MRC | Biostatistics Unit

Network meta-analysis

LSHTM, 31st January 2014

Ian White
MRC Biostatistics Unit, Cambridge, UK

MRC Biostatistics Unit

Assessing inconsistency in multiple treatments meta-analysis

LSHTM
Wednesday 24 November 2010

Ian White
MRC Biostatistics Unit, Cambridge

MRC | Medical Research Council

Have you heard it all before?

No, thanks to:

Maurice Belz
(1897–1975)

SSAI
Statistical Society of Australia Inc.

Plan

- Systematic review
- Meta-analysis
- Indirect comparisons
- Network meta-analysis
  - models allowing for heterogeneity
  - models allowing for inconsistency
  - model estimation
  - examples
  - controversies

Systematic review

- Define a clinical question
  - typically: how good is this intervention? (often drugs, but also e.g. psychological therapy)
- Obtain all papers relevant to the question using a systematic search strategy
  - typically restricted to randomised controlled trials (RCTs)
- Record study characteristics including quality
- Extract quantitative study results
- If appropriate, perform a statistical summary

statistician's roles

this talk

See e.g.

- Egger M, Davey Smith G, Altman DG, editors. *Systematic Reviews in Healthcare: Meta-analysis in Context* (2001)
- The Cochrane Collaboration, http://www.cochrane.org/

Pairwise meta-analysis: data from 15 randomised trials

study	dA	nA	dC	nC
1	9	140	23	140
6	75	731	363	714
7	2	106	9	205
8	58	549	237	1561
9	0	33	9	48
10	3	100	31	98
11	1	31	26	95
12	6	39	17	77
13	95	1107	134	1031
14	15	187	35	504
15	78	584	73	675
16	69	1177	54	888
17	64	642	107	761
18	5	62	8	90
19	20	234	34	237

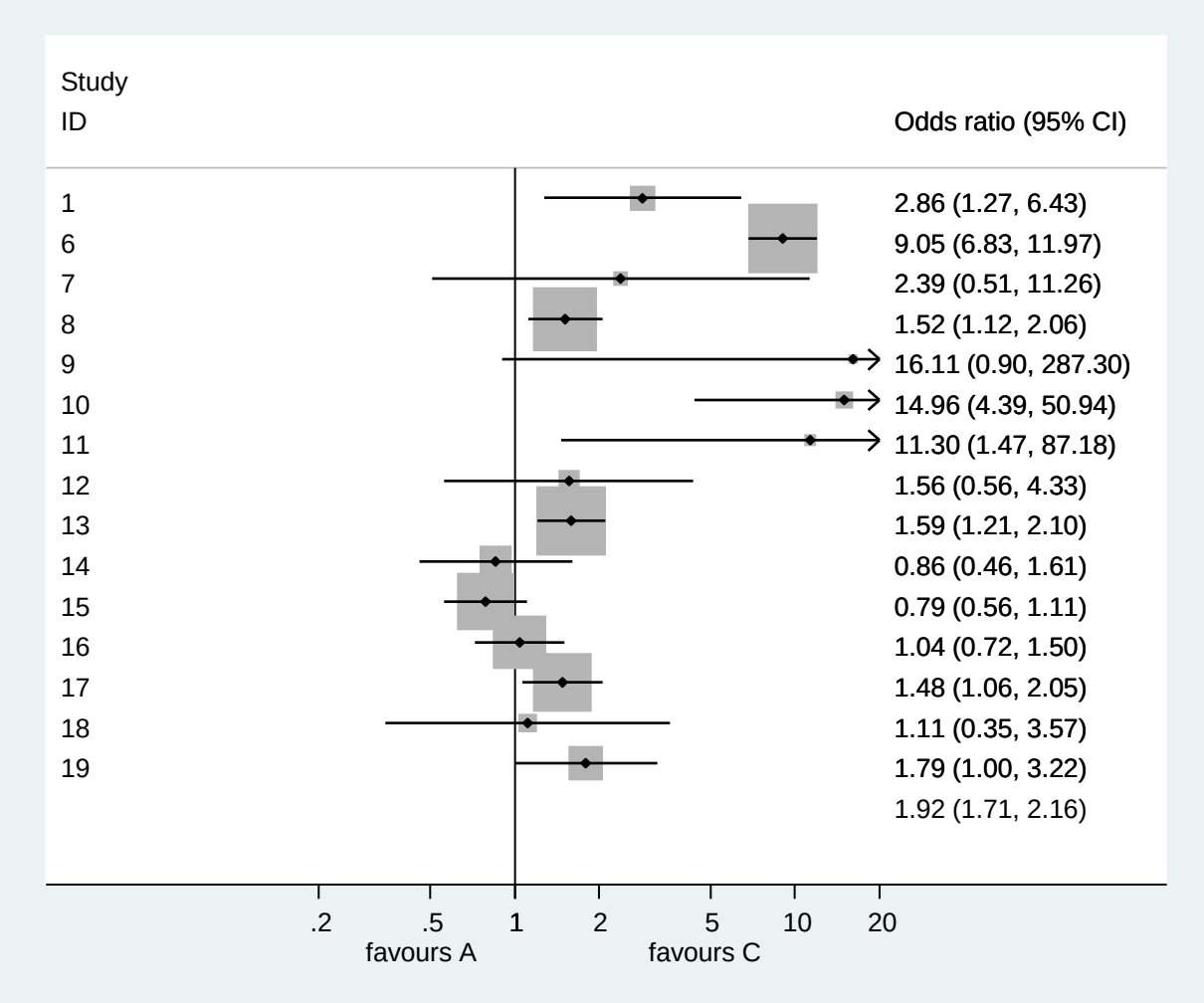
Aim is to compare effectiveness of individual counselling ("C") with no contact ("A") in helping smokers to quit.

Data in arm A, C:

- dA, dC = # who quit smoking
- nA, nC = # randomised

Source: Lu & Ades, JASA 2006; 101: 447–459.

Data display: Forest plot



Forest plot shows odds ratio (95% confidence interval) for C vs. A for each of the 15 studies.

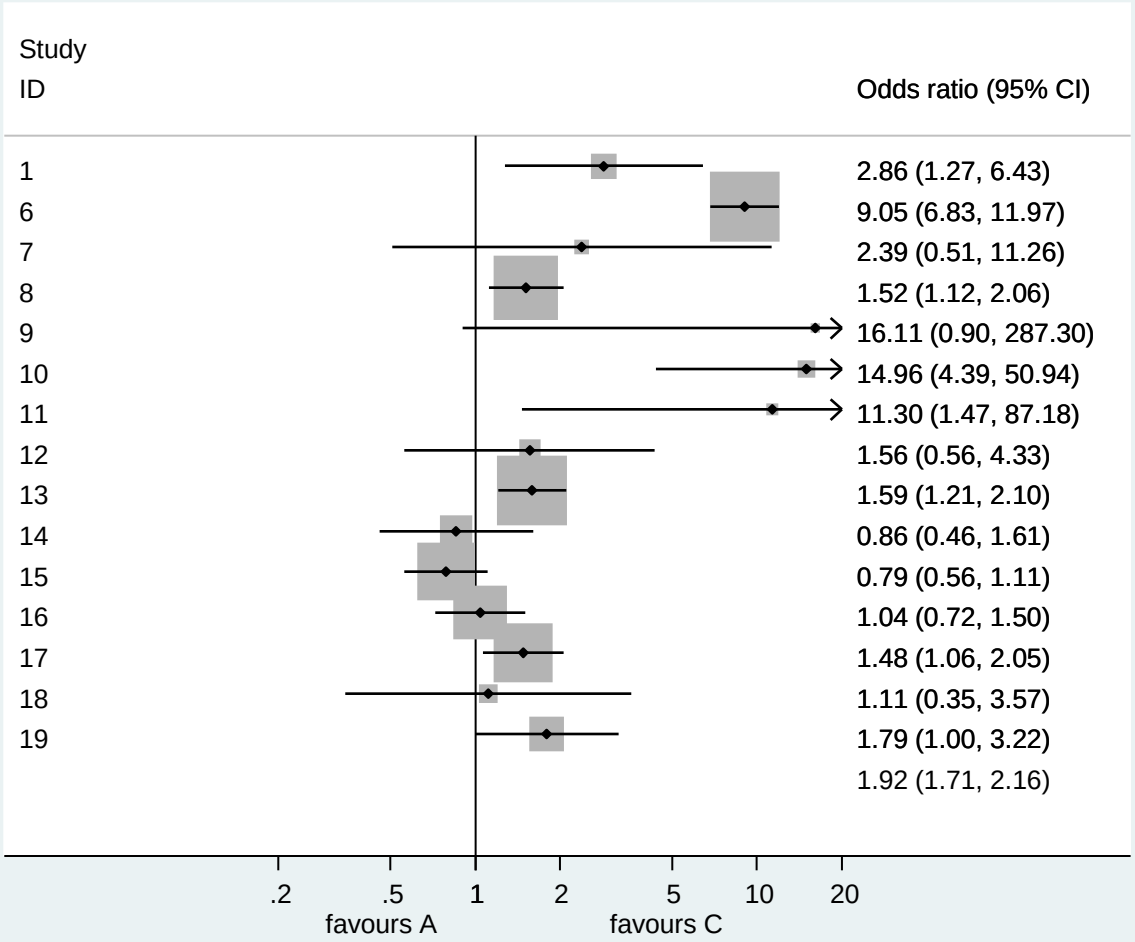
Shaded blocks represent amount of information (area $\propto 1/se^2$)

Pairwise meta-analysis: "fixed-effect" model

- Say we're interested in the log odds ratio
- Assume there is a "true log odds ratio" μ
- Express the results from study i as
  - y_i = estimated log odds ratio
  - s_i = its standard error
- Model: $y_i \sim N(\mu, s_i^2)$
  - approximation, valid for moderate/large counts

- (We are using a two-stage estimation procedure: compute the y_i, then estimate μ. We can also do one-stage estimation – see later.)

