Loopy comparisons:
when can more than two treatments be ranked?

Thomas Lumley,

(with Bruce Psaty, Gina Schellenbaum,

Noel Weiss, Curt Furberg;

Dan Gillen, Scott Emerson)

Three-way comparisons

- For three treatments there are three pairwise comparisons, with $2^3=8$ possible results
- There are only 3!=6 ways to order three treatments, so 2 of the sets of results are not consistent with any ordering because they violate transitivity.

- Can these occur in real statistics, and what are the implications?

Causes of non-transitivity

- Pairwise comparisons can be inconsistent because they are conducted on different groups of patients. This is the problem of indirect comparisons, addressed in part by network meta-analysis.

- Pairwise comparisons can be inconsistent because the statistical tests are optimized for comparing two treatments, not for comparisons among larger sets. A surprisingly large range of statistical tests are non-transitive.

Problems with the data

Motivating data

- High blood pressure is common in western countries and leads to higher risk of stroke, coronary disease, and congestive heart failure.
- Blood pressure is naturally regulated through
  - changing volume of blood
  - constricting or relaxing blood vessels
  - altering how hard the heart muscle works
- Drugs exist that affect each of these mechanisms, or combinations of them

Motivating data

Large prevention trials with real clinical outcomes exist comparing (at least)

- Placebo or no treatment
- Diuretics (low & high dose, thiazide and other)
- β-blockers
- α-blockers
- Angiotensin Converting Enzyme (ACE) inhibitors
- Calcium channel blockers (dihydropyridine and other)
- Angiotensin II receptor blockers

Criteria other than cost and effectiveness in prevention are important for a minority of patients

Why indirect comparisons?

- There is more than one possible comparison
  - vs placebo
  - vs standard treatment
  - vs another standard treatment…
- The most interesting comparisons may not be done
  - Regulatory requirements may force comparisons to placebo
  - Strong incentives to compare new drugs to the least effective standard treatment in active control trials
  - There just isn't enough time and money

NY Times opinion

…For the most part, drugs in this country are not tested against other drugs in the same class. Instead they are tested against a placebo, and if shown to be comparatively safe and effective are approved for marketing. That leaves both patients and their doctors uncertain which approved drugs are better than their competitors and whether high-priced drugs warrant their added cost compared with lower-cost alternatives.

Editorial, Sunday 2003-11-16

Statistical model

- Individual-level mean model

$$\log P(\text{event}) = \alpha_{trial} + \beta_{drug}$$

- Trial-level mean model

$$E[Y_{drug1,drug\ 2,trial}] = \log \text{Relative Risk} = \beta_{drug1} - \beta_{drug2}$$

The challenge is specifying the error terms

Why not indirect comparisons?

- If A is equivalent to B, and B is equivalent to C, in randomised comparisons, is A equivalent to C?
  - Not necessarily. Equivalence is only up to some tolerance (eg 10-15%), so it isn't an equivalence relation
- If A is better than B, and B is better than C, is A better than C?
  - Not necessarily. A and B may have been compared in people for whom B didn't work, and B and C in people for whom C didn't work.

Regulators' opinion

The International Council on Harmonisation says

> Placebo-controlled trials lacking an active control give little useful information about comparative effectiveness, information that is of interest and importance in many circumstances. Such information cannot reliably be obtained from cross-study comparisons, as the conditions of the studies may have been quite different.
>
> (ICH E10 2.7.1.4)

