II INTERNATIONAL SUMMER SCHOOL

Rare disease and orphan drug registries

Day 3
17.09.2014

Epidemiologic analyses, confounders, sample stratification

Michele SANTORO
National Council of Research
Pisa, Italy

Organised by Istituto Superiore di Sanità
Rome (Italy), September 15-19, 2014

EPIDEMIOLOGY

"the study of the distribution and determinants of disease frequency"

(K *Rothman*)

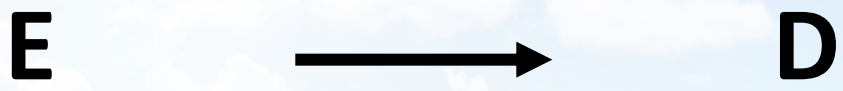
"Epidemiology is the study of the distribution and determinants of health-related states or events (including disease), and the application of this study to the control of diseases and other health problems"

(WHO)

EPIDEMIOLOGY

- the study of the distribution of disease frequency
- the study of the determinants of disease frequency

To investigate and measure the association between an exposure factor **E** and a disease outcome **D**



Study of a cause-effect relationship

EXPOSURE

In epidemiology denotes any of a subject's attributes or any agent that may be relevant to health

OUTCOME

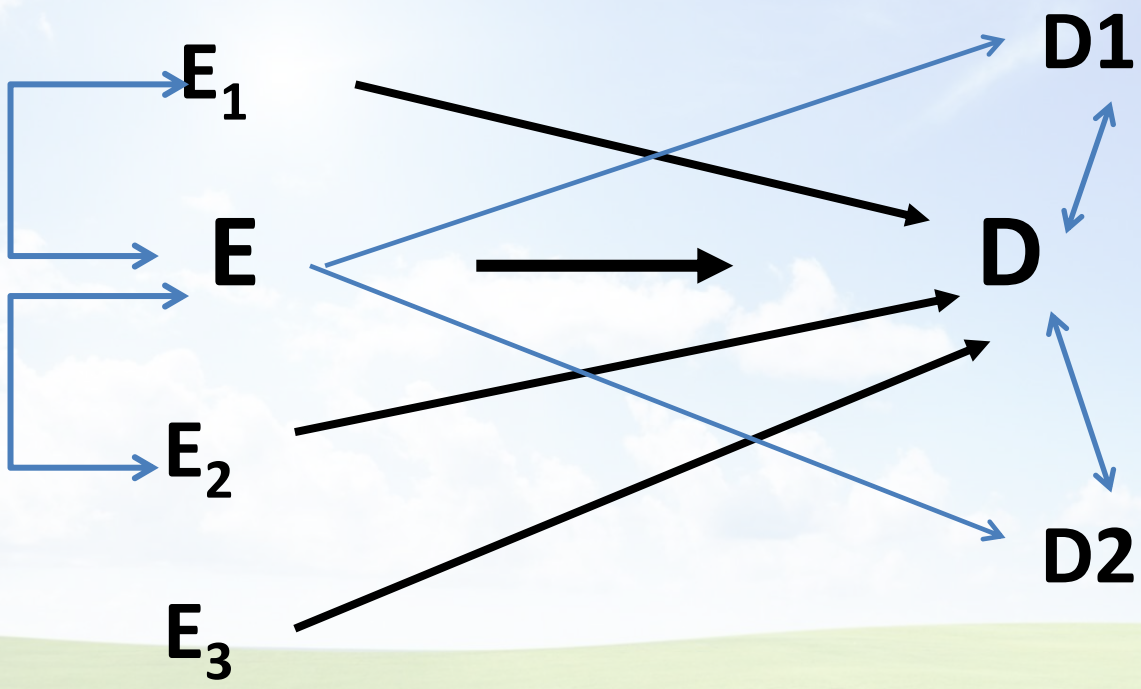
Any biological, health or health related consequent of exposure(s)

EXPOSURE (cause)

- Occupation
- Environmental
- SES
- Smoking
- Treatment
- Drug
- ...

OUTCOME (effect)

- Incidence
- Prevalence
- Mortality
- Survival
- ...



Probabilistic context

MEASURES OF OCCURRENCE

15 cases of heart birth defects have been occurred

This is a partial information

To define a measure of occurrence:

1. Number of events (change or condition of disease) observed
2. Number and characteristics of persons in the population under study
3. The time period during which the events are observed

Risk

$$\text{Risk} = \frac{\text{n.events observed during a time period t}}{\text{population observed in time period t}}$$

$$R = \frac{E}{N}$$

E = number of events

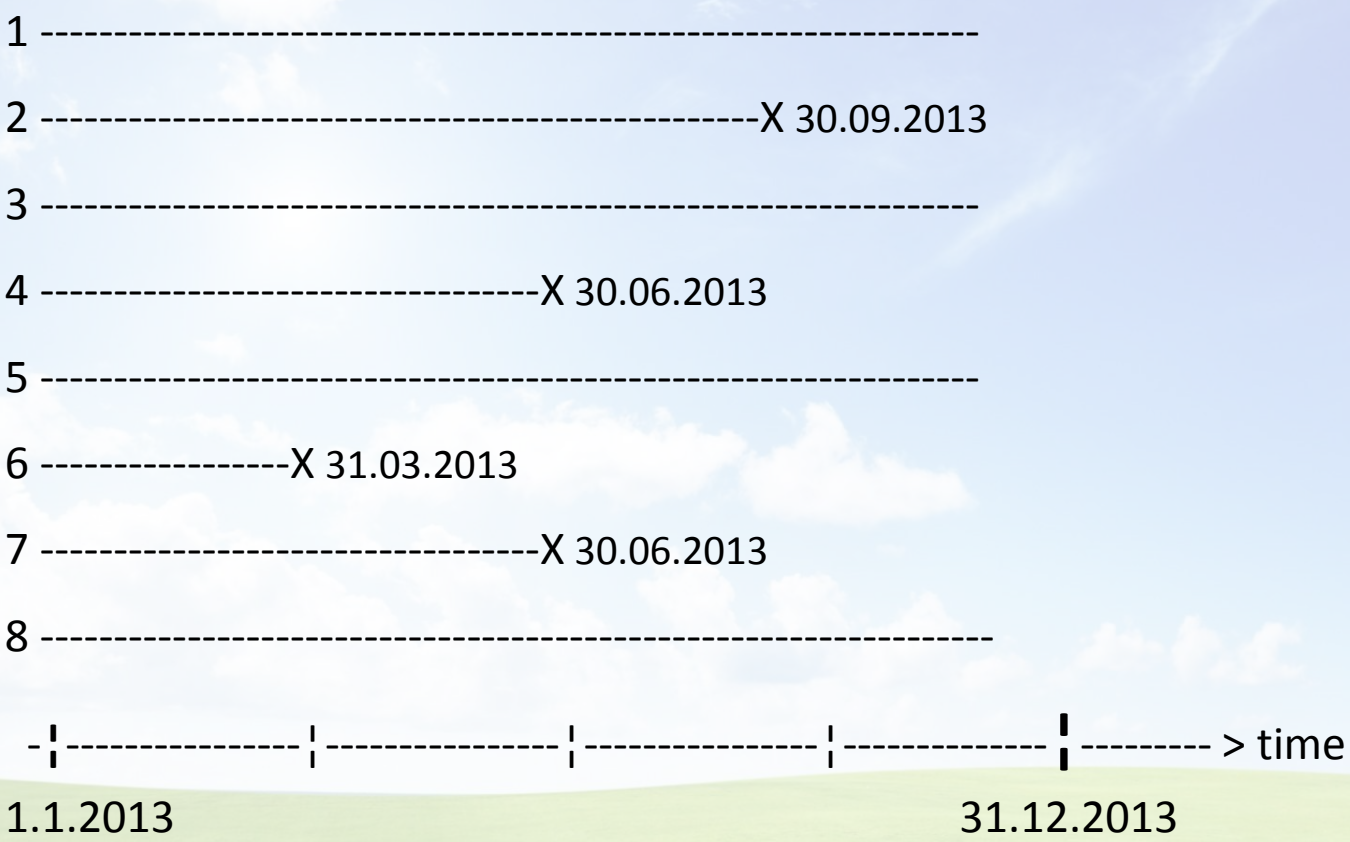
N = number of subjects followed for t

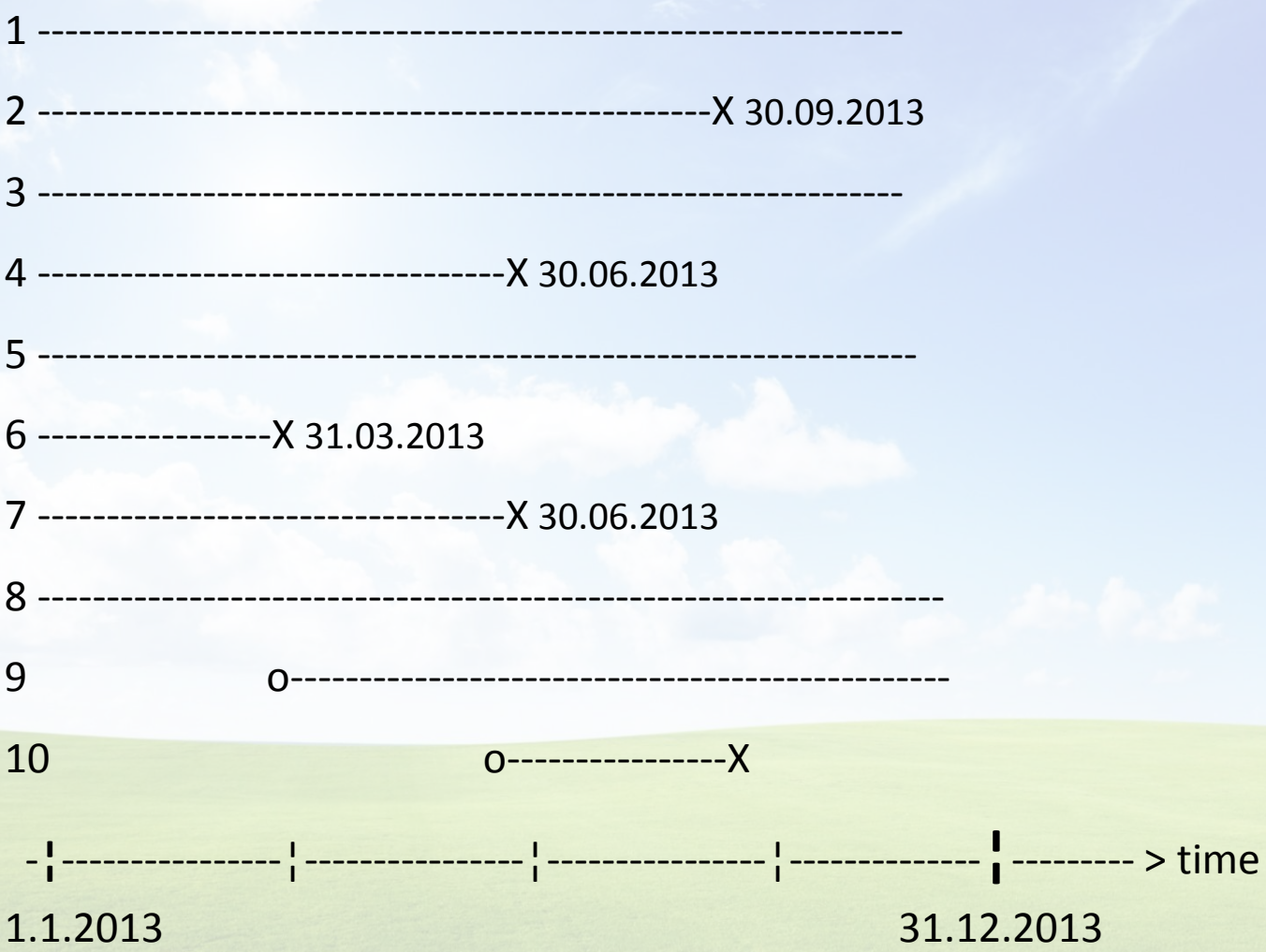
Example:

