Evidence-Based Medicine and Statistics

Introductory Overview Lecture

JSM 2011

Christopher Schmid

Tufts University

1 August 2011

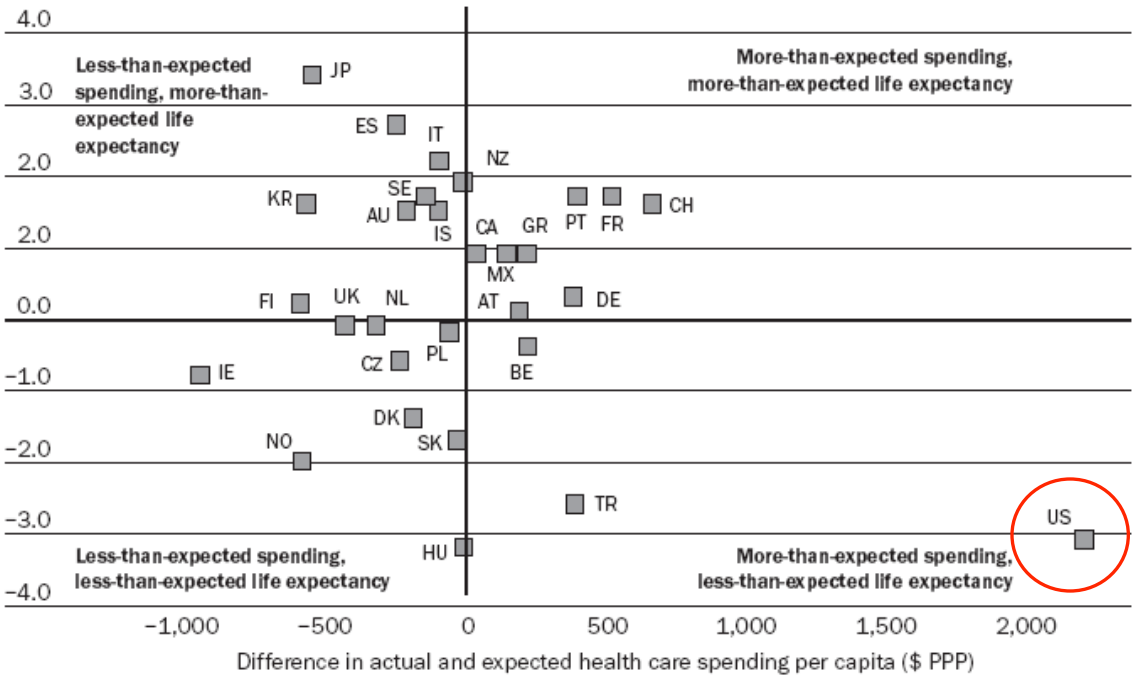
Evidence-Based Medicine

Evidence-based medicine is the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients

Sackett et al. Oxford. *CEBM*, *BMJ*. 1996;312:71-2.

Healthcare Spending and Quality

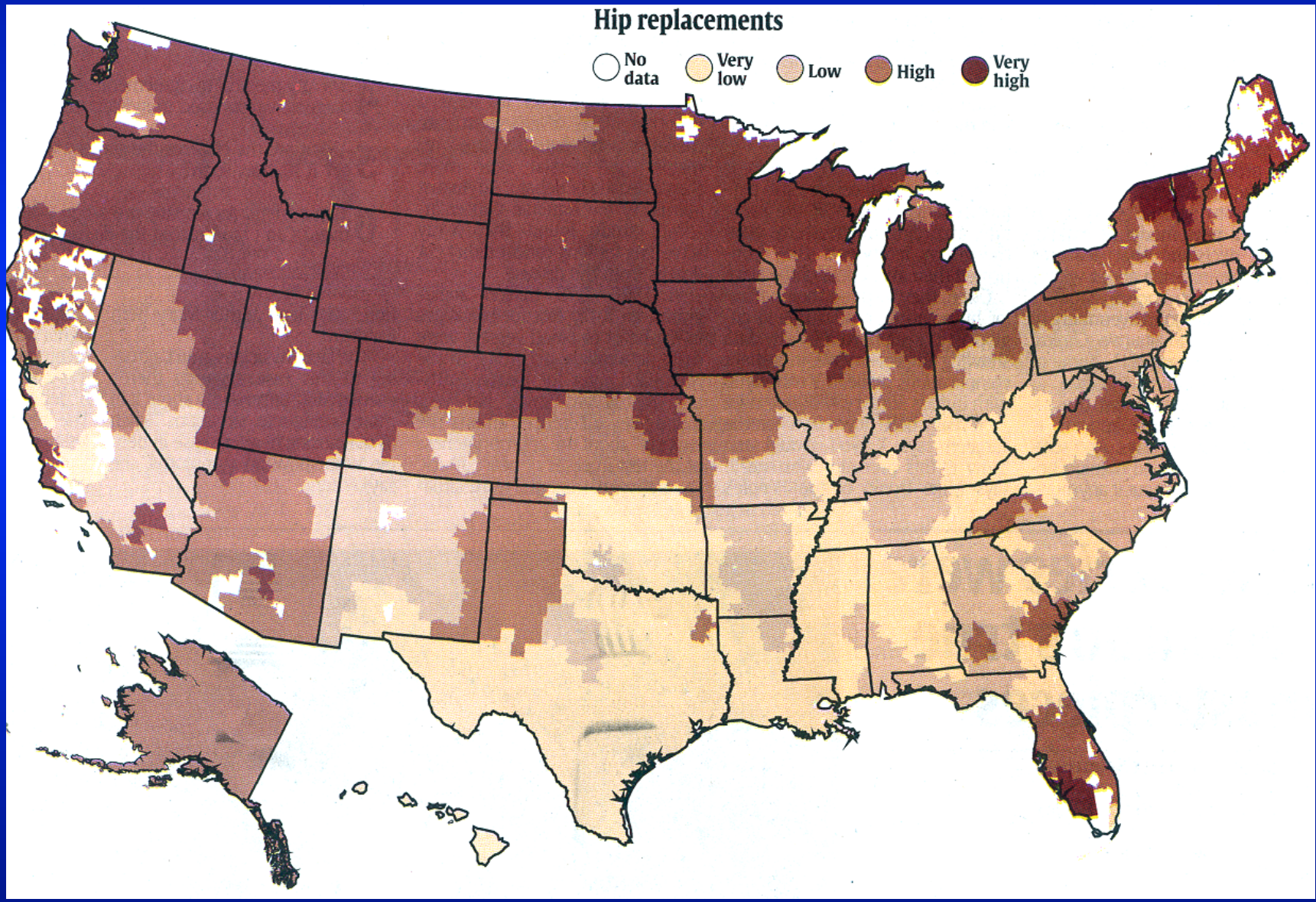
Difference Between Actual And Expected Health Care Spending Per Capita And Actual And Expected Life Expectancy In Organization For Economic Cooperation And Development (OECD) Countries, 2005



SOURCE: Organization for Economic Cooperation and Development, *OECD Health Data, 2007* (Paris: OECD, 2007).
NOTES: Regression equation for expected health spending is $y = 0.1174x - 706.35$ with $R_2 = 0.79$, where y is health care spending per capita ($ purchasing power parity, or PPP) in 2005 and x is gross domestic product (GDP) per capita ($ PPP) in 2005. Regression equation for expected life expectancy is $y = 0.0002x - 72.503$ with $R_2 = 0.57$, where y is life expectancy in years in 2005 and x is GDP per capita ($ PPP) in 2005. For details, see Notes 15, 16, and 18 in text. For Australia, Hungary, Japan, and the Netherlands, health spending data for 2004 are used. For Canada and the United States, life expectancy data for 2004 are used. Country abbreviations are spelled out in Exhibit 2. Luxembourg (LX) is omitted from this analysis.

- $2,197 per capita more than expected
- 3.1 life years less than expected

How health care varies by region



Percentage of Acute Otitis Media Patients Given Antibiotics

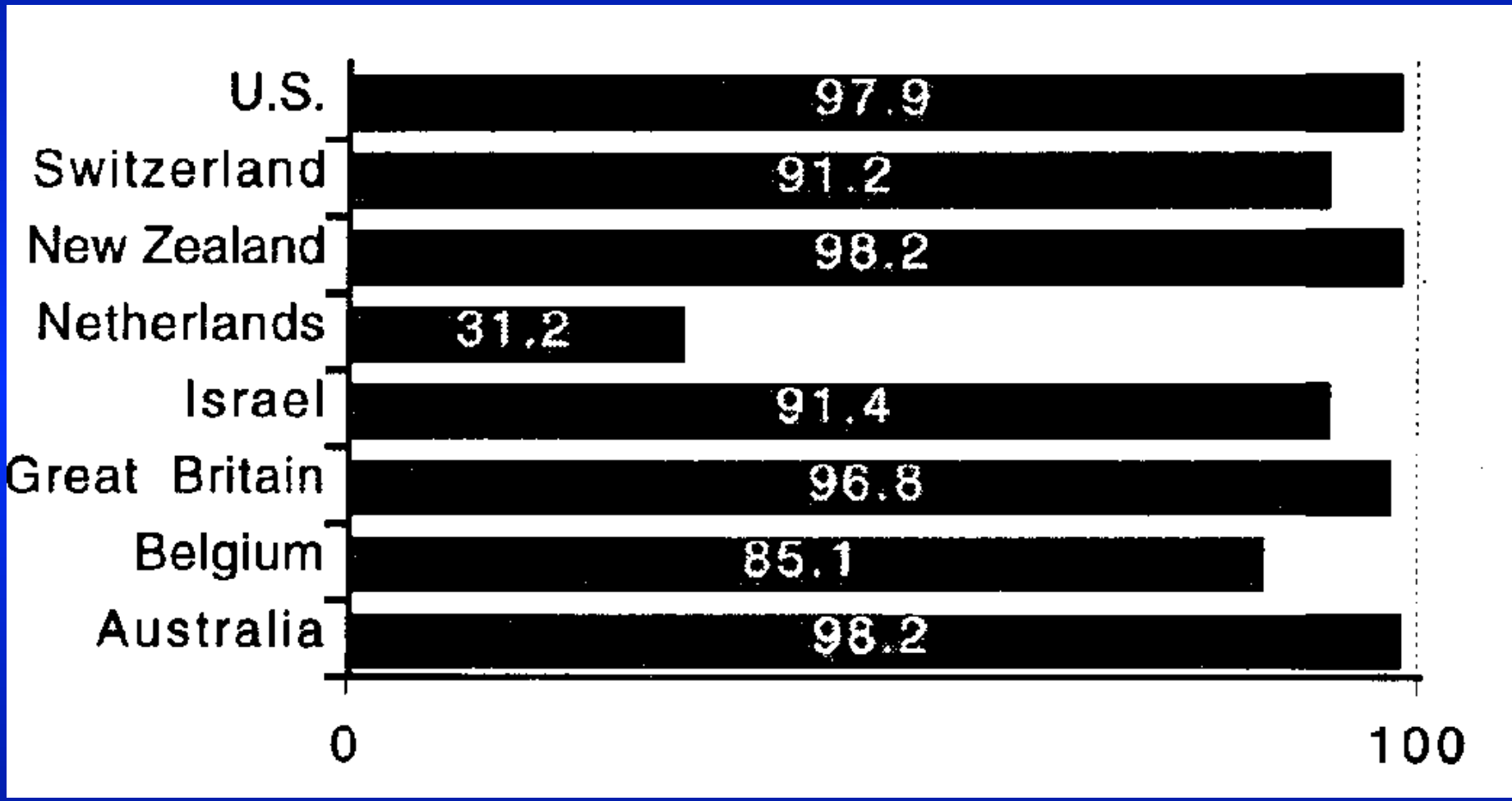


Figure taken from Froom J et al. Diagnosis and antibiotic treatment of acute otitis media: report from International Primary Care Network. BMJ 1990;300:582-6.

Tradition-based Medicine

- Emphasizes
  - primacy of knowledge
  - experience
  - intuition in exercising good clinical judgment
- Observational
- Susceptible to bias
- Individual experiences limited and problems heterogeneous
- Lack of conceptual framework for synthesizing evidence
- Lack of conceptual framework for clinical decision making

Evidence-Based Medicine

Stresses

- examination of evidence from clinical research
- systematic collection of evidence
- synthesis of evidence

De-emphasizes

- intuition
- unsystematic experience
- pathophysiological rationale (surrogates)

Broad View of Clinical Research

- Improve health outcomes of individual patients and society
- Translate (basic) science discoveries into clinical practice
- Optimize use and delivery of healthcare technologies in society
- Provide information to guide
  - Patient management
  - Individual decision making
  - Policy decision making
    - Public health
    - Reimbursement
  - Research agenda of funding agencies

Limitations of Current Best Evidence

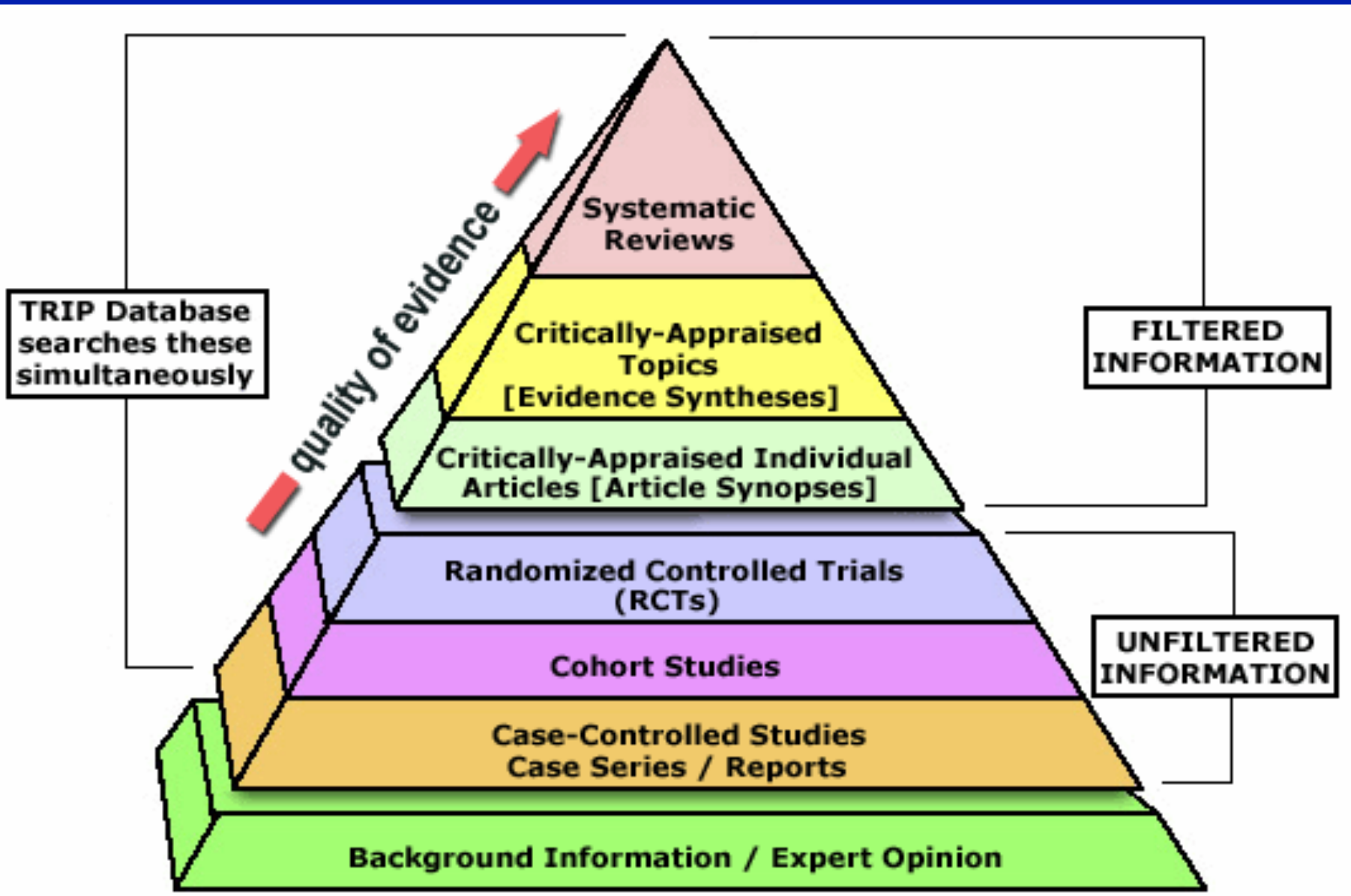
Little evidence about which treatments work best for which patients

- Summary results
- Trial and study exclusions
- Poor comparators

Little evidence about whether the benefits of more expensive therapies warrant their additional costs

- Few RCTs include a cost study
- Poor data
- Skepticism about cost effectiveness analysis, simulation and other decision analysis methods that incorporate cost information

Hierarchy of Evidence



Comparative Effectiveness Research (CER)

Institute of Medicine Definition

CER is the generation and synthesis of evidence that compares the benefits and harms of alternative methods to prevent, diagnose, treat, and monitor a clinical condition or to improve the delivery of care. The purpose of CER is to assist consumers, clinicians, purchasers, and policy makers to make informed decisions that will improve health care at both the individual and population levels.

Evidence-Based Science

“evidence-based”	35,200,000
“evidence-based medicine”	1,880,000
“evidence-based practice”	1,390,000
“evidence-based nursing”	525,000
“evidence-based healthcare”	374,000
“evidence-based mental health”	168,000
“evidence-based nutrition”	467,000
“evidence-based dentistry”	156,000
“evidence-based pediatrics”	33,900
“evidence-based surgery”	33,700
“evidence-based veterinary medicine”	362,000
“evidence-based management”	4,280,000
“evidence-based social”	2,200,000
“evidence-based education”	66,800
“evidence-based marketing”	1,270,000
“evidence-based politics”	44,100
“clinical practice guideline”	867,000
“systematic review”	1, 970,000
“meta-analysis”	3,880,000

Source: Google – December 12, 2010

Evidence-Based Medicine

1) Systematic Reviews and Meta-Analyses
2) Randomized Controlled Clinical Trials
3) Observational Studies
4) Case reports

Special case: decision modeling, including simulations and cost effectiveness analysis

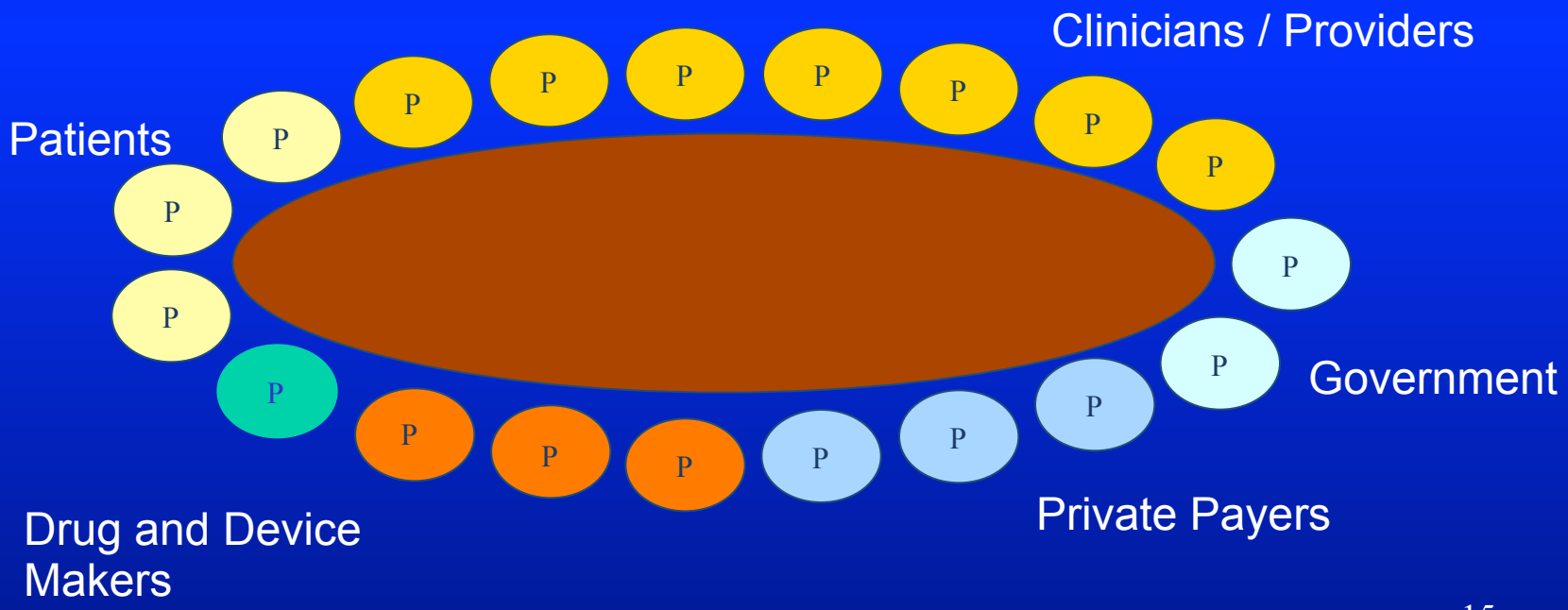
Patient Centered Outcomes Research Institute (PCORI)