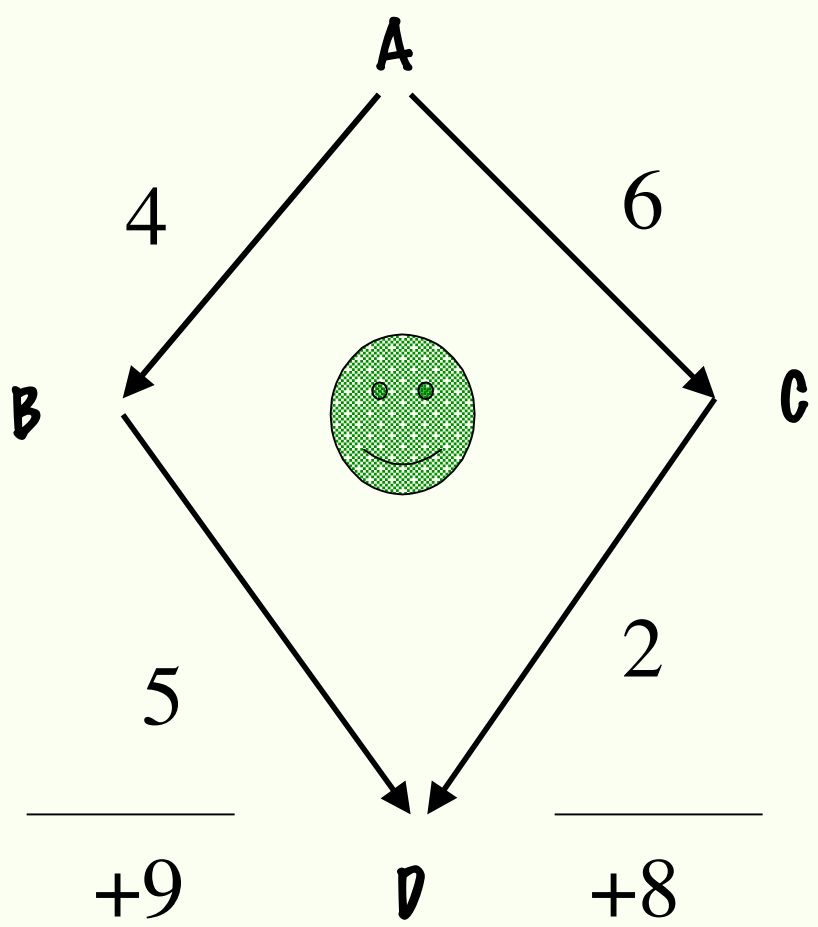
Objective

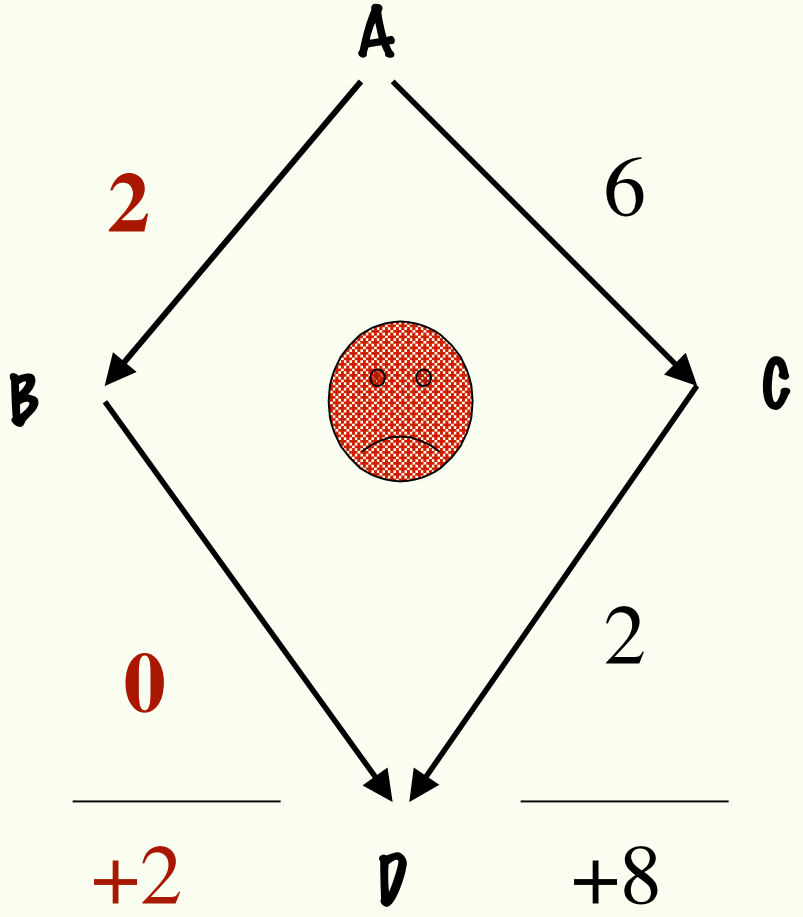
- Indirect comparisons are often reliable but sometimes unreliable (Bucher H. J Clin Epi 1997; Song F. BMJ 2003)
- It is hard to guess which case is which
- Estimation using indirect comparisons is easy. The challenge is ensuring that estimation fails when it should
- Need a data-based assessment of trial consistency.

Networks of trials

- Data can be represented as a graph with nodes for each treatment and (directed) edges for each trial.
- Each edge is labelled with the estimated treatment difference in that trial.
- The analysis will depend only on the randomised estimates of treatment differences

Networks of trials





Inconsistency between trials

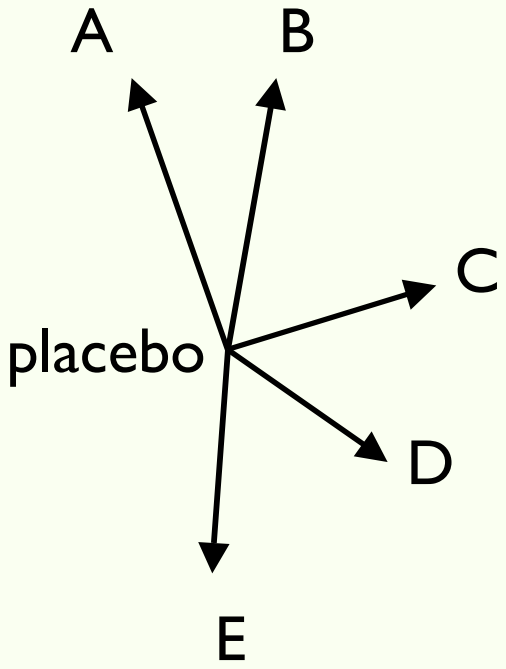
Compute the sum of treatment effects $Y_{i,i-1}$ around any loop in the graph. The result should be zero

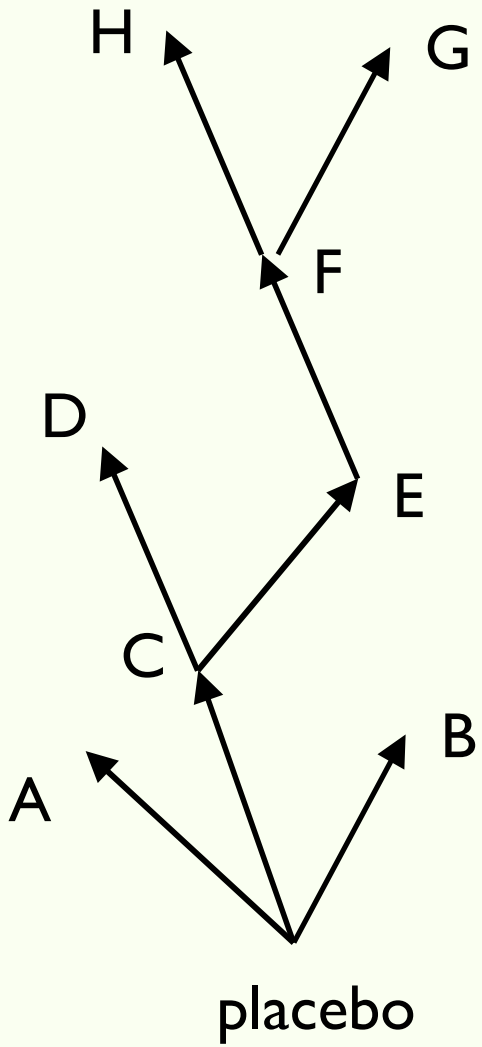
$$S = \sum_i Y_{i,i-1} \sim N\left(0, \sum_i \sigma^2_{i,i-1}\right)$$

If S is large compared to its variance we have evidence of inconsistency. Large inconsistencies rule out a meta-analysis, small inconsistencies should add uncertainty to the results

Inconsistency and loops

- The inconsistency can be estimated only for loops, so more loops allow better diagnosis of consistency.
- Consistency cannot be assessed for a `star' design comparing everything to placebo, or for a `ladder' design where new treatments are always compared to current standard.
- It is not always possible to isolate which trials are responsible for inconsistency.


