Systematic Approaches to Assessing the Internal and External Validity of Randomized Controlled Trials

John P.A. Ioannidis, MD, DSc

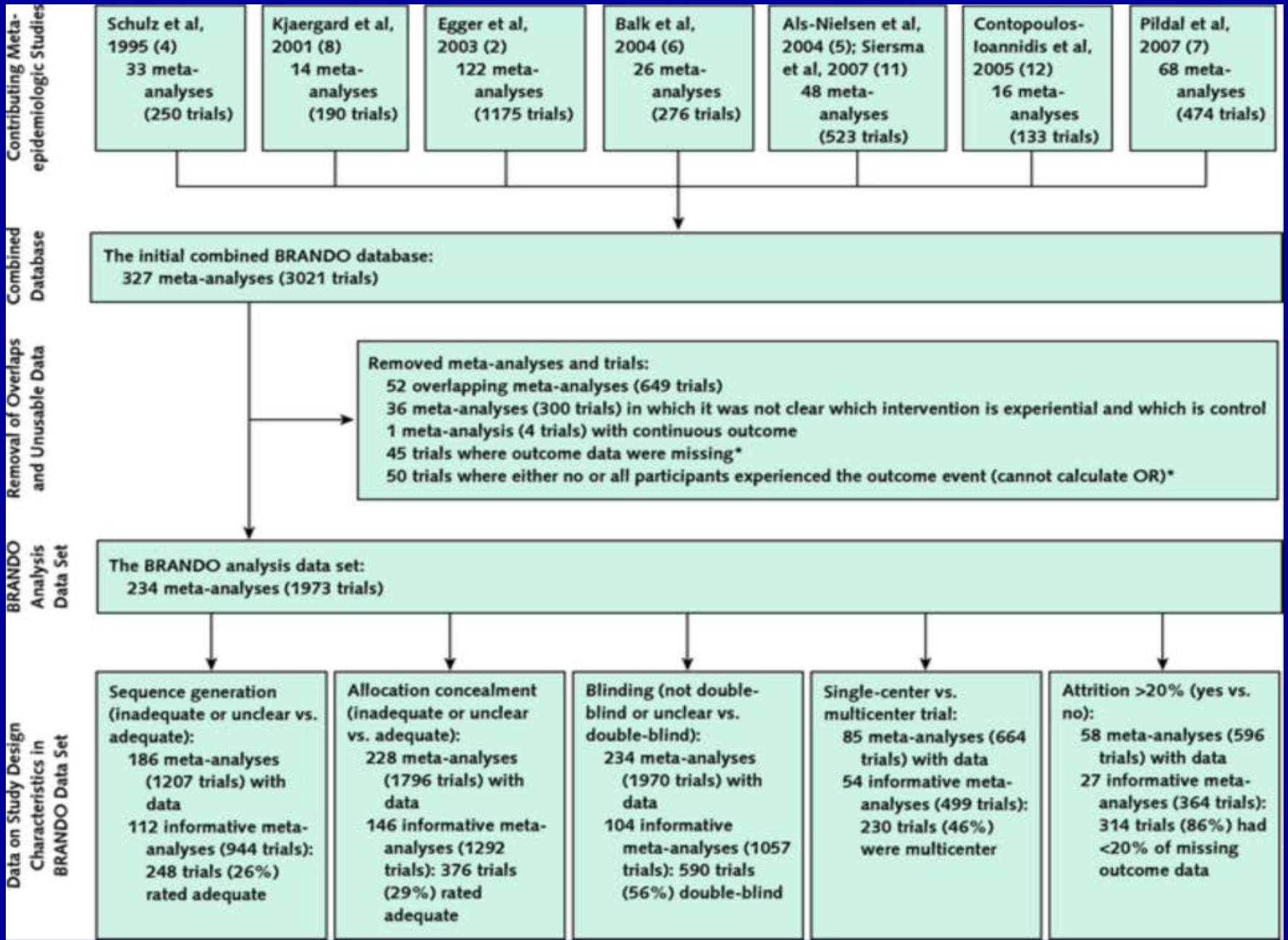
Appraising the quality of studies

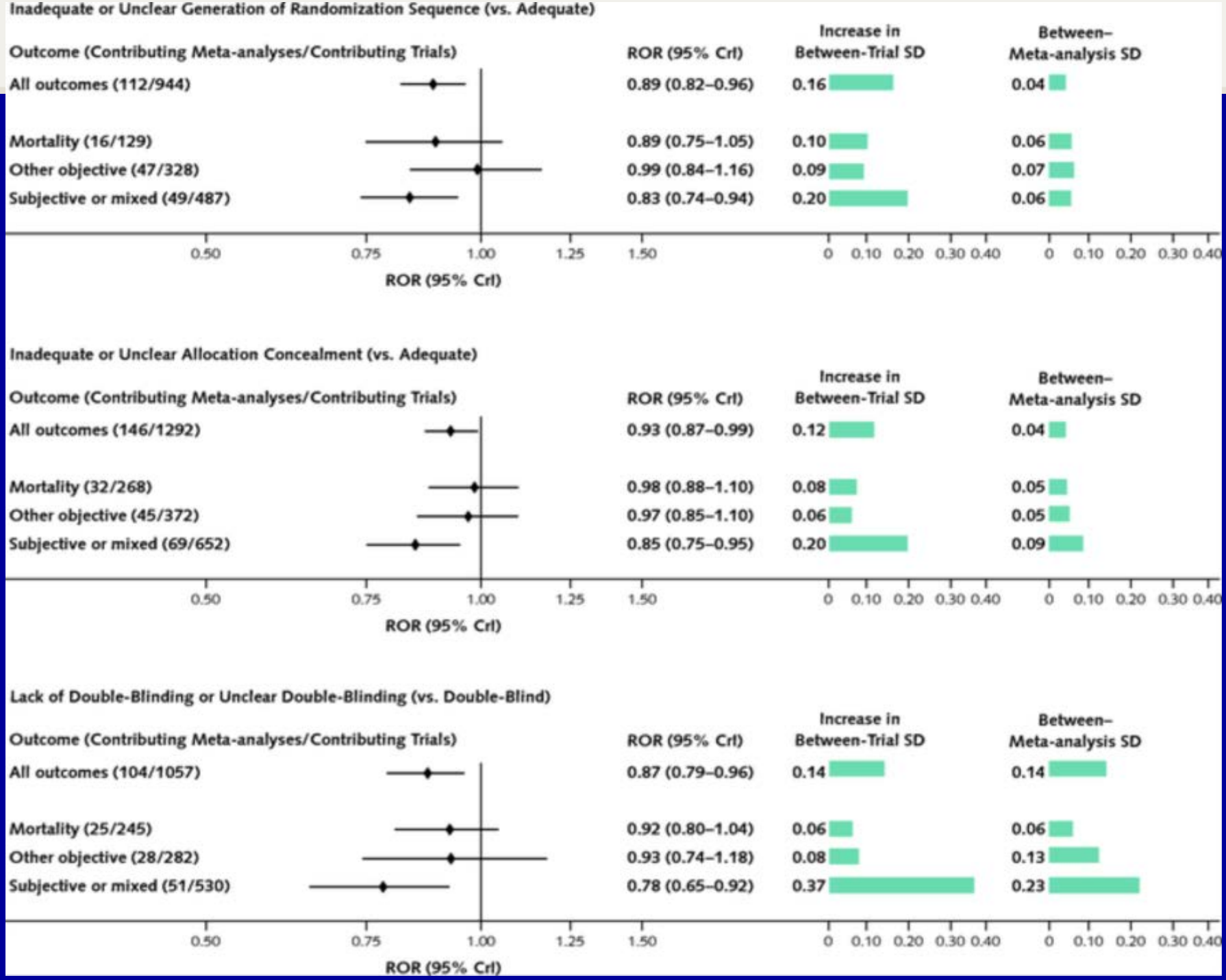
- Early empirical evaluations suggested that effect sizes in randomized trials may depend on aggregate quality scores; this has been dismissed, since there are so many quality scores, that inferences are widely different
- Other empirical evaluations suggested that specific quality items such as lack of blinding and lack of allocation concealment in RCTs may inflate treatment effects (e.g. Shultz et al. JAMA 1995)
- Now it seems more likely that such quality deficits may be associated either with inflated or with deflated treatment effects
- Poor quality indicators may be associated with larger heterogeneity of effects, especially for outcomes that are subjective.