- Independent agency outside US Government

- Roles and responsibilities
  - Set research priorities
  - Determine project agenda and methods to be used
  - Award contracts with preference to NIH and AHRQ
  - Appoint expert advisory panels
  - Develop methods and methods standards
  - Conduct peer review
  - Disseminate research findings

PCORI Governing Board

- AHRQ Director
- NIH Director
- 19 Stakeholders – clinicians, patients, researchers, consumers



An early Clinical Trial (N = 2)

In the late 18th century, King Gustav III of Sweden decided that coffee was poison and ordered a clinical trial

Intervention: Convicted murderer to drink coffee daily

Control: Another murderer to drink tea daily

Outcome: Death

Outcome Assessment: 2 physicians to determine outcome

Results

- Two doctors died first
- King was murdered
- Both convicts enjoyed long life until tea drinker died at age 83 (Age of coffee drinker not reported)

Discussion

- One should not rely on such a small sample size
-
- Perhaps the end point was too hard
- Outcome of trial had no effect on decision makers
- Coffee was forbidden in Sweden in 1794 and again in 1822

Conclusions

- None possible regarding the effect of coffee
- External events and other biases may have confounded result
- Kings shouldn’t mess with clinical trials

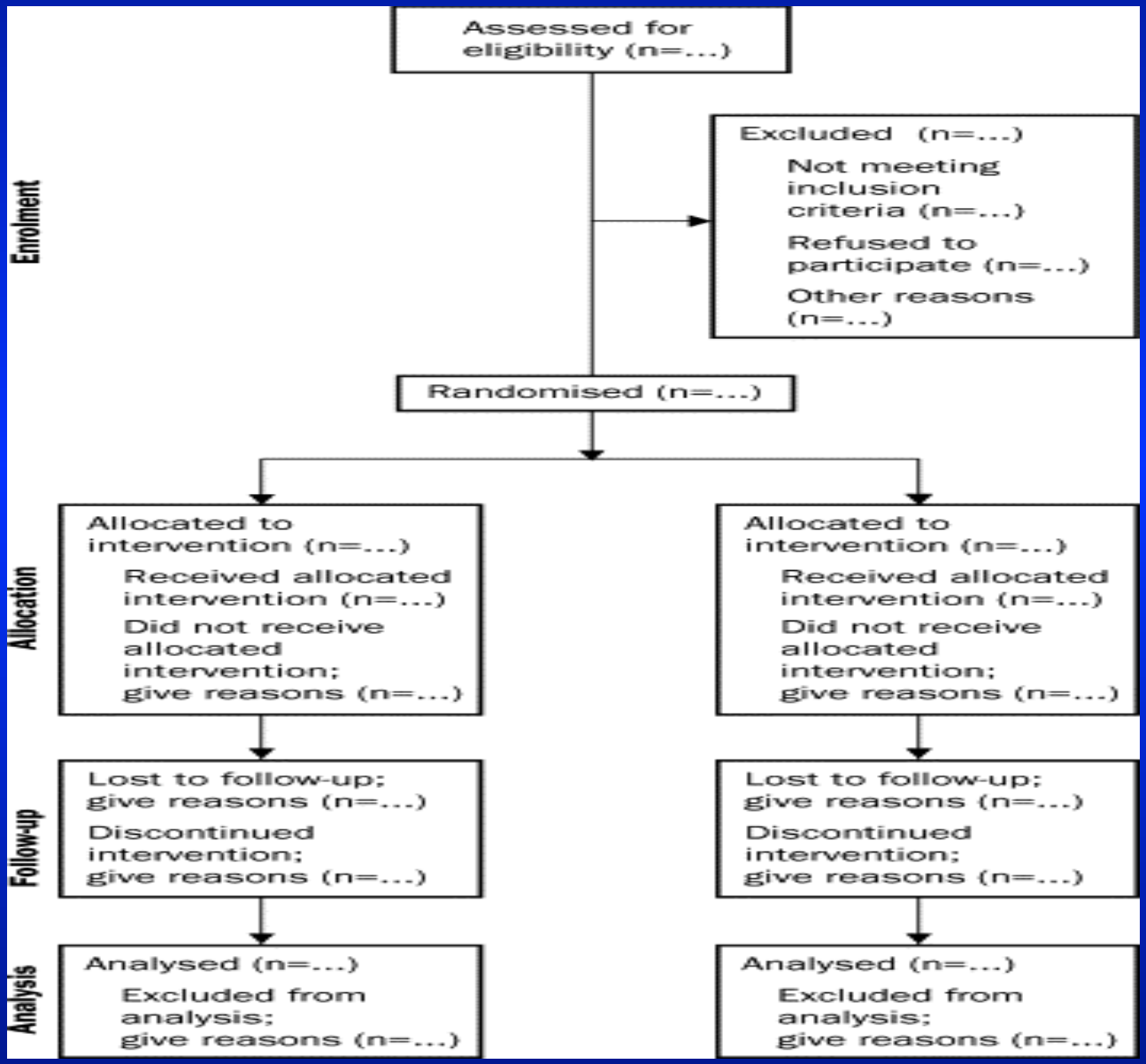
Randomized Clinical Trials (RCTs)

- Use random treatment assignment to determine efficacy of intervention under ideal circumstances
- Patients are randomly assigned to treatment or control groups with pre-and post treatment measurement, double blinding and closely followed treatment protocols
- 1993 conference reviewing quality of publications reporting clinical trials found considerable variation in quality and issued new standard for measuring quality of RCT reports

CONSORT Statement (Consolidated Standards of Reporting Trials)

- Checklist for reporting of 25 items:
  - Title and Abstract
  - Scientific background and rationale
  - Methods
  - Results
  - Discussion
- Flow diagram to describe patient flows through enrollment, intervention allocation, follow-up and data analysis

CONSORT Flow Diagram



Advantages of RCTs

- *A priori* hypothesis
- Internal validity if randomized and controlled
- Near-certain test of efficacy of intervention vs. placebo

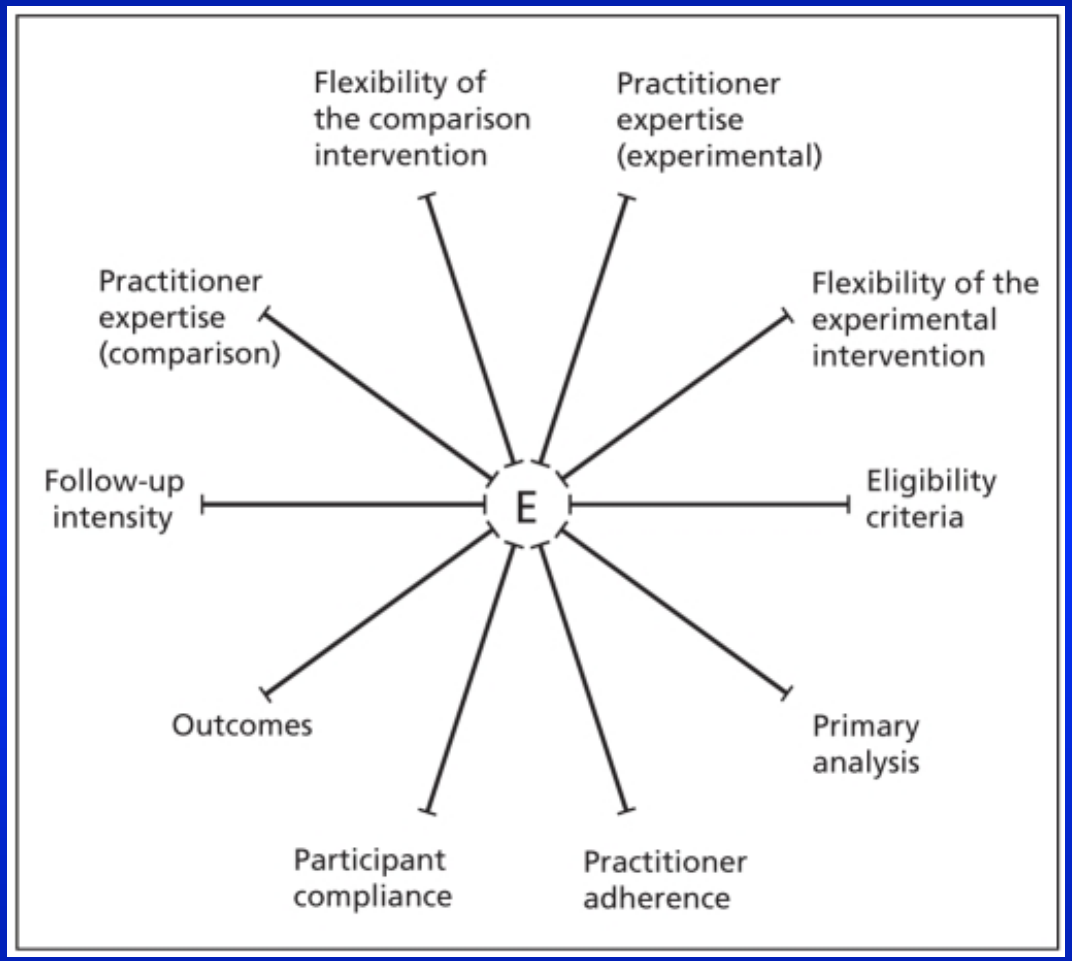
Well-designed clinical trials excel at testing an *a priori* causal hypotheses, typically comparing the effect of an intervention against placebo, for an ideal population, in a controlled setting

Limitations of RCTs

- Limited external validity
- Uncertain effectiveness of the intervention
- Uncertain comparison to alternatives
- Difficult to apply summary results to individual patients
- RCTs are often very slow to produce results

Even well-designed trials may not be very good at determining the effects of an intervention, compared to existing alternatives under the usual conditions in which they are be applied

Pragmatic Trials



Practical or pragmatic trial designed to determine effects of intervention under usual conditions in which it will be applied

ClinicalTrials.Gov

- Website to register RCT protocols and results
- Required by many journals, US funding agencies and FDA
- May reduce problems of publication and reporting bias

Publication Bias

- Negative studies are more likely than positive studies to remain unpublished
- Negative" studies are likely to be small
- In general, not concerned about unpublished "positive" studies.
- Negative studies might invalidate meta-analysis results
- Publication bias is only a part of the bigger "missing data" problem in meta-analysis (and clinical research)
- Selective reporting bias may be a bigger problem

Case Reports (Case Series)

- Detailed report of diagnosis, treatment, and follow-up of individual patient

- Contain some demographic information about patient

Advantages

- Helpful in medical education to describe unusual occurrences
- Development of clinical judgement

Limitations

- Anecdotal evidence
- Limited (to no) generalizibility

Observational Studies

Case-control and cohort designs typically use existing population data, a hypothesis and statistical controls to evaluate a problem or identify associations between an “intervention” and an “outcome”

Other non-randomized designs

- Cross sectional studies
- Surveillance studies using registry data

Advantages

- For retrospective approaches, readily available data
- Faster results
- Hypothesis-generating

Limitations

- Confounding
- Limited causal inference
- Limited external validity (often, not always)

Major Impacts of Non-randomized Evidence

- Lind, 1747, 6 pairs of sailors with scurvy
- Jenner, smallpox, late 18th century
- Fleming, penicillin, 1928-1940s

Observational Study Findings Later Disproved

- Hormone therapy / cardio-protective effects of estrogen
- β carotene and α-tocopherol and cancer
- Fiber and colon cancer

Major Impacts of Randomized Evidence

- Streptomycin for tuberculosis
- Polio vaccine
- Treatments for acute myocardial infarction
- Estrogen Replacement Therapy

Observational vs. Randomized Evidence

- Treatment effects in RCTs and observational studies on same topic tend to be highly correlated
  - Discrepancies occur in about 1 out of 6 cases, even when accounting for between-study heterogeneity
  - Discrepant pairs tend to show more favorable results in observational studies

- Discrepancies in magnitude of effect very common
- Observational studies exhibit larger variability in treatment effects than RCTs
- Discrepancies more common with retrospective designs

Systematic Review

- Scientific discipline to combine information across studies using defined protocol to answer focused research question(s)
- Formulate well-focused study question
- Establish eligibility criteria (study, patient, and disease characteristics, intervention, comparator, outcomes)
- Review literature comprehensively
- Identify relevant studies
- Extract data
- Critically appraise study quality and conclusions

Meta-Analysis

- Quantitative analysis of data from systematic review
- Estimate effect size and uncertainty (treatment effect, association, test accuracy) by statistical methods
- Combine “under-powered” studies to give more definitive conclusion
- Explore heterogeneity / explain discrepancies
- Identify research gaps and need for future studies

Systematic Reviews and Meta-Analyses

Advantages

- Resolve inconsistent studies
- Guide clinical research w/ new hypotheses
- Identify effects earlier through cumulative analysis

Limitations

- Difficult to identify all relevant studies (limitations of electronic searches + publication bias)
- Difficult to judge the quality of all identified studies
- Difficult to apply summary results to individual patients
- Difficult to account for between-trial differences

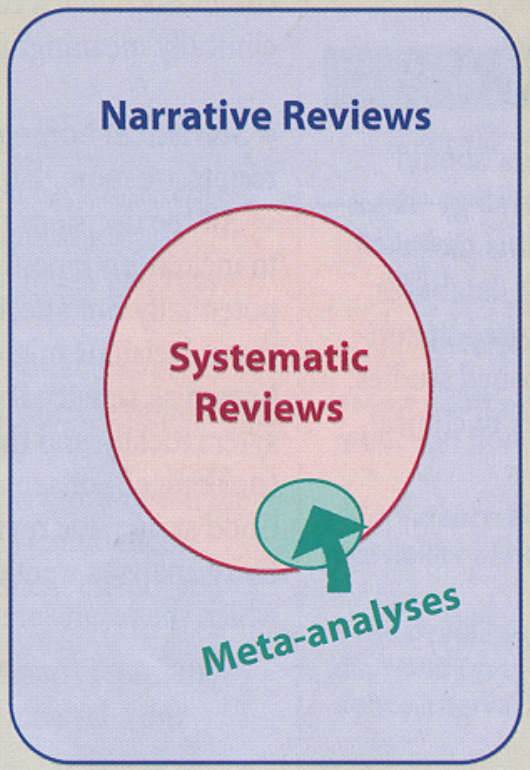
Comparing systematic reviews with narrative "non-systematic" reviews

Narrative Reviews

Give panoramic view, usually cover whole topic. Example: textbook chapters

Emphasize "background" knowledge:
What causes the disorder?
What are are the clinical manifestations?
What treatment options are available?

Susceptible to bias in selecting, appraising and combining studies to answer questions



Systematic Reviews

Give telescopic view, usually address one question or a few questions

Focus on "foreground" knowlege: For example, in treating patients with this disorder, which of the two available treatments is better at improving clinical outcomes safely?

Use rigorous methods to minimize bias and help improve reliability and accuracy of conclusions

Can provide pooled estimates of treatment benefits and risks

Semi-annual newsletter• VA Center of Excellence•San Antonio and Charleston•Fall 1998

Applying SR and MA in Healthcare

- Interventions (most common) estimate efficacies and harms of treatments
- Epidemiologic (many) to provide more reliable estimates of risks, associations
- Diagnostic tests (increasing) provide more reliable estimates of diagnostic accuracy of tests
- Genomics (rapidly increasing) estimate effects of microarray and GWAS studies
- Health economics

An Early Meta-Analysis

The British Medical Journal Nov. 5, 1904. pp. 1243-46.

REPORT ON CERTAIN ENTERIC FEVER INOCULATION STATISTICS.

PROVIDED BY LIEUTENANT-COLONEL R. J. S. SIMPSON, C.M.G., R.A.M.C.

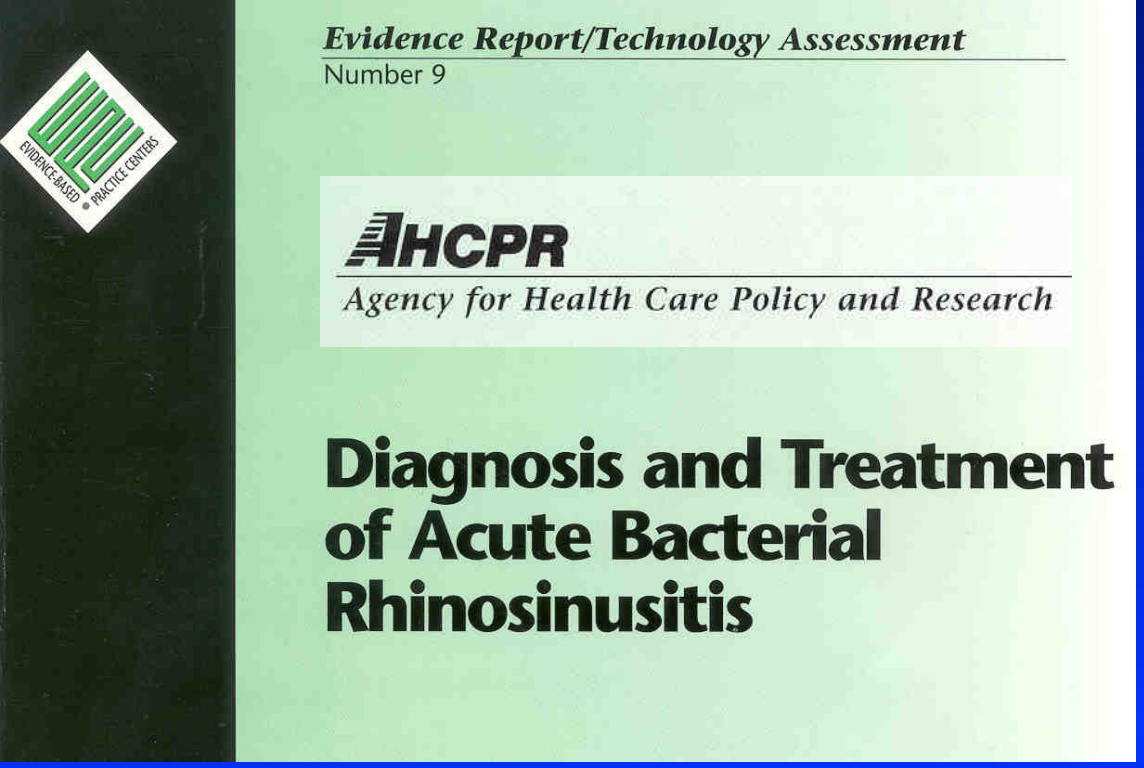
BY KARL PEARSON, F.R.S.,

Professor of Applied Mathematics, University College, London.

THE statistics in question were of two classes: (A) Incidence (B) Mortality Statistics. Under each of these headings the data belonged to two groups: (i) Indian experience; (ii) South African War experience. These two experiences were of a somewhat different character. That for India covered apparently the European army, of whatever branch and wherever distributed; that for South Africa was given partly by locality, partly by column, and partly by special hospital. Thus the Indian and South African experiences seem hardly comparable. Many of the groups in the South African experience are far too small to allow of any definite opinion being formed at all, having regard to the size of the probable error involved. Accordingly, it was needful to group them into larger series. Even thus the material appears to be so heterogeneous, and the results so irregular, that it must be doubtful how much weight is to be attributed to the different results.

