Precision in Meta-Analysis
-
Are we being misled by the play of chance?

Presentation objectives

1. Intro to precision in meta-analysis (MA), overview of current research

2. Debate on future challenges and required efforts

Outline

I: Intro

II: Preparation for debate

III: Overview of current research

IV: Debate and questions

Part I:

Intro

Inferences in meta-analysis

We perform significance tests because we want to control the risk that we are being mislead by the play of chance

Inferences in meta-analysis

When a meta-analysis becomes 'statistically significant', authors, journal editors, policy makers, etc. typically put strong confidence in the estimated effect

Inferences in meta-analysis

This is all fine if the statistical methods do what they are supposed to do.

Inferences in meta-analysis

But if they don't....

- how often are we actually being mislead by the play of chance?
- how often will the implications for clinical practice be serious?

Statistical tests in meta-analysis

We typically use Z-statistics to test for 'statistical significance'

$$Z = \frac{\text{Estimated treatment difference}}{\text{Standard Error}}$$

Z is subsequently transformed to a p-value

Statistical tests in meta-analysis

We conclude some treatment effect is statistically significant when our p-value crosses below the overall type I error, α (or when $|Z|$ crosses above Z_α)

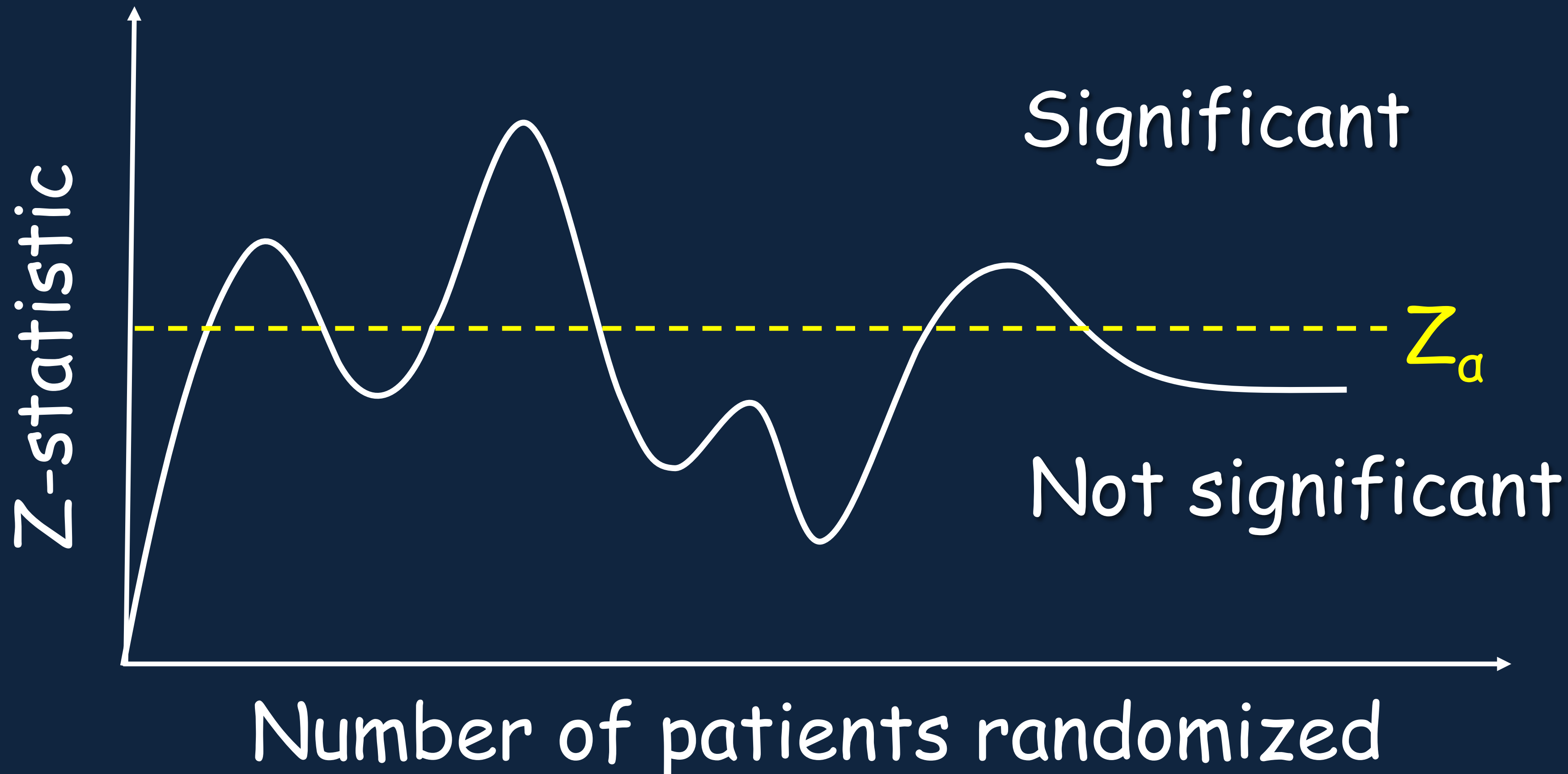
We assume the pooled treatment effect estimate is reliable

Statistical tests in meta-analysis

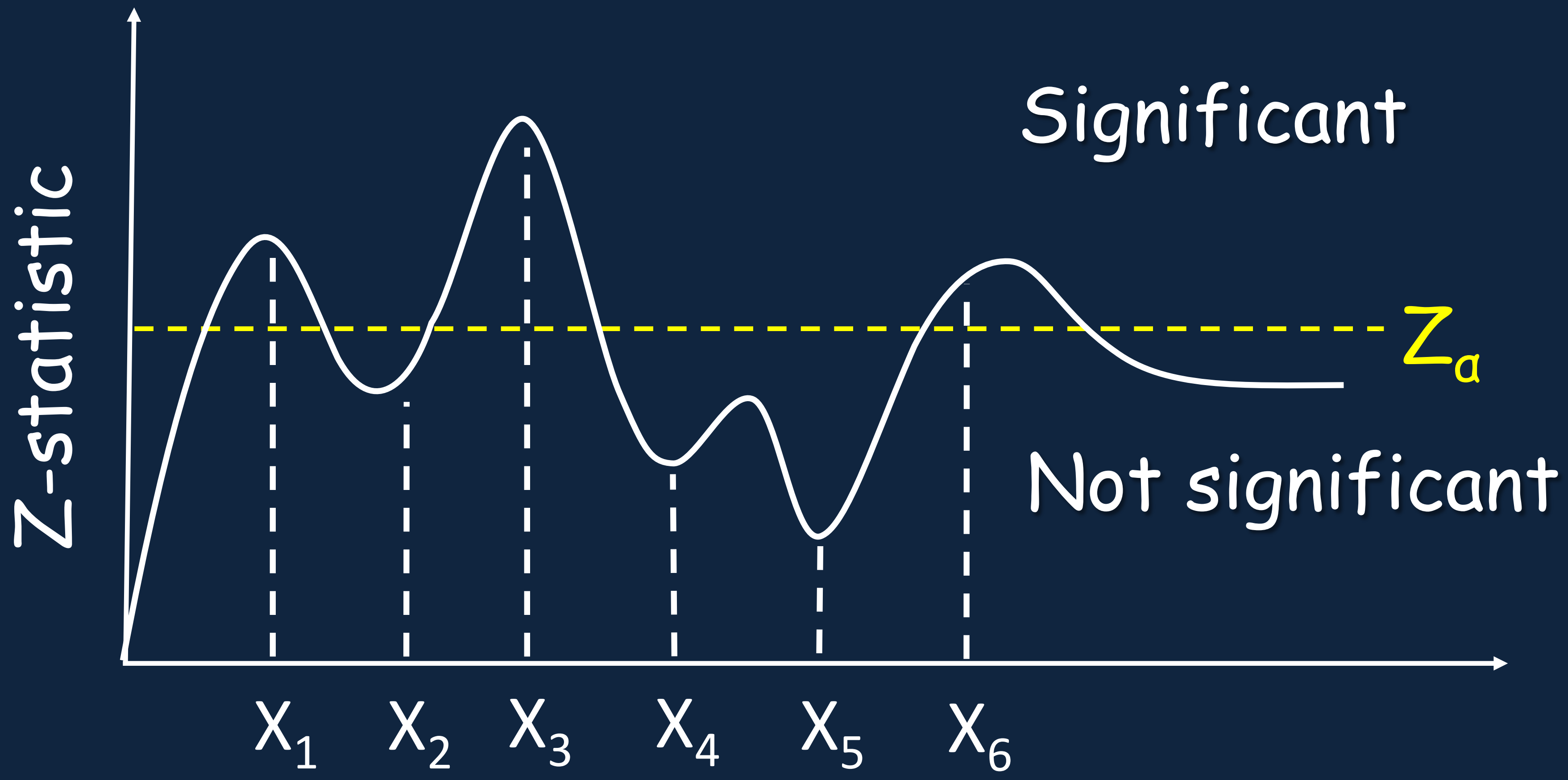
This conduct can lead to

1. exaggeration of type I error (multiplicity)
2. overestimation of treatment effects