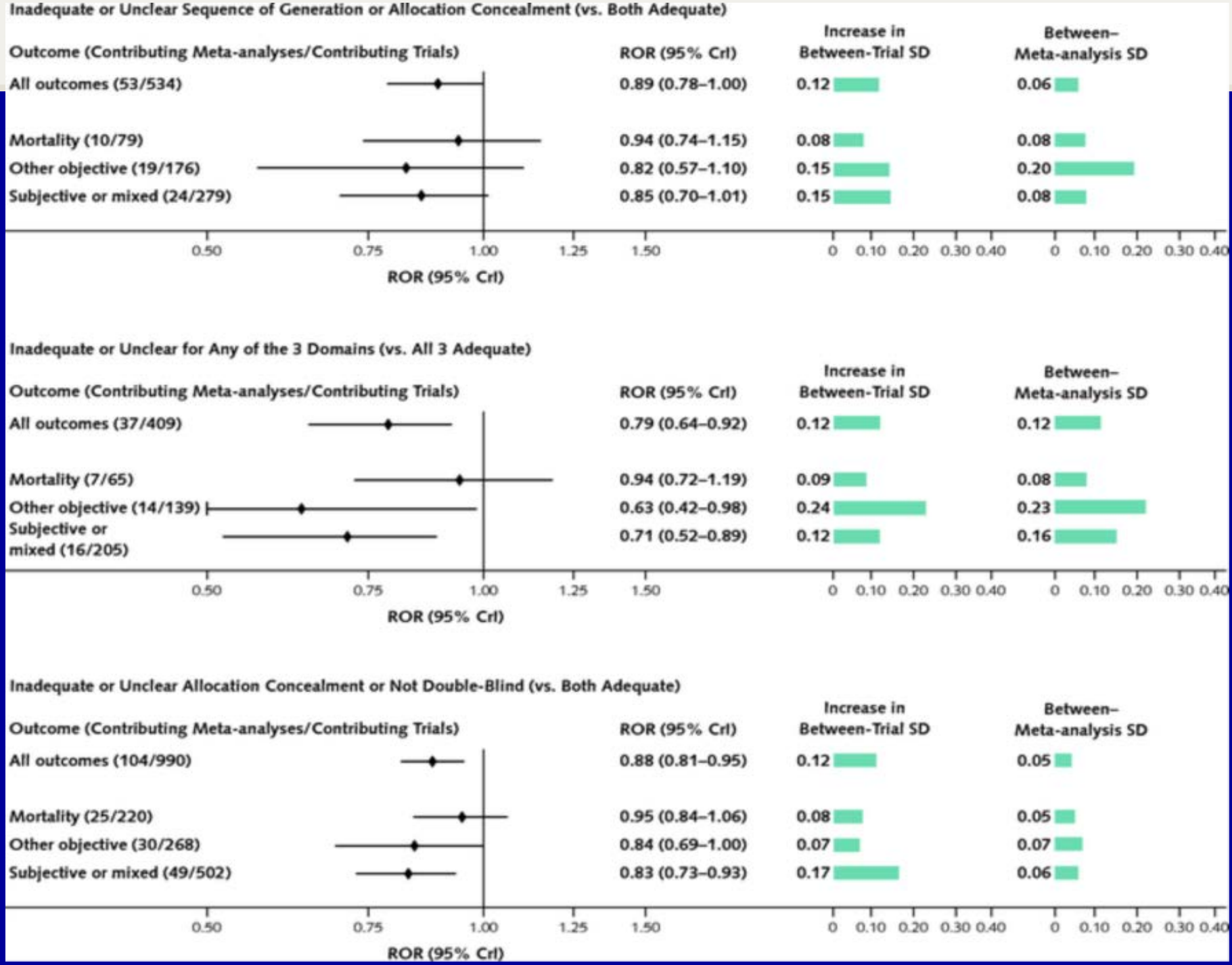
The two kinds of bad quality

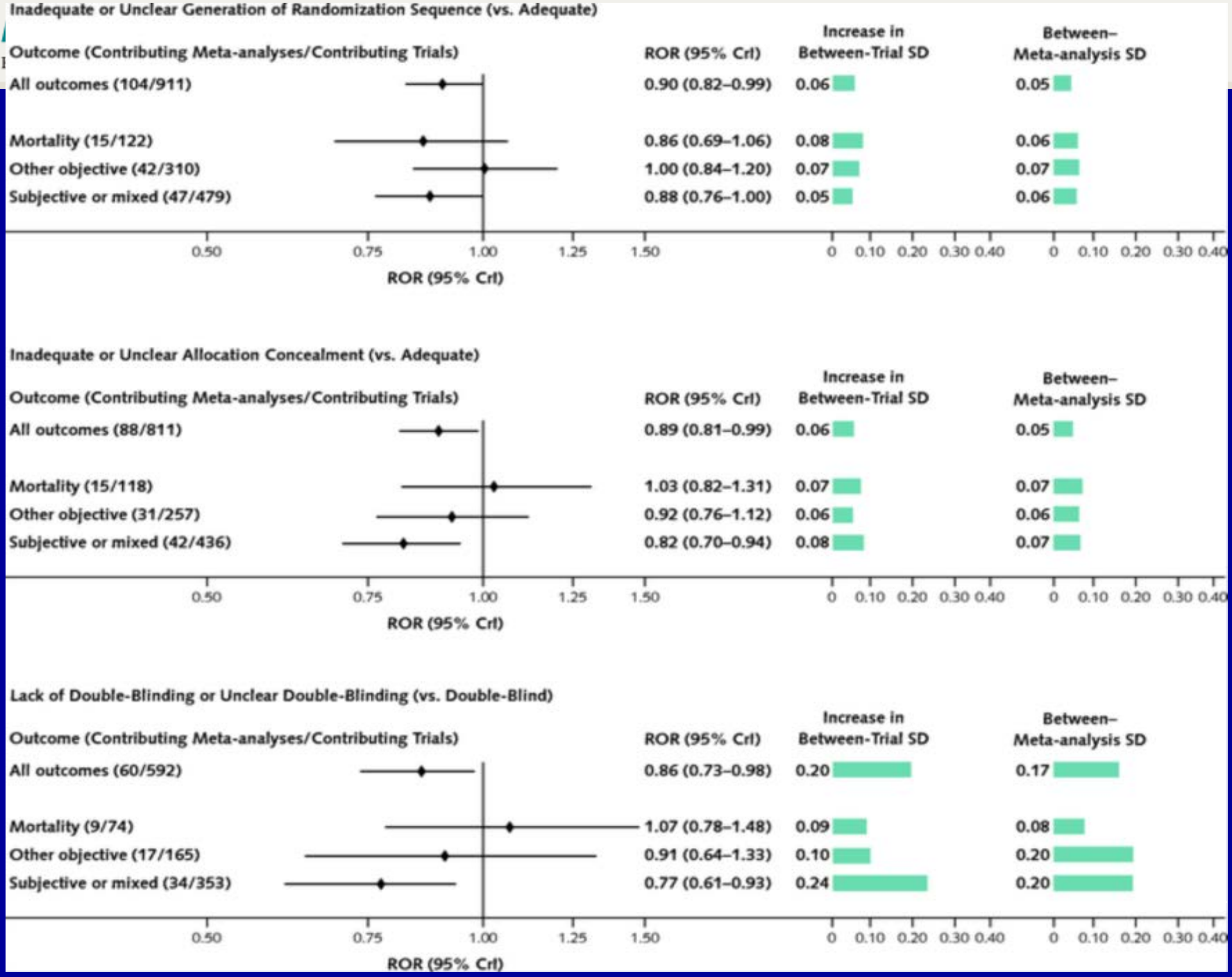
- Quality is bad on (evil) purpose = the effect sizes are almost always inflated
- Quality is bad because of stupidity = the effect sizes may be anything; usually, but not always, they are deflated

Ann Intern Med. 2012;157(6):429-438. doi:10.7326/0003-4819-157-6-201209180-00537




Inadequate or Unclear Generation of Randomization Sequence (vs. Adequate)
Outcome (Contributing Meta-analyses/Contributing Trials)
ROR (95% CrI)
Increase in Between-Trial SD
Between-Meta-analysis SD
All outcomes (112/944)
0.89 (0.82–0.96)
0.16
0.04
Mortality (16/129)
0.89 (0.75–1.05)
0.10
0.06
Other objective (47/328)
0.99 (0.84–1.16)
0.09
0.07
Subjective or mixed (49/487)
0.83 (0.74–0.94)
0.20
0.06
0.50 0.75 1.00 1.25 1.50
0 0.10 0.20 0.30 0.40
0 0.10 0.20 0.30 0.40
ROR (95% CrI)
Inadequate or Unclear Allocation Concealment (vs. Adequate)
Outcome (Contributing Meta-analyses/Contributing Trials)
ROR (95% CrI)
Increase in Between-Trial SD
Between-Meta-analysis SD
All outcomes (146/1292)
0.93 (0.87–0.99)
0.12
0.04
Mortality (32/268)
0.98 (0.88–1.10)
0.08
0.05
Other objective (45/372)
0.97 (0.85–1.10)
0.06
0.05
Subjective or mixed (69/652)
0.85 (0.75–0.95)
0.20
0.09
0.50 0.75 1.00 1.25 1.50
0 0.10 0.20 0.30 0.40
0 0.10 0.20 0.30 0.40
ROR (95% CrI)
Lack of Double-Blinding or Unclear Double-Blinding (vs. Double-Blind)
Outcome (Contributing Meta-analyses/Contributing Trials)
ROR (95% CrI)
Increase in Between-Trial SD
Between-Meta-analysis SD
All outcomes (104/1057)
0.87 (0.79–0.96)
0.14
0.14
Mortality (25/245)
0.92 (0.80–1.04)
0.06
0.06
Other objective (28/282)
0.93 (0.74–1.18)
0.08
0.13
Subjective or mixed (51/530)
0.78 (0.65–0.92)
0.37
0.23
0.50 0.75 1.00 1.25 1.50
0 0.10 0.20 0.30 0.40
0 0.10 0.20 0.30 0.40
ROR (95% CrI)


Inadequate or Unclear Sequence of Generation or Allocation Concealment (vs. Both Adequate)
Outcome (Contributing Meta-analyses/Contributing Trials)
ROR (95% CrI)
Increase in Between-Trial SD
Between-Meta-analysis SD
All outcomes (53/534)
0.89 (0.78–1.00)
0.12
0.06
Mortality (10/79)
0.94 (0.74–1.15)
0.08
0.08
Other objective (19/176)
0.82 (0.57–1.10)
0.15
0.20
Subjective or mixed (24/279)
0.85 (0.70–1.01)
0.15
0.08
0.50 0.75 1.00 1.25 1.50
0 0.10 0.20 0.30 0.40
0 0.10 0.20 0.30 0.40
ROR (95% CrI)
Inadequate or Unclear for Any of the 3 Domains (vs. All 3 Adequate)
Outcome (Contributing Meta-analyses/Contributing Trials)
ROR (95% CrI)
Increase in Between-Trial SD
Between-Meta-analysis SD
All outcomes (37/409)
0.79 (0.64–0.92)
0.12
0.12
Mortality (7/65)
0.94 (0.72–1.19)
0.09
0.08
Other objective (14/139)
0.63 (0.42–0.98)
0.24
0.23
Subjective or mixed (16/205)
0.71 (0.52–0.89)
0.12
0.16
0.50 0.75 1.00 1.25 1.50
0 0.10 0.20 0.30 0.40
0 0.10 0.20 0.30 0.40
ROR (95% CrI)
Inadequate or Unclear Allocation Concealment or Not Double-Blind (vs. Both Adequate)
Outcome (Contributing Meta-analyses/Contributing Trials)
ROR (95% CrI)
Increase in Between-Trial SD
Between-Meta-analysis SD
All outcomes (104/990)
0.88 (0.81–0.95)
0.12
0.05
Mortality (25/220)
0.95 (0.84–1.06)
0.08
0.05
Other objective (30/268)
0.84 (0.69–1.00)
0.07
0.07
Subjective or mixed (49/502)
0.83 (0.73–0.93)
0.17
0.06
0.50 0.75 1.00 1.25 1.50
0 0.10 0.20 0.30 0.40
0 0.10 0.20 0.30 0.40
ROR (95% CrI)