15 cases observed in 2 years in a population of 100 persons

$$R = \frac{E}{N} = \frac{15}{100} = 0.15 = 15\%$$

Risk *is the probability that an event will occur, e.g. that an individual will become ill or die, within a stated period of time*


1 ----------
2 ----------X 30.09.2013
3 ----------
4 ----------X 30.06.2013
5 ----------
6 ----------X 31.03.2013
7 ----------X 30.06.2013
8 ----------
-¦----------¦----------¦----------¦----------¦--------- > time
1.1.2013
31.12.2013


1 ----------
2 ----------X 30.09.2013
3 ----------
4 ----------X 30.06.2013
5 ----------
6 ----------X 31.03.2013
7 ----------X 30.06.2013
8 ----------
9 o----------
10 o----------X
---> time
1.1.2013
31.12.2013

Rate

$$\text{Rate} = \frac{\text{n.events observed in time period t}}{\text{n. person-years in time period t}}$$

Rate

$$\text{Rate} = \frac{\text{n.events observed in time period t}}{\text{n. person-years in time period t}}$$

$$R = \frac{E}{\Sigma t_i} \times 10^K$$

E = number of events

t_i = time experienced for subject i

10^K = multiplicative constant

Person years= 1 + 0.75 + 1 + 0.5 + 1 + 0.25 + 0.5 + 1 = 6

Number of events E is 2:

$$\text{Risk} = \frac{E}{N} = \frac{2}{8} = 0.25 = 25\%$$

$$\text{Rate} = \frac{E}{\Sigma ti} \times 10^{K} = \frac{2}{6} \times 1{,}000 = 333.3$$

*The **Rate** does not express the probability of disease but the average of the observed cases in a defined time (e.g. year) in a defined population*

Examples of rates

$$\text{Incidence Rate} = \frac{\text{n. of new cases in period of time t}}{\text{n. person-years in period of time t}} \times 10^k$$

~

$$\text{Incidence Rate} = \frac{\text{n. of new cases in period of time t}}{\text{total population}} \times 10^k$$

("approximate")

$$\text{Mortality Rate} = \frac{\text{n. of deaths in time period } t}{\text{population}} \times 10^k$$

$$\text{Letality Rate} = \frac{\text{n. of deaths in time period } t}{\text{n. subjects affected by the disease D}} \times 10^k$$

$$\text{Prevalence Rate} = \frac{\text{n. of subjects with a disease in a point of time}}{\text{population in the same point of time}} \times 10^k$$

Odds

$$\text{Odds} = \frac{\text{n.events observed in period of time t}}{\text{n. of not events observed in period of time t}}$$

$$Odds = \frac{E}{N - E}$$

E = number of events

N = number of subjects observed

Example:

15 cases observed in 2 years in a population of 100 persons

$$\text{Odds} = \frac{E}{N-E} = \frac{15}{85} = 0.18$$

$$\text{Risk} = \frac{E}{N} = \frac{15}{100} = 0.15 = 15\%$$

Risk and Odds

Risk	Odds
1/10 = 0.1	1/9
0.2	2/8
0.3	3/7
0.4	4/6
0.5	5/5 = 1

Risk= Odds / (1+Odds)

MEASURES OF ASSOCIATIONS

Epidemiology is the study of the determinants of disease frequency

How?

- Studying a population exposed (to a factor of interest) and one not exposed and investigate on the health differences

- Studying a population healthy (not affected by the disease of interest) and a population diseased and investigate on the exposure differences

We are interested in comparing 2 groups

compairing the risk or the occurrence (rate) of the disease in 2 groups

The measure of association can be interpreted as a measure of the strength of association between exposure and disease.

	exposed	not exposed	
cases	a	b	*a+b*
not cases	c	d	*c+d*
	a+c	*b+d*	*a+b+c+d = N*

a = cases exposed

b = cases not exposed

c = not cases exposed

d = not cases not exposed

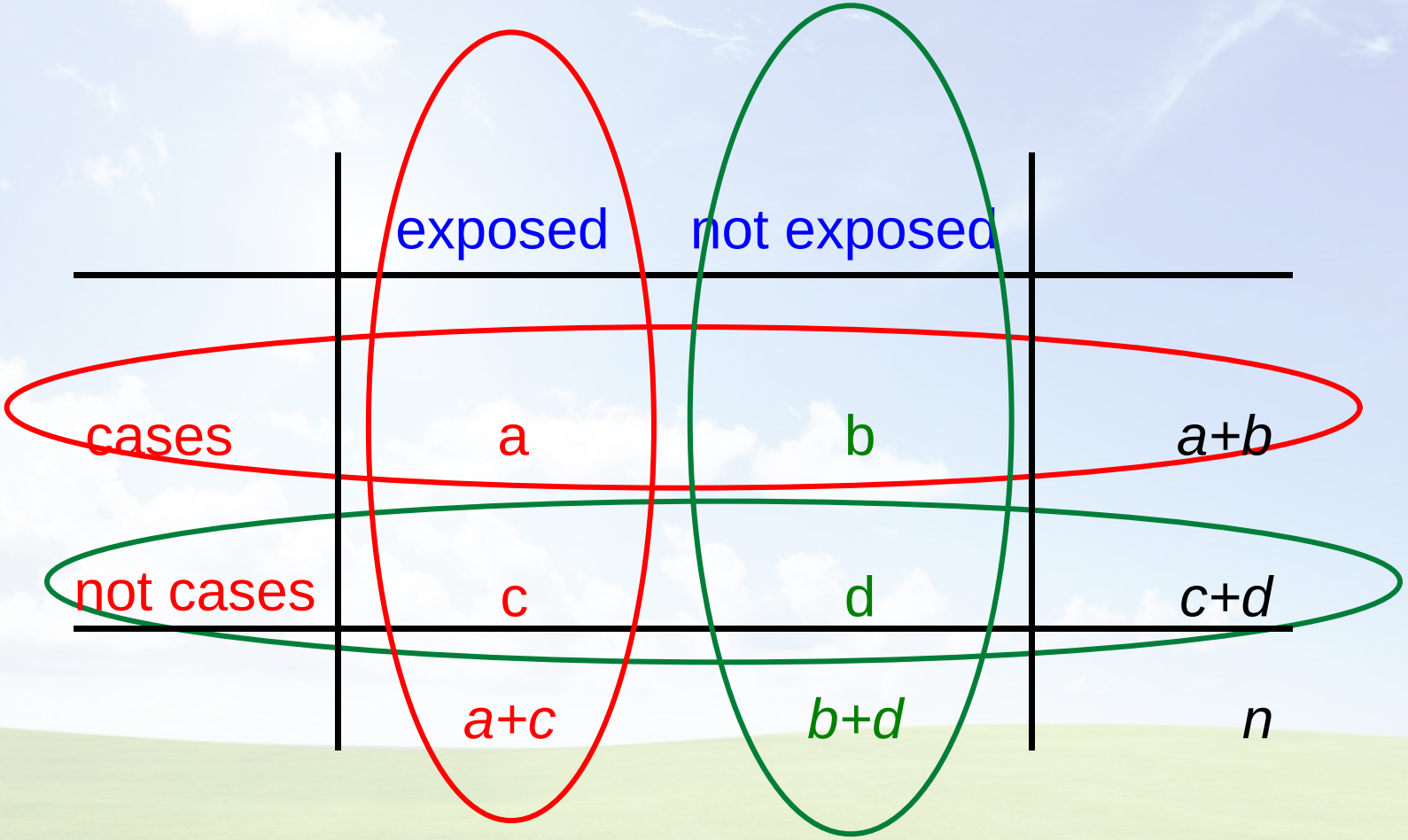
a+b = total cases

c+d = total not cases

a+c = total exposed

b+d = total not exposed

N = population in study


exposed
not exposed
cases
a
b
a+b
not cases
c
d
c+d
a+c
b+d
n

Relative Risk

$$RR = \frac{\text{Risk in the exposed group}}{\text{Risk in the not exposed group}}$$

	exposed	not exposed	
cases	a	b	*a+b*
not cases	c	d	*c+d*
	a+c	*b+d*	*n*

$$RR = \frac{a/(a+c)}{b/(b+d)}$$

Example:

	exposed	not exposed	
cases	80	40	
not cases	20	40	
	100	80	N=180

$$RR = \frac{80/100}{40/80} = \frac{0.8}{0.5} = 1.6$$

Interpretation of Relative Risk

RR > 1 → Exposed subjects have a higher risk than not exposed

RR < 1 → Exposed subjects have a lower risk than not exposed

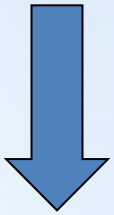
RR =1 → Exposed subjects have the same risk of not exposed

Relative effect= RR – 1

RR = 1.6 Relative effect = 0.6 = 60%

RR = 0.8 Relative effect = - 0.2 = -20%

RR = 1.6



The exposed subjects are 60% more likely to develop the disease than not exposed ones

RR = 0.8

The exposed subjects are 20% less likely to develop the disease than not exposed ones

Rate Ratio

$$RR = \frac{\text{Rate of the exposed group}}{\text{Rate of the not exposed group}}$$

Example:

	exposed	not exposed
cases	80	40
not cases	24,920	15,960
	---	---
person years	*25,000*	*16,000*

$$RT = \frac{\frac{80 \times 10{,}000}{25{,}000}}{\frac{40 \times 10{,}000}{16{,}000}} = \frac{32}{25} = 1.28$$

Odds Ratio

$$OR = \frac{\text{Odds of disease in the exposed group}}{\text{Odds of the disease in the not exposed group}}$$

	exposed	not exposed	
cases	a	b	*a+b*
not cases	c	d	*c+d*
	a+c	*b+d*	*n*

$$OR = \frac{a/c}{b/d} = \frac{ad}{bc}$$

Example:

	exposed	not exposed
cases	80	40
not cases	20	40
	100	80

N=180

$$OR = \frac{80/20}{40/40} = \frac{4}{1} = 4$$

OR is similar to RR when the number of events is small (rare diseases)

	exposed	not exposed
cases	6	2
not cases	94	78
	100	80

RR = 2.40 **OR = 2.49**

STUDY DESIGN

A study design is a specific plan or protocol for conducting the study, which translates the conceptual hypothesis into an operational one.

