Understanding & Using Systematic Reviews

W. Scott Richardson, M.D.

AU/UGA Partnership, Athens

Three Owl Learning Institute

Conflicts of Interest

- I have no financial ties with industry that pose a conflict of interest regarding the content of this presentation
- I will not be discussing “off label” uses of any medications or devices
- Image copyrights are retained by their original creators, publishers, etc.

Session Aims

Structure and function of systematic reviews of treatment trials

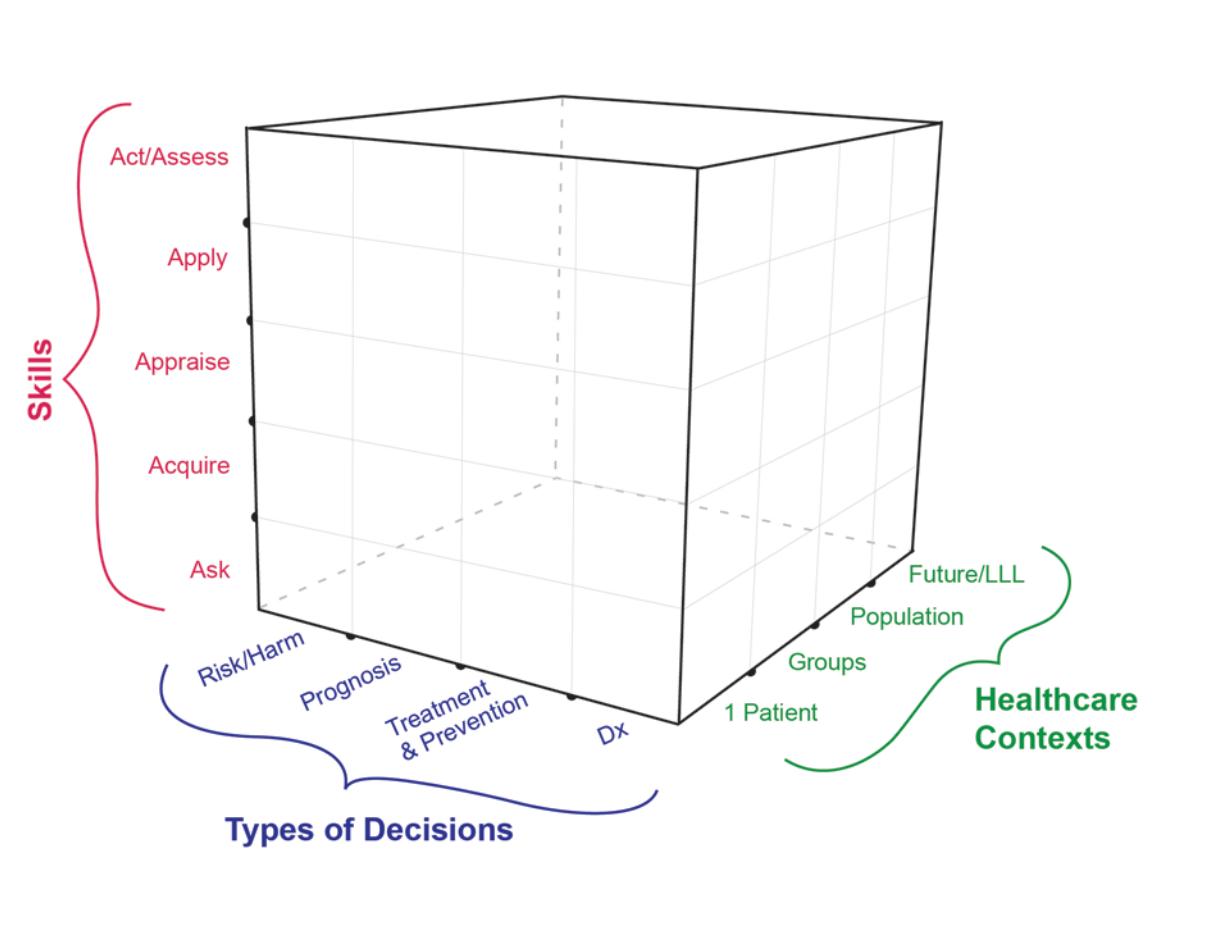
Appraise SR methods

Understand SR results

EBM: Why Bother

- We can’t make well-informed decisions without information
- Not all information is created equal
- Misinformation can be worse than no information
- Better information → better informed decisions → better outcomes

EBM Curriculum Overview



Decisions affecting treatment

- Whether condition warrants treatment
- What treatment options are available, affordable, acceptable?
- For each, what is balance of benefits versus harms and costs?
- What are this patient's values and preferences?
- How can we make a wise decision and provide kind and careful treatment?