Forest plot again



Note the high degree of heterogeneity between studies.

Ideally we'd explain it – e.g. if study 6 was in people who had just had a major diagnosis.

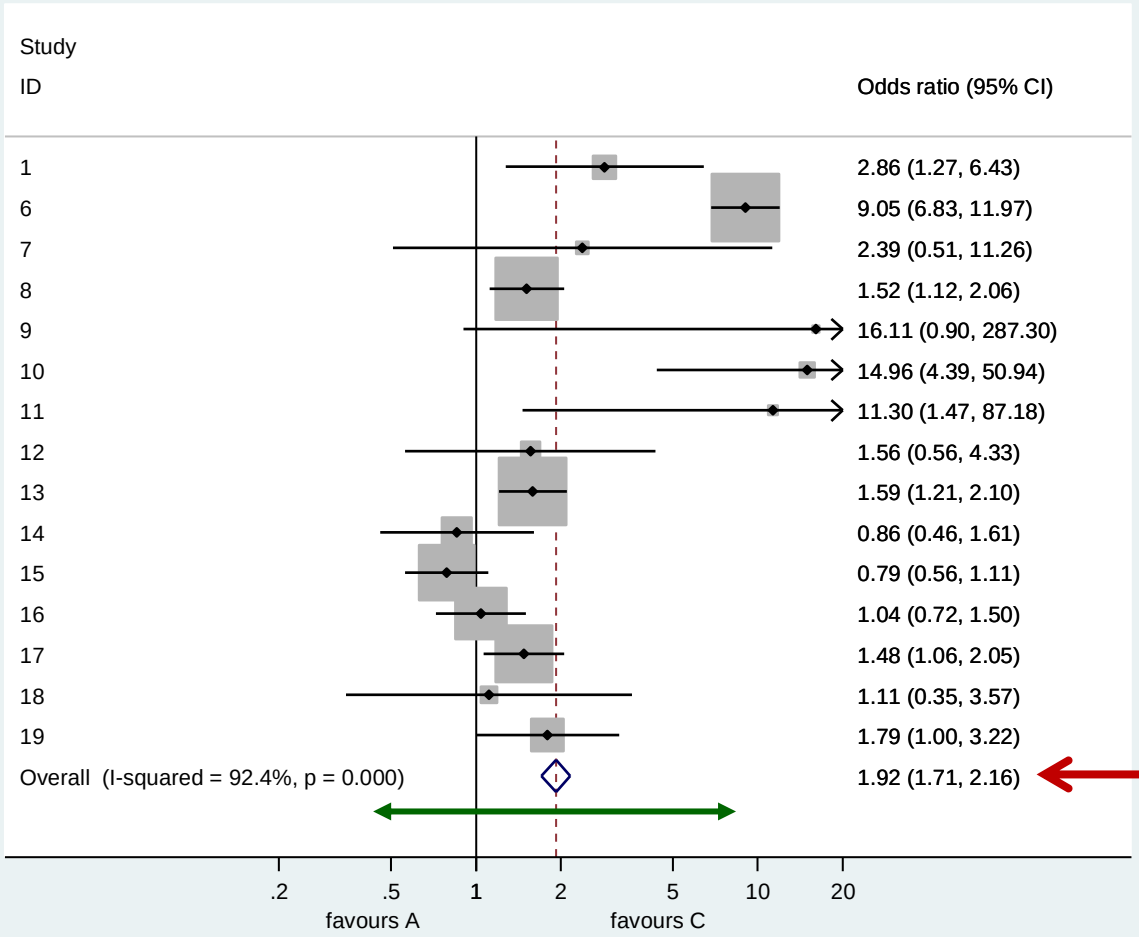
But often we need to model it instead.

Pairwise meta-analysis: random-effects model

- Model for "true log odds ratio in study i": $\mu_i \sim N(\mu, \tau^2)$
- Parameters of interest:
  - μ is the overall mean treatment effect
  - τ^2 is the between-studies (heterogeneity) variance

- Two-stage estimation procedure
- Model for point estimate: $y_i \sim N(\mu_i, s_i^2)$
  - y_i = estimated log odds ratio in study *i*
  - s_i = its standard error
- Estimate τ^2 (and hence μ) by
  - method of moments – very popular
  - or restricted maximum likelihood (REML)

Forest plot showing meta-analysis result



The random-effects analysis gives an estimate of the overall mean allowing for heterogeneity

and a prediction interval (effect in a new study)

A note on terminology

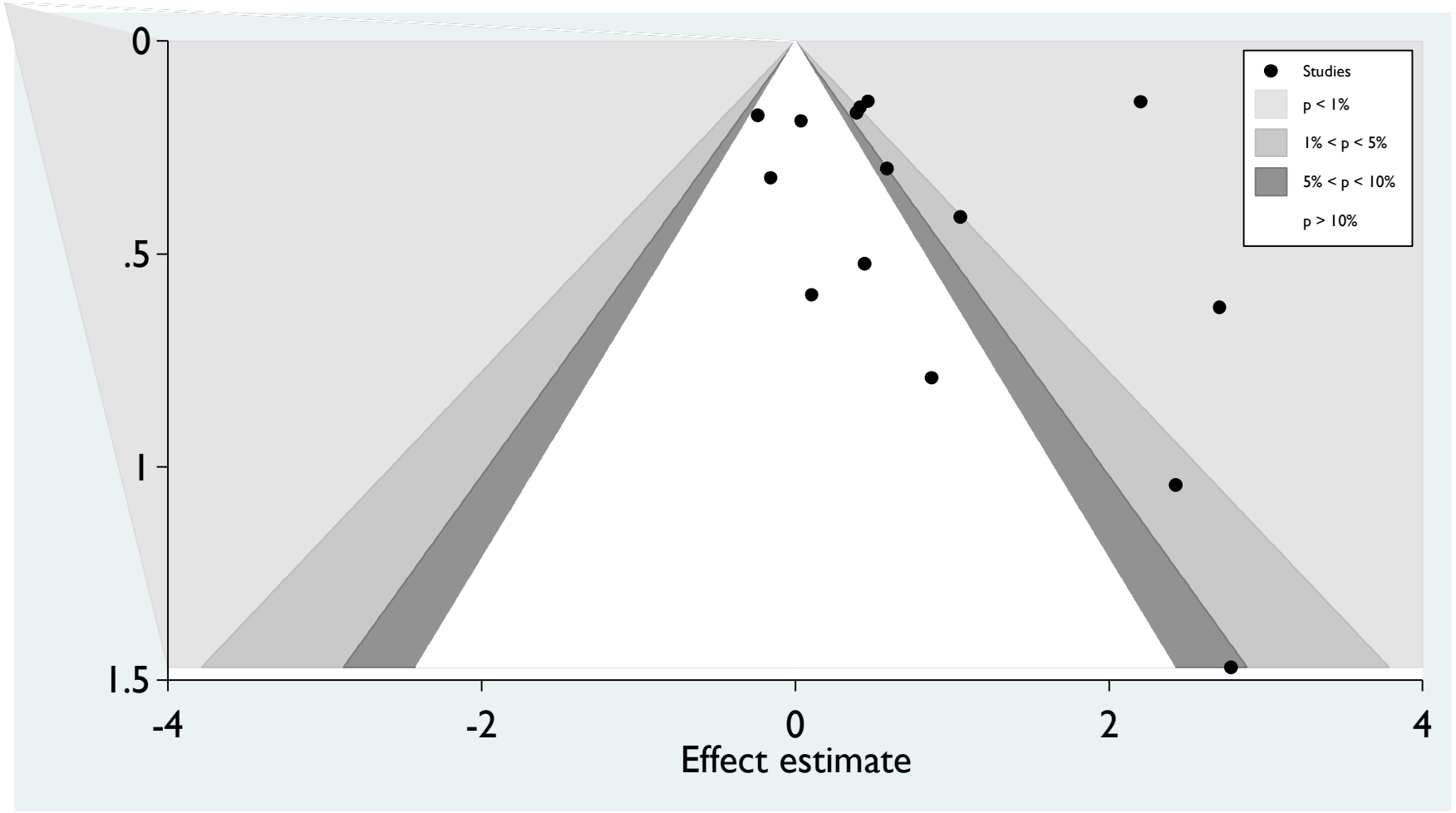
- "Fixed-effect" is unfortunate terminology
- Elsewhere in statistics, "fixed effects" means lots of free parameters
- Could consider $y_i \sim N(\mu_i, s_i^2)$ with these 3 models:

Model	Standard name	Better name?
$\mu_i = \mu$	Fixed-effect	Common effect
$\mu_i \sim N(\mu, \tau^2)$	Random-effects	Random effects
μ_i all separate	(not used)	Fixed effects

Other issues in (pairwise) meta-analysis

- Study quality
- Study-level covariates → "meta-regression"
- Publication bias
  - small trials more likely to be published if they show statistically significant effects?
  - see next

Exploring publication bias: “funnel plot”