Types of studies

- **Descriptive studies**

describe the occurrence of the disease

- **Analytic studies**

describe the association between exposure and disease

Descriptive

Case report

Case series

Descriptive Epidemiology

Analytic

RCT

Cohort study

Case-Control

Cross-sectional

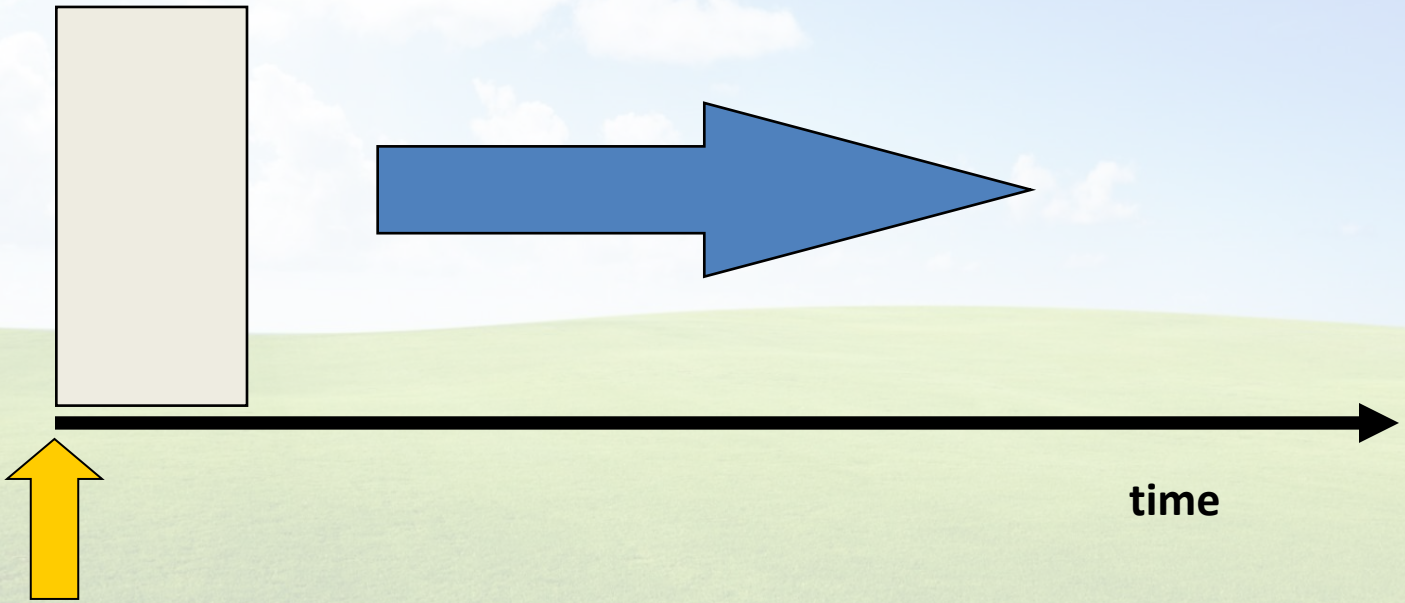
Ecologic study

Case-Crossover

Before-After

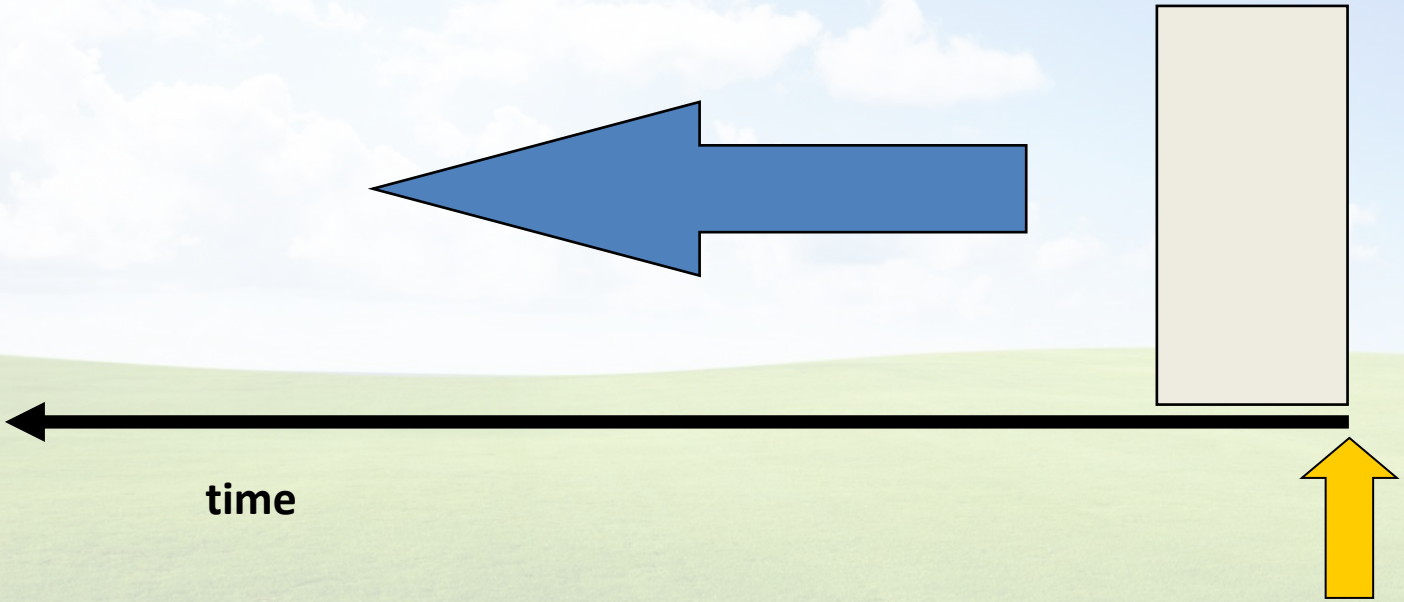
Timeframe of Studies

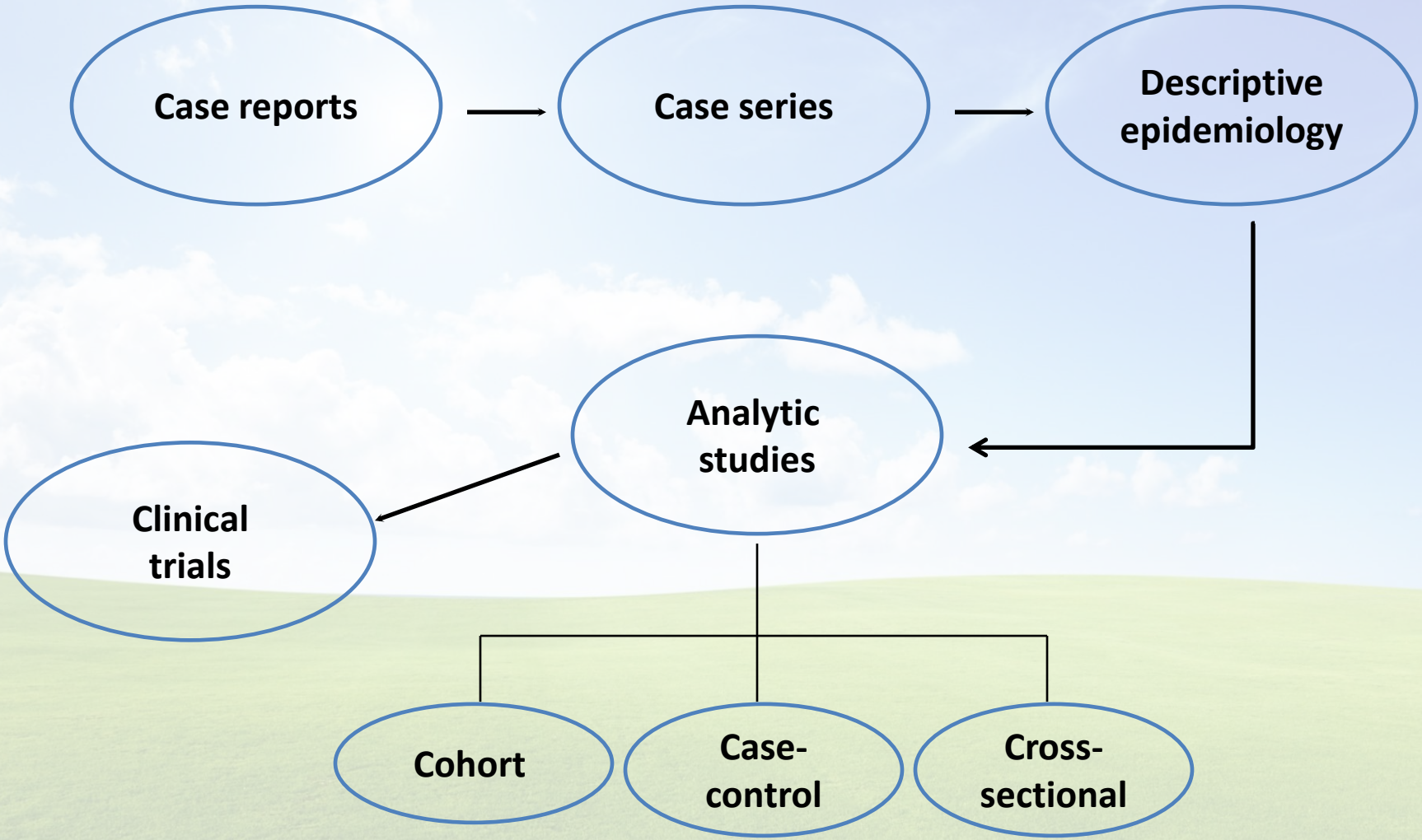
Prospective Study : looks forward, new events



Timeframe of Studies

Retrospective Study : looks back, events already occurred




Case reports
Case series
Descriptive epidemiology
Analytic studies
Clinical trials
Cohort
Case-control
Cross-sectional

STUDY DESIGNS

Observational

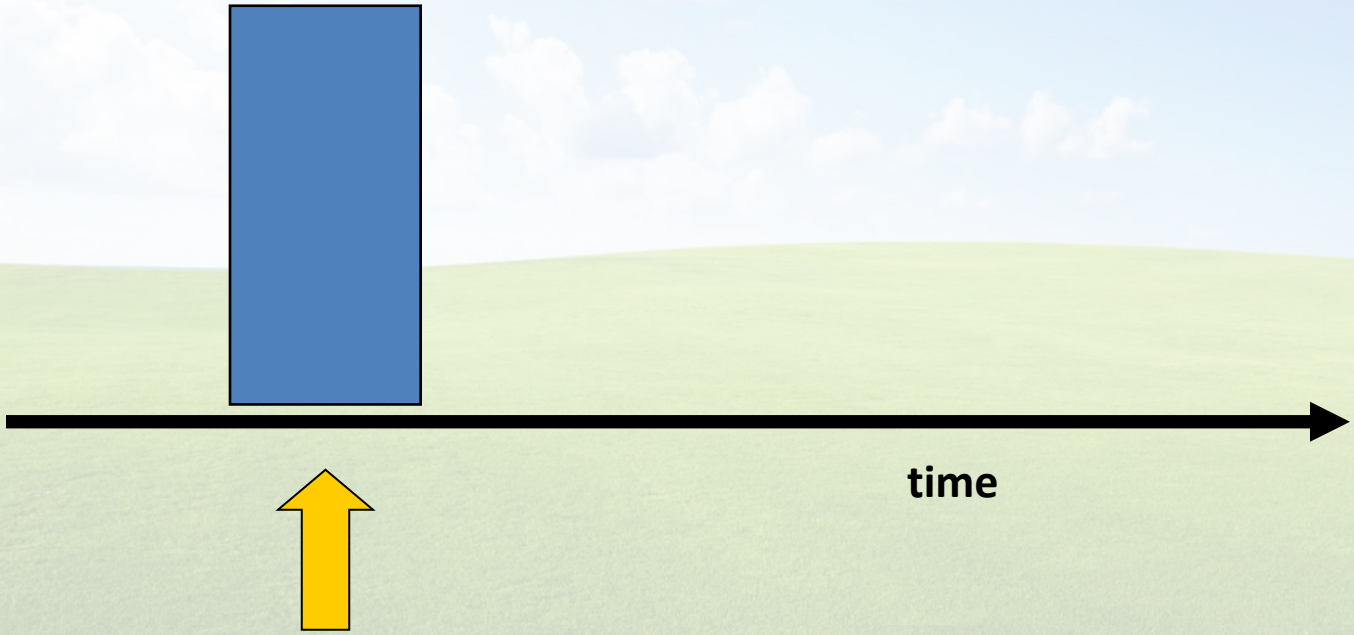
The researcher studies but does not alter what occurs

