Commonly used statistical models in pharmacoepidemiology

Klaus Kaae Andersen

May 2, 2019

What model to use?

- The choice of statistical model depends on
  - Available data
  - Applied design
  - Research question

- In this presentation we are assuming that we are working with registry data as we know them in Denmark.

- As a starting point when introducing the models we will assume that no cofounders are present.
- We furthermore assume that we are modelling an event, e.g. getting cancer or dying from cancer.
- As seen in the previous lecture on designs, we often establish models by considering
  - Case control data
  - Cohort data

Reporting

THE FOUR MAIN STEPS IN DATA ANALYSIS AND REPORTING FOR CLINICAL TRIALS

1 What to include in result tables and figures

Characteristic
Age (yrs)
Female, n (%)
Previous myocardial infarction (MI), n (%)
Race, n (%)
Black, white, asian, other...

Table of Baseline Data

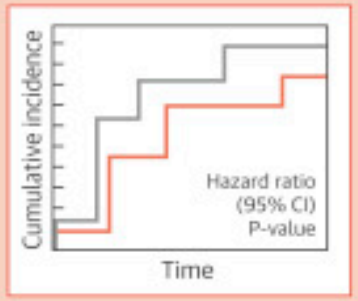
First table for any clinical trial report

- Total nos. of patients per group
- Key demographic variables
- Related medical history
- Other endpoint-related variables

Endpoint
Cardiovascular death
Death from any cause
MI
Ischemic stroke
Repeat hospitalization
Hospitalization for heart failure

Table of Main Outcome Events

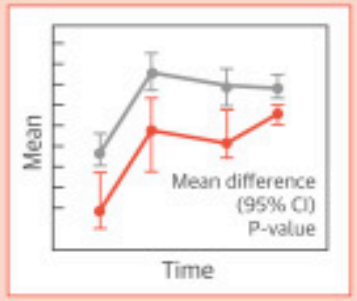
- Main outcome by group
- Nos. (%) experiencing endpoint by group
- For composite endpoints report nos. (%) experiencing each component event
- Analysis of first and subsequent events



Kaplan-Meier Plot of cumulative incidence over time, by group

Common figure in major trial reports

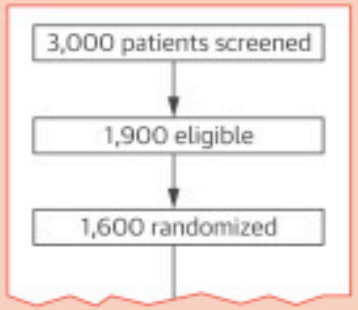
- Focus on cumulative incidence
- Sensible vertical axis range
- Report number at risk over follow-up time



Repeated Measures Over Time

Figure to show change in mean over time by group

- Standard error bars to express uncertainty



Trial Profile

Flow of patients through trial

- Nos. of eligible patients identified
- Nos. randomized into trial
- Nos. lost to follow-up
- Nos. included in analysis

2 Quantify associations

Estimate treatment effect (numerous methods):

- Relative risk/relative odds for binary outcomes
- Relative risk reduction
- Absolute difference in percentage
- Number Needed to Treat (NNT)
- Hazard ratio for time-to-event outcomes
- Mean difference using ANCOVA for quantitative outcomes

3 Express uncertainty

Confidence interval

Estimates will always have built-in imprecision because of the finite sample of patients studied

- Always acknowledge a degree of uncertainty (95% confidence interval, "95% CI")
- Larger studies provide more reliable estimates with tighter confidence intervals (i.e., 99% CI)

4 Assess evidence

P values and interpretation

Determine whether there is real treatment effect

The *smaller* the value of P the *stronger* the evidence to contradict the null hypothesis of no true treatment difference

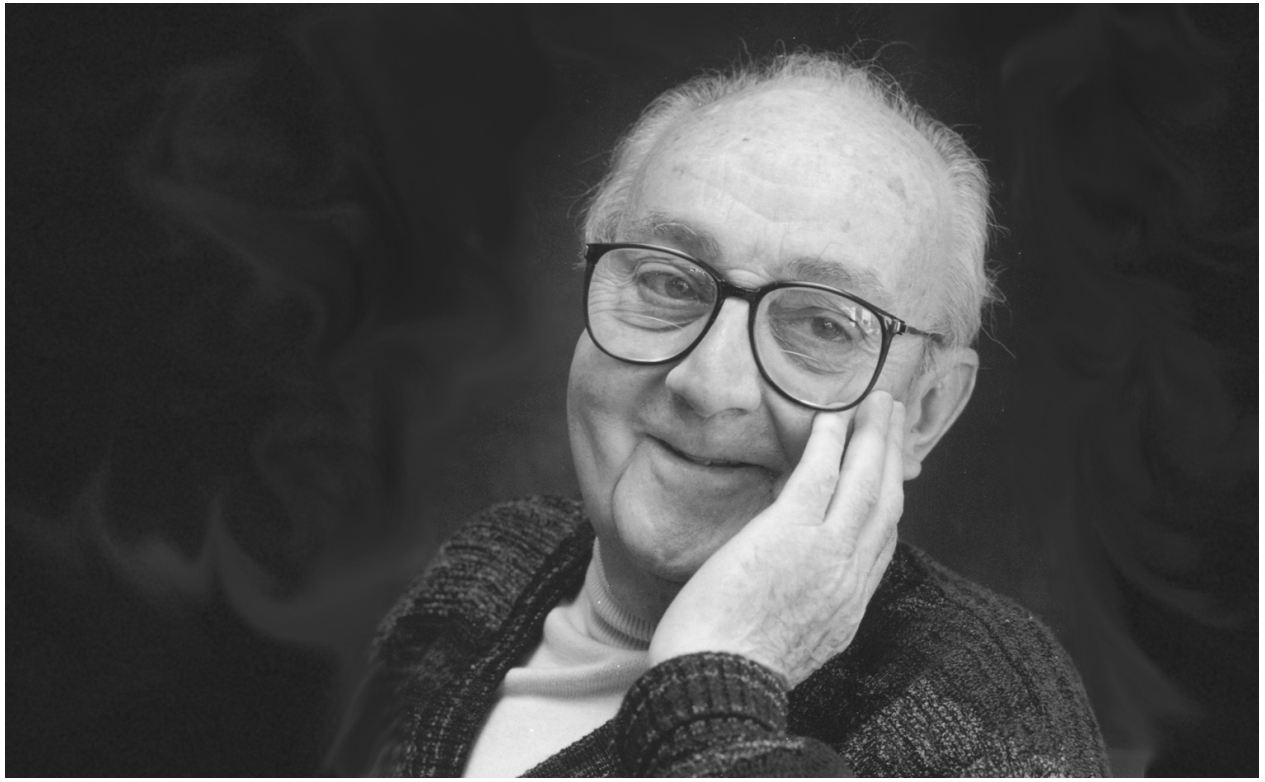
- Report actual p value, i.e., p = 0.042
- Note if p value meets significance level ($p < 0.05$)
- Use two-sided p values