Actually the data are more complicated …

study	dA	nA	dB	nB	dC	nC	dD	nD
1	9	140			23	140	10	138
2			11	78	12	85	29	170
3	79	702	77	694				
4	18	671	21	535				
5	8	116	19	146				
6	75	731			363	714		
7	2	106			9	205		
8	58	549			237	1561		
9	0	33			9	48		
10	3	100			31	98		
11	1	31			26	95		
12	6	39			17	77		
13	95	1107			134	1031		
14	15	187			35	504		
15	78	584			73	675		
16	69	1177			54	888		
17	64	642			107	761		
18	5	62			8	90		
19	20	234			34	237		
20	0	20					9	20
21			20	49	16	43		
22			7	66			32	127
23					12	76	20	74
24					9	55	3	26

24 trials compared 4 different interventions to help smokers quit:

A="No contact"

B="Self help"

C="Individual counselling"

D="Group counselling"

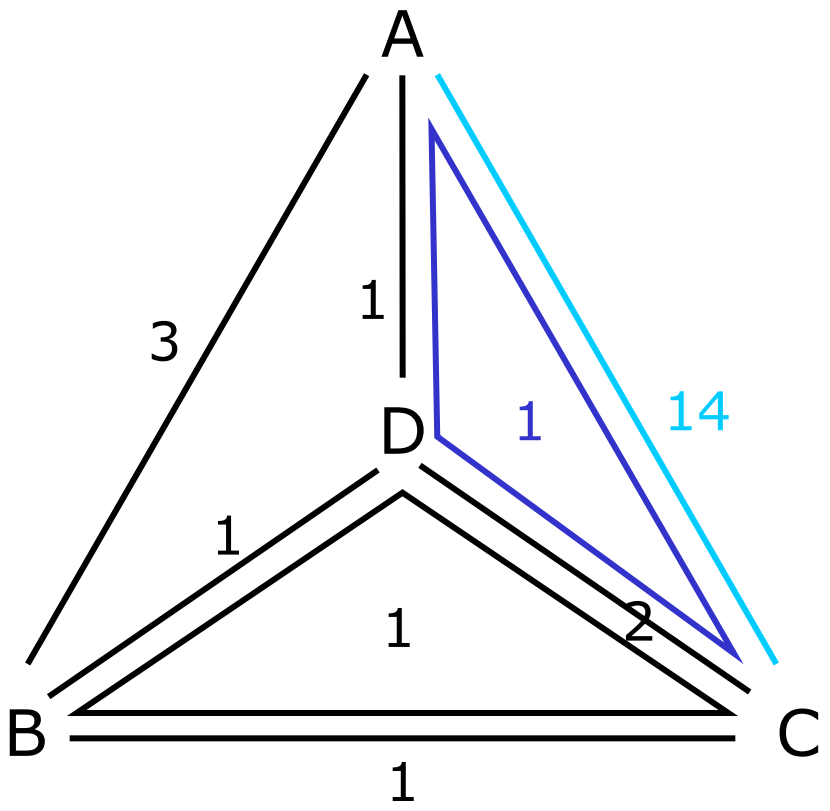
Actually the data are more complicated …

study	dA	nA	dB	nB	dC	nC	dD	nD
1	9	140			23	140	10	138
2			11	78	12	85	29	170
3	79	702	77	694				
4	18	671	21	535				
5	8	116	19	146				
6	75	731			363	714		
7	2	106			9	205		
8	58	549			237	1561		
9	0	33			9	48		
10	3	100			31	98		
11	1	31			26	95		
12	6	39			17	77		
13	95	1107			134	1031		
14	15	187			35	504		
15	78	584			73	675		
16	69	1177			54	888		
17	64	642			107	761		
18	5	62			8	90		
19	20	234			34	237		
20	0	20					9	20
21			20	49	16	43		
22			7	66			32	127
23					12	76	20	74
24					9	55	3	26

We have trials of different designs:

- A vs C vs D
- B vs C vs D
- A vs B (x3)
- A vs C (x14)
- A vs D
- B vs C
- B vs D
- C vs D (x2)

Evidence network: the smoking data



14 trials compared A with C

"design AC"

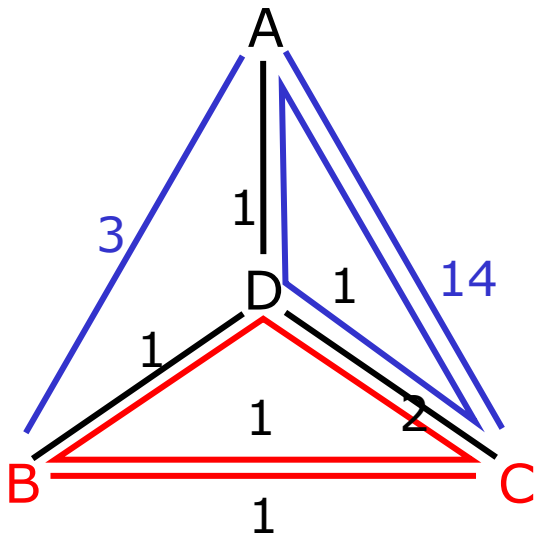
1 trial compared A, C and D

"design ACD"

etc.

Indirect comparisons

- Let's now focus on comparing B with C
- Evidence from B vs C and B vs C vs D trials is "direct evidence"
- Can we use indirect evidence to compare B with C?
  - e.g. combining A vs B trials with A vs C trials
- The maths is easy: using indirect evidence only, $\hat{\delta}_{BC} = \hat{\delta}_{AC} - \hat{\delta}_{AB}$ with variance $var(\hat{\delta}_{BC}) = var(\hat{\delta}_{AC}) + var(\hat{\delta}_{AB})$
  - where $\hat{\delta}_{BC}$ = effect of C compared to B, etc.
- But the assumptions are tricky: must assume the 3 designs (A vs B, A vs C, B vs C) are comparable



Bias in indirect comparisons (1)

- Suppose B and C are equally beneficial compared to A
  - B was trialled in the 1990s in a wide range of smokers
  - C was trialled in the 2000s in smokers who had failed in previous quit attempts
- So C is likely to show smaller benefit than B
- Quit rates might be:

Trial	A	B	C
1990s	10%	20%	20%
2000s	10%	15%	15%

- But what if all 3 interventions had been tried?
- Can regard C in 1990s and B in 2000s as "missing groups" – and data are missing not at random

Bias in indirect comparisons (2)

- If the overall event rates differ, then there are also problems with the scale on which intervention effects are measured. Suppose:

	Smoking quit rates			Comparison with A		
Trial	A	B	C	Risk difference	Risk ratio	Odds ratio
A vs. B	20%	30%		+10%	1.50	1.71
A vs. C	10%		18%	+8%	1.80	1.98
				B best	C best	C best

- Extrapolation problem – no easy answer

Network meta-analysis

- Despite these problems, I'll proceed to combine all the evidence – indirect and direct – in order to get our best estimates of the value of all the interventions
- This is called network meta-analysis
  - multiple treatments meta-analysis
  - mixed treatment comparisons
- Network meta-analysis addresses the real clinical question: which intervention is best for the patient?
  - may additionally require modelling covariates
- Much used by NICE (National Institute for Clinical Excellence) in comparing interventions

- See e.g. Salanti G, Higgins JP, Ades A, Ioannidis JP. Evaluation of networks of randomized trials. Statistical Methods in Medical Research 2008; 17: 279–301.

Aims of network meta-analysis

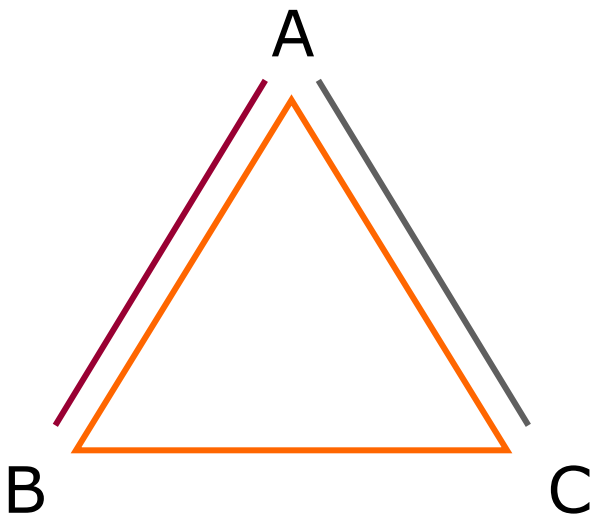
1. Use all the data & thus get
   - better estimates of treatment effects
   - opportunity to identify the best treatment
2. Assess whether the evidence is consistent
   - i.e. does the indirect evidence agree with the direct evidence?