Experimental

The researcher intervenes to change reality, then observes what happens

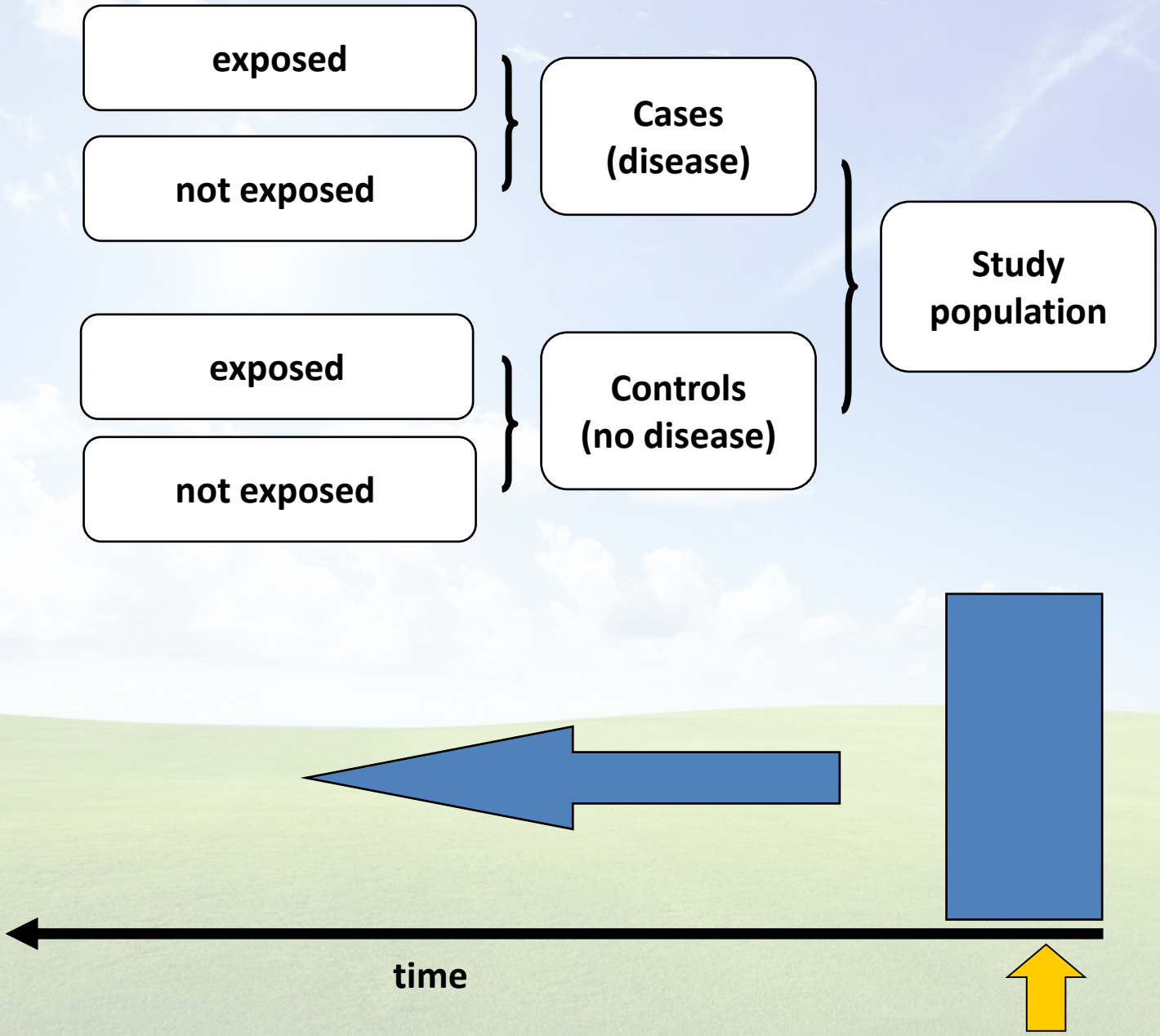
Cross-sectional study

An observational design that surveys exposures and disease status at a single point in time (a cross-section of the population)



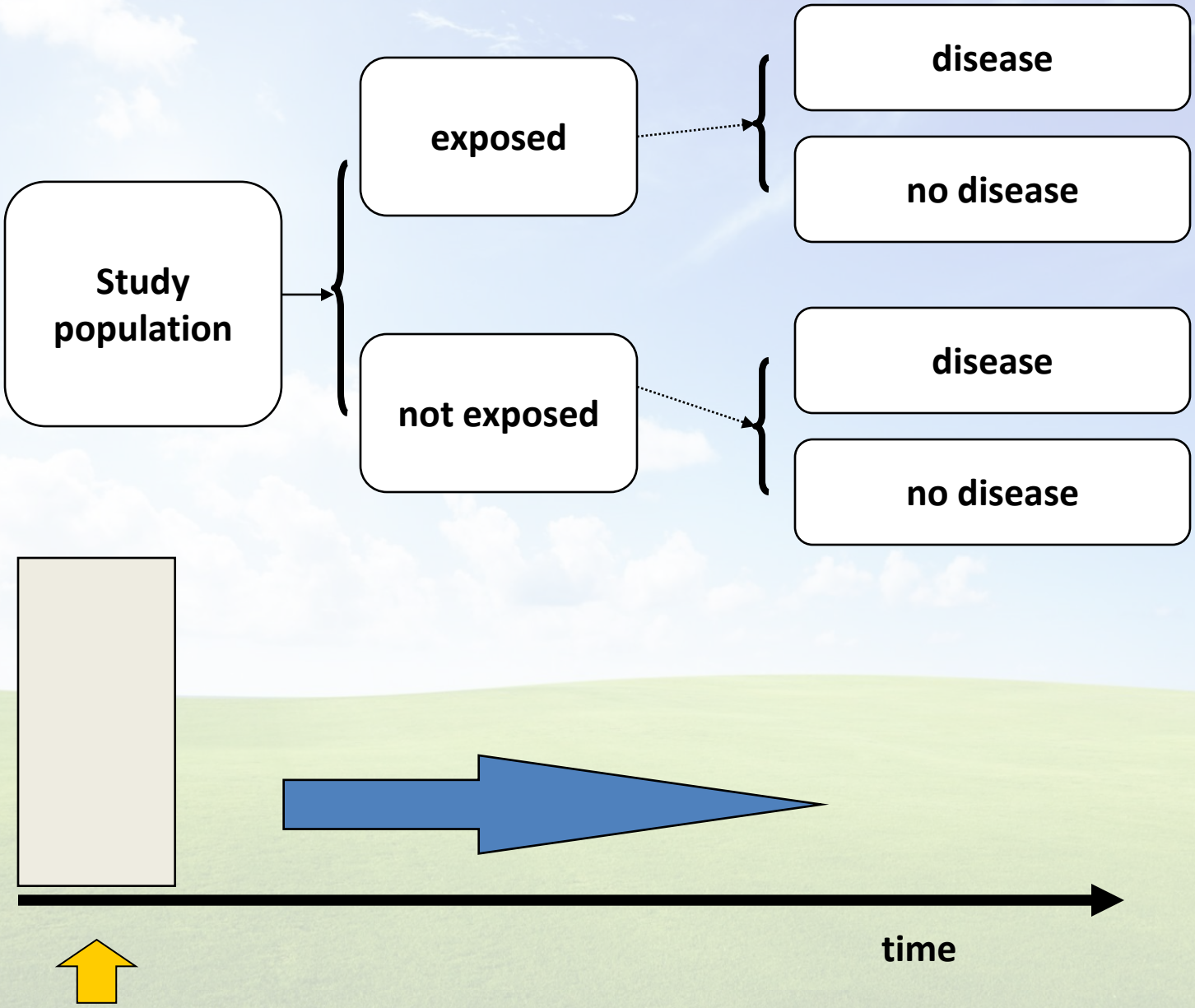
Case-control study

- An observational design comparing exposures in diseased cases vs. healthy cases (controls) from the same population
- exposure data collected retrospectively
- most feasible design where disease outcomes are rare

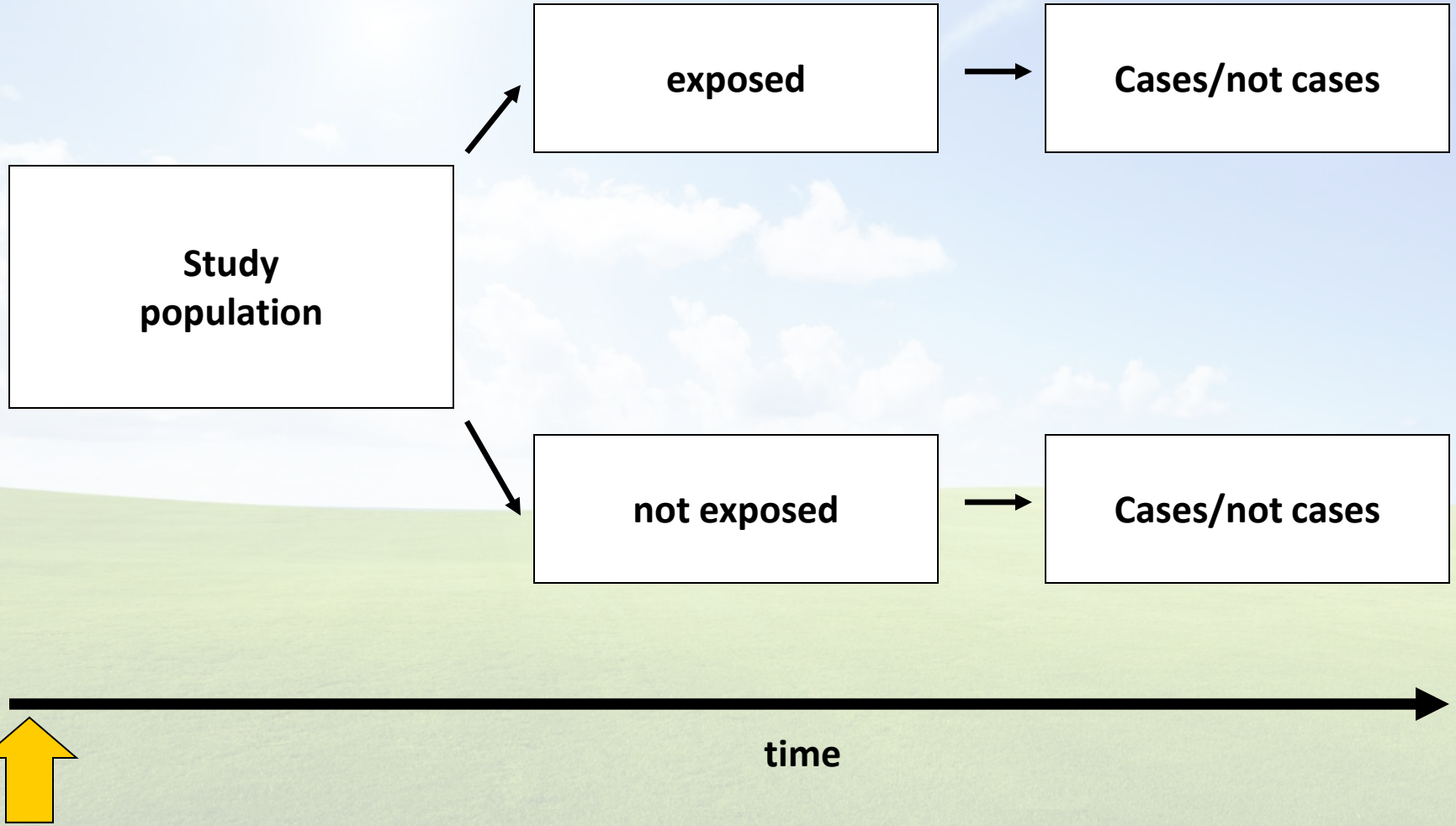

exposed
not exposed
Cases
(disease)
exposed
not exposed
Controls
(no disease)
Study
population
time