Pocock, S.J. et al. J Am Coll Cardiol. 2015; 66(22):2536-49.

Remember

“All models are wrong, but some are useful”

George Box



The principle of parsimony

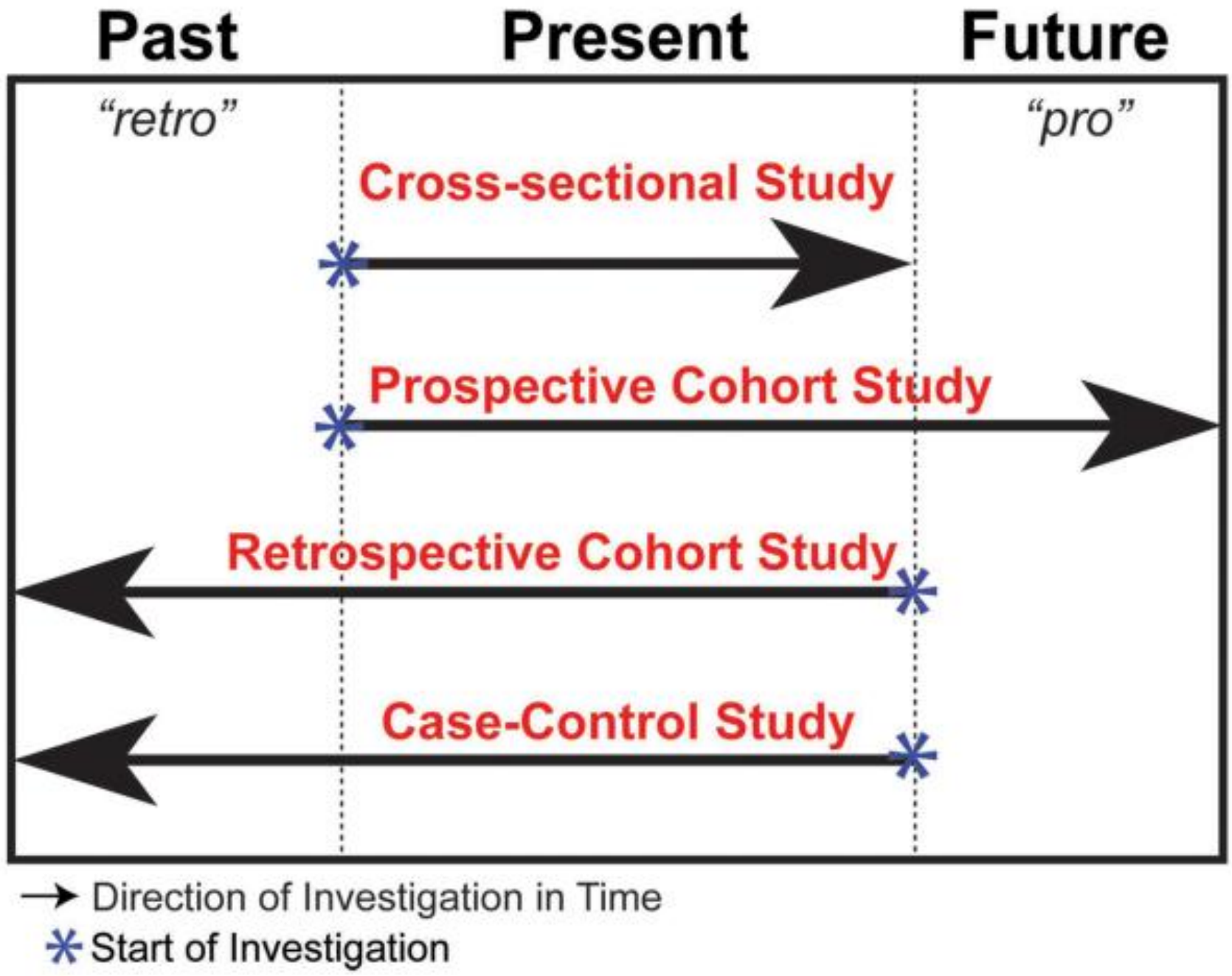
- The parsimony principle is basic to all science and tells us to choose the simplest scientific explanation that fits the evidence.

