ratio	1.00		25.0		5.02	
Incidence difference	0.0048		0.0048		0.6680	
Incidence odds ratio	25.1		25.1		25.2	

when you don't have the rarity the OR is not a good approximation of incidence ratio.

Greenland et al., 1988

COLLAPSIBILITY OR HETEROGENEITY IR NON-COLLAPSIBILITY ID

collapsibility OR comparability IP="

incidence proportion if exposure had been absent: $\frac{\frac{995}{1000} \times 1000 + \frac{1}{1000} \times 1000}{2000} = 0.498$

Odds Ratio Noncollapsibility

Incidence proportion in unexposed group = $\frac{996}{2000} = 0.498$

TABLE 2

Hypothetical cohort data, no effect modification; parameter is odds ratio; exposure and strata independent in joint source population

	Stratum 1		Stratum 2		Summary	
	Exposed	Nonexposed	Exposed	Nonexposed	Exposed	Nonexposed
Diseased	999	995	5	1	1,004	996
Nondiseased	1	5	995	999	996	1,004
Total	1,000	1,000	1,000	1,000	2,000	2,000
	OR $\theta_1 = 5.02$		$\theta_2 = 5.02$		$\theta_s = \theta^* = 1.02$	

IPR $\frac{999/1000}{995/1000} = 1.004$ $\neq$ $\frac{5/1000}{1/1000} = 5$ $\neq$ $\frac{1004/2000}{996/2000} = 1.008$

Grayson, 1987

IPD $\frac{999}{1000} - \frac{995}{1000} = 0.004$ = $\frac{5}{1000} - \frac{1}{1000} = 0.004$ = $\frac{1004}{2000} - \frac{996}{2000} = 0.004$

NON-COLLAPSIBILITY OR HETEROGENEITY IPR

HOMOGENEITY IPD NO CONFOUNDING!

Confounding? Effect modification?

TABLE 3

Hypothetical cohort data; effect modification present; exposure and strata independent; parameter is odds ratio

	Stratum 1		Stratum 2		Summary	
	Exposed	Nonexposed	Exposed	Nonexposed	Exposed	Nonexposed
Diseased	999	996	6	1	1,005	997
Nondiseased	1	4	994	999	995	1,003
Total	1,000	1,000	1,000	1,000	2,000	2,000
	$\Theta_1 = 4.01$		$\Theta_2 = 6.03$		$\Theta_s = \Theta^* = 1.02$	

OR: noncollapsibility, heterogeneity (note disease incidence)
RR: heterogeneity
RD: heterogeneity
SMR crude = SMR adjusted = 1.008

STANDARIZATION OF RATES.

- **Incidence** tells us about the occurrence of new disease in a population (remember: numerator, denominator and specification of time are required).
- **Incidence rate** **(sometimes just called incidence density rate, hazard rate, or hazard for short)**: the denominator consists of the sum of different times each individual was at risk: that is, the "person-time" in the denominator. The numerator same as for cumulative incidence for same time period.

Age	Age proportion of both populations	Population 1 mortality rate per 100,000	Population 2 mortality rate per 100,000
Young	33.3%	7.75	26.25
Middle aged	33.3%	18.5	18.5
Elderly	33.3%	26.25	7.75
Overall	**100%**	**17.5**	**17.5**

- The age group-specific mortality rates vary enormously by population.
- But because although age is not associated with the populations (they have the same age distribution/structure), the differences cancel out.
- **In this case, it would make more sense to report age-specific mortality rates for each population than to report overall mortality rates.**

Age	Population 1		Population 2	
	Age proportion	Mortality rate per 100,000	Age proportion	Mortality rate per 100,000
Young	10%	10	70%	15
Middle aged	20%	20	20%	25
Elderly	70%	65	10%	70
Overall	**100%**	**50.5**	**100%**	**22.5**

- Note that in each of the three age groups, the mortality rate in Population 2 is greater than Population 1.
  - Young: 15 vs. 10 deaths per 100,000
  - Middle aged: 25 vs. 20 deaths per 100,000
  - Elderly: 70 vs. 65 deaths per 100,000
- *Nevertheless*, the overall mortality rate in Population 1 (50.5 deaths per 100,000) is greater than in Population 2 (22.5 deaths per 100,000).

AGE is a CONFOUNDER

- Notice that AGE is associated with population: Population 1 is much older (70% elderly) than Population 2 (70% young).
- Notice that AGE is predictive of mortality: the rates in young individuals are much less than older individuals.

