Summary statistics for   binary data

Tabriz University of Medical Sciences
Standard Workshop on Systematic Reviews _ October 2012

Dr. Shayesteh Jahanfar,
University of British Columbia

Outline

- identify binary outcomes
- be familiar with ways of expressing chance of an event when using binary outcomes
- understand how to express and interpret the relative chance of an event when comparing groups
- select effect measure

What is a binary outcome?

- e.g. dead or alive, pain free or in pain, smoking or not smoking
- each participant is in one of two possible, mutually exclusive, states

What were the chances of that?

- risks and odds are just ways of expressing chance in numbers
- for binary events, they just express the chance of being in one of the two states

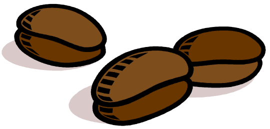
Risk

- 24 people drank an espresso, and 6 fell asleep
- risk of falling asleep

= 6 asleep/24 who could have fallen asleep

= 6/24 = ¼ = 0.25 = 25%

$$\text{risk} = \frac{\text{number of events of interest}}{\text{total number of observations}}$$



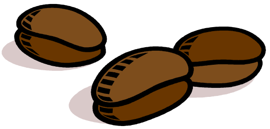
Odds

- 24 people drank an espresso, and 6 fell asleep
- odds of falling asleep

  = 6 asleep/18 did not fall asleep

  = 6/18 = 1/3 = 0.33 (not usually expressed as %)

$$\text{odds} = \frac{\text{number of events of interest}}{\text{number without the event}}$$



Expressing it in words

- risk
  - the chances of falling asleep were one in four, or 25%
- odds
  - the chances of falling asleep were one third of the chances of staying awake
  - one person fell asleep for every three that stayed awake
  - the chances of falling asleep were 3 to 1 against

Do risks and odds differ much?

2 examples from caffeine trials

- 130 people ‘still awake’ out of 164
- chance of still being awake
  - **risk = 130/164 = 0.79;** **odds = 130/34 = 3.82**

- 4 people with ‘headaches’ out of 63
- chance of having a headache
  - **risk = 4/63 = 0.063;** **odds = 4/59 = 0.068**

Comparing groups – 2x2 table

	Asleep	Awake	Total (by group)
Caffeine	12	48	60
Decaf	16	33	49
Total (by event)	28	81	109

- to express the relative chance of an event

Meta-analysis of binary data

- calculate a single summary statistic to represent the effect found in each study
- 3 options
  - risk ratio (relative risk)
  - odds ratio
  - risk difference

Risk ratio

- risk of event on treatment

  = **12/60**

- risk of event on control

  = **16/49**

	Asleep	Awake	Total
Caffeine	**12**	48	**60**
Decaf	**16**	33	**49**
Total	28	81	109

- risk ratio $= \dfrac{\text{risk on treatment}}{\text{risk on control}}$

$$= \frac{\mathbf{12/60}}{\mathbf{16/49}} = \frac{0.2}{0.327} = 0.61$$

Where risk ratio = 1, this implies no difference in effect

Expressing risk ratios in words

- risk ratio 0.61
  - the risk of falling asleep on treatment (caffeine) was about 61% of the risk on placebo (decaf)
  - caffeine reduced the risk to about 60% of what it was
  - the risk of falling asleep on caffeine is 39% lower compared to decaf
  - caffeine reduced the risk by 39%

Odds ratio

- odds of event on treatment

  = **12/48**
- odds of event on control

  = **16/33**

	Asleep	Awake	Total
Caffeine	**12**	**48**	60
Decaf	**16**	**33**	49
Total	28	81	109

- odds ratio

$$\text{odds ratio} = \frac{\text{odds on treatment}}{\text{odds on control}}$$

$$= \frac{\mathbf{12/48}}{\mathbf{16/33}} = \frac{0.25}{0.485} = 0.52$$

Where odds ratio = 1, this implies no difference in effect

Expressing odds ratios in words

- odds ratio 0.52
  - caffeine reduced the odds of falling asleep to 52% of what they were
  - the odds of falling asleep on caffeine is 48% lower compared to decaf
  - caffeine reduced the odds by 48%

Risk difference

- risk of event on treatment

  = **12/60**

- risk of event on control

  = **16/49**

	Asleep	Awake	Total
Caffeine	**12**	48	**60**
Decaf	**16**	33	**49**
Total	28	81	109

- risk difference = risk on control - risk on treatment

  = 16/49 - 12/60 = 0.327 - 0.2 = 0.127

- usually expressed as a %, 13%

Expressing risk difference in words

- risk difference 13%
  - caffeine reduced the risk of falling asleep by about 13 percentage points

Number needed to treat

- this is often expressed as how many we expect to treat, on average, before one extra person is helped
- NNT = 1/RD
- e.g. = 1/0.127 = 8 (*round up* to whole people)
- we would need to give 8 people caffeine to keep one extra person from falling asleep
- not used directly for meta-analysis as there is no useful variance formula

Choosing the effect measure

Criteria to consider when selecting a summary statistic

1. communication of effect
2. consistency of effect across studies
3. mathematical properties

Summary

	OR	RR	RD
Communication	-	+	++
Consistency	+	+	_
Mathematics	++	_	_

Take home message

- risks and odds are just ways of expressing chance
- risk ratios and odds ratios are ways of comparing chances in more than one setting/group
- RR and OR differ when the event is common

Take home message

- risk difference shows the amount of change from baseline in absolute terms
- NNT communicates how many people would need to be treated for one extra to be helped
- ALL these estimates of treatment effect are uncertain, and should be presented with a confidence interval

Summary statistics for continuous data

Outline

- identify continuous outcomes
- understand how to summarise continuous data and pool studies with:
  - measures on the same scale
  - measures on different scales
- recognise some of the challenges of continuous data

Types of data

- Binary data
- Counts of infrequent events (e.g. number of strokes)
- Short ordinal scales (e.g. pain grades: none/mild/moderate/severe)
- Long ordinal scales (e.g. disability scales)
- Continuous data (e.g. blood pressure)
- Censored data (e.g. survival times)

What are continuous data?

- data with an infinite number of values that are equally spaced
- example: height - it can be measured along a numerical continuum of centimetres, metres or inches, feet
  - a person can be 175.24678cm tall, assuming the measurement instrument is accurate enough
  - the difference between 160 and 161cm, and 180 and 181cm, is the same

Long ordinal scales

- sometimes treated as continuous data
- but not true continuous because
  - they have a finite number of distinct values
  - there are gaps in the continuum
- have multiple, ordered categories which imply magnitude
  - e.g. one category is greater or lesser than another
- spacing between categories is not numerically equivalent
- approach ‘continuous’ with increasing categories

What continuous data can we combine?

- data represent continuous measures
- the mean value is in the middle (distribution is roughly symmetrical)
- measurements are made on all participants (not censored or survival type data)
- data are available for both groups in each trial

What data is needed?

	Mean	SD	Sample size
Treatment	m_t	sd_t	n_t
Control	m_c	sd_c	n_c

Meta-analysis of continuous data

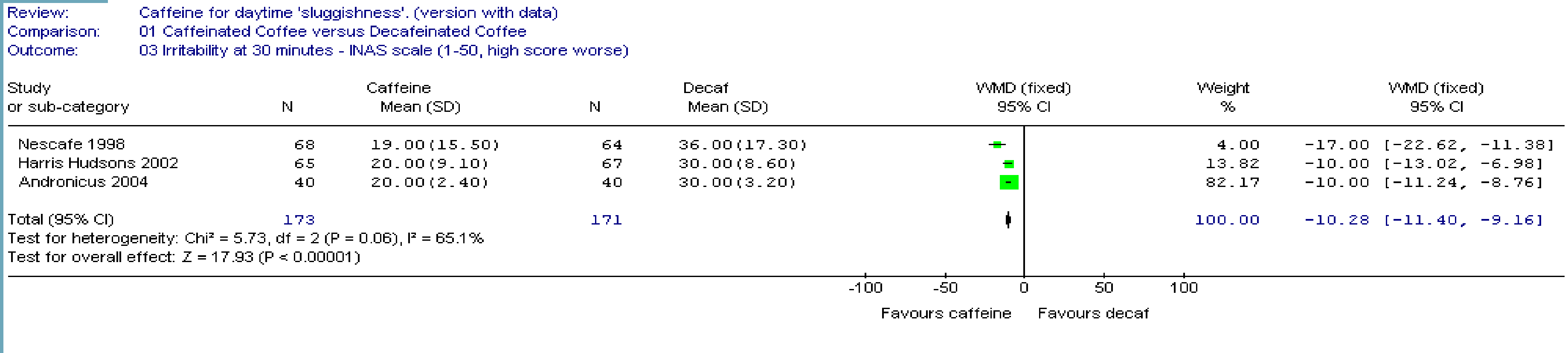
- calculate a single summary statistic to represent the effect found in each study
- Summary statistics combined in meta-analysis
- 2 options
  - mean difference
  - standardised mean difference