Law of the instrument

TIME WELL SPENT™

marketoonist

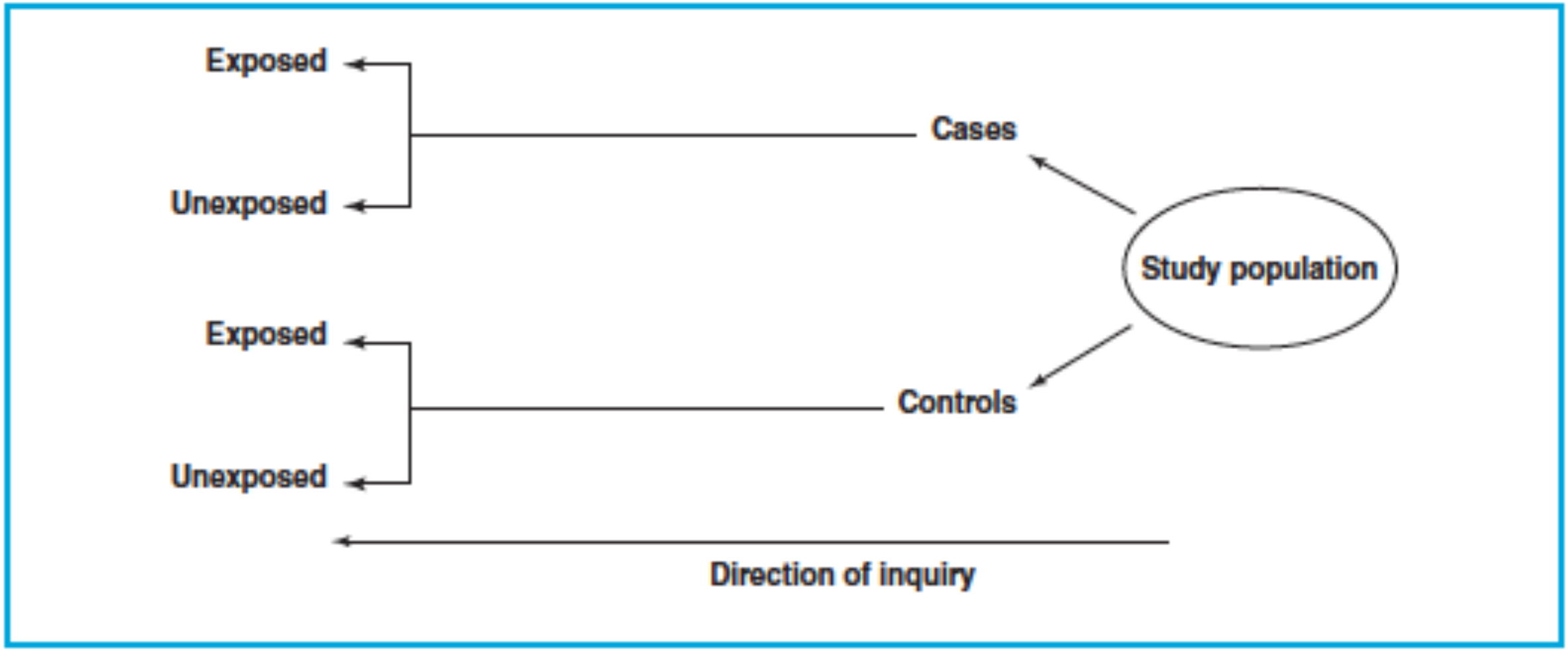
KRONOS.COM/TIMEWELLSPENT


Past
Present
Future
"retro"
"pro"
Cross-sectional Study
Prospective Cohort Study
Retrospective Cohort Study
Case-Control Study
Direction of Investigation in Time
Start of Investigation

Case–control studies

Case–control studies are observational studies in which the starting point is the identification of ‘cases’ of the disease (or condition) of interest, and of suitable ‘controls’ without that disease (or condition).

Cases and controls are then compared to assess whether there were any differences in their past exposure to possible risk factors.


Exposed
Unexposed
Cases
Study population
Exposed
Unexposed
Controls
Direction of inquiry

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta \cdot T$$

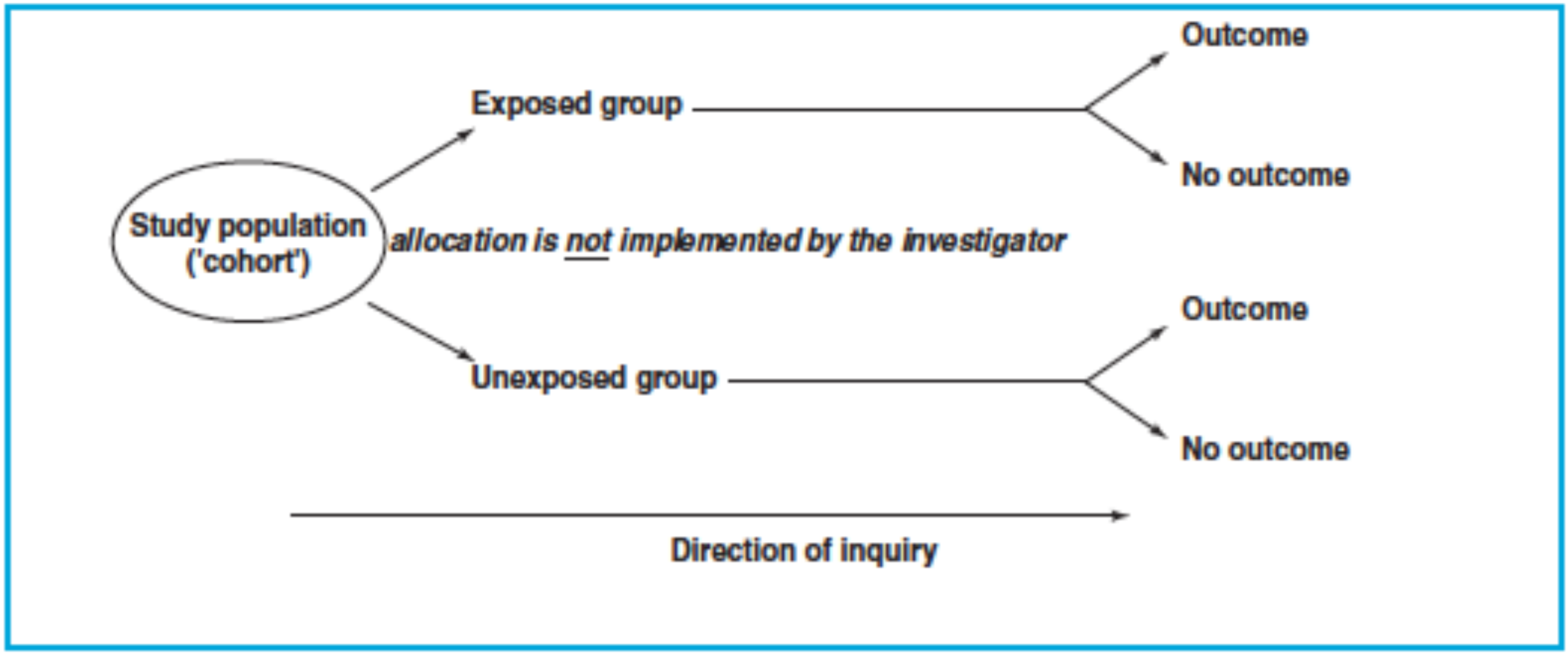
$$p = \frac{e^{\alpha+\beta\cdot T}}{1+e^{\alpha+\beta\cdot T}}$$

