

Case-kontrol studier og genetiske associationsmodeller

www.biostat.ku.dk/~bxc/SDC-courses

Bendix Carstensen

Steno Diabetes Center &
Biostatistisk afdeling, KU

bxc@steno.dk

www.biostat.ku.dk/~bxc

Claus Thorn Ekstrøm

Inst. f. Matematik og Fysik, KVL &
Steno Diabetes Center

ekstrom@dina.kvl.dk

www.matfys.kvl.dk/~ekstrom

December 2002

Logarithms and exponentials

$$10^2 = 10 \times 10$$

$$10^3 = 10 \times 10 \times 10$$

$$10^2 \times 10^3 = 10^5$$

$$10^3 / 10^2 = 10^1$$

$$(10^3)^2 = 10^6$$

$$10^2 / 10^2 = 10^0 = 1$$

$$10^2 / 10^3 = 10^{-1} = 1/10$$

$$10^{1/2} \times 10^{1/2} = 10^1$$

$$10^{1/2} = \sqrt{10}$$

$$10^{0.3010} = 2$$

$$\log_{10}(2) = 0.3010$$

$$10^{0.4771} = 3$$

$$\log_{10}(3) = 0.4771$$

$$10^1 = 10$$

$$\log_{10}(10) = 1$$

Multiplication and division

$$2 \times 3 = 6$$

$$\log_{10}(2) = 0.3010$$

$$\log_{10}(3) = 0.4771$$

$$0.3010 + 0.4771 = 0.7781$$

$$\log_{10}(6) = 0.7781$$

$$10^{0.3010} \times 10^{0.4771} = 10^{0.7781}$$

$$10^{0.7781} = 6$$

In general:

$$\log(xy) = \log(x) + \log(y)$$

$$\log(x/y) = \log(x) - \log(y)$$

$$\log(x^a) = a \log(x)$$

$$\log(1/x) = -\log(x)$$

Natural logarithms: $e = 2.7183$

$$\log_e(e) = 1$$

$$e^{0.6931} = 2$$

$$\log_e(2) = 0.6931$$

$$e^{1.0986} = 3$$

$$\log_e(3) = 1.0986$$

$$2 \times 3 = 6$$

$$e^{0.6931} \times e^{1.0986} = e^{1.7918}$$

$$e^{1.7918} = 6$$

In general:

$$e^x = \exp(x)$$

$$e^x \times e^y = e^{x+y}$$

$$e^x / e^y = e^{x-y}$$

$$(e^x)^y = e^{x \times y}$$

$$1/e^x = e^{-x}$$

Names for the logarithms

Engineers and calculators:

\log is the logarithm to base 10.

\ln is the logarithm to base e , the natural log

Mathematicians:

\log is the logarithm to base e , the natural log

\log_{10} is the logarithm to base 10.

Why natural logarithms?

For small values of x (relative to 1):

$$\begin{array}{lcl} e^x & \approx & 1 + x \\ e^{-x} & \approx & 1 - x \\ \ln(1 + x) & \approx & x \\ \ln(1 - x) & \approx & -x \end{array} \quad \Rightarrow \quad \begin{array}{l} \ln(1.01) = 0.01 \\ \ln(0.99) = -0.01 \\ \ln(1.04) \approx 0.04 \\ \ln(1.20) = 0.182 \neq 0.20 \end{array}$$

But:

$$\begin{array}{l} \log_{10}(1.01) = 0.4343 \times 0.01 \\ \log_{10}(0.99) = 0.4343 \times -0.01 \\ \\ \log_{10}(x) = 0.4343 \times \ln(x) \end{array}$$

Hypothesis tests in statistical analysis

For two populations the hypothesis of equal means is normally formulated as:

$$H_0 : \delta = 0 \quad \Leftrightarrow \quad \mu_1 = \mu_2$$

Statisticians would consider two models:

$$\begin{array}{ll} 1: & \begin{array}{l} x_{i1} \sim \mathcal{N}(\mu_1, \sigma^2) \\ x_{i2} \sim \mathcal{N}(\mu_2, \sigma^2) \end{array} & 2: & \begin{array}{l} x_{i1} \sim \mathcal{N}(\mu, \sigma^2) \\ x_{i2} \sim \mathcal{N}(\mu, \sigma^2) \end{array} \end{array}$$

H_0 would in this context then be:

Can model 1 be reduced to model 2 ?

Hypothesis testing is comparison of models.

Comparing statistical models

- Can a complicated model be reduced to one describing data in a simpler fashion?

This is the kind of model that one would like to see accepted.

- Can a model be reduced to a model that describes data as not varying with exposure / treatment?

This is the kind of model that one would like to see rejected.

Relevance of $p < 0.05$ depends on context.

Probability

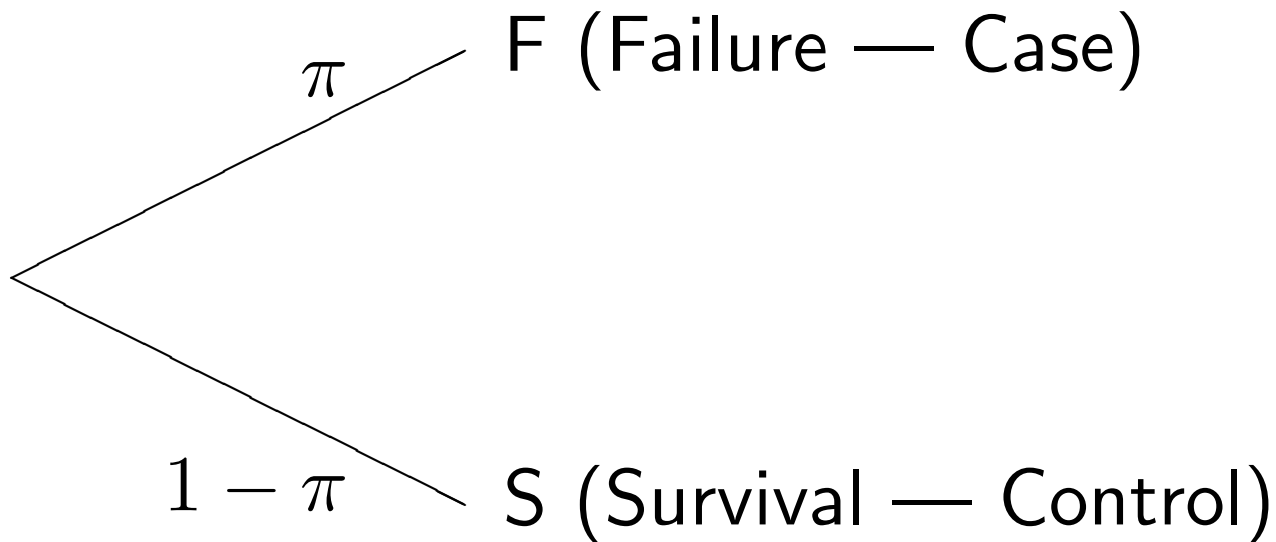
In all scientific studies the outcome is subject to random variation.

