Applications of GWAS:
Mendelian Randomization

Christian Werner

(Quantitative geneticist and biostatistician) **EiB, CIMMYT**, Texcoco (Mexico)

Filippo Biscarini

HerrFalloppio

(Biostatistician, bioinformatician and quantitative geneticist) **CNR-IBBA**, Milan (Italy)

Oscar González-Recio

OscarGenomics

(Computational biologist and quantitative geneticist) **INIA-UPM**, Madrid (Spain)

The curse of correlations (and associations)

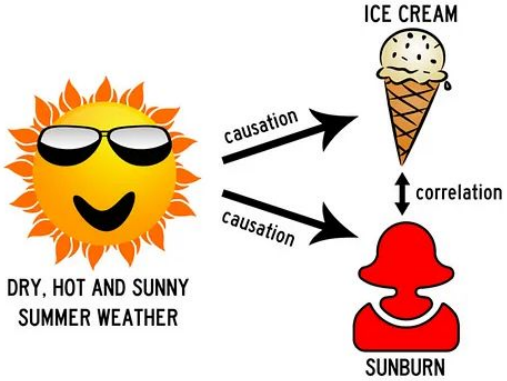
Correlation is not causation

<u>Reverse causation</u>

- n. of firemen←→size of the fire (the more the firemen, the bigger the fire?)
- smoking ← →depression (smoking causes depression?)

<u>Missing variable</u>

- ice cream consumption ← →n. of sunburn cases
- buying lighters ← →lung cancer

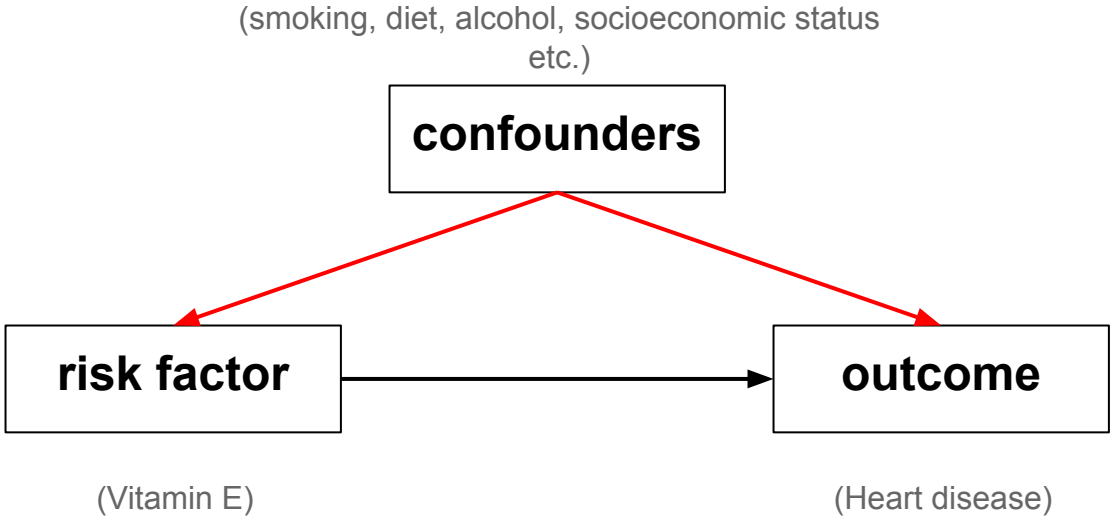


[https://towardsdatascience.com/]

Causation or not causation, that is the question

- post hoc, ergo propter hoc
- strength of correlation
- consistency (different datasets, studies etc.)
- dose-response relationship
- analysis “ceteris paribus” (controlling for confounders, stratifiers etc.)
- **experiments**

Causation from observational studies



- Does the use of vitamin E supplementation reduce the risk of heart disease?
- Is there a causal relationship between dietary assumption of vitamin E and the risk of heart disease?