Inadequate or Unclear Generation of Randomization Sequence (vs. Adequate)
Outcome (Contributing Meta-analyses/Contributing Trials)
ROR (95% CrI)
Increase in Between-Trial SD
Between-Meta-analysis SD
All outcomes (104/911)
0.90 (0.82–0.99)
0.06
0.05
Mortality (15/122)
0.86 (0.69–1.06)
0.08
0.06
Other objective (42/310)
1.00 (0.84–1.20)
0.07
0.07
Subjective or mixed (47/479)
0.88 (0.76–1.00)
0.05
0.06
0.50 0.75 1.00 1.25 1.50
0 0.10 0.20 0.30 0.40
0 0.10 0.20 0.30 0.40
ROR (95% CrI)
Inadequate or Unclear Allocation Concealment (vs. Adequate)
Outcome (Contributing Meta-analyses/Contributing Trials)
ROR (95% CrI)
Increase in Between-Trial SD
Between-Meta-analysis SD
All outcomes (88/811)
0.89 (0.81–0.99)
0.06
0.05
Mortality (15/118)
1.03 (0.82–1.31)
0.07
0.07
Other objective (31/257)
0.92 (0.76–1.12)
0.06
0.06
Subjective or mixed (42/436)
0.82 (0.70–0.94)
0.08
0.07
0.50 0.75 1.00 1.25 1.50
0 0.10 0.20 0.30 0.40
0 0.10 0.20 0.30 0.40
ROR (95% CrI)
Lack of Double-Blinding or Unclear Double-Blinding (vs. Double-Blind)
Outcome (Contributing Meta-analyses/Contributing Trials)
ROR (95% CrI)
Increase in Between-Trial SD
Between-Meta-analysis SD
All outcomes (60/592)
0.86 (0.73–0.98)
0.20
0.17
Mortality (9/74)
1.07 (0.78–1.48)
0.09
0.08
Other objective (17/165)
0.91 (0.64–1.33)
0.10
0.20
Subjective or mixed (34/353)
0.77 (0.61–0.93)
0.24
0.20
0.50 0.75 1.00 1.25 1.50
0 0.10 0.20 0.30 0.40
0 0.10 0.20 0.30 0.40
ROR (95% CrI)

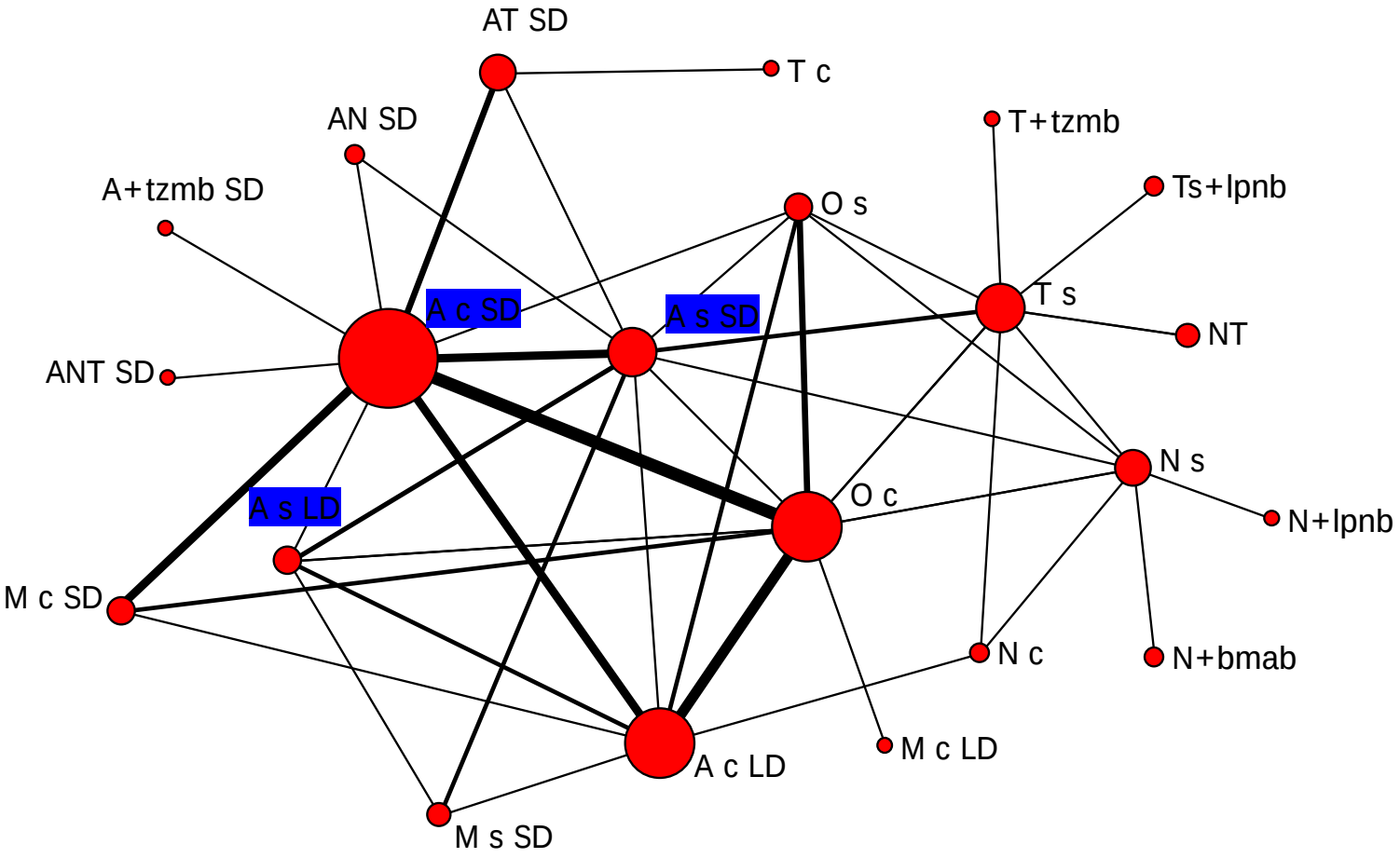
Randomized trials published in major journals, 2009

TABLE Comparison of the nominal significance of effect between the unadjusted and the adjusted estimates in randomized trials that reported both types of estimate (n=40).

		Adjusted estimate	
		Significant	Not-significant
		N (%)	N (%)
Unadjusted estimate	Significant	17 (42.5%)	3 (7.5%)
	Not significant	4 (10.0%)	16 (40.0%)

Saquib, Saquib, and Ioannidis, BMJ 2013

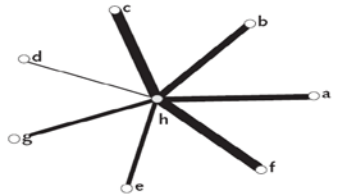
Networks of randomized evidence



Mauri et al, JNCI 2008

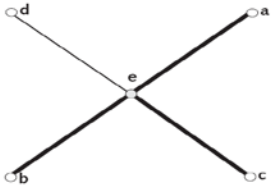
A

Second-generation antiepileptic drugs in partial epilepsy



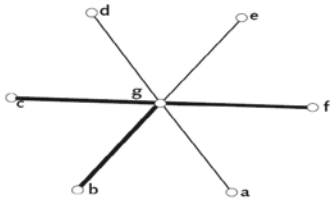
a: levetiracetam, b: gabapentin, c: lamotrigine, d: oxcarbazepine, e: tiagabine, f: topiramate, g: zonisamide, h: placebo

Biologic treatments for rheumatoid arthritis



a: adalimumab, b: infliximab, c: etanercept, d: anakinra, e: placebo

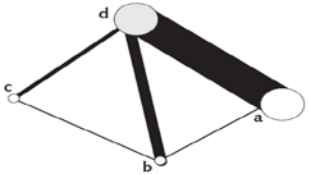
Intracoronary drug-eluting stents vs. bare-metal stents



a: AES, b: apolymeric PES, c: polymeric SES, d: MES, e: polymeric EES, f: polymeric PES, g: BMS

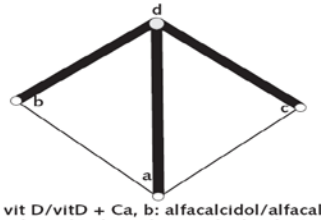
B

Smoking cessation therapies



a: nicotine replacement therapy, b: buproprion, c: varenicline, d: placebo/no treatment

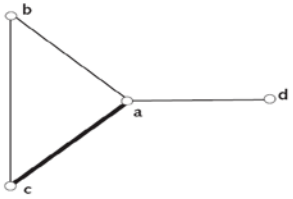
C Vitamin D and analogues to prevent bone loss and fractures



a: vit D/vitD + Ca, b: alfacalcidol/alfacalcidol + Ca, c: calcitriol/calcitriol + Ca, d: placebo/Ca

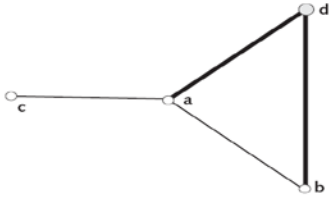
C *continued*

Self-monitoring of glucose in type 2 diabetes



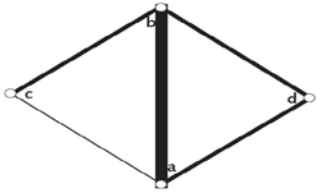
a: self-monitoring of blood glucose, b: self-monitoring of urine glucose, c: no self-monitoring, d: self-monitoring of blood glucose with regular feedback

Antiretroviral resistance testing in treatment-experienced patients



a: PART, b: GART, c: vPART, d: empiric

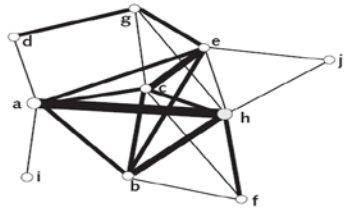
Prophylaxis for *Pneumocystis carinii* in HIV-infected patients



a: trimethoprim–sulfamethoxazole, b: pyrimethamine, c: dapsone, d: dapsone–pyrimethamine

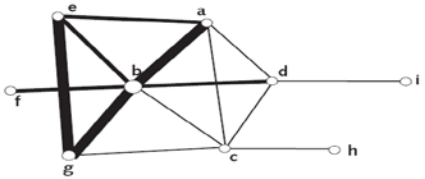
D

First-line antihypertensive therapy



a: diuretics, b: β-blockers, c: CCB, d: nonhydropiridine CCB, e: ACE-i, f: ARB, g: diuretics or β-blockers, h: placebo/not treated, i: α-blocker, j: ACE-i + diuretics