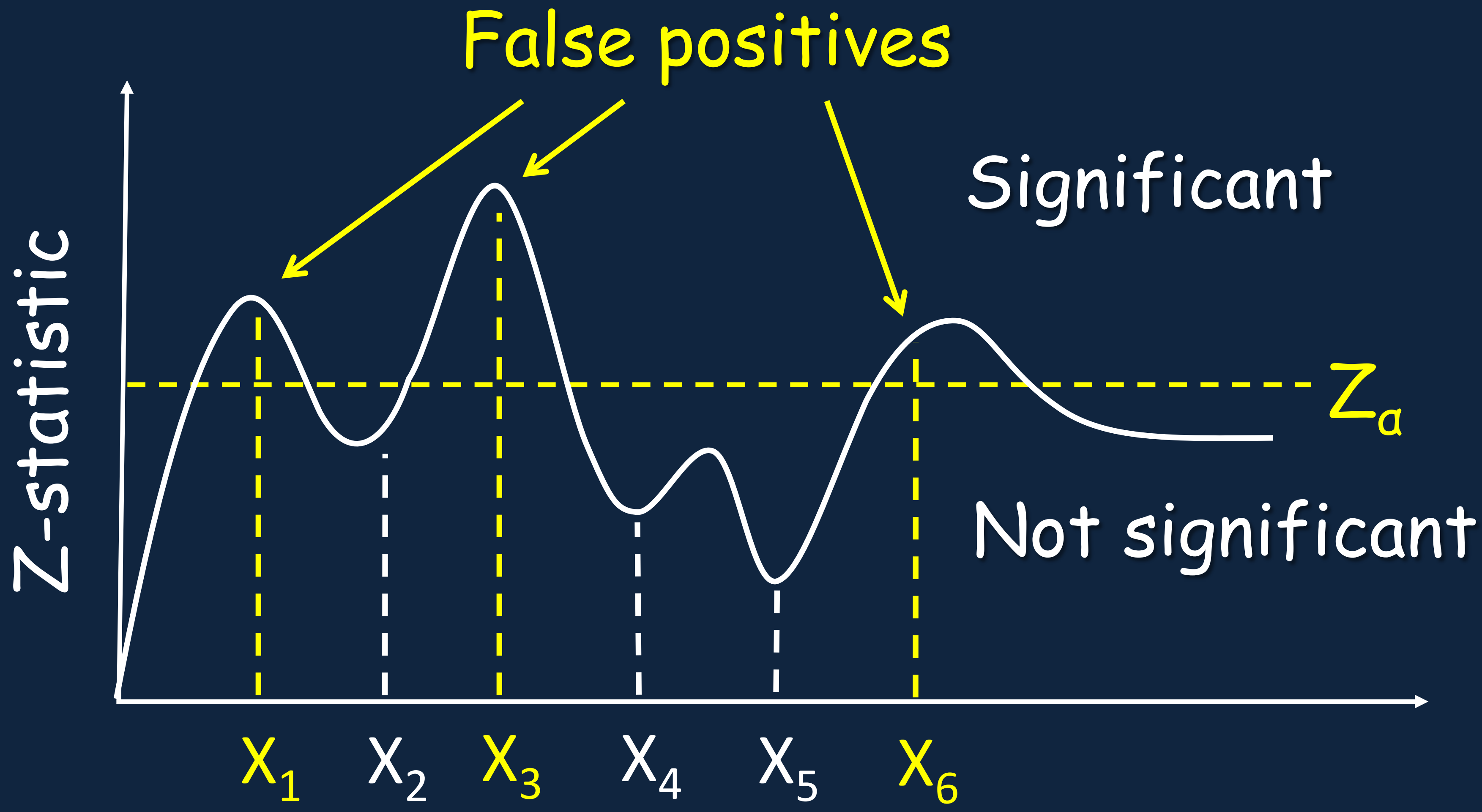
Issue 1: Multiplicity



Issue 1: Multiplicity



Issue 1: Multiplicity



Issue 1: Multiplicity

Every time we test for statistical significance over time we increase the risk of type I error (multiplicity)

This problem occurs in meta-analysis due to updating

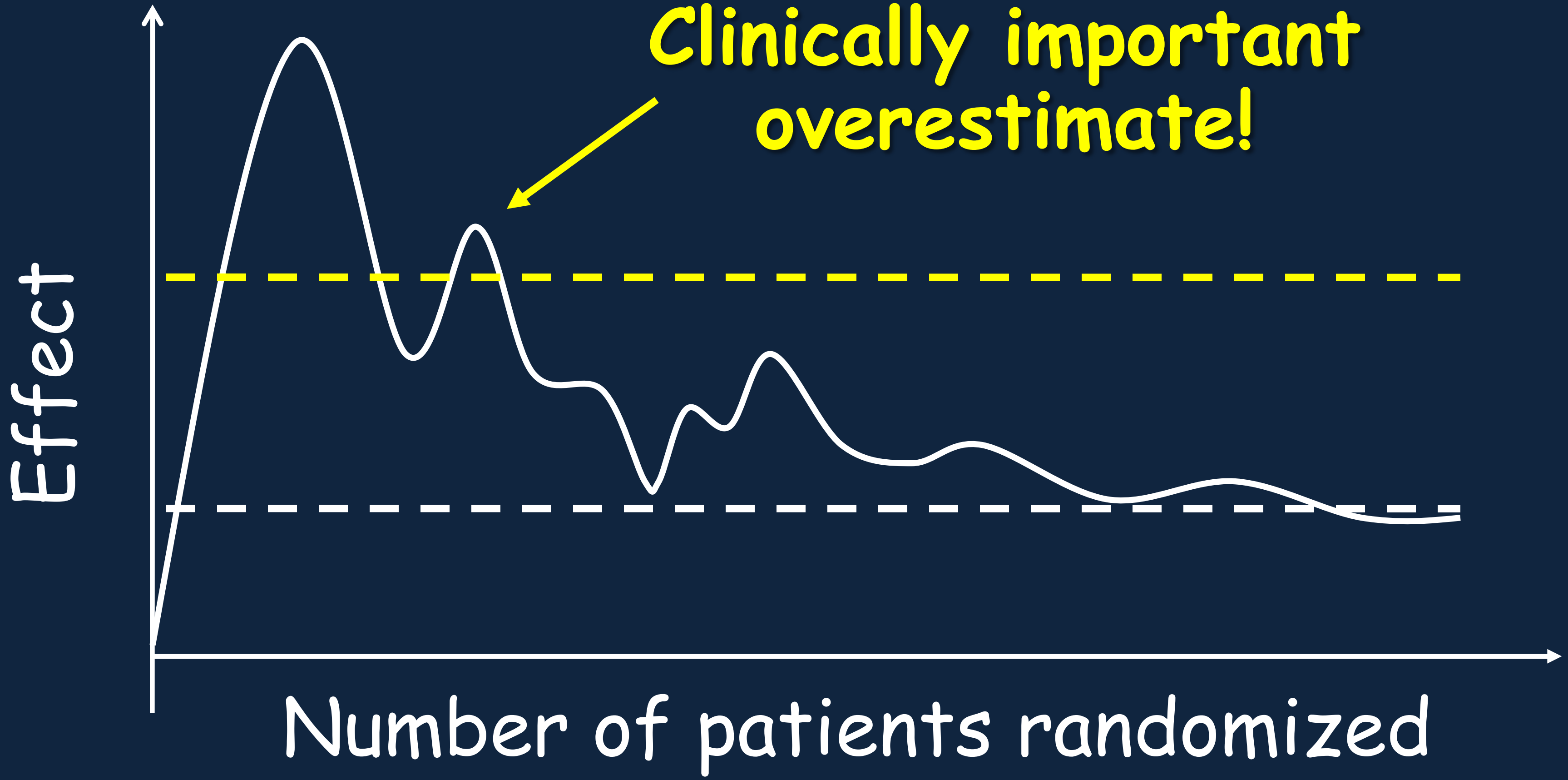
Issue 2: Overestimation

In 'early' meta-analysis precision is low, and thus, the standard error is large.

$$Z = \frac{\text{Estimated treatment difference}}{\text{Standard Error}}$$

In 'early' statistically significant MAs, the est. treatment difference are large

Issue 2: Overestimation



Clinically important overestimate!

Part II:

Preparation for debate

INPUT WANTED

As you will see in the following slides, both false positives (multiplicity) and overestimation is often problematic in meta-analysis that draw 'conventional' statistical inferences

INPUT WANTED

To the extent policy makers and clinicians rely on meta-analyses, the implications may be serious

INPUT WANTED

Potential solutions comprise

1. Adjustment of thresholds or tests
2. Setting some yardstick for when 'the answer is in' (e.g. required sample size)

Questions?

Research efforts so far have focused on superiority testing for binary outcome meta-analysis...

Questions?

- What more is needed to seal the deal?
- What about inferiority testing?
- What about other types of data?
  - continuous (HRQL)
  - time-to-event/survival

Questions?

Clinicians are notorious for relying on thresholds, statisticians for making things too complex.

For decision-making, can we meet in the middle?

Questions?

How do we (largely) avoid misuse of proposed methodologies?