In case-control studies and association studies outcomes and exposures are discrete:

- Case / Control
- Genotype: aa / aA / AA

“Measurement”-error described by probabilities for each possible outcome.

The binary probability model

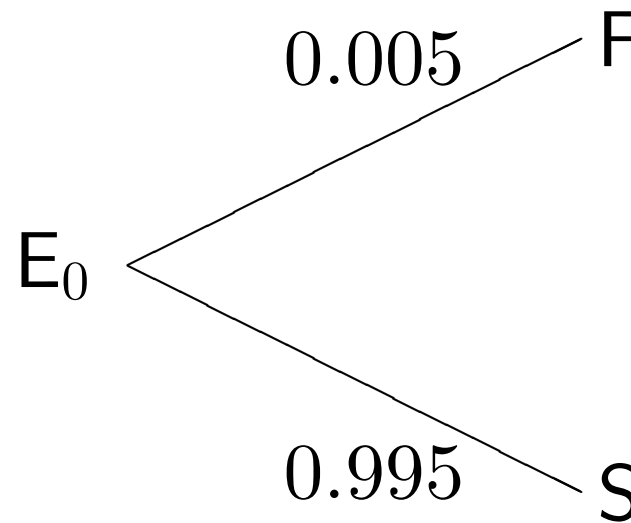
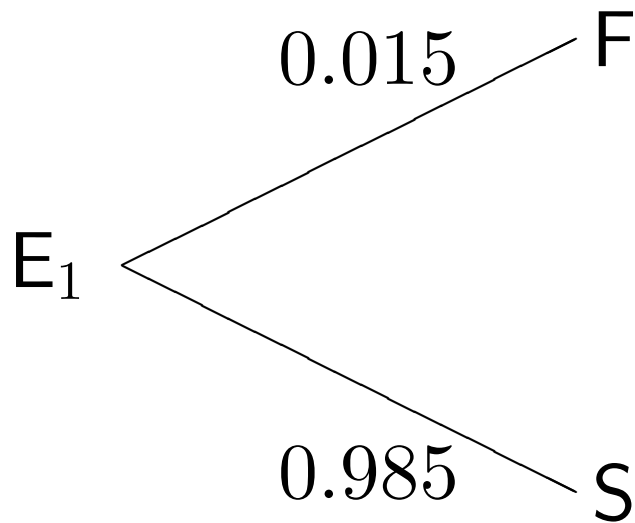


The **risk**
parameter:
 π (pi).

The **odds**
parameter:
 ω (omega).

$$\omega = \frac{\pi}{1 - \pi} \quad \Leftrightarrow \quad \pi = \frac{\omega}{1 + \omega}$$

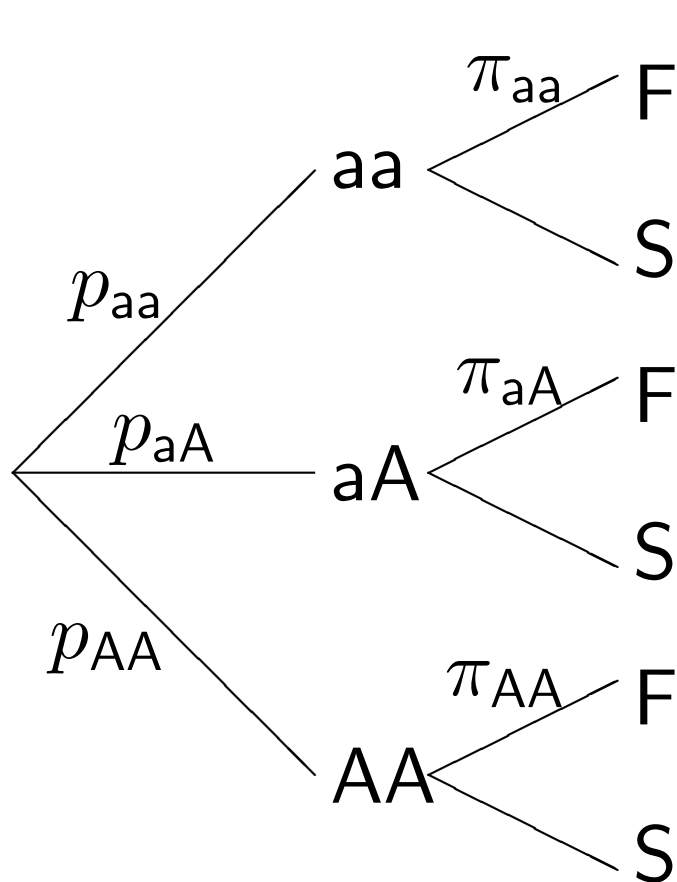
Conditional probabilities of failure



$$P \{F \mid E_1\} = 0.015 \quad P \{F \mid E_0\} = 0.005$$

Risk for exposed individuals is increased by a factor of $0.015/0.005 = 3.0$, relative to unexposed

Conditional probabilities of failure



p_{aa} is the probability that a person has genotype aa .

π_{aa} is the conditional probability of failure **given** genotype aa .

$p_{aa} \times \pi_{aa}$ is the probability that a person has genotype aa **and** fails.

Relationship between follow–up studies and case–control studies

In a **cohort study**, the relationship between exposure and disease incidence is investigated by following the entire cohort and measuring the rate of occurrence of new cases in the different exposure groups.