Mean difference

- outcomes measured in same unit using same scale (e.g. blood pressure as mmHg)
- pooled analysis in “natural units” and therefore easy to interpret
- studies weighted according to the inverse of the variance (a function of size and SD)

MD = mean on treatment – mean on control

Mean difference: example



Review: Caffeine for daytime 'sluggishness'. (version with data)
Comparison: 01 Caffeinated Coffee versus Decafeinated Coffee
Outcome: 03 Irritability at 30 minutes - INAS scale (1-50, high score worse)

Study or sub-category	N	Caffeine Mean (SD)	N	Decaf Mean (SD)	WMD (fixed) 95% CI	Weight %	WMD (fixed) 95% CI
Nescafe 1998	68	19.00(15.50)	64	36.00(17.30)		4.00	-17.00 [-22.62, -11.38]
Harris Hudsons 2002	65	20.00(9.10)	67	30.00(8.60)		13.82	-10.00 [-13.02, -6.98]
Andronicus 2004	40	20.00(2.40)	40	30.00(3.20)		82.17	-10.00 [-11.24, -8.76]
Total (95% CI)	173		171			100.00	-10.28 [-11.40, -9.16]

Test for heterogeneity: Chi² = 5.73, df = 2 (P = 0.06), I² = 65.1%
Test for overall effect: Z = 17.93 (P < 0.00001)

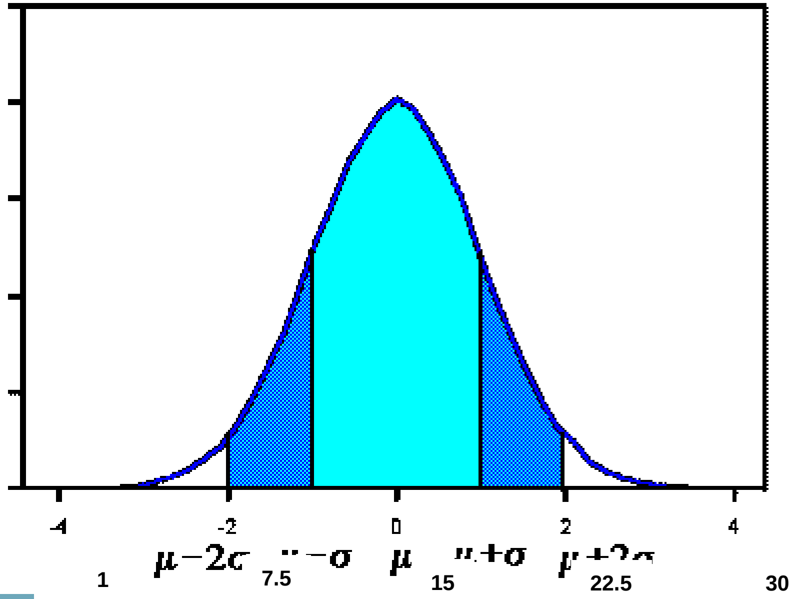
Standardised mean difference

- Outcome is same concept measured on different scales, the values must be transformed to a common scale before pooling
- Sometimes scale factors are known and transformations are made directly (e.g weight)
- Standardised mean difference calculated as:

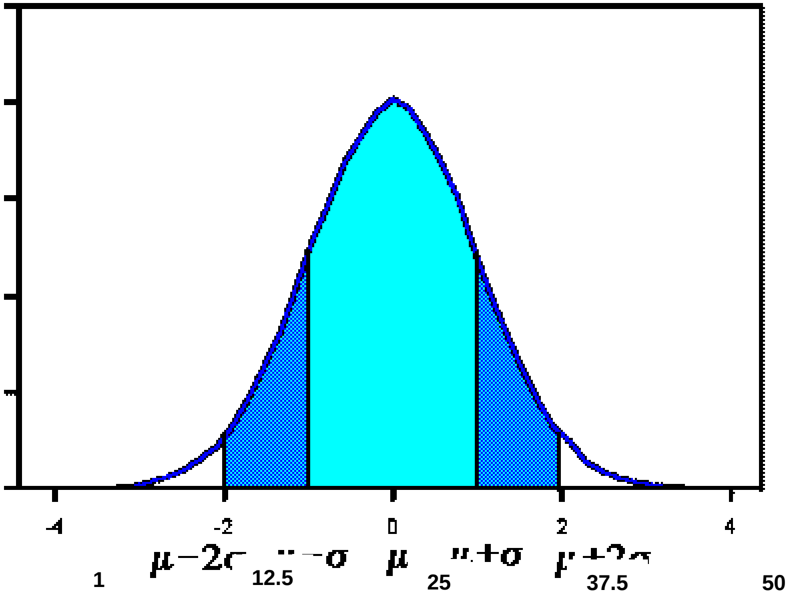
$$\frac{\text{Difference in means between groups}}{\text{Average standard deviation}}$$

Standardised mean difference

Beck Irritability Scale (1-30)



Irritability Negativity Affectivity Subscale (1-50)



Different scales but averages mean the same thing
(i.e. average person is just as irritable!)

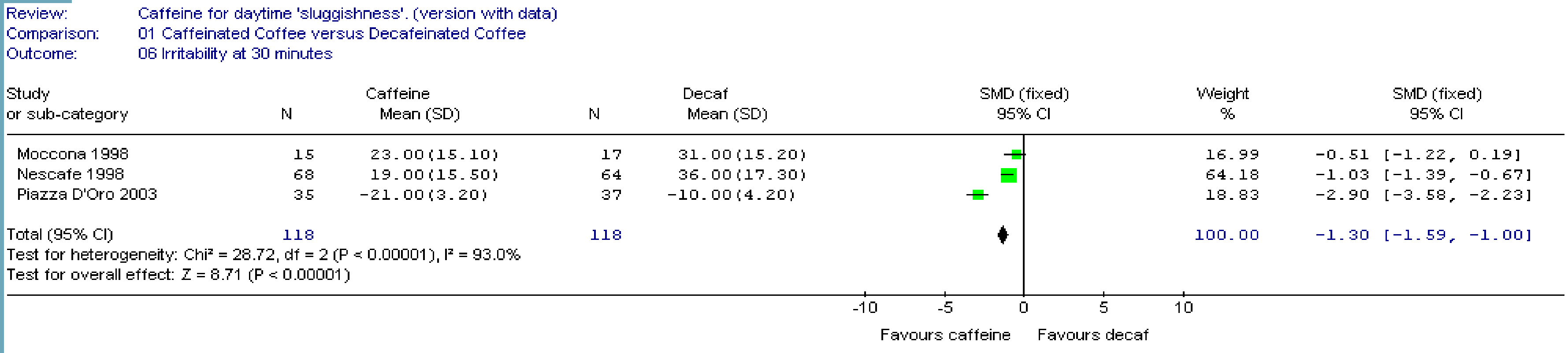
Measurements on different scales

Comparing irritability at 30 minutes between caffeinated coffee and decafe coffee

Trial	**Caffeinated**	**Decafe**	**Irritability**
	N. mean (SD)	**N. mean (SD)**	scale
Moccona 1998	15 23.0 (15.1)	17 31.0 (15.2)	INAS
Nescafe 1998	68 19.0 (15.5)	64 36.0 (17.3)	INAS
Piazza D’oro 2003	35 21.0 (3.2)	37 10.0 (4.20)	BII

High scores on the Beck Irritability Scale (BII) (1-30) good outcomes, while high scores on the Irritability Negative Affectivity Subscale (INAS) (1-50) are poor outcomes