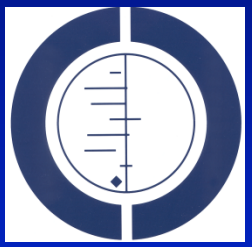
Systematic Review Products

- Journal publications
- Evidence reports
- Comparative effectiveness reviews (CER)
- Technology assessments
- Horizon scans
- Future research needs documents
- Feeders into clinical practice guidelines, coverage, and policy decision making



- About 20 studies with usable primary data for pediatric population
- 450 reports on complication of sinusitis
- 233 narrative reviews

The Cochrane Collaboration



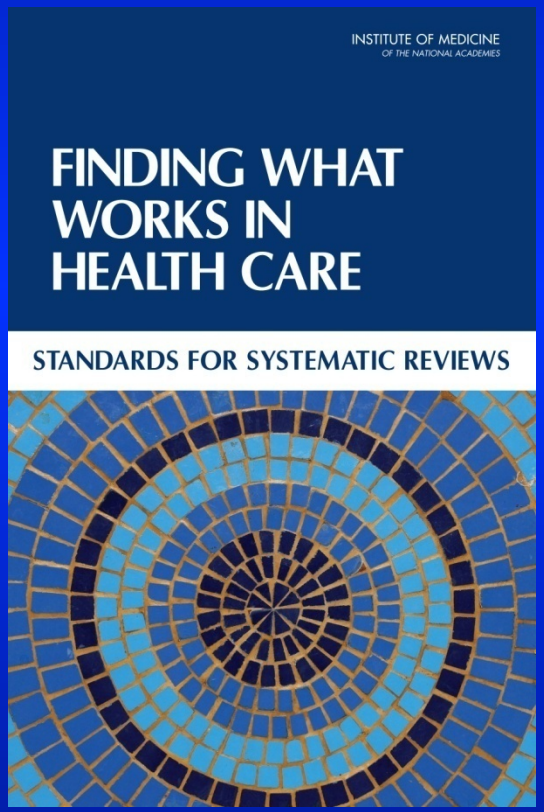
- International collaboration to promote research synthesis
- National centers (one in USA)
- Collaborative review groups organized by clinical area
- Over 2000 meta-analyses published
- Also has register of randomized controlled trials

PRISMA Statement

Checklist of 27 topics to present in Systematic Reviews

1) Background and Methods
2) Data Collection
3) Analysis Plan
4) Results
5) Summary
6) Synthesis
7) Conclusions

Institute of Medicine (IOM) Standards for Systematic Reviews

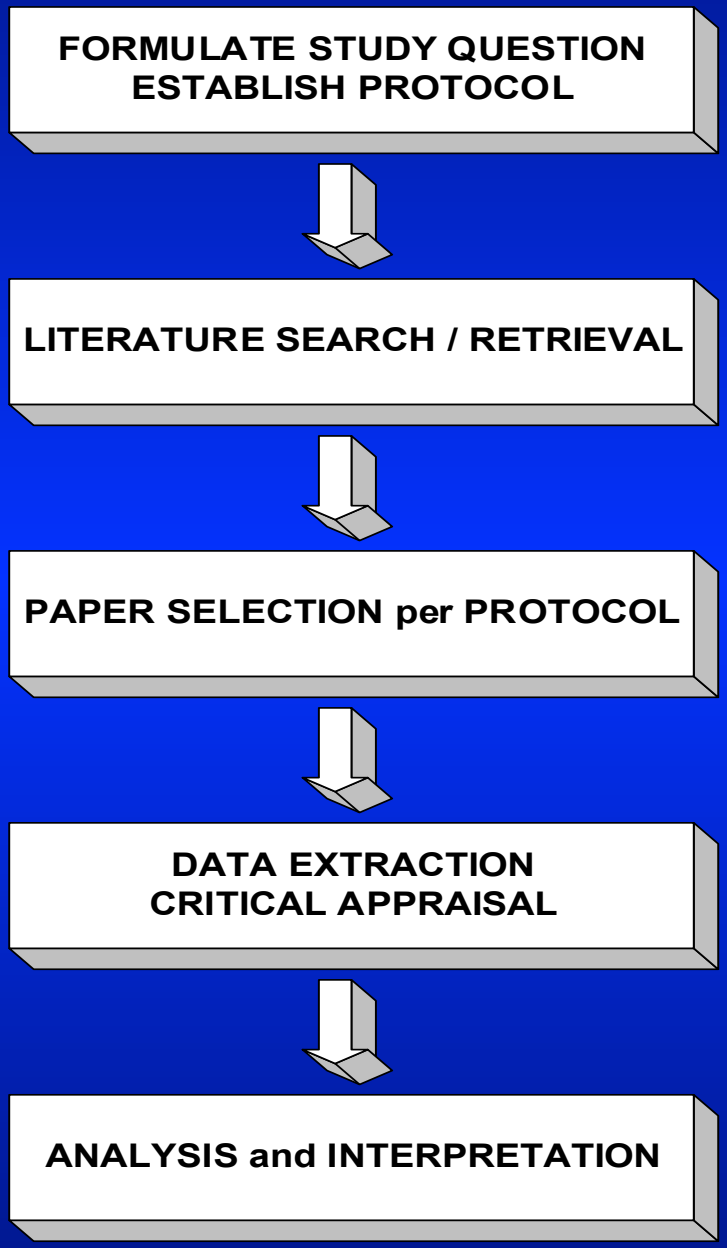


For more information about the report go to www.iom.edu/srstandards

or

www.nap.edu

STEPS OF PERFORMING A META-ANALYSIS



Formulating Answerable SR Questions

- Who is SR for and how will results be interpreted and used?
- Narrow versus broad question (e.g., for individual or population)
- Clinically meaningful and useful (based on sound biological and epidemiological principles)
- Very broadly defined questions may be criticized for mixing apples and oranges
- Very narrowly focused questions may have no data or have limited generalizability and sometimes may lead to misinterpretations
- Include stakeholders, clinicians, methodologists

PICO(TS) Formulation

- Population
- Interventions
- Comparators
- Outcomes
- Timing
- Study design

- Eligibility criteria

Example: The Well - Formulated Question

The Cochrane Collaboration "How to Conduct a Cochrane Systematic Review" 1996

Intervention	Outcomes	Population Setting	Condition of interest

Does ***drug therapy*** decrease ***long-term morbidity and mortality*** in ***older persons*** with ***mild to moderate hypertension***?

ACE inhibitors Angiotensin Receptor Antagonists Combined Alpha and Beta Blockers Calcium-Channel Blockers Diuretics Alpha Adrenergic Blockers Central Sympatholytics Direct Vasodilators Peripheral Adrenergic Antagonist	> 1 year	Fatal and non-fatal strokes Fatal and non-fatal Coronary Heart Disease (MI, sudden death) Cardiovascular events (above plus aneurysm, congestive heart failure, transient ischemic attacks) Total Mortality	> 60 yrs old outpatients	Systolic 140-179 Diastolic 90-109

Identifying the Literature

- Guided by key questions and eligibility criteria
- Comprehensive but practical
  - Search multiple databases
  - Balance between feasibility, resources, and needs
- Minimize selection bias
  - Language: English only?
  - Include unpublished studies?
  - Multiple (overlapping) publications of same data
- Minimize errors
- Often iterative process with question formulation

18,000 citations were screened for the cancer pain evidence report

Principles of Data Extraction

- Extract data needed to survey literature
- Extract data needed to critically appraise study
- Extract data needed to conduct meta-analyses
- Take steps to minimize data extraction errors
  - Data extraction requires methods and domain knowledge
  - Create and test data collection form
  - Train and calibrate data extractors
  - Perform double independent data extraction or extract by one and verify by another

Some Data Extraction Problems

- Data reporting errors
- Non-uniform outcomes (different measurements in different studies)
- Incomplete data (frequent problem: no standard error or confidence interval)
- Discrepant data (different parts of same report gave different numbers)
- Confusing data (can't figure out what authors reported)
- Non-numeric format (reported as graphs)
- Missing data (only conclusion reported)
- Multiple (overlapping) publications of same study

Example of Data Reporting Problem

TABLE 1 Demographic Characteristics of Clinically Evaluable Patients and Overall Description of Pathologies Treated

	Treatment Group	
	Roxithromycin	Clarithromycin
No. of patients	100	95
Sex		
Female	43	44
Male	57	51
Age (years)		
Mean	39.30	40.06
Range	47–92	48–95
Weight (kg)		
Mean	66.42	66.98
Range	47–92	48–95

Another Example of Data Reporting Problem

Data for the 40 patients who were given all four doses of medications were considered evaluable for efficacy and safety.

The overall study population consisted of ten (44%) men and 24 (56%) women, with a racial composition of 38 (88%) whites and five (12%) blacks.

Rationale for Quality Appraisal

- Assess risk of bias and potential effect on conclusions
- Set threshold for inclusion and exclusion of studies in review
  - Use in sensitivity analysis (test robustness)
- Potentially explain differences in results between trials
- Weight statistical analysis of results
  - Quality scores not recommended
- Establish strength of recommendation in guidelines
- But poor reporting may be mistaken for poor quality

Commonly Assessed Quality Features

- Allocation concealment
- Blinding
- Description of intervention
- Withdrawals
- Statistical analysis
- Accuracy of reporting

Types of Data to Combine

- Dichotomous (events, e.g. deaths)
- Measures (odds ratios, correlations)
- Continuous data (mmHg, pain scores)
- Effect size
- Survival curves
- Diagnostic test (sensitivity, specificity)
- Individual patient data

Effect Size

- Dimensionless metric
- Basic idea is to combine standard deviations of diverse types of related effects
- However, availability and selection of reported effects may be biased, variable importance of different effects
- Frequently used in education, social science literature
- Infrequently used in medicine, difficulty in interpreting results

What is the average difference in DBP?

Study	Sample Size	Δ mmHg	95% CI
ANBP	554	-6.2	-6.9 to -5.5
EWPHE	304	-7.7	-10.2 to -5.2
Kuramoto	39	-0.1	-6.5 to 6.3

Simple Average

$$\frac{(-6.2) + (-7.7) + (-0.1)}{3} = -4.7 \text{ mmHg}$$

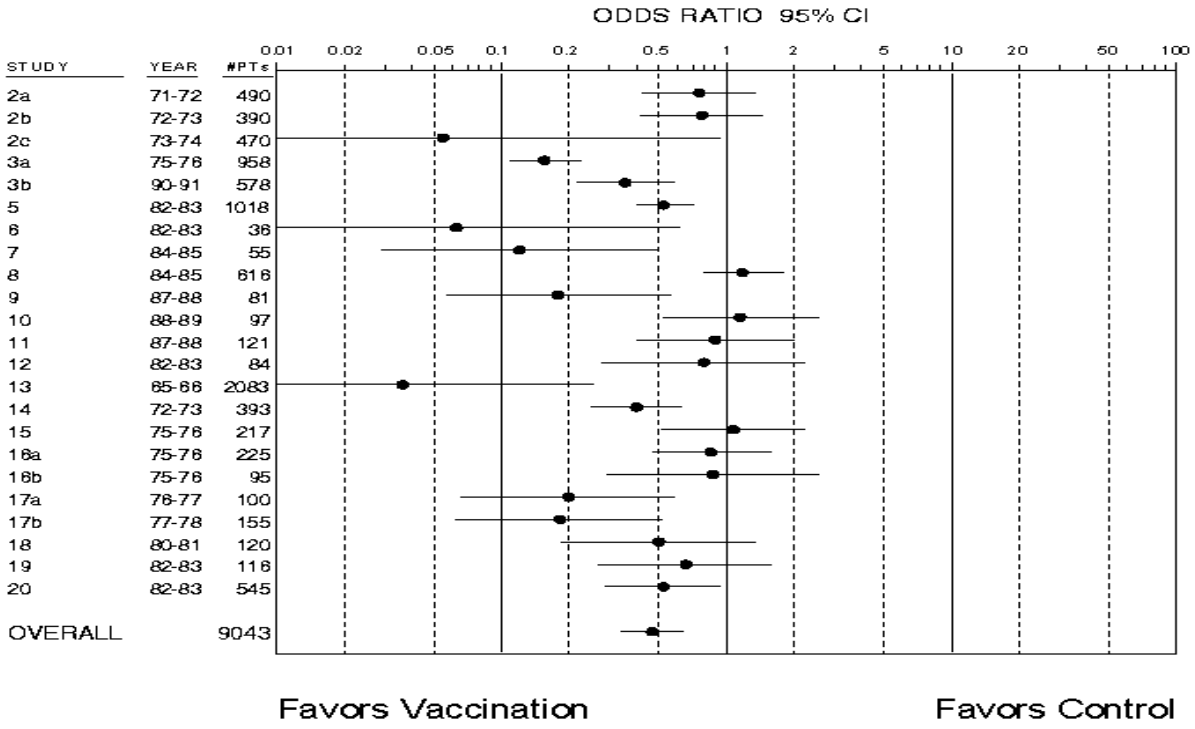
Study	Sample Size	Δ mmHg	95% CI
ANBP	554	-6.2	-6.9 to -5.5
EWPHE	304	-7.7	-10.2 to -5.2
Kuramoto	39	-0.1	-6.5 to 6.3

Average Weighted by Sample Size

$$\frac{(554 \times -6.2) + (304 \times -7.7) + (39 \times -0.1)}{554 + 304 + 39} = -6.4 \text{ mmHg}$$

Study	Sample Size	Δ mmHg	95% CI
ANBP	554	-6.2	-6.9 to -5.5
EWPHE	304	-7.7	-10.2 to -5.2
Kuramoto	39	-0.1	-6.5 to 6.3

Forest Plot: Influenza Vaccine Efficacy



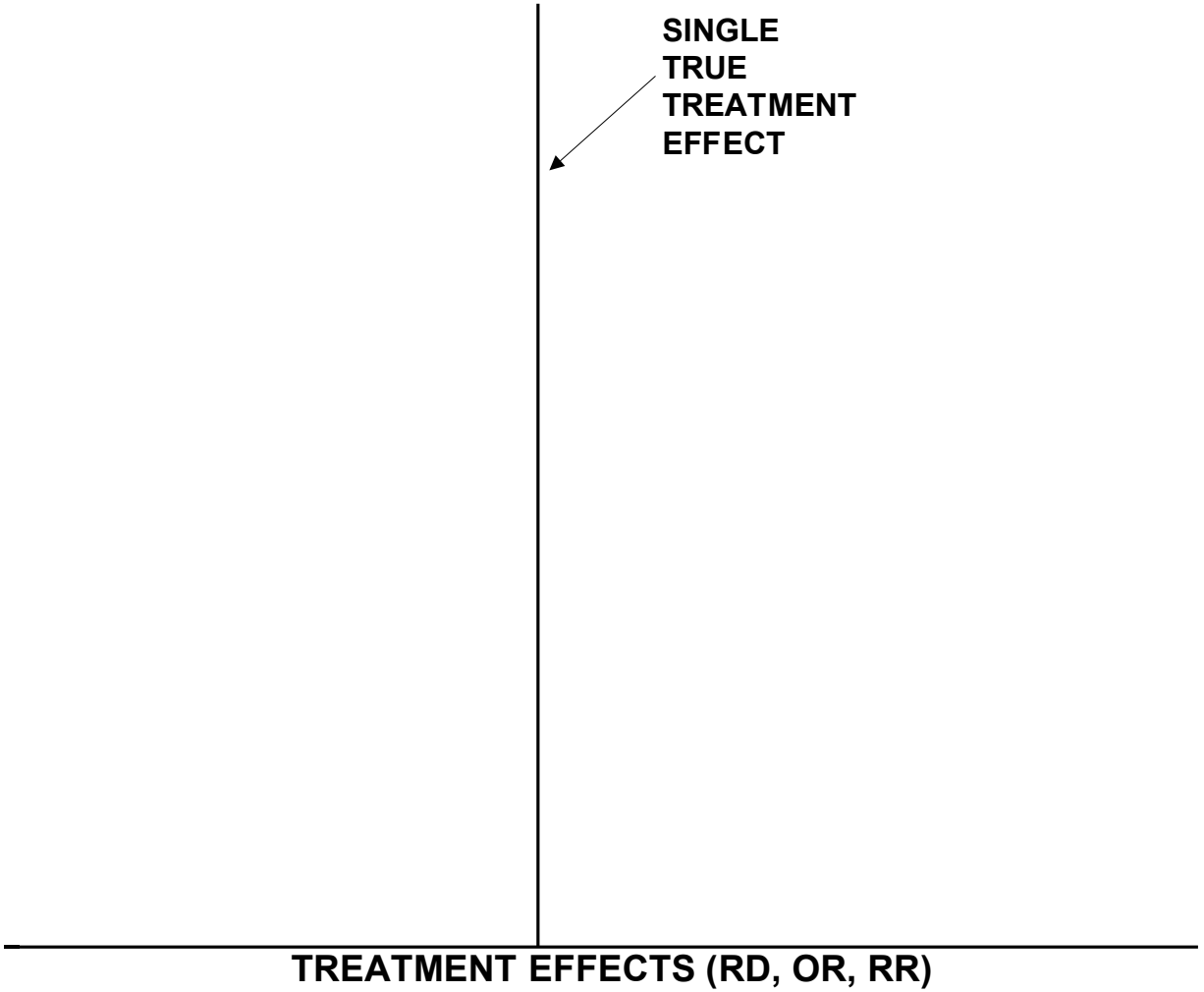
Heterogeneity (diversity)

- Is it reasonable (are studies and effects sufficiently similar) to estimate an average effect?

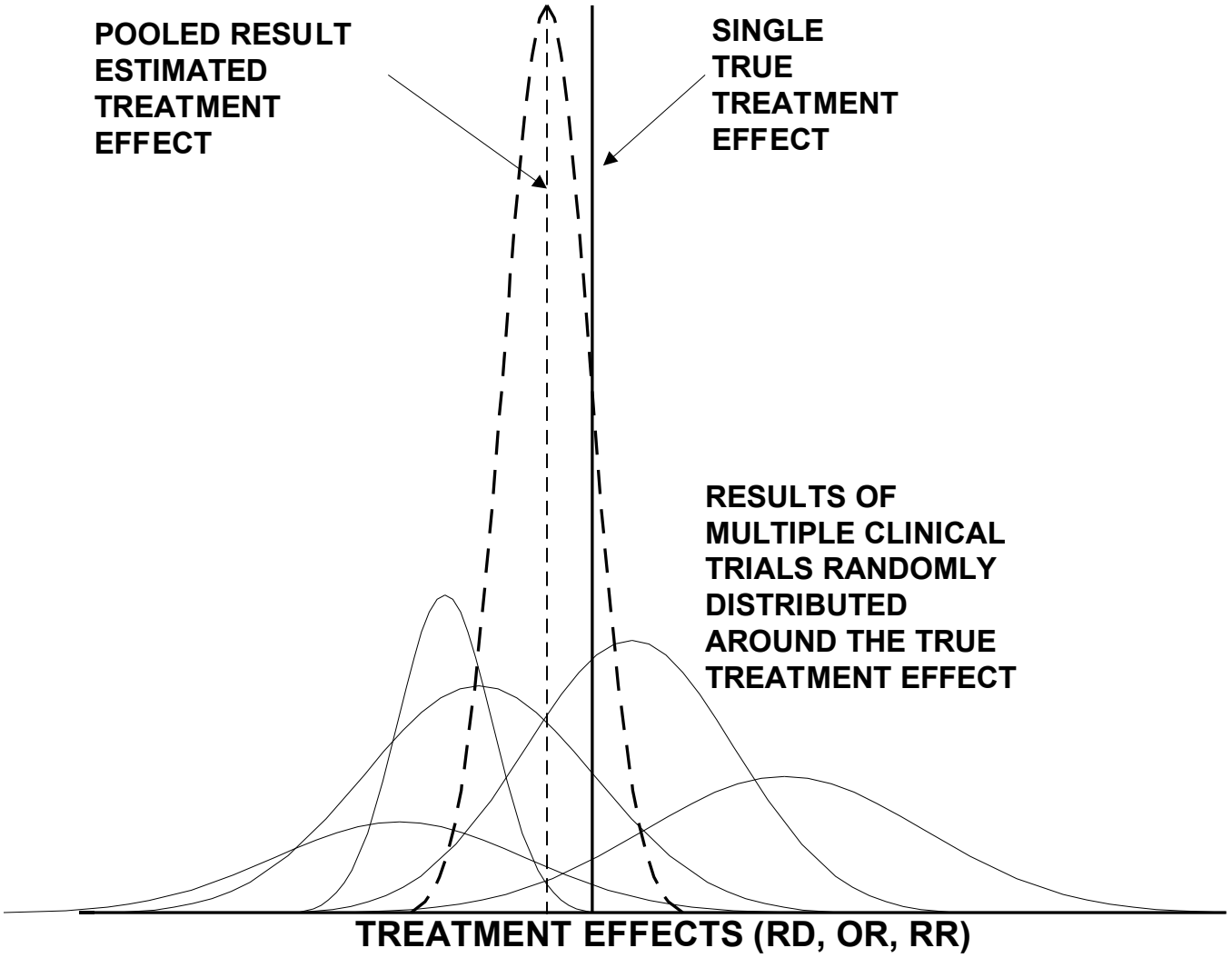
- Types of heterogeneity
  - *Conceptual* (clinical) heterogeneity: Are studies of similar treatments, populations, settings, design, etc., such that an average effect would be clinically meaningful?