Chemotherapy regimens for ovarian cancer

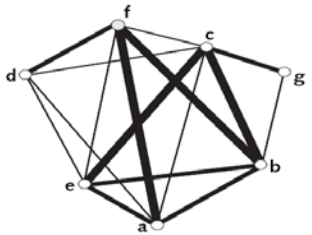


a: platinum monotherapy, b: platinum-based combination, c: taxane monotherapy, d: platinum + taxane-based combination, e: nonplatinum/nontaxane monotherapy, f: platinum-based combination (ip), g: nonplatinum/nontaxane combination, h: taxane-based combination, i: platinum/taxane-based combination (ip)

Shaded nodes indicate placebo or no active treatment. The thickness of the lines is proportional to the number of trials addressing each specific comparison. **A.** Star-shaped networks. **B.** Non–star-shaped networks with limited diversity and significant co-occurrence. **C.** Non–star-shaped networks with limited diversity and nonsignificant co-occurrence. **D.** Networks with considerable diversity and significant co-occurrence. **E.** Networks with considerable diversity and nonsignificant co-occurrence. ACE-I = angiotensin-converting enzyme

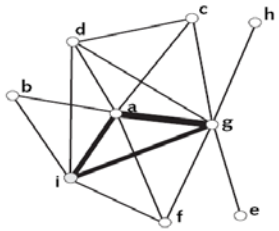
E

Antihypertensive treatment (incidence of diabetes)



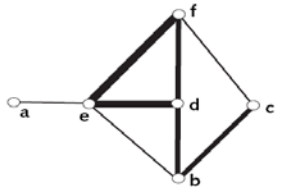
a: diuretic, b: ACE-i, c: CCB, d: ARB, e: β-blocker, f: placebo, g: β-blocker or diuretic

Stroke prevention in nonrheumatic atrial fibrillation



a: aspirin, b: alternate-day aspirin, c: fixed low-dose-warfarin, d: fixed low-dose warfarin and aspirin, e: indobufen, f: adjusted low-dose warfarin, g: adjusted standard-dose warfarin, h: ximelagatran, i: placebo/control

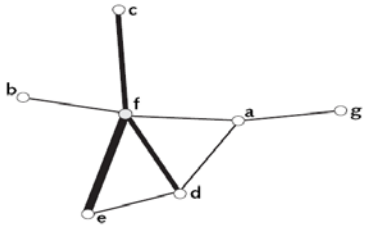
Treatments for acute myocardial infarction



a: anistreplase, b: accelerated t-PA, c: reteplase, d: angioplasty, e: streptokinase, f: t-PA

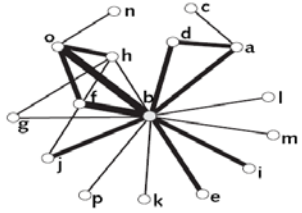
F

α-1 antagonists in lower urinary tract symptoms



a: alfuzocin, b: alfuzocin SR, c: doxazocin, d: tamsulocin, e: terazocin, f: placebo, g: prazosin

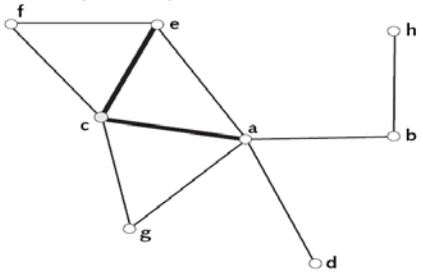
Topical nonsteroidal anti-inflammatory drugs for acute pain



a: indomethacin, b: placebo, c: indomethacin + placebo, d: piroxicam, e: niflumic acid, f: ibuprofen, g: ketorolac, h: etofenamate, i: diclofenac, j: felbinac, k: fentiazac, l: naproxen, m: meclofenamic acid, n: flunoxaprofen, o: ketoprofen, q: flurbiprofen

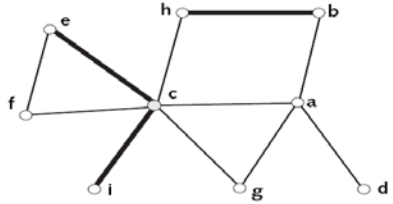
G

Antifungal prophylaxis in liver transplant recipients



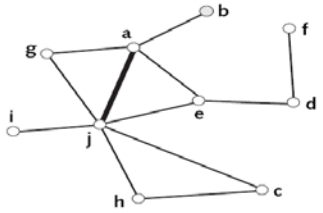
a: fluconazole, b: nystatin, c: placebo, d: clotrimazole, e: itraconazole, f: fluconazole + itraconazole, g: amphotericin, h: liposomal amphotericin B

Antifungal prophylaxis in solid organ transplant recipients



a: fluconazole, b: nystatin, c: placebo, d: amphotericin B, e: liposomal amphotericin B, f: fluconazole + itraconazole, g: itraconazole, h: clotrimazole, i: ketoconazole

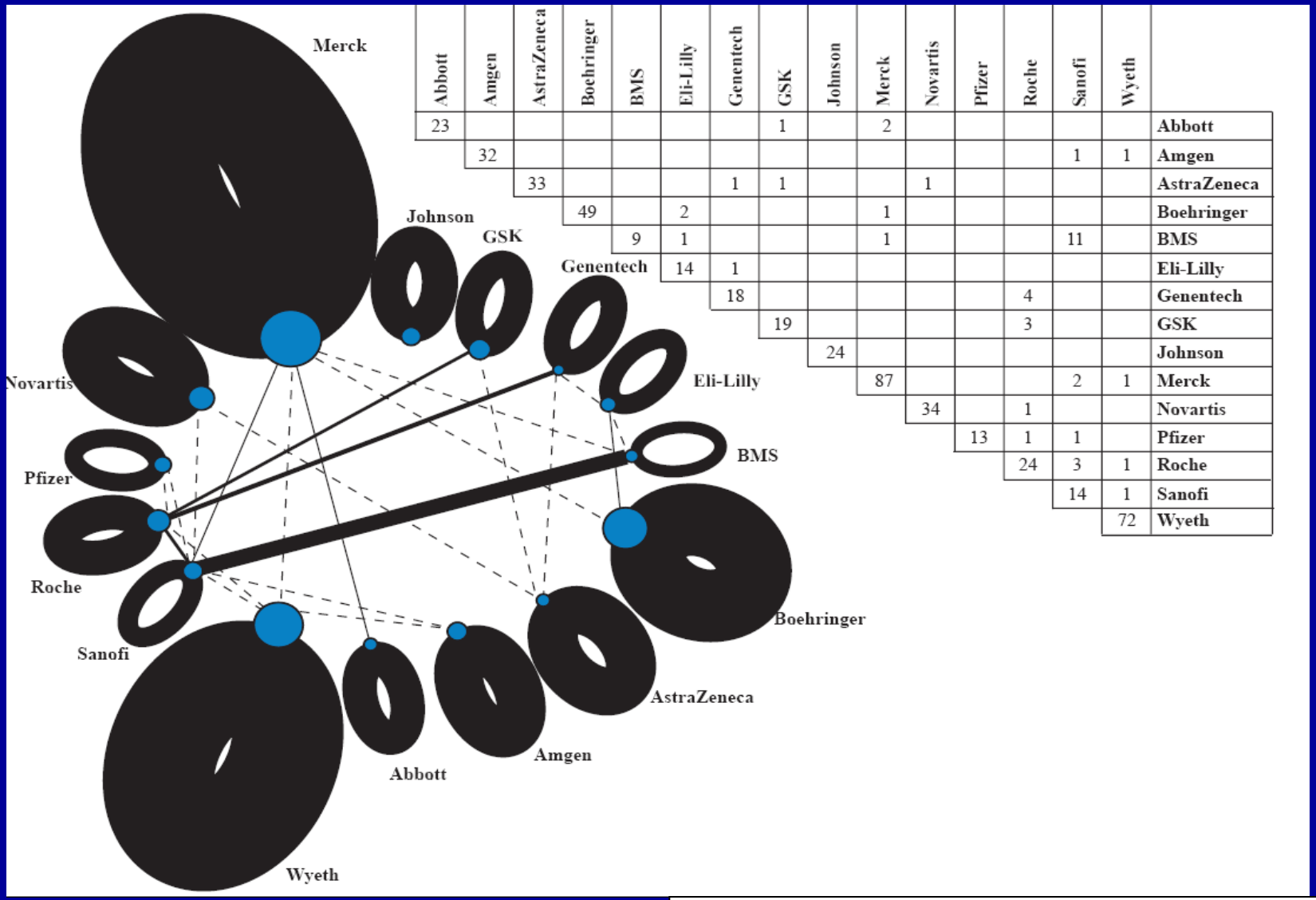
Topical antibiotics without steroids for chronic ear discharge without eardrum perforation



a: ciprofloxacin, b: placebo, c: ofloxacin, d: TSP, e: gentamicin, f: TP, g: tobramycin, h: neomycin–polymyxin, i: chloramphenicol/gentamycin, j: antiseptic