Part III:

Overview of current research

Part III:

From 1992 to present day

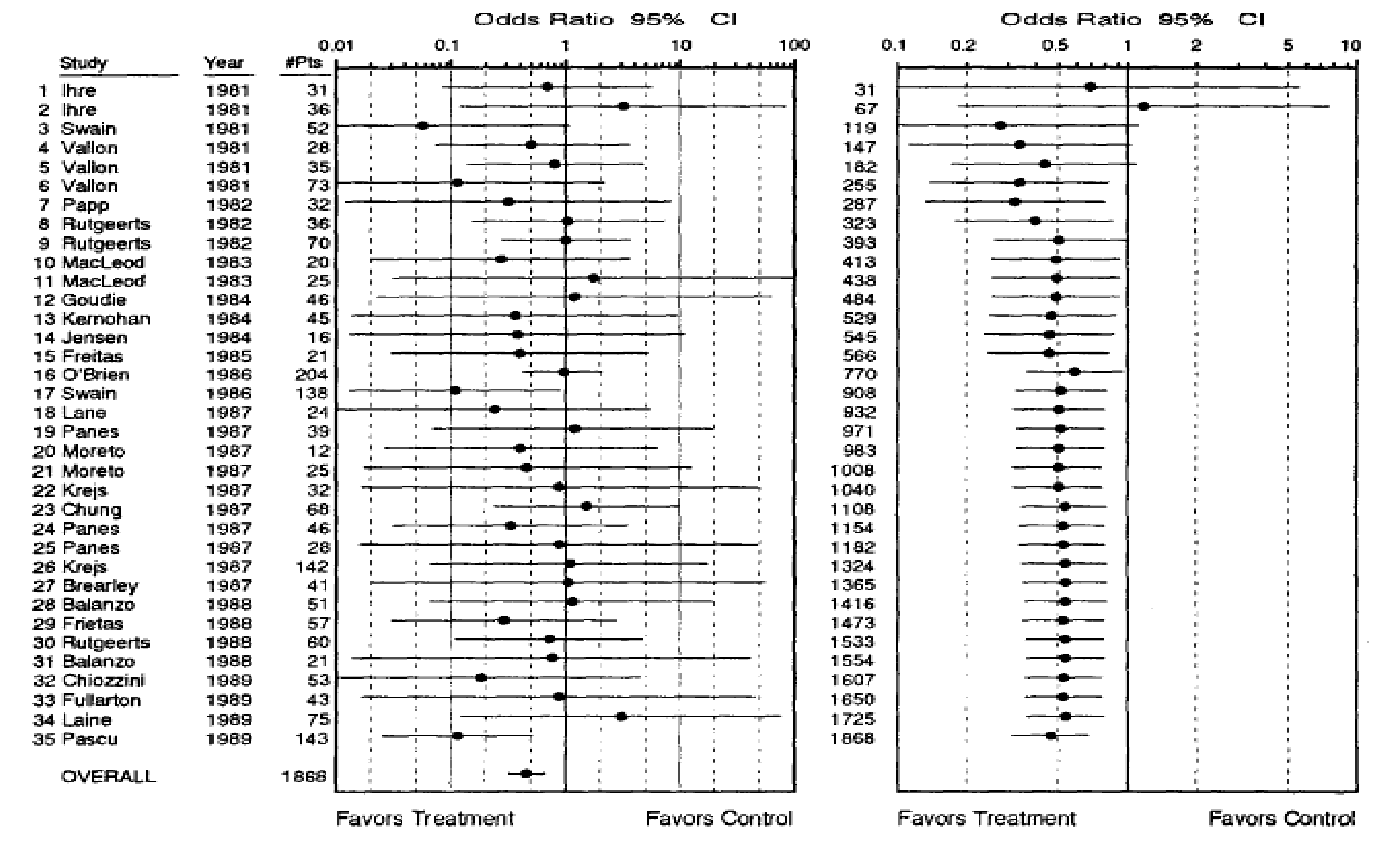
1992-1995

Lau et al. proposed cumulative meta-analysis

NEJM 1992; 327(4):248-254
J Clin Epi 1995; 48(1):357-371

1992-1995

Lau et al. proposed cumulative meta-analysis



	Study	Year	#Pts	
1	Ihre	1981	31	31
2	Ihre	1981	36	67
3	Swain	1981	52	119
4	Vallon	1981	28	147
5	Vallon	1981	35	182
6	Vallon	1981	73	255
7	Papp	1982	32	287
8	Rutgeerts	1982	36	323
9	Rutgeerts	1982	70	393
10	MacLeod	1983	20	413
11	MacLeod	1983	25	438
12	Goudie	1984	46	484
13	Kernohan	1984	45	529
14	Jensen	1984	16	545
15	Freitas	1985	21	566
16	O'Brien	1986	204	770
17	Swain	1986	138	908
18	Lane	1987	24	932
19	Panes	1987	39	971
20	Moreto	1987	12	983
21	Moreto	1987	25	1008
22	Krejs	1987	32	1040
23	Chung	1987	68	1108
24	Panes	1987	46	1154
25	Panes	1987	28	1182
26	Krejs	1987	142	1324
27	Brearley	1987	41	1365
28	Balanzo	1988	51	1416
29	Frietas	1988	57	1473
30	Rutgeerts	1988	60	1533
31	Balanzo	1988	21	1554
32	Chiozzini	1989	53	1607
33	Fullarton	1989	43	1650
34	Laine	1989	75	1725
35	Pascu	1989	143	1868
	OVERALL		1868	

1992-1995

Maybe this is where is went wrong...

"The cumulative aspect of the [frequentist] meta-analysis ... are calculated and interpreted in through the Bayesian paradigm"

Conclusion: no need to adjust for multiplicity

1996-1998

Berkey et al simulated 'uncertainty of first time to significance' with associated power and type I error using real MA data (MCMC)

- With α=5%, the actual type I error after 15 trials was 15% (update for every trial)

Contr Clin Trials 1996; 17(5):357-371

1996-1998

Pogue&Yusuf and Whitehead seperately proposed use of 'information size' considerations and sequential monitoring boundaries to control type I error

Stat Med 1997; 16(24):2901-2913
Contr Clin Trials 1997; 18(6):580-93
Lancet 1998; 351(9095):47-52

1996-1998

Whitehead proposed use of the *triangular test* for random-effects meta-analysis and performed a simulation study