A
B
C
placebo
D
E


H
G
F
D
E
C
B
A
placebo

Statistical model

- Trial results for kth trial comparing treatments i and j is Y_{ijk} with internally estimated variance σ^2_{ijk}.
- Sampling errors and heterogeneity

$$\epsilon_{ijk} \sim N(0, a(\sigma_{ijk} + b)^2)$$

- Random effects $\xi_{ij} \sim N(0, \omega^2)$ model inconsistency between treatment pairs, ω is called **incoherence**

$$Y_{ijk} = \mu_i - \mu_j + \xi_{ij} + \epsilon_{ijk}$$

Estimation

- A linear mixed model with random intercept for treatment pair *(i,j)*
- Only differences μ_i-μ_j are identifiable, so pick a reference drug and drop one column of design matrix.
- Fit with `lme()` in R or S-PLUS
- Incoherence is reported as the random intercept standard deviation: identifiable only when loops are present.
- Confidence intervals for fixed effects incorporate incoherence (and heterogeneity)

Interpretation

- Estimating functions for fixed effects in a linear mixed model are weighted averages of differences in responses
  - Treatment difference is a weighted average of sums along all paths connecting the treatments
  - A long path is always downweighted relative to a direct comparison, because the incoherence contributes for each link in the path.
  - Estimation is not possible without closed loops in the graph.
  - If incoherence is present, standard errors will not go to zero in the absence of direct comparisons for that contrast.

