

# Moving from Observational Studies to Clinical Trials: Why do We Sometimes Get It Wrong?

Joseph Lau, MD  
Rapporteur

# Evaluating Study Outcomes: Biomarkers, Intermediate Endpoints, and Surrogate Endpoints

- Ross Prentice, PhD
  - Surrogate Endpoint Definition and Application
- Stuart Baker, ScD
  - Recent Approaches to Surrogate Endpoint Validation
- David Ransohoff, MD
  - New Complexity: The “Omics” Revolution
- Daniel Hayes, MD
  - Methods of Biomarker Validation

# Ross Prentice, PhD: Surrogate Endpoint Definition and Application

- Surrogate outcome definition
- Conceptual framework for associations of treatment, surrogate, and true endpoint
- Proposed meta-analysis approach of borrowing information in prior studies of similar treatments in similar populations

# Stuart Baker, ScD: Recent Approaches to Surrogate Endpoint Validation

- Process of validating markers or endpoints
  - Hypothesis testing framework
  - Estimation framework
- Recommended meta-analysis estimation approach to validate surrogate endpoint
- Real examples?
- Has this method been validated empirically?
- Other approaches? Bayesian method?

# David Ransohoff, MD: New Complexity: The “Omics” Revolution

- Promises and disappointments of cancer markers
- Rules of evidence not well developed
- Current overly optimistic interpretation of “omics” data
- Bias as threat to validity

# Daniel Hayes, MD: Methods of Oncology Biomarker Validation

- Many proposed tumor markers
- Most inadequately validated

# A few comments on other sessions

- Current concepts
- How best to evaluate existing evidence?
- How to design better future studies?

# Issues evaluating evidence: an EBM-er's perspective

- Evidence is seldom single sourced (basic science, animal, human observations, human experiments)
- Observational studies vs RCTs
- Surrogates vs clinical outcomes
- Mega-trials vs (meta-analyses) small trials
- Large RCTs vs large RCTs
- Methodological quality of the studies
- Publication bias

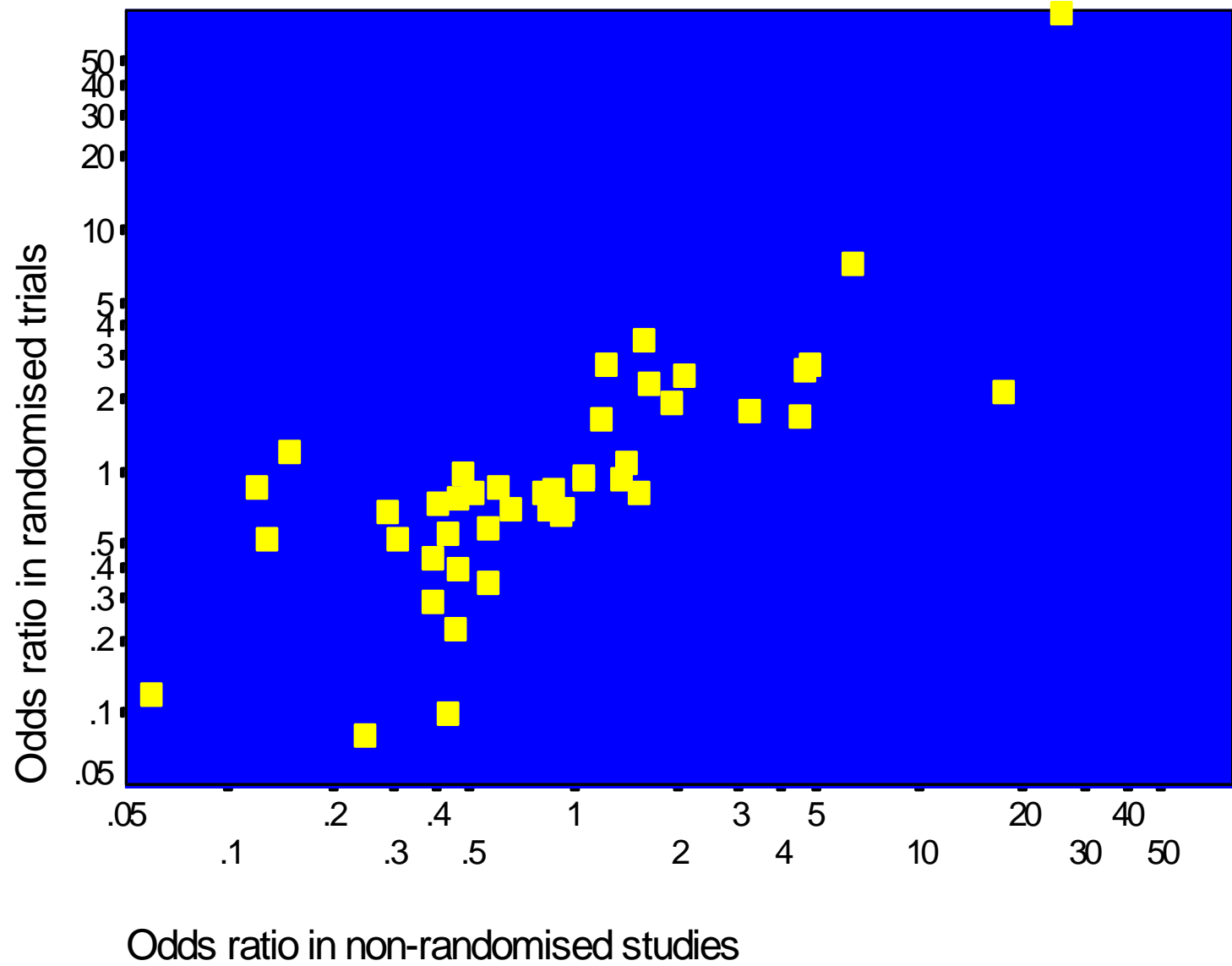
# Comparisons of RCTs with NROS

- BMJ 1998; Oxman et al.
- NEJM 2000; Concato et al.
- NEJM 2000; Benson et al.
- JAMA 2001; Ioannidis et al.