  - *Statistical* heterogeneity: Is observed variability of effects greater than that expected by chance alone?

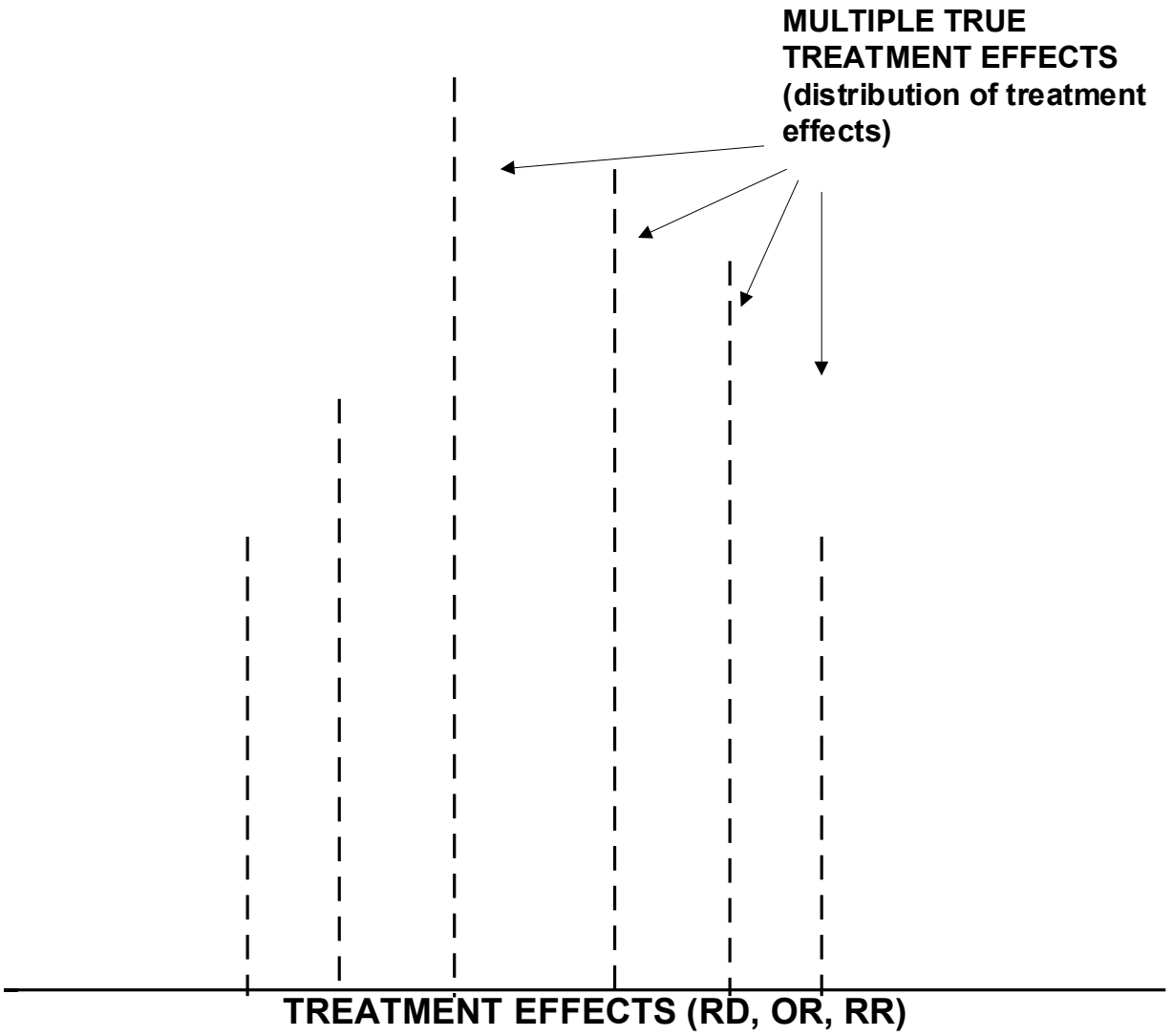
FIXED EFFECTS MODEL



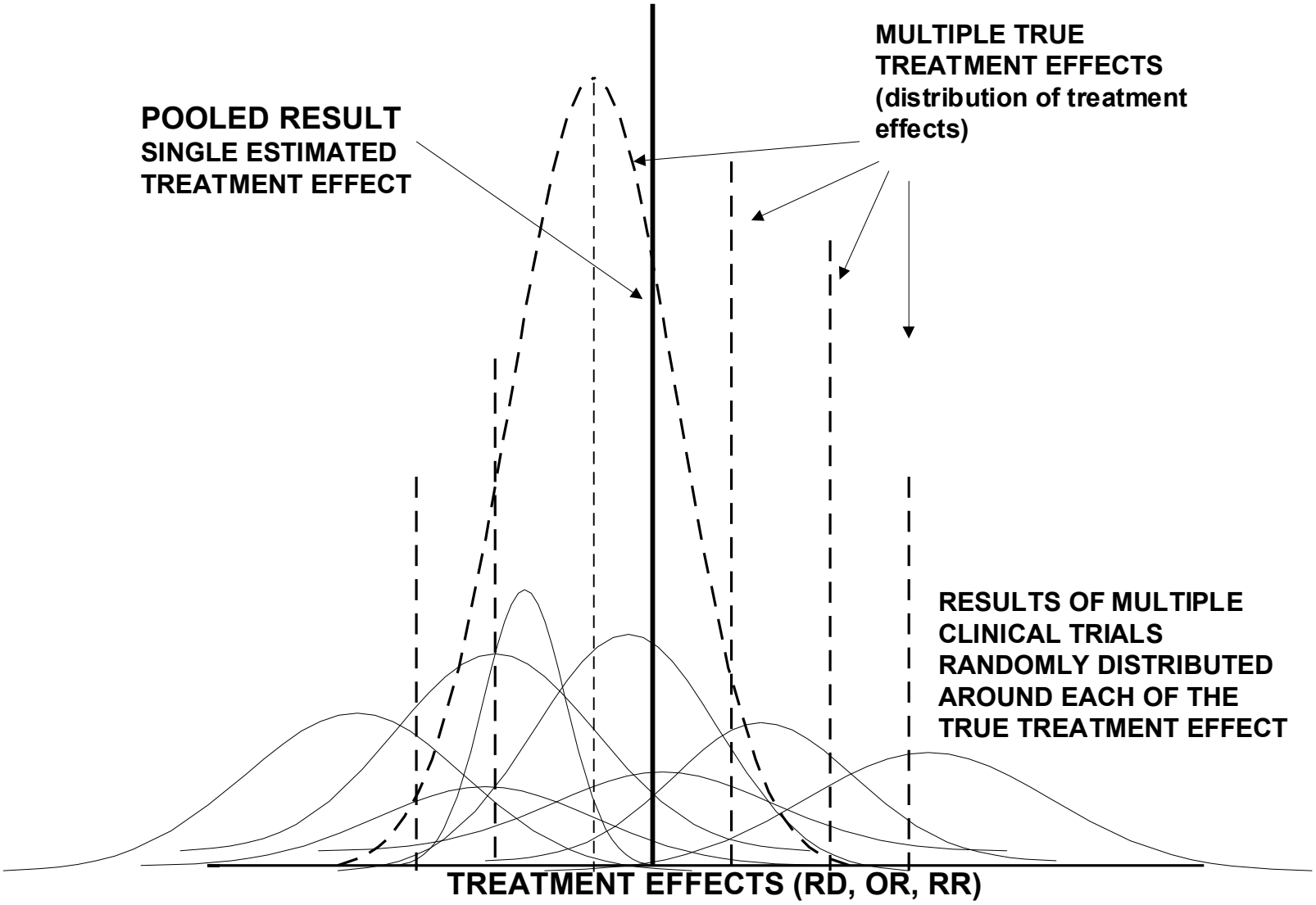
FIXED EFFECTS MODEL



RANDOM EFFECTS MODEL



RANDOM EFFECTS MODEL



General Formula - Weighted Average Effect Size

$$d_+ = \frac{\sum_{i=1}^{k} w_i d_i}{\sum_{i=1}^{k} w_i}$$

d_i = effect size of study i

w_i = weight of study i

k = number of studies

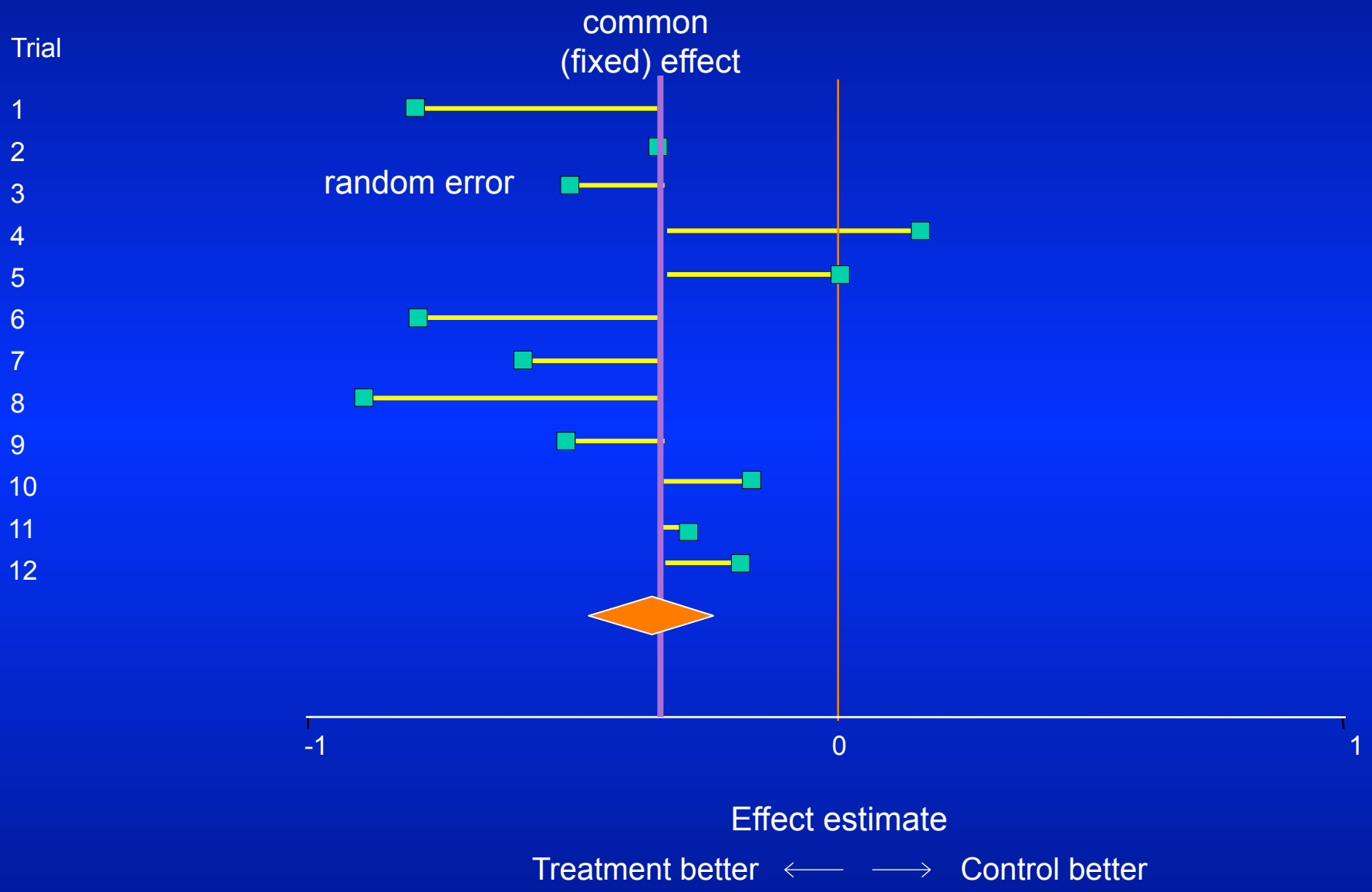
s_i = within study variance

τ^2 = between study variance

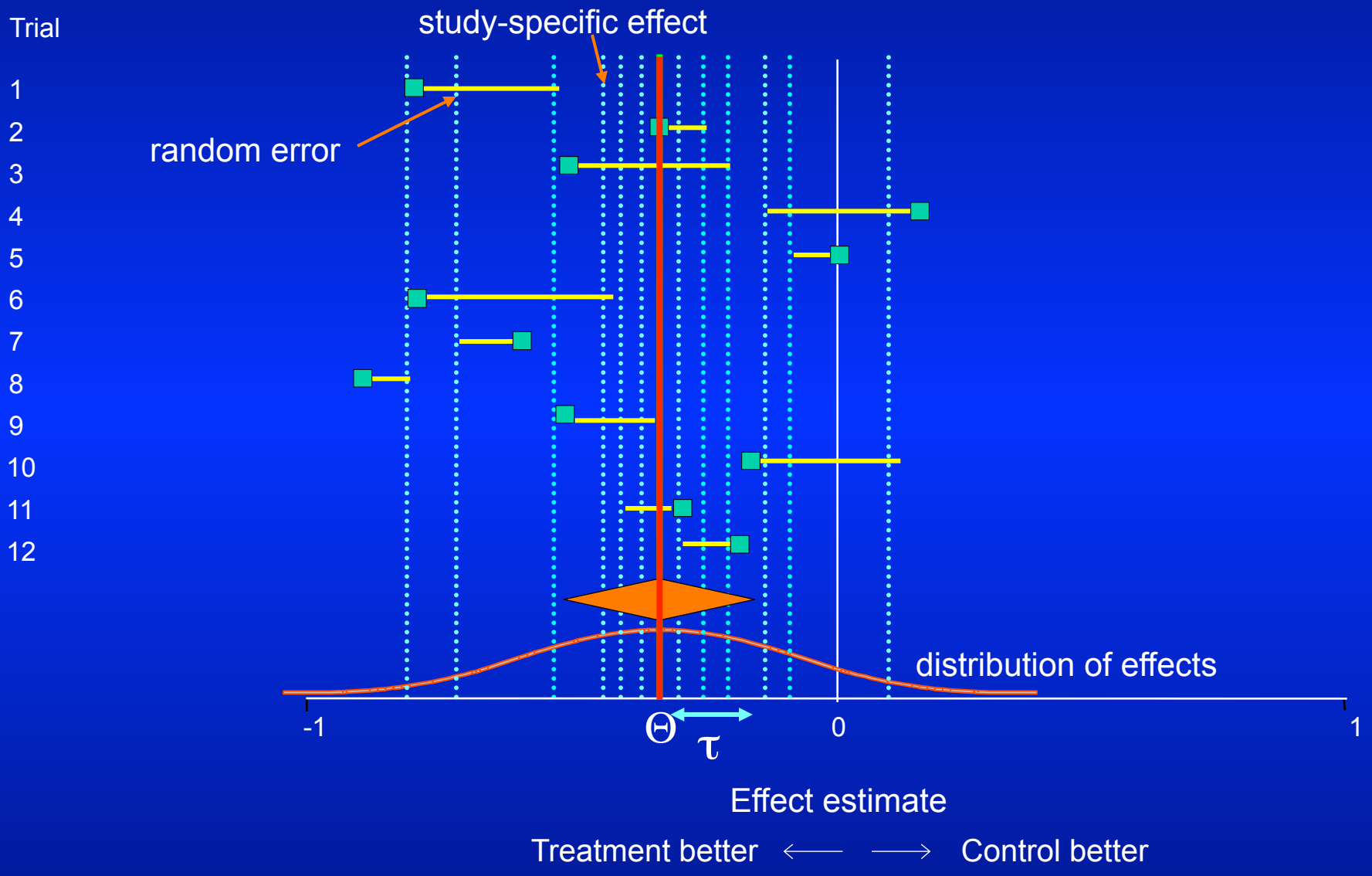
Fixed Effect Weight $W_i = 1/s_i$

Random Effect Weight $W_i = 1/[s_i + \tau^2]$

Fixed Effects Meta-Analysis



Random Effects Meta-Analysis



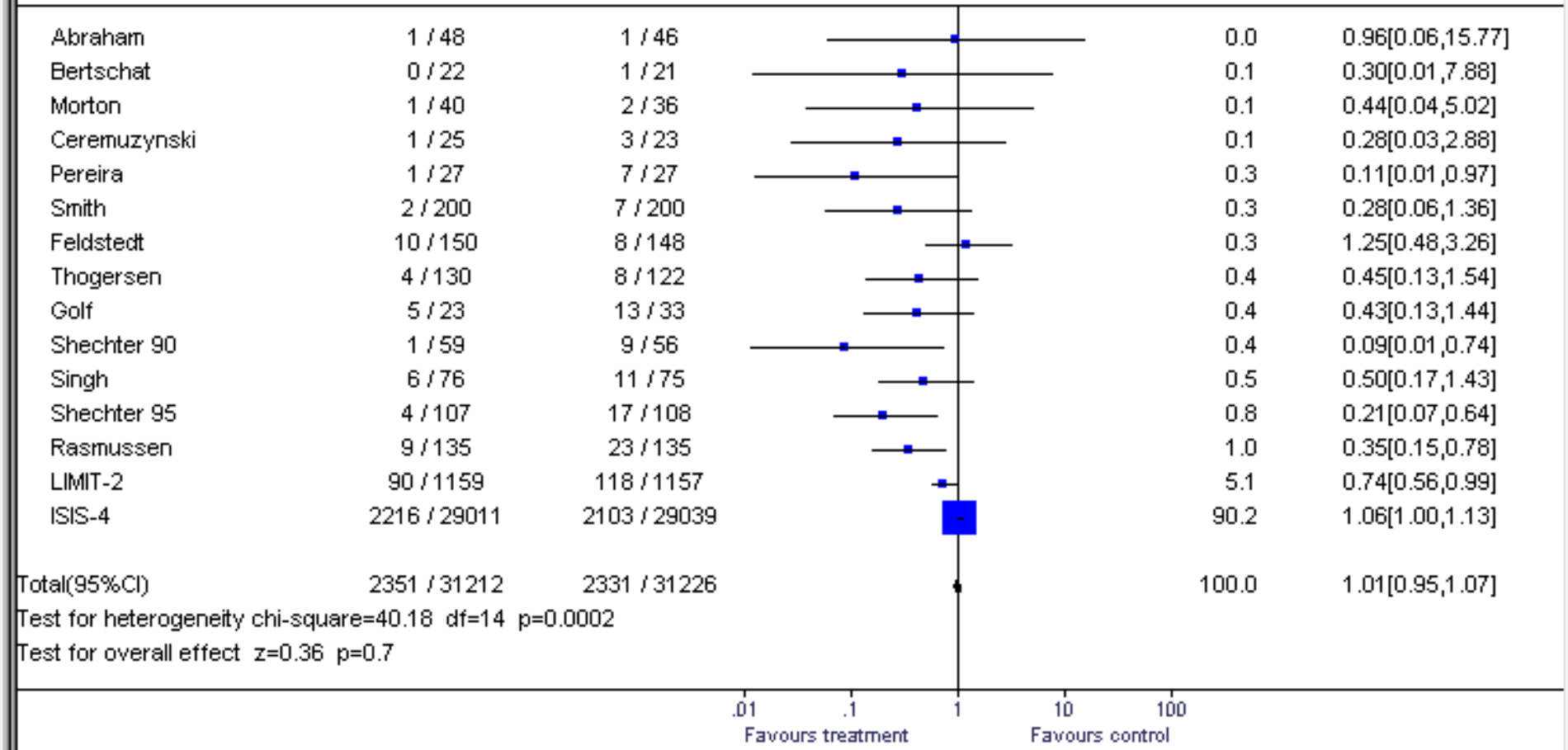
METAVIEW 4.1 [Review: Magnesium for acute AI]