Clinical trials: 1967 to 1985

Hi-dose diuretics

Placebo

Alpha-blockers

Beta-blockers

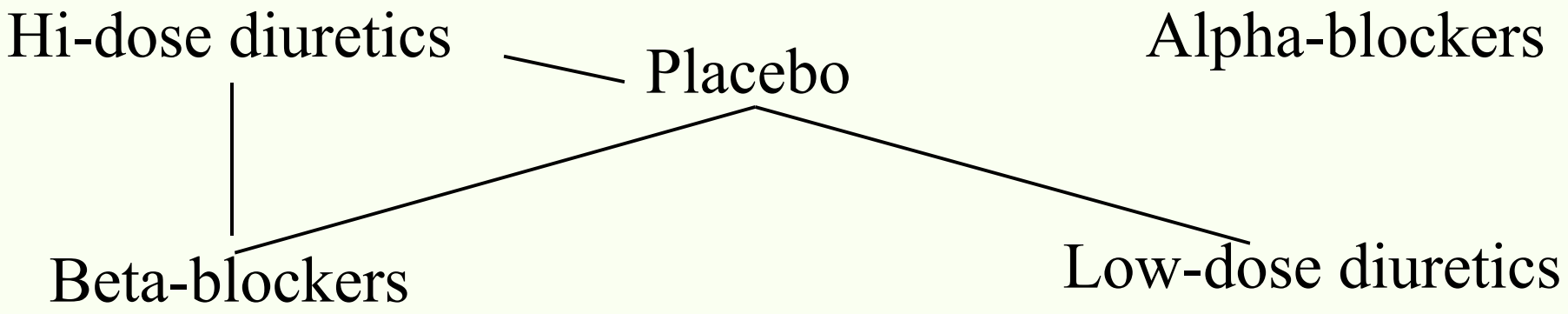
Low-dose diuretics

Calcium-channel blockers

Angiotensin converting enzyme inhibitors

Angiotensin receptor blockers

Clinical trials: 1985 to 1992

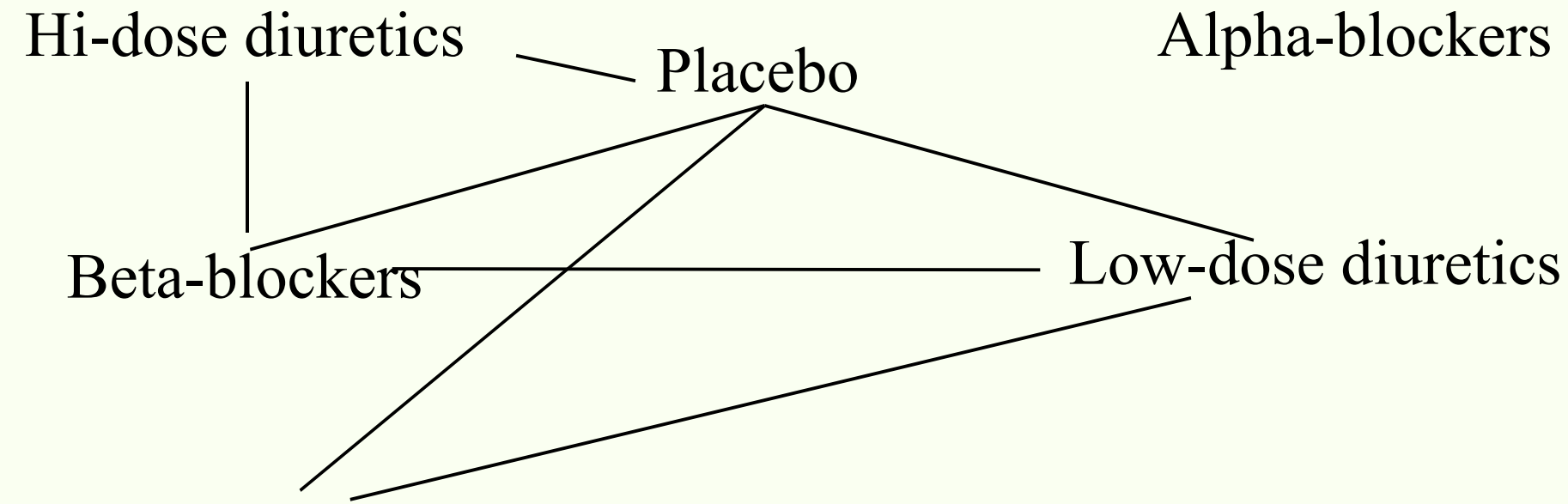


Calcium-channel blockers

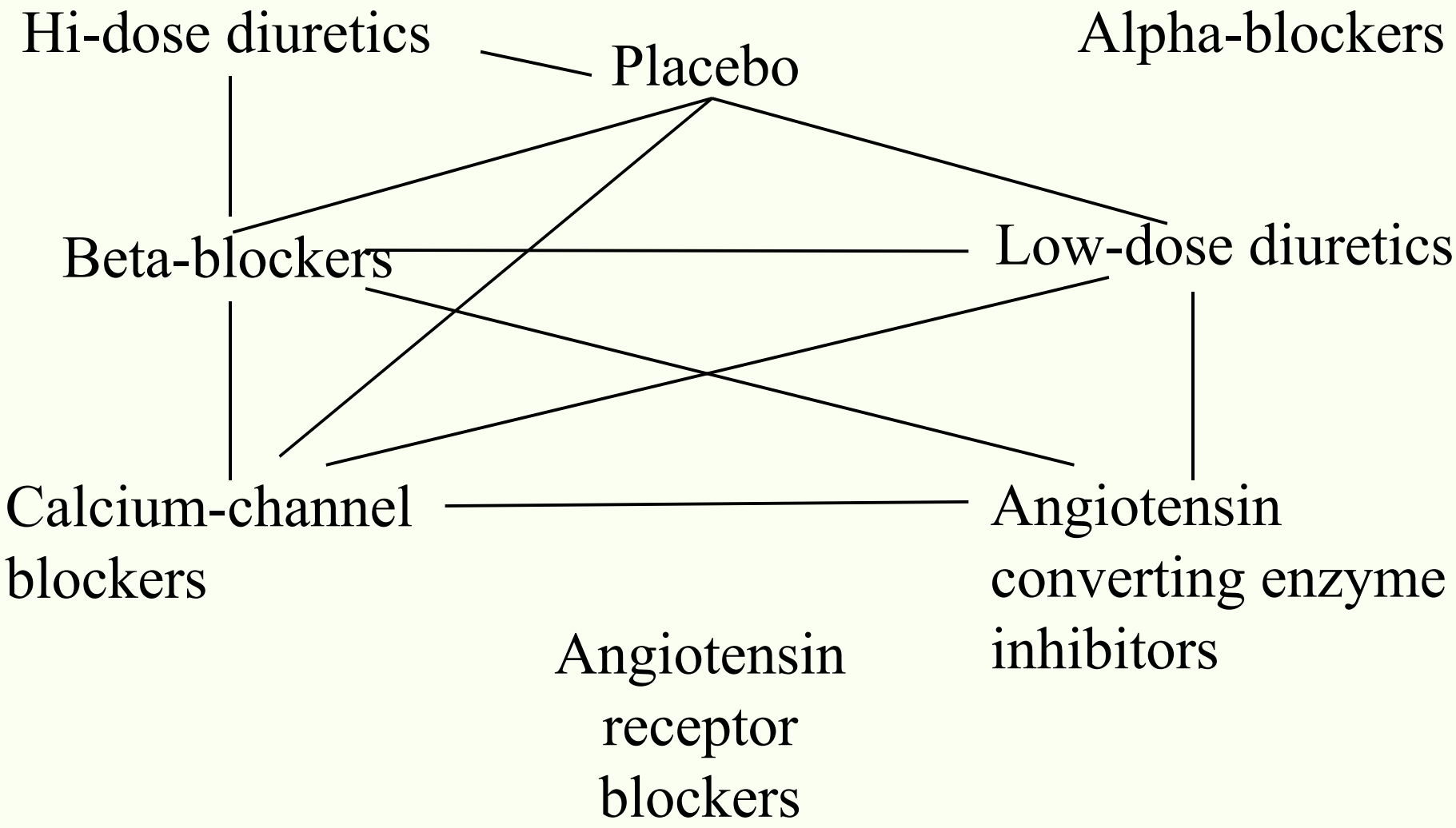
Angiotensin receptor blockers

Angiotensin converting enzyme inhibitors

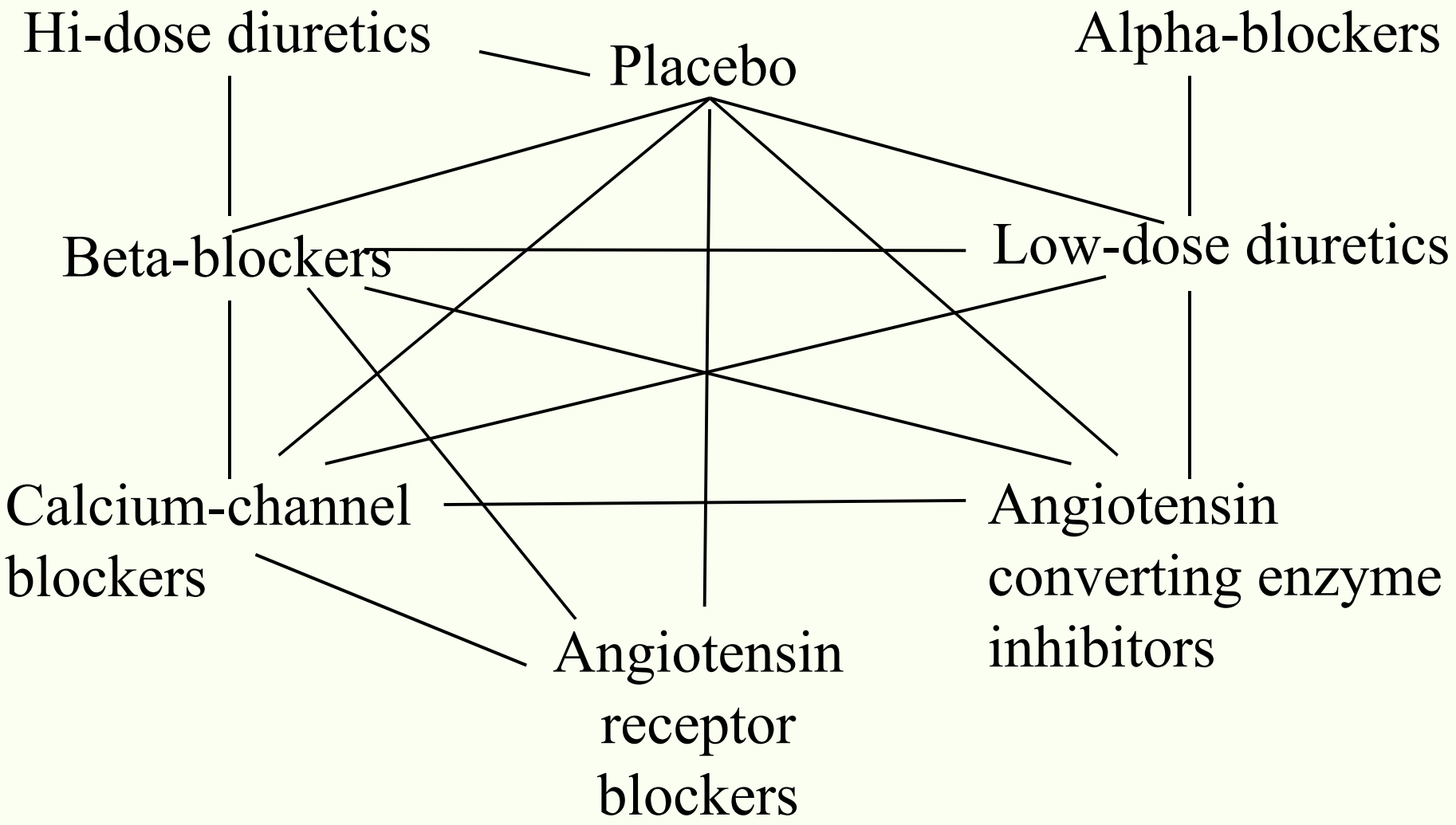
Clinical trials: 1992 to 1997