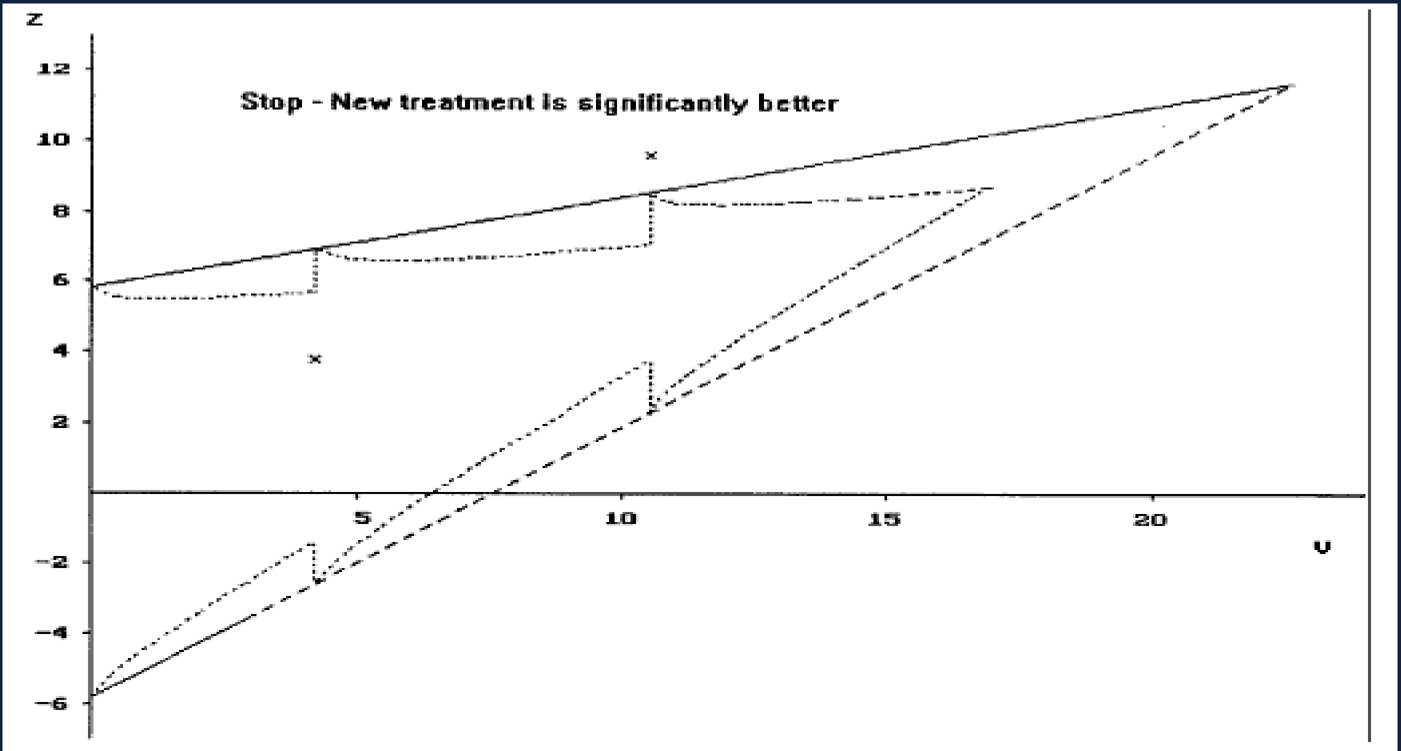


Figure 2. Sample path of the example

1996-1998

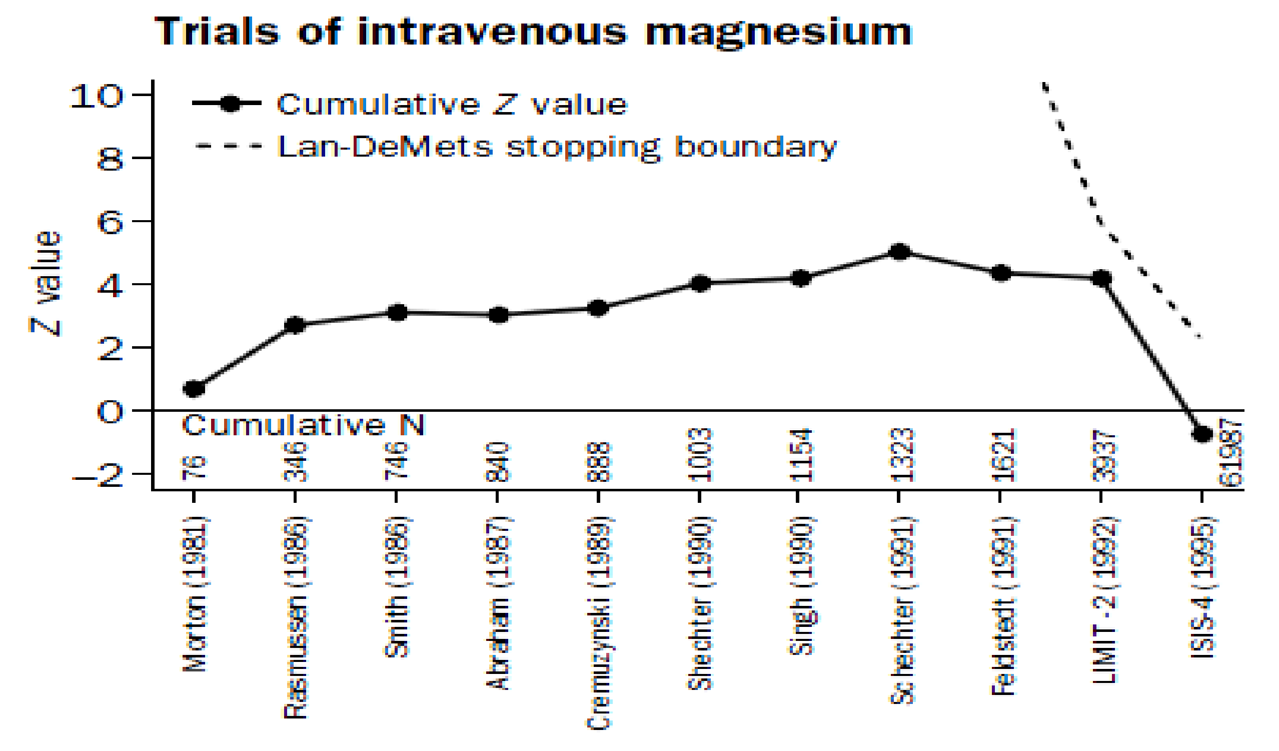
Whitehead proposed use of the *triangular test* for random-effects meta-analysis and performed a simulation study

- Only found satisfactory control of type I error when heterogeneity was mild

- Information scale (Fischer) is only reliable with good heterogeneity estimates

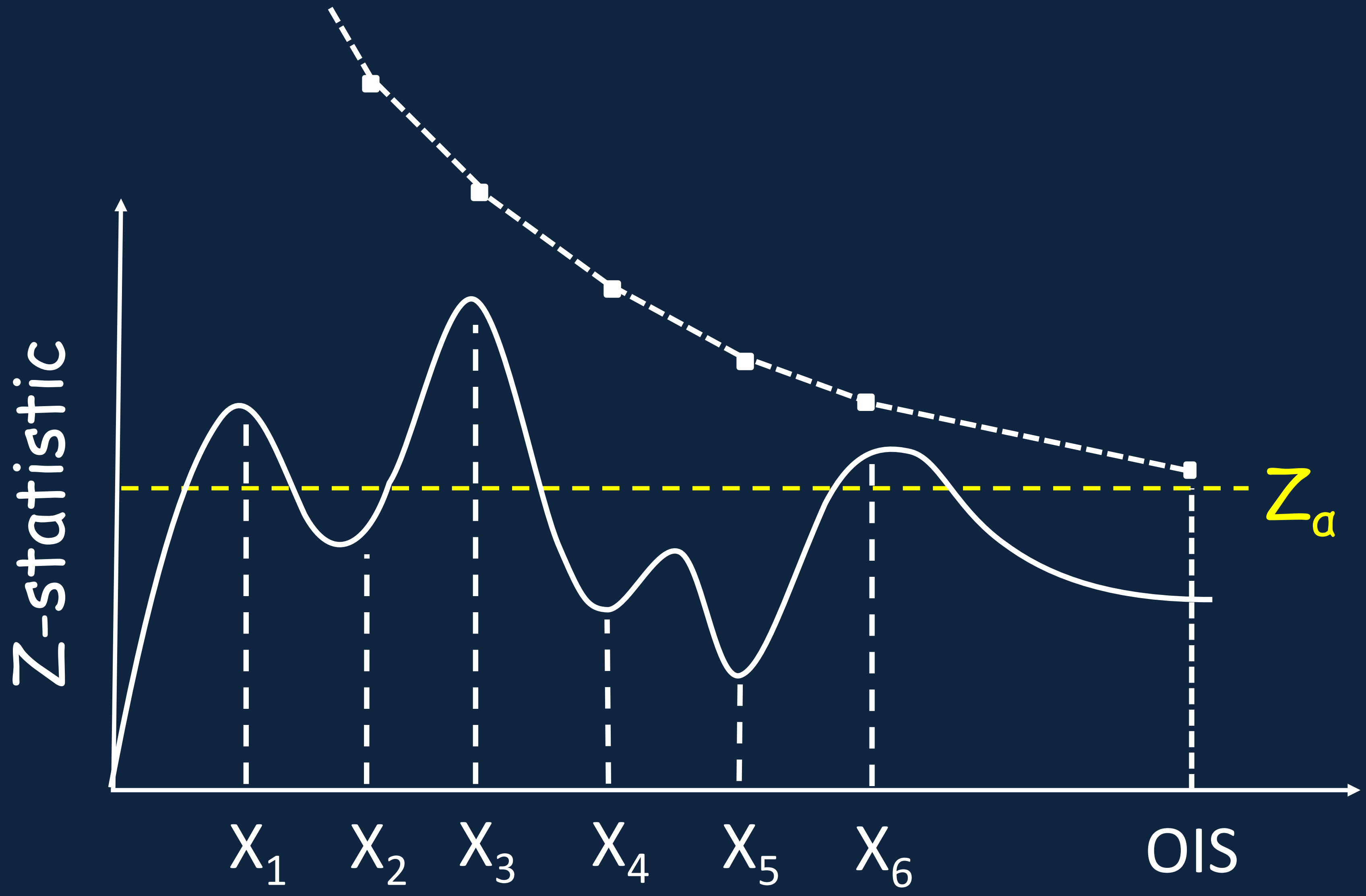
1996-1998

Pogue&Yusuf proposed use of Lan-DeMets α-spending monitoring boundaries and the Optimum Information Size (OIS)



Trials of intravenous magnesium

1996-1998



2001 - 2005

First signs of empirical evidence

First actual application

Proc Natl Acad Sci USA 2001; *98*(3):831-836
J Clin Epi 2004; 57(11):1124-1130
BMJ 2005; 331(7512):313-321

2001-2005

First signs of empirical evidence:

Ioannidis et al looked at the 'Evolution of treatment effects over time' in 60 meta-analyses on interventions in perinatal medicine and for myocardial infarction