# Comparison of RCTs and NROS in meta-analyses

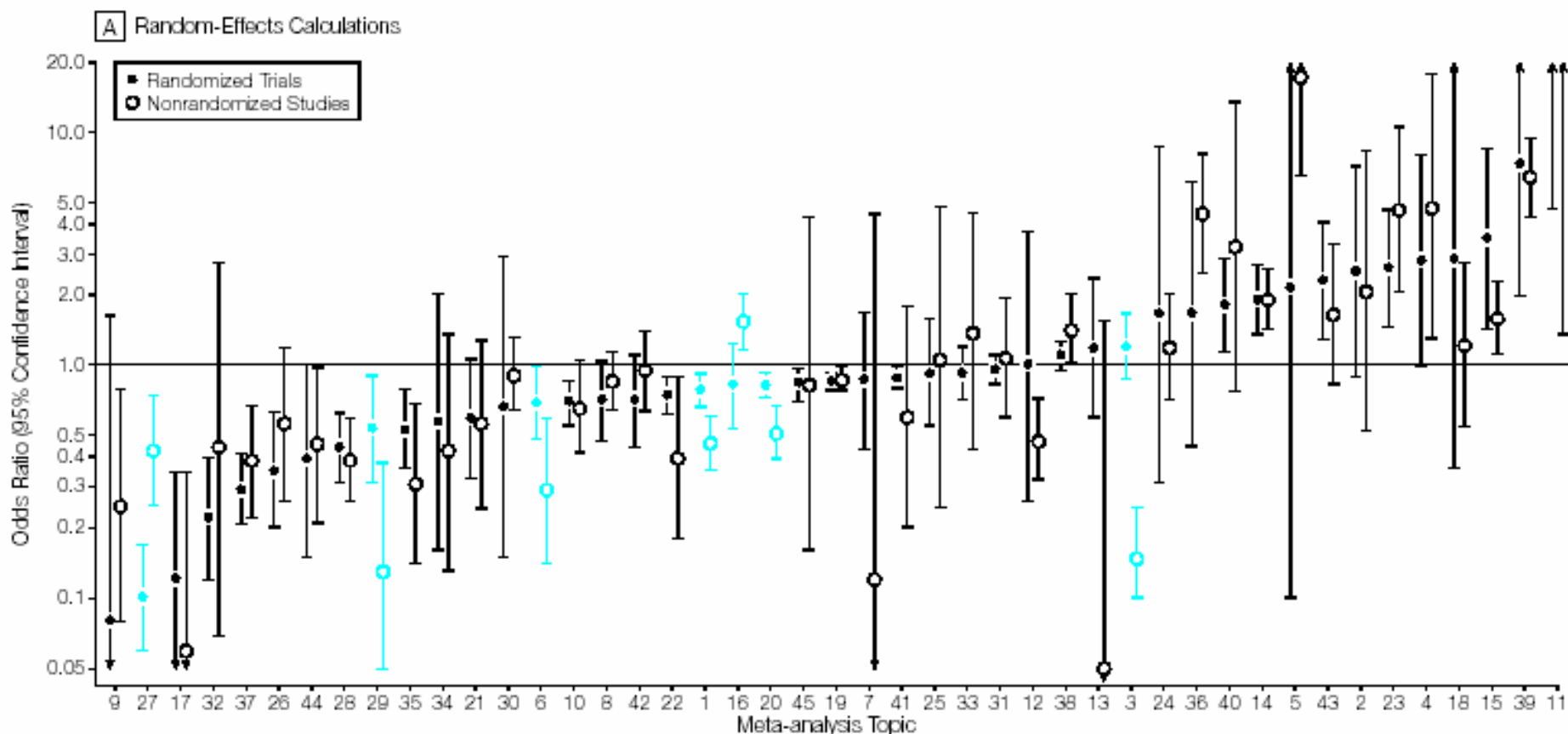
Ioannidis et al. JAMA 2001;286:821-830

- A total of 45 topics were considered.
- They were identified from comprehensive searches of MEDLINE, The Cochrane Library, previous relevant publications and personal archives – c. 3,000 meta-analyses were screened.
- The 45 topics included 408 primary studies with available binary data (240 RCTs and 168 NROS)
- NROS included 71 prospective studies, 40 retrospective cohort studies, 25 case-control studies, 29 studies with historical controls, and 3 studies with unclear designs



# Comparisons between randomized and non-randomized evidence.

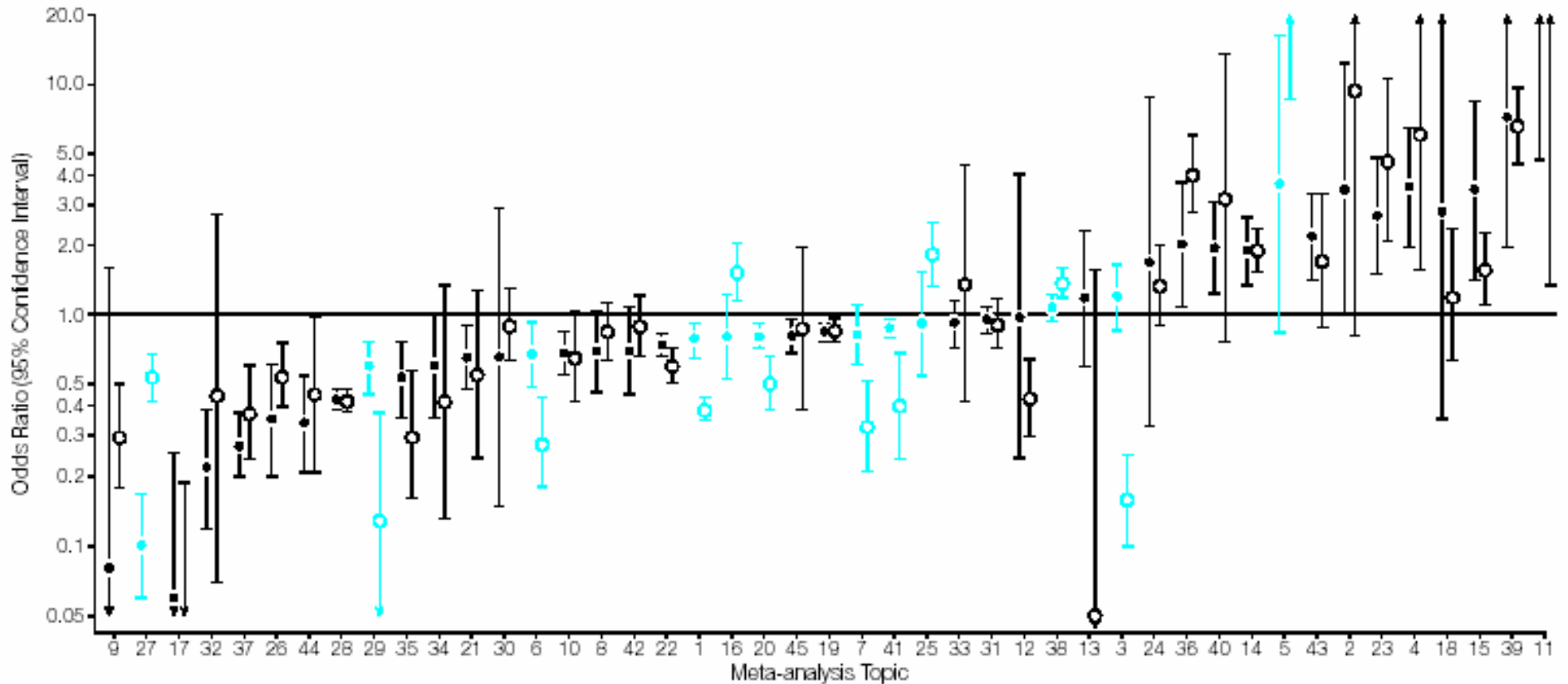
Ioannidis J. et al. JAMA 2001;286:821-830.