[Eidelman et al 2004]

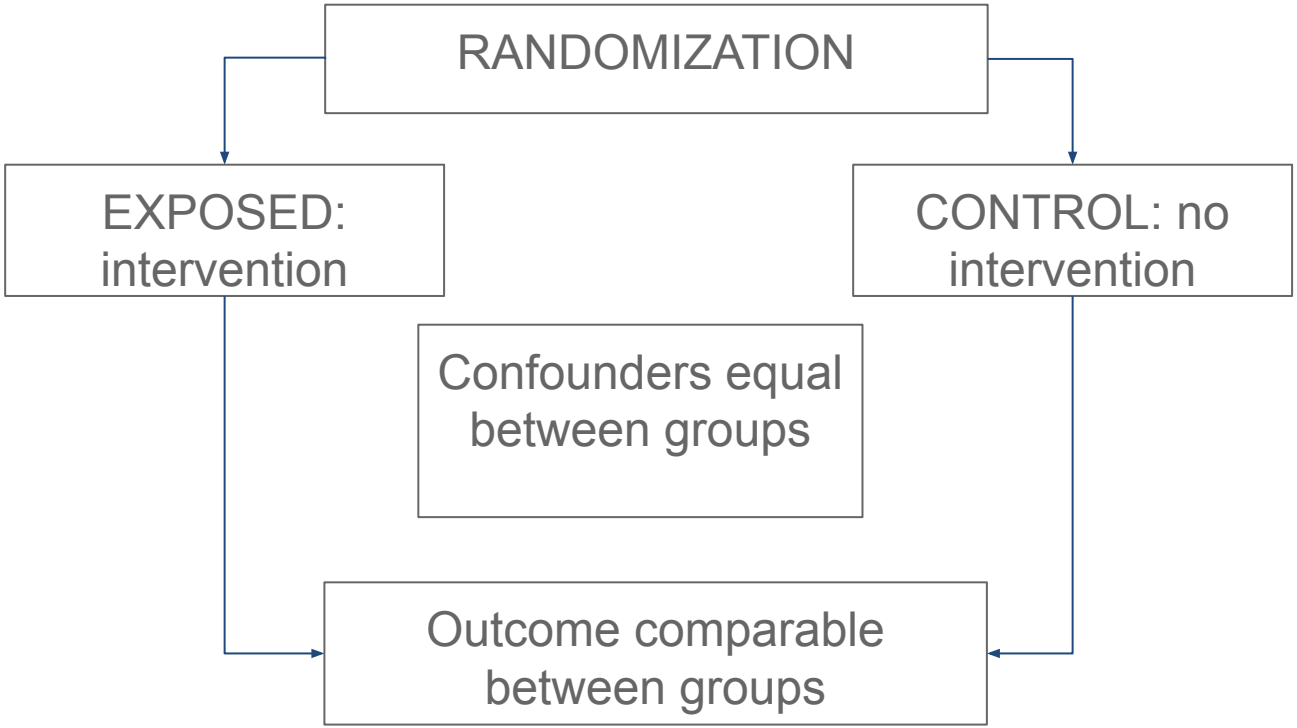
Causation from observational studies

- Confounding: missing variable
- Reverse causation: B causes A, not the other way around as designed/hypothesized
- Bias: the observed sample is not representative of the target population (e.g. different sex ratio, age distribution etc.) [different from random error -variance- which can be reduced by increasing the sample size]

SNP from GWAS: causal relationship?

- probably not causal mutation, but LD
- post hoc, ergo propter hoc: gene → phenotype
- strength of correlation (p-value, multiple testing adjustments etc.)
- consistency (different datasets, studies etc.)
- dose-response relationship (allele coding)
- analysis “ceteris paribus” (controlling for confounders, stratifiers etc.)
- experiments (follow-up studies, post-GWAS, validation, etc.)