Clinical trials: 1998 to 1999



Clinical trials: 2000 to 2002

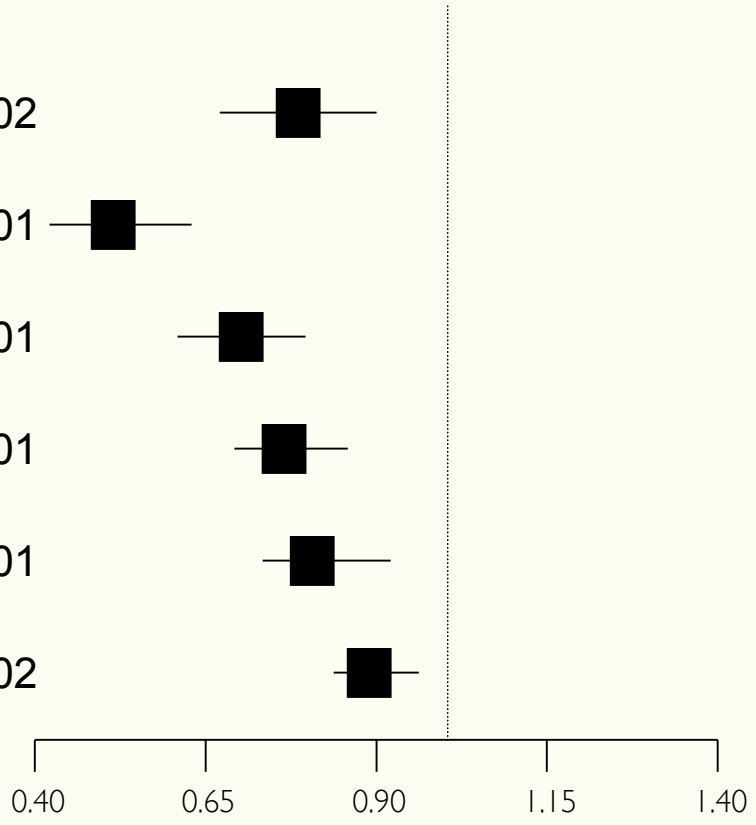


Example: antihypertensives

- Previous meta-analyses either looked at pairwise direct comparisons (losing information), or grouped drugs into `old' and `new' (losing plausibility and relevance)
- Three sets of results from our meta-analysis:
  - Diuretics vs placebo (largely direct)
  - Diuretics vs calcium channel blockers (substantially indirect)
  - Diuretics vs α-blockers (a single large trial)