The follow–up allows the investigator to register those subjects who develop the disease during the study period and to identify those who remain free of the disease.

In a **case-control study** the subjects who develop the disease (the cases) are registered by some other mechanism than follow-up, and a group of healthy subjects (the controls) is used to represent the subjects who do not develop the disease.

Rationale behind case-control studies

- In a follow-up study, rates among exposed and non-exposed are estimated by:

$$\frac{D_1}{Y_1} \quad \frac{D_0}{Y_0}$$

where D are no. events and Y person-years.

The rate ratio is estimated by:

$$\frac{D_1}{Y_1} / \frac{D_0}{Y_0} = \frac{D_1}{D_0} / \frac{Y_1}{Y_0}$$

Necessary to classify both cases and person-years by exposure.

- In a case-control study we use the same cases, but select controls to represent the distribution of risk time between exposed and unexposed:

$$\frac{H_1}{H_0} \approx \frac{Y_1}{Y_0}$$

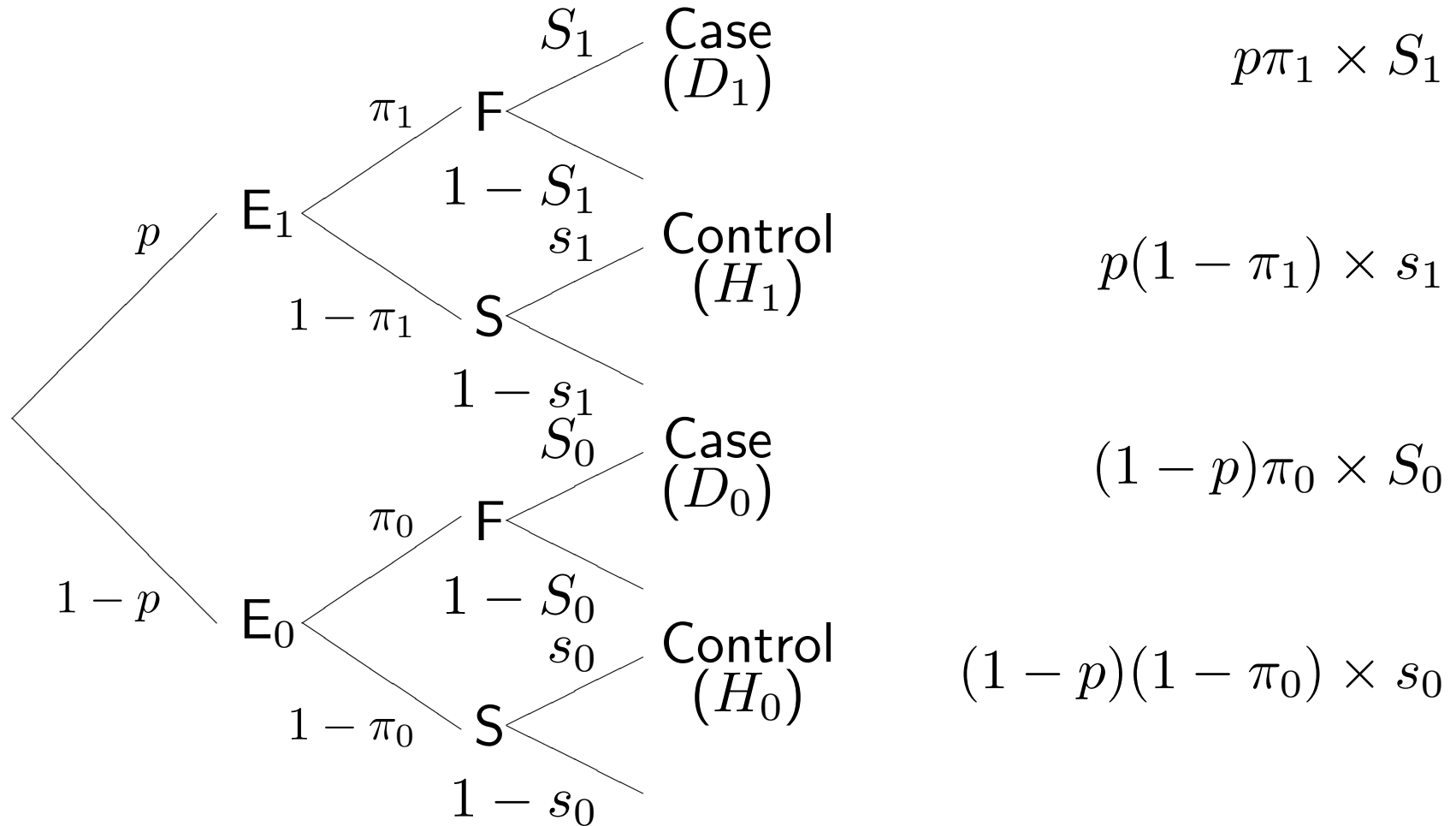
Therefore the rate ratio is estimated by:

$$\frac{D_1}{D_0} / \frac{H_1}{H_0}$$

- Controls represent risk time, **not** disease-free persons.

Case-control probability tree

Exposure Failure Selection Probability



What is estimated by the case-control ratio?

$$\frac{D_1}{H_1} = \frac{S_1}{s_1} \times \frac{\pi_1}{1 - \pi_1} \qquad \frac{D_0}{H_0} = \frac{S_0}{s_0} \times \frac{\pi_0}{1 - \pi_0}$$

$$\text{OR}_{\text{study}} = \frac{D_1/H_1}{D_0/H_0} = \frac{\pi_1/(1 - \pi_1)}{\pi_0/(1 - \pi_0)} = \text{OR}_{\text{population}}$$

but only if $S_1/s_1 = S_0/s_0$, i.e. if sampling fractions are independent of exposure:

$$S_1 = S_0 \quad \text{and} \quad s_1 = s_0$$

S sampling fraction for cases — large

s sampling fraction for controls — small

Estimation from case-control study

Odds-ratio of disease between exposed and unexposed *given inclusion in the study*:

$$\text{OR} = \frac{\omega_1}{\omega_0} = \frac{\pi_1}{1 - \pi_1} \bigg/ \frac{\pi_0}{1 - \pi_0}$$

is the same as the odds-ratio of disease between exposed and unexposed *in the “study base”*, **provided**:

- sampling fraction is the same for exposed and unexposed cases.
- sampling fraction is the same for exposed and unexposed

controls.