Individual randomized trials of treatment …

- Each trial is one experiment, one new chance to get closer to the ‘truth’
- One trial ~ one race
- Often, more than one trial is done
- Will all trial results agree (even by chance)?

As trials accumulate …

- Seldom is one trial definitive (“One ring to rule them all …”)
- In science, as experiments accrue, knowledge is built cumulatively
- Is there a scientific way to combine results of individual trials?
- Yes! Systematic reviews (we’ll abbreviate “SRs”)

‘Narrative’ vs. ‘Systematic’

- Address disorder as a whole – overview
- Or, tell a ‘story’
- Variety of questions
- No methods section
- No formal pooling
- Thus, may be cumulative but not comprehensive

- Address focused question (e.g. effect of therapy, accuracy of diagnostic test)
- Methods section
- Formal pooling, when appropriate
- Thus, cumulative *and* comprehensive

SR Methods

- Formulate questions
- Define eligibility criteria for study inclusion
- Develop *a priori* hypotheses to explain heterogeneity
- Conduct search
- Screen titles, abstracts for inclusion, exclusion
- Review full text
- Assess the risk of bias
- Abstract data
- When meta-analysis is performed:
  - Summary estimates, confidence intervals
  - Explain heterogeneity
  - Rate confidence in estimates of effect
- Report results
- Update review as needed

‘PRISMA’

- ‘Preferred Reporting Items for Systematic reviews and Meta-Analyses’
- Incorporates evolutionary advances
- Specifies 27 item checklist for reporting, e.g. standardizes figures, etc.
- Since 2009, has replaced ‘QUOROM’, has been adopted by many journals
- *Ann Intern Med 2009; 151: 264 – 269*

Finding SRs

- Cochrane Library
  - CDSR – Cochrane Database of Systematic Reviews
  - DARE – Database of Abstracts of Reviews of Effects
- PubMed
  - Publication types
  - Clinical queries
- Work with your team to find SRs

Critical Appraisal of SRs

Credibility:

- Sensible question?
- Exhaustive search?
- Selection, assessments reproducible?
- Present results ready for application?
- Address confidence in estimates of effect?

Confidence in Estimates:

- Risk of bias?
- Consistent across studies?
- Effect: RR, OR, WMD
- Precision: 95% CI
- Apply to my patient?
- Reporting bias?
- Reasons to increase confidence rating?

‘Risk of bias’

The Cochrane Collaboration’s tool for assessing risk of bias in randomised trials

OPEN ACCESS

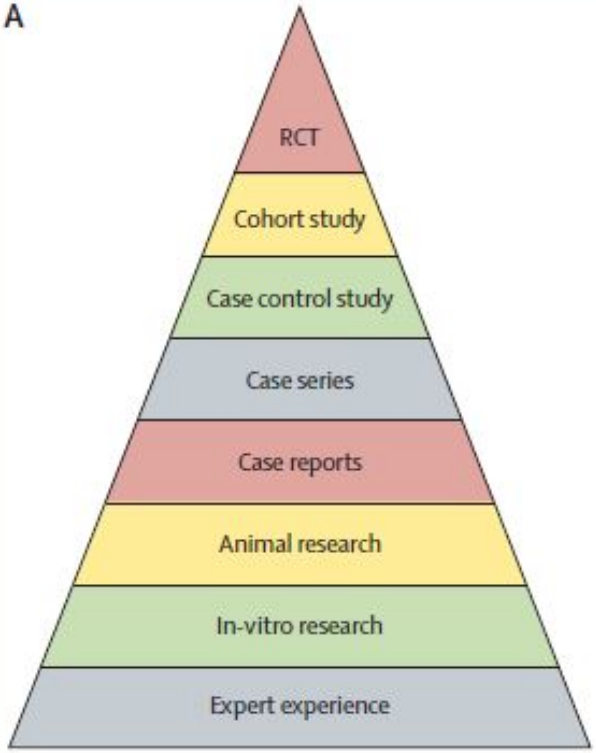
Flaws in the design, conduct, analysis, and reporting of randomised trials can cause the effect of an intervention to be underestimated or overestimated. The Cochrane Collaboration’s tool for assessing risk of bias aims to make the process clearer and more accurate

Julian P T Higgins *senior statistician*[1], Douglas G Altman *director*[2], Peter C Gøtzsche *director*[3], Peter Jüni *head of division*[4], David Moher *senior scientist*[5 6], Andrew D Oxman *senior researcher*[7], Jelena Savović *postdoctoral fellow*[8], Kenneth F Schulz *vice president*[9], Laura Weeks *research associate*[5], Jonathan A C Sterne *professor of medical statistics and epidemiology*[8], Cochrane Bias Methods Group, Cochrane Statistical Methods Group