2001-2005

Ioannidis 60 meta-analyses: relative change in estimates per added trial

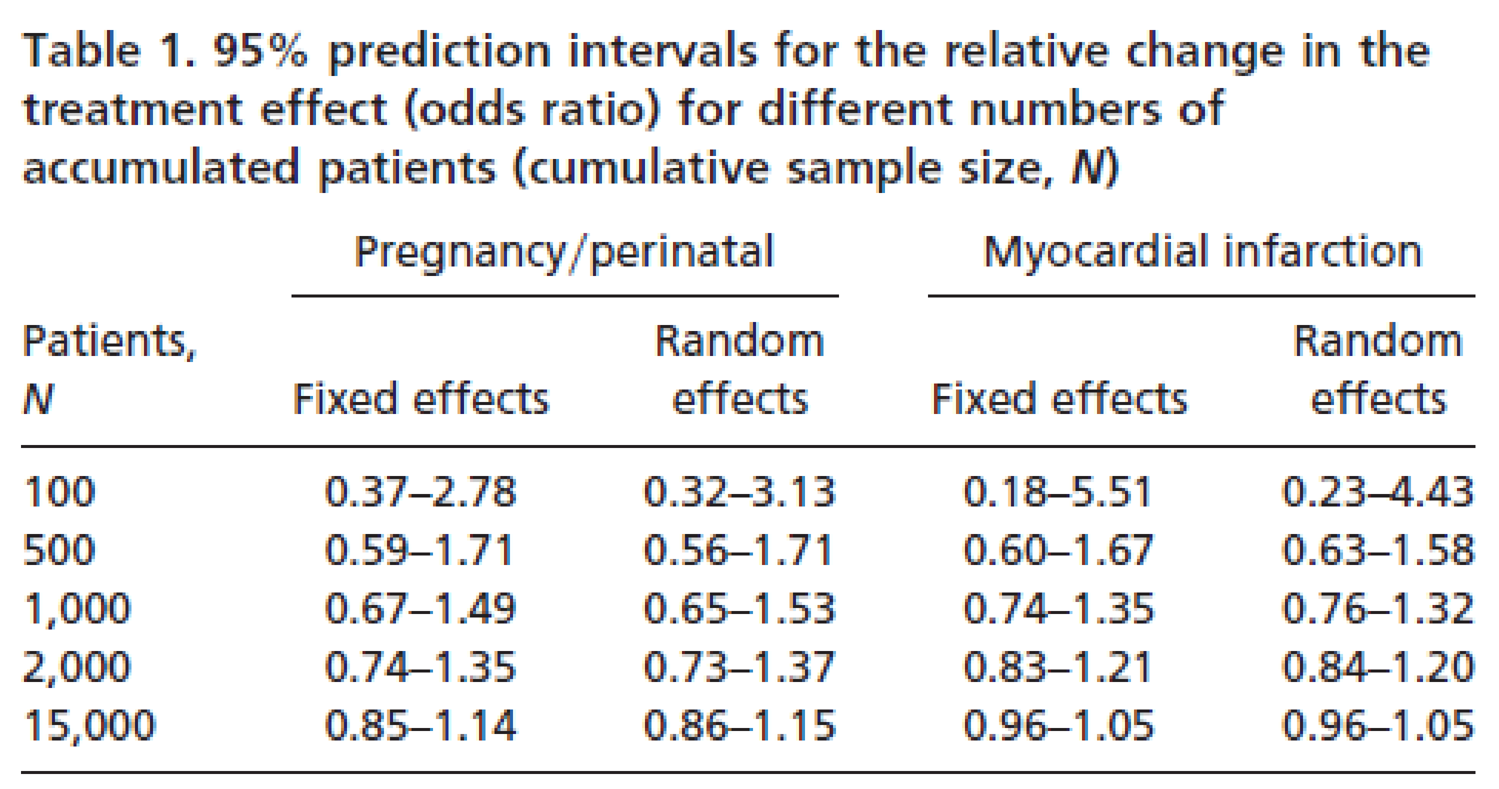


Table 1. 95% prediction intervals for the relative change in the treatment effect (odds ratio) for different numbers of accumulated patients (cumulative sample size, *N*)

	Pregnancy/perinatal		Myocardial infarction	
Patients, *N*	Fixed effects	Random effects	Fixed effects	Random effects
100	0.37–2.78	0.32–3.13	0.18–5.51	0.23–4.43
500	0.59–1.71	0.56–1.71	0.60–1.67	0.63–1.58
1,000	0.67–1.49	0.65–1.53	0.74–1.35	0.76–1.32
2,000	0.74–1.35	0.73–1.37	0.83–1.21	0.84–1.20
15,000	0.85–1.14	0.86–1.15	0.96–1.05	0.96–1.05

2001-2005

Ioannidis 60 meta-analyses: relative change in estimates per added trial

“More than 10,000 patients are required to relieve uncertainty about the first decimal point in the odds ratio of a treatment effect reported by a meta-analysis.”

2001-2005

First sign of empirical evidence:

Trikalinos et al looked at the evolution of treatment effects and statistical significance in 100 mental health meta-analyses

2001 - 2005

100 patients: subsequent changes in odds ratios of 3- to 5-fold were common

500 patients, changes >1.5-fold were only observed in 5% of the meta-analyses

>2000 patients randomised, subsequent changes were unlikely

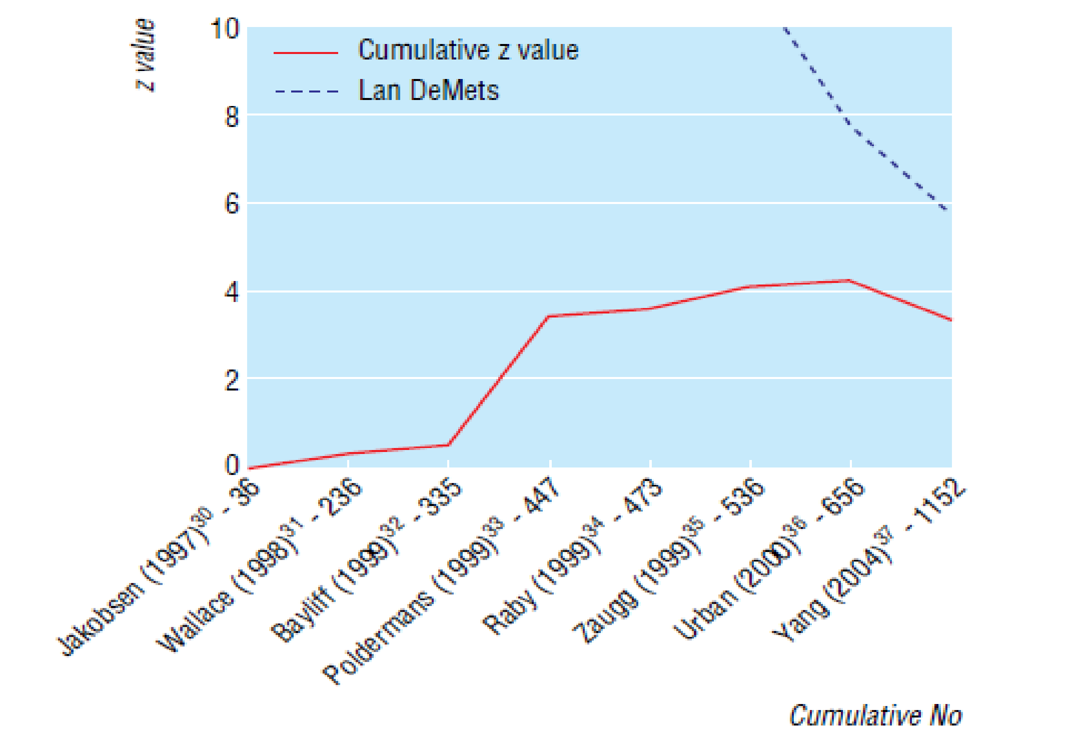
2001-2005

In the interim, 8 of the 44 meta-analyses showing no effect were temporarily statistically significant

(=18.2% max type I error)

2001-2005

First application by PJ Devereaux et al: Meta-analysis of periop beta-blockade in non-cardiac surgery



2006-2009

- More compelling evidence from empirical and simulation studies

- Methodological proposals

- Increasing awareness (e.g. GRADE)

- Applications in systematic reviews

2006-2009

Copenhagen Trial Unit group:
Applied monitoring boundaries and heterogeneity adjusted OIS to all Cochrane neonatal and other meta-analyses

J Clin Epi 2008; 61(1):64-75
J Clin Epi 2008; 61(8):763-769
Int J Epi 2009; 38(1):276-86
Int J Epi 2009; 38(1):287-98
BMC Med Res Meth 2009; 9:86