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta \cdot T + \sum \beta_j f(x_j).$$

Cohort studies

- Cohort studies are observational studies in which the starting point is the selection of a study population, or cohort. Information is obtained to determine which members of this cohort are exposed to the factor of interest. The entire population is then followed up over time and the incidence of the disease in the exposed individuals is compared with the incidence in those not exposed .

- This type of observational study is the one that most closely resembles intervention studies, except that allocation of subjects to the exposure is not controlled by the investigator.


Outcome
Exposed group
No outcome
Study population
('cohort')
allocation is not implemented by the investigator
Outcome
Unexposed group
No outcome
Direction of inquiry

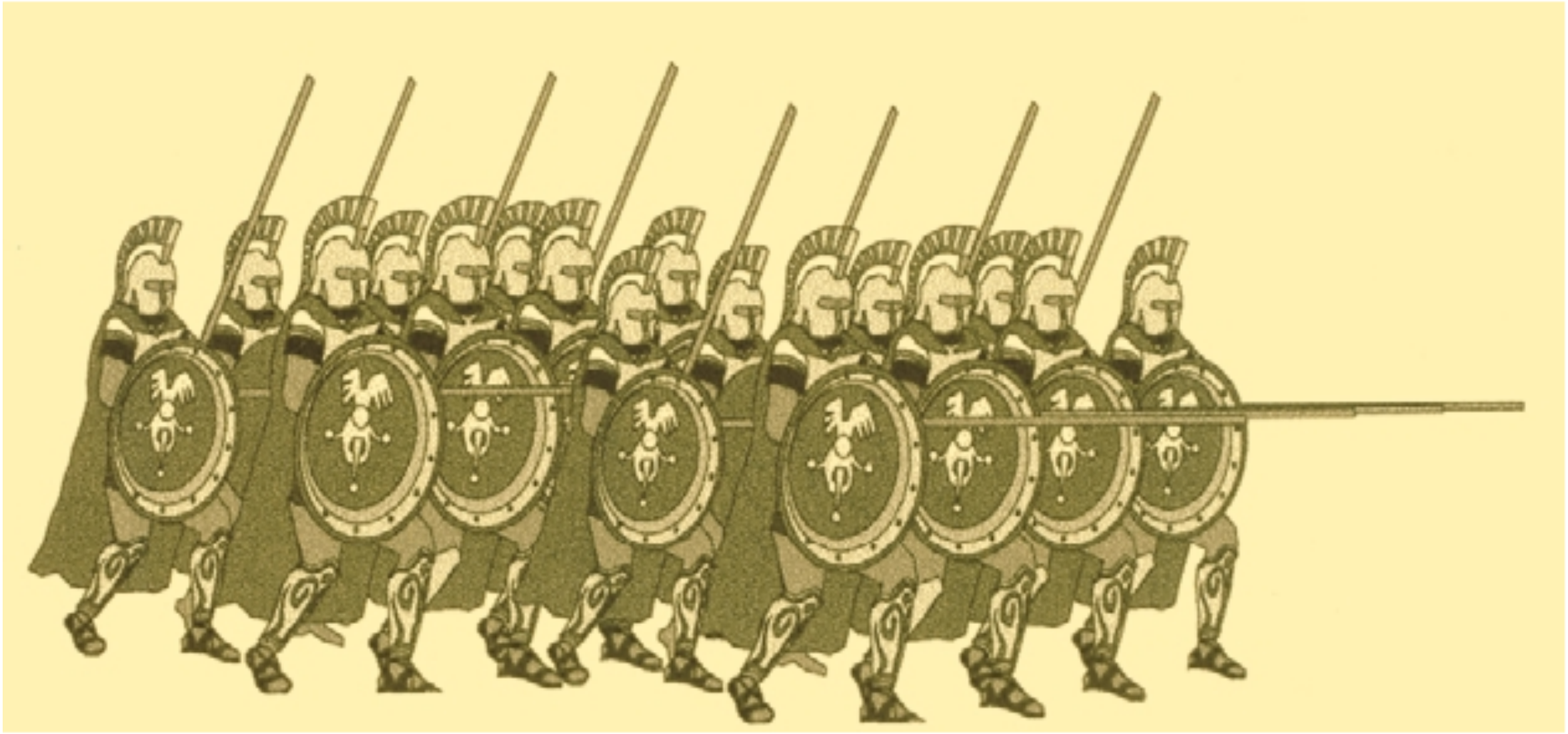
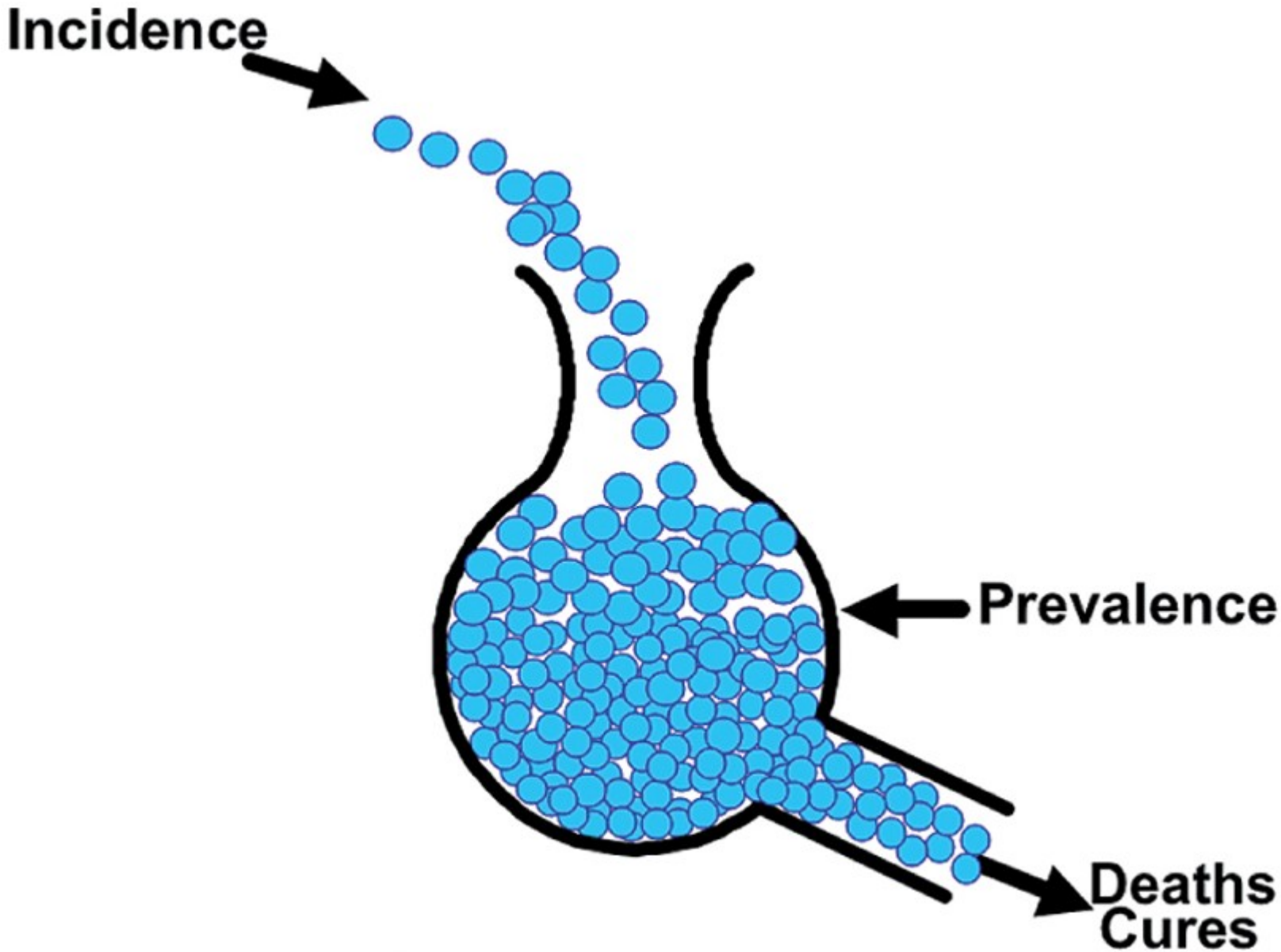