SMD: example



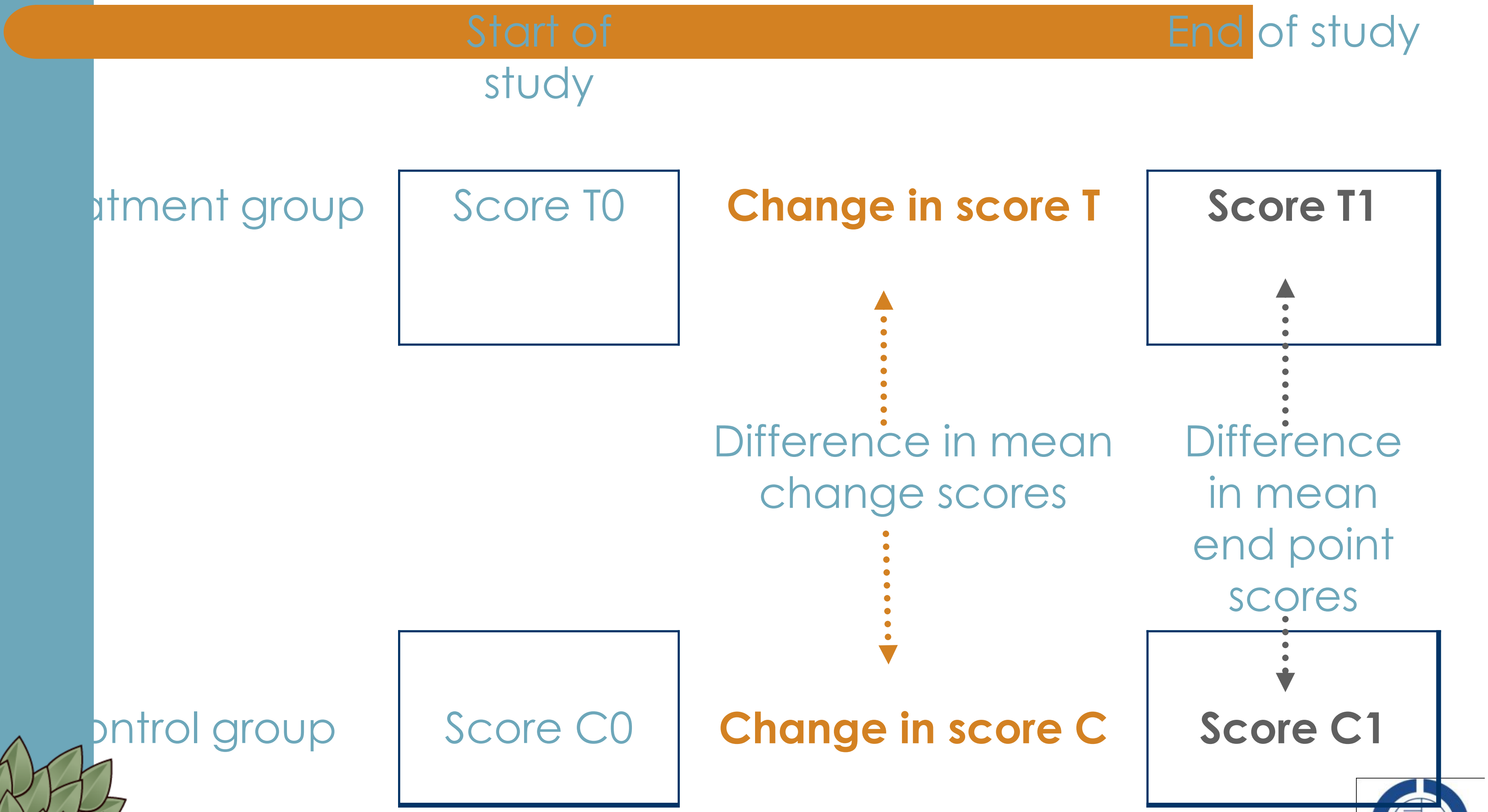
Review: Caffeine for daytime 'sluggishness'. (version with data)
Comparison: 01 Caffeinated Coffee versus Decafeinated Coffee
Outcome: 06 Irritability at 30 minutes

Study or sub-category	N	Caffeine Mean (SD)	N	Decaf Mean (SD)	SMD (fixed) 95% CI	Weight %	SMD (fixed) 95% CI
Moccona 1998	15	23.00(15.10)	17	31.00(15.20)		16.99	-0.51 [-1.22, 0.19]
Nescafe 1998	68	19.00(15.50)	64	36.00(17.30)		64.18	-1.03 [-1.39, -0.67]
Piazza D'Oro 2003	35	-21.00(3.20)	37	-10.00(4.20)		18.83	-2.90 [-3.58, -2.23]
Total (95% CI)	118		118			100.00	-1.30 [-1.59, -1.00]

Test for heterogeneity: Chi² = 28.72, df = 2 (P < 0.00001), I² = 93.0%
Test for overall effect: Z = 8.71 (P < 0.00001)

RevMan exercise

Change vs endpoint scores



Start of study

End of study

atment group

Score T0

Change in score T

Score T1

Difference in mean change scores

Difference in mean end point scores

ontrol group

Score C0

Change in score C

Score C1

Problems with MD and SMD

- what constitutes a clinically important change?
- restrictive eligibility criteria results in smaller standard deviations; therefore these trials given more weight
- mean difference
  - measurements on the same scale are not always comparable (e.g. health care costs in different places, process of care measures)
- standardised mean difference
  - difficult to interpret outcomes in units of SD, but can transform back to units of the scale
  - estimates of variation may not always be comparable making the SD a poor scaling factor

Take home message

- pooling continuous data – use mean difference or standardised mean difference
- check data for skewness
- can calculate SDs from other statistics
- can use either endpoint or change scores

Heterogeneity

Outline

- what is heterogeneity?
- causes of heterogeneity
- identifying heterogeneity
- dealing with heterogeneity
- fixed and random effects meta-analysis

What is heterogeneity?

- Heterogeneity is variation between the results of a set of studies

Causes of heterogeneity: clinical

Differences between studies with respect to:

- participants
  - conditions under investigation, eligibility criteria for trials, geographical variation
- interventions
  - e.g. type of drug, intensity, dose, duration, mode of administration, experience of practitioners, nature of control (placebo, none, standard care)
- outcomes
  - e.g. type, follow-up duration, ways of measuring outcomes, definition of an event

Causes of heterogeneity: methodological

Differences between studies with respect to:

- design
  - e.g. randomised vs non-randomised, parallel group vs crossover vs cluster randomised, length
- conduct
  - e.g. allocation concealment, blinding, approach to analysis, imputation methods for missing data