Cohort study

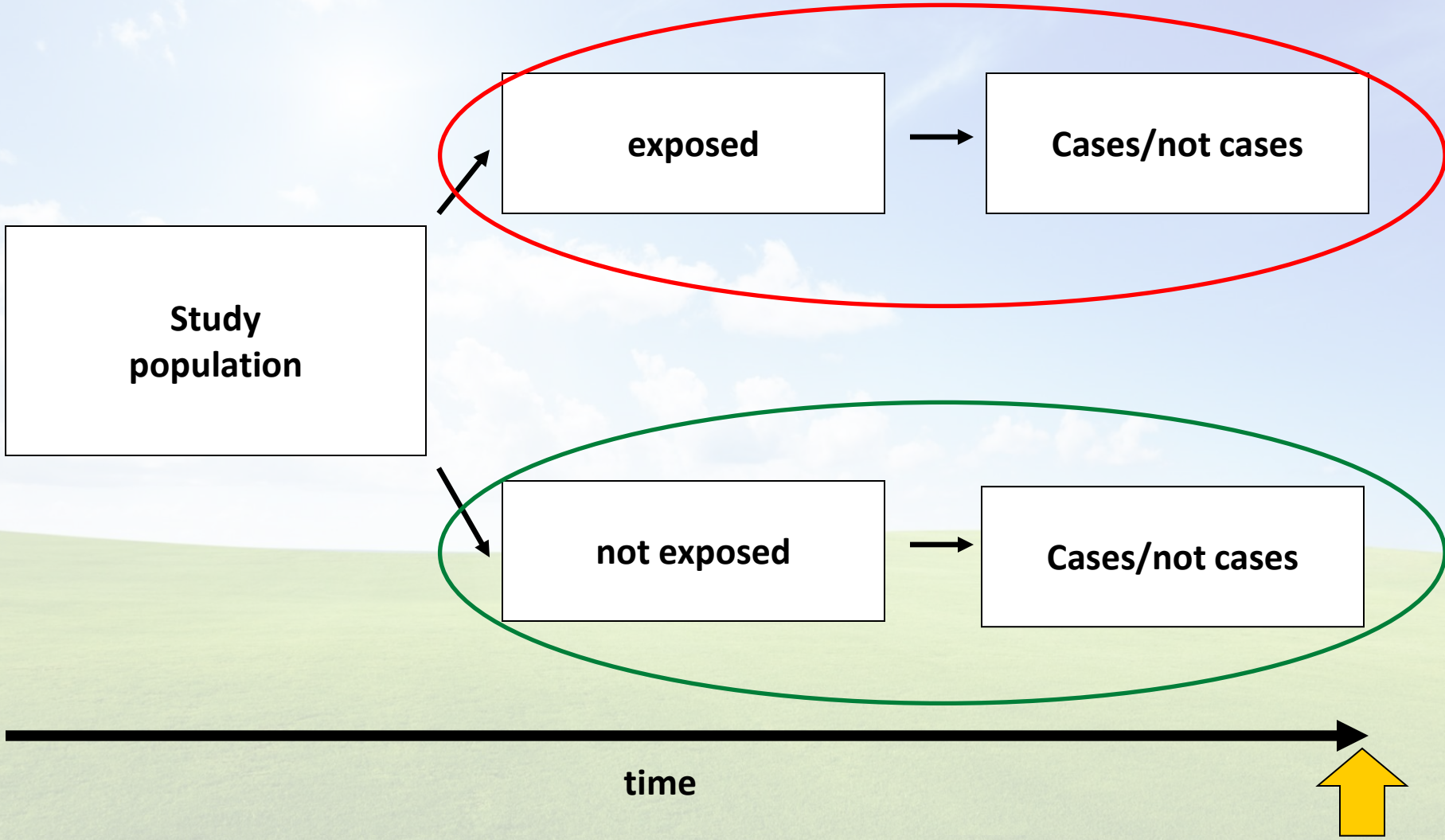
- an observational design comparing individuals with a known exposure with others without the exposure
- looking for a difference in the risk (incidence) of a disease over time
- data usually collected prospectively (sometimes retrospectively)
- Inefficient for rare diseases and diseases with long latency


Study population
exposed
not exposed
disease
no disease
disease
no disease
time

Prospective Cohort study

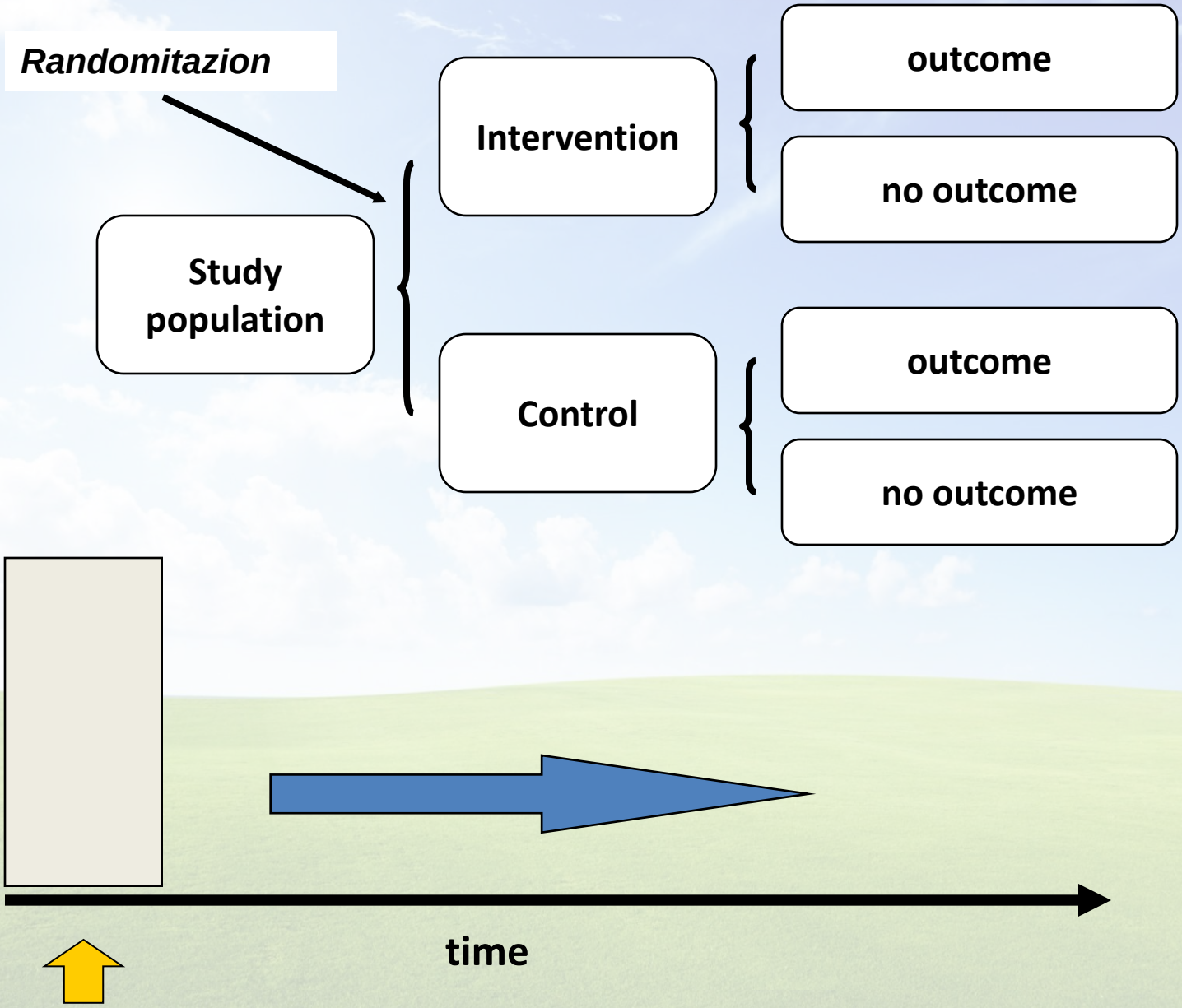


Retrospective Cohort study



Randomized Controlled Trials

- An experimental design in which we are interested in the consequences of some specific treatment on some specific outcomes
- compare subjects randomly assigned to treatment (treatment group) and comparison groups
- provides most convincing evidence of relationship between exposure and effect
- not possible to use RCTs to test effects of exposures that are expected to be harmful, for ethical reasons


Randomitazion
Study population
Intervention
Control
outcome
no outcome
outcome
no outcome
time

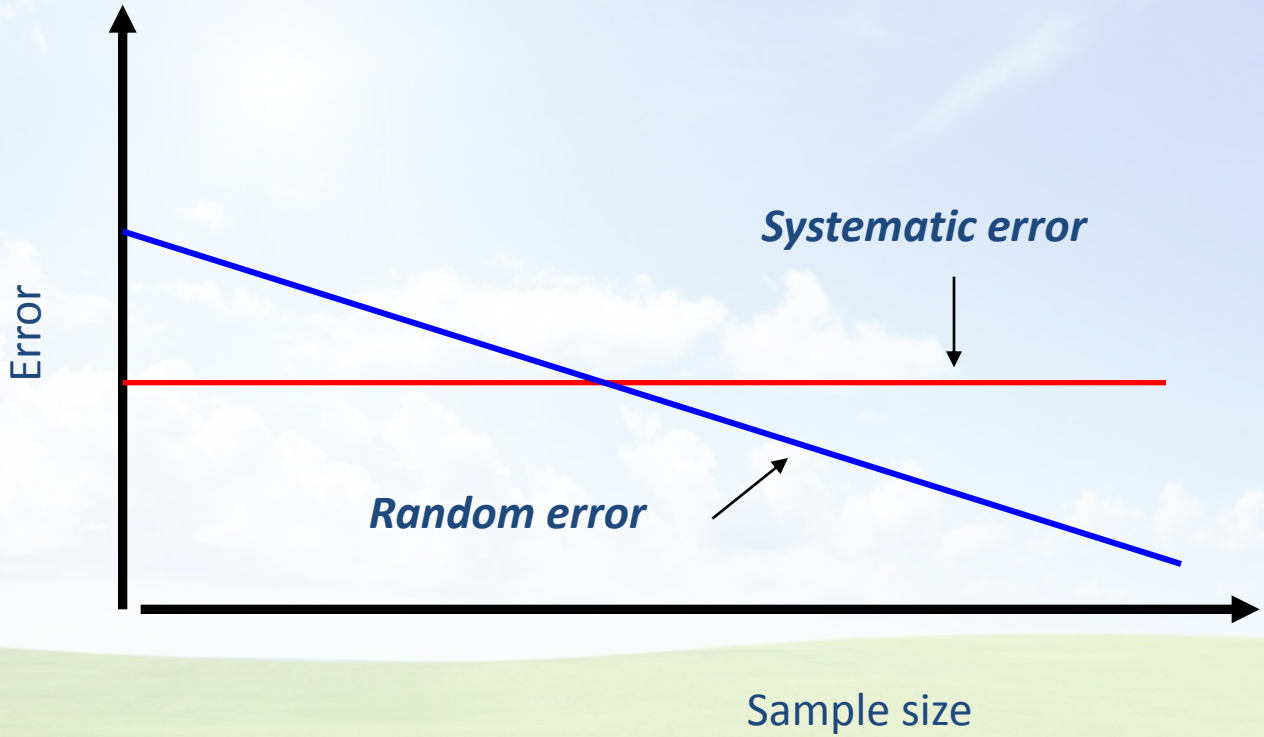
BIAS

Epidemiology is a quantitative discipline.
The objective of the epidemiological studies is to measure (exposure and outcome).