2006-2009

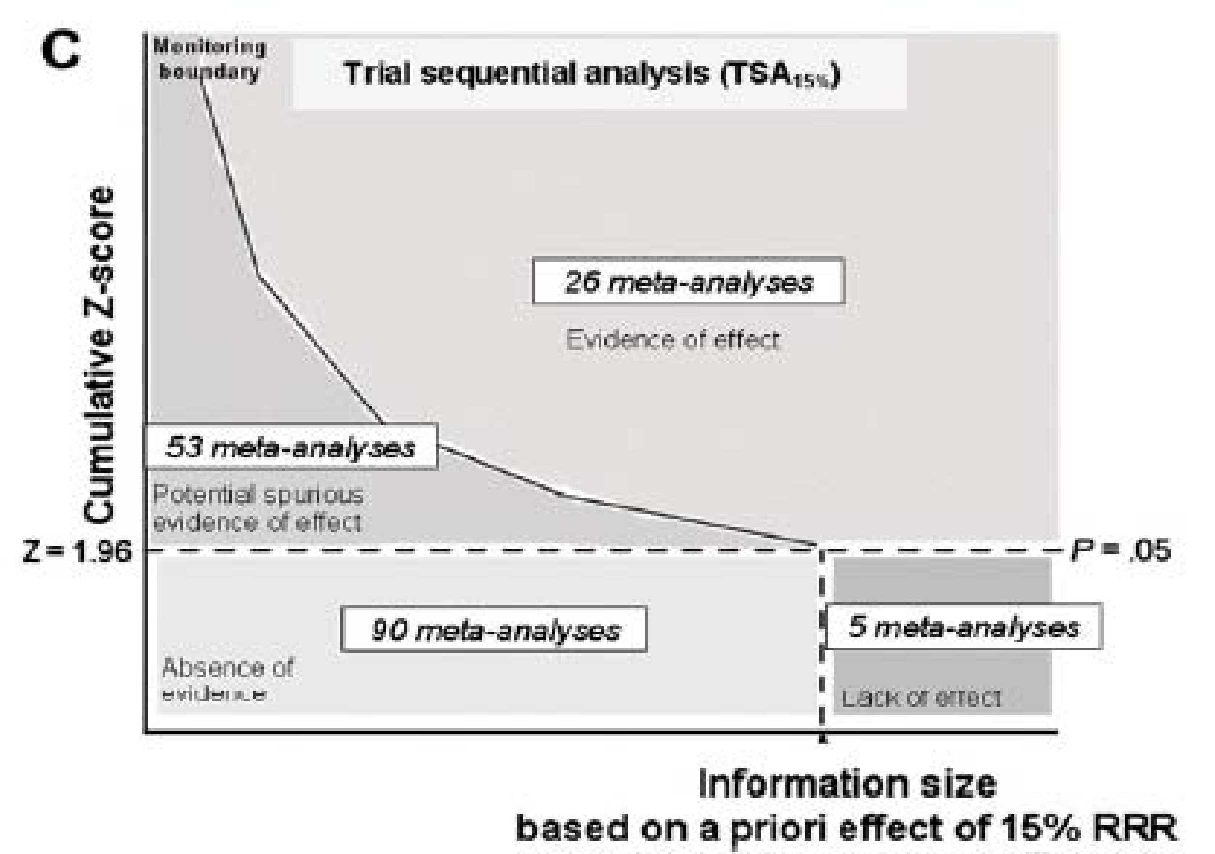
Copenhagen Trial Unit group:
Proposed heterogeneity adjusted
Optimum Information Size (OIS)

$$OIS=N/(1-H)$$

N is the required sample size for a RCT
H is the degree of heterogeneity (0-100%)

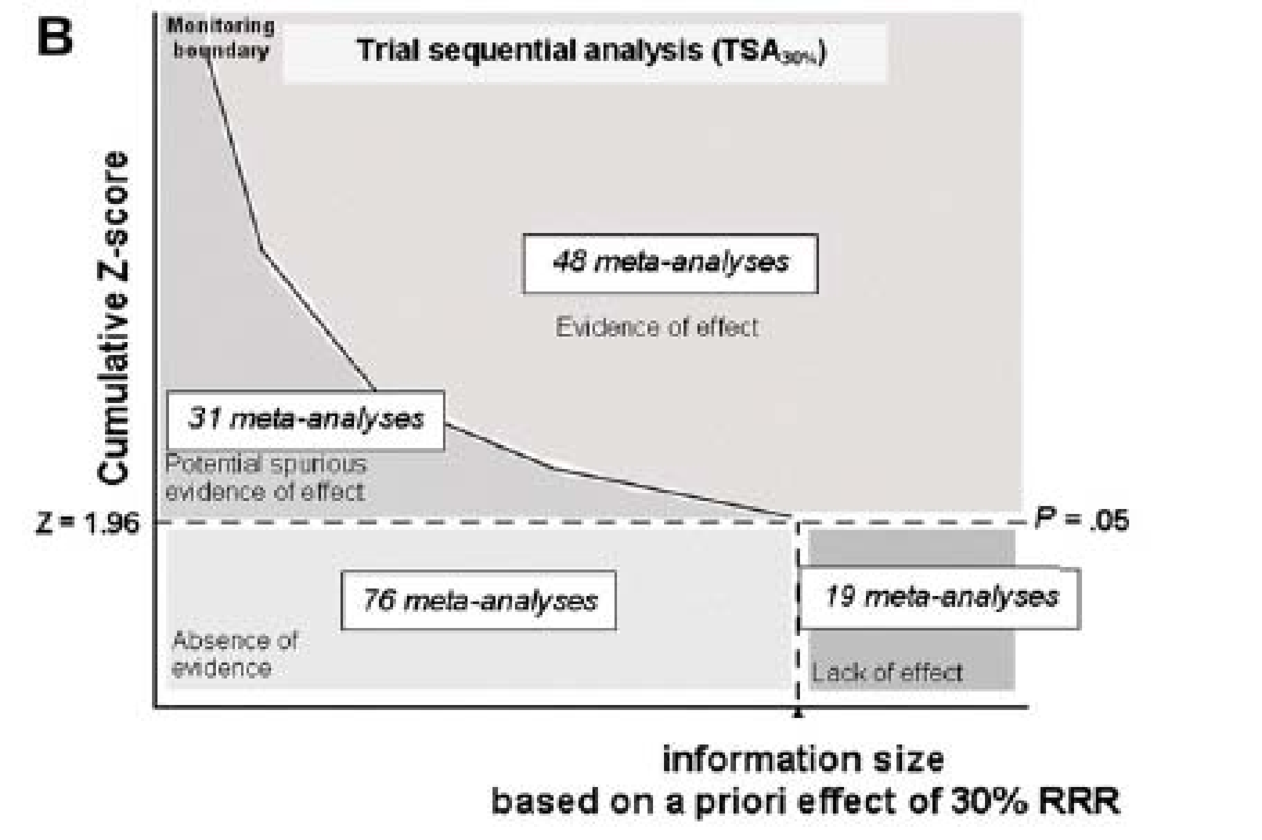
2006-2009

Copenhagen Trial Unit group:
Application to Cochrane neonatal reviews



2006-2009

Copenhagen Trial Unit group:
Application to Cochrane neonatal reviews



2006-2009

Copenhagen Trial Unit group:
Application to Cochrane neonatal reviews

Conclusion: Most meta-analysis are inconclusive and potentially spurious

The recommendations in current guidelines is discrepant with the strength of evidence

2006-2009

Copenhagen Trial Unit group:
Application to 33 meta-analysis surpassing their heterogeneity adjusted OIS

6 out of 21 'positive' meta-analysis yielded important overestimates at *first time of statistical significance*

2006-2009

Copenhagen Trial Unit group:
Application to 33 meta-analysis surpassing their heterogeneity adjusted OIS

3 out of 12 'negative' meta-analysis yielded false positive results in the interim

2006-2009

Copenhagen Trial Unit group:
Application to 33 meta-analysis surpassing their heterogeneity adjusted OIS

Monitoring boundaries eliminated all spurious inferences

2006-2009

Compelling evidence from simulations and other proposed methods

Clin Trials 2007; 4:329
J Clin Epi 2009; 62:825-830

2006-2009

Hu et al proposed a *penalized* Z-statistic based on *the law of the iterated logarithm* and simulation

$$Z^*(k) = Z(k)/(\sqrt{\lambda} \cdot \log(\log(I(k)))$$

Where k is the number of trials and $I(k)$ is the inverse of the pooled variance

2006-2009

Hu et al simulations

- Repeated significance testing (with α=5%) yields an actual type I error of 15-35%
- The penalized Z-statistics exhibits good control of the type I error (reasonable values for λ is provided in the paper)

2006-2009

Borm et al proposed a *k-fail-safe* adjustment of the p-value based on the max number of MA updates and simulation

Regression on *Type I error* = $\alpha \cdot f$, where *f* is some function of the max number of updates in the meta-analysis

2006-2009

Borm et al proposed a *k-fail-safe* adjustment of the p-value based on the max number of MA updates and simulation

$$P^* = P \cdot \sqrt{(6 \cdot Max\ no.\ updates - 1.5)}$$

2006-2009

Borm et al simulations

- Repeated significance testing (with one-sided α=2.5%) yields 2- to 7-fold inflation of the type I error
- The adjusted P-value exhibits good control of the type I error

2006-2009

GRADEProfiler recommends downgrading of the quality of the overall evidence if a meta-analysis does not surpass its OIS (or has less than 300 events)

For that reason, a good number of systematic reviews now consider the meta-analysis sample size

2006-2009

Copenhagen Trial Unit group:
A number of applications to systematic reviews as well as empirical studies

2006-2009

Copenhagen Trial Unit group - applications:

Whenever one of us was invited to co-author a SR we have used OIS and monitoring boundaries to achieve reliable statistical inferences

2010 -

More simulation studies

More empirical studies

More applications

Abuse of the methodology?