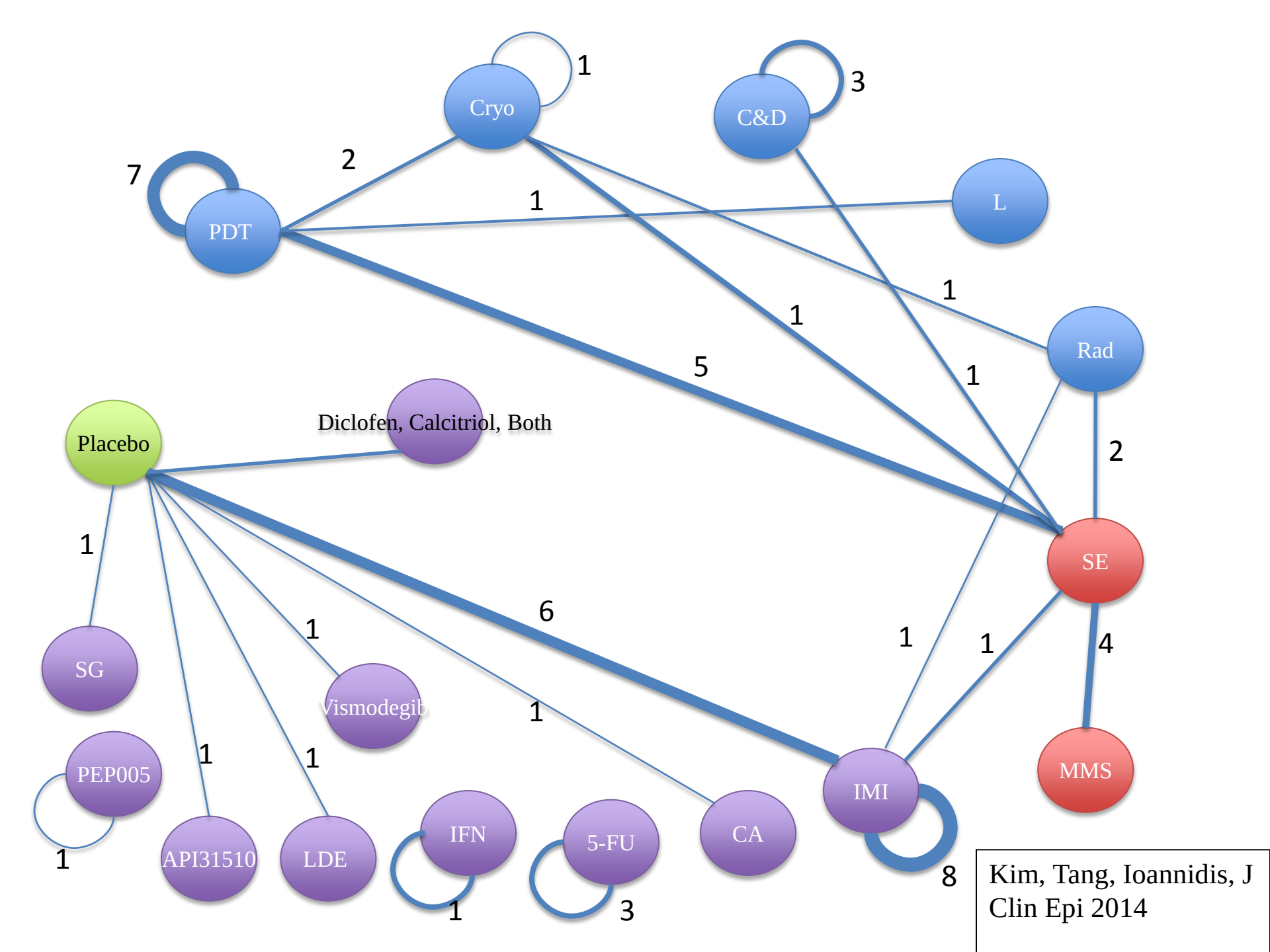
Auto-looping

Design of clinical research: an open world or isolated city-states (company-states)?



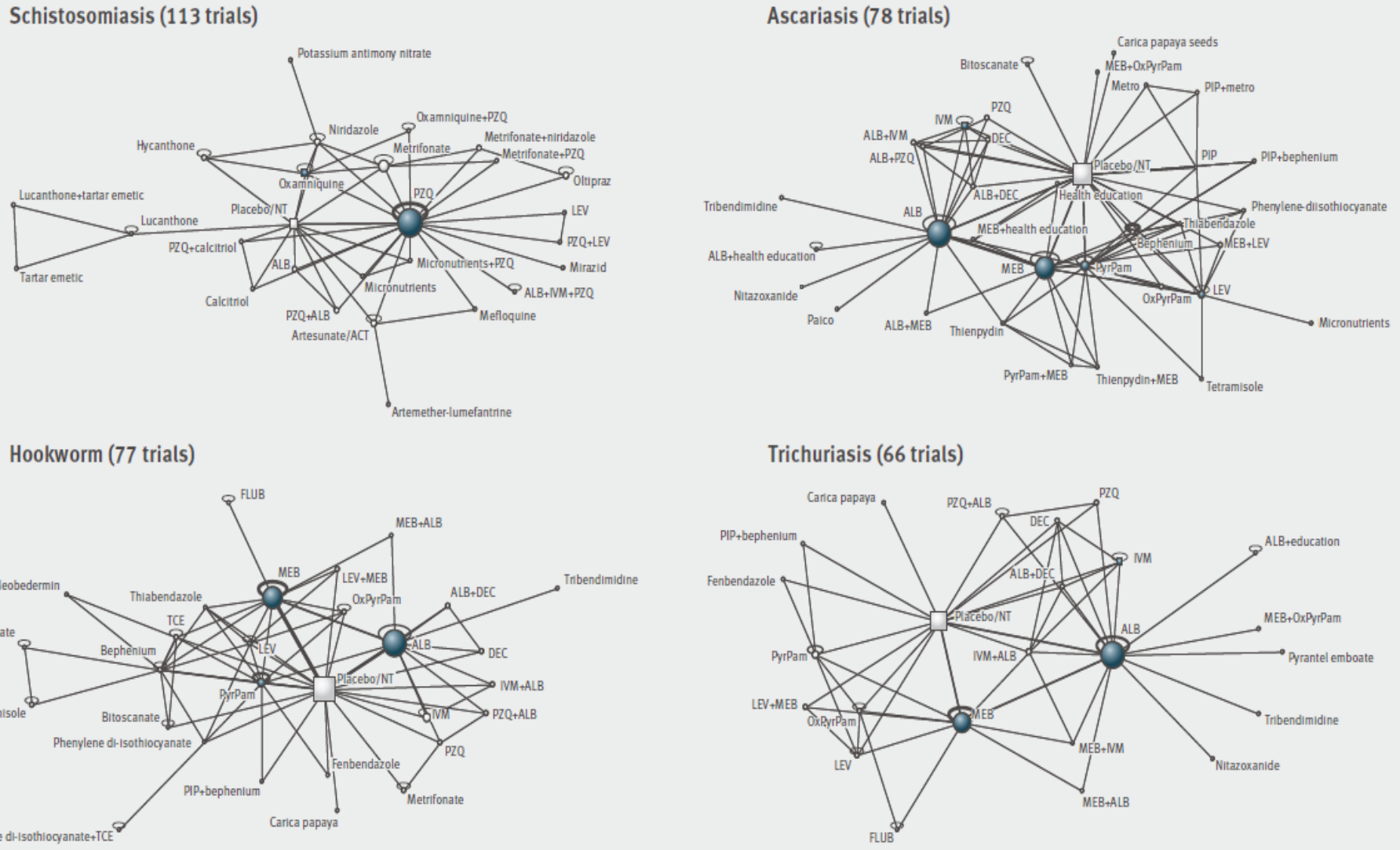
	Abbott	Amgen	AstraZeneca	Boehringer	BMS	Eli-Lilly	Genentech	GSK	Johnson	Merck	Novartis	Pfizer	Roche	Sanofi	Wyeth
Abbott	23							1		2					
Amgen		32												1	1
AstraZeneca			33				1	1			1				
Boehringer				49		2				1					
BMS					9	1				1				11	
Eli-Lilly						14	1								
Genentech							18						4		
GSK								19					3		
Johnson									24						
Merck										87				2	1
Novartis											34		1		
Pfizer												13	1	1	
Roche													24	3	1
Sanofi														14	1
Wyeth															72

Lathyris et al., Eur J Clin Invest, 2010



Kim, Tang, Ioannidis, J Clin Epi 2014

Even the most simple research agendas are complex



Kappagoda and Ioannidis, BMJ 2012

Are large treatment effects generalizable?

ORIGINAL CONTRIBUTION

Empirical Evaluation of Very Large Treatment Effects of Medical Interventions

Tiago V. Pereira, PhD
Ralph I. Horwitz, MD
John P. A. Ioannidis, MD, DSc

Context Most medical interventions have modest effects, but occasionally some clinical trials may find very large effects for benefits or harms.

Objective To evaluate the frequency and features of very large effects in medicine.

Data Sources Cochrane Database of Systematic Reviews (CDSR, 2010, issue 7).

Study Selection We separated all binary-outcome CDSR forest plots with comparisons of interventions according to whether the first published trial, a subsequent trial (not the first), or no trial had a nominally statistically significant ($P<.05$) very large effect (odds ratio [OR], ≥5). We also sampled randomly 250 topics from each group for further in-depth evaluation.

Data Extraction We assessed the types of treatments and outcomes in trials with very large effects, examined how often large-effect trials were followed up by other trials on the same topic, and how these effects compared against the effects of the respective meta-analyses.