— that is the selection mechanism is **only** depending on case/control status.

Log-likelihood for case-control studies

Likelihood: Probability of observed data given the statistical model.

Log-Likelihood (conditional on being included) is a binomial likelihood with odds ω_0 and ω_1

$$D_0 \ln(\omega_0) - N_0 \ln(1 + \omega_0) + D_1 \ln(\theta\omega_0) - N_1 \ln(1 + \theta\omega_0)$$

Odds-ratio (θ) is the ratio of ω_1 to ω_0 , so:

$$\ln(\theta) = \ln(\omega_1) - \ln(\omega_0)$$

Estimates of $\ln(\omega_1)$ and $\ln(\omega_0)$ are:

$$\ln \left(\frac{D_1}{H_1} \right) \quad \text{and} \quad \ln \left(\frac{D_0}{H_0} \right)$$

with standard errors:

$$\sqrt{\frac{1}{D_1} + \frac{1}{H_1}} \quad \text{and} \quad \sqrt{\frac{1}{D_0} + \frac{1}{H_0}}$$

Exposed and unexposed form two independent bodies of data, so the estimate of $\ln(\theta)$ [= $\ln(\text{OR})$] is

$$\ln \left(\frac{D_1}{H_1} \right) - \ln \left(\frac{D_0}{H_0} \right), \quad \text{s.e.} = \sqrt{\frac{1}{D_1} + \frac{1}{H_1} + \frac{1}{D_0} + \frac{1}{H_0}}$$

Computing c.i. for odds-ratios

$$\hat{\text{OR}} = \frac{D_1/H_1}{D_0/H_0} \quad \text{s.e.}[\ln(\text{OR})] = \sqrt{\frac{1}{D_1} + \frac{1}{H_1} + \frac{1}{D_0} + \frac{1}{H_0}}$$

95% c.i. for $\ln(\text{OR})$:

$$\ln(\text{OR}) \pm 1.96 \times \text{s.e.}[\ln(\text{OR})]$$

95% c.i. for OR by taking the exponential:

$$\text{OR} \times \underbrace{\exp(1.96 \times \text{s.e.}[\ln(\text{OR})])}_{\text{error factor}}$$

Kir 6.2 homozygotes and diabetes

Genotype	Diabetes cases	Population controls
KK	134	124
EE/EK	669	738

What is the odds-ratio of diabetes associated with being homozygous for the K-allele?

This compares KK genotypic persons with EE and EK seen as one group.

How precisely is this odds-ratio determined?

$$\text{OR} = \frac{D_1/H_1}{D_0/H_0} = \frac{134/124}{669/738} = \frac{1.081}{0.907} = 1.192 = 1.19$$

$$\begin{aligned} \text{s.e.}(\ln[\text{OR}]) &= \sqrt{\frac{1}{D_1} + \frac{1}{H_1} + \frac{1}{D_0} + \frac{1}{H_0}} \\ &= \sqrt{\frac{1}{134} + \frac{1}{124} + \frac{1}{669} + \frac{1}{738}} = 0.136 \end{aligned}$$

The 95% limits for the odds-ratio are:

$$\text{OR} \times \exp(1.96 \times 0.136) = 1.192 \times 1.304 = (0.91 - 1.55)$$

K-carriers and diabetes: your turn!

Genotype	Diabetes cases	Population controls
EK/KK	516	532
EE	287	330

What is the odds-ratio of diabetes associated with being a carrier for the K-allele?

This compares KK/EK persons with EE persons.

How precisely is this odds-ratio determined — give a 95% c.i.

Solution to exercise

$$\text{OR} = \frac{D_1/H_1}{D_0/H_0} = \frac{516/532}{287/330} = \frac{0.970}{0.870} = 1.115$$

$$\text{s.e.}(\ln[\text{OR}]) = \sqrt{\frac{1}{516} + \frac{1}{532} + \frac{1}{287} + \frac{1}{330}} = 0.102$$

The 95% limits for the odds-ratio are:

$$\text{OR} \times \exp(1.96 \times 0.102) = 1.115 \times 1.22 = (0.91 - 1.22)$$

More levels of exposure — genotypes

Genotype	Diabetes cases	Population controls	case/ control odds	OR relative to (0)
EE (0)	287	330	0.870	1.000
EK (1)	382	408	0.936	1.077
KK (2)	134	124	1.081	1.243

The **relationship** of case-control ratios is what matters.

Odds-ratio of diabetes for EK vs. EE is 1.08

Odds-ratio of diabetes for KK vs. EE is 1.24

Odds-ratio of diabetes for KK vs. EK is

Odds-ratio and rate ratio (RR)

- If the disease probability, π , in the study period is small:

$$\pi = \text{cumulative risk} \approx \text{cumulative rate} = \lambda T$$

with λ the rate and T the study period.

- For small π , $1 - \pi \approx 1$, so:

$$\text{OR} = \frac{\pi_1 / (1 - \pi_1)}{\pi_0 / (1 - \pi_0)} \approx \frac{\pi_1}{\pi_0} \approx \frac{\lambda_1}{\lambda_0} = \text{RR}$$