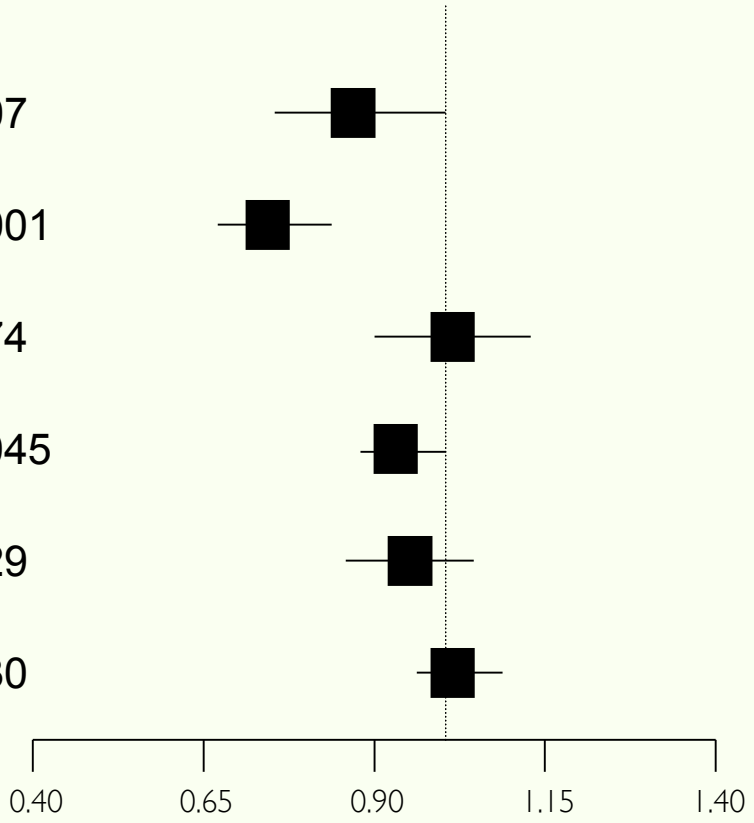
Diuretics vs placebo

Outcome	RR	95% CI	p
CHD	0.79	0.69-0.92	0.002
Heart failure	0.51	0.42-0.62	<0.001
Stroke	0.71	0.63-0.81	<0.001
CVD events	0.76	0.69-0.83	<0.001
CVD mortality	0.81	0.73-0.92	0.001
Total mortality	0.90	0.84-0.96	0.002



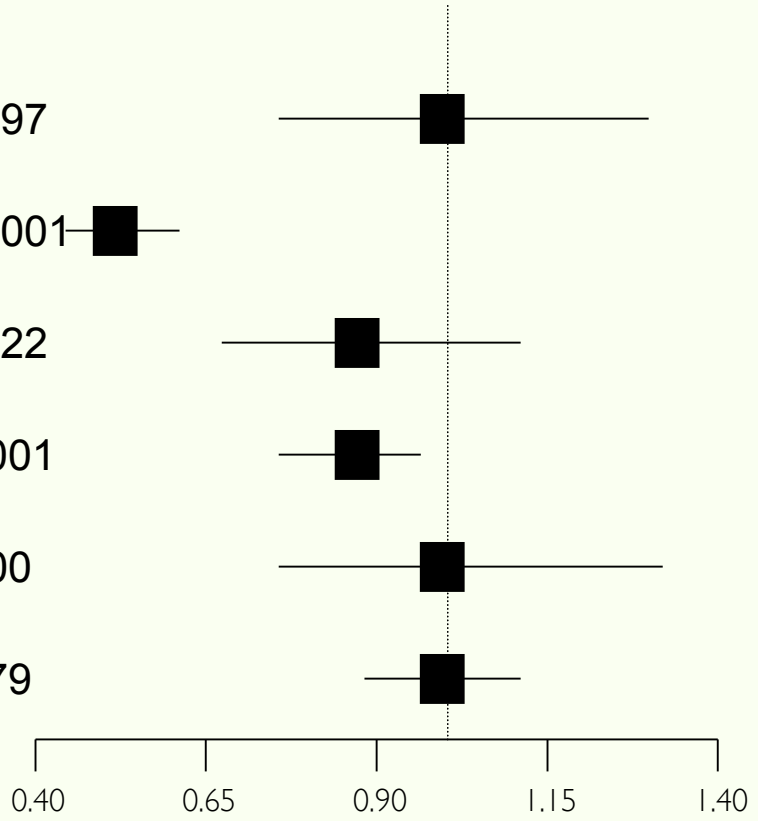
Diuretics vs Ca channel blockers

Outcome	RR	95% CI	p
CHD	0.89	0.76-1.01	0.07
Heart failure*	0.74	0.67-0.81	<0.001
Stroke*	1.02	0.91-1.14	0.74
CVD events*	0.94	0.89-1.00	0.045
CVD mortality*	0.95	0.87-1.04	0.29
Total mortality*	1.03	0.98-1.08	0.30



Diuretics vs α-blockers

Outcome	RR	95% CI	p
CHD	0.99	0.75-1.31	0.97
Heart failure	0.51	0.43-0.60	<0.001
Stroke	0.85	0.66-1.10	0.22
CVD events	0.84	0.75-0.93	0.001
CVD mortality	1.00	0.75-1.34	1.00
Total mortality	0.98	0.88-1.10	0.79