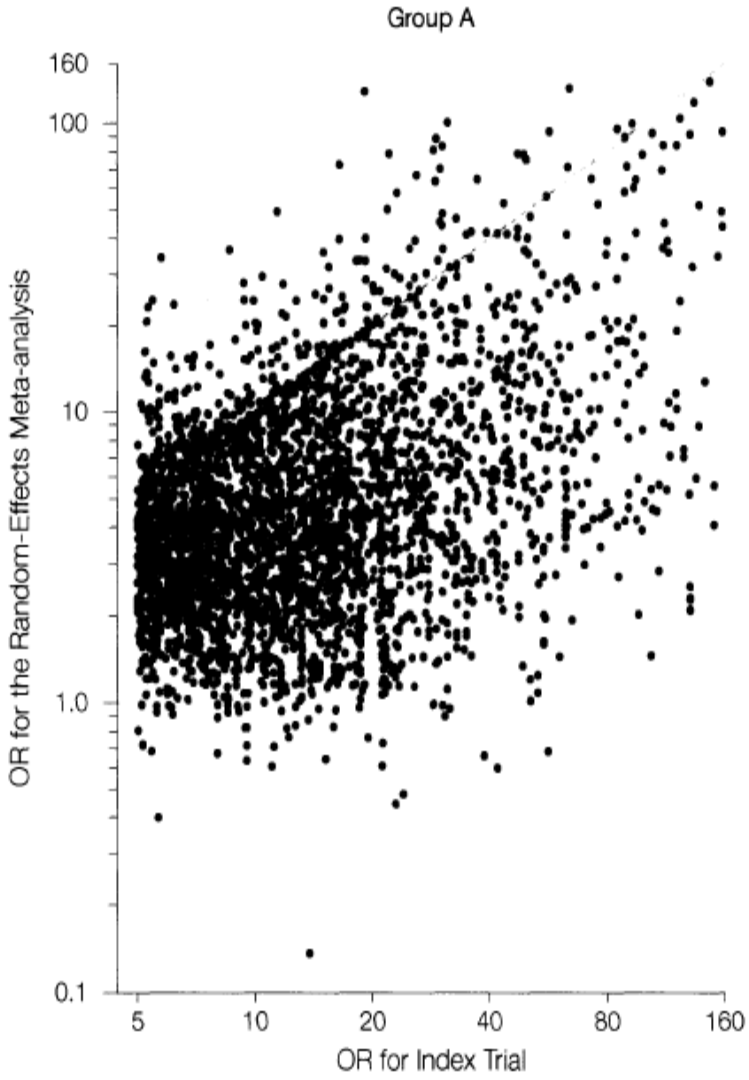
Results Among 85 002 forest plots (from 3082 reviews), 8239 (9.7%) had a significant very large effect in the first published trial, 5158 (6.1%) only after the first published trial, and 71 605 (84.2%) had no trials with significant very large effects. Nominally significant very large effects typically appeared in small trials with median number of events: 18 in first trials and 15 in subsequent trials. Topics with very large effects were less likely than other topics to address mortality (3.6% in first trials, 3.2% in subsequent trials, and 11.6% in no trials with significant very large effects) and were more likely to address laboratory-defined efficacy (10% in first trials,10.8% in subsequent, and 3.2% in no trials with significant very large effects). First trials with very large effects were as likely as trials with no very large effects to have subsequent published trials. Ninety percent and 98% of the very large effects observed in first and subsequently published trials, respectively, became smaller in meta-analyses that included other trials; the median odds ratio decreased from 11.88 to 4.20 for first trials, and from 10.02 to 2.60 for subsequent trials. For 46 of the 500 selected topics (9.2%; first and subsequent trials) with a very large-effect trial, the meta-analysis maintained very large effects with $P<.001$ when additional trials were included, but none pertained to mortality-related outcomes. Across the whole CDSR, there was only 1 intervention with large beneficial effects on mortality, $P<.001$, and no major concerns about the quality of the evidence (for a trial on extracorporeal oxygenation for severe respiratory failure in newborns).

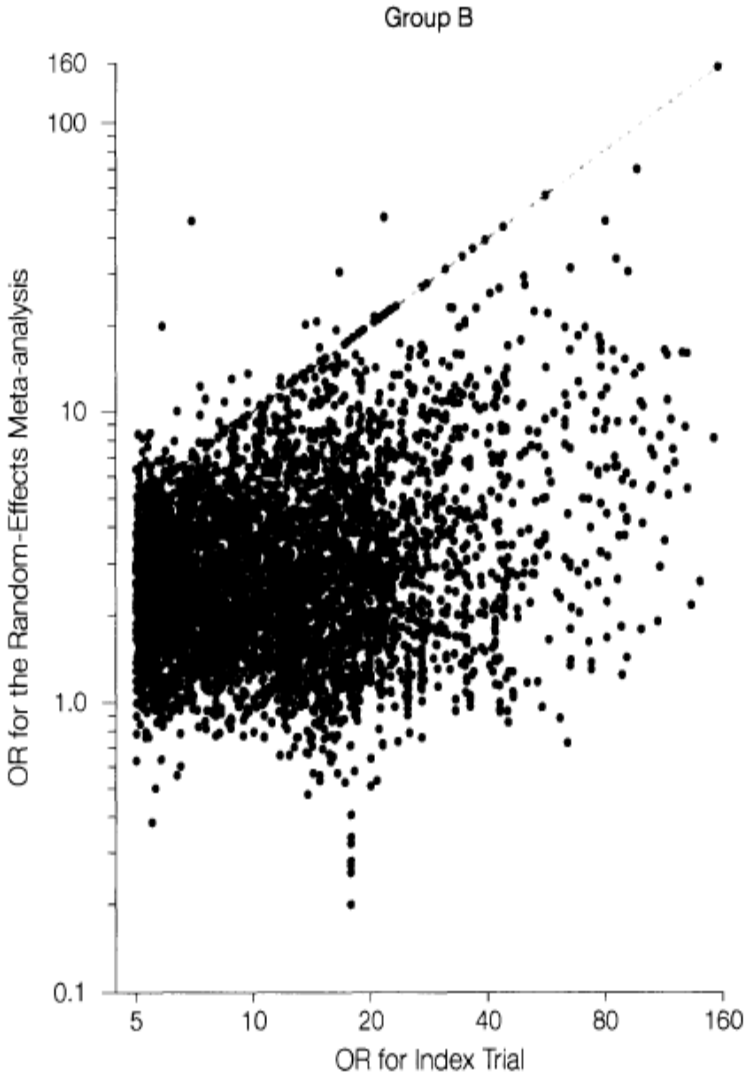
Conclusions Most large treatment effects emerge from small studies, and when additional trials are performed, the effect sizes become typically much smaller. Well-validated large effects are uncommon and pertain to nonfatal outcomes.

JAMA. 2012;308(16):1676-1684 www.jama.com

MOST EFFECTIVE INTERVENTIONS in health care confer modest, incremental benefits.[1,2] Randomized trials, the gold standard to evaluate medical interventions, are ideally conducted under the principle of equipoise[3]: the compared groups are not perceived to have a clear advantage; thus, very large treatment effects are usually not anticipated. However, very large treatment effects are observed occasionally in some trials. These effects may include both anticipated and unexpected treatment benefits, or they may involve harms.

Large effects are important to document reliably because in a relative scale they represent potentially the cases in which interventions can have the most impressive effect on health outcomes and because they are more likely to be adopted rapidly and with less evidence. Consequently, it is important to know whether, when observed, very large effects are reliable and in what sort of experimental outcomes they are commonly observed. The importance of very large effects has drawn attention mostly in observational studies[4,5] but has not been well studied in randomized evidence. It is unknown how often very large effects are replicated in





Pereira, Horwitz, Ioannidis, JAMA 2012

Adjusting effects downwards

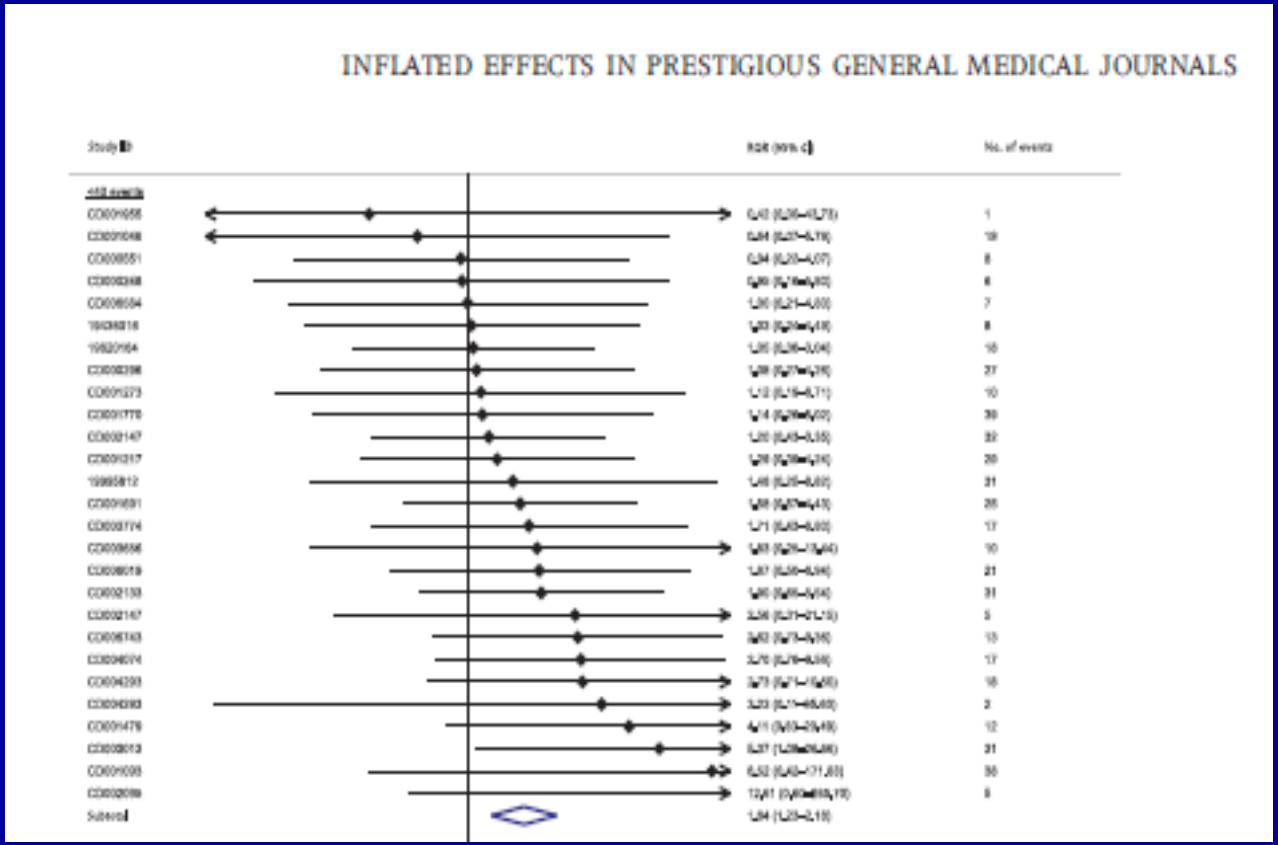
Published by Oxford University Press on behalf of the International Epidemiological Association
© The Author 2011; all rights reserved. Advance Access publication 8 September 2011

International Journal of Epidemiology 2011;**40**:1280–1291
doi:10.1093/ije/dyr095

METHODOLOGY

Magnitude of effects in clinical trials published in high-impact general medical journals

Konstantinos CM Siontis,[1] Evangelos Evangelou[1] and John PA Ioannidis[1,2,3,4]*



How to Use an Article Reporting a Multiple Treatment Comparison Meta-analysis

Edward J. Mills, PhD, MSc
John P. A. Ioannidis, MD, DSc
Kristian Thorlund, PhD, MSc
Holger J. Schünemann, MD, PhD, MSc
Milo A. Puhan, MD, PhD
Gordon H. Guyatt, MD, MSc

Multiple treatment comparison (MTC) meta-analysis uses both direct (head-to-head) randomized clinical trial (RCT) evidence as well as indirect evidence from RCTs to compare the relative effectiveness of all included interventions. The methodological quality of MTCs may be difficult for clinicians to interpret because the number of interventions evaluated may be large and the methodological approaches may be complex. Clinicians and others evaluating an MTC should be aware of the potential biases that can affect the interpretation of these analyses. Readers should consider whether the primary studies are sufficiently homogeneous to combine; whether the differ-

CLINICAL SCENARIO

You are seeing a 45-year-old patient for whom, 6 weeks previously, you pre-

BMJ

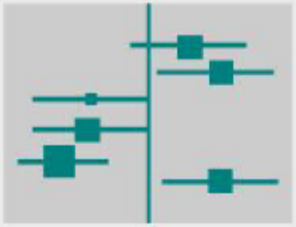
BMJ 2013;346:f2914 doi: 10.1136/bmj.f2914 (Published 14 May 2013)

RESEARCH METHODS & REPORTING

Demystifying trial networks and network meta-analysis

Networks of randomized clinical trials can be evaluated in the context of a network meta-analysis, a procedure that permits inferences into the comparative effectiveness of interventions that may or may not have been evaluated directly against each other. This approach is quickly gaining popularity among clinicians and guideline decision makers. However, certain methodological aspects are poorly understood. Here, we explain the geometry of a network, statistical and conceptual heterogeneity and incoherence, and challenges in the application and interpretation of data synthesis. These concepts are essential to make sense of a network meta-analysis.