File Display Sort Statistics Previous Outcome Next Outcome Help

Comparison: 13 Magnesium vs placebo
Outcome: 03 Mortality

Study	Treatment n/N	Control n/N	OR (95%CI Fixed)	Weight %	OR (95%CI Fixed)
Abraham	1 / 48	1 / 46		0.0	0.96[0.06,15.77]
Bertschat	0 / 22	1 / 21		0.1	0.30[0.01,7.88]
Morton	1 / 40	2 / 36		0.1	0.44[0.04,5.02]
Ceremuzynski	1 / 25	3 / 23		0.1	0.28[0.03,2.88]
Pereira	1 / 27	7 / 27		0.3	0.11[0.01,0.97]
Smith	2 / 200	7 / 200		0.3	0.28[0.06,1.36]
Feldstedt	10 / 150	8 / 148		0.3	1.25[0.48,3.26]
Thogersen	4 / 130	8 / 122		0.4	0.45[0.13,1.54]
Golf	5 / 23	13 / 33		0.4	0.43[0.13,1.44]
Shechter 90	1 / 59	9 / 56		0.4	0.09[0.01,0.74]
Singh	6 / 76	11 / 75		0.5	0.50[0.17,1.43]
Shechter 95	4 / 107	17 / 108		0.8	0.21[0.07,0.64]
Rasmussen	9 / 135	23 / 135		1.0	0.35[0.15,0.78]
LIMIT-2	90 / 1159	118 / 1157		5.1	0.74[0.56,0.99]
ISIS-4	2216 / 29011	2103 / 29039		90.2	1.06[1.00,1.13]
Total(95%CI)	2351 / 31212	2331 / 31226		100.0	1.01[0.95,1.07]

Test for heterogeneity chi-square=40.18 df=14 p=0.0002
Test for overall effect z=0.36 p=0.7



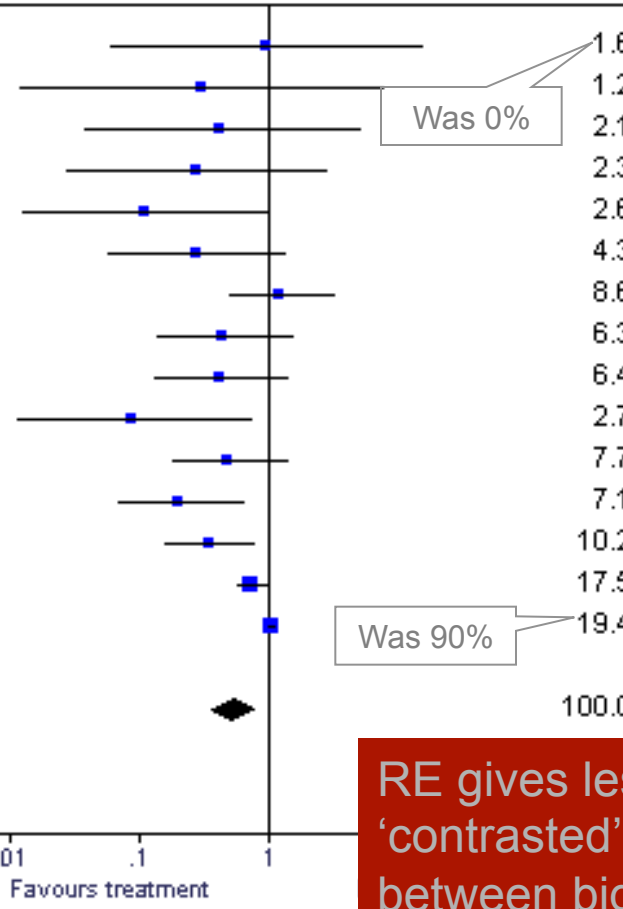
METAVIEW 4.1 [Review: Magnesium for acute AI]

File Display Sort Statistics Previous Outcome Next Outcome Help

Comparison: 13 Magnesium vs placebo
Outcome: 03 Mortality

Study	Treatment n/N	Control n/N	OR (95%CI Random)	Weight %	OR (95%CI Random)
Abraham	1 / 48	1 / 46		1.6	0.96[0.06,15.77]
Bertschat	0 / 22	1 / 21		1.2	0.30[0.01,7.88]
Morton	1 / 40	2 / 36		2.1	0.44[0.04,5.02]
Ceremuzynski	1 / 25	3 / 23		2.3	0.28[0.03,2.88]
Pereira	1 / 27	7 / 27		2.6	0.11[0.01,0.97]
Smith	2 / 200	7 / 200		4.3	0.28[0.06,1.36]
Feldstedt	10 / 150	8 / 148		8.6	1.25[0.48,3.26]
Thogersen	4 / 130	8 / 122		6.3	0.45[0.13,1.54]
Golf	5 / 23	13 / 33		6.4	0.43[0.13,1.44]
Shechter 90	1 / 59	9 / 56		2.7	0.09[0.01,0.74]
Singh	6 / 76	11 / 75		7.7	0.50[0.17,1.43]
Shechter 95	4 / 107	17 / 108		7.1	0.21[0.07,0.64]
Rasmussen	9 / 135	23 / 135		10.2	0.35[0.15,0.78]
LIMIT-2	90 / 1159	118 / 1157		17.5	0.74[0.56,0.99]
ISIS-4	2216 / 29011	2103 / 29039		19.4	1.06[1.00,1.13]
Total(95%CI)	2351 / 31212	2331 / 31226		100.0	0.53[0.36,0.77]

Test for heterogeneity chi-square=40.18 df=14 p=0.0002
Test for overall effect z=-3.34 p=0.0008



RE gives less 'contrasted' weights between big and small studies

Identifying Heterogeneity

- Visualize data
- Statistical test
  - Low power since usually very few studies
  - But has excessive power to detect clinically unimportant heterogeneity with many studies

Quantify amount of heterogeneity

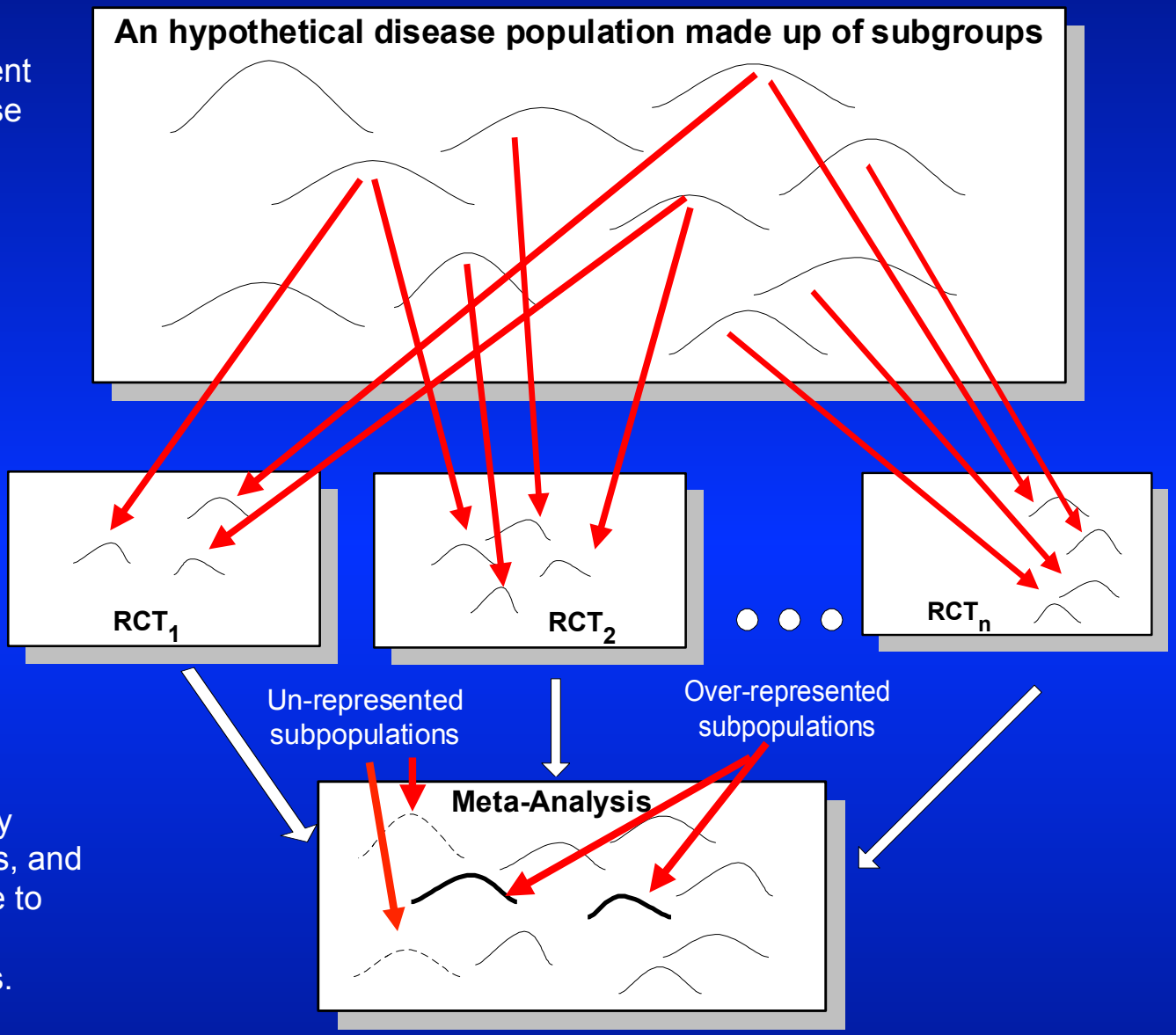
- Between-study variance
- Test Statistic
- Percent of Total Variation Between Studies

Heterogeneity in a disease population, RCTs, and meta-analysis of the trials

Different subgroups representing various patient characteristics and disease manifestations may have different responses to a treatment.

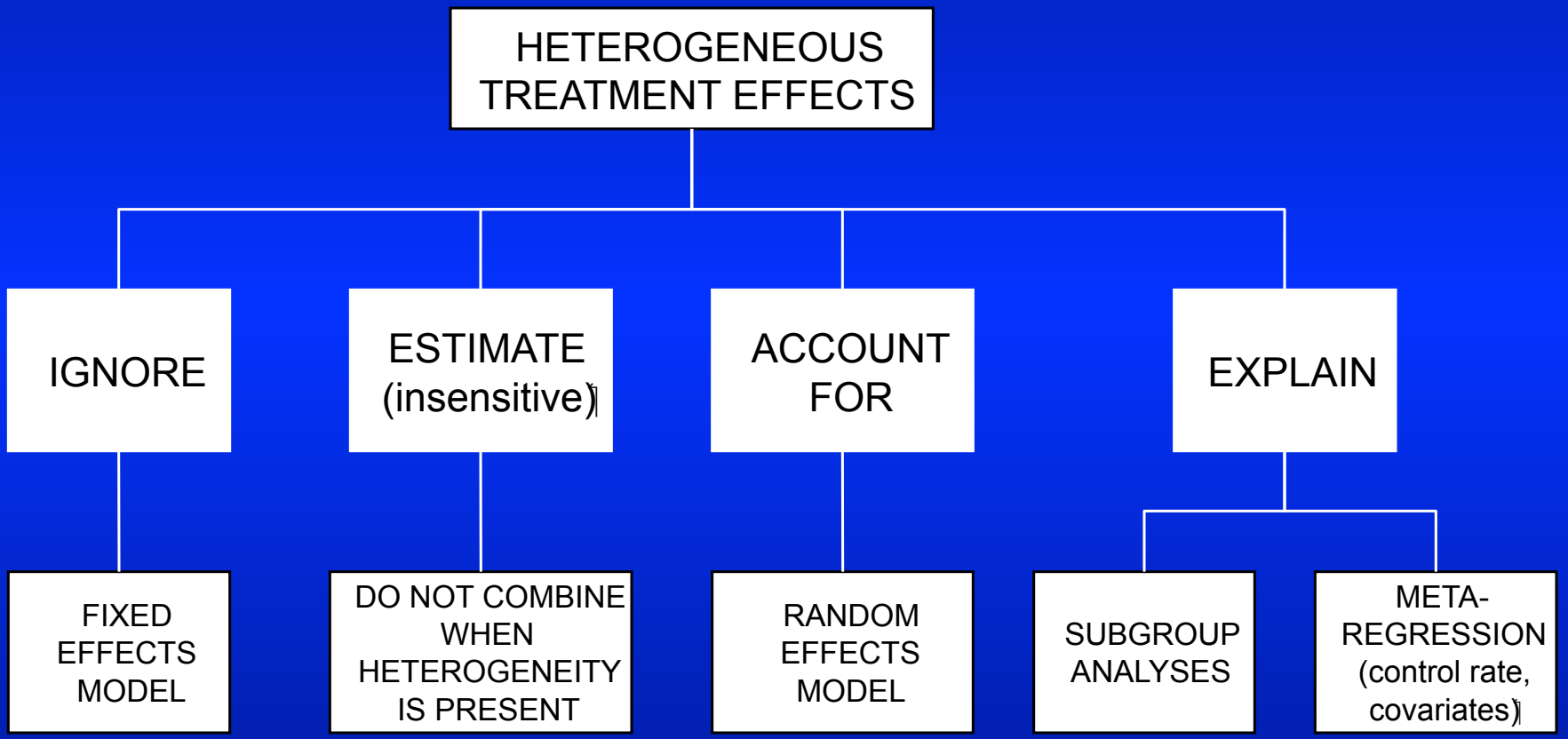
Different inclusion criteria, patient recruitment, and random variations may result in study cohorts consisting or different distributions and combinations of subgroups in RCTs.

Protocol differences, study design and reporting flaws, and publication bias contribute to bias or exclusion of some studies in a meta-analysis.



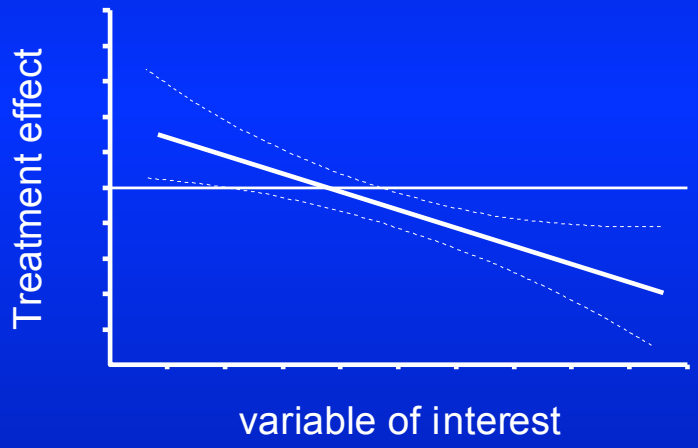
Interpreting the results of meta-analysis of RCTs depends on how the data are synthesized: weighted average, regression, or individual patient data modeling.

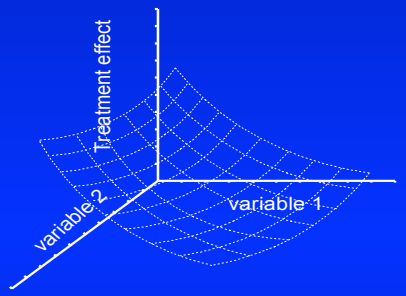
Dealing With Heterogeneity



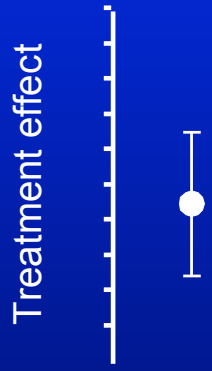
RESPONSE SURFACE
modeling individual patient data