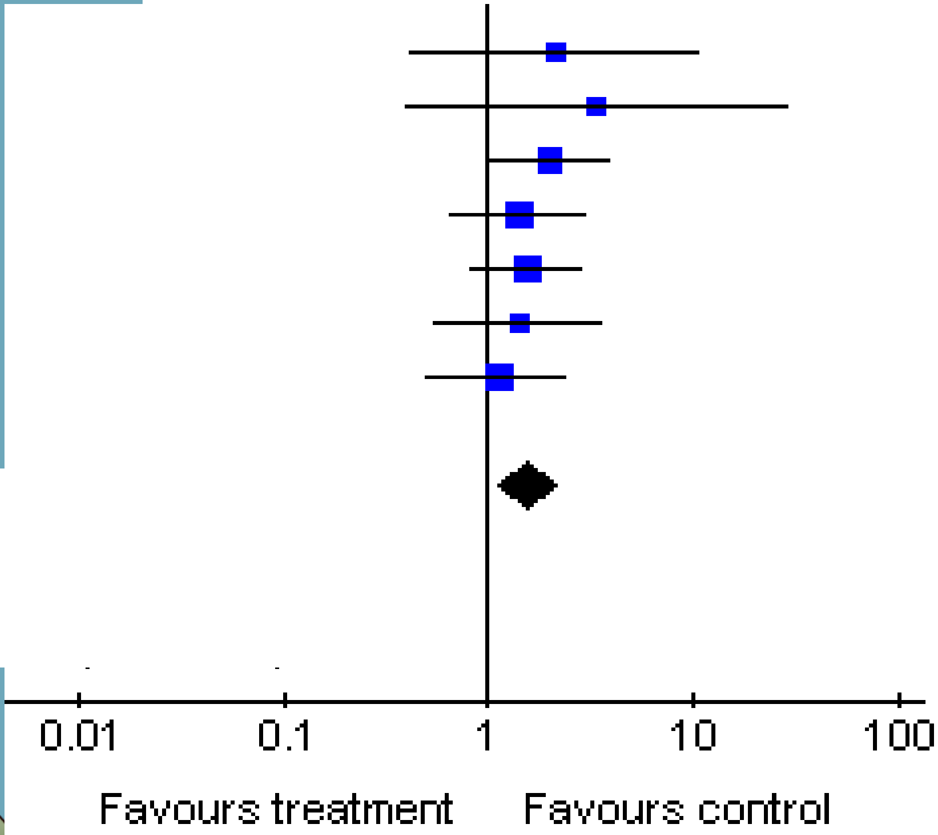
Statistical heterogeneity

- excessive variation in the results of studies above that expected by chance

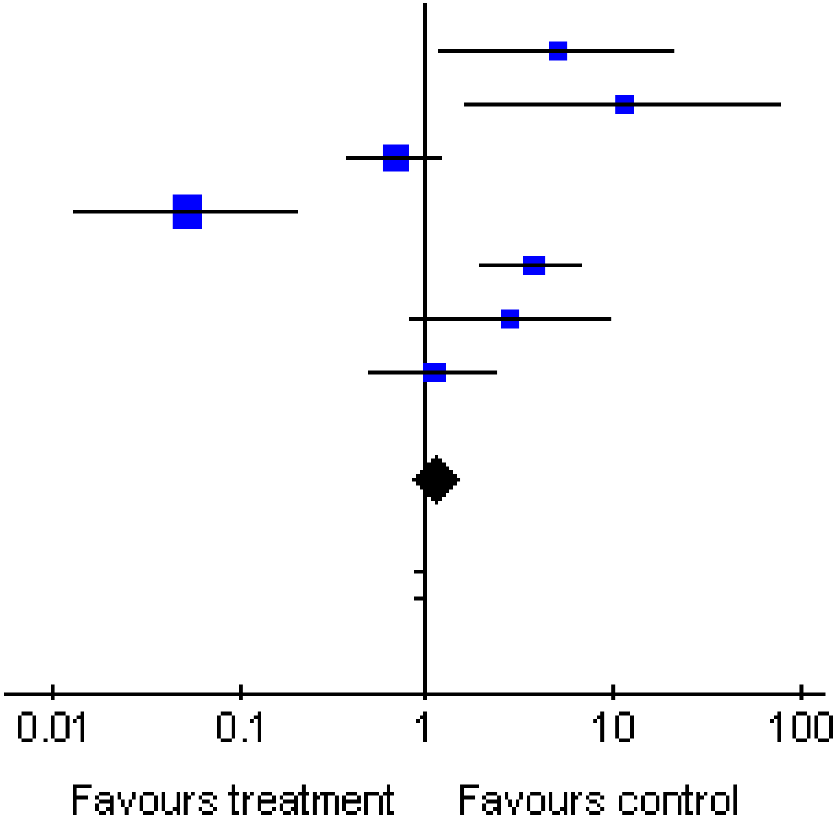
Identifying heterogeneity

1. graphically – the eyeball test
2. numerically – the I^2 test

Forest plot A



Forest plot B



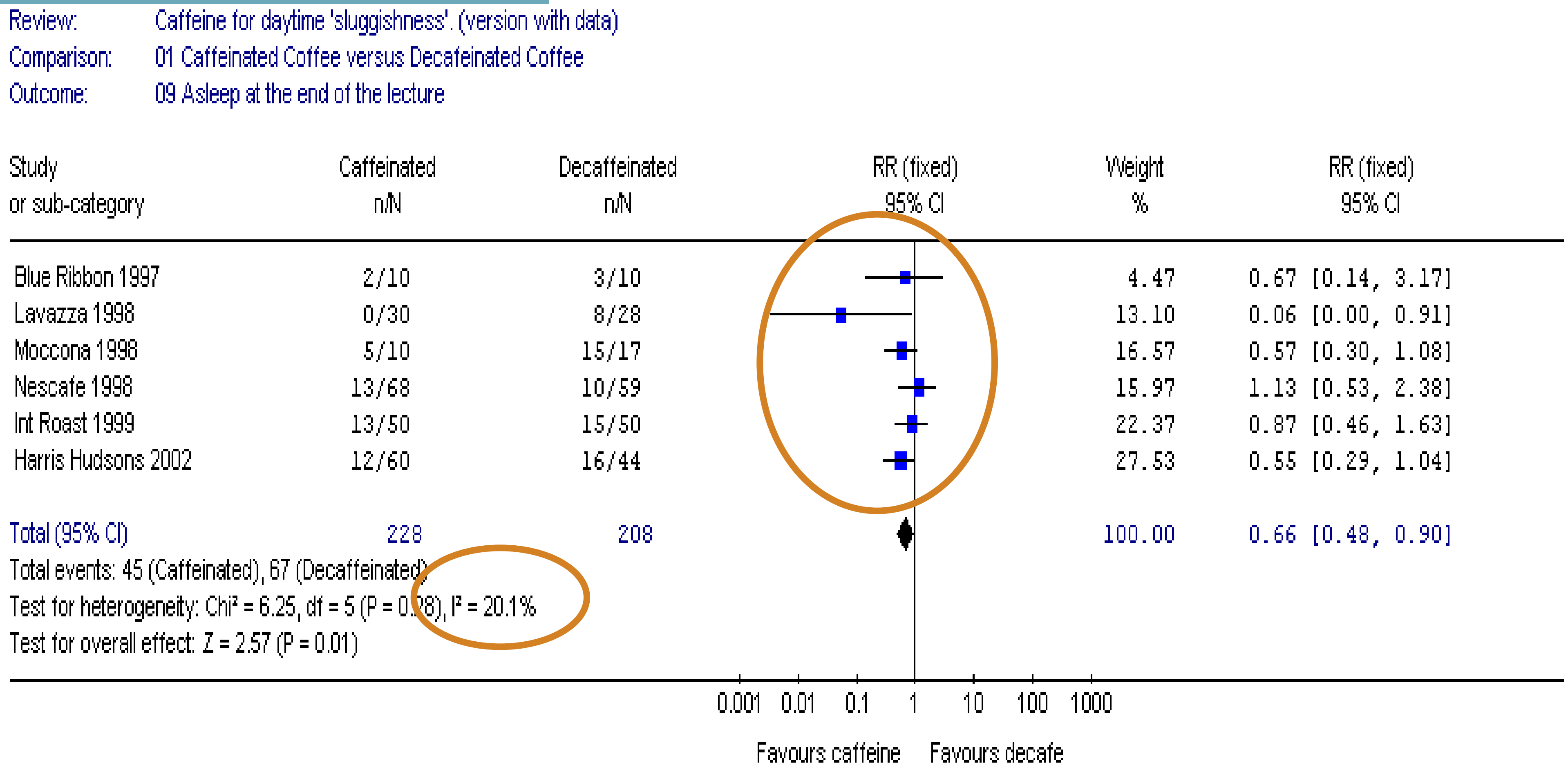
Quantifying heterogeneity

- I^2 describes the proportion of total variation across studies that is due to heterogeneity rather than chance
- based on Cochran Q test and its degrees of freedom

- $I^2 = \frac{(Q - df)}{Q} \times 100\%$ (df = the number of studies minus 1)

Quantifying heterogeneity

- low (and negative) values of I^2 indicate no, or little, heterogeneity
- larger values of I^2 show increasing heterogeneity
- roughly, values of of 25%, 50% and 75% correspond to low, moderate and high levels of heterogeneity (Higgins et al 2003, BMJ)



Review: Caffeine for daytime 'sluggishness'. (version with data)
Comparison: 01 Caffeinated Coffee versus Decafeinated Coffee
Outcome: 09 Asleep at the end of the lecture

Study or sub-category	Caffeinated n/N	Decaffeinated n/N	RR (fixed) 95% CI	Weight %	RR (fixed) 95% CI
Blue Ribbon 1997	2/10	3/10		4.47	0.67 [0.14, 3.17]
Lavazza 1998	0/30	8/28		13.10	0.06 [0.00, 0.91]
Moccona 1998	5/10	15/17		16.57	0.57 [0.30, 1.08]
Nescafe 1998	13/68	10/59		15.97	1.13 [0.53, 2.38]
Int Roast 1999	13/50	15/50		22.37	0.87 [0.46, 1.63]
Harris Hudsons 2002	12/60	16/44		27.53	0.55 [0.29, 1.04]
Total (95% CI)	228	208		100.00	0.66 [0.48, 0.90]

Total events: 45 (Caffeinated), 67 (Decaffeinated)
Test for heterogeneity: Chi² = 6.25, df = 5 (P = 0.28), I² = 20.1%
Test for overall effect: Z = 2.57 (P = 0.01)