Edward J Mills *associate professor*[1 2], Kristian Thorlund *associate professor*[2 3], John P A Ioannidis *professor*[2 4]



Effective ⟷ Harmful

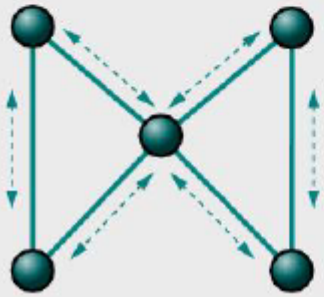
Statistical heterogeneity
Are the treatment effects observed in individual trials on the same comparison similar or dissimilar?

Population

Designs

Outcomes

Conceptual heterogeneity and incoherence
Across the pairwise comparisons and across the network, are the individual trials importantly different in terms of populations included, study designs, outcomes, etc?



Statistical incoherence
Are the treatment effects consistent or inconsistent across indirect and direct evidence estimates in the network

Posterior distributions of effects and corresponding predictive distributions of effects

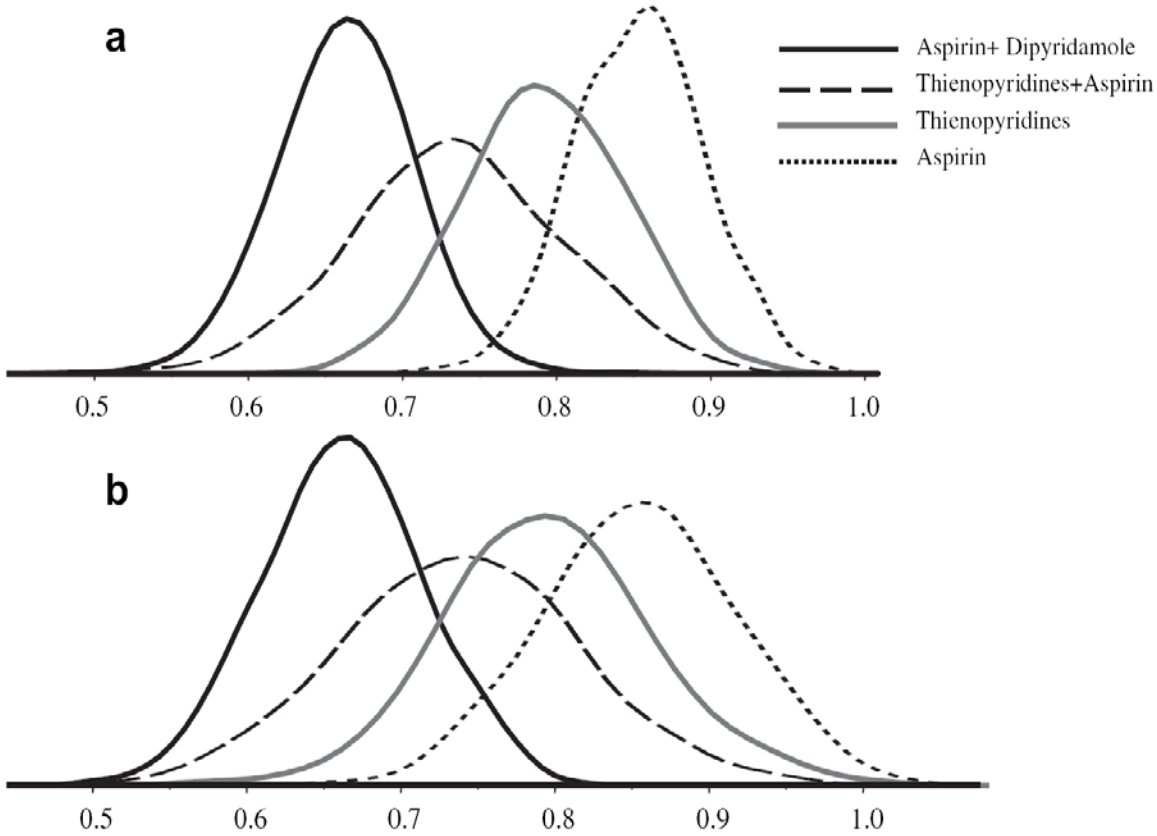
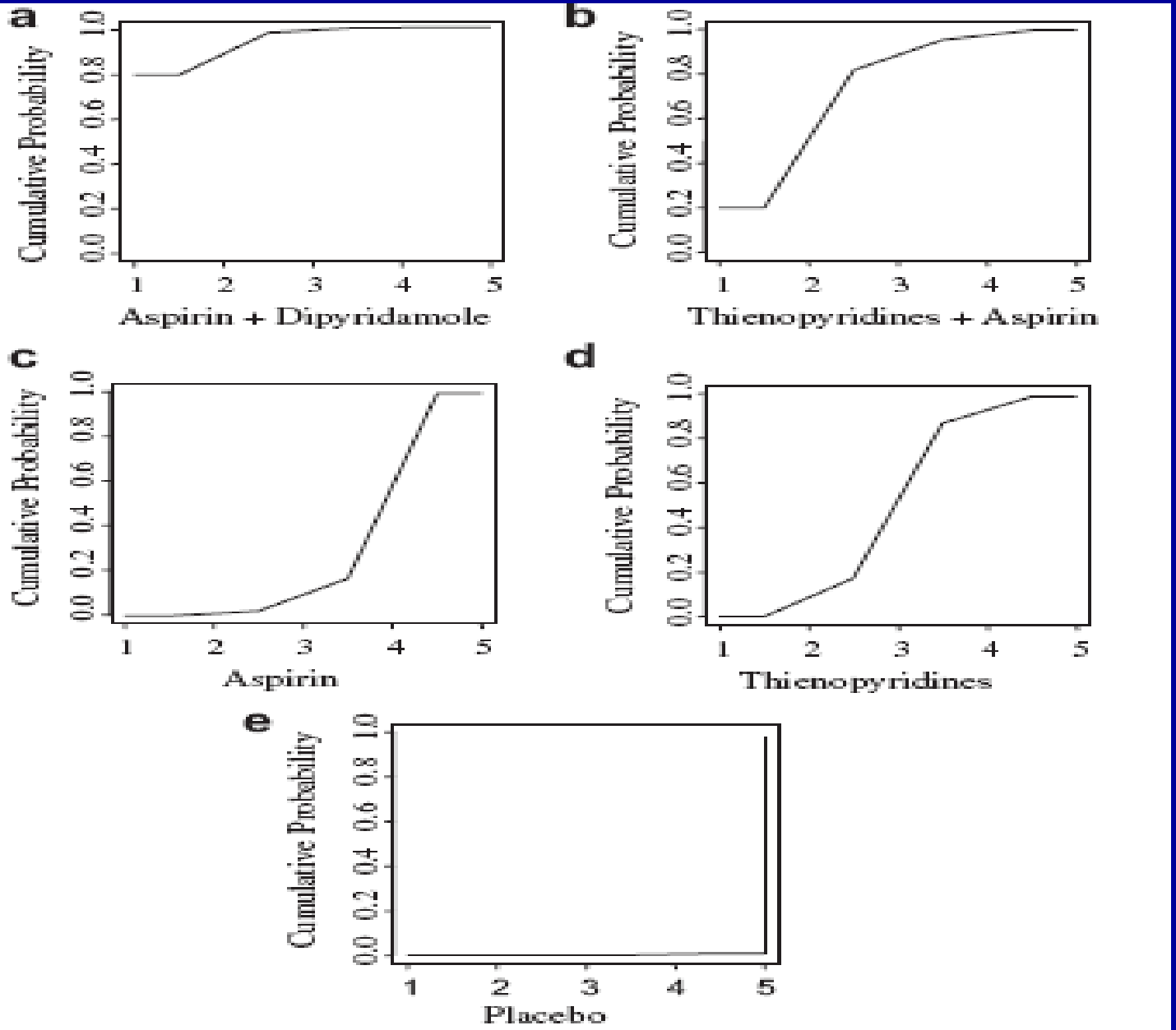


Fig. 2. Posterior distributions of mean odds ratios for serious vascular events with antiplatelet treatments compared with placebo (a) and the corresponding predictive distributions of effects within which the effect size of a new study is expected to be found with 95% probability (b).