Randomized trial



Randomized trial

Randomized trials are the **gold standard for causal inference**

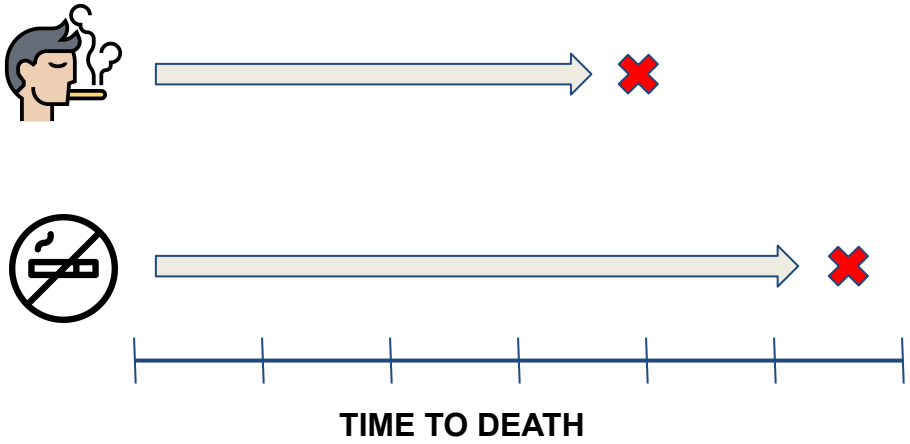
However:

- sometimes not ethical or practical (e.g. toxic exposures)
- expensive and time-consuming
- long follow-up times

Enter Mendelian Randomization

- Genes are inherited at the time of conception
- They follow Mendel rules of transmission: **random segregation** and **independent assortment** (e.g. 50% probability that one of two alleles is passed on from parents to their offspring)
- This is not affected by the lifestyle choices of persons, not related to confounding factors

Enter Mendelian Randomization

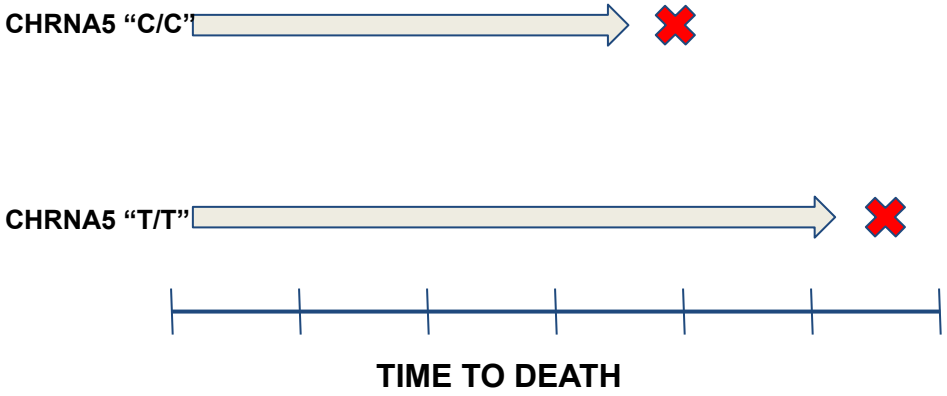


Observation: smokers die sooner than non-smokers

Confounders: smokers on average drink more alcohol, less healthy diets

Reverse causation: when smokers get ill they quit smoking: no smoking →higher risk of death (apparently)