The main statistical challenges are

- formulating and fitting models that allow for heterogeneity and inconsistency
- assessing inconsistency and (if found) finding ways to handle it

Less-statistical challenges include defining the scope of the problem: which treatments to include, what patient groups, what outcomes

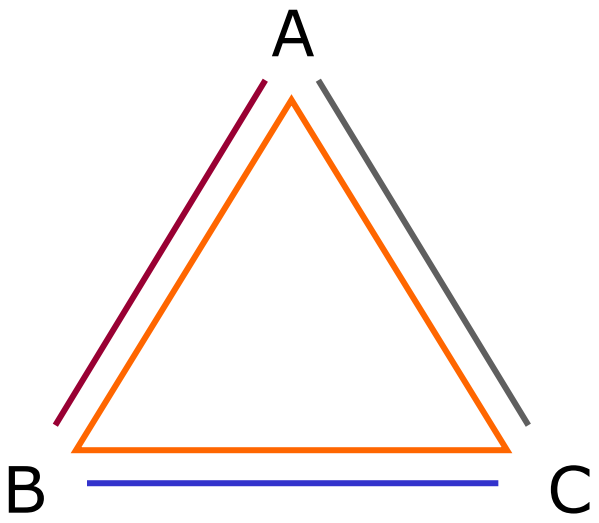
Models for network meta-analysis: consistency model (1)



True log odds in each group in trial *i*			
Design	*A*	*B*	*C*
ABC	α_i	$\alpha_i + \mu_{iB}$	$\alpha_i + \mu_{iC}$
AB	α_i	$\alpha_i + \mu_{iB}$	-
AC	α_i	-	$\alpha_i + \mu_{iC}$

- Trials have different baseline risks: no assumptions on α_i ("fixed effects" for trial)
- Between-trials model: $\mu_i = (\mu_{iB}, \mu_{iC}) \sim N(\mu, \Sigma)$
  - heterogeneity (variation between trials): $\Sigma \neq 0$ ("random effects" for treatment*trial)
- Consistency: μ_i has same mean $\mu = (\mu_B, \mu_C)$ in each design, where μ_B, μ_C = average effect of B, C vs A

Models for network meta-analysis: consistency model (2)



True log odds in each group in trial *i*			
Design	*A*	*B*	*C*
ABC	α_i	$\alpha_i + \mu_{iB}$	$\alpha_i + \mu_{iC}$
AB	α_i	$\alpha_i + \mu_{iB}$	-
AC	α_i	-	$\alpha_i + \mu_{iC}$
BC	α_i	$\alpha_i + \mu_{iB}$	$\alpha_i + \mu_{iC}$

- What about trials with no arm A?
- Easiest to regard arm A in BC trials as "missing data"
- Design BC still contributes to estimating $\mu_C - \mu_B$

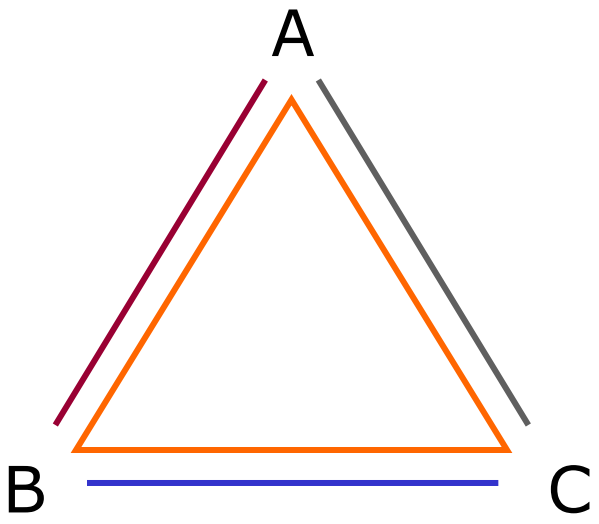
Full consistency model

- Notation:
  - interventions A (reference), B, C, D, …
  - effect of intervention J vs. A:
    - » estimate (from data) y_{iJ} → $y_i = (y_{iB}, y_{iC}, y_{iD}, \dots)$
    - » study-specific mean μ_{iJ} → $\mu_i = (\mu_{iB}, \mu_{iC}, \mu_{iD}, \dots)$
    - » overall mean μ_J → $\mu = (\mu_B, \mu_C, \mu_D, \dots)$
  - estimated variance-covariance matrix of y_i is S_i
- Within-trial model: $y_i \sim N(\mu_i, S_i)$
- Between-trials model: $\mu_i \sim N(\mu, \Sigma)$
- Doesn't matter that some y_{iJ} are missing

we'll come back to Σ later

- This is a *contrast-based* model, cf. an *arm-based* model for summary outcomes $y_i^* = (y_{iA}^*, y_{iB}^*, y_{iC}^*, y_{iD}^*, \dots)$

Inconsistency model



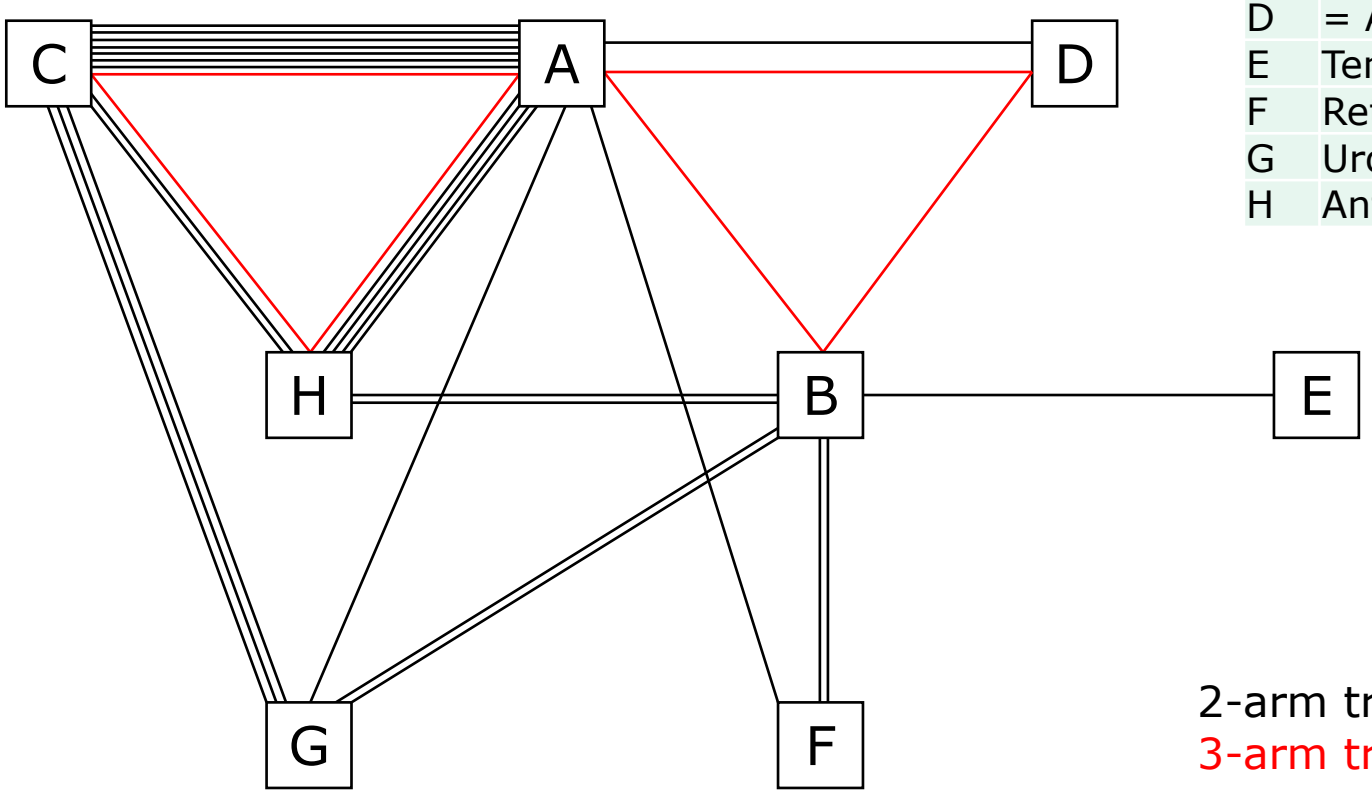
True log odds in each group in trial *i*			
Design	*A*	*B*	*C*
ABC	α_i	$\alpha_i + \mu_{iB}$	$\alpha_i + \mu_{iC}$
AB	α_i	$\alpha_i + \mu_{iB} + \omega_1$	-
AC	α_i	-	$\alpha_i + \mu_{iC} + \omega_2$
BC	α_i	$\alpha_i + \mu_{iB}$	$\alpha_i + \mu_{iC} + \omega_3$

- Inconsistency: treatment effects differ across designs
  - "design-by-treatment interaction"
  - regard the ω's as fixed effects

Higgins JPT, Jackson D, Barrett JL, Lu G, Ades AE, White IR. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. *Research Synthesis Methods* 2012; **3**: 98–110.

Heterogeneity

- Many networks are sparse
- e.g. a network meta-analysis of 8 thrombolytic treatments for AMI:

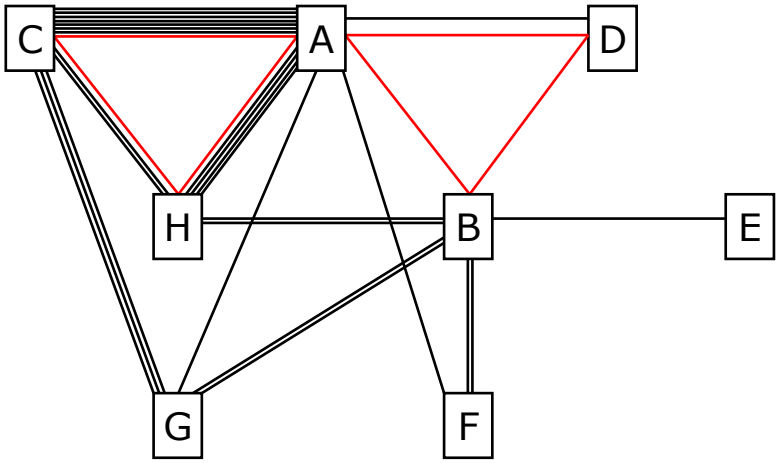


A	Streptokinase
B	Accelerated alteplase
C	Alteplase
D	= A + C
E	Tenecteplase
F	Reteplase
G	Urokinase
H	Anti-streptilase

2-arm trials
3-arm trials

Heterogeneity models

- Why does sparseness matter?
- Because between-trials variance $\Sigma = var(\mu_i)$ includes unidentified terms
  - e.g. $var(\mu_{iD} - \mu_{iE})$ and hence $\text{cov}(\mu_{iD}, \mu_{iE})$ isn't identified without a D-E trial
  - nor is $\text{cov}(\mu_{iB}, \mu_{iE})$ with only 1 B-E trial