Figure 1: **An early cohort in search of favourable outcomes**


Incidence
Prevalence
Deaths
Cures

$$\lambda(t,x) = \lambda_0 e^{\beta \cdot T}$$

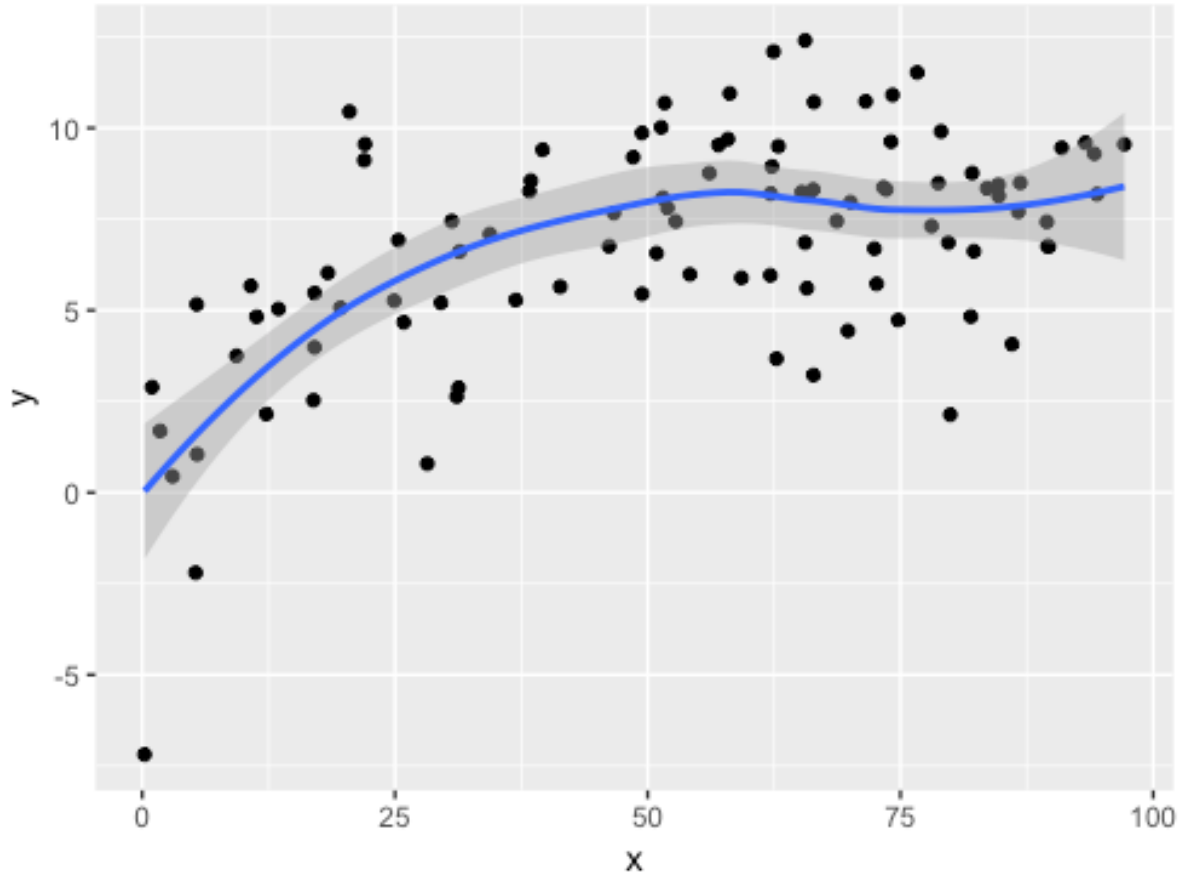
$$\lambda(t,x) = \lambda_0 e^{\beta \cdot T + \sum \beta_j f(x_j)}$$

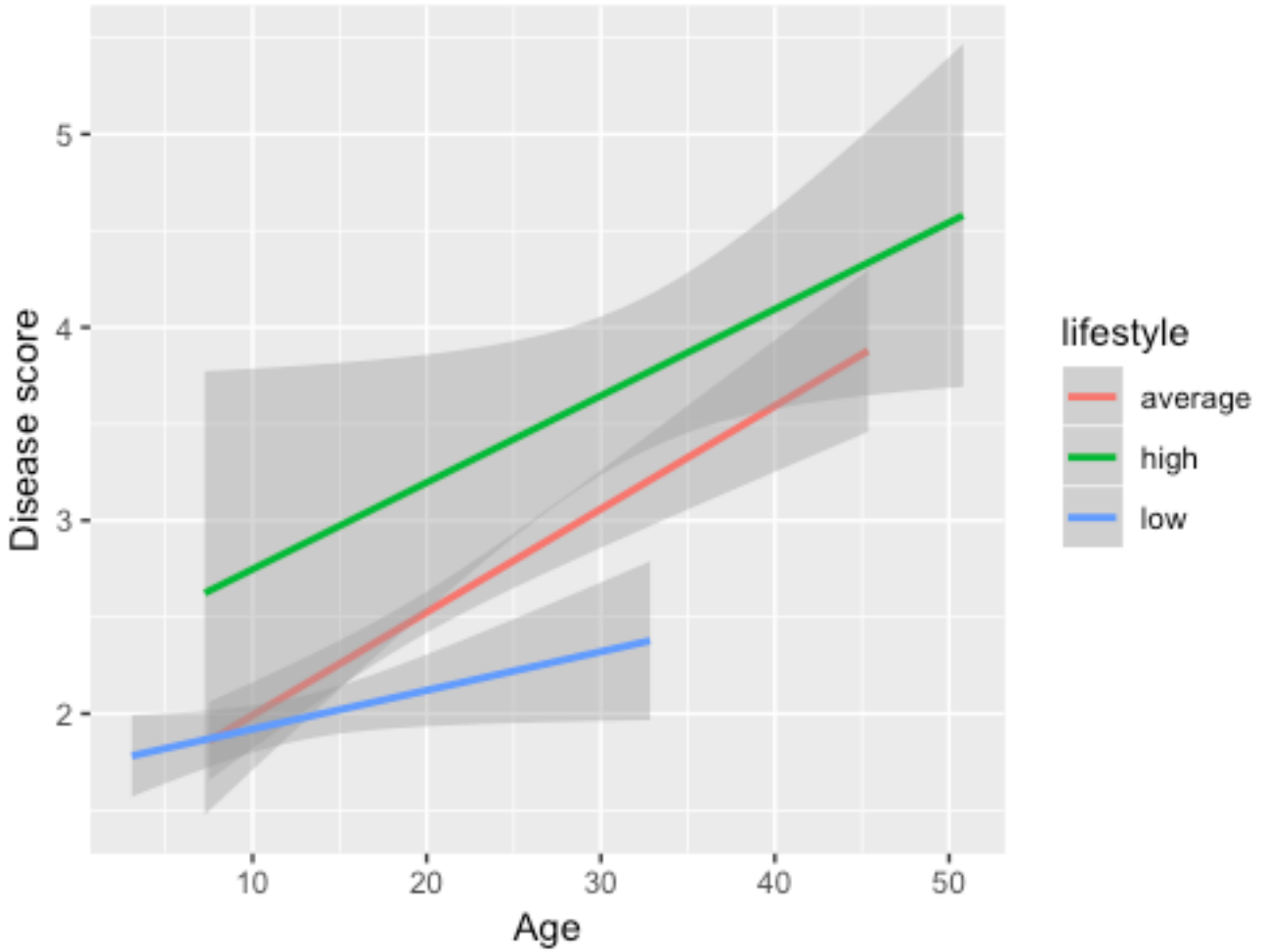
$$\log\left(\frac{\lambda(t,x)}{\lambda_0}\right) = \beta \cdot T + \sum \beta_j f(x_j)$$