Enter Mendelian Randomization



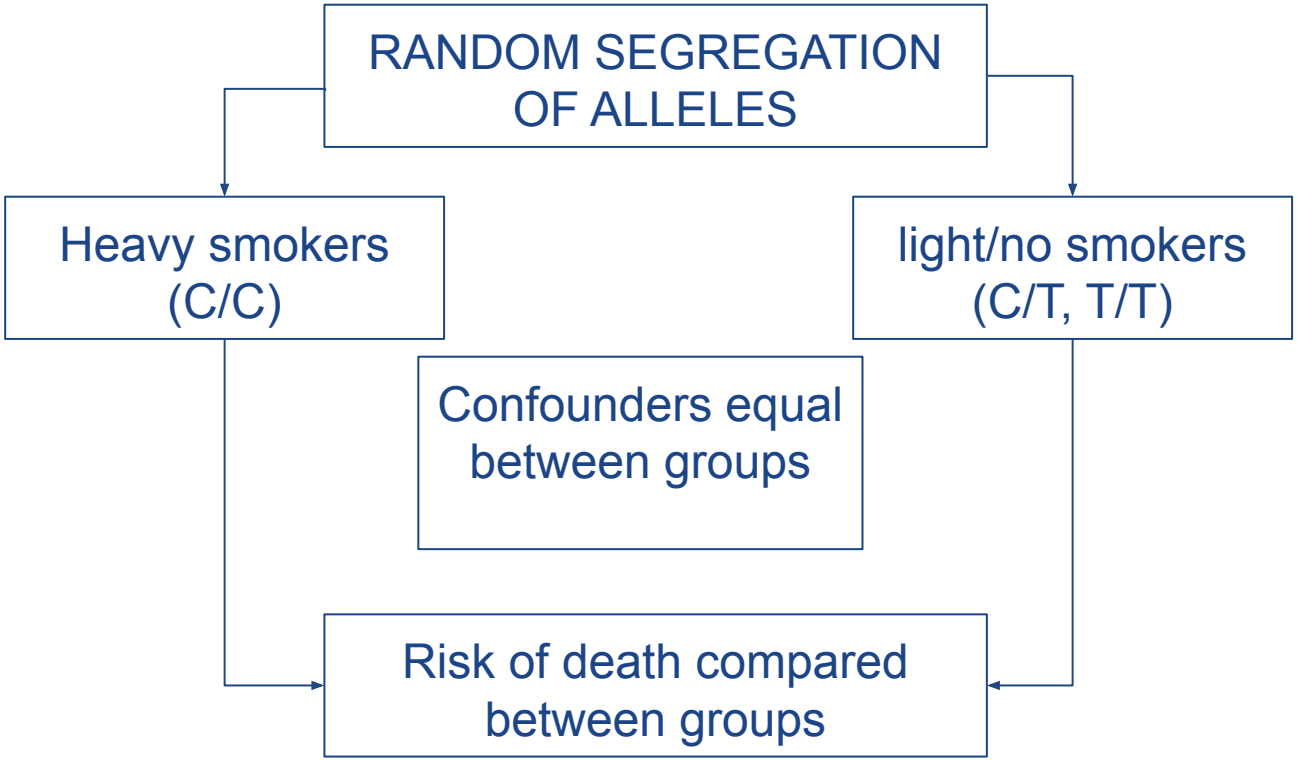
CHRNA5: acetylcholine receptor subunit alpha 5

→associated with smoking behavior:

- allele “C” →more smoking (recessive/codominant)
- allele “T” → less smoking

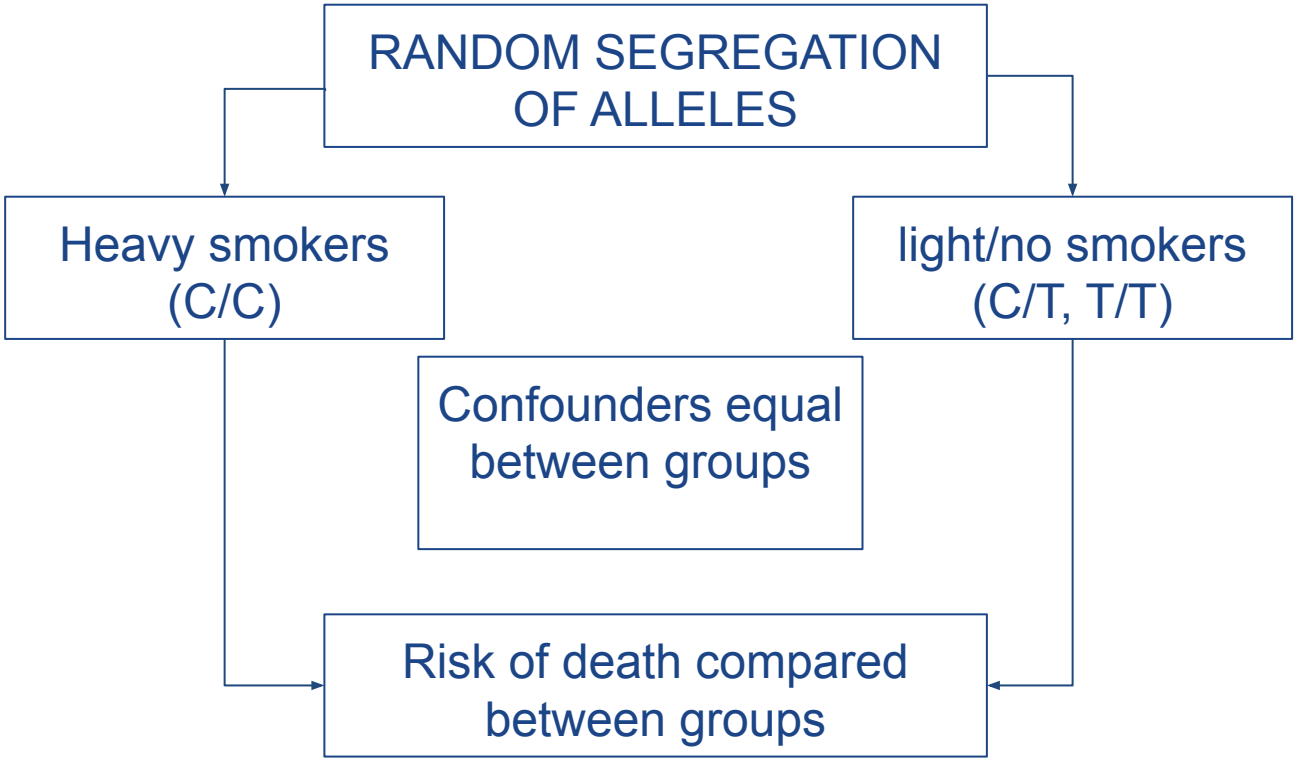
Persons with the allele associated with more smoking die younger (on average)