META-REGRESSION
modeling summary data



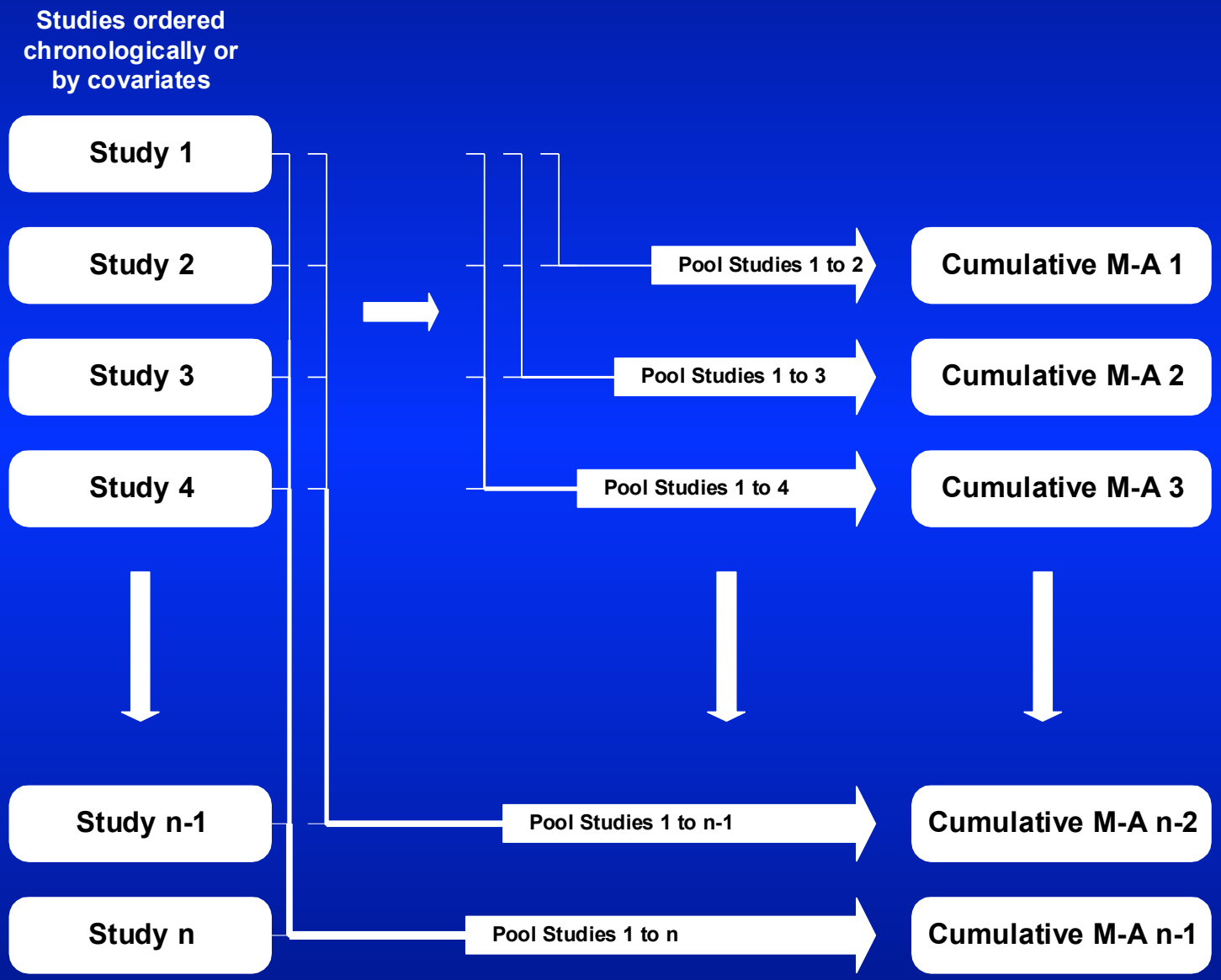


OVERALL ESTIMATE
combining summary data

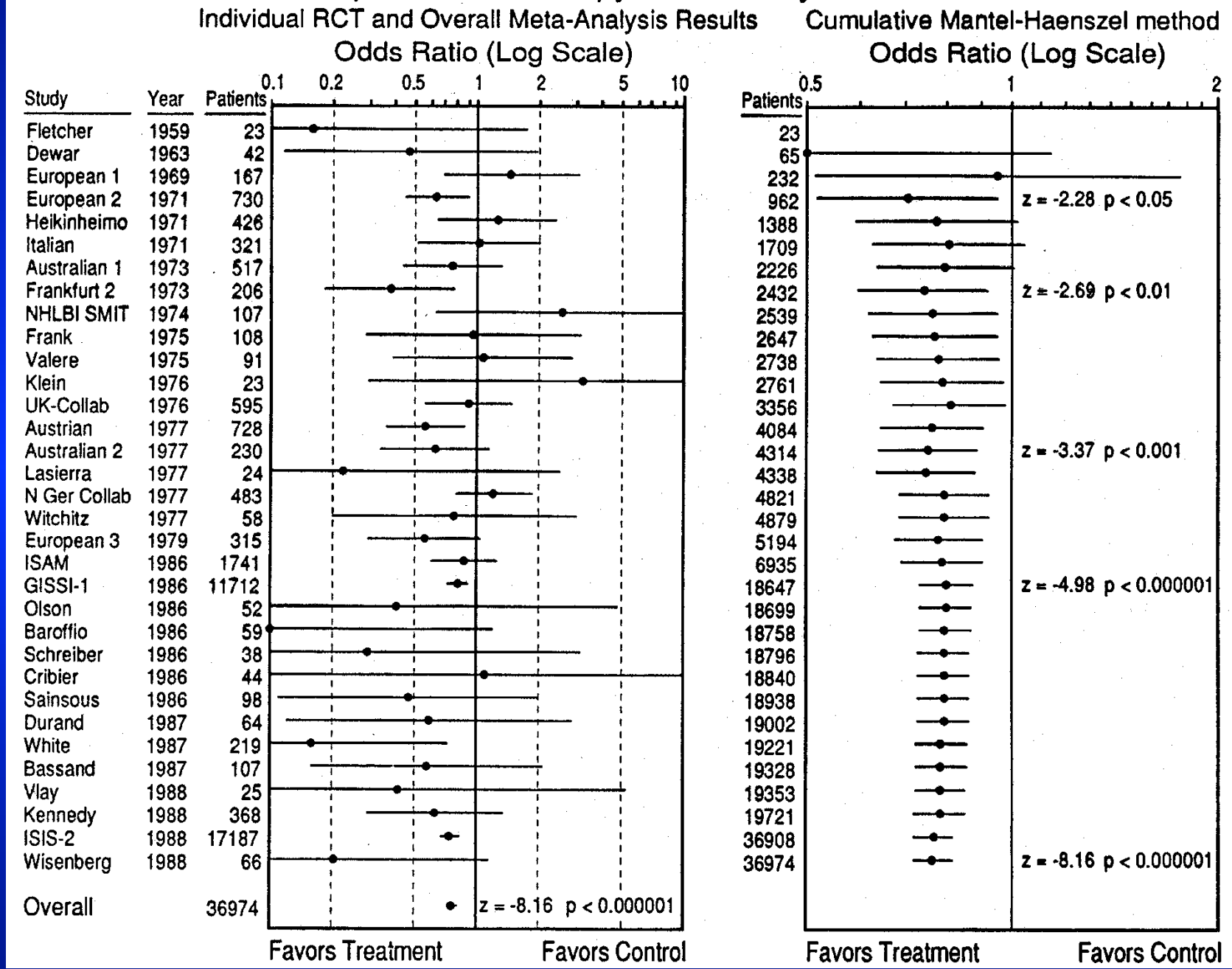


76

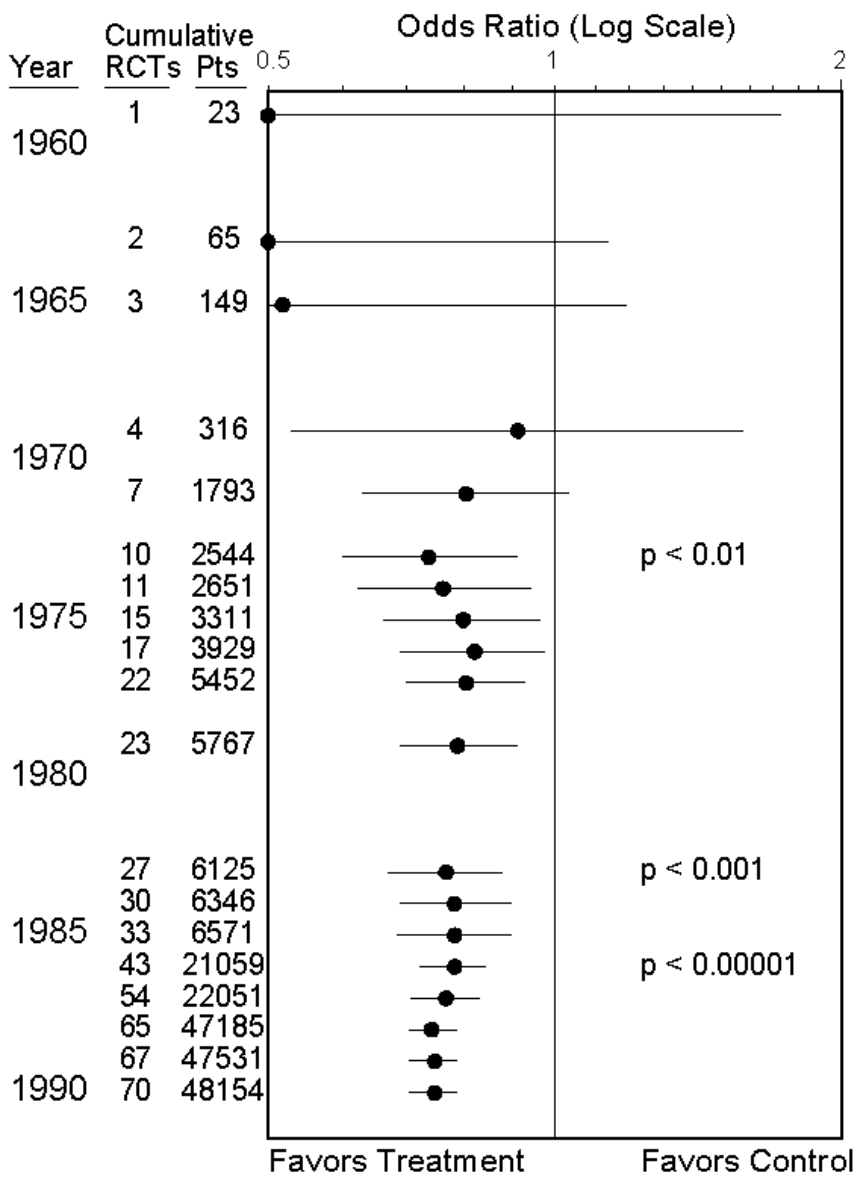
Basic Concept of Cumulative Meta-Analysis



Intravenous Streptokinase Therapy in Acute Myocardial Infarction



Thrombolytic Therapy for AMI



Textbook/Review Recommendations

Routine	Specific	Rare/Never	Experimental	Not Mentioned
				21
				5
			1	10
			1	2
			2	8
				7
				8
	1			12
	1		8	4
	1		7	3
5	2		2	1
15	8			1
6	1			

Antman EM, Lau J, et al. JAMA 1992

Findings of Cumulative Meta-analysis

- Clinical experts' recommendations often are unreliably synchronized with developing RCT evidence.
- Large clinical trials often echo findings from meta-analyses of several smaller studies.
- Trends established by cumulative meta-analyses of previous studies are unlikely to be reversed
- Cumulative meta-analysis is an example of Bayesian updating

Hierarchical Meta-Analysis Model

- Y_i observed treatment effect (e.g. odds ratio) and θ_i unknown true treatment effect from i^{th} study
- First level describes variability of Y_i given θ_i

$$Y_i \sim N\left(\theta_i, \sigma_i^2\right)$$

- Within-study variance often assumed known
- But could use common variance estimate if studies are small
- If data are binary, use binomial distribution here

Hierarchical Meta-Analysis Model

Second level describes variability of study-level parameters θ_i

$$\theta_i \sim N(\theta, \tau^2)$$

in terms of population level parameters: θ and τ^2

Fixed Effects $\theta_i = \theta \ (\tau^2 = 0)$

Random Effects $\theta_i \sim N(\theta, \tau^2)$

$$\Rightarrow Y_i \sim N(\theta_i, \sigma_i^2 + \tau^2)$$

Bayesian Hierarchical Model

Placing priors on *hyperparameters* (θ, τ^2) makes Bayesian model

Posterior distribution of random effects is

$$\theta_i | y_i, \theta, \sigma^2 \sim N(\theta_i^* , V_i(1 - B_i))$$

where

$$\theta_i^* = (1 - B_i)y_i + B_i\theta$$
$$B_i = V_i/(V_i + \tau^2)$$

Each study's conditional mean is weighted average of observed study mean and overall mean

- Inferences sensitive to prior on τ^2

Shrinkage

$B_i = V_i/(V_i + \tau^2)$ are *shrinkage factors*

- Larger B_i shrink θ_i^* more back to the grand mean θ
- Well-estimated studies (small within-study variances) weighted most
- Bigger within-study variances lead to more shrinkage
- Smaller within-study variances lead to less shrinkage
- Increased between-study variance weights studies more evenly

Example: Magnesium for AMI

- Infamous because random effects and fixed effects analysis lead to different conclusions

  Random effects OR = 0.59
  Fixed effects OR = 1.02

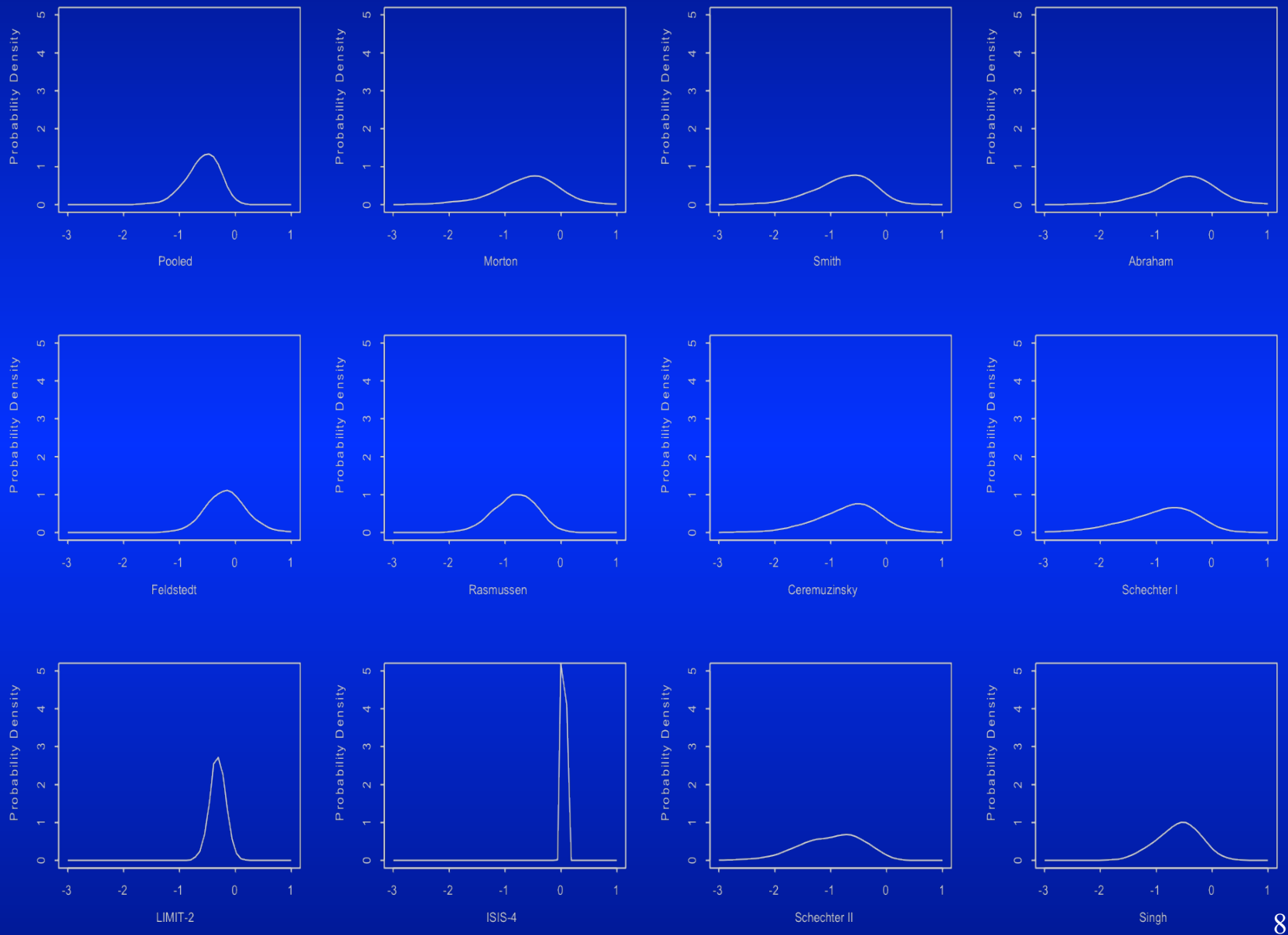
- Very large, influential clinical trial showed no treatment benefit

- Contradicted earlier MA with large trial showing large benefit

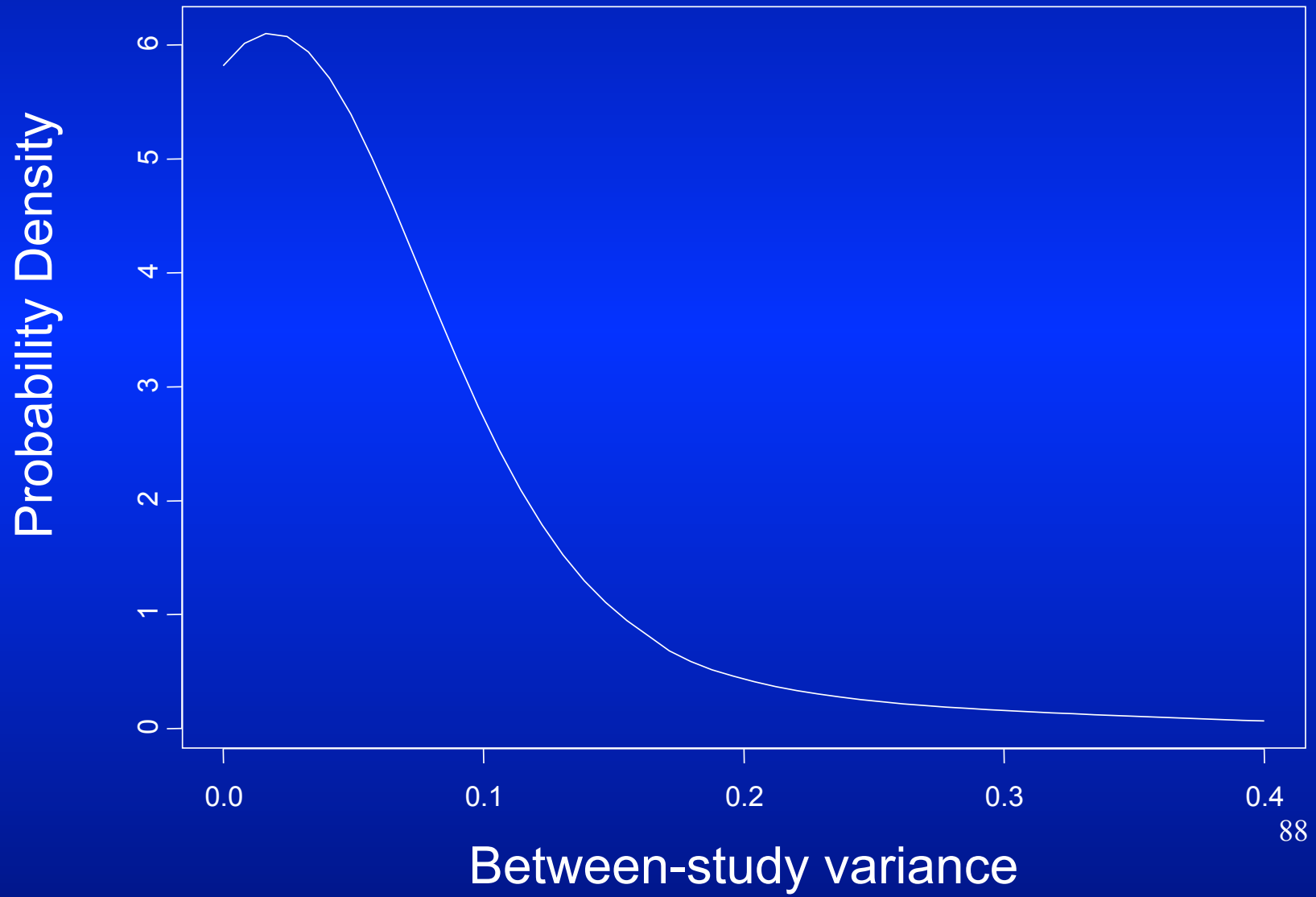
Meta-analysis for Magnesium Studies

	Mortality		Pooled Odds Ratio				
			Observed		Posterior		
Study	**Treated**	**Control**	**Est**	**95% PI**	**Est**	**95% PI**	**Pr(OR<1)**
Morton	1/40	2/36	0.44	0.0, 5.0	0.54	0.2, 1.6	0.89
Smith	2/200	7/200	0.28	0.1, 1.5	0.46	0.1, 1.1	0.96
Abraham	1/48	1/46	0.96	0.1, 15.8	0.61	0.2, 1.9	0.84
Feldstedt	10/150	8/148	1.25	0.5, 3.3	0.86	0.4, 1.9	0.70
Rasmussen	9/135	23/135	0.35	0.2, 0.8	0.43	0.2, 0.9	0.99
Ceremuz.	1/25	3/23	0.28	0.0, 2.9	0.49	0.1, 1.4	0.92
Shechter I	1/59	9/56	0.09	0.0, 0.7	0.38	0.1, 1.4	0.97
LIMIT 2	90/1159	118/1157	0.74	0.6, 1.0	0.73	0.6, 1.0	0.99
ISIS-4	2216/29011	2103/29039	1.06	1.0, 1.1	1.06	1.0, 1.1	0.04
Shechter II	2/89	12/80	0.13	0.0, 0.6	0.36	0.1, 0.9	0.99
Singh	6/76	11/75	0.50	0.2, 1.4	0.54	0.2, 1.1	0.95
Pooled			0.59	0.4, 0.9	0.55	0.3, 0.9	0.99

Kernel Density Plots of Posteriors



Distribution of Between-Study Variance



Meta-Regression

- Investigate sources of heterogeneity in meta-analysis
- Regression analysis to identify correlations between treatment effects (outcomes) and covariates of interest (predictors)
- Estimates *interaction* between covariate and treatment effect, i.e. how treatment effect is *modified* by covariate
- Unit of analysis is the individual study
- Correlation implies treatment interaction

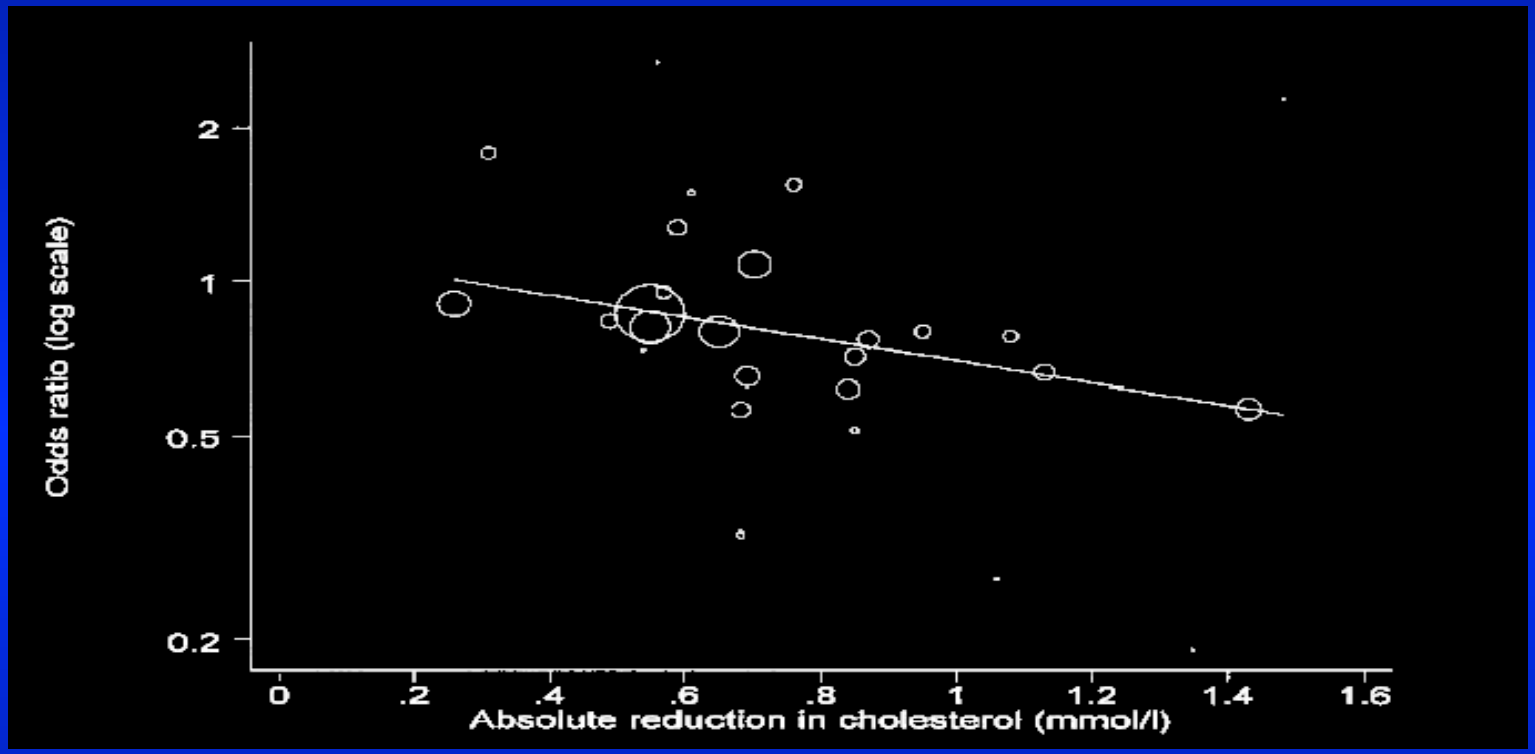
Meta-Regression

- Factors may be study-level or subject-level
- Study-level factors: blinding, randomization, dosage, protocol
- Subject-level factors: age, gender, race, blood pressure
- Study effect is no longer a single value, but is a function of predictors

$$\theta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + u_i$$

- Or can use baseline risk level (control rate)