# Comparisons between randomized and non-randomized evidence.

Ioannidis J. et al. JAMA 2001;286:821-830.

B Fixed-Effects Calculations



# Heterogeneity in RCTs and in NROS

Ioannidis et al. JAMA 2001;286:821-830.

- Statistically significant heterogeneity between randomized trials was seen in 9 of 39 topics with at least 2 RCTs included
- Statistically significant heterogeneity between the non-randomized studies was seen in 13 of 32 topics with at least 2 NROS included
- The estimated between-study heterogeneity tended to be smaller among RCTs than among NROS ( $p=0.032$ )

# Comparison of the magnitude of treatment effects

Ioannidis J. et al. JAMA 2001;286:821-830.

- In 25 of 45 cases, the non-randomized studies showed a larger treatment effect for the experimental treatment than the randomized trials. The opposite occurred in 14 cases, but it was a data artifact in 3 of them. In 6 topics there was either no clear-cut experimental arm or the effects were similar ( $p=0.009$ ).

# Discrepancies between RCTs and NROS

Ioannidis J. et al. JAMA 2001;286:821-830.

- Discrepancies beyond chance were observed in 12 of 45 cases by fixed effects and in 7 of 45 cases by random effects
- In these discrepancies, almost always the treatment effect was more favorable in NROS
- When limiting analyses to prospective studies, there were disagreements in 2 of 26 topics (8%)

# Conclusions

Ioannidis J. et al. JAMA 2001;286:821-830.

- Treatment effects in RCTs and observational studies on the same topic tend to be highly correlated
- Nevertheless, discrepancies do occur in about 1 out of 6 cases, even when between-study heterogeneity is accounted for
- Typically, discrepant pairs tend to show more favorable results in observational studies
- Discrepancies in the absolute magnitude of effect (=“how much it works”) are very common

# Conclusions (cont)

Ioannidis J. et al. JAMA 2001;286:821-830.

- Observational studies exhibit larger variability in their treatment effects than RCTs
- Discrepancies are more common when retrospective observational designs are considered
- Both RCTs and NROS must be carefully scrutinized for sources of genuine heterogeneity and bias
- RCTs and NROS should not be seen as mutually exclusive domains of research

# Comparisons of Large RCTs with Meta-analyses of small trials

- Villar et al. Lancet 1995
- Cappelleri et al. JAMA 1996
- LeLorier et al. NEJM 1997

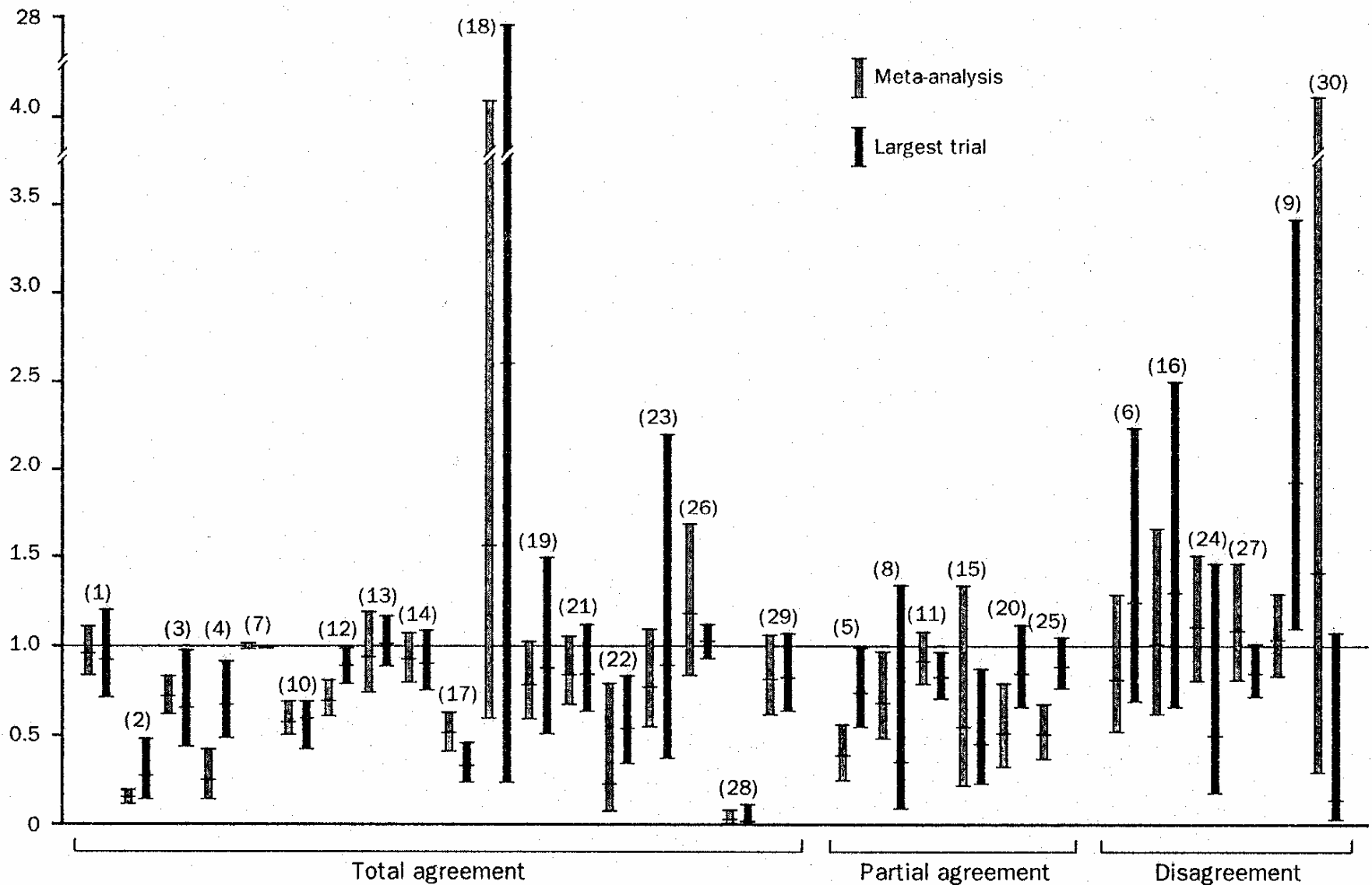
# Some Issues in the Comparisons of Meta-Analysis and Large Trial

Ioannidis et al. JAMA 1998

- Definition of large (arbitrary, power)
- Source of meta-analyses (why done?)
- Source of large trials
- Types of outcomes ( 1<sup>o</sup>, 2<sup>o</sup> )
- Meta-analysis statistics (FEM, REM)
- Definition of agreement (p-value, corr.)
- Reasons for disagreement

# Comparison of 30 meta-analyses of RCTs with largest corresponding trial

Villar J, Carroll G, Belizan JM. Lancet 1995; 345:772-76.