Enter Mendelian Randomization



- Two groups based on alleles → random segregation
- “Genetic” smokers vs “genetic” non-smokers
- These two groups do not differ for confounders (diet, alcohol, wealth etc.)

Enter Mendelian Randomization



Maybe CHRNA5 affects life expectancy in other ways?

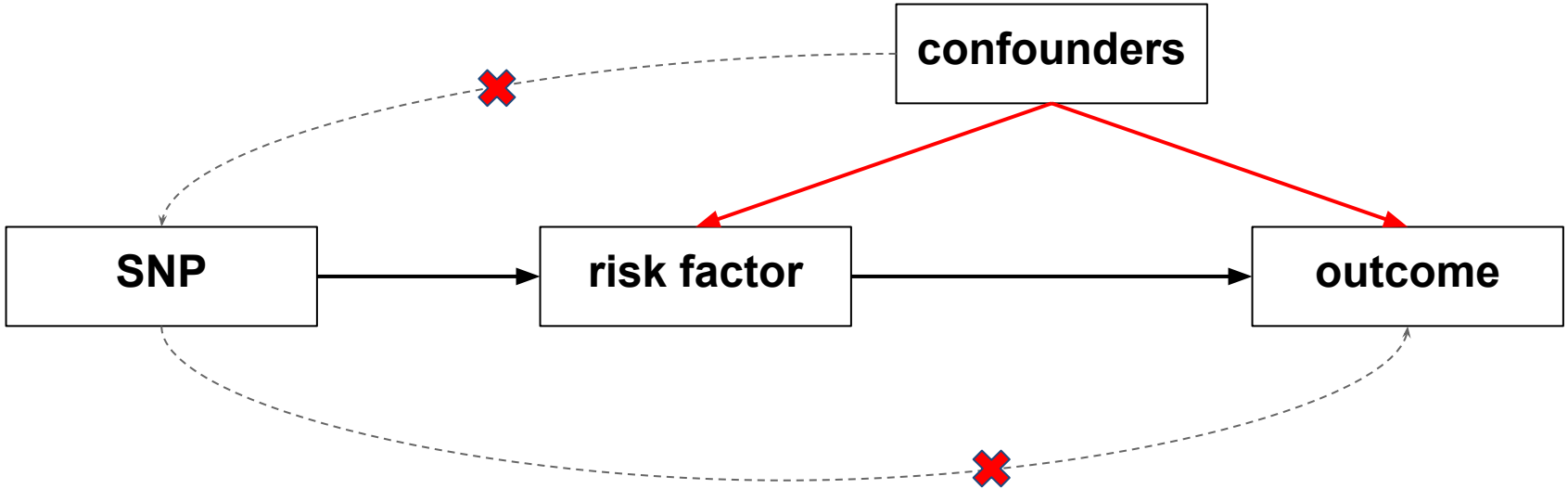
If we look only at non smokers, C/C vs C/T + T/T, we observe no effect on the risk of death