Dealing with heterogeneity

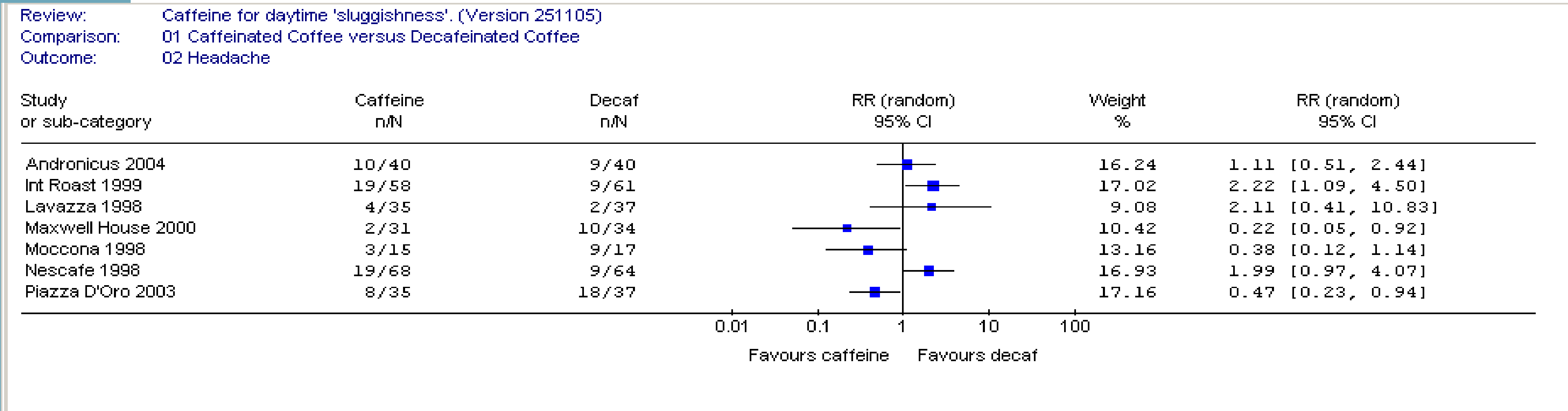
Options available to you:

1. check the data
2. don't pool studies
3. ignore heterogeneity: use fixed effect model
4. investigate reasons for heterogeneity
5. incorporate heterogeneity: use random effects model

Option 1: Check the data

- Check extracted data
- Check analyses of individual studies

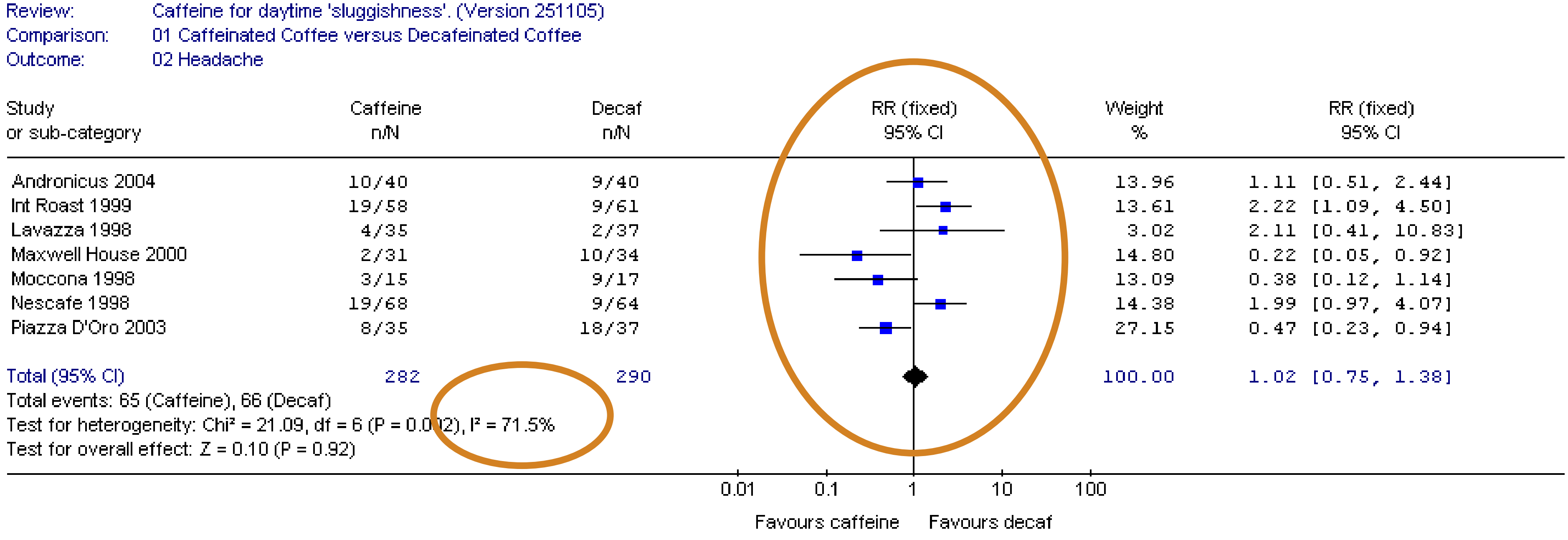
Option 2: Don't pool studies



Review:	Caffeine for daytime 'sluggishness'. (Version 251105)
Comparison:	01 Caffeinated Coffee versus Decafeinated Coffee
Outcome:	02 Headache

Study or sub-category	Caffeine n/N	Decaf n/N	Weight %	RR (random) 95% CI
Andronicus 2004	10/40	9/40	16.24	1.11 [0.51, 2.44]
Int Roast 1999	19/58	9/61	17.02	2.22 [1.09, 4.50]
Lavazza 1998	4/35	2/37	9.08	2.11 [0.41, 10.83]
Maxwell House 2000	2/31	10/34	10.42	0.22 [0.05, 0.92]
Moccona 1998	3/15	9/17	13.16	0.38 [0.12, 1.14]
Nescafe 1998	19/68	9/64	16.93	1.99 [0.97, 4.07]
Piazza D'Oro 2003	8/35	18/37	17.16	0.47 [0.23, 0.94]

Option 3: Ignore heterogeneity



Review: Caffeine for daytime 'sluggishness'. (Version 251105)
Comparison: 01 Caffeinated Coffee versus Decafeinated Coffee
Outcome: 02 Headache

Study or sub-category	Caffeine n/N	Decaf n/N	RR (fixed) 95% CI	Weight %	RR (fixed) 95% CI
Andronicus 2004	10/40	9/40		13.96	1.11 [0.51, 2.44]
Int Roast 1999	19/58	9/61		13.61	2.22 [1.09, 4.50]
Lavazza 1998	4/35	2/37		3.02	2.11 [0.41, 10.83]
Maxwell House 2000	2/31	10/34		14.80	0.22 [0.05, 0.92]
Moccona 1998	3/15	9/17		13.09	0.38 [0.12, 1.14]
Nescafe 1998	19/68	9/64		14.38	1.99 [0.97, 4.07]
Piazza D'Oro 2003	8/35	18/37		27.15	0.47 [0.23, 0.94]
Total (95% CI)	282	290		100.00	1.02 [0.75, 1.38]

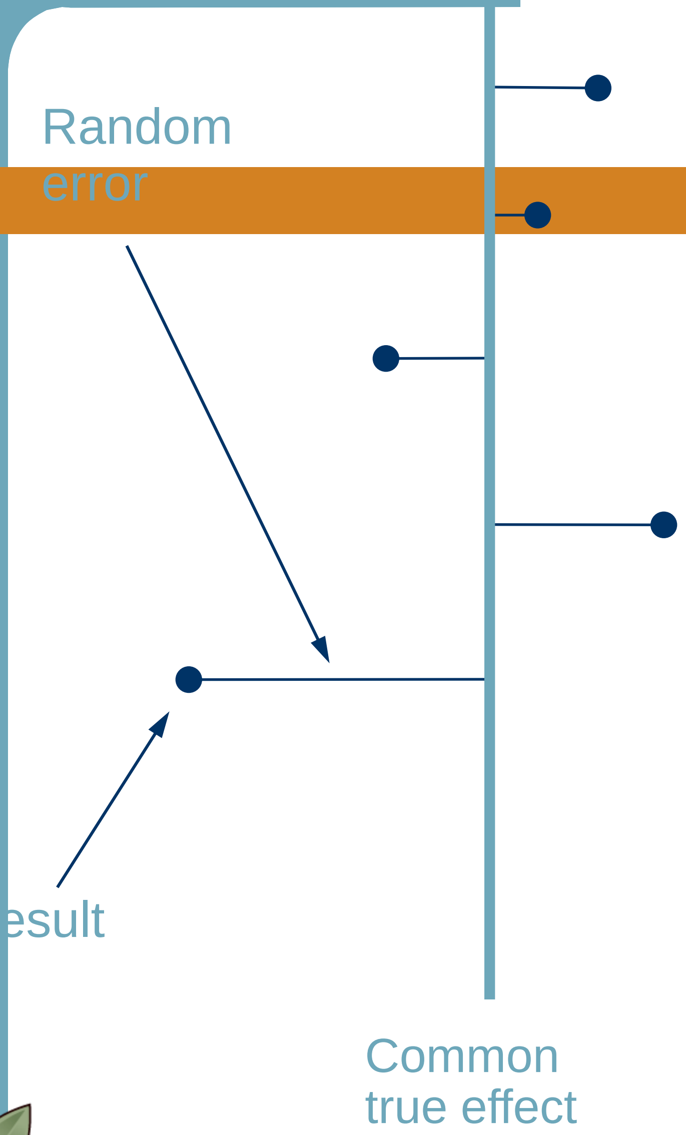
Total events: 65 (Caffeine), 66 (Decaf)
Test for heterogeneity: Chi² = 21.09, df = 6 (P = 0.002), I² = 71.5%
Test for overall effect: Z = 0.10 (P = 0.92)