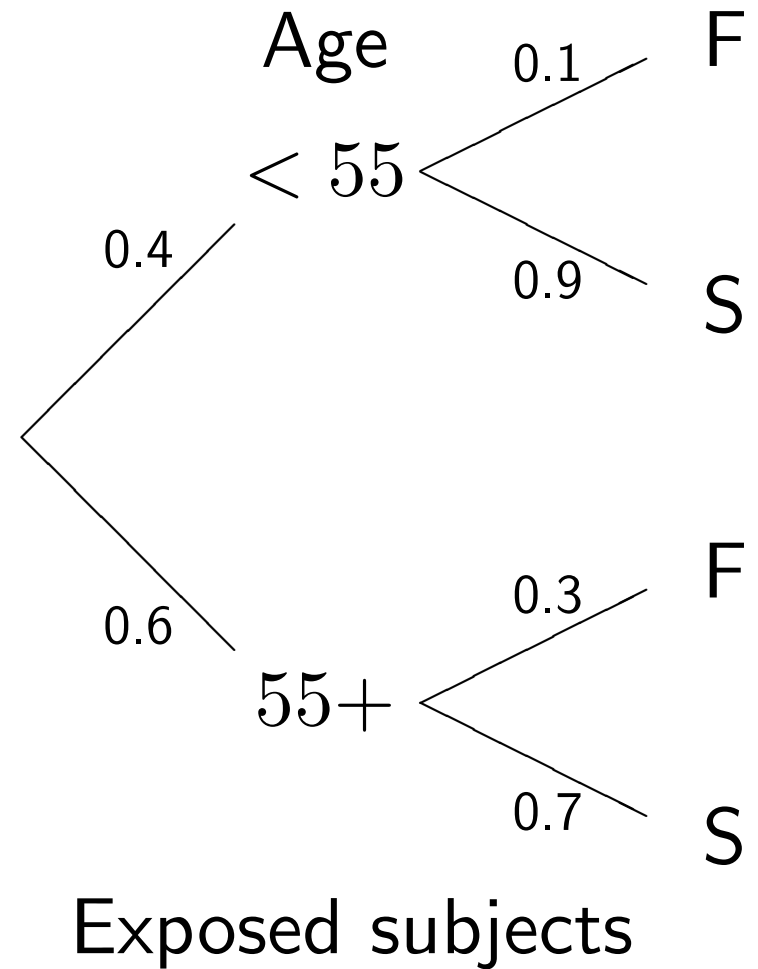
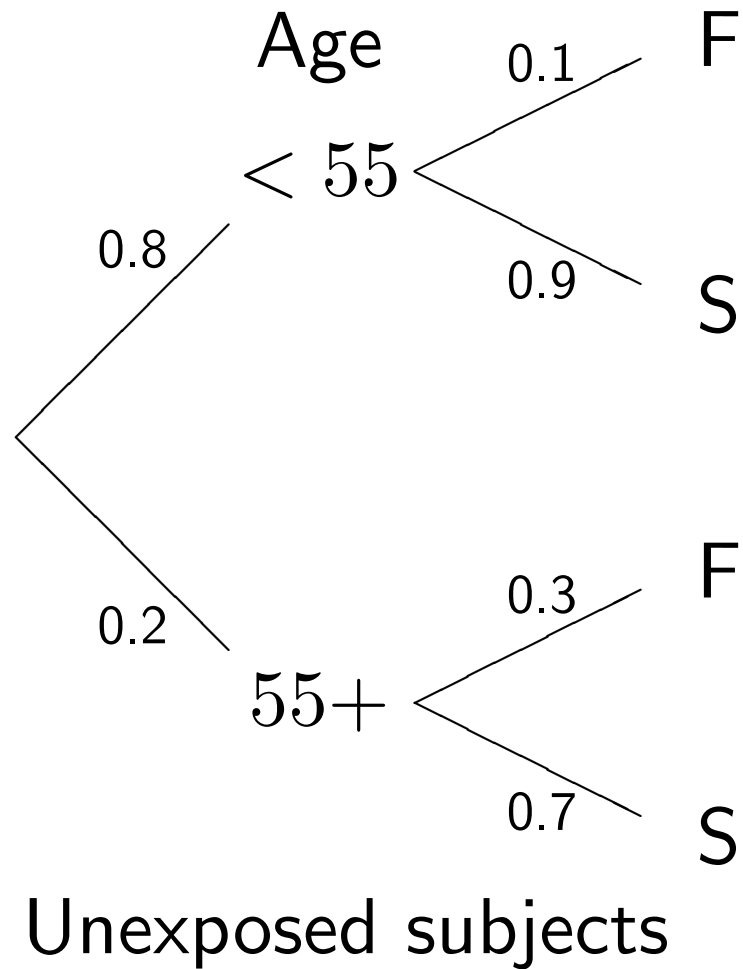
π small \Rightarrow OR estimate of RR.

Confounding

- Epidemiology relies on *observational studies* of *experiments of nature*
- Often these are poor experiments
 - no control for *confounding* by extraneous influences
- Definition:

A confounder is a variable whose influence we would have controlled if we had been able to design the natural experiment.

Example: confounding by age



- Probability of failure for **unexposed**:

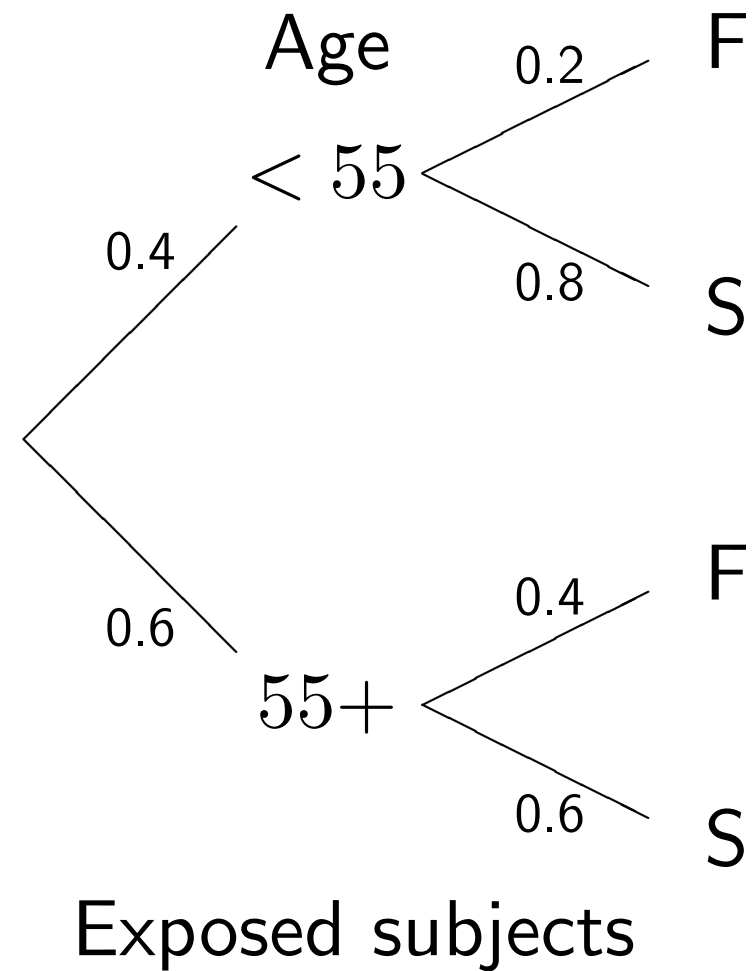
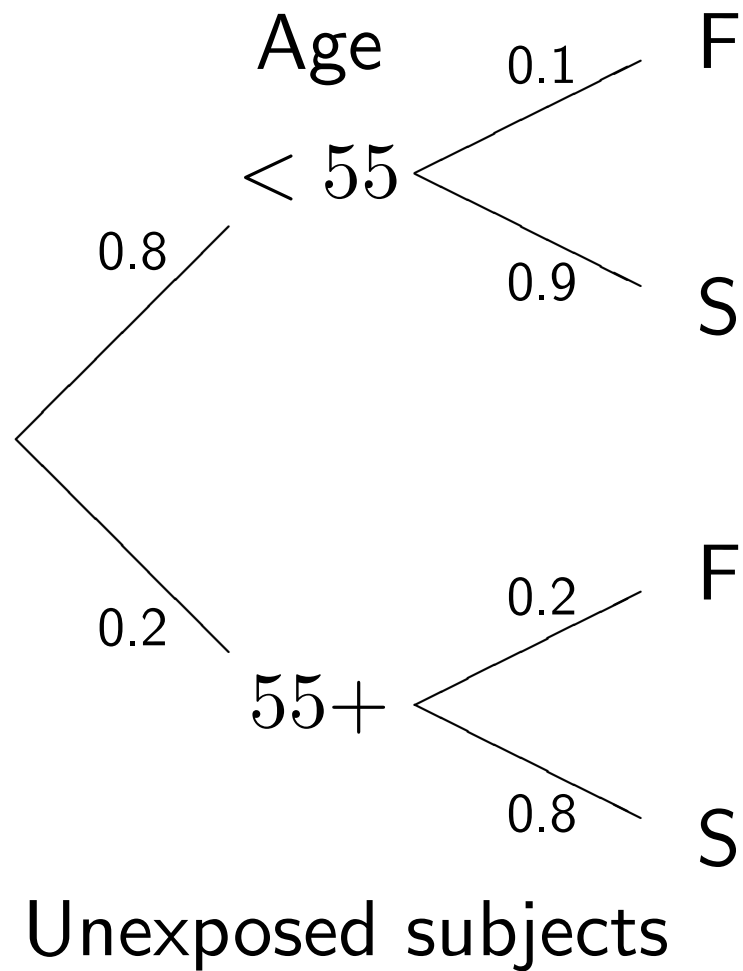
$$(0.8 \times 0.1) + (0.2 \times 0.3) = 0.14$$

- Probability of failure for **exposed**:

$$(0.4 \times 0.1) + (0.6 \times 0.3) = 0.22$$

- Difference entirely due to difference in age structure.
- When there is a true effect, its magnitude can be distorted by such influences.

Confounding when $RR = 2$



- The true relative risk, $RR_T = 0.2/0.1 = 0.4/0.2 = 2$

- Probability of failure for **unexposed**:

$$(\quad \times \quad) + (\quad \times \quad) =$$

- Probability of failure for **exposed**:

$$(\quad \times \quad) + (\quad \times \quad) =$$

- The apparent relative risk:

$$RR_O =$$

- The true relative risk, $RR_T = 0.2/0.1 = 0.4/0.2 = 2$

- Probability of failure for **unexposed**:

$$(0.8 \times 0.1) + (0.2 \times 0.2) = 0.12$$

- Probability of failure for **exposed**:

$$(0.4 \times 0.2) + (0.6 \times 0.4) = 0.32$$

- The apparent relative risk:

$$RR_O = 0.32/0.12 = 2.67$$

Confounding

A confounder is:

- Associated with outcome:
The older persons have higher disease probability.
- Associated with the exposure:
The older persons are more / less likely to be exposed.
- Is not a result of either exposure or disease.
Not a statistical property. Cannot be seen from tables.
- Common sense is required!

Controlling confounding

In **controlled experiments** there are two ways of controlling confounding:

1. **Randomization** of subjects to experimental groups so that the *distributions* of the confounder are the same.
2. Hold the confounder **constant**.

Standardization is a statistical technique for controlling for extraneous variables in the analysis of an observational study:

- **Direct** standardization simulates randomization by equalising the distribution of extraneous variables.
- **Indirect** standardization simulates the second method: holding extraneous variables constant.

The latter is the preferred technique.
It leads to proper statistical modelling.

Indirect standardization

- Aim is to hold age (the confounder) constant.
- Compare exposed and unexposed *within age strata*
- But this leads to *several* experiments, each one rather small, hence imprecise.
- Calculate a single *combined* estimate of the exposure effect over all strata.
- This procedure implies a **model** in which there is no (systematic) variation of effect over strata.

Meta-analysis

If several case-control studies are conducted in different populations, they cannot be regarded as one because:

- Study may be associated with exposure — in this case genotype distribution.
- Study may be associated with outcome — in this case occurrence of diabetes.

Thus study population should be regarded as a confounder.

Model for confounder control

Assumption of similar effect across studies in different populations: $OR_p = \theta$ independent of p , so for odds of disease ω_{p1} :

$$\omega_{p1} = \theta\omega_{p0}$$

Odds of disease increase by the same amount, θ , by exposure, regardless of study.

But the disease odds among unexposed, ω_{p0} , may vary between studies.

On the log-scale:

$$\ln \left(\frac{\pi_{p1}}{1 - \pi_{p1}} \right) = \ln(\omega_{p1}) = \ln(\theta) + \ln(\omega_{p0})$$

Model for case-control studies

Case-control studies has different sampling fractions for cases (S , large) and controls (s , small):

$$\begin{aligned}\ln[\text{odds}(\text{case} \mid \text{incl.})] &= \ln \left[\frac{\pi_{p1}}{1 - \pi_{p1}} \times \frac{S}{s} \right] \\ &= \ln \left[\frac{\pi_{p1}}{1 - \pi_{p1}} \right] + \ln \left[\frac{S}{s} \right] \\ &= \underbrace{\ln(\theta) + \ln(\omega_{p1})}_{\text{intercept, population}} + \ln \left[\frac{S_p}{s_p} \right]\end{aligned}$$

Logistic regression model with effects of exposure and study population. Estimates for effect of population is irrelevant, since sampling fractions most likely depends on population.