# Meta-analyses vs. Mega-trials

Cappelleri JC, Ioannidis JPA, deFerranti SD, Schmid CH, Aubert M, Chalmers TC, Lau J. Large trials versus meta-analyses of smaller trials: How do their results compare? JAMA 1996; 276:1332-38.

# Large trials versus meta-analysis of smaller trials

## Data source

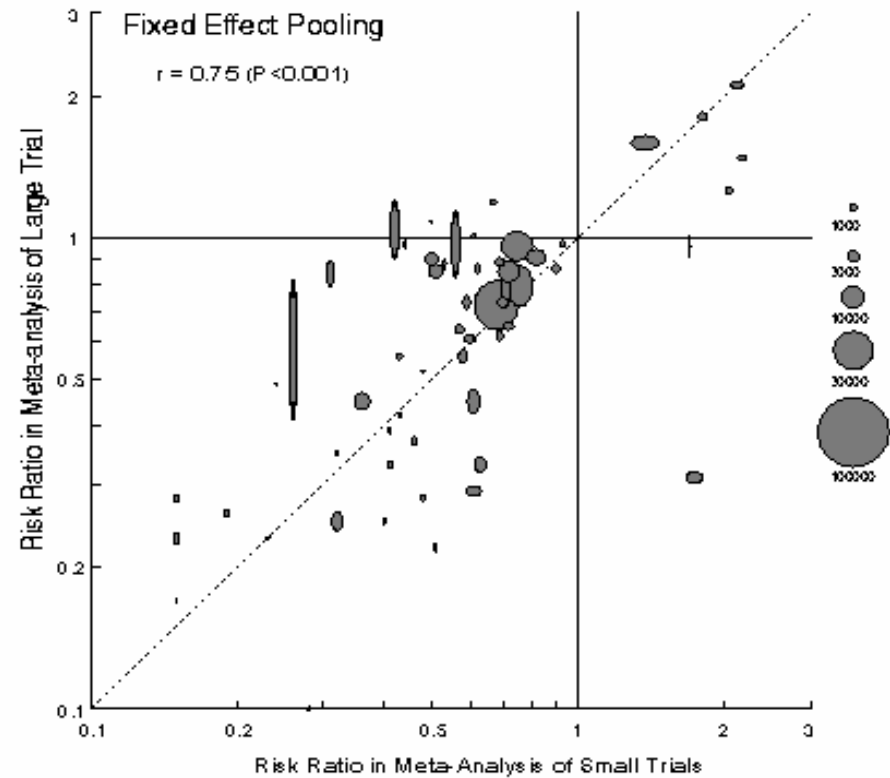
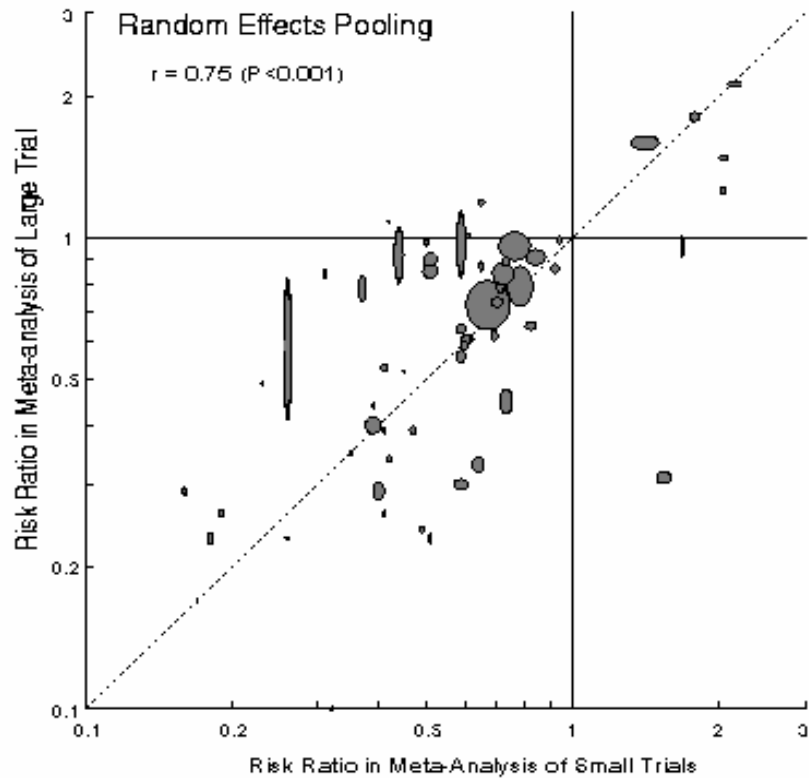
	Large sample size ( $\geq 1000$ pts)	Sufficient Power ( $\geq 80\%$ )
Cochrane Perinatal Database (1994)	33	43
Medline (1966 – 1995)	46	18
<b>TOTAL</b>	<b>79</b>	<b>61</b>

# Large trials vs meta-analysis of smaller trials: How do their results compare ?

- By random effect calculations, agreements found between large and smaller trials in:  
90% selected by sample size approach (1,000); 82% by statistical power approach
- Twice as many disagreements appeared when the variability among large studies and the variability among smaller studies was not considered (fixed effects calculations).