- Need modelling assumptions for Σ
- Commonest is "common heterogeneity assumption":
  $var(\mu_{iJ} - \mu_{iI}) = \tau^2$ for all treatment pairs (I, J)

Network meta-analysis: standard model

- Let y_{di}^{IJ} be the estimated log odds ratio (or other measure) for treatment J vs. I in study i with design d
- Let s_{di}^{IJ} be its standard error
- Consistency model: $y_{di}^{IJ} \sim N(\mu_{di}^{IJ}, (s_{di}^{IJ})^2)$ ← approximation
  where $\mu_{di}^{IJ} \sim N(\delta^J - \delta^I, \tau^2)$
- δ^J is the mean effect of J vs. reference treatment A
  - we make sure that results don't depend on the choice of reference treatment
- τ^2 is the common heterogeneity (between-studies) variance
- Inconsistency model: $\mu_{di}^{IJ} \sim N(\delta^J - \delta^I + \omega_d^{IJ}, \tau^2)$
  - true treatment effects are different in every design

Network meta-analyses: estimation

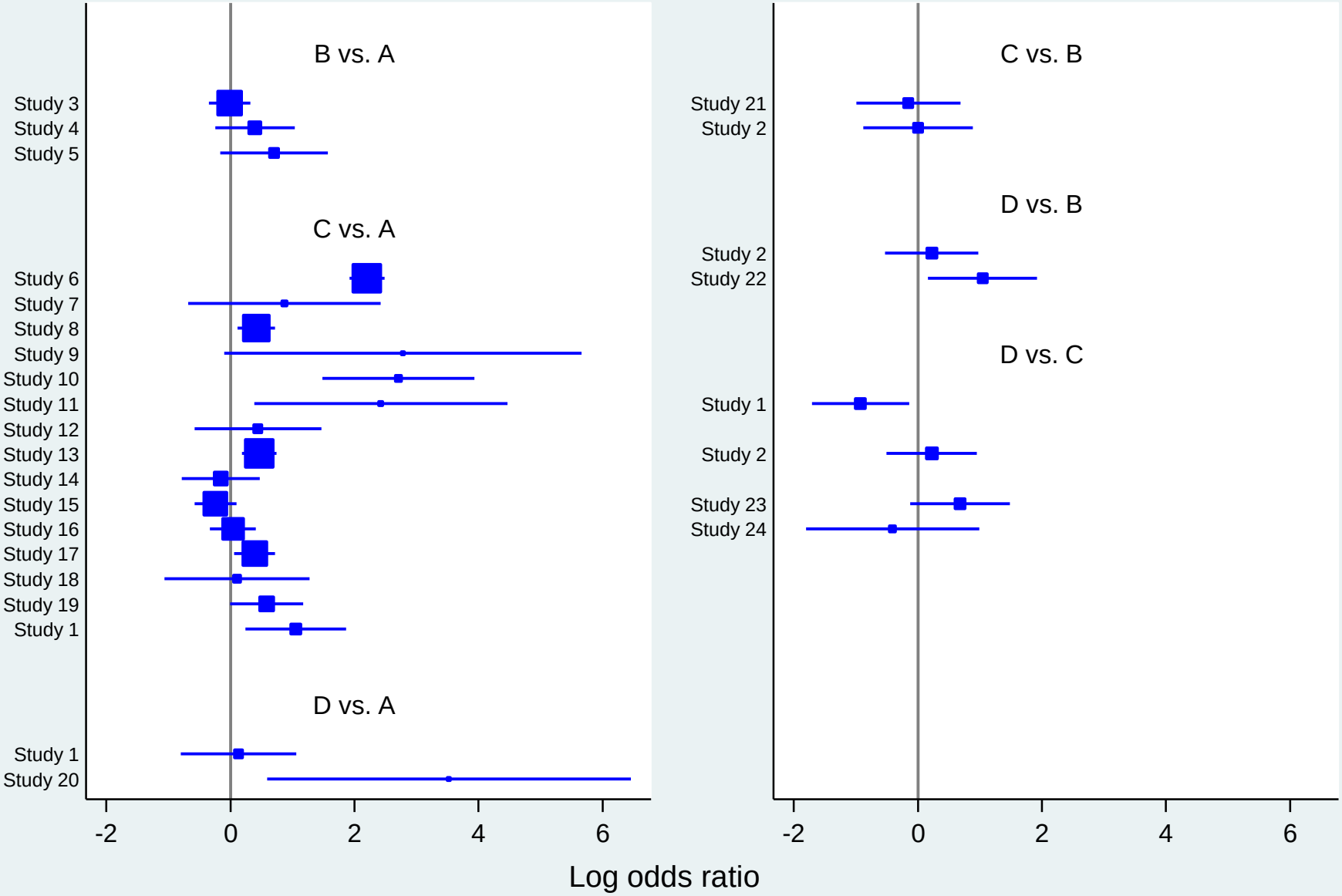
- In the past, the models have been fitted using WinBUGS
  - because frequentist alternatives have not been available
  - has made network meta-analysis difficult for non-statisticians
- Now, consistency and inconsistency models can be fitted using multivariate meta-analysis and multivariate meta-regression
- Trials without the reference intervention are handled
  - by a trial-specific baseline intervention (complicates code); or
  - by "augmenting" these trials with a very small reference arm (e.g. 0.0001 successes out of 0.001)

Network meta-analysis: multi-arm trials

- Multi-arm trials contribute >1 log odds ratio
  - need to allow for their covariance
  - mathematically straightforward but complicates programming
- With only 2-arm trials, we can fit models using standard "meta-regression"
- Multi-arm trials complicate this – need suitable data formats and multivariate analysis

Example analyses

Smoking network

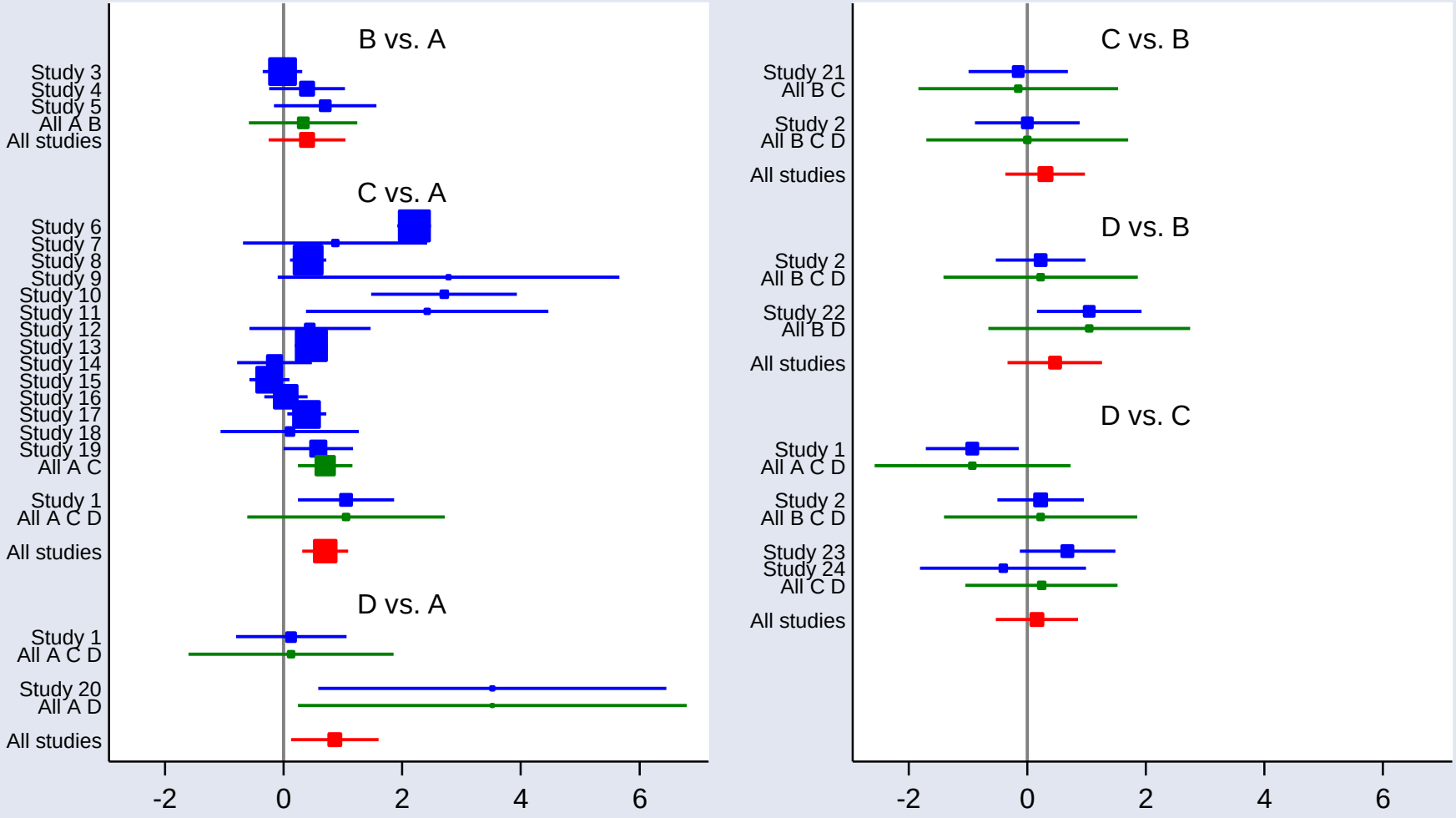


Smoking network: results

Intervention	Odds ratio (95% CI)	P(best)
A (no contact)	1 (reference)	0.0%
B (self help)	1.49 (0.78-2.85)	3.1%
C (individual counselling)	2.02 (1.37-2.98)	31.9%
D (group counselling)	2.38 (1.14-4.97)	65.0%