Smoking is the driver for the difference in risk of dying between the two groups

GWAS and Mendelian Randomization

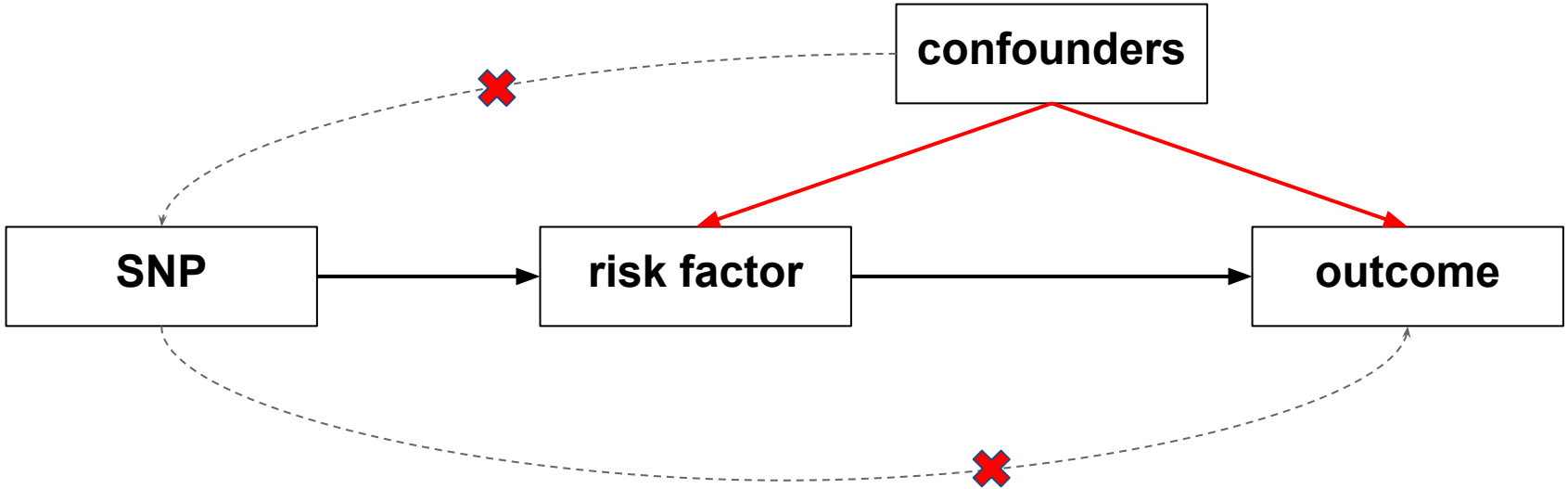
- GWAS studies produce lots of SNPs associated with phenotypes → IVs!
- This is an important application of GWAS experiments
- Important: the associations must be **true** (not spurious) and **strong**

Mendelian Randomization



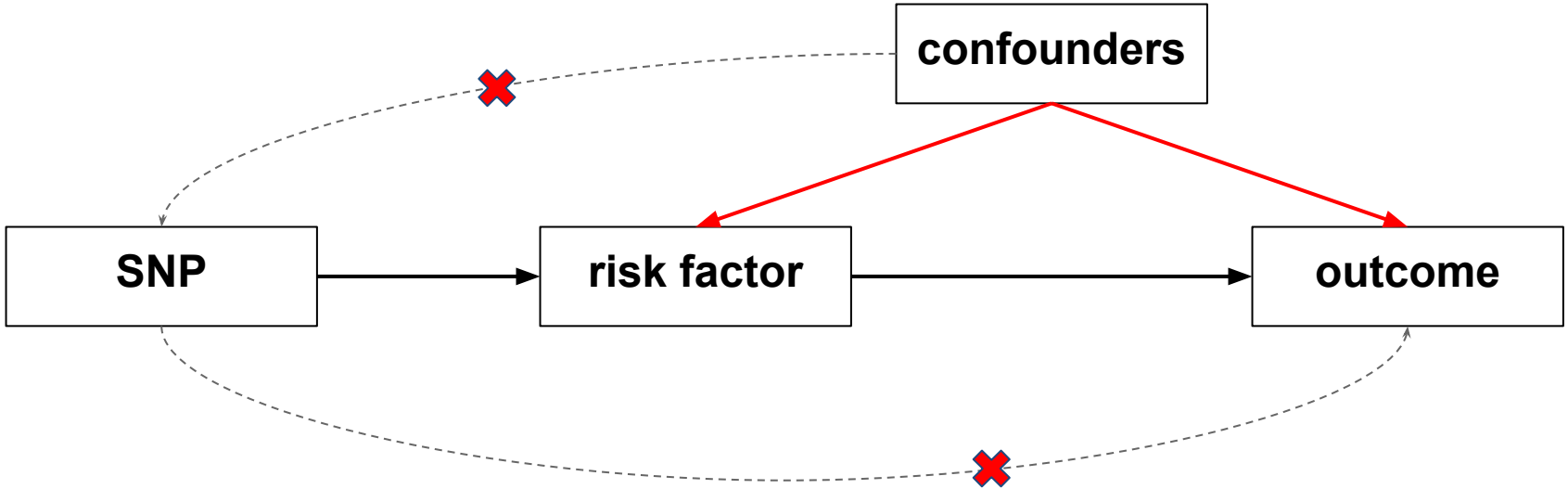
1. The SNP is associated with the risk factor