- Moves away from dichotomous “yes/no” to explicit rating of risk of bias
- At both study-level and outcome-level
- *BMJ 2011; 343: d5928 doi*

Pyramid vs GRADE

A

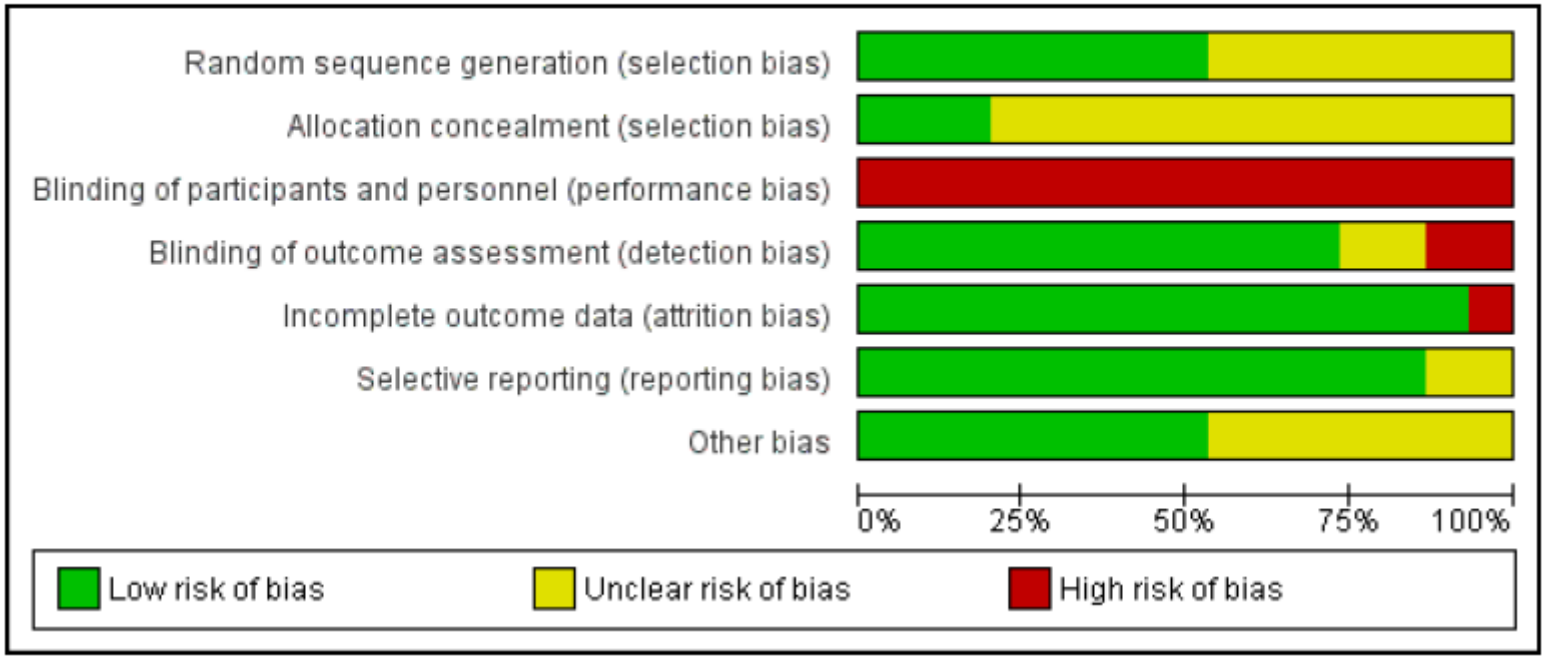


B

Quality of evidence	Study design	Lower quality if*	Higher quality if†
High	Randomised trial	Study limitations - 1 serious - 2 very serious	Large effect + 1 large + 2 very large
Moderate		Inconsistency - 1 serious - 2 very serious	Dose response + 1 evidence of a gradient
Low	Observational study	Indirectness - 1 serious - 2 very serious	All plausible confounders + Would reduce a demonstrated effect or + Would suggest a spurious effect when results show no effect
Very low		Imprecision - 1 serious - 2 very serious	
		Publication bias - 1 likely - 2 very likely	

Risk of bias graphs

Figure 1. Risk of bias graph: review authors' judgements about each risk of bias item presented as percentages across all included trials.