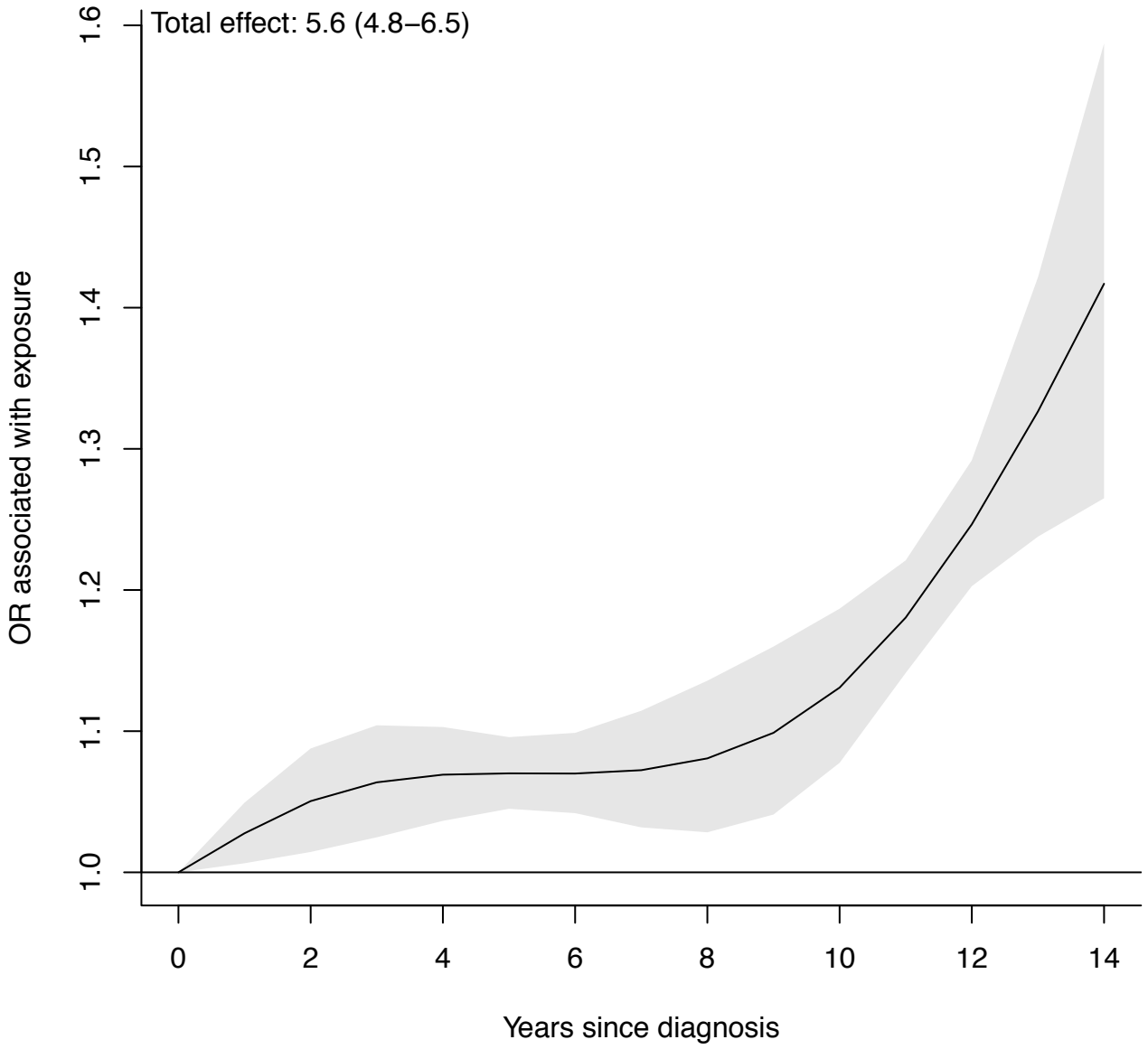
Common for the modelling approaches

- Models for case-control data and for cohort data can easily be extended (and in a similar manner) in order to
  - Include covariates / confounders (the parametrization of confounders depends on the research question / purpose).
  - Include effect modification (interaction terms).
  - Distributed lags
  - Ect
- Stratification* is easy and very useful

*Stratification in the model, not in epidemiological sense


10
5
0
-5
y
0
25
50
75
100
x


5
4
3
2
Disease score
10
20
30
40
50
Age
lifestyle
average
high
low


Total effect: 5.6 (4.8–6.5)
OR associated with exposure
1.0
1.1
1.2
1.3
1.4
1.5
1.6
0
2
4
6
8
10
12
14
Years since diagnosis