Meta-Regression with Study-Level Summary of Patient Level Covariates



- Weighted regression (need to make adjustment to program because weights known exactly)

- Data points proportional to study size

Problems with Meta-Regression

- Number of studies usually small
- Number of potential predictors may be large
- Data may be unavailable (not conceived or not reported)
- Covariates pre-selected (biased?)
- Little variation in range of mean predictor
- Subject-level factors can be affected by ecological bias
- Causality uncertain

Ecological Bias

- Group averages don't represent individuals well
  - –E.g., what does percentage male/female mean?
- Averages have little between-study variation
- Averages do not account for within-study variation e.g. 40 year average age can mean different things

Study	Mean Age	%> 60	Odds Ratio
A	40	0	1.0
B	40	10	0.8

- Events concentrated in high-risk subgroup
  - – May want to construct group-level variable to represent this E.g., percentage of elderly, rather than mean age

Baseline Risk Meta-Regression

- Control group event rate reflects multiple risk factors
  - different populations
  - underlying baseline risk of patients
  - length of study follow-up
  - treatment delivery
- Related to severity of illness but not interpretable for individual
- Data always available
- May signal multiple causes
- Standard weighted LS biased
  - ignores correlated measurement error

Meta-Regression vs. Individual Patient Regression

	Meta-Regression	Individual Patient
Cost	Cheap	Expensive
Data Available	Usually	Infrequently
Factors	Study	Patient and Study
Outcomes	Reported	Updated, Complete
Data Cleaning	Impossible	Possible
Bias	Reporting, Ecological	Reporting, Retrieval
Interpretation	Study-specific	Patient-specific

ACE Inhibitors for Non-Diabetic Renal Disease

- Meta-analysis of 11 RCTs of ACE inhibitors (ACEI) published in 1997 showed treatment effective for non-diabetics in preventing progression of disease
- Is ACEI effect completely explained by its effect to lower blood pressure and urine protein?
- Do ACEI work equally well for all nondiabetic renal patients or are there treatment interactions?
- What is the optimal dosing of ACEI and what concomitant medications might improve its efficacy?
- ***With only 10 studies, need patient-level data to answer all these questions***

Meta-Regression with Summary Data

$$\bar{Y}_{.j1} - \bar{Y}_{.j0} \sim N(\beta_j, \omega_j^2)$$

$$\beta_j \sim N(\beta_0 + \beta_1 X_j, \tau_\beta^2)$$

- Fixed Effects if $\tau_\beta^2 = 0$
- Can fit with standard weighted linear regression model
- With individual patient data, can fit by two-step process

Individual Patient Data Regression Model

$$Y_{ij} \sim N(\alpha_j + \beta_j T_{ij} + \gamma Z_{ij} + \delta Z_{ij} * T_{ij}, \sigma_j^2)$$

$$\alpha_j \sim N(\alpha_0 + \alpha_1 X_j, \tau_\alpha^2)$$

$$\beta_j \sim N(\beta_0 + \beta_1 X_j, \tau_\beta^2)$$

- Multilevel model without aggregate effects

$$\tau_\alpha^2 = \tau_\beta^2 = 0$$

- Can also assume common study variance σ^2

Combining IPD and Summary Data

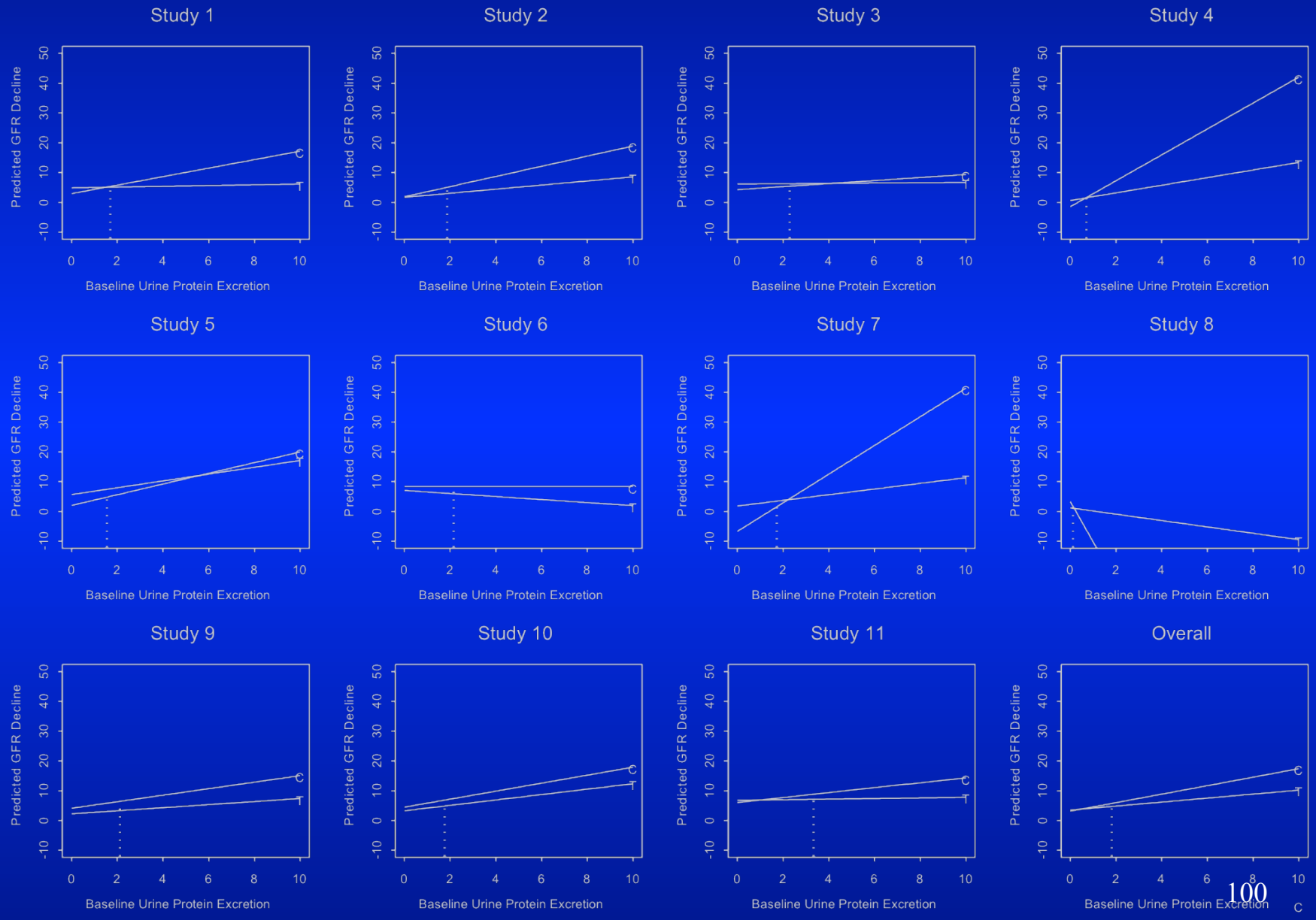
$$Y_{ij}^{*} \sim N(D_j \alpha_j + \beta_j T_{ij}, V_j^{*})$$

$$\alpha_j \sim N(\alpha_0 + \alpha_1 X_j, \tau_\alpha^2)$$

$$\beta_j \sim N(\beta_0 + \beta_1 X_j, \tau_\beta^2)$$

	IPD	Summary Data
Y_{ij}^{*}	Y_{ij}	$\hat{\beta}$
V_j^{*}	σ_j^2	$V(\hat{\beta})$
D_j	1	0

Within-Study Interaction



Issues With Meta-Analysis of Observational Studies

- Need to adjust for potential confounders
- Different studies may adjust for different confounders or may use different adjustment techniques
  - Some variables uncollected in original studies
- May want data on individual participants
- Misclassification and measurement of exposure
- Selection of subjects for control group may differ
- Lack of knowledge of study design characteristics

Studies of Maternal Obesity & Stillbirth

Cohort

							BMI Categories		
Study	Country	Time	Type	Source	Size	Normal	Overweight	Obese	Severely Obese
Cedergren (2004)	Sweden	1992-2001	Prospective	National birth registry	621221	19.8-26	NA	29.1-35	>35
Djorlo (2002)	France	1999	Retrospective	Clinic records	323	20-24.9	25-29.9	>=30	NA
Kristensen (2005)	Denmark	1989-1996	Prospective	Registry, Records	24505	18.5-24.9	25-29.9	>=30	NA
Kumari (2001)	UAE	1996-1998	Retrospective	Clinic records	488	22-28	NA	NA	>=40
Nohr (2005)	Denmark	1998-2001	Retrospective	National birth registry	54505	18.5-24.9	25-29.9	>=30	NA
Sebire (2001)	UK	1989-1997	Retrospective	Clinic records	325395	20-24.9	25-29.9	>=30	NA

Case-control

						BMI Categories			
Study	Country	Time	Type	Source	Cases/Controls	Normal	Overweight	Obese	Severely Obese
Froen (2001)	Norway	1986-1995	Retrospective	National Birth Registry	291/582	20-24.9	25-29.9	³30	NA
Little (1993)	USA	1980	Retrospective	Birth/death certificates	1590/1565	18.1-22	22.1-30	>30	NA
Stephansson (2001)	Sweden	1987-1996	Retrospective	National Birth Registry	649/690	20-24.9	25-29.9	³30	NA

Network of 12 Antidepressants

paroxetine —— reboxetine

duloxetine —— mirtazapine

escitalopram —— fluvoxamine

milnacipran —— citalopram

sertraline —— venlafaxine

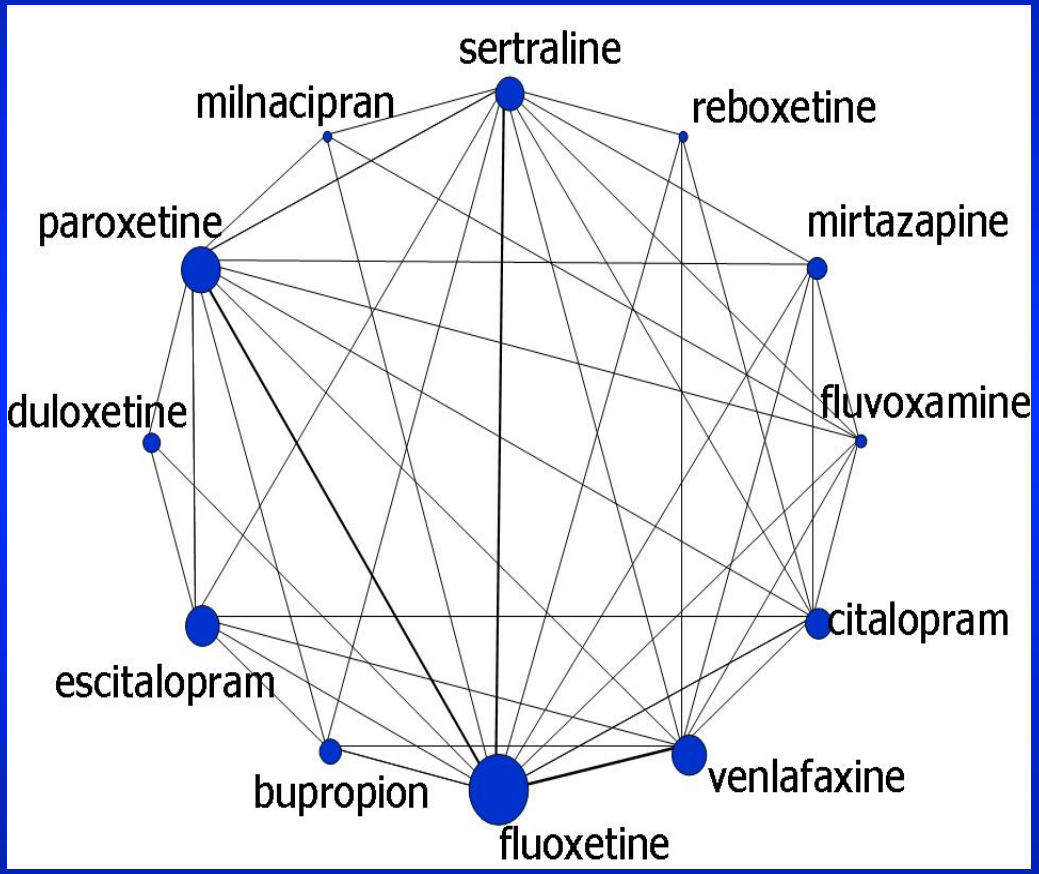
bupropion —— fluoxetine

milnacipran —— paroxetine

sertraline ? duloxetine

bupropion —— escitalopram

fluvoxamine —— milnacipran



19 **meta-analyses** published in the last two years

Indirect Comparisons of Multiple Treatments

Trial			
1	A	B	
2	A	B	
3		B	C
4		B	C
5	A		C
6	A		C
7	A	B	C

- Want to compare A vs. B
  Direct evidence from trials 1, 2 and 7
  Indirect evidence from trials 3, 4, 5, 6 and 7
- Combining all “A” arms and comparing with all “B” arms destroys randomization
- Use indirect evidence of A vs. C and B vs. C comparisons as additional evidence to preserve randomization and within-study comparison

Indirect Comparisons

How do we make the indirect comparisons:

Calculate effect of A vs. C and B vs. C separately

$$T_{AB} = T_{AC} - T_{BC}$$

with SE = square root of sum of variances

<u>Strong Assumptions:</u>

- All trials comparing pairs of tx arms estimate same effect
- Different sets of trials being used are similar

Measuring Inconsistency

Suppose we have AB, AC, BC direct evidence

Indirect estimate $\hat{d}_{BC}^{indirect} = \hat{d}_{AC}^{direct} - \hat{d}_{AB}^{direct}$

Measure of inconsistency: $\hat{\omega}_{BC} = \hat{d}_{BC}^{indirect} - \hat{d}_{BC}^{direct}$

Approximate test (normal distribution):

$$z_{BC} = \frac{\hat{\omega}_{BC}}{\sqrt{V\left(\hat{\omega}_{BC}\right)}}$$

with variance

$$V\left(\hat{\omega}_{BC}\right) = V\left(d_{BC}^{direct}\right) + V\left(d_{AC}^{direct}\right) + V\left(d_{AB}^{direct}\right)$$

Example

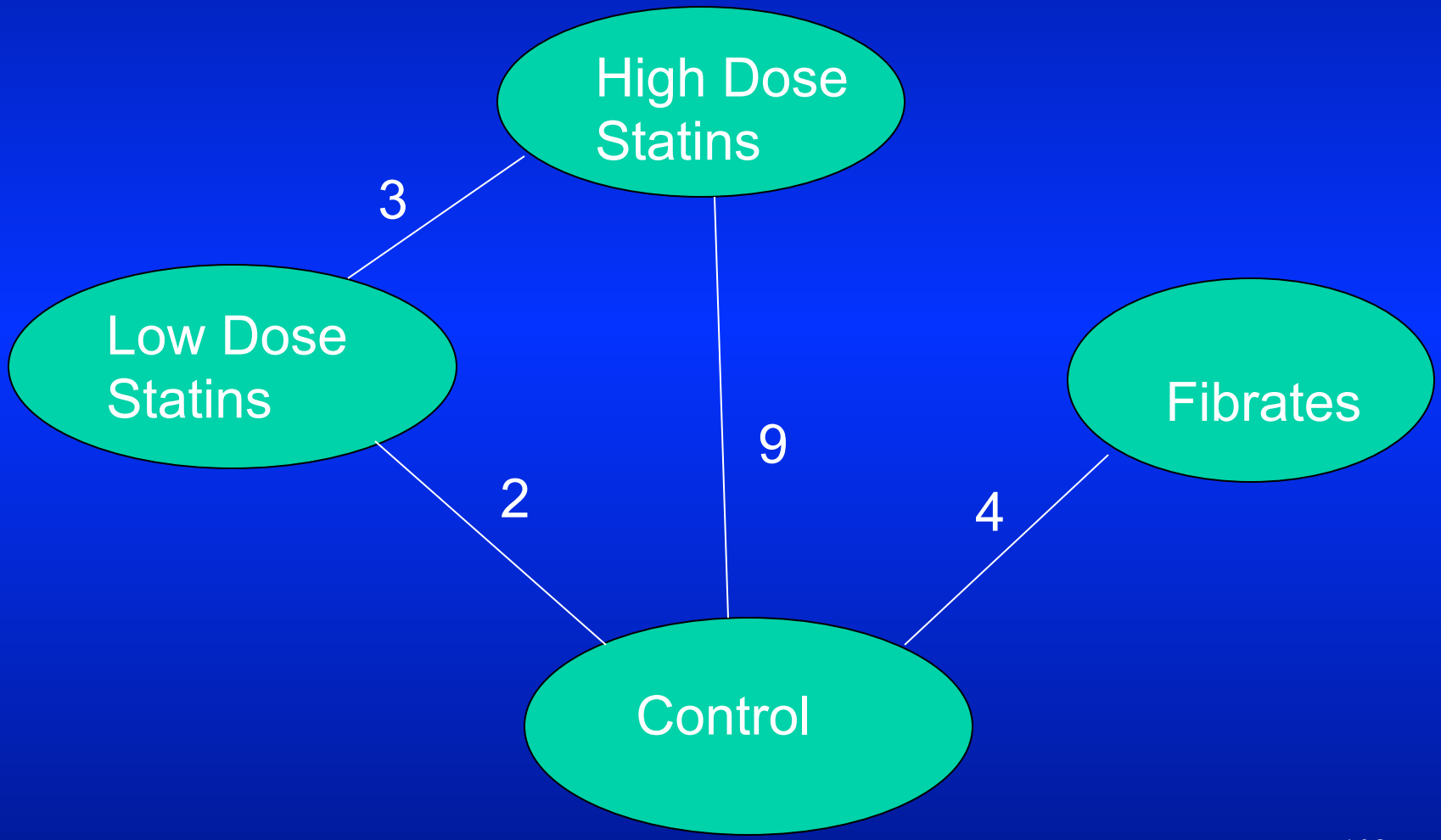
- Population: Patients with cardiovascular disease

- Treatments: Statin treatment (different doses), fibrate
- Comparator: Conventional care or placebo
- Covariates: Baseline cholesterol, triglycerides

- Outcomes:
  - Myocardial infarction (fatal or non-fatal)
  - Stroke (fatal or non-fatal)
  - Death from all other causes