Age	Population 1		Population 2	
	Age proportion	Mortality rate per 100,000	Age proportion	Mortality rate per 100,000
Young	10%	10	10%	15
Middle aged	20%	20	20%	25
Elderly	70%	65	70%	70
Overall	**100%**	**50.5**	**100%**	**55.5**

- Now note the mortality ratios comparing Population 2 to Population 1:
  - 15/10 = 1.500
  - 25/20 = 1.250
  - 70/65 = 1.0769
- The crude mortality ratio is 55.5/50.5 = 1.099, correctly reflecting the higher mortality in Population 2 compared to Population 1.

AGE IS A CONFOUNDER

ELIMINATION OF CONFOUNDING.

- ELIMINATE ASSOCIATION BETWEEN (CONFOUNDER / RATE)
  "
  STANDARIZATION

- DIRECT STANDARIZATION :

- The reason for standardization is so one can compare populations using an overall summary (synoptic) rate measurement that are inherently "non-comparable" because of their inherent differences that influence the outcome rates.
- Standardization makes non-comparable populations "comparable".

- The directly standardized rate is motivated by the idea of determining what the crude disease or death rate would have been in a cohort if it had an age distribution like that of a standard population
- That is, we apply age-specific cohort rates to the standard age distribution.
- w_j represents the proportion of individuals in the standard population in the j^{th} age group.

Directly standardized rate $\hat{\lambda} = \sum_{j=1}^{J} w_j^* \hat{\lambda}_j$

d_j : nº death/dis in j^{th} category

j : vble we standardise on (age)

n_j : nº person/time in j^{th} cat.

w_j : nº or proportion of ind. in the standard pop in the j^{th} category

$\hat{\lambda}_j$: j^{th} category-stratum-spp incidence rate in cohort.

- The result of direct standardization is an actual rate.
- The weights are supplied by the standard population.
- The rates are supplied by the populations (or cohorts) that are being compared.

- It is important to recognize that direct standardization involves taking *weighted* averages of age (and/or other stratum)-specific rates.

- If one is interested in the effect of some exposure in a particular population, then the "exposed" population should provide the weighting for standardization.

Example

standard pop.

Age (yrs)	Dairy 1			Dairy 2		
	Population	Number of mastitis cases	Age-specific mastitis rate	Population	Number of mastitis cases	Age-specific mastitis rate
2-3	200	14	0.070	1,000	20	0.020
4-5	1,200	36	0.030	100	14	0.140
6-7	800	12	0.015	400	12	0.030
All ages	2,200	62	0.028 (crude)	1,500	46	0.031 (crude)

Age (yrs)	Dairy 1	Dairy 2	Combined	Percent (w_j)
2-3	200	1,000	1,200	0.324
4-5	1,200	100	1,300	0.351
6-7	800	400	1,200	0.324
	2,200	1,500	3,700	1

- Note that by using the SAME age distribution (weights) to get a weighted average of each population's age-specific mastitis rates, we *remove* the association between age and population, and hence remove confounding by age.
- The populations can now be compared.

Age (yrs)	Dairy 1			Dairy 2		
	Standard weight	Age-specific mastitis rate	$w_j^* \lambda_j$	Standard weight	Age-specific mastitis rate	$w_j^* \lambda_j$
2-3	0.324	0.070	0.0227	0.324	0.020	0.0065
4-5	0.351	0.030	0.0105	0.351	0.140	0.0491
6-7	0.324	0.015	0.0049	0.324	0.030	0.0097
Total			0.0381			0.0653

- Note that the ratio of the two age-adjusted (standardized) rates comparing Dairy 2 to Dairy 1 = 0.065 / 0.038 = 1.71.
- In contrast, the ratio of the crude mortality rates = 1.11.
- The difference in these ratios is a consequence of confounding by age: Dairy 2 has younger cows, and the incidence of mastitis varies by age.

- INDIRECT STANDARDIZATION: when you want to compare the incidence rate in a cohort to the incidence rate in a ★ standard pop, controlling for confounders.

j th STRATUM-SPP RATE ★ :

$\lambda_j^* = d_j^* / n_j^*$

CRUDE MORTALITY RATE ★ :

$\underline{Ed}^* = \underline{D}$ = #death/dis ev time

$E n_j^*$ $E n_j^*$ = ind/time of all ind in jth stratum

EXPECTED Nº DEATHS / DIS: nº of ind we would expect to die in the cohort if it had the same death rates as ★?

$$E^{\star} = \sum n_j \lambda_j^{\star}$$

OBSERVED Nº DEATHS

$$D = \sum d_j$$

$$\frac{\text{observed}}{\text{Expected}} = \frac{D}{E^{\star}} = \frac{\sum d_j}{\sum \frac{n_j}{\lambda_j^{\star}}} = \frac{\sum d_j / \sum n_j}{\frac{n_j}{\lambda_j^{\star}} / \sum n_j} = \frac{\text{crude rate in cohort}}{\text{weighted average of category-spp rates}}$$

weights are applied by the cohort $\sum n_j$ → you can't compare across cohorts

rate applied by the standard pop ★

$\lambda_j^{\star}$

> "Of course, there is no guarantee that this process will identify those diseases or causes of death that are most closely associated with the exposures. Cause-of-death specific rates for unexposed cohort members may be less than those in the general population, whereas rates for the exposed members are higher, and the two effects may cancel each other out when averaged over the entire cohort."