But population **must** be in the model.

Meta-analysis

Analysis with study population as controlling variable.

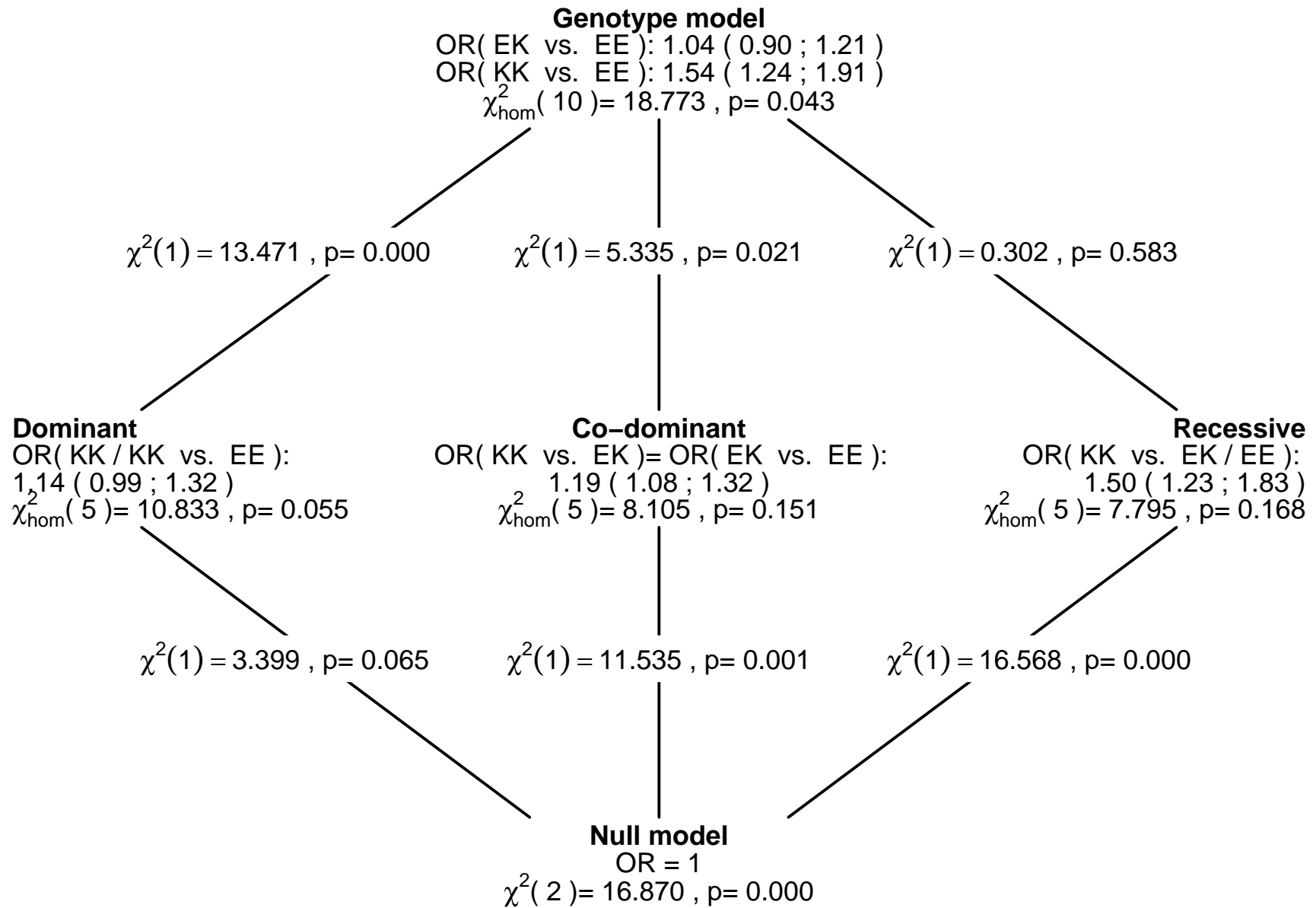
Two things to consider:

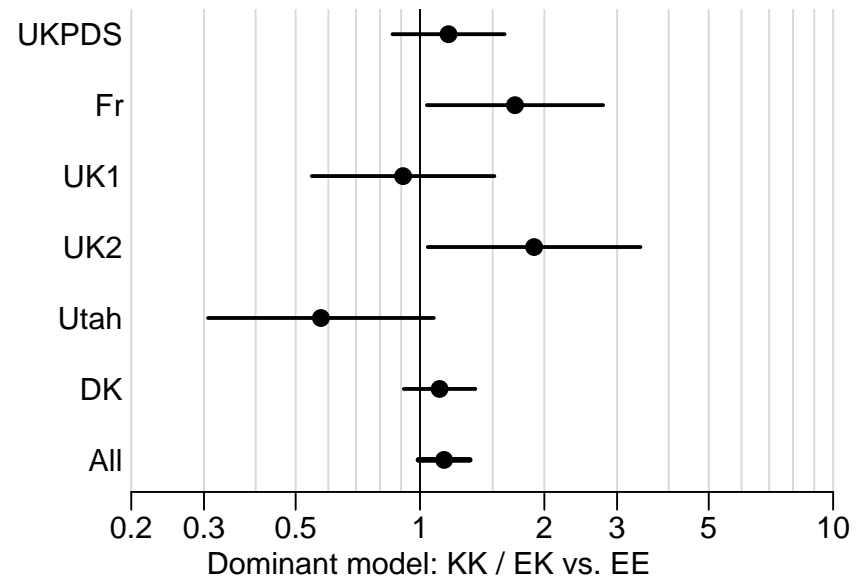
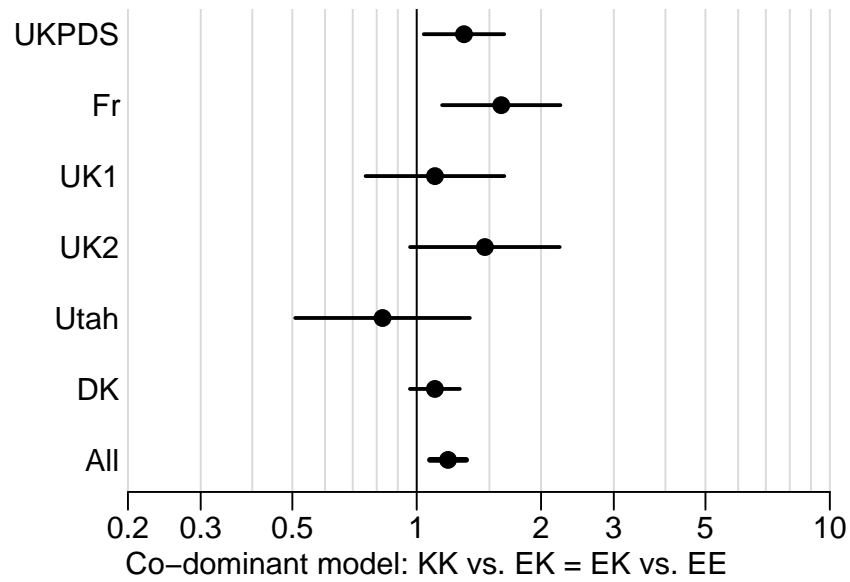
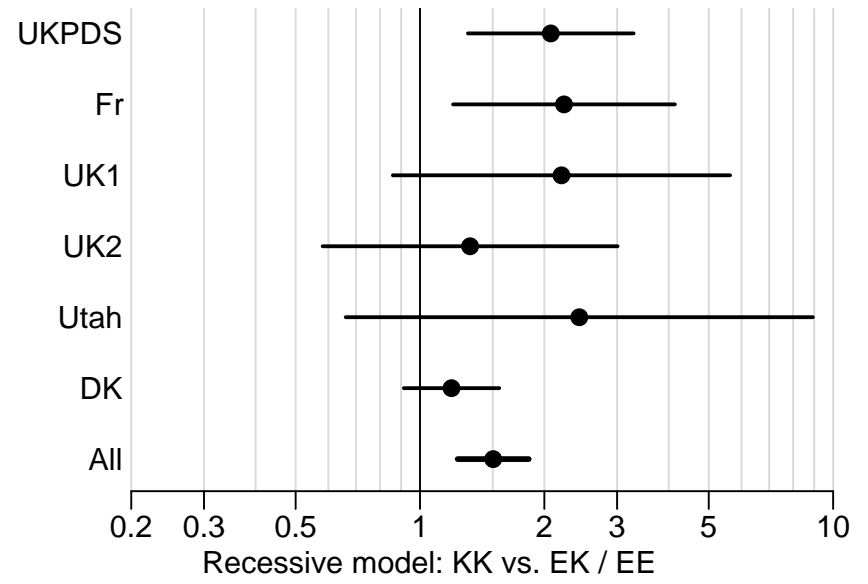
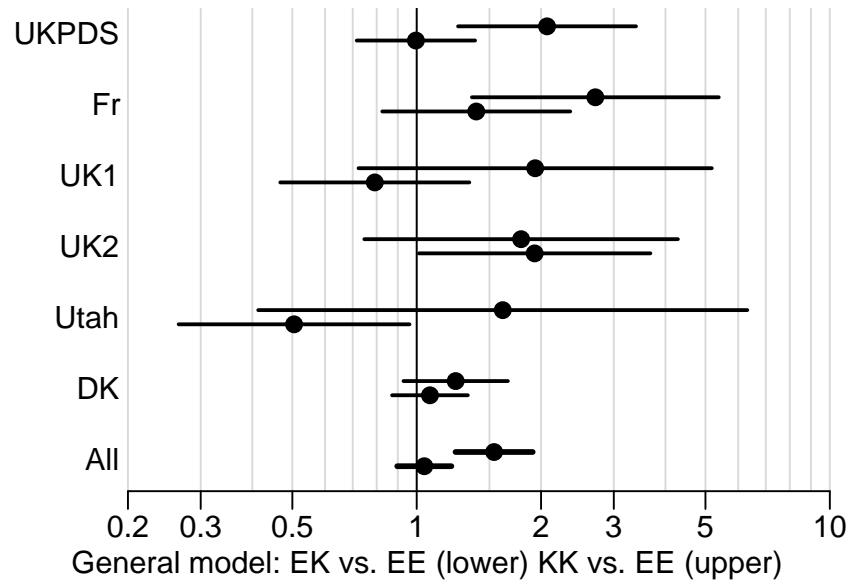
- How is the genotype effect: dominant, co-dominant or recessive?

Similar to the analysis for one population. But in stratified model.

- Is the effect same across populations?

Test for homogeneity of effect (interaction)





Do the studies actually show the same?

Apart from the visual inspection of the diagram, formal tests for the models separately may be of interest.

If these χ^2 -statistics are added up, they will form a χ^2 -statistic which simultaneously tests for the given model, but allowing for different effect sizes.

Formally a hypothesis, but hard to attach any simple biological meaning to.

Test for models, single studies

χ^2	Model			d.f.
	Dominant	Co-dominant	Recessive	
UKPDS	9.133	4.759	0.001	1
Fr	4.116	0.451	1.543	1
UK1	3.655	3.521	0.758	1
UK2	0.027	1.234	4.051	1
Utah	3.480	5.923	4.456	1
DK	0.999	0.115	0.470	1
All	21.411	16.003	11.280	6

Test for models, single studies

p-values	Model		
	Dominant	Co-dominant	Recessive
UKPDS	0.003	0.029	0.979
Fr	0.042	0.502	0.214
UK1	0.056	0.061	0.384
UK2	0.869	0.267	0.044
Utah	0.062	0.015	0.035
DK	0.317	0.735	0.493
All	0.002	0.014	0.080

Two different kinds of tests for Dominant / Co-dominant / Recessive:

- Test in stratified model, assuming **same** effect in all populations.

This is the test shown in the diagram.

- Test in separate models added up.

Tests for mode of action, allowing for separate effects between populations.

This is the test in the last line of the table.