Measurement → Error

Errors in Epidemiology

- Systematic error (bias)
  - Selection bias
  - Information bias
  - Confounding
- Random error


Error
Systematic error
Random error
Sample size

- Systematic error (bias)
  - **Selection bias**
  - **Information bias**
  - Confounding
- Random error

- **Selection bias**

  It is a systematic error due to a wrong procedure used to select subjects and from factors that influence study participation → *non-representative sample*

- **Information bias**

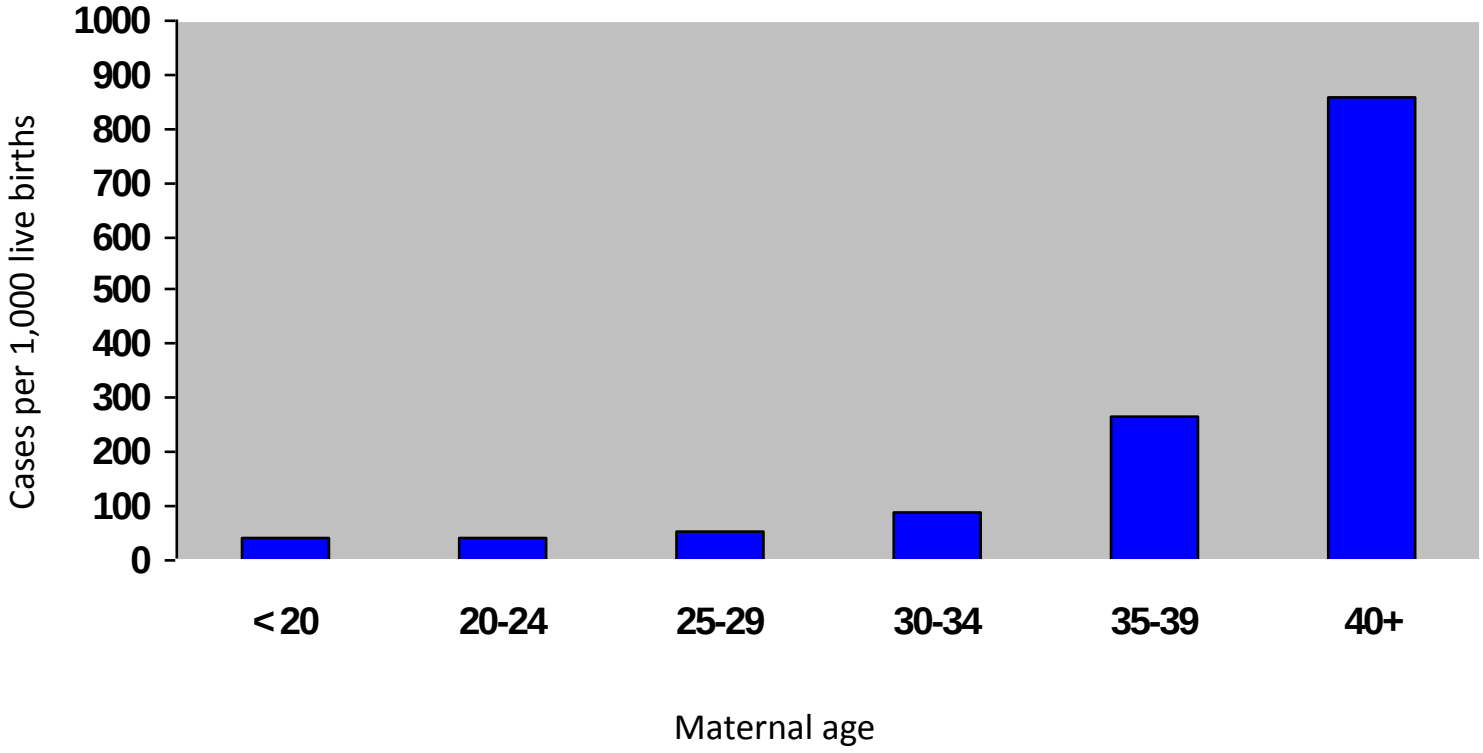
  Systematic error due to inaccurate measurement or classification of disease, exposure or other variables (i.e. Misdiagnosis, Recall bias, Missing data,etc..)

Prevalence of Down syndrome at birth by birth order

(Source:Stark et al. 1966)

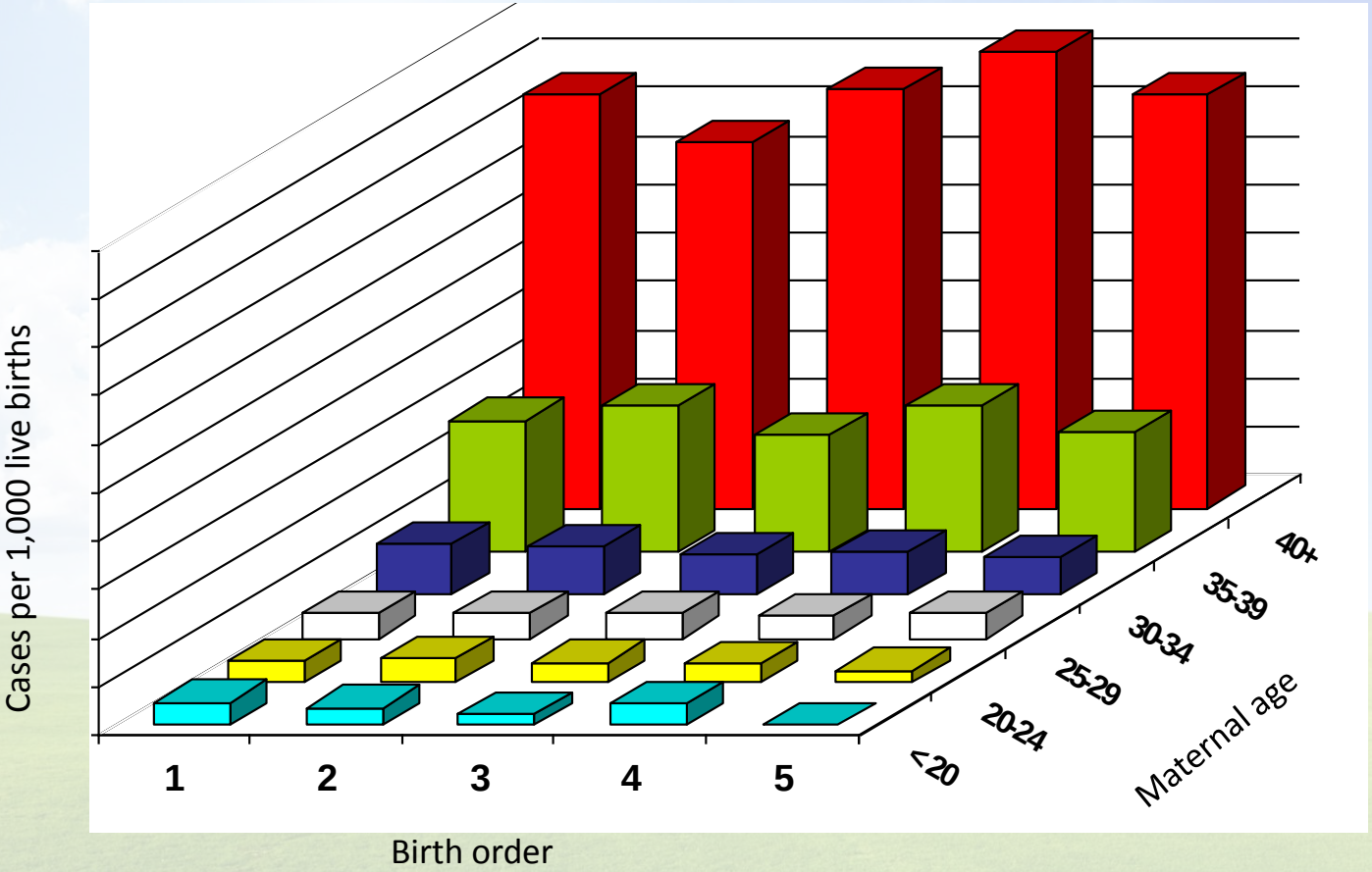
Prevalence of Down syndrome at birth by mother's age

(Stark et al. 1966)



Prevalence of Down syndrome by birth and mother's age

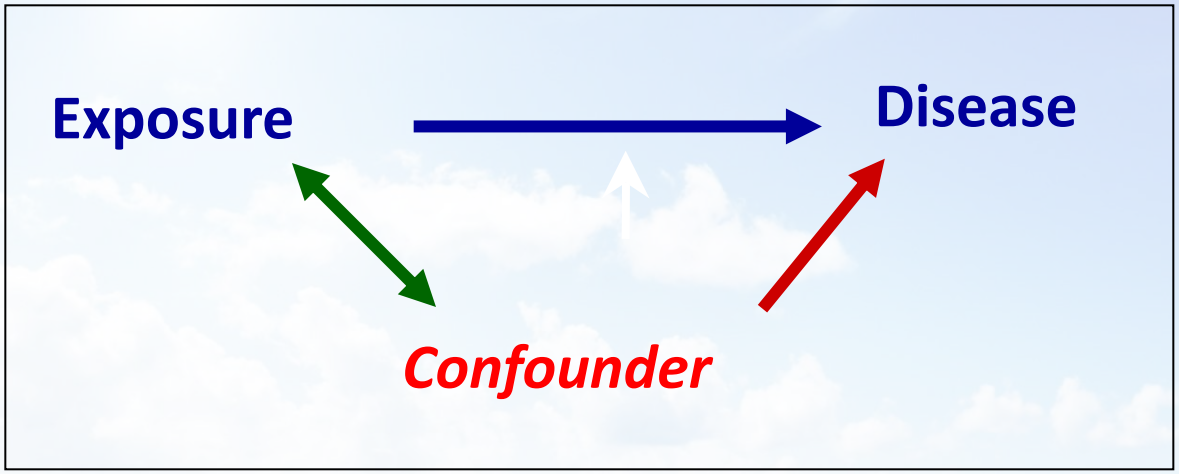
(Stark et al. 1966)

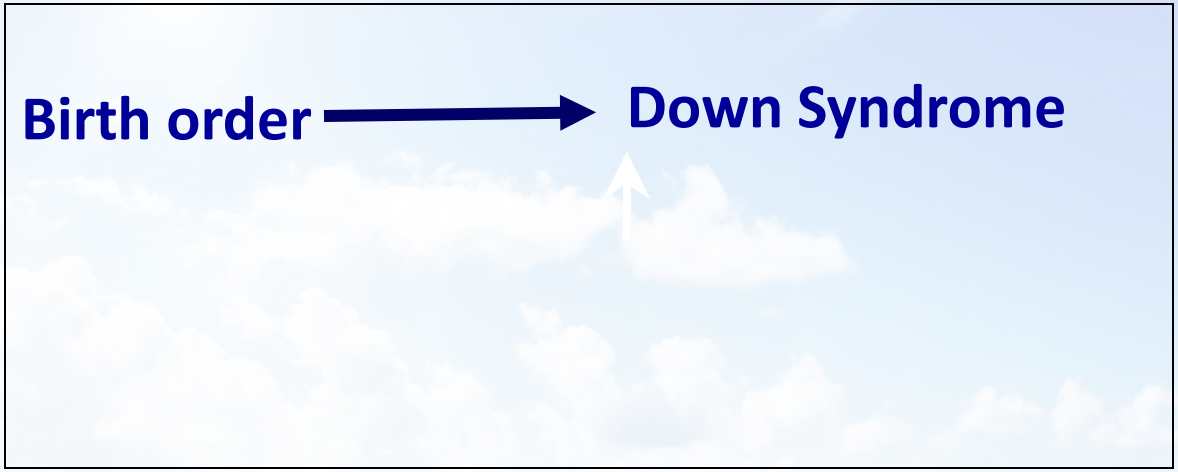


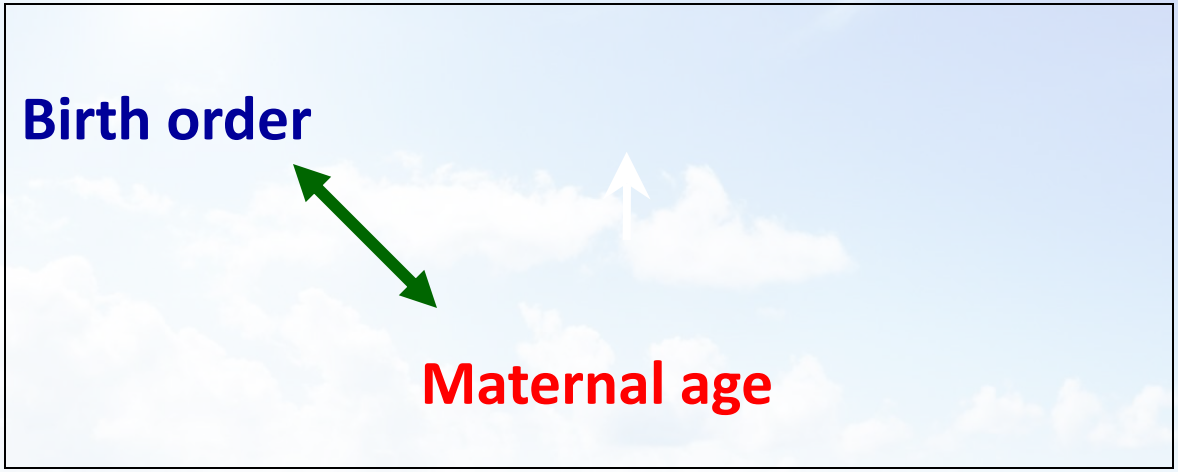
- Systematic error (bias)
  - Selection bias
  - Information bias
  - **Confounding**
- Random error