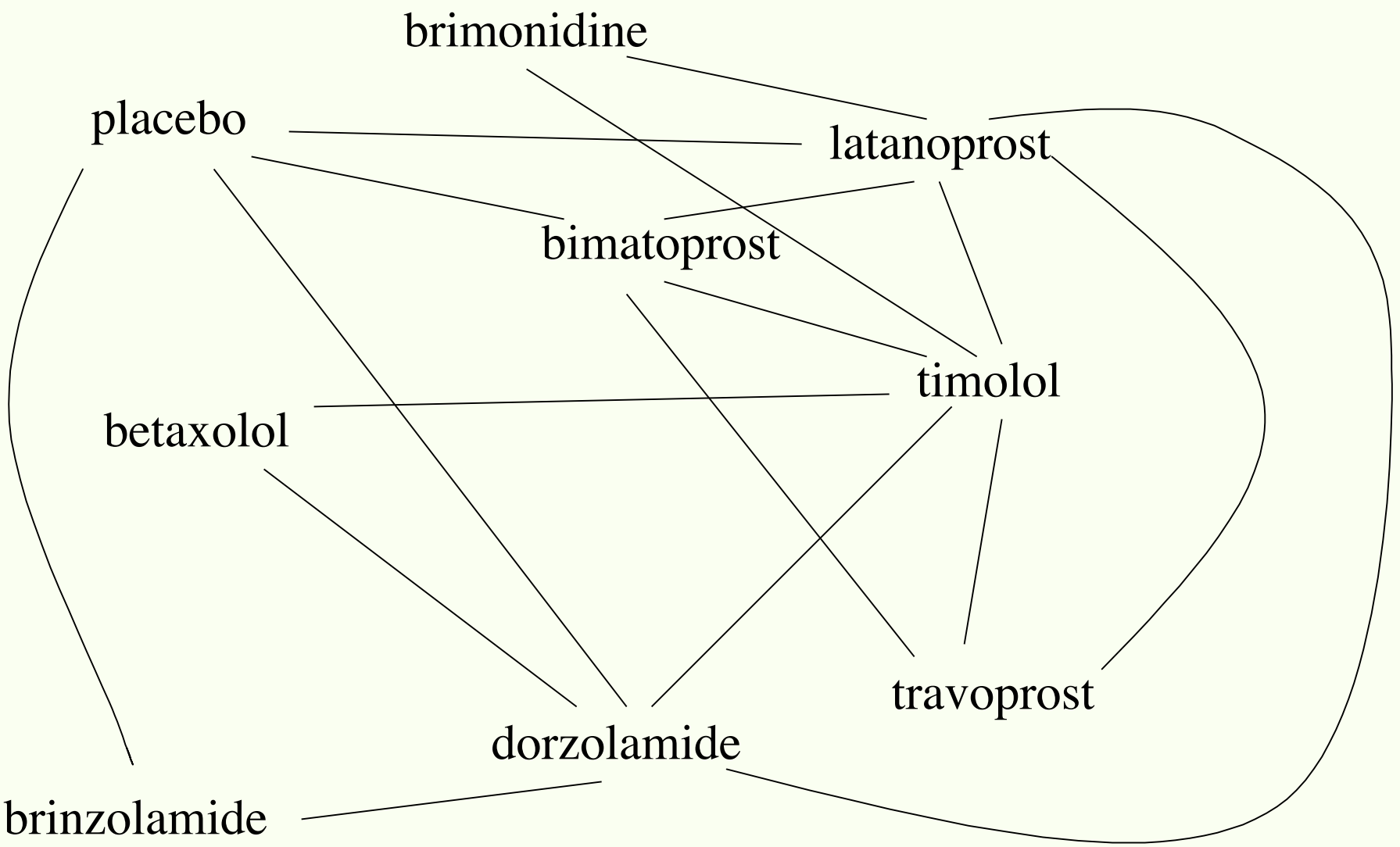


Direct vs indirect comparisons

- The estimates are weighted averages of randomised contrasts and so should be about as reliable as ordinary meta-analyses.
- Our estimated incoherence was very low except for the outcome of congestive heart failure.
- Two large trials came out while our paper was under review: ALLHAT and ANBP2. Their results were consistent with our estimates for the outcomes and comparisons they evaluated.

Example: glaucoma

- Glaucoma: increase in pressure in the eye, can lead to reduction in peripheral vision, eventual blindness.
  - β-blockers reduce resistance to outflow: *timolol, betaxolol,*
  - α_2-agonists reduce production of fluid: *brimonidine*
  - carbonic anhydrase inhibitors reduce production of fluid: *dorzolamide, brinzolamide*
  - prostaglandins increase outflow of fluid: *latanoprost, travoprost, bimatoprost.*


brimonidine
placebo
latanoprost
bimatoprost
timolol
betaxolol
travoprost
dorzolamide
brinzolamide

Problems with the test

Incoherent tests

- We assumed that the systematic part of the model is a difference $\delta_{ij}=\mu_i-\mu_j$
- If a trial uses rank tests (Wilcoxon, logrank, Cox likelihood ratio) then this is not true
  - It is possible with all the standard rank tests to have treatment A better than B better than C better than A *in a single 3-armed trial*
  - Phenomenon is called non-transitivity and was known for the Wilcoxon test before the test was invented.

Example

Consider a clinical syndrome with three possible causes: I, II, III. Two treatments A and B each improve some causes, make others worse.

Cause	Prevalence	Duration untreated	Duration with A	Duration with B
I	40%	3	2	0
II	20%	3	2	4
III	40%	3	5	4

Wilcoxon test, based on Pr(X<Y), says A better than untreated, B better than A, untreated better than B.

Three Wilcoxon tests

- If we know a priori that the distributions are related by location shift, we have a transitive test and can estimate a shift parameter and confidence interval
- If we know a priori that the distributions are stochastically ordered, the test is transitive and consistent for determining the ordering
- Without assumptions the test cannot be used to order distributions (even though that is what it is used for)

Non-transitivity

- Say that a test comparing X and Y is for a parameter T if for every α there are two rejection regions, one with T(X)>T(Y) and one with T(X)<T(Y).
- The test is unbiased if it has level α whenever T(X)=T(Y).
- Theorem: A test that has power >α and <1 is transitive iff it is a test for some parameter.
- Theorem: Tests for single quantiles are the only small-sample unbiased, transitive location tests.
- Theorem: Any parameter has large-sample unbiased tests

Outline of proof

In concrete examples it is usually easy to find a parameter for a transitive test, the trick is showing that this can always be done.