Fixed effect model

Philosophy behind model:

- there is one real value for the treatment effect
- all trials are estimating this common treatment effect

Fixed effect model



- assumes that all studies are evaluating the same treatment effect
- *i.e.* if they were all infinitely large they'd produce an identical result

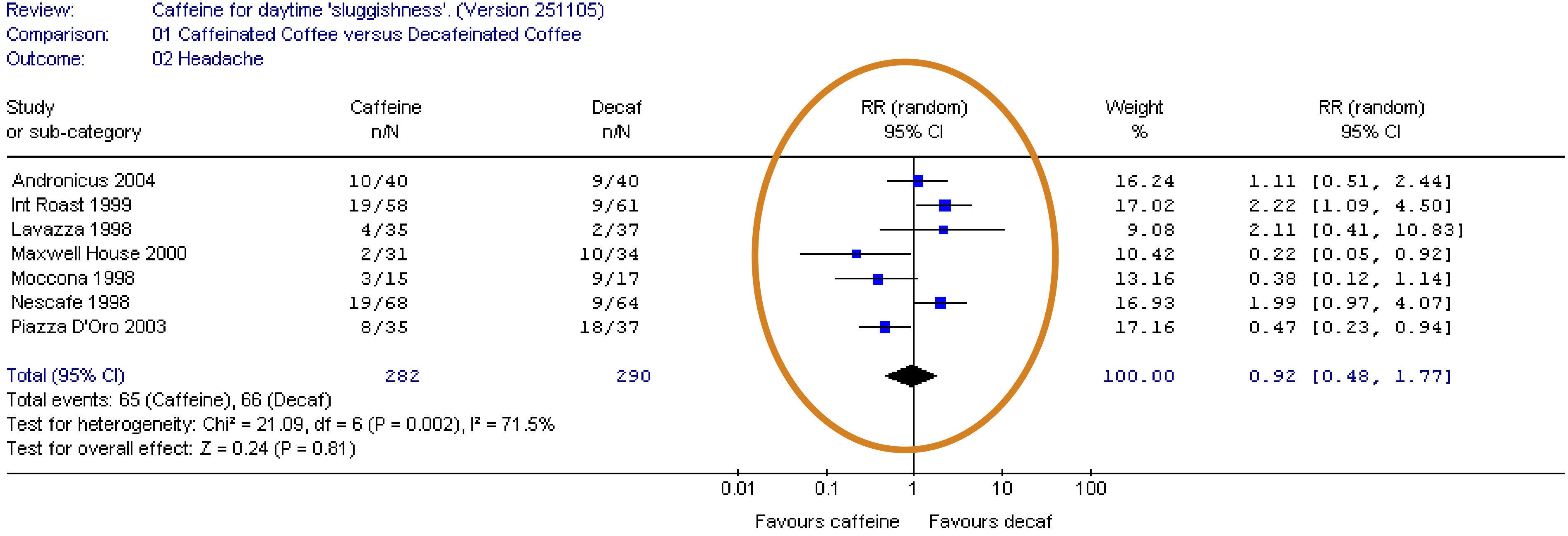
Option 4: Investigating heterogeneity

- as an objective of your review (should be pre-specified in your protocol)
- to determine causes of unexpected statistical heterogeneity
  - note. post hoc investigations should be reported as such and are hypothesis-generating at best

nvestigating heterogeneity: tools

- subgroup analysis
  - get answers to secondary questions concerning subsets of participants or interventions
  - can yield spurious findings if not used carefully
- meta-regression
  - examine relationship between treatment effect and a particular characteristic of the study (not patients)
    - not available in RevMan
- individual patient data (IPD) meta-analysis
  - investigate patient-level characteristics
  - time consuming and expensive

Option 5: Incorporate heterogeneity



Review: Caffeine for daytime 'sluggishness'. (Version 251105)
Comparison: 01 Caffeinated Coffee versus Decafeinated Coffee
Outcome: 02 Headache

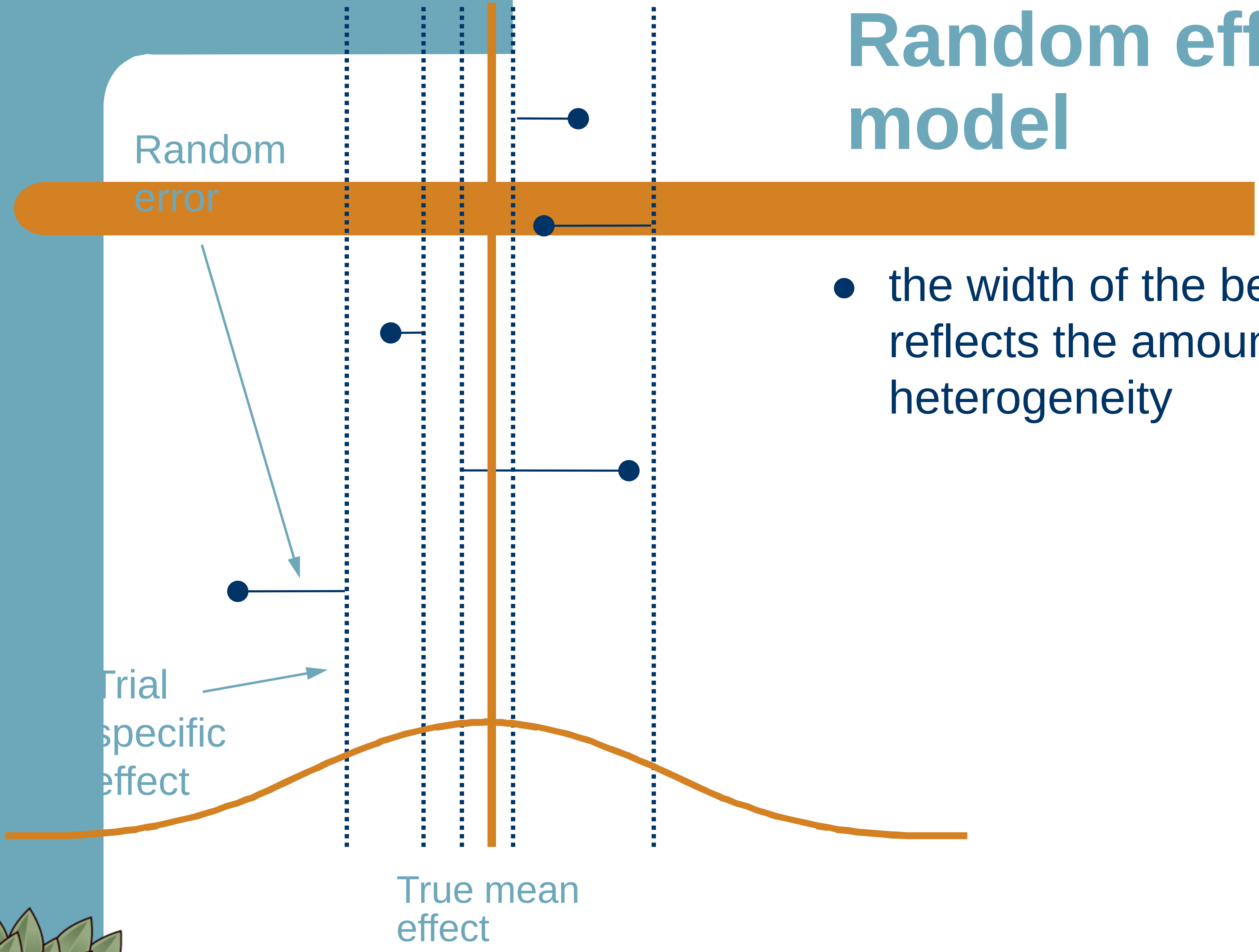
Study or sub-category	Caffeine n/N	Decaf n/N	Weight %	RR (random) 95% CI
Andronicus 2004	10/40	9/40	16.24	1.11 [0.51, 2.44]
Int Roast 1999	19/58	9/61	17.02	2.22 [1.09, 4.50]
Lavazza 1998	4/35	2/37	9.08	2.11 [0.41, 10.83]
Maxwell House 2000	2/31	10/34	10.42	0.22 [0.05, 0.92]
Moccona 1998	3/15	9/17	13.16	0.38 [0.12, 1.14]
Nescafe 1998	19/68	9/64	16.93	1.99 [0.97, 4.07]
Piazza D'Oro 2003	8/35	18/37	17.16	0.47 [0.23, 0.94]
Total (95% CI)	282	290	100.00	0.92 [0.48, 1.77]