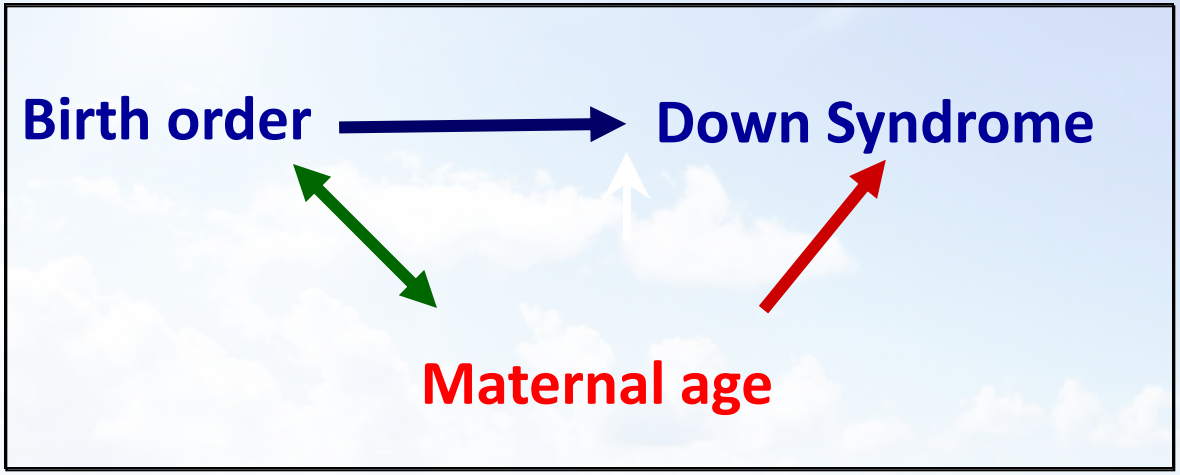
Confouding

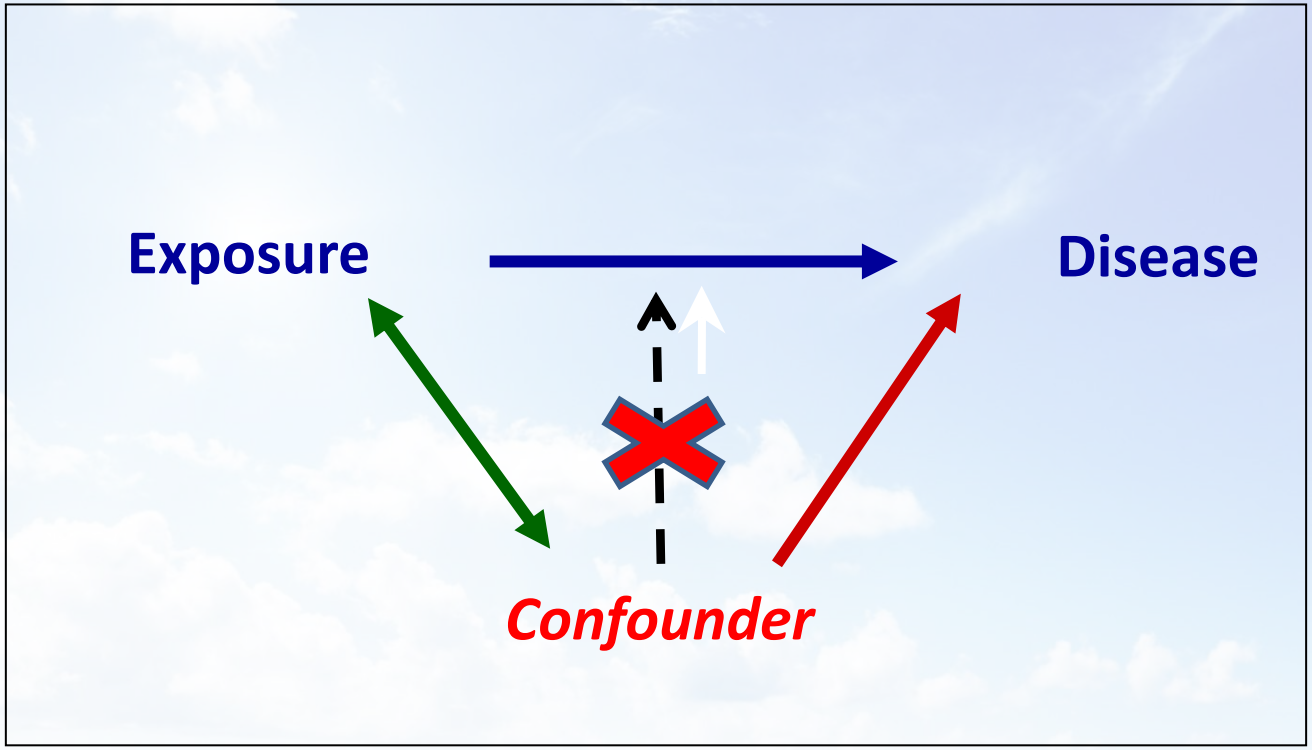
A situation in which the association between an exposure and an outcome is distorted by the presence of another variable (confounder)


Exposure
Disease
Confounder


Birth order
Down Syndrome


Birth order
Maternal age

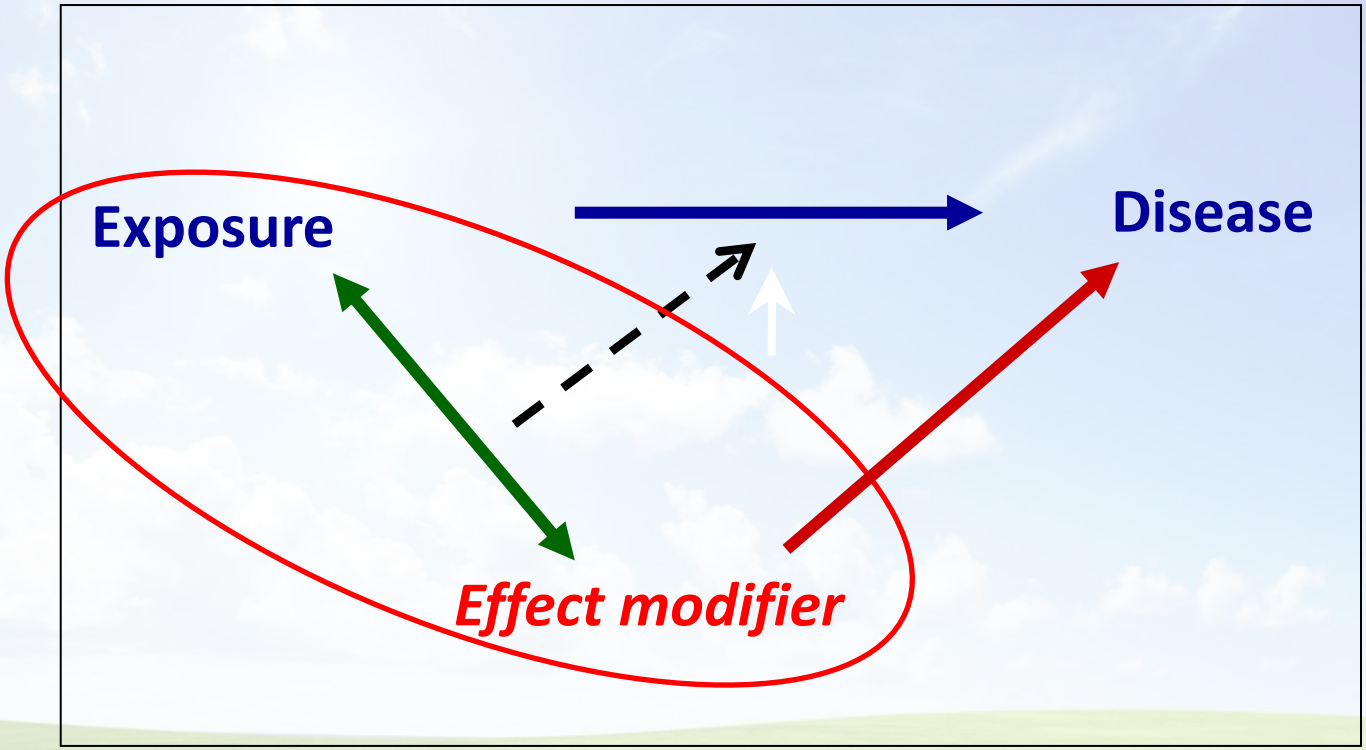

Birth order
Down Syndrome
Maternal age



Conditions for a factor to be a Confounder:

- Must be a risk factor for the disease
- Must be associated with the exposure
- Must not be an intermediate step in the causal path

Effect modification



A variable that differentially modifies the effect of an exposure factor on disease: different levels of the *effect modifier* change the magnitude of the association between the primary exposure and the outcome