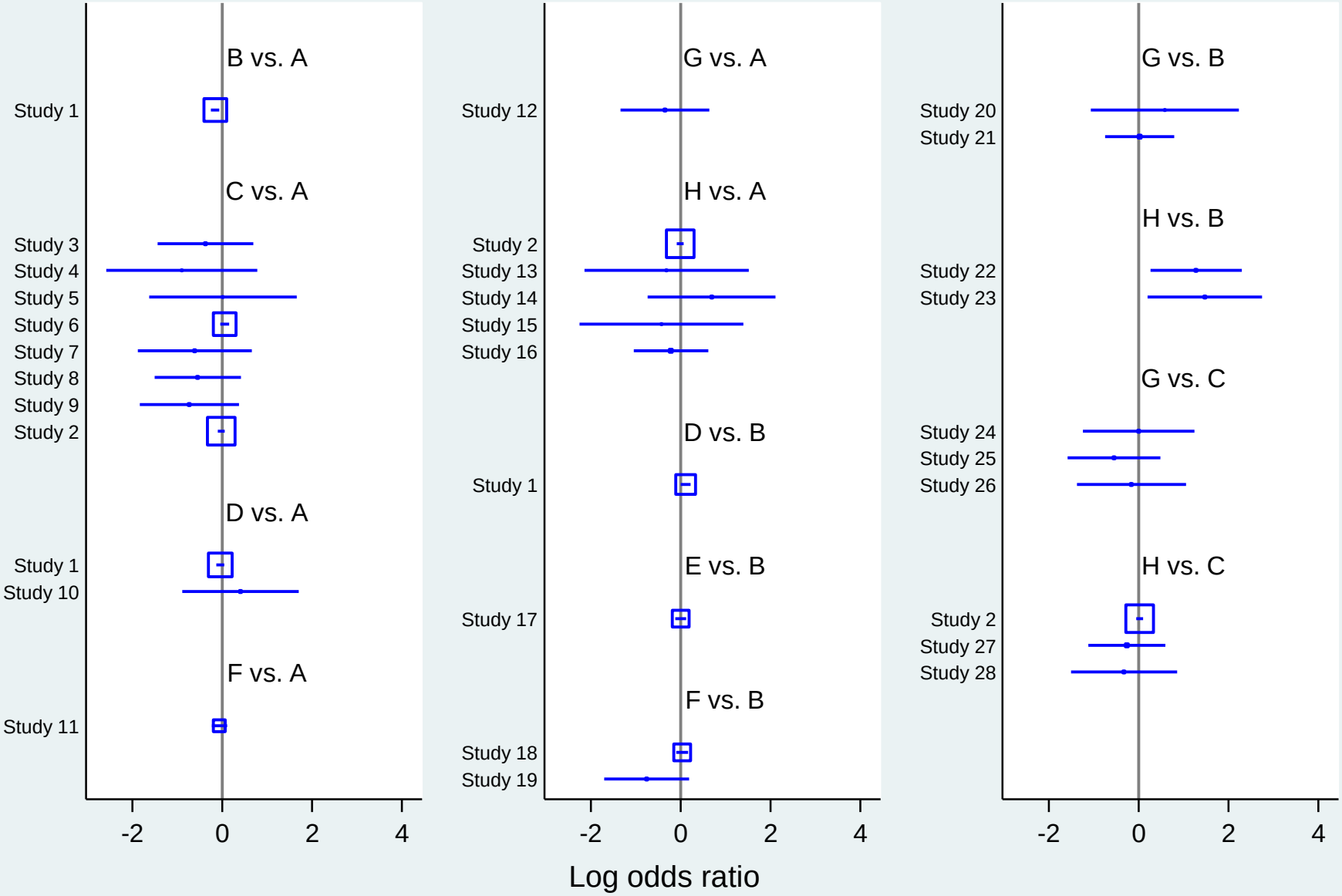
- Between-trials SD on log OR scale: $\hat{\tau}$ = 0.674 (large)
- D or C is likely to be best
- Test of inconsistency (Wald test in design-by-treatment interaction model): χ^2=5.11 on 7 df, p=0.65

Smoking network



Test of consistency: chi2(7)=5.11, P=0.646

Thrombolytics network

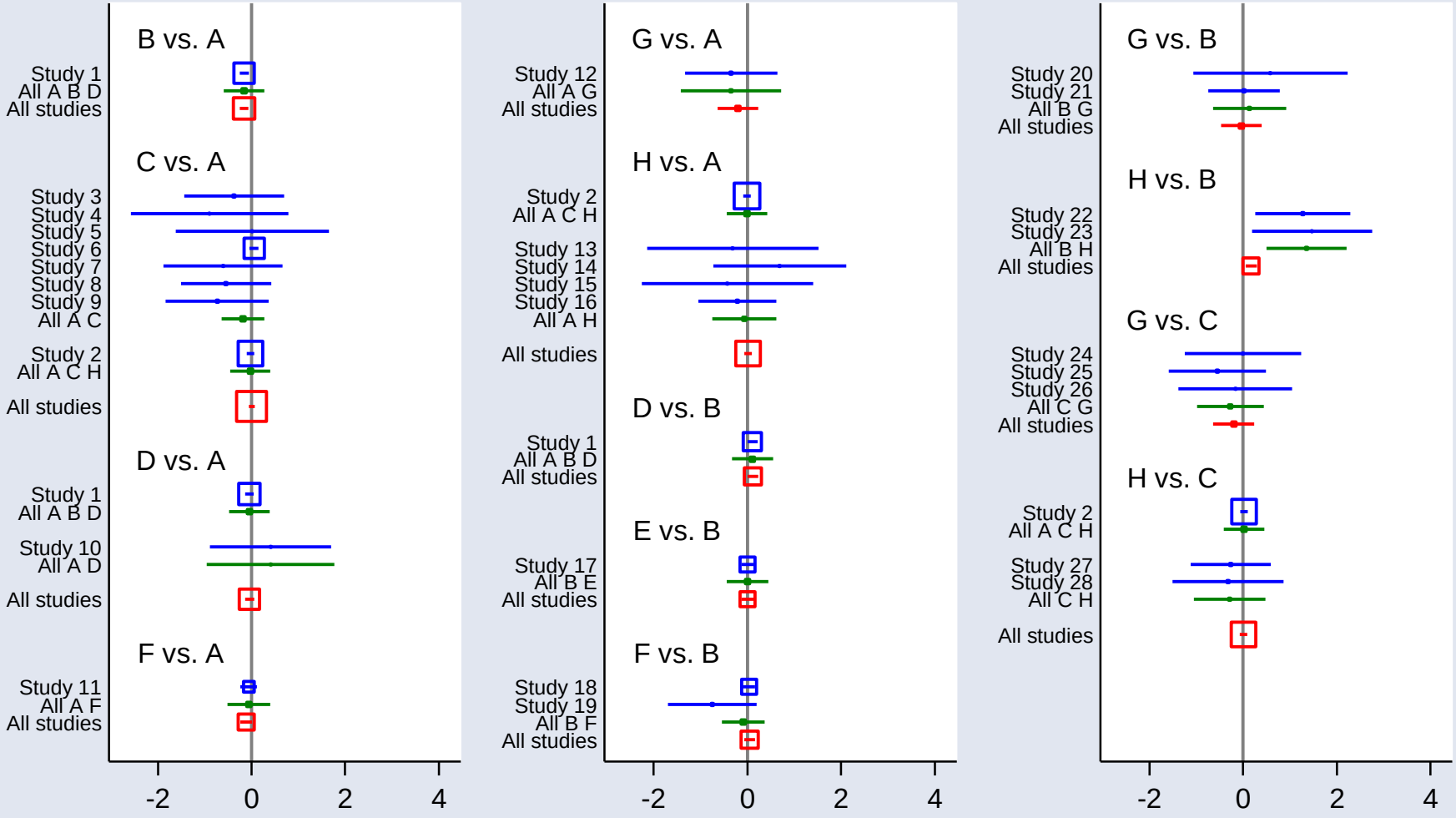


Thrombolytics network: results

Intervention		Odds ratio (95% CI)	P(best)
A	(streptokinase)	1 (reference)	0.0%
B	(accelerated alteplase)	0.85 (0.78-0.93)	19.3%
C	(alteplase)	1.00 (0.94-1.07)	0.1%
D	(=A+C)	0.96 (0.87-1.05)	0.5%
E	(tenecteplase)	0.86 (0.73-1.00)	22.4%
F	(reteplase)	0.89 (0.79-1.01)	6.8%
G	(urokinase)	0.82 (0.53-1.27)	50.9%
H	(anti-streptilase)	1.01 (0.94-1.10)	0.0%

- Between-trials SD on log OR scale: $\hat{\tau}$=0.015 (small)
- B, E or G is probably best
- Test of inconsistency: χ^2=8.61 on 8 df, p=0.38

Thrombolytics network



Test of consistency: chi2(8)=8.61, P=0.377

Some controversies

Controversies: what data to extract?