Mendelian Randomization



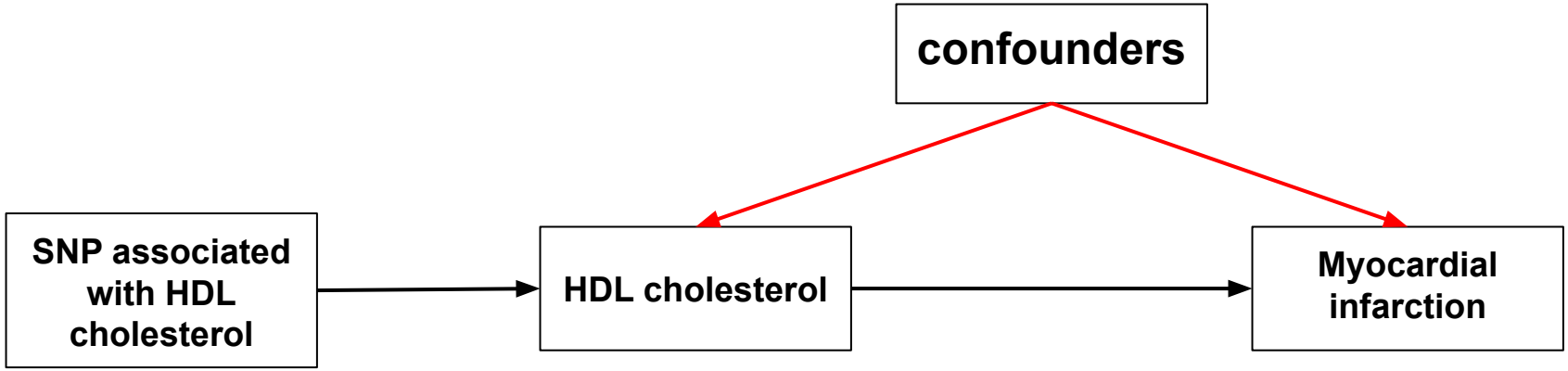
2. The SNP is NOT associated with confounders

Mendelian Randomization



3. The SNP is associated with the outcome ONLY through exposure

Mendelian Randomization: another example



raising HDL cholesterol does not always lower risk of myocardial infarction: interventions (lifestyle or pharmacological) that raise plasma HDL cholesterol cannot be assumed to lead to a corresponding benefit with respect to risk of myocardial infarction (Voight et al. 2012)

Mendelian Randomization: concluding remarks

- **robust to confounding** due to Mendel's laws:
  - Law of segregation: inheritance of an allele is random and independent of environment etc
  - Law of independent assortment: genes for different traits segregate independently (assuming not in LD)
- the **direction of causality is known** – always from SNP to trait
- genetic variants (SNPs) are potentially very **good instrumental variables**
- in most cases, not one single SNP is used, but multiple SNPs assembled into a combined score/IV

Mendelian Randomization: concluding remarks

- In a different flavor of MR, SNPs (IVs) may also be associated with the outcome to study variation in risk factors (still to study the causal relationship between the two)
- Limitations:
  - population stratification (bias, generalization)
  - pleiotropy
  - power (SNP may explain only a small fraction of the phenotypic variance)

NEXT COURSE

Deep learning for biologists