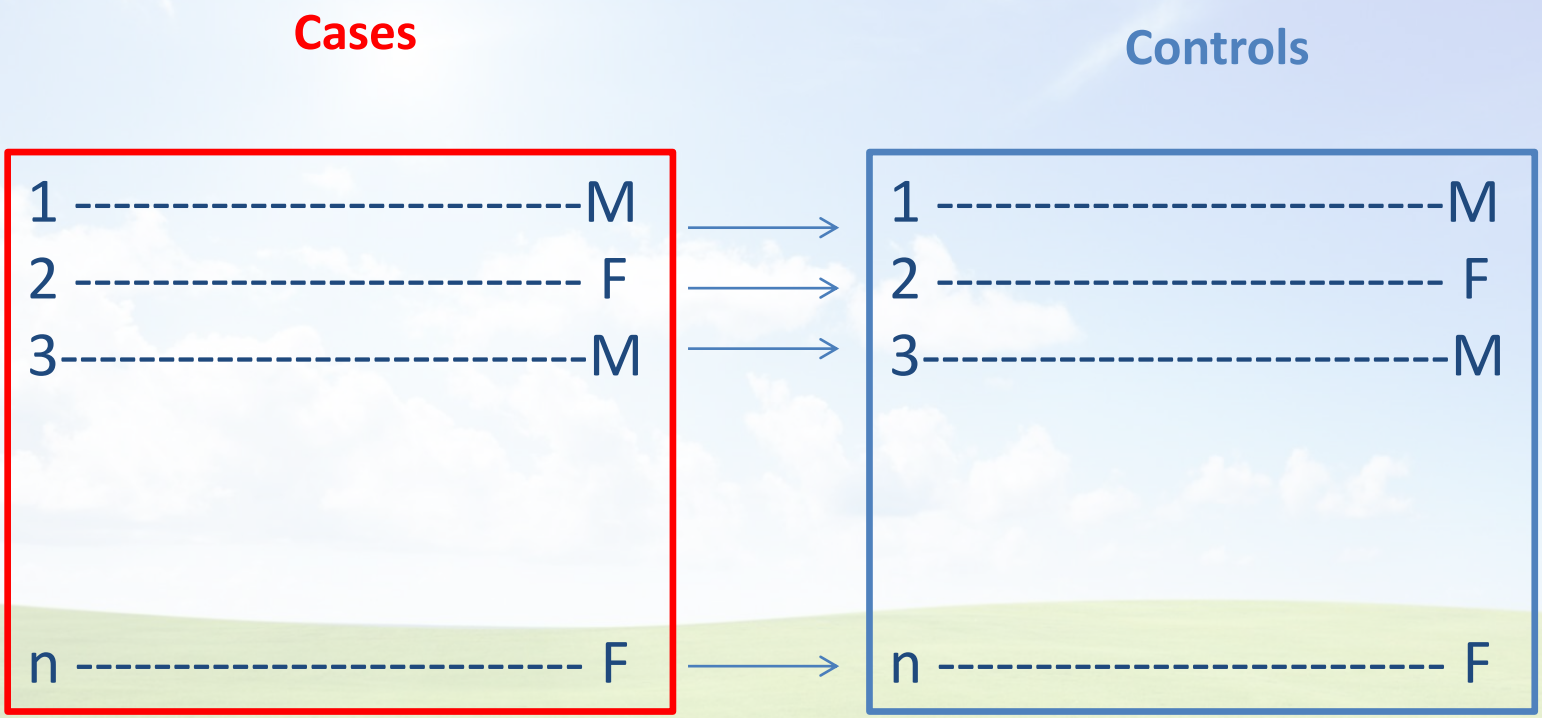
How to control for confounding

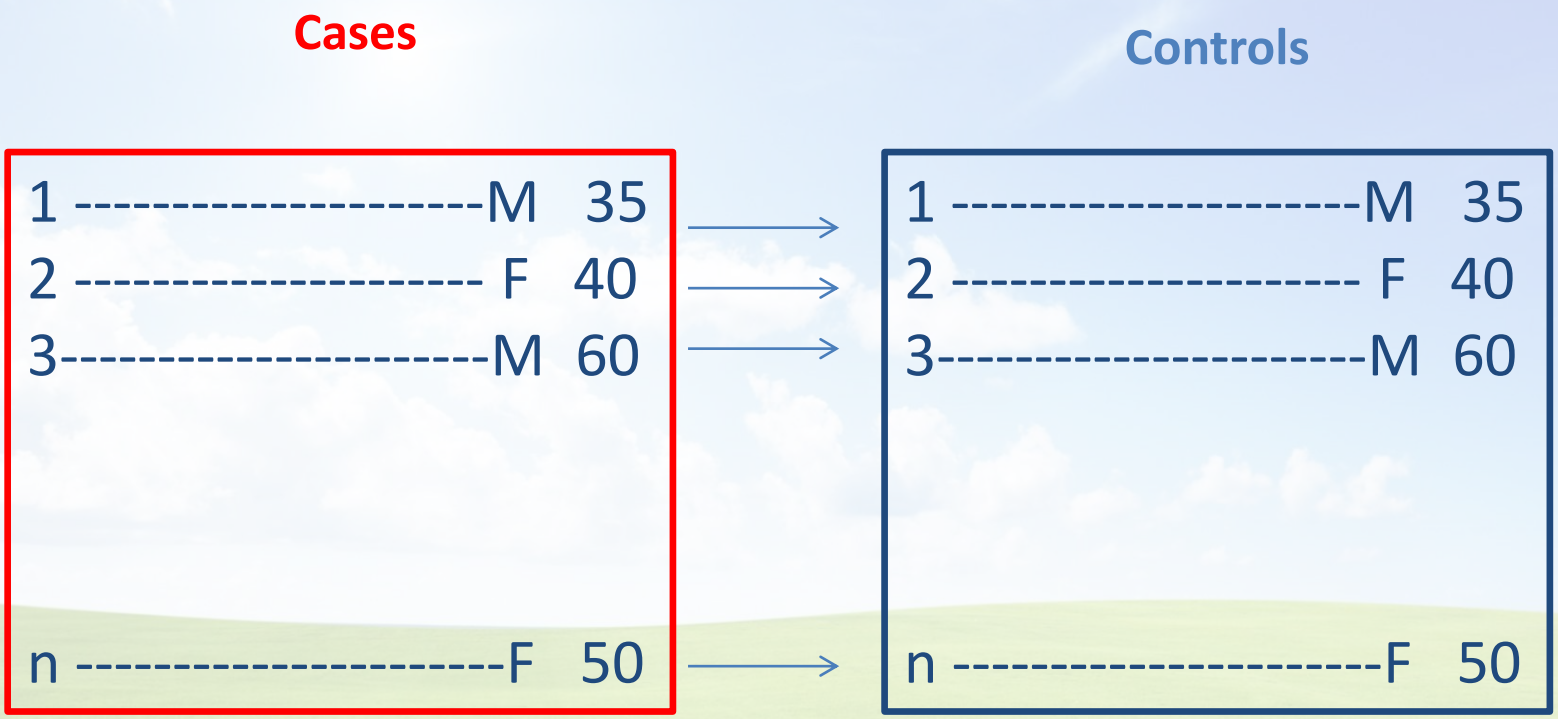
- At the design phase
  - Randomization
  - Matching
- At the analysis phase
  - Stratification
  - Standardization
  - Multivariable adjustment (Linear regression, Logistic regression, Poisson regression, Cox regression modeling)

Sample matching

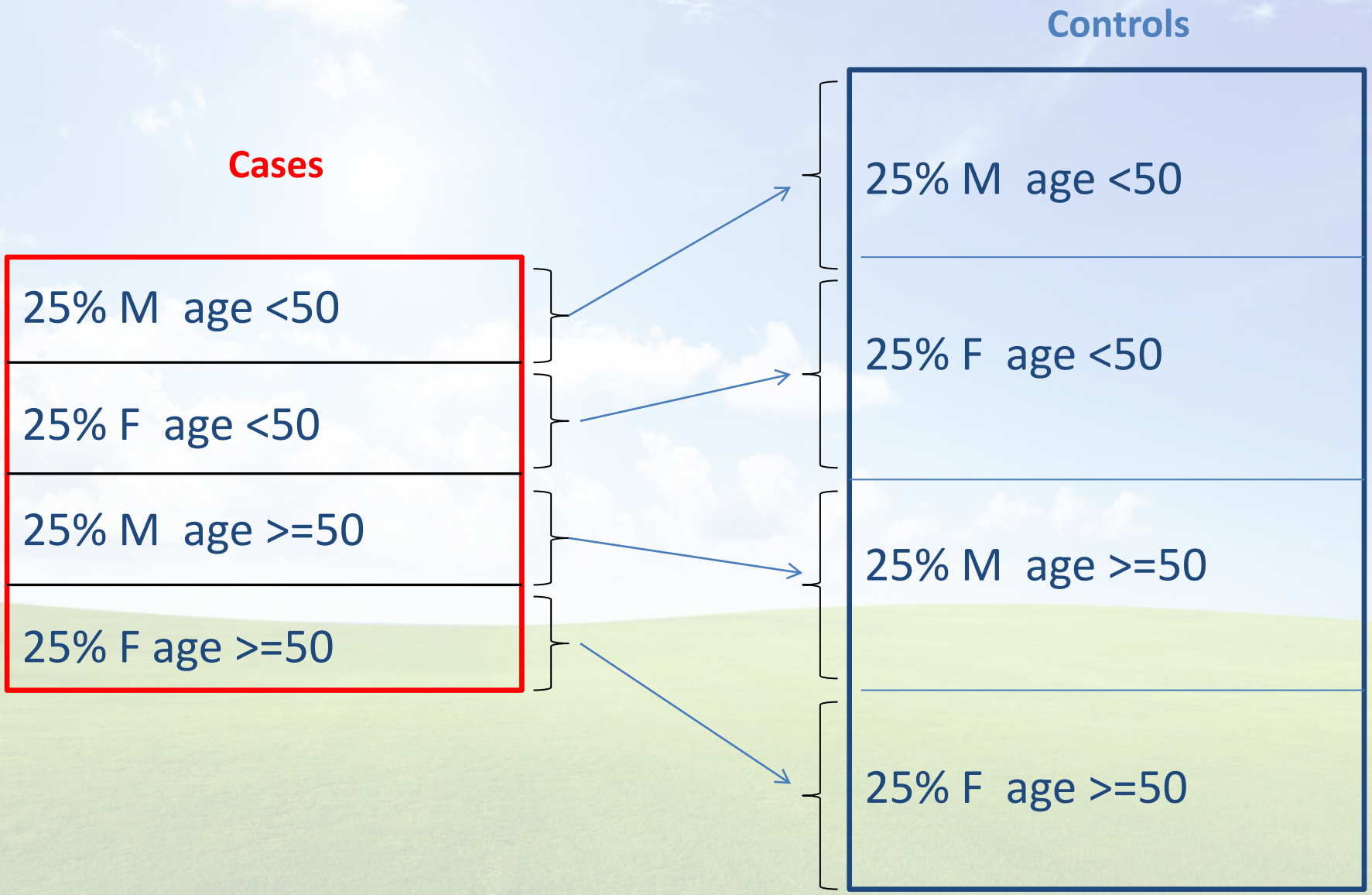
A pair of matched samples are those in which each member of a sample is matched with a corresponding member in the other sample using by reference variables other than those immediately under investigation.

Individual matching




Cases
Controls
1 ----------M 35
2 ---------- F 40
3----------M 60
n ----------F 50
1 ----------M 35
2 ---------- F 40
3----------M 60
n ----------F 50

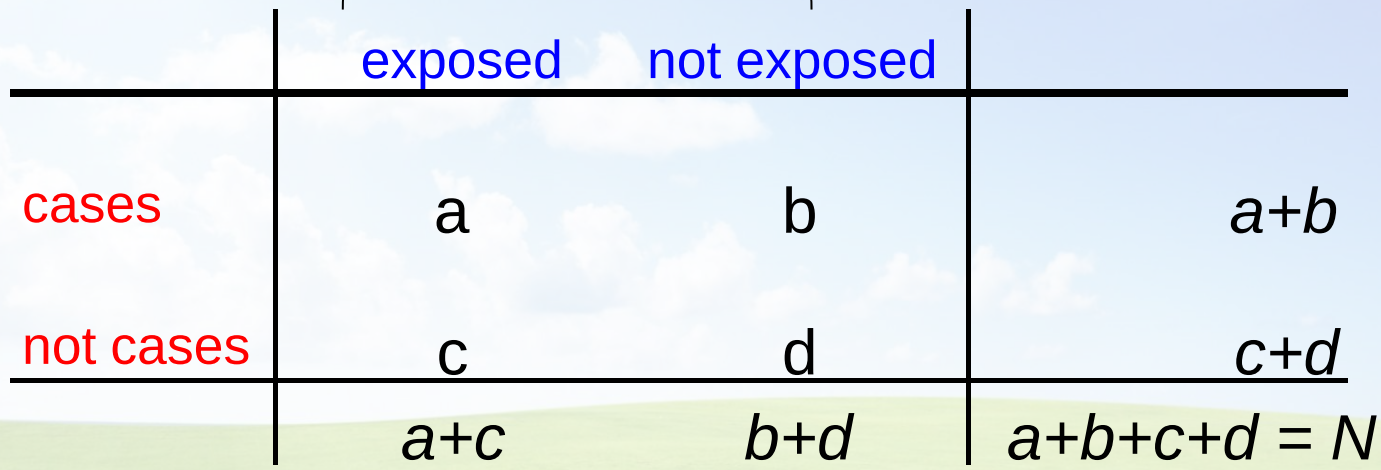
Frequency matching



How to control for confounding

- At the design phase
  - Randomization
  - Matching
- At the analysis phase
  - Stratification
  - Standardization
  - Multivariable adjustment (Linear regression, Logistic regression, Poisson regression, Cox regression modeling)

Confounder



Standardization

Method of combining category-specific rates into a single summary value by taking a weighted average of them. It weights category-specific rates using weights that come from a standard population.

The result represents the “behaviour” of the groups in study if they had the same distribution of the confounding variable

Multiple regression model

$$Y = aX_1 + bX_2 + cX_3 + dX_4 + k$$

Using multiple regression models in epidemiologic analysis

$$Y = aX_1 + bX_2 + cX_3 + dX_4 + k$$

Y = outcome

X_1 = primary exposure

X_2; X_3; X_4 = Confounders

Example: Logistic regression model

$$Y = aX_1 + bX_2 + cX_3 + dX_4 + k$$

a = Odds Ratio of X_1 adjusted for other variables

- Systematic error (bias)
  - Selection bias
  - Information bias
  - Confounding
- **Random error**

After bias is eliminated, the error remained is the Random error: it arises from an unpredictable process (chance)

Statistic is used to evaluate the Random error

"An epidemiologic study can be viewed as an exercise in measurement. As in any measurement the goal is to obtain an accurate result, with as little error as possible"

(K. Rothman)