JCE, 2011

Cumulative ranking probability



Probability of not being worse than threshold t from the best treatment

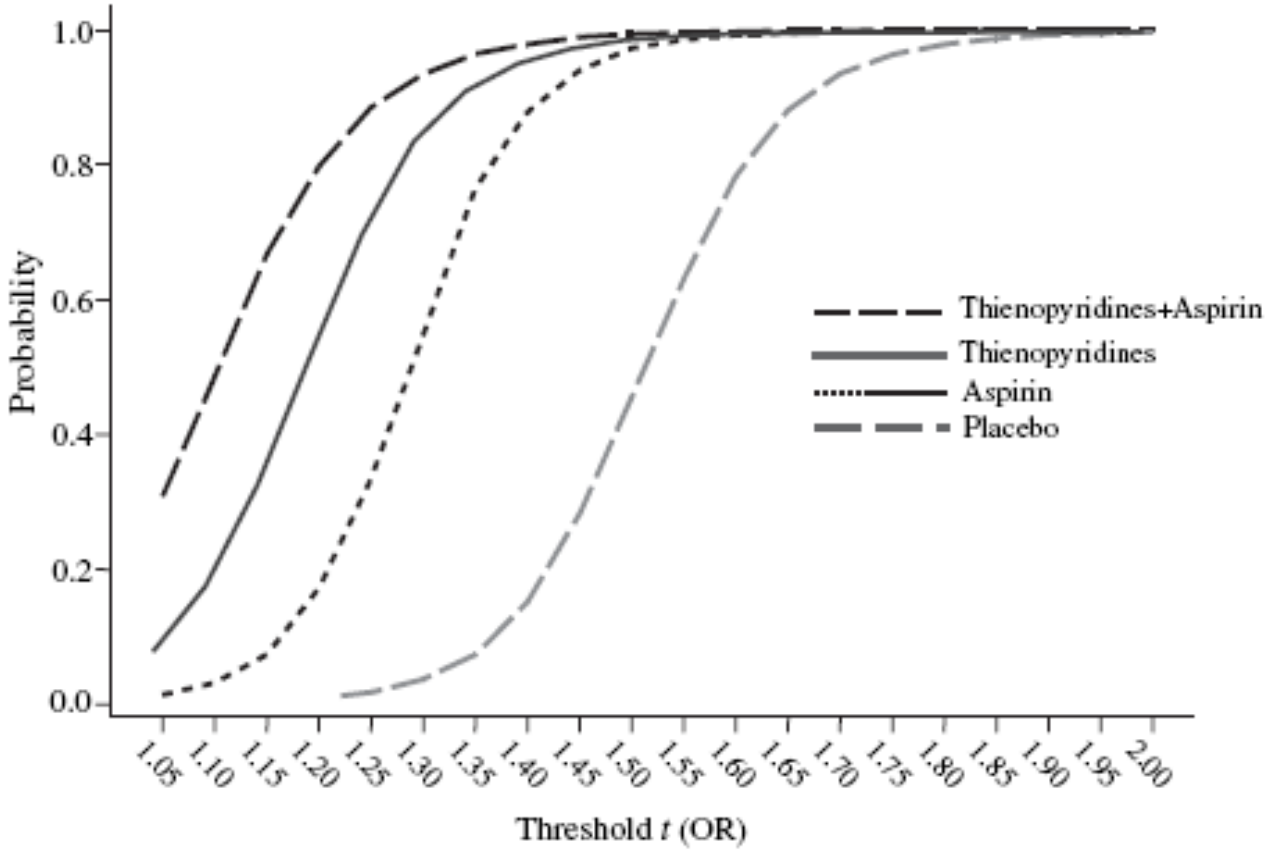


Fig. 6. Probabilities for each treatment to be no worse than the combination of aspirin and dipyridamole regarding the incidence of serious vascular events by a certain threshold *t* (on the horizontal axis) measured in odds ratio scale.

Modeling bias

Research Article

Statistics in Medicine

Received 15 December 2009, Accepted 26 May 2010 View this article online at wileyonlinelibrary.com

(wileyonlinelibrary.com) DOI: 10.1002/sim.4001

Evaluating novel agent effects in multiple-treatments meta-regression

Georgia Salanti,[a*†] Sofia Dias,[b] Nicky J. Welton,[b] AE Ades,[b] Vassilis Golfinopoulos,[c] Maria Kyrgiou,[d] Davide Mauri[a,e] and John P. A. Ioannidis[a,f,g]

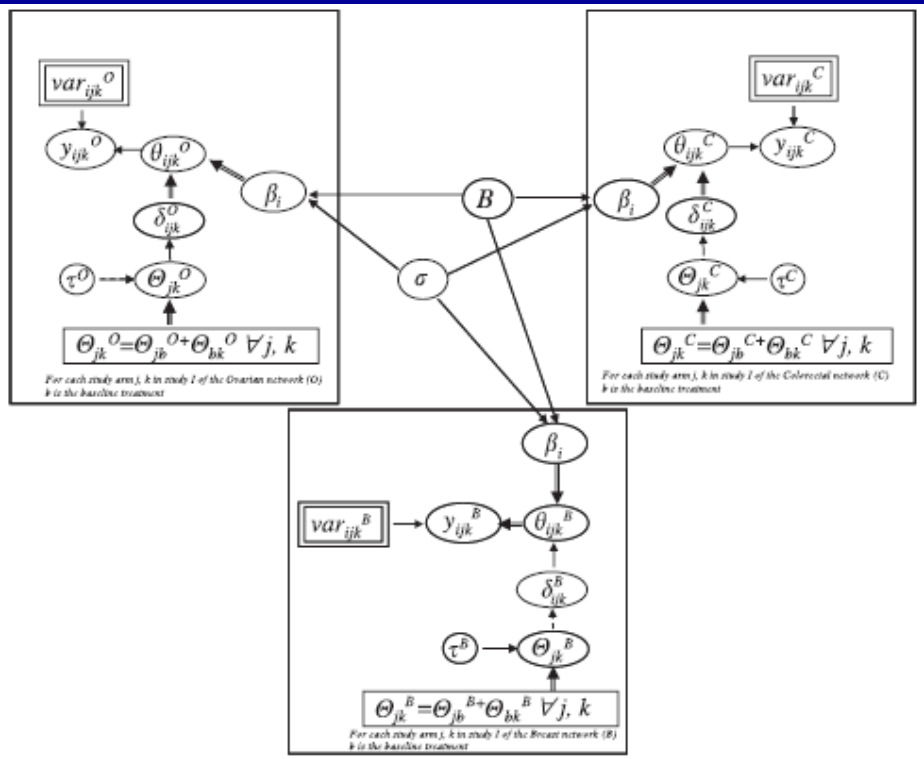


Figure 2. Graphical representation of the jointly adjusted model. Stochastic nodes (associated with distributions) and deterministic nodes (logical functions of parameters) are presented in oval shapes and data are presented in rectangular shapes. Single-line arrows represent distributions and double-line arrows represent logical functions.

Changes in cumulative ranking

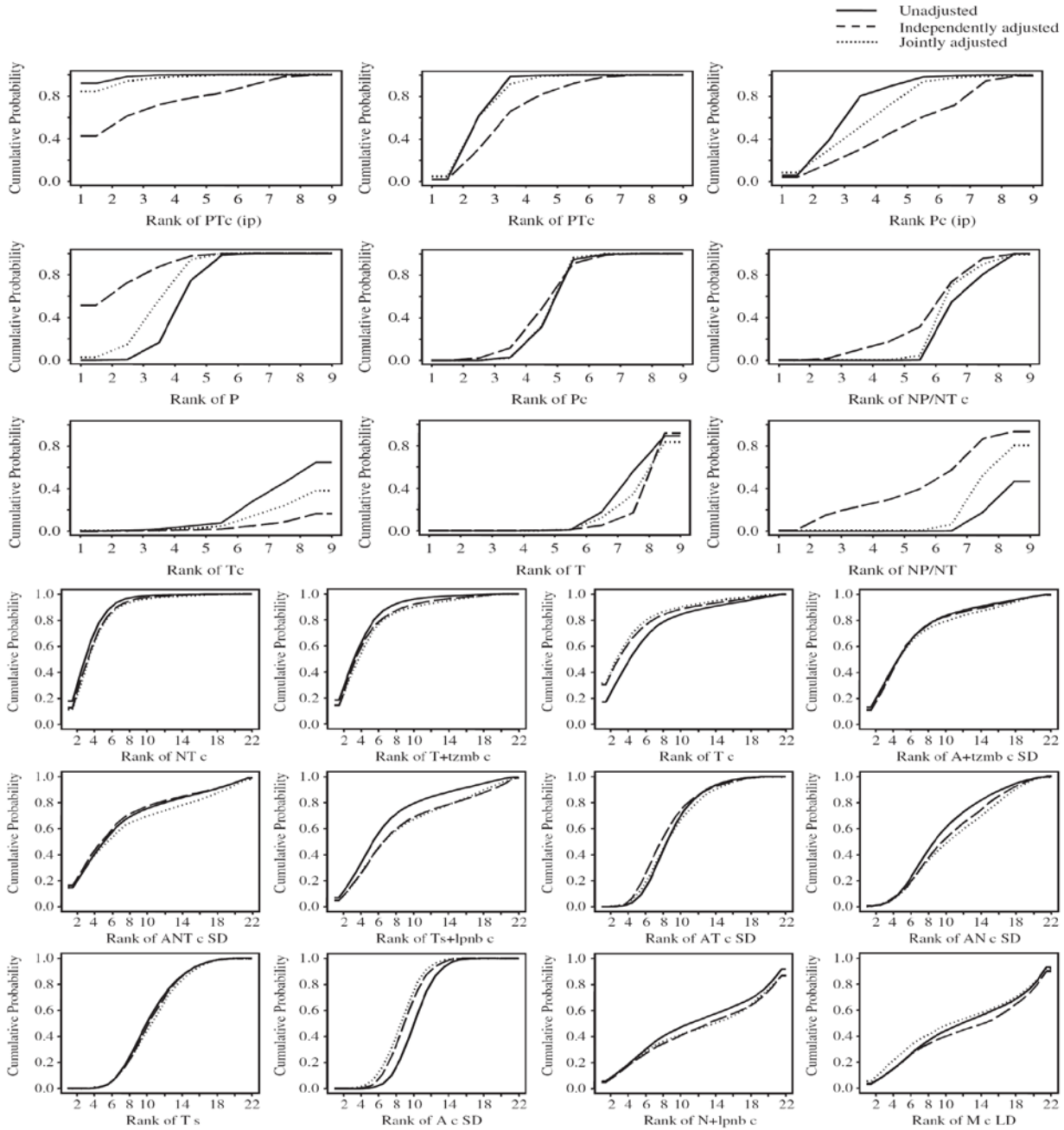


Figure A1. Cumulative ranking curves for the ovarian and breast cancer treatments obtained from the unadjusted model (solid line), the independently adjusted (dashed line) and jointly adjusted model (dotted line). The surface under each cumulative curves expressed as percentage is presented in Appendix Table AII.