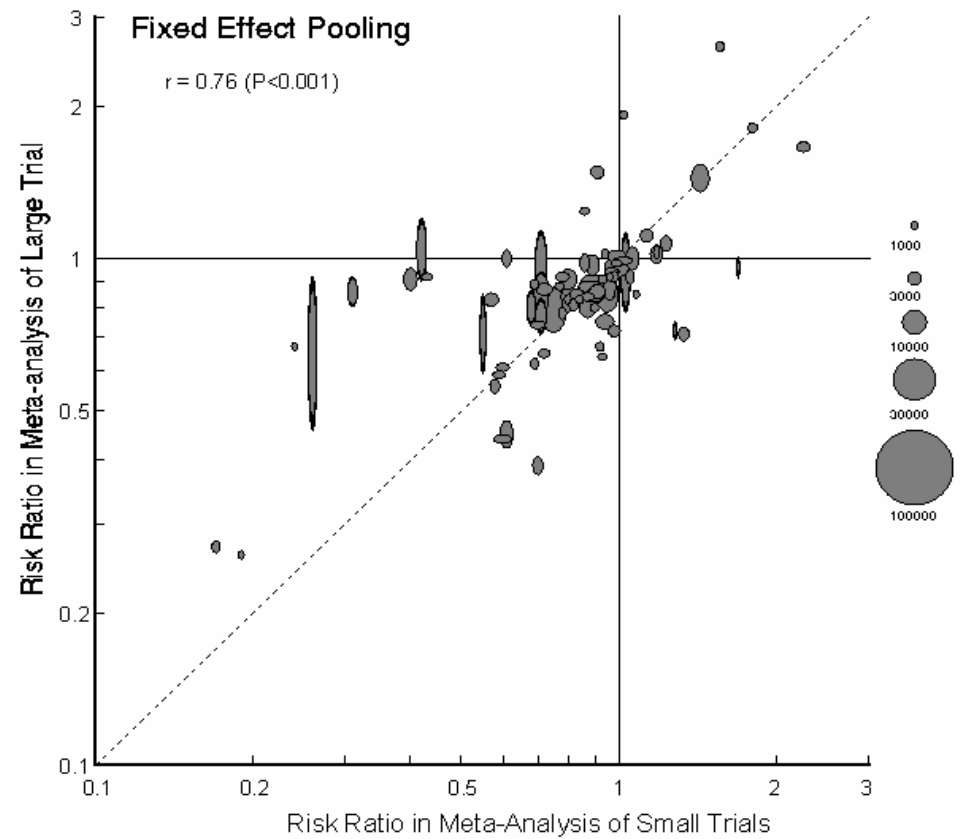
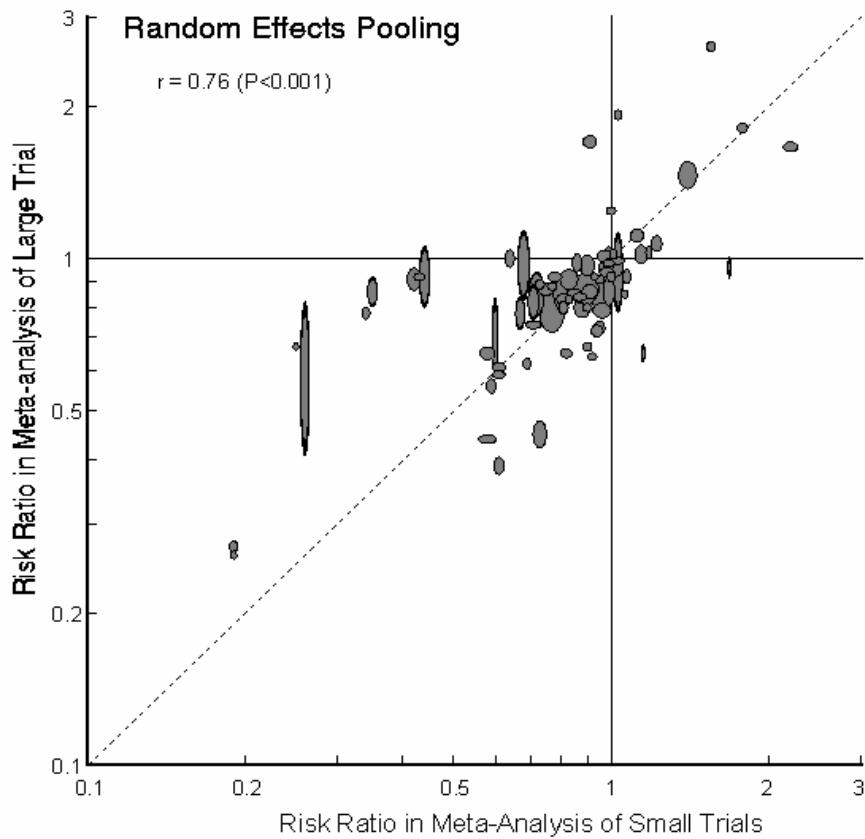
## Comparisons between Large Trials and Meta-Analyses of Small Trials

### Cappelleri Protocol - Statistical Power Rule (61 comparisons)



# Comparisons between Large Trials and Meta-Analyses of Small Trials

## Cappelleri Protocol - 1000 Size Rule (79 comparisons)



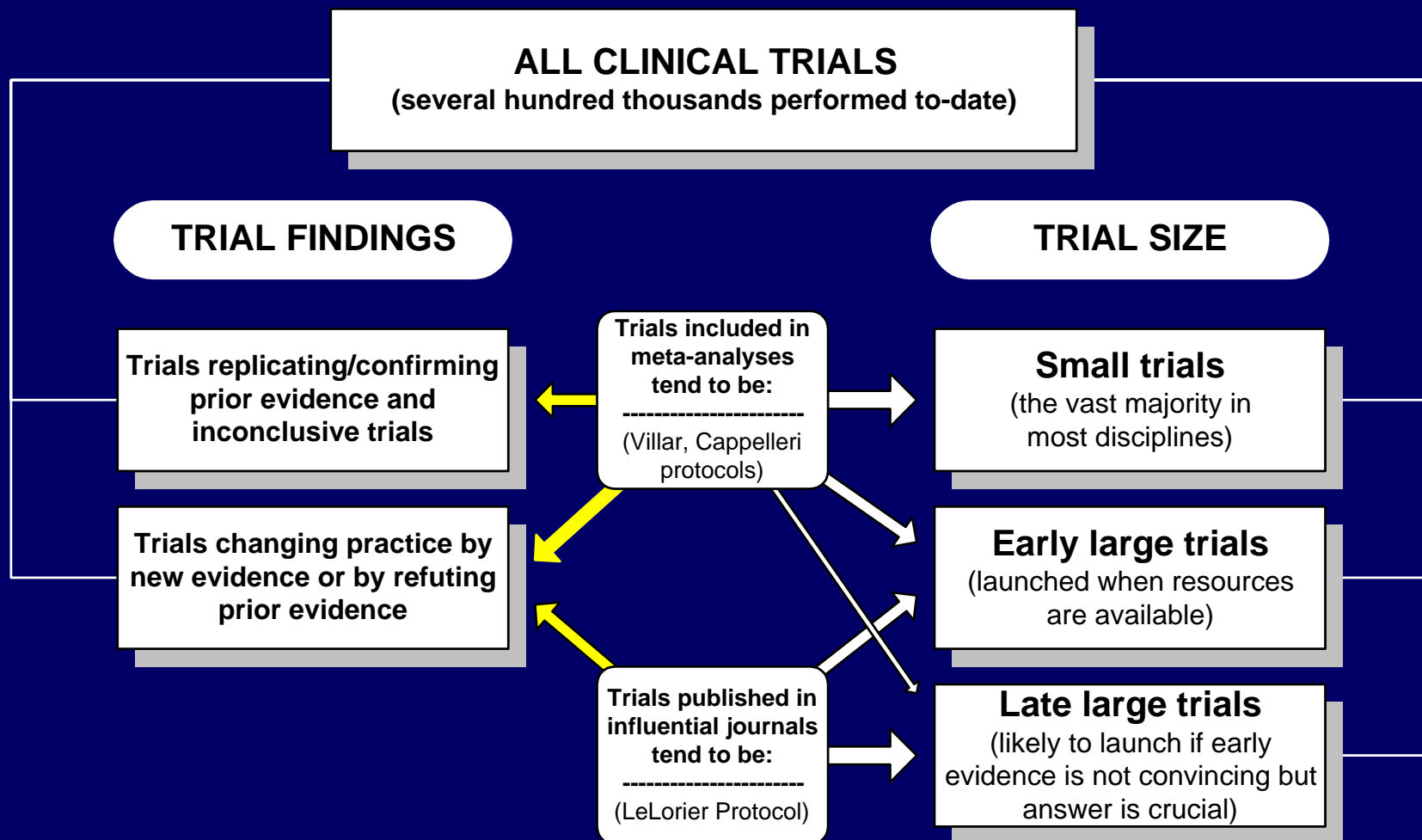
# Large Trials vs Meta-Analysis of Smaller Trials: How do their results compare ? (cont.) Cappelleri et al, JAMA 1996