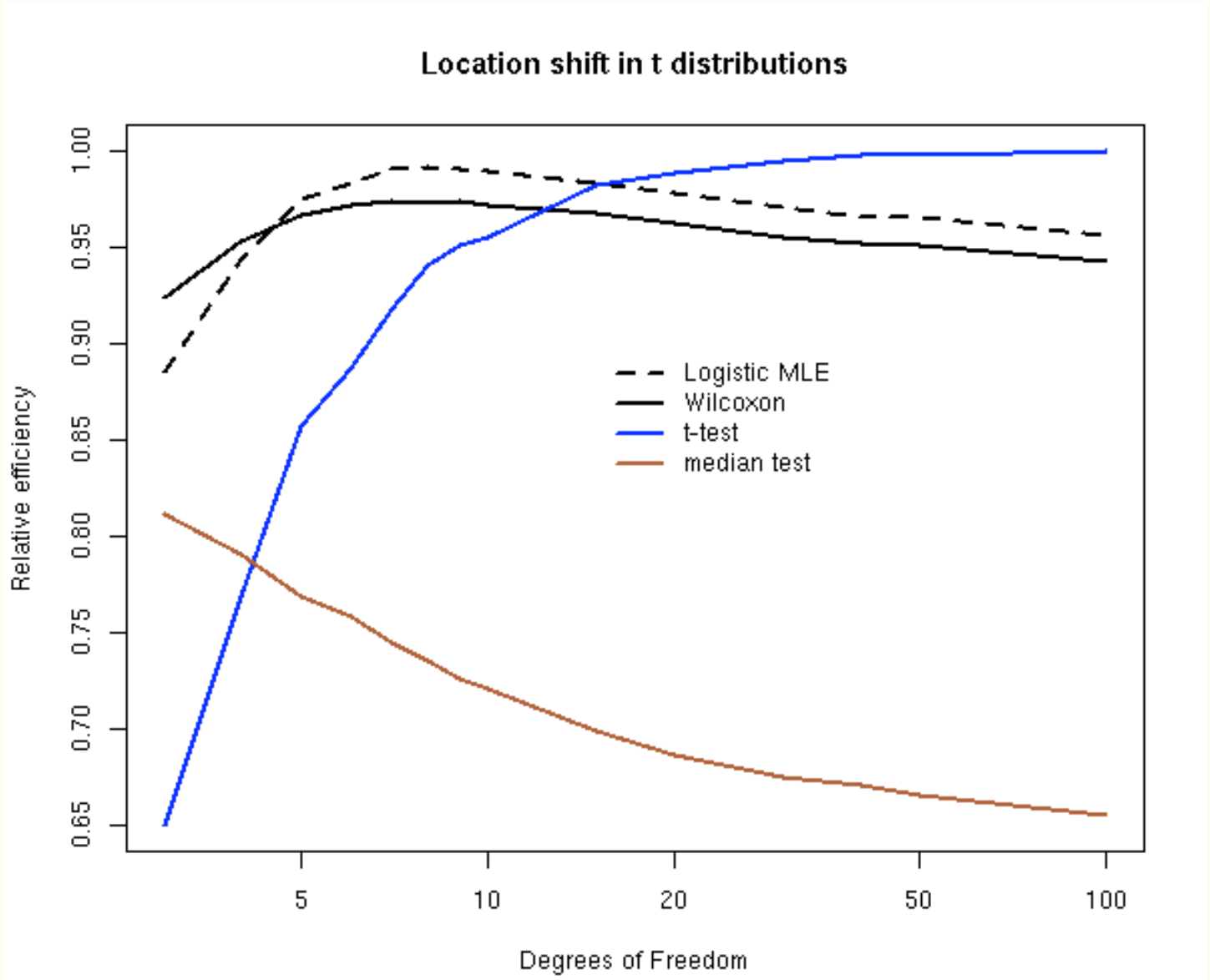
- A transitive test defines a weak linear order on distributions
- A weak linear order defines a linear order on equivalence classes of distributions
- A linearly ordered set can be labelled with real numbers iff the order topology is separable
- We can then define a parameter for the test by the labels on the equivalence classes.
- Under mild conditions on the power of the test, the order topology is metrisable and thus separable.

Fixing non-transitivity

- Some non-transitive tests are very useful in particular circumstances
  - The Wilcoxon test has good power for location shifts in symmetric, moderately heavy-tailed distributions
  - The logrank test has good power for proportional hazards
- Can we find tests with similar operating characteristics in these specialised circumstances that are transitive unconditionally?

Fixing non-transitivity

- Wilcoxon test is optimal for logistic distributions, so is equivalent to likelihood ratio test for location parameter there.
- This defines a parameter that we can use in other distributions, and that will give a test similar to the Wilcoxon in situations where the Wilcoxon is powerful
- Bootstrap or sandwich estimators of variance then give an unbiased large-sample test.


Location shift in t distributions
Relative efficiency
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00
5
10
20
50
100
Degrees of Freedom
Logistic MLE
Wilcoxon
t-test
median test

Non-transitivity

- Most rank tests are inconsistent with any ordering of all distributions. Most statisticans do not know this. (informal sampling estimate about 2%)
- Rank tests can be modified to be transitive, with little impact on power in the cases where power is good, but the resulting tests are not distribution-free in small samples.
- Essentially all transitive tests are tests for equality of some univariate summary statistic (eg mean, median, proportion, variance,…)

Should we care?

- Survival analyses in randomised trials: we are already cautious in the absence of stochastic ordering
- Survival analysis with late entry: stochastic ordering is not enough, so some care is needed
- Wilcoxon test: often recommended as a statistical garbage disposal with no assumptions. Not true.
- Math. stat classes: teach transitivity along with other optimality properties of tests?

Conclusions — methods

- Considering whole networks of clinical trials makes the usefulness of indirect comparisons an empirical question
- The estimates are weighted averages of randomised comparisons and so should have similar evidential weight to ordinary meta-analyses
- The main challenge in meta-analysis is making estimation fail when appropriate
- Separating testing from estimation can lead to confusing results. But we knew that.

Conclusions — hypertension example

- **The literature on hypertension trials is more consistent than it initially appears, in particular ALLHAT results were not surprising.**

- "If you are taking something other than low-dose diuretics for high blood pressure, it's reasonable to ask your physician why." — *Bruce Psaty*

References

Network meta-analysis

- Lumley T. *Stat Med* 2002; 21:2313-2324
- Psaty BM, et al *JAMA* 2003; 289:2534-2544

Literature on indirect comparisons

- Bucher H, et al. *J Clin Epi* 1997; 50: 683
- Song F, et al. *BMJ* 2003; 326: 472.
- ICH E10 at http://www.ich.org
- Catrin Tudur Smith (talk at RSS 2004)
- Gould AL, *Drug Information Journal* 1991 25:369-80

Related meta-analysis problems

- Surrogate outcomes: Hughes et al, AIDS. 1998 12:1823-32.
- Cross-species comparisons: DuMouchel et al, Health Phys. 1989;57:411-8.

Non-transitivity

- Brown BM, Hettmansperger TP. ANZ J Stat 2002 44:427-438
- Dan Gillen's PhD dissertation
- Martin Gardner's *Mathematical Games* column October 1974, December 1970