2010 -

Simulation study sneak peak

Plausible cardiology meta-analysis scenario
(based on survey of Cochrane Heart Group meta-analysis on mortality)

2010 -

Plausible cardiology meta-analysis scenario (random-effects model simulation)

True effect: RR=0.80

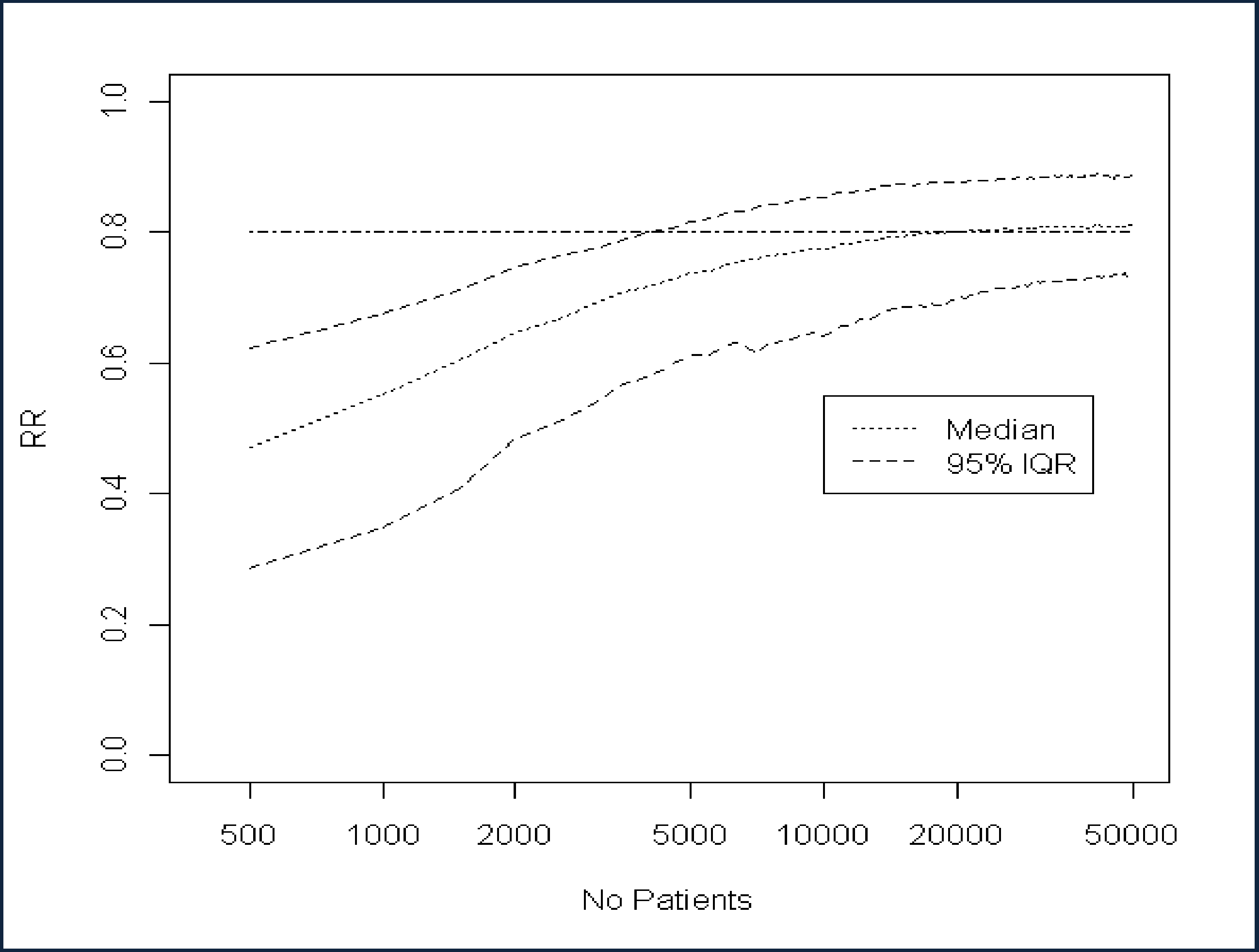
Event rate: 1%-15%

Moderate heterogeneity (RR, 0.60-1.05)

Trial sizes: 40-400(25%), 401-1000(65%) and 1001-10000(10%)

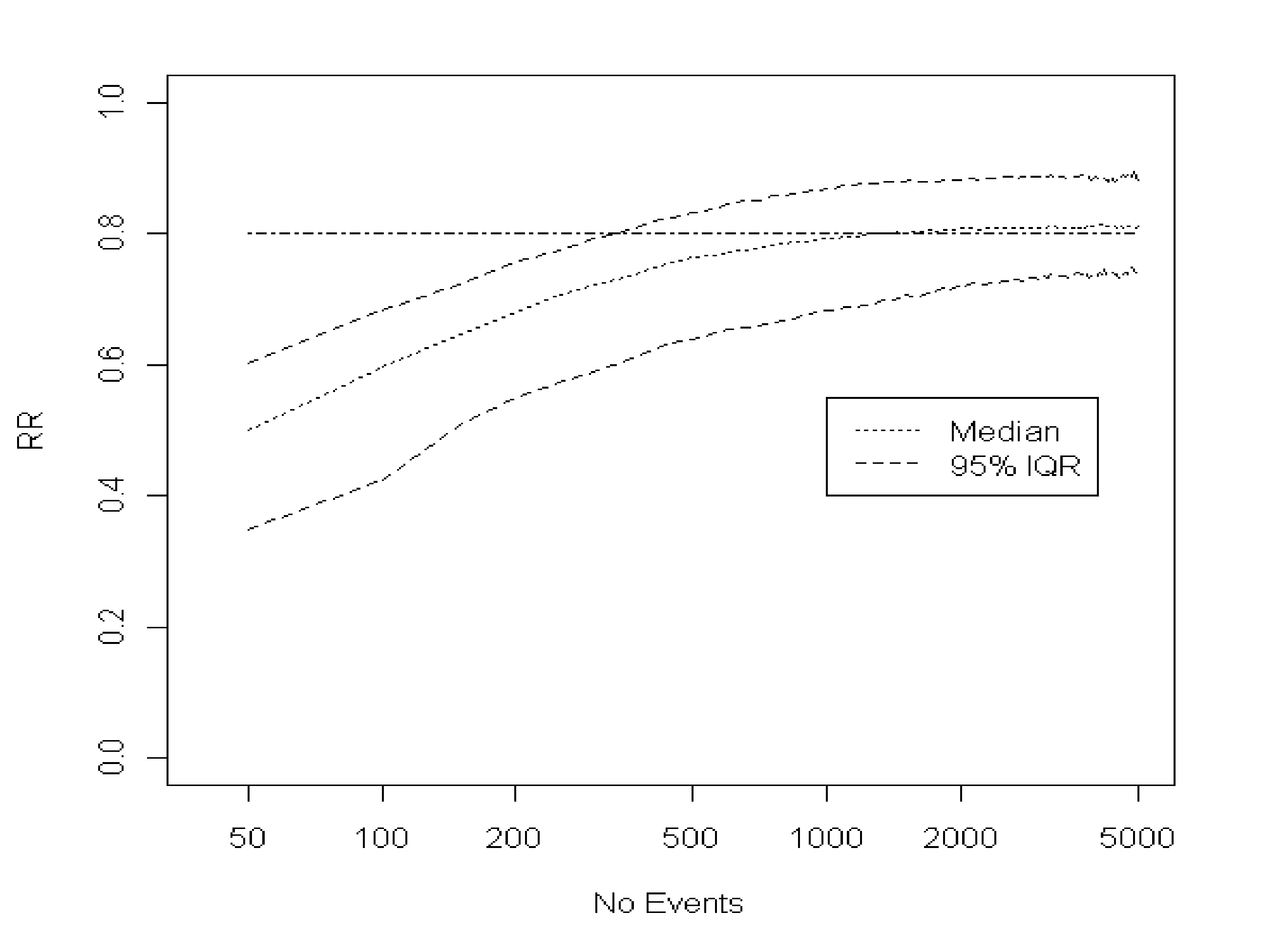
2010 -

Plausible cardiology meta-analysis scenario



2010 -

Plausible cardiology meta-analysis scenario



2010 -

Alternative applications:
Isoniazid chemoprophylaxis (IHZ) for preventing tuberculosis (TB) among purified protein derivative negative (PPD-) HIV infected individuals