- Both my examples have summarised each study as a 2x2 table: successes/total in each arm
  - the standard in Cochrane systematic reviews
  - has the advantage of avoiding authors' tendency to "cherry-pick" the best results
- An alternative is to use the estimated treatment effect(s) in each trial's report
  - may be adjusted for prognostic factors (increases power in RCTs)
  - essential in observational studies (where we have to trust the authors to adjust for confounders)

Controversies: are published data enough?

- Published data have limitations
- The ideal is to get the raw data from all studies (individual participant data, IPD)
- IPD is especially valuable when exploring phenomena which tend to be inconsistently analysed / reported:
  - interactions (subgroup effects)
  - adjustment for confounding in observational studies
- But it is much slower and much more expensive…

Controversies: the common heterogeneity model

- The common heterogeneity model assigns heterogeneity even when a contrast is estimated in a single study (e.g. B-E in thrombolytics) – must be good.
- But homogeneous parts of the network may become "contaminated" by more heterogeneous parts.
  - could in principle have:

Pairwise, B vs A:
OR = 0.8 (95% CI, 0.7-0.9)
$\hat{\tau}^2 = 0$

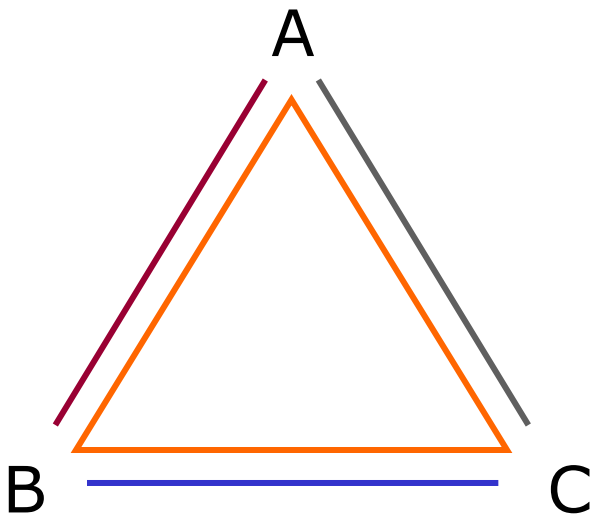
Pairwise, C vs A:
OR = 1 (95% CI, 0.5-2.0)
$\hat{\tau}^2 = 2$

Network, B vs A:
OR = 0.8 (95% CI, 0.5-1.3)
$\hat{\tau}^2 = 1$ "unfair"?!

Network, C vs A:
OR = 1 (95% CI, 0.6-1.7)
$\hat{\tau}^2 = 1$

Ideally want a model with τ^2 exchangeable across comparisons

Controversies: defining inconsistency



True log odds in each group in trial *i*			
Design	*A*	*B*	*C*
ABC	α_i	$\alpha_i + \mu_{iB}$	$\alpha_i + \mu_{iC}$
AB	α_i	$\alpha_i + \mu_{iB} + \omega_1$	-
AC	α_i	-	$\alpha_i + \mu_{iC} + \omega_2$
BC	α_i	$\alpha_i + \mu_{iB}$	$\alpha_i + \mu_{iC} + \omega_3$

- Our "design-by-treatment interaction model" has 3 inconsistency parameters
- Intuitively, should be only one per "loop"
  - but we haven't found a sensible way to define it
  - model of Lu & Ades (2006) isn't symmetrical with multi-arm trials

Controversies: testing for inconsistency

- Test for inconsistency is a global test on many degrees of freedom
  - likely to have low power in practice
- Can we use substantive knowledge to define more targetted tests?
- Should we accept that inconsistency is present even when test is non-significant?

Controversies: allowing for inconsistency

What do we do if we decide we have inconsistency?

Obviously we first try to explain it – "did the A-B trials recruit more severely ill patients?", etc.

If we fail, then do we

- refuse to draw conclusions about treatment comparisons? (maybe we asked the wrong question?)
- infer treatment comparisons from the consistency model, with appropriate caveats?
- treat inconsistency as another random effect?
  - we've proposed a model for this (Jackson et al, under review)
  - it inflates std errors to "account for" inconsistency
  - just as the standard random-effects model inflates std errors to "account for" heterogeneity.

Controversies: estimation

- Network meta-analysis was in the past done using Bayesian methods (1-stage analysis, arm-based model, full binomial likelihood)
  - WinBUGS
  - rank treatments, give p(treatment C is best) etc.
- I've proposed frequentist methods based on multivariate meta-analysis (2-stage analysis, contrast-based model, Normal approximation to the likelihood)
  - faster and more accessible
  - don't allow well for sparse binary data (e.g. smoking trial 9: 0/33 vs 9/48)
- Next slide compares the methods in the smoking data…

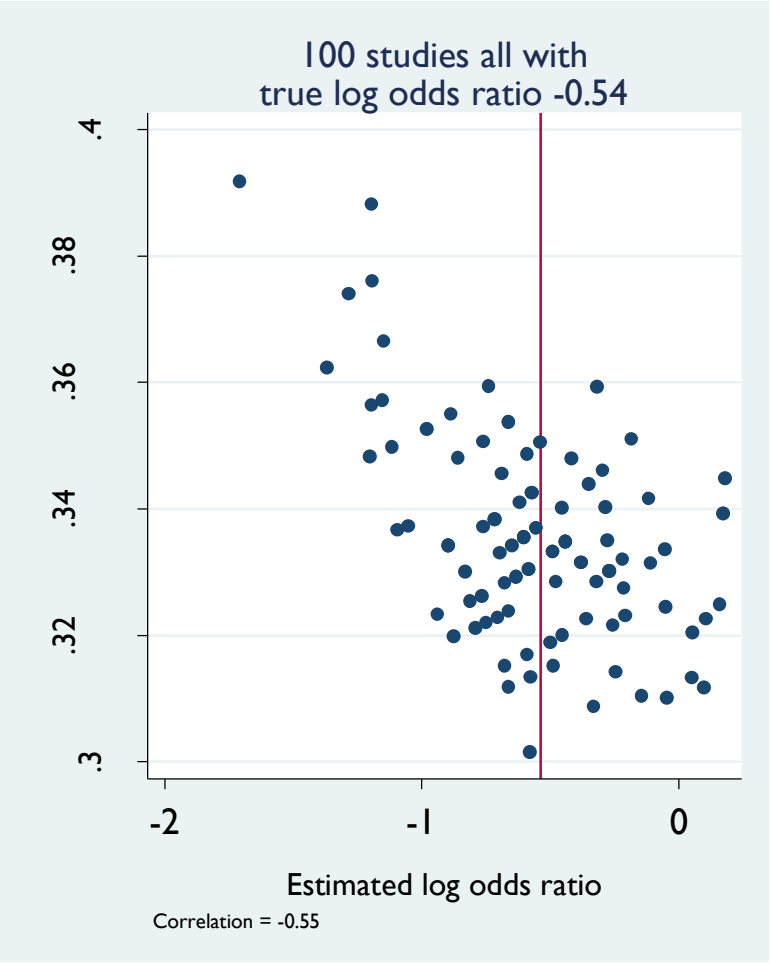
Smoking network: method comparison

log OR: treatment vs. A	Two-stage frequentist			One-stage Bayesian		
	Est.	std err	P(best)	Est.	std err	P(best)
A (ref)	-	-	0.0%	-	-	0.0%
B	0.398	0.331	3.1%	0.494	0.399	5.7%
C	0.702	0.199	31.9%	0.844	0.236	23.5%
D	0.866	0.376	65.0%	1.101	0.437	70.8%
τ: between trials SD	0.674	0.140		0.731	.	

- One-stage Bayesian analysis taken from Lu & Ades, JASA 2006; **101**: 447–459.
- Differences between methods are mainly attributable to the approximation in the two-stage method

Why is the two-stage method inaccurate?

- Because the standard error is correlated with the point estimate
  - more extreme estimates are down-weighted, causing bias towards null
- Problem appears to be restricted to binary data

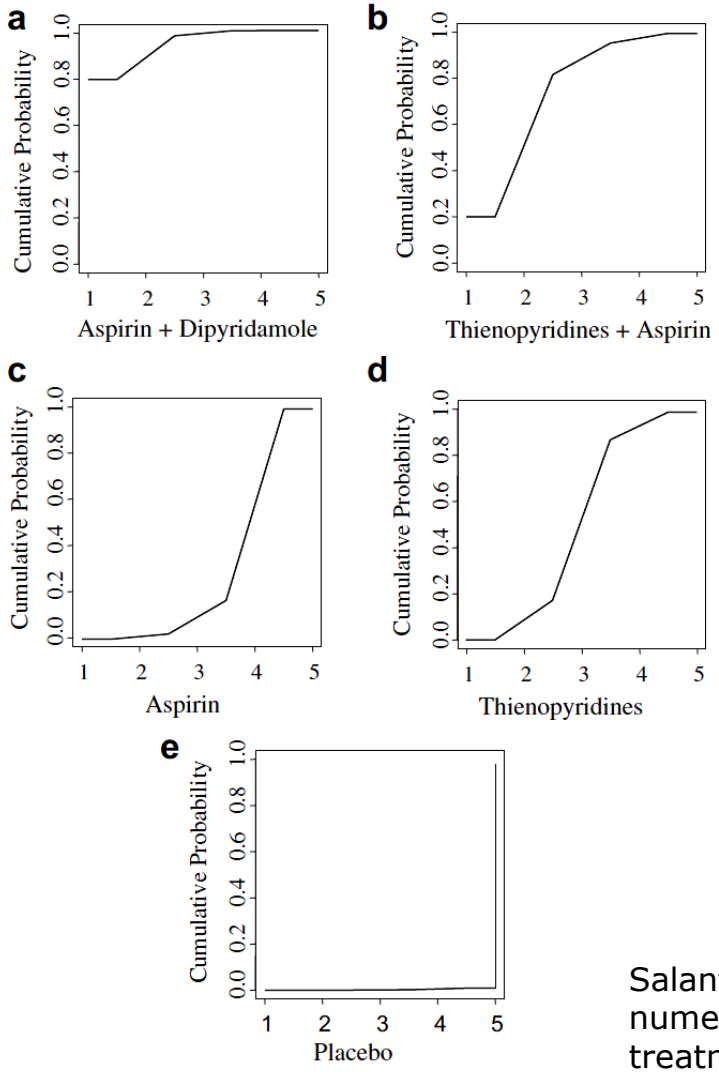


A frequentist one-stage method for binary data?