Example: healthy worker effect

EXAMPLE : compare crude mortality rates and adjust for age.

	Davis			Walnut Creek		
Age	Pop□	Deaths	Rate	Pop□	Deaths	Rate
0-9	6,000	54	0.009	9,000	45	0.005
10-19	10,500	33	0.003	16,500	36	0.002
20-29	13,500	60	0.004	3,000		
30-39	18,000	36	0.002	9,000		
40-49	4,500	36	0.008	7,500	54	0.007
50-59	3,000	75	0.025	4,500	42	0.009
60+	4,500	126	0.028	7,500	45	0.006
All ages	60,000	420	0.007	57,000	315	0.0055

Walnut creek = Exposed

Davis: standard pop.

CRUDE MORTALITY RATE RATIO : $\frac{0.0055}{0.007} = 0.79$

Observed = 315

Expected = $\sum n_j \lambda_j^*$

SMR = $\frac{315}{543} = 0.58 \neq 0.79$

	Davis	Walnut Creek	Age-specific expected values
Age	Rate (λ_j^*)	Pop□ (n_j)	$\lambda_j^* n_j$
0-9	0.009	9,000	81
10-19	0.003	16,500	49.5
20-29	0.004	3,000	12
30-39	0.002	9,000	18
40-49	0.008	7,500	60
50-59	0.025	4,500	112.5
60+	0.028	7,500	210
All ages	0.007	57,000	543

Confounding by age.

	Zimbabwe			United States		
Age (yrs)	Deaths	Popn	Rate	Deaths	Popn	Rate
0-4	?	1,899,204	?	44,000	19,204,000	0.00229
5-24	?	5,537,992	?	45,000	72,244,000	0.00062
25-44	?	2,386,079	?	147,700	82,197,000	0.00180
45-64	?	974,235	?	368,800	46,751,000	0.00789
65-74	?	216,387	?	478,600	18,280,000	0.02618
75+	?	136,109	?	1,084,900	13,484,000	0.08046
Total	98,808	11,150,006	0.00886	2,169,000	252,160,000	0.00860

Not directly comparable ⇒ ≠ age dist. can't do direct standardization bc we don't have the age-sp rates in Z.

standard pop.

Age (yrs)	US rate (λ_j^*)	Zimbabwe popn (n_j)	Expected = $\sum n_j \lambda_j^*$
0-4	0.00229	1,899,204	4,349
5-24	0.00062	5,537,992	3,434
25-44	0.00180	2,386,079	4,295
45-64	0.00789	974,235	7,687
65-74	0.02618	216,387	5,665
75+	0.08046	136,109	10,951
Total			36,381

$$\frac{\text{Observed}}{\text{expected}} = \frac{98,808}{36,381} = 2.72$$

The overall mortality rate in Zimbabwe would be 2.72 times greater than the mortality rate in the US if Zimbabwe had the same age-specific mortality rates that the US does.

The number of deaths observed in Zimbabwe is 2.72 times greater than the number of deaths that would be expected in Zimbabwe if Zimbabwe had the same age-specific mortality rates that the US does.

The effects of exposure are identical in both pop
The difference in the overall SMR is that their age distributions/weights are ≠

Exercise 3 (Rothman)

Age	General pop^n	Exposed Cohort 1		Exposed Cohort 2
Young				
cases	50	50		5
person-time	100,000	10,000		1,000
incidence	0.0005/yr	0.005/yr	=	0.005/yr
SMR		10	=	10
rate ratio			1	
Old				
cases	400	4		40
person-time	200,000	1,000		10,000
incidence	0.002/yr	0.004/yr	=	0.004/yr
SMR		2	=	2
rate ratio			1	
Overall SMR	$\frac{50+4}{(0.005 \cdot 10{,}000)+(0.002 \cdot 1000)}$ =	7.7	≠ (why!)	2.2
	$\frac{50+4}{(0.005 \cdot 1000)+(0.002 \cdot 10{,}000)}$ =			