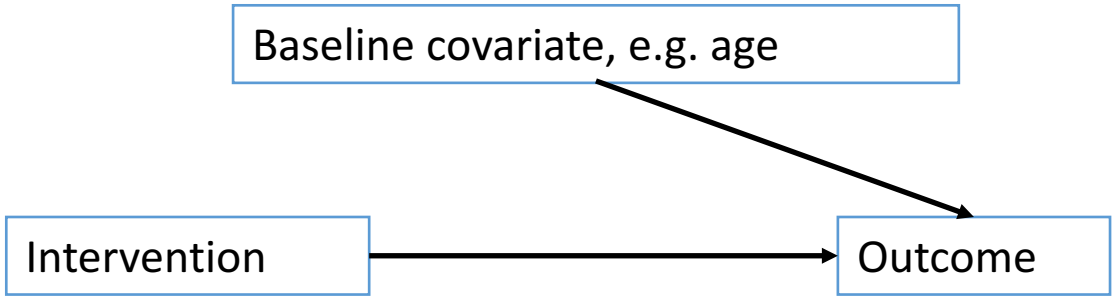
$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta \cdot T + \sum \beta_j f(x_j).$$

$$\log\left(\frac{\lambda(t,x)}{\lambda_0}\right) = \beta \cdot T + \sum \beta_j f(x_j)$$

Pros and cons

Adjustment for baseline covariates

- In epidemiology we have to think a lot about confounders this is different in RCT.
- It can still be a good idea to adjust for strong predictors of the outcome.
  - For continuous outcomes we get increased precision of the estimates.
  - For binary or time-to-event outcomes the point estimates tend to move further away from the null.



Null findings

MENU nature International journal of science

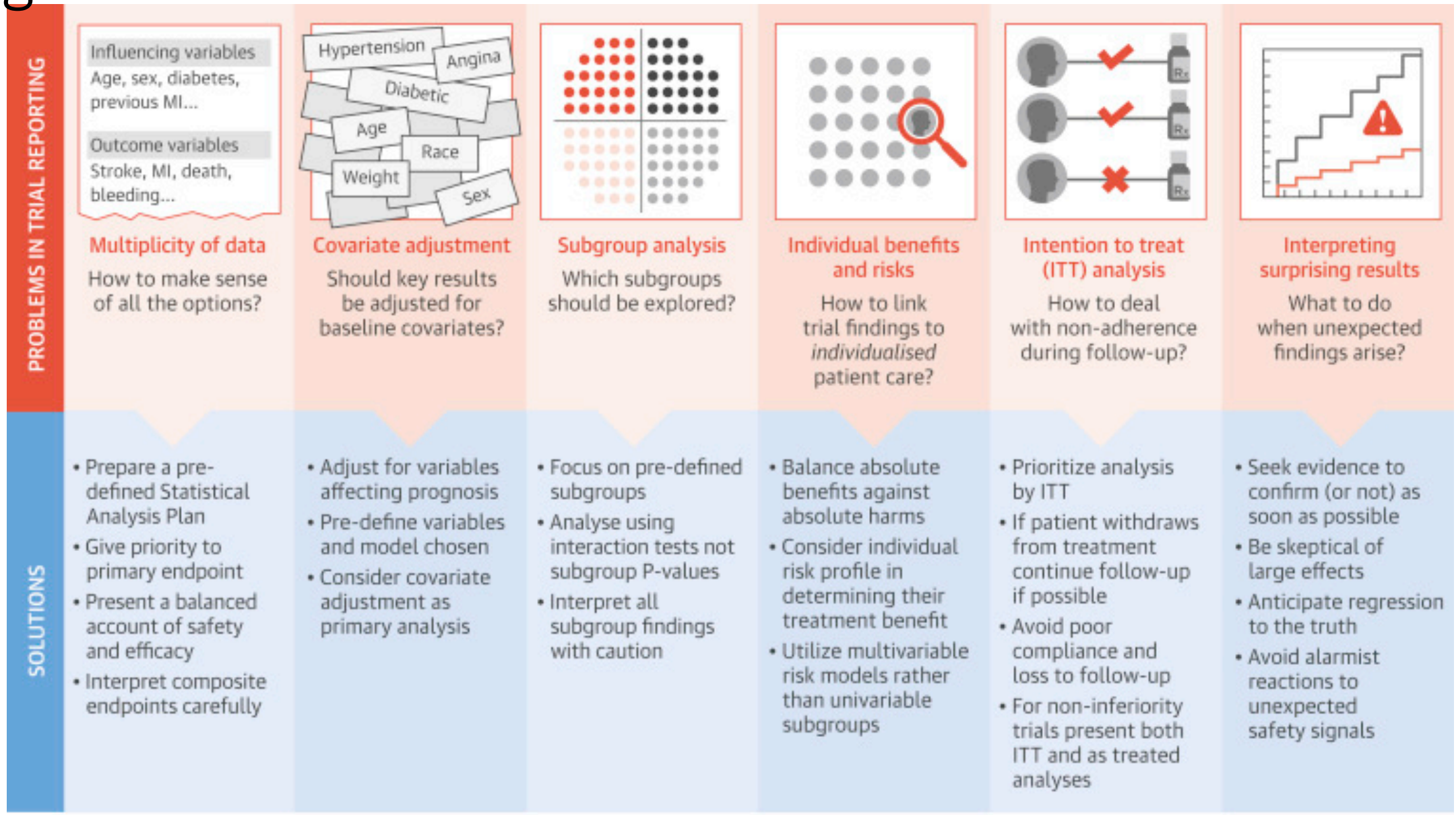
Search E-alert Su

NEWS · 24 OCTOBER 2018

First analysis of 'pre-registered' studies shows sharp rise in null findings

Logging hypotheses and protocols before performing research seems to work as intended: to reduce publication bias for positive results.

Challenges



Pocock, S.J. et al. J Am Coll Cardiol. 2015; 66(23):2648-62.