Total events: 65 (Caffeine), 66 (Decaf)
Test for heterogeneity: Chi² = 21.09, df = 6 (P = 0.002), I² = 71.5%
Test for overall effect: Z = 0.24 (P = 0.81)

Random effects model

- if heterogeneity cannot be explained by characteristics of the studies, it may be incorporated into the meta-analysis using the random-effects model
- the true treatment effects underlying the studies are allowed to differ and are assumed to be distributed around a central (mean) value
- weights are adjusted to account for both within-study and between-study variation

Random effects model



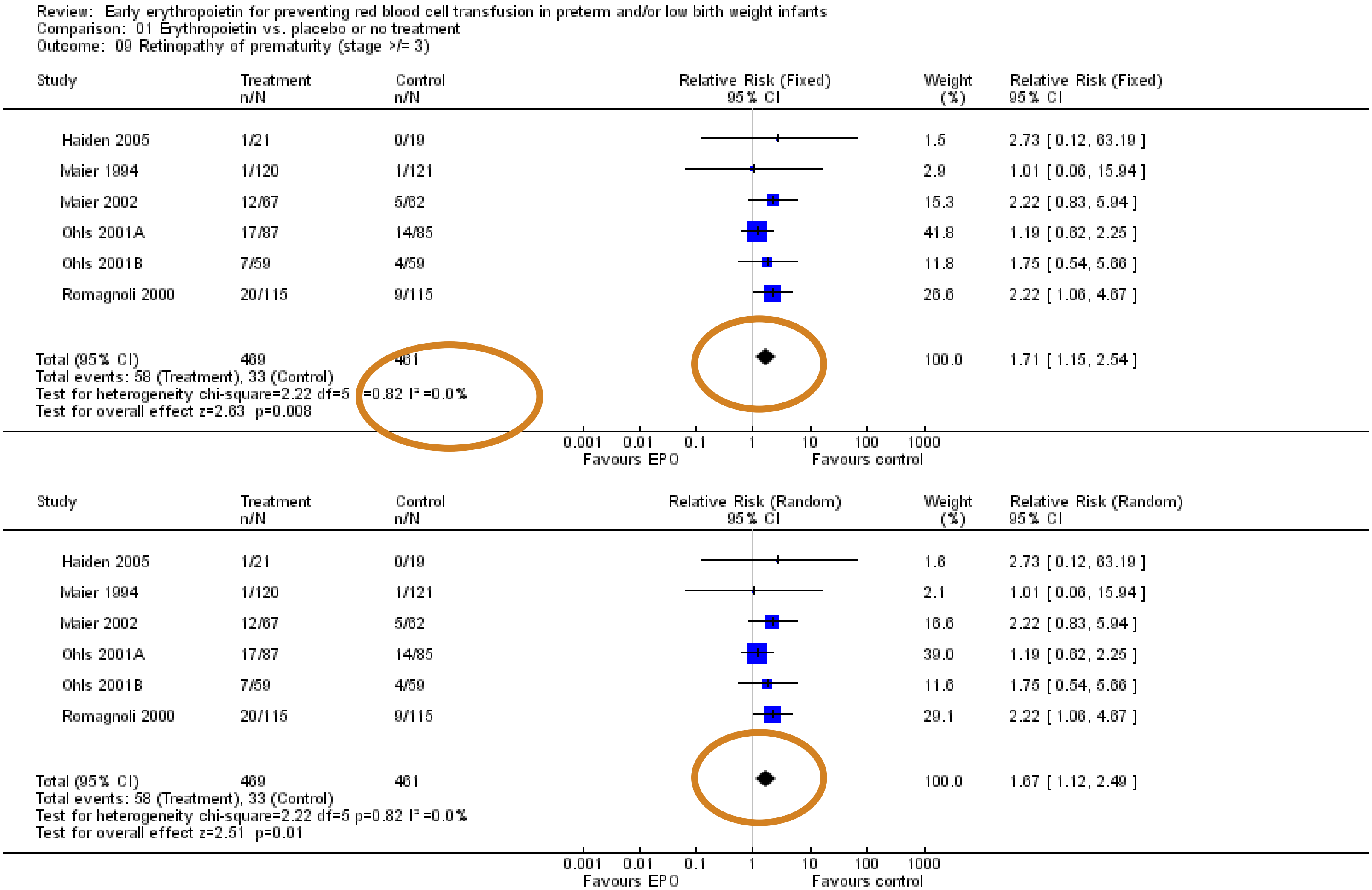
- the width of the bell shape reflects the amount of heterogeneity

nterpreting random effects meta-analyses

Random effects meta-analyses are...

- **identical** to fixed effect analyses when there is no clear heterogeneity
- **similar** to fixed effect meta-analyses but *with wider confidence intervals* when there is heterogeneity
- **different** from fixed effect meta-analyses when there is publication bias (or funnel plot asymmetry)
  - random effects analyses give relatively more weight to smaller studies

Fixed versus random effects



Review: Early erythropoietin for preventing red blood cell transfusion in preterm and/or low birth weight infants
Comparison: 01 Erythropoietin vs. placebo or no treatment
Outcome: 09 Retinopathy of prematurity (stage >/= 3)

Study	Treatment n/N	Control n/N	Weight (%)	Relative Risk (Fixed) 95% CI
Haiden 2005	1/21	0/19	1.5	2.73 [0.12, 63.19]
Maier 1994	1/120	1/121	2.9	1.01 [0.06, 15.94]
Maier 2002	12/67	5/62	15.3	2.22 [0.83, 5.94]
Ohls 2001A	17/87	14/85	41.8	1.19 [0.62, 2.25]
Ohls 2001B	7/59	4/59	11.8	1.75 [0.54, 5.66]
Romagnoli 2000	20/115	9/115	26.6	2.22 [1.06, 4.67]
Total (95% CI)	469	461	100.0	1.71 [1.15, 2.54]