- Design: RCTs

Network



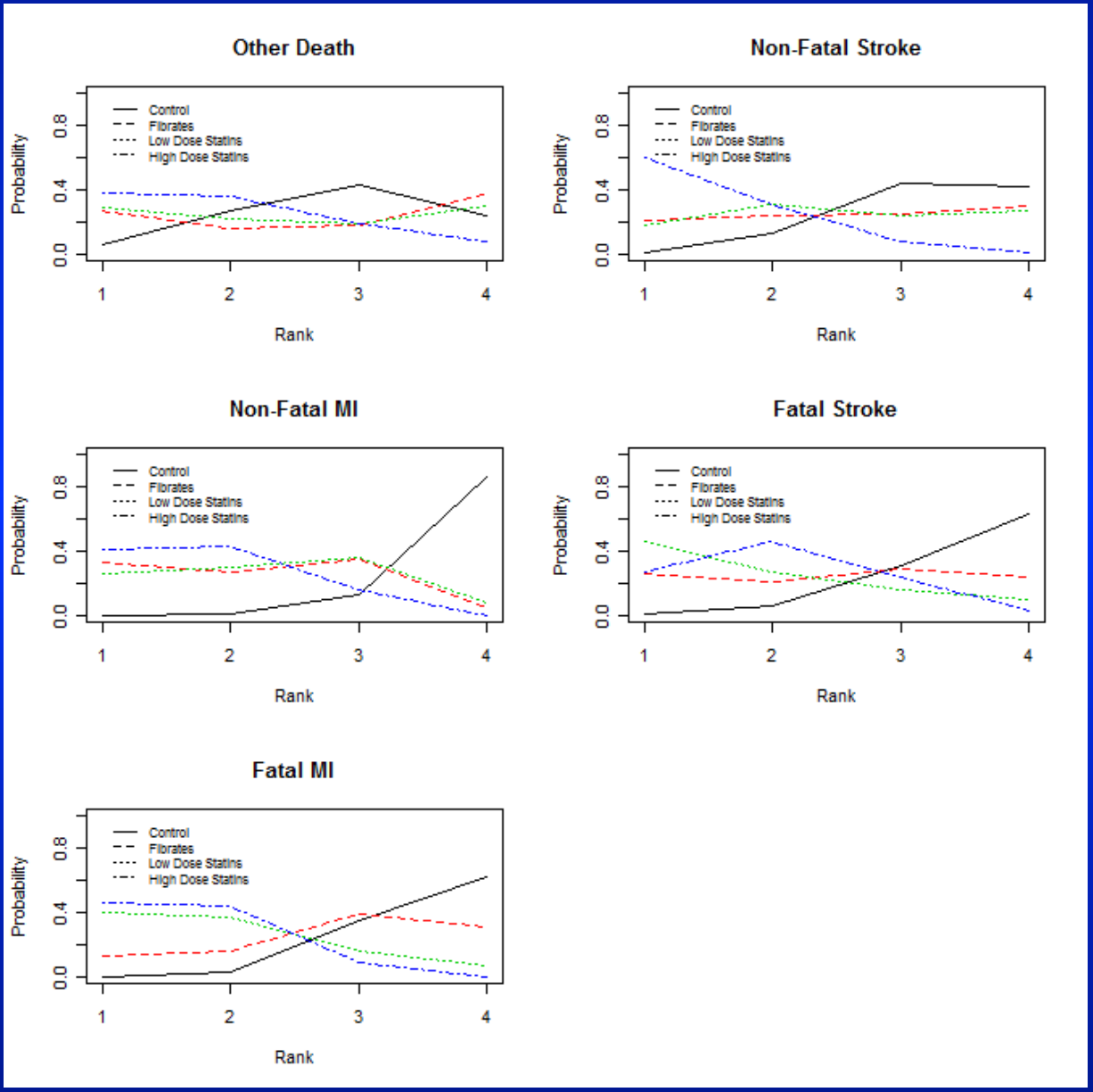
Data Setup

- Each study has 6 possible outcomes and 4 possible tx's
- Not all tx's carried out in each study
- Not all outcomes observed in each study
- Incomplete data with partial information from summary categories
- Can use available information to impute missing values
- Can build into Bayesian algorithm using multinomial model

Odds Ratios

	Other death	Non-fatal Stroke	Non-fatal MI	Fatal Stroke	Fatal MI
Fibrate v Control	1.03 (0.63 – 1.80)	0.90 (0.57 – 1.33)	0.69 (0.44 – 0.98)	0.80 (0.40 – 1.49)	0.94 (0.50 – 1.53)
LDS v Control	0.93 (0.60 – 1.41)	0.87 (0.56 – 1.42)	0.76 (0.46 – 1.08)	0.72 (0.32 – 1.69)	0.64 (0.39 – 1.04)
HDS v Control	0.84 (0.69 – 1.15)	0.72 (0.50 – 0.94)	0.66 (0.53 – 0.81)	0.74 (0.41 – 1.13)	0.64 (0.45 – 0.83)
LDS v Fibrate	0.88 (0.49 – 1.57)	0.95 (0.69 – 1.15)	1.11 (0.62 – 1.88)	0.88 (0.38 – 2.59)	0.69 (0.37 – 1.40)
HDS v Fibrate	0.81 (0.54 – 1.51)	0.80 (0.45 – 1.26)	0.97 (0.66 – 1.47)	0.92 (0.40 – 2.41)	0.67 (0.37 – 1.26)
HDS v LDS	0.94 (0.62 – 1.36)	0.80 (0.50 – 1.25)	0.87 (0.64 – 1.35)	1.01 (0.53 – 1.89)	0.93 (0.690– 1.74)

Rank Plot


Other Death
Non-Fatal Stroke
Non-Fatal MI
Fatal Stroke
Fatal MI
Control
Fibrates
Low Dose Statins
High Dose Statins
Probability
0.0
0.4
0.8
1
2
3
4
Rank

Multivariate Model

Assume two outcomes Y_{i1}, Y_{i2} observed in I studies

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix}_i \sim N\left(\begin{bmatrix} \theta_{i1} \\ \theta_{i2} \end{bmatrix}, \begin{bmatrix} s_{i1}^2 & \rho_{W_i} s_{i1} s_{i2} \\ \rho_{W_i} s_{i1} s_{i2} & s_{i2}^2 \end{bmatrix} \right)$$

$$\begin{pmatrix} \theta_{i1} \\ \theta_{i2} \end{pmatrix}_i \sim N\left(\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}, \begin{bmatrix} \tau_1^2 & \rho_B \tau_1 \tau_2 \\ \rho_B \tau_1 \tau_2 & \tau_2^2 \end{bmatrix} \right)$$

- May be difficult to estimate within-study correlations
- Instead, could reformulate problem to estimate only single correlation for marginal model adding within and between-study
- Or could estimate each outcome separately
- Ignoring within-study correlation gives biased estimates

Longitudinal Model

Each study has K measurements taken over time

θ_i is vector of K treatment effects at each time for i^{th} study

θ is vector of average treatment effects at each time

$$Y_i \sim MVN\left(\theta_i, \Sigma_i\right)$$

$$\theta_i \sim MVN(X_i\theta, Z_i D Z_i)$$

- Often reporting times differ across studies
- Can aggregate

Longitudinal Model:Variance Structure

- Σ_i usually assumed known
- May not have information reported on correlations
- Could assume Σ_i diagonal or take $\Sigma_i = W_i^{-1/2} C W_i^{-1/2}$
- W_i is diagonal matrix holding known within-study variances
- C is correlation matrix constant across studies and estimated from data
- Could use autoregressive structure or allow different random effects at each time
  - E.g. D is AR(1) with unequal variances

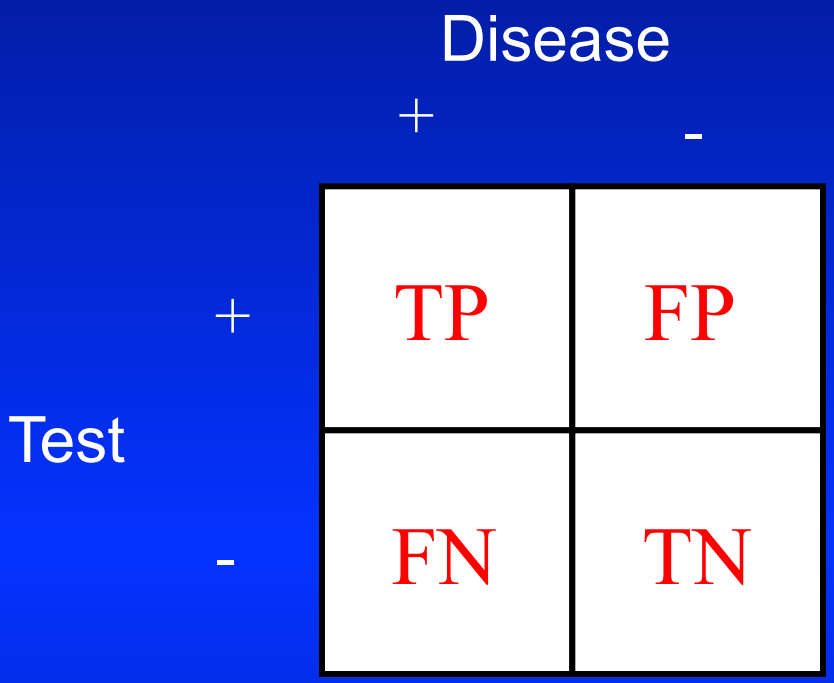
Uses of Diagnostic Tests

- Screen (mammography for breast cancer)
- Diagnose (ECG for acute myocardial infarction)
- Grade (stage of cancer)
- Monitor progression (recurrence)
- Monitor therapy (blood drug level) and therapeutic response (regression of tumor size)
- Guide treatments (arteriography for CABG)

False positive results may lead to unnecessary tests and treatments and possible harms

False negative results may prevent proper treatment

Defining Test Performance



Prevalence =(TP + FN) / (TP + FP + FN + TN)

Accuracy= (TP+TN)/(TP+FP+FN+TN)

Sensitivity (TPR)= TP/(TP+FN)

Specificity (TNR)= TN(TN + FP)

Predictive Value + = TP/(TP + FP)

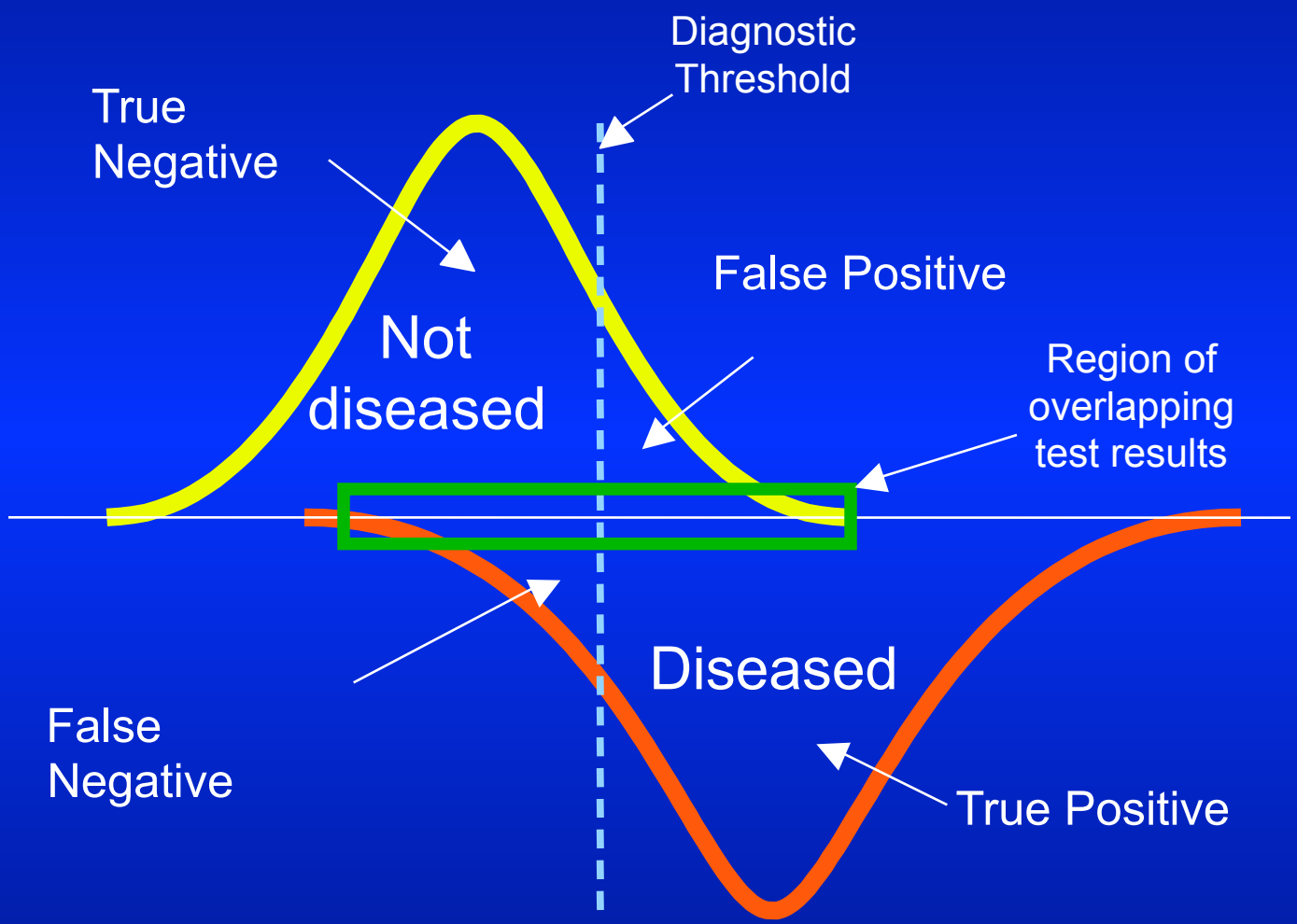
Predictive Value - = TN(TN + FN)

Odds Ratio = (TP x TN)/(FP x FN)
= {Se / (1 - Se)} / {(1-Sp) / Sp}
= LR + / LR -

Likelihood Ratio + = {TP/(TP + FN)}/{FP/(FP + TN)}/

Likelihood Ratio - = {FN/(TP + FN)}/{TN(FP + TN)}/

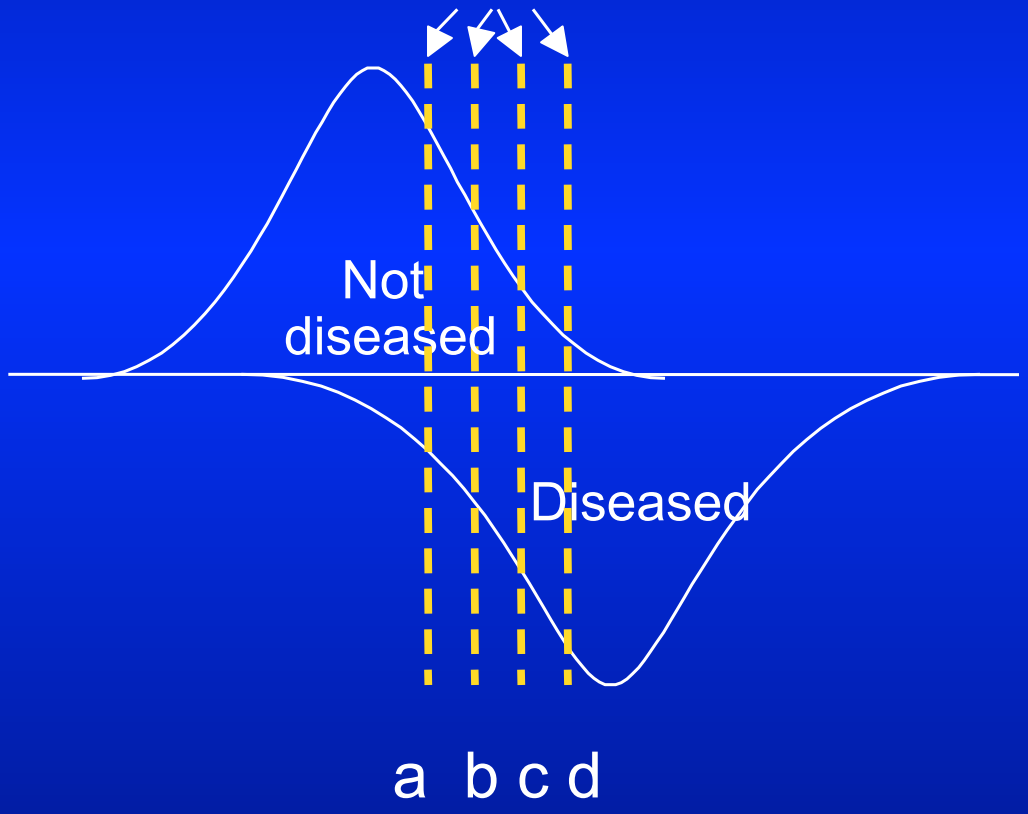
Test Performance

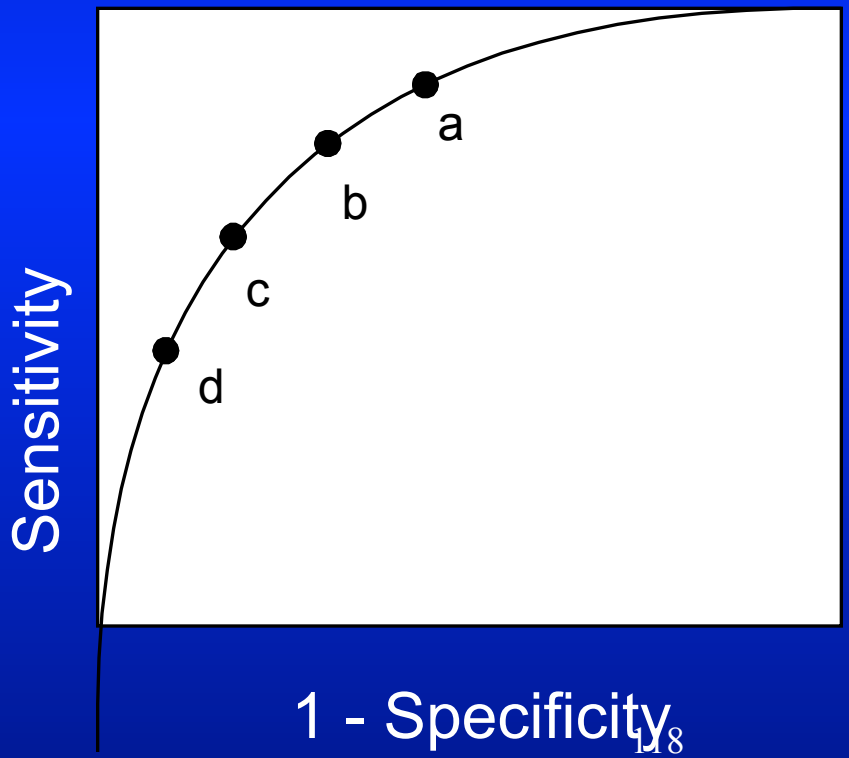


Changing diagnostic threshold or disease spectrum changes test performance

Making ROC Curve from Multiple Test Thresholds

Multiple thresholds evaluated in test





Full Cycle of Diagnostic Test Evaluation: Magnetic Resonance Spectroscopy for Brain

Level	Example of study purpose	# studies	# patients
1: Technical feasibility	Ability to produce consistent spectra	85	2434
2: Test accuracy	Sensitivity and specificity	8	461
3:Diagnostic impact	Percentage of times clinicians' subjective assessment of diagnostic probabilities changed after the test	2	32
4:Therapeutic impact	Percentage of times therapy planned before MRS changed after the test	2	105
5:Clinical outcomes	Percentage of patients who improved with MRS diagnosis compared with those without MRS	0	0
6:Societal Impact	CEA: use of test in asymptomatic patients	0	0

Diagnostic Technology Controversy: Screening Mammography RCTs

- 1999 study found no decrease in breast cancer mortality in Sweden, where screening has been recommended since 1985
- Reviewed methodological quality of mammography trials and repeated a meta-analysis

Relative Risk of Death from Breast Cancer

	Number randomized		# of deaths from breast CA		Relative risk
	Screening	Control	Screening	Control	(95% CI)
Randomization adequate					
Malmo	21088	21195	63	66	0.96 (0.68-1.35)
Canada	44925	44910	120	111	1.08 (0.84–1.40)
Total	66013	66105	183	177	**1.04 (0.84-1.27)**
Randomization not adequate					
Goteberg	11724	14217	18	40	0.55 (0.31-0.95)
Stockholm	40318	19943	66	45	0.73 (0.50-1.06)
Kopparberg	38589	18582	126	104	0.58 (0.45-0.76)
Ostergotland	38491	37403	135	173	0.76 (0.61-0.95)
New York	30131	30565	153	196	0.79 (0.64-0.98)
Edinburgh	22926	21342	156	167	0.87 (0.70-1.08)
Total	182179	142052	654	725	**0.75 (0.67-0.83)**

Policy Results

- Switzerland decided to not cover screening mammography
- National Cancer Institute wavered on value of screening mammograms
- Women and doctors more confused about value of test

Conclusions

- Evidence-based medicine requires collaboration of doctors, statisticians, librarians, epidemiologists and other experts
- Goal is to provide scientific basis for clinical decisions
- Often requires sifting through extensive literature
- Systematic reviews more scientific than narrative reviews
- Determines validity of evidence and identifies research gaps
- Discovery of heterogeneity can improve interventions