- Of 15 disagreements between results of large and smaller trials using the random effects model, plausible explanations were identified in 10 meta-analyses:
  - 5 with differences in the control rate between large and smaller trials
  - 4 with specific protocol or study differences
  - 1 with potential publication bias
  
- 2 other disagreements were not clinically important tentative reasons could be identified for 2 of the remaining 3 disagreements

## Large trials vs meta-analysis of smaller trials: How do their results compare ?

- Meta-analyses of smaller studies are generally comparable with results from large studies.
- Differences can be attributed to insufficient sample sizes, control rates, or protocols.
- These reasons are not mutually exclusive.
- Publication bias is a possibility but has never been proven to be a factor.
- Need to explore reasons for heterogeneity.

# Some characteristics of clinical trials used in the protocols of comparison of large trials and meta-analyses of small trials



# Discrepancies between megatrials.

Furukawa et al. J Clin Epidemiol 2000;53:1193-99.

Why should large trials be the reference standard?

What do we know about the agreements among large trials on the same problem?

# Discrepancies between megatrials.

Furukawa et al. J Clin Epidemiol 2000;53:1193-99.

- “megatrial” defined as >1,000 patients
- 289 pairs identified in Cochrane Library
  - 79/289 (27%) pairs were statistically significantly different from each other
- 133 comparisons in LeLorier article
  - 36/133 (27%) were statistically significantly different

# Discrepancies between megatrials.

Furukawa et al. J Clin Epidem 2000;53:1193-99.

- Agreement among megatrials was approximately as large as that reported between meta-analyses and megatrials
- If we were to base the recommendation for the treatment in question on the primary outcome, 53% (Cochrane set) and 31% (LeLorier set) of the treatment recommendation by a megatrial was not confirmed by a later megatrial.
- On the other hand, 30% to 47% of the treatments once found ineffective or harmful in a megatrial were shown to be beneficial by a later megatrial.

## Insights from these empirical studies

- Heterogeneity of treatment effects is common among clinical trials, whether they are large or small; RCTs or observational studies
- Meta-analysis of small trials (dis)agree with large trials approximately as often dis(agreement) among large trials themselves
- We need to understand the cause of heterogeneity in clinical trials and learn how to handle them in meta-analysis

# Controversy due to quality assessment: **Screening mammography RCTs**

- Gotszche and Olsen. Lancet 2000;355:129
- A 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

## Controversy : Screening Mammography RCTs

- 8 trials identified
- Baseline imbalances were found in 6 of 8 trials
- 2 **adequately** randomized trials found no effect of screening on on breast cancer mortality
  - pooled risk ratio 1.04 (95% CI 0.84 - 1.27)
- 6 **inadequately** randomized trials found significant effect
  - Pooled risk ratio 0.75 (95% CI 0.67 – 0.83)

# Relative risk of death from breast cancer in screening versus control groups

	Number randomized		# of deaths from breast CA		Relative risk
	Screening	Control	Screening	Control	(95% CI)
<b>Randomization adequate</b>					
Malmö	21088	21195	63	66	0.96 (0.68-1.35)
Canada	44925	44910	120	111	1.08 (0.84-1.40)
<b>Total</b>	<b>66013</b>	<b>66105</b>	<b>183</b>	<b>177</b>	<b>1.04 (0.84-1.27)</b>
<b>Randomization not adequate</b>					
Göteborg	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)
Ostergötland	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)
<b>Total</b>	<b>182179</b>	<b>142052</b>	<b>654</b>	<b>725</b>	<b>0.75 (0.67-0.83)</b>

# Mammography screening trials according to methodological quality

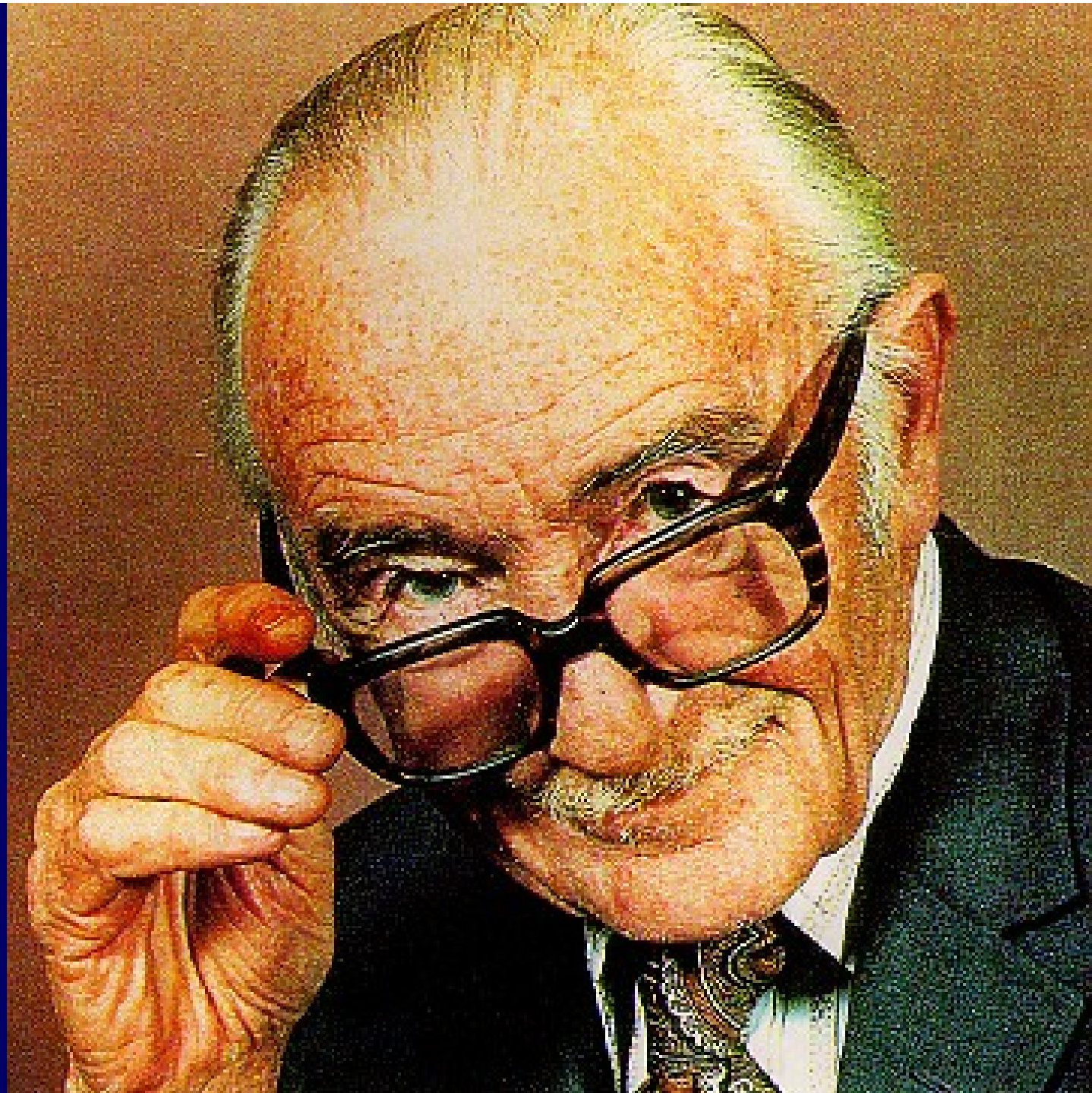
	Randomization produced similar groups	Account of number of patients consistent
Malmo	Yes	Yes
Canada	Yes	Yes
Goteberg	No	Yes
Stockholm	No	No
Kopparberg	No	No
Ostergotland	No	No
New York	No	No
Edinburgh	No	Yes