- Should be able to fit a generalised linear mixed model (Stata **`melogit`**)
  - random effect for study*treatment interaction
  - (± fixed or random effect for design*treatment interaction)
- How do we handle main effect of study?
  - fixed effect? → one parameter per study → may underestimate heterogeneity variance & std error
  - random effect? but then results are contaminated by between-study information
  - eliminate it by conditioning on study margins? may be ideal but computationally difficult

    Stijnen T, Hamza TH, Özdemir P. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Statistics in Medicine* 2010; **29**: 3046–3067.

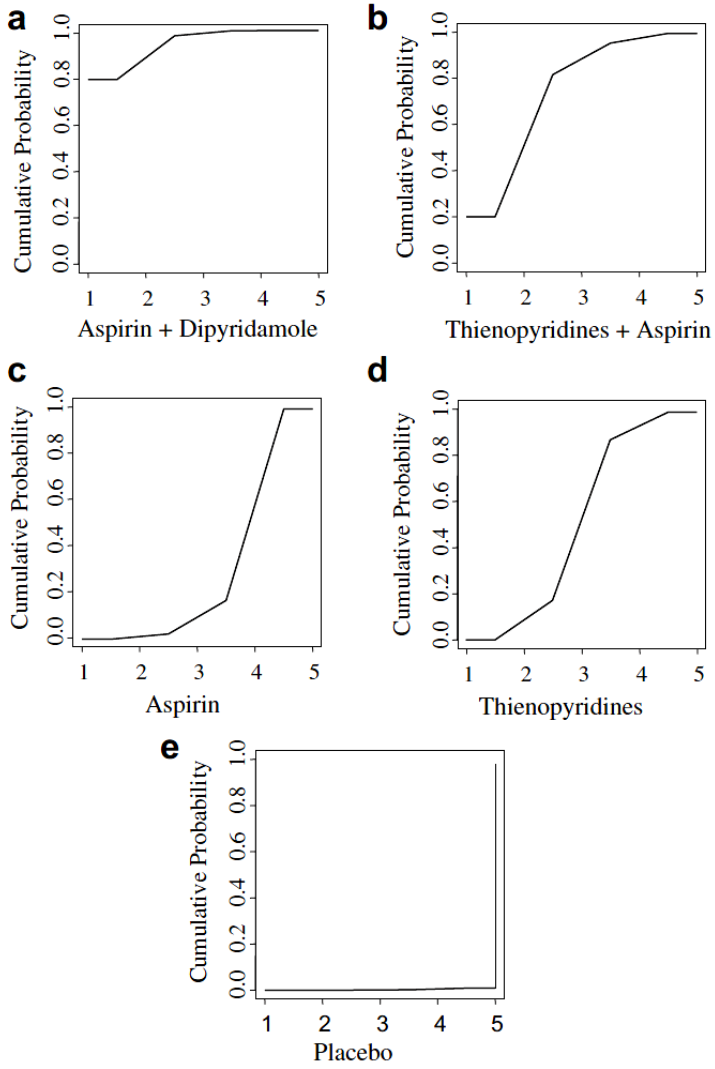
Controversies: ranks



- Rankogram displays the posterior probability that each treatment is
  - ranked 1 (the best), ≤2, ≤3 etc.
- The argument is
  - a clinician wants to use the best treatment, so we maximise their chances
  - if best treatment isn't available, want to maximise their chance of getting the 2nd best

Salanti G, Ades A, Ioannidis J. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *Journal of Clinical Epidemiology* 2011; **64**: 163–171.

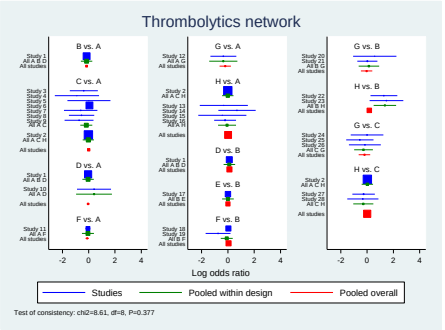
Controversies: ranks



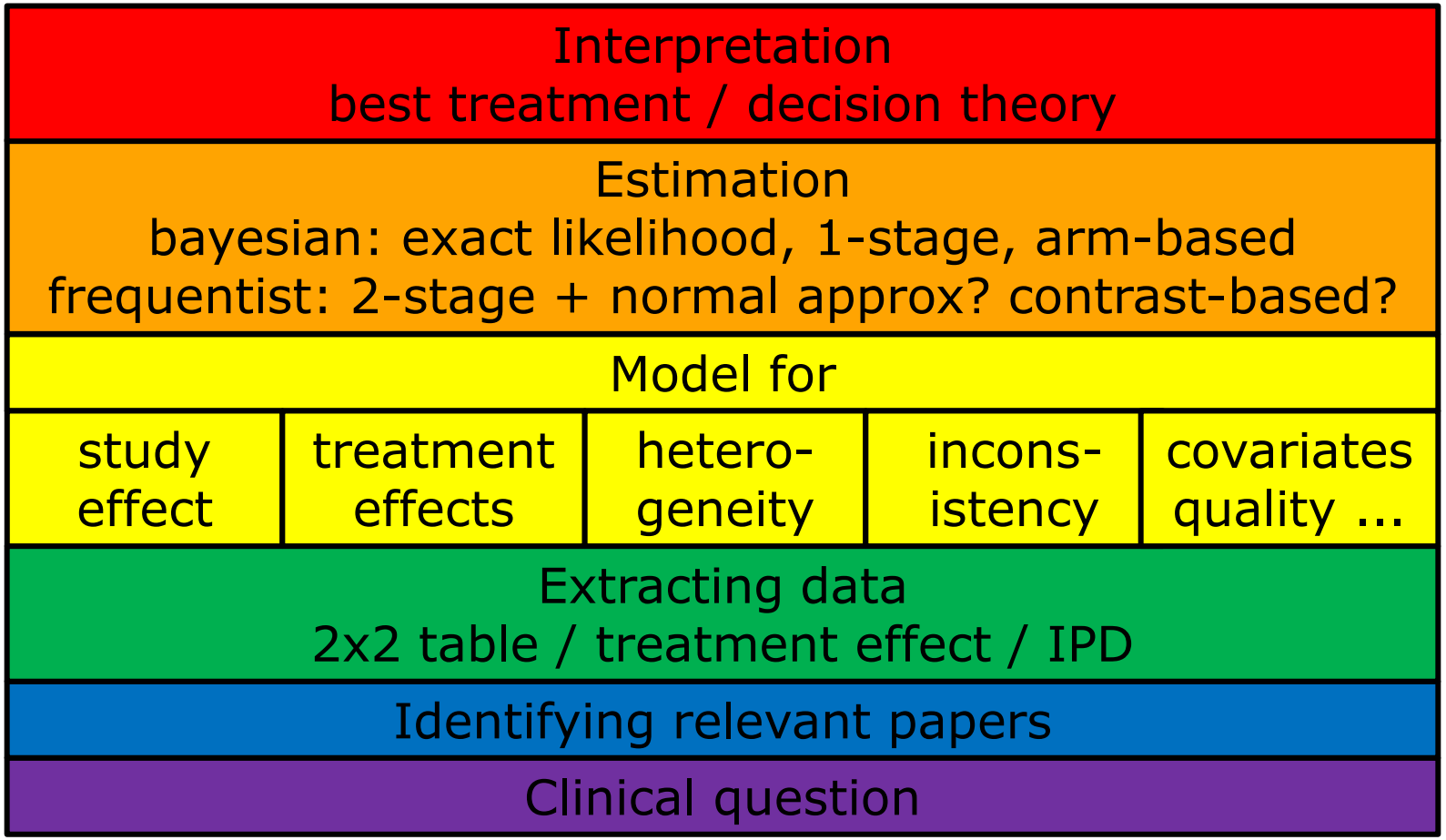
- But is this the right way to choose a treatment?
- Decision theory suggests choosing the treatment which maximises the expected utility, e.g. p(quit smoking | treatment)
  - would take account of uncertainty
  - best would depend on "baseline risk" p(quit smoking | no treatment)

Resources

- Bayesian approach using WinBUGS: the NICE decision support unit has a series of useful documents at http://www.nicedsu.org.uk/Evidence-Synthesis-TSD-series%282391675%29.htm
- Frequentist approach using Stata: I have written **`network`**, a suite of programs to read in data, fit consistency and inconsistency models, and graph results
  - the consistency and inconsistency models are expressed as multivariate meta-analyses / meta-regressions and fitted using my **`mvmeta`**
  - **`net from http://www.mrc-bsu.cam.ac.uk/IW_Stata/`**
- Frequentist approach using R: Antonio Gasparrini has written an R counterpart to mvmeta



Network meta-analysis: summary



Thanks to Julian Higgins (U of Bristol), Dan Jackson (BSU) and Jessica Barrett (U of Cambridge) who worked with me on this.