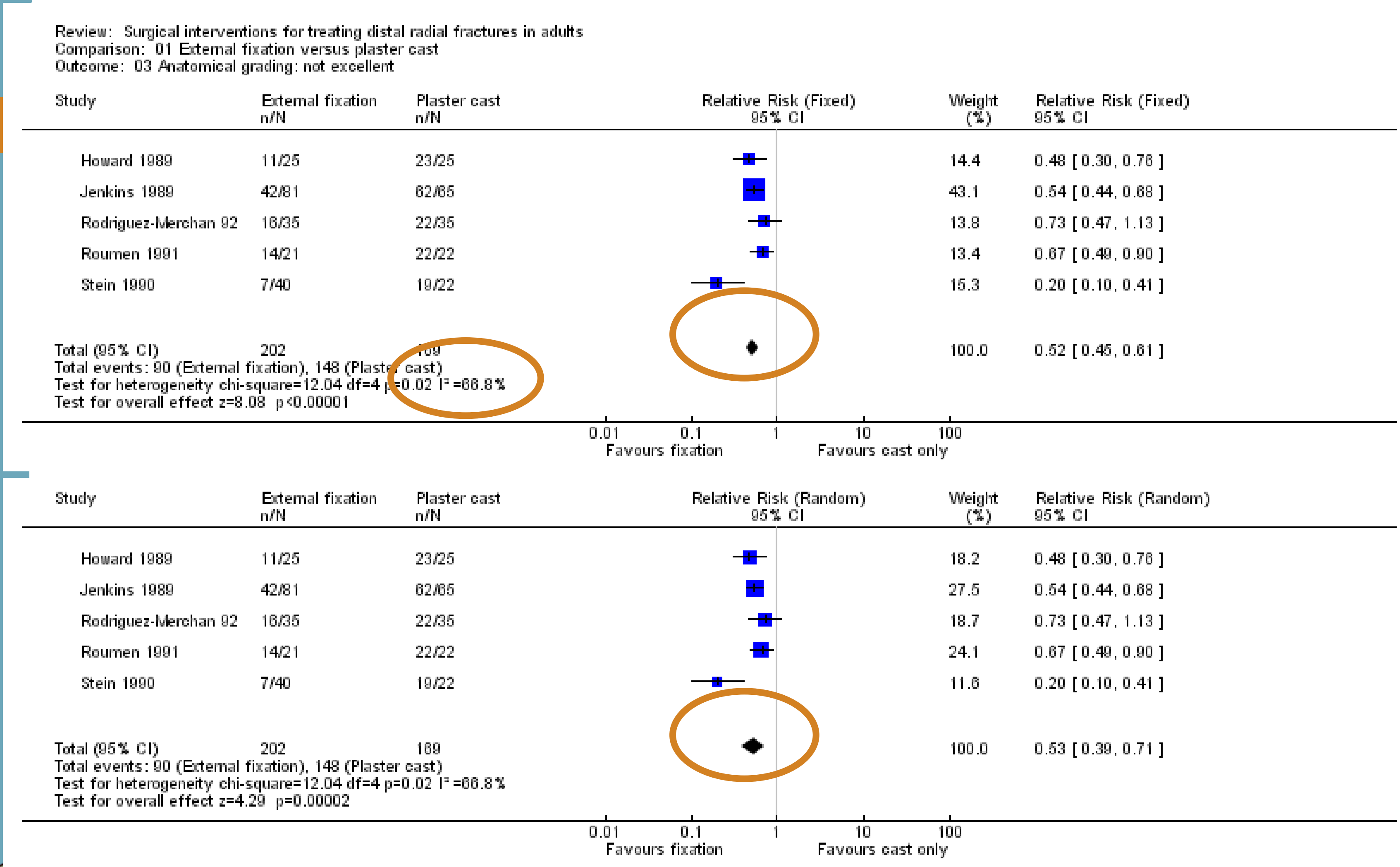
Total events: 58 (Treatment), 33 (Control)
Test for heterogeneity chi-square=2.22 df=5 p=0.82 I² =0.0%
Test for overall effect z=2.63 p=0.008

Study	Treatment n/N	Control n/N	Weight (%)	Relative Risk (Random) 95% CI
Haiden 2005	1/21	0/19	1.6	2.73 [0.12, 63.19]
Maier 1994	1/120	1/121	2.1	1.01 [0.06, 15.94]
Maier 2002	12/67	5/62	16.6	2.22 [0.83, 5.94]
Ohls 2001A	17/87	14/85	39.0	1.19 [0.62, 2.25]
Ohls 2001B	7/59	4/59	11.6	1.75 [0.54, 5.66]
Romagnoli 2000	20/115	9/115	29.1	2.22 [1.06, 4.67]
Total (95% CI)	469	461	100.0	1.67 [1.12, 2.49]

Total events: 58 (Treatment), 33 (Control)
Test for heterogeneity chi-square=2.22 df=5 p=0.82 I² =0.0%
Test for overall effect z=2.51 p=0.01

almost identical

Fixed versus random effects



Review: Surgical interventions for treating distal radial fractures in adults
Comparison: 01 External fixation versus plaster cast
Outcome: 03 Anatomical grading: not excellent

Study	External fixation n/N	Plaster cast n/N	Weight (%)	Relative Risk (Fixed) 95% CI
Howard 1989	11/25	23/25	14.4	0.48 [0.30, 0.76]
Jenkins 1989	42/81	62/65	43.1	0.54 [0.44, 0.68]
Rodriguez-Merchan 92	16/35	22/35	13.8	0.73 [0.47, 1.13]
Roumen 1991	14/21	22/22	13.4	0.67 [0.49, 0.90]
Stein 1990	7/40	19/22	15.3	0.20 [0.10, 0.41]
Total (95% CI)	202	169	100.0	0.52 [0.45, 0.61]

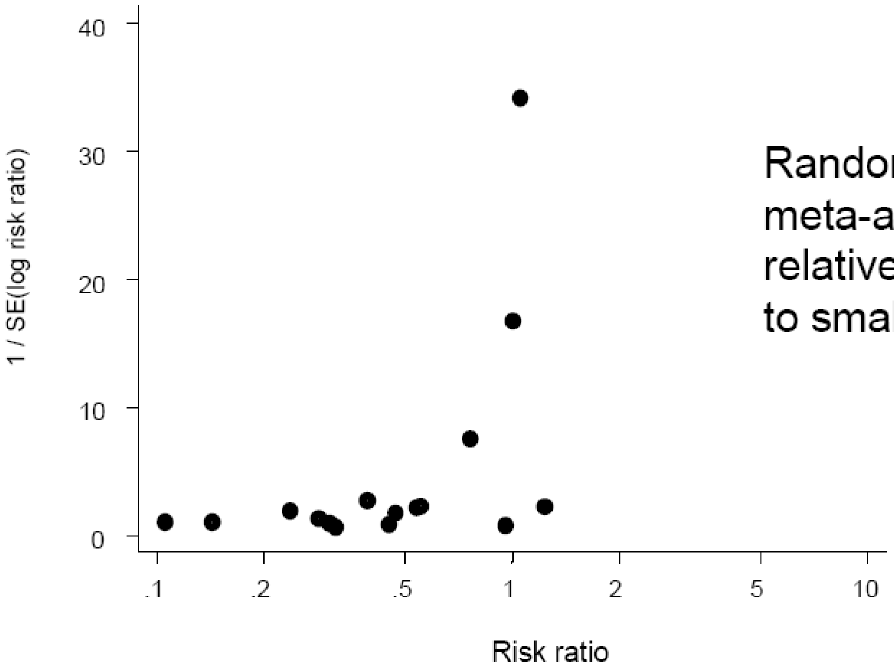
Total events: 90 (External fixation), 148 (Plaster cast)
Test for heterogeneity chi-square=12.04 df=4 p=0.02 I² =66.8%
Test for overall effect z=8.08 p<0.00001

Study	External fixation n/N	Plaster cast n/N	Weight (%)	Relative Risk (Random) 95% CI
Howard 1989	11/25	23/25	18.2	0.48 [0.30, 0.76]
Jenkins 1989	42/81	62/65	27.5	0.54 [0.44, 0.68]
Rodriguez-Merchan 92	16/35	22/35	18.7	0.73 [0.47, 1.13]
Roumen 1991	14/21	22/22	24.1	0.67 [0.49, 0.90]
Stein 1990	7/40	19/22	11.6	0.20 [0.10, 0.41]
Total (95% CI)	202	169	100.0	0.53 [0.39, 0.71]

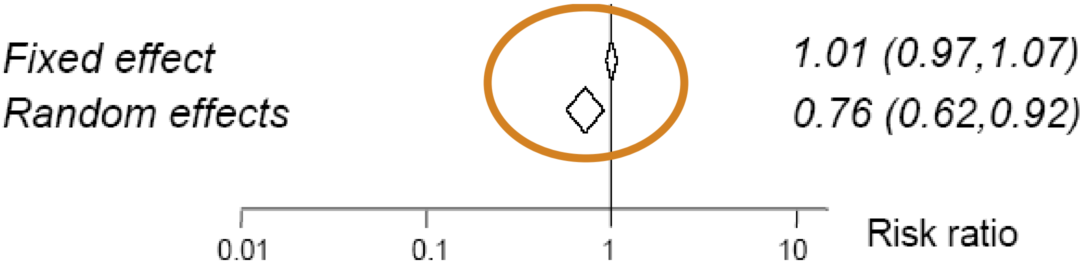
Total events: 90 (External fixation), 148 (Plaster cast)
Test for heterogeneity chi-square=12.04 df=4 p=0.02 I² =66.8%
Test for overall effect z=4.29 p=0.00002

similar, but wider CIs

Fixed versus random effects



Random-effects meta-analyses give relatively more weight to smaller studies



very different results

source: with thanks to Julian Higgins

Take home messages

- heterogeneity should be assessed and addressed
- statistical heterogeneity occurs when studies are not all evaluating the same treatment effect
- looking at overlap of confidence intervals on forest plot is a good way to identify statistical heterogeneity
- I^2 can quantify the degree of inconsistency across studies
- there are several options for dealing with heterogeneity
- methods to investigate heterogeneity should be pre-specified in the protocol
- random effects meta-analyses are useful for incorporating unexplained variability into a summary
- but random effects meta-analyses are not a panacea