# Definition of Poor Quality

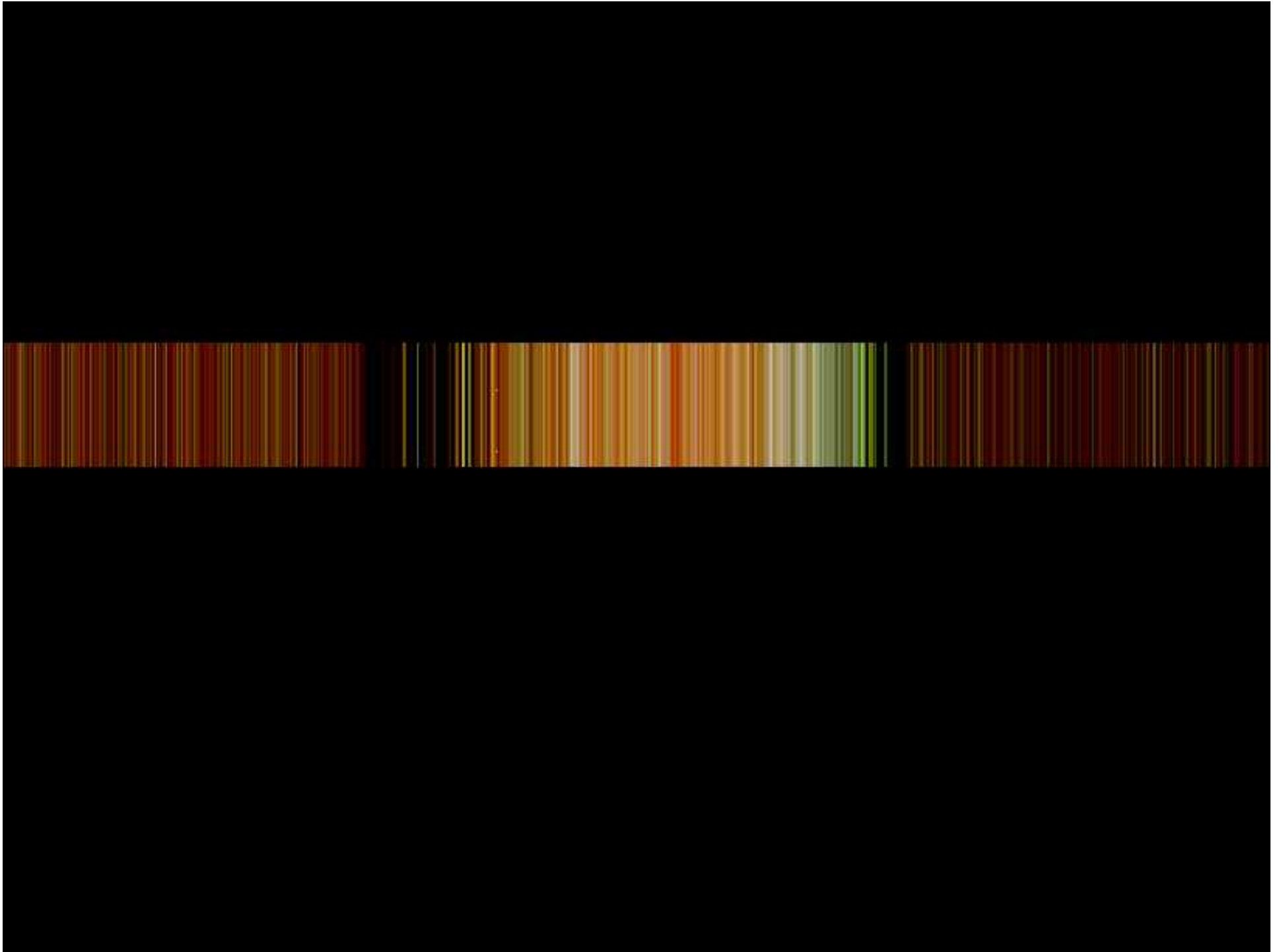
- Based on Randomization adequacy
  - Based on minor differences in mean age
  - Failed to consider other explanations for difference in mean ages
  - Failed to consider other measures of quality