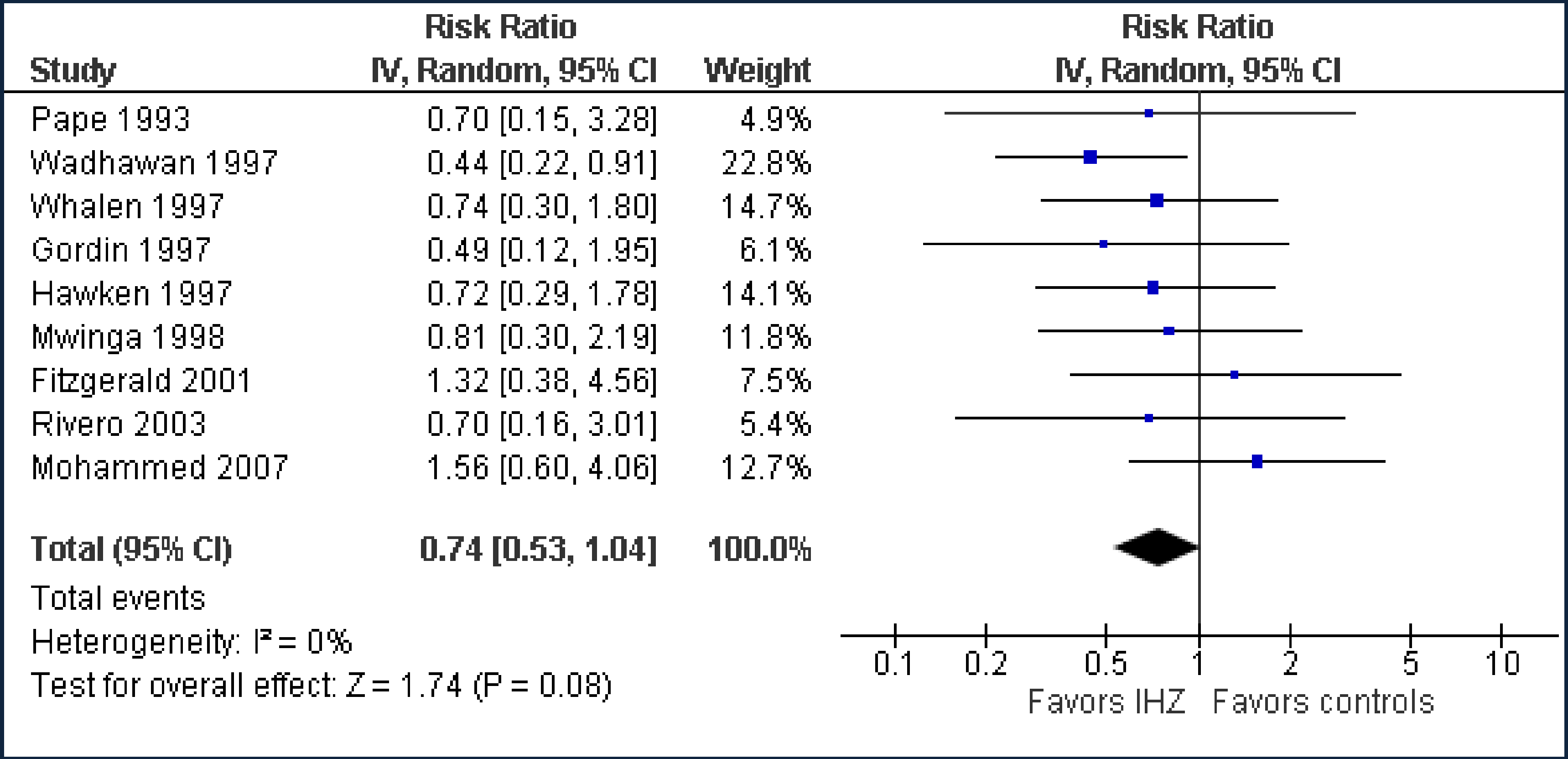
IHZ: antibiotic
PPD+/-: test for active/past TB

Thorlund K, Aranka A, Mills E. *Clin Epi* (in press)

2010 -

Alternative applications:

9 trials, 2911 patients, event rate 2-12%



Study	Risk Ratio IV, Random, 95% CI	Weight
Pape 1993	0.70 [0.15, 3.28]	4.9%
Wadhawan 1997	0.44 [0.22, 0.91]	22.8%
Whalen 1997	0.74 [0.30, 1.80]	14.7%
Gordin 1997	0.49 [0.12, 1.95]	6.1%
Hawken 1997	0.72 [0.29, 1.78]	14.1%
Mwinga 1998	0.81 [0.30, 2.19]	11.8%
Fitzgerald 2001	1.32 [0.38, 4.56]	7.5%
Rivero 2003	0.70 [0.16, 3.01]	5.4%
Mohammed 2007	1.56 [0.60, 4.06]	12.7%
Total (95% CI)	**0.74 [0.53, 1.04]**	**100.0%**
Total events		

Heterogeneity: $I^2 = 0\%$

Test for overall effect: Z = 1.74 (P = 0.08)

2010 -

Alternative applications:
9 trials, 2911 patients, event rate 2-12%

Relative risk 0.74 (95% CI, 0.53 to 1.04)
OIS=10500 patients
(α=5%, β=80%, PC=5%, RRR=25%)

The answer is not in!

2010 -

Need an additional 7500 patients (too much?)

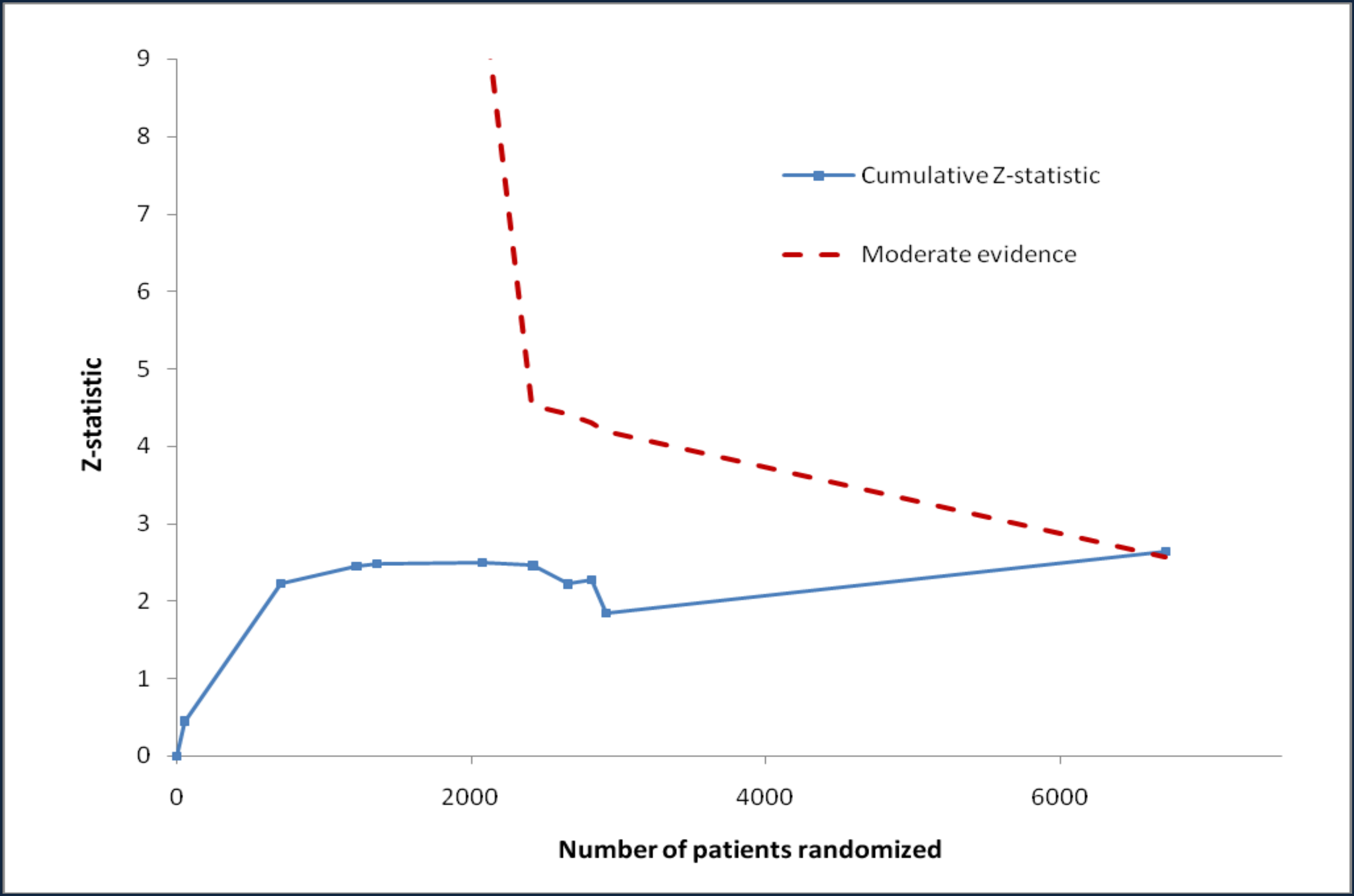
Applied monitoring boundaries and played around with numbers to approximate:

'How large must a new trial be for the MA to cross the monitoring boundaries?'

(assuming RRR=25% and PC=5%)

2010 -

Topping up the sample size (3800 pts)



Summary

Type I error due to multiplicity

- Simulation studies: 2- to 7-fold

- Empirical studies 4 to 5 fold

Summary

Treatment effect estimates

- 'Early' = unrealiable

- 'Early' significant = overestimate

- Definition of 'early' differs across medical areas

Summary

Methodology

- OIS and monitoring boundaries
- adjustment/penalization of Z-statistics and p-values

Summary

OIS and monitoring boundaries have some support from studies and applications

Other methods have never been applied to real MA data

Part IV:

Questions and debate

Questions?

Research efforts so far have focused on superiority testing for binary outcome meta-analysis...

Questions?

- What more is needed to seal the deal?
- What about inferiority testing?
- What about other types of data?
  - continuous (HRQL)
  - time-to-event/survival
- Other…

Questions?

Clinicians are notorious for relying on thresholds, statisticians for making things too complex.

For decision-making, can we meet in the middle?

Questions?

How do we (largely) avoid misuse of proposed methodologies?

Bring it on...