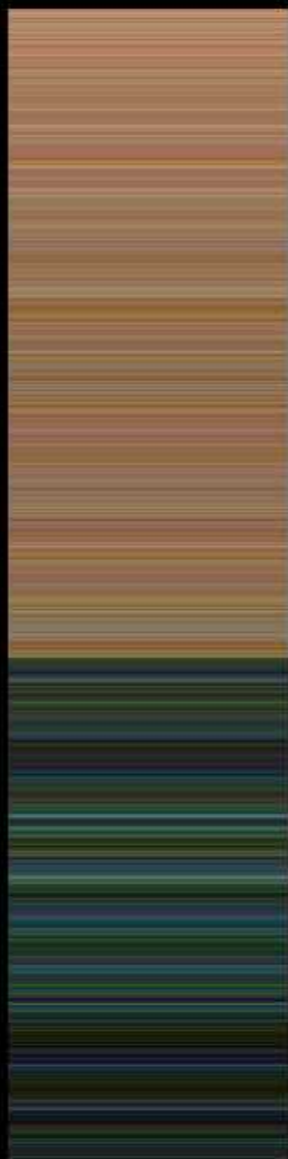
# Policy Results

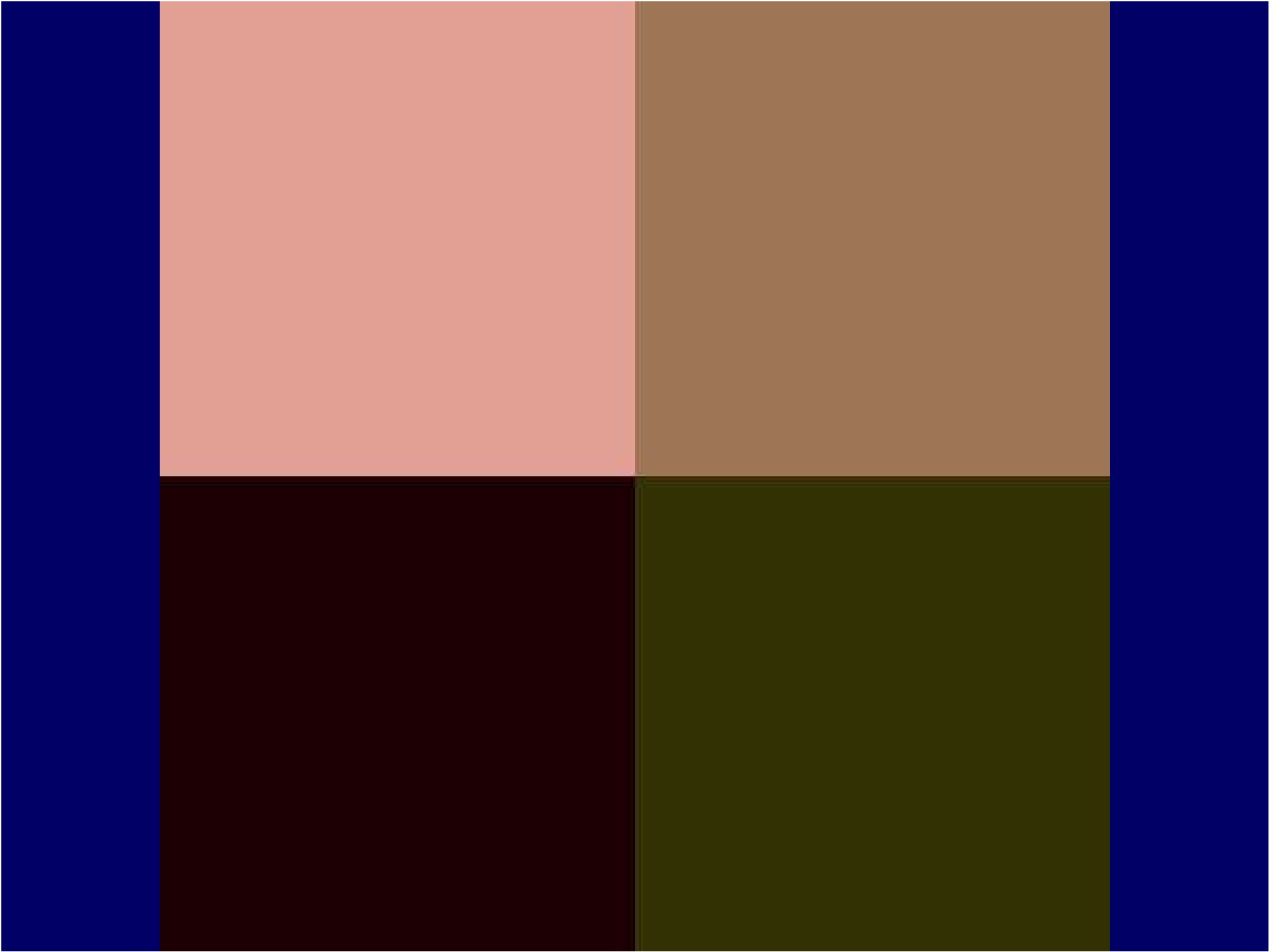
- Switzerland decided to not cover screening mammography
- NCI wavers on value of screening mammograms
- Women and doctors more confused about value of test







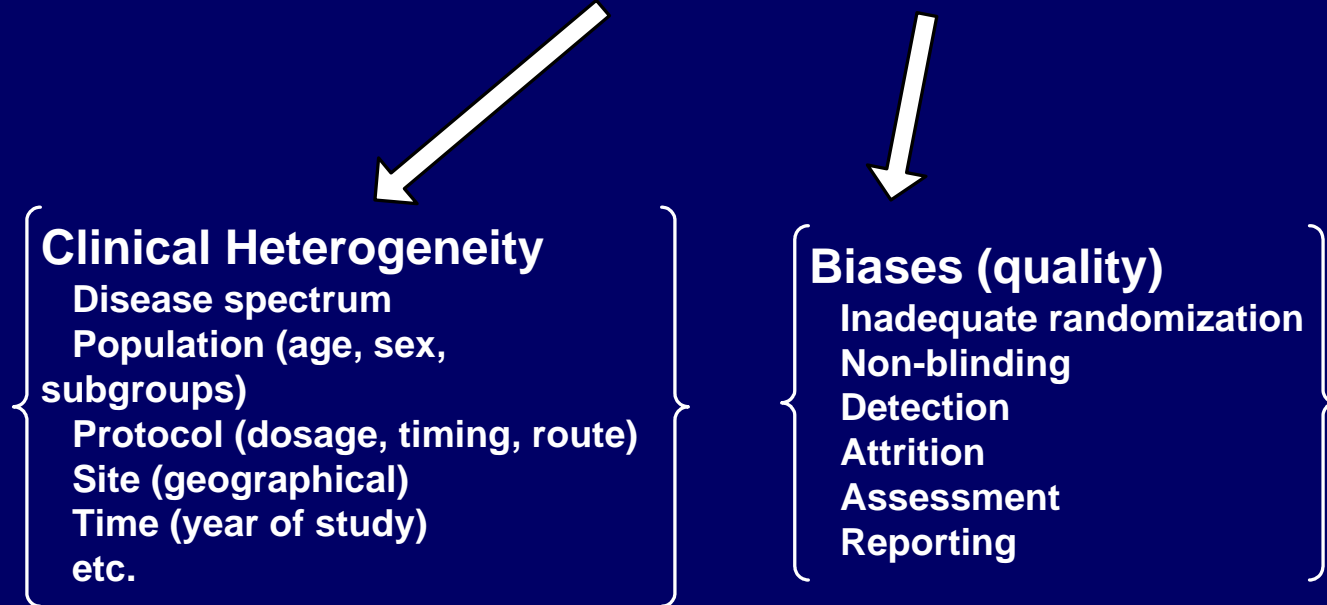






# Treatment effect observed (reported) in a RCT

Observed Effect = True Effect + Biases + Random Errors



$$TE_{\text{obs}} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_i x_i + \beta_j x_j + \dots + \varepsilon$$

# Estimating a single treatment effect across multiple trials in a meta-analysis

Weighted Average

$$\hat{\Omega} = \frac{\sum w_i TE_i}{\sum w_i}$$

$\theta_i$  = true effect of individual study

Biases<sub>M-A</sub>

publication bias  
selection bias  
etc.

$$\hat{\Omega}_{adj} = \frac{\sum w_i (\theta_i + \text{biases}_i + \varepsilon_i)}{\sum w_i} + \text{Biases}_{M-A}$$