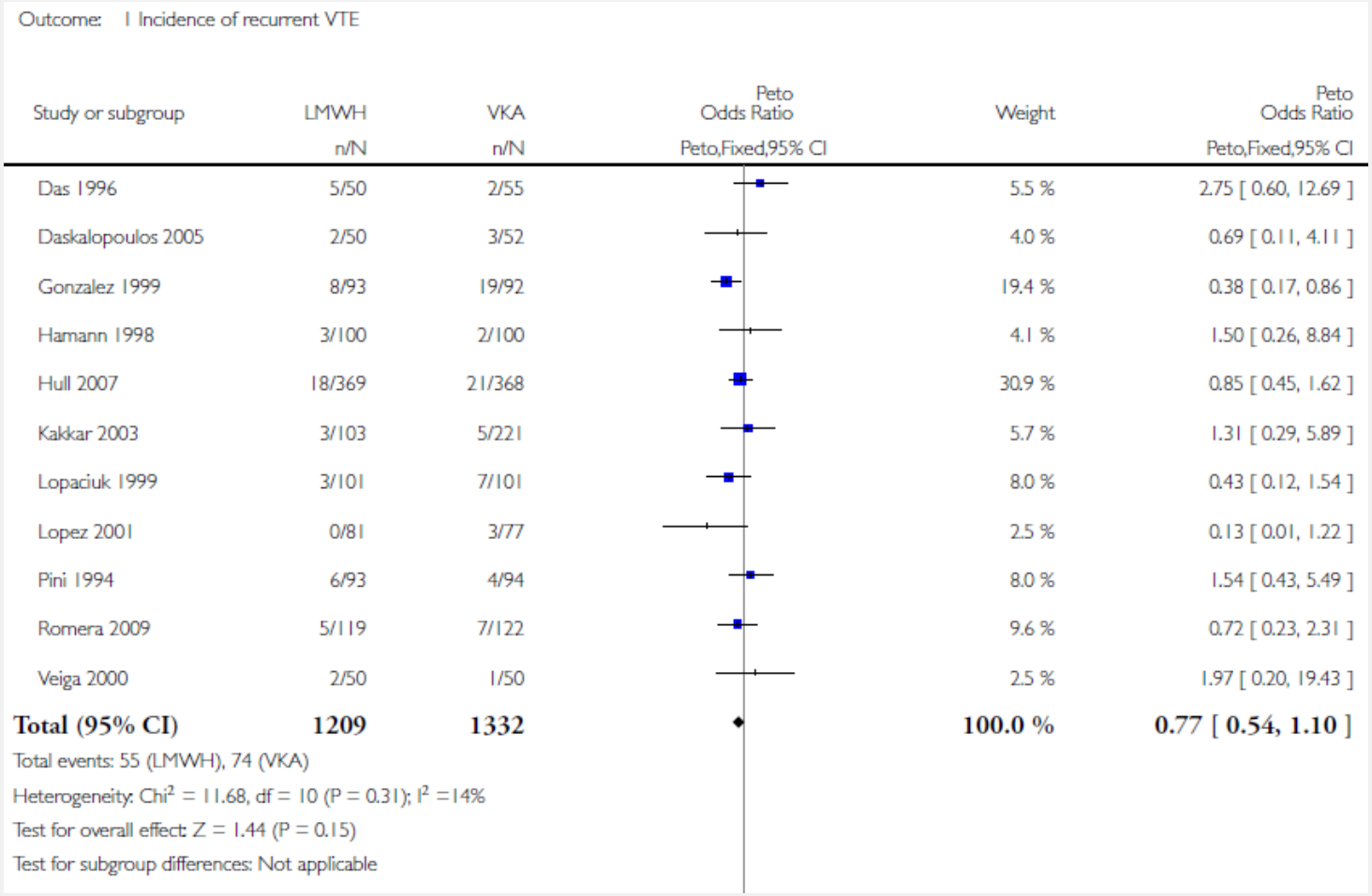


Reporting Biases

- Selective reporting of studies
  - Delayed (or never)
  - Location, language
- Selective reporting of outcomes, times
- Selective reporting of analyses
- *UG 3/e Box 23-2*

- Empirical evidence
- Distort the ‘body of evidence’ in the literature
- Can lead to wrong conclusions about the benefits and harms

Forest Plot – a



Outcome: 1 Incidence of recurrent VTE

Study or subgroup	LMWH n/N	VKA n/N	Peto Odds Ratio Peto,Fixed,95% CI	Weight	Peto Odds Ratio Peto,Fixed,95% CI
Das 1996	5/50	2/55		5.5 %	2.75 [0.60, 12.69]
Daskalopoulos 2005	2/50	3/52		4.0 %	0.69 [0.11, 4.11]
Gonzalez 1999	8/93	19/92		19.4 %	0.38 [0.17, 0.86]
Hamann 1998	3/100	2/100		4.1 %	1.50 [0.26, 8.84]
Hull 2007	18/369	21/368		30.9 %	0.85 [0.45, 1.62]
Kakkar 2003	3/103	5/221		5.7 %	1.31 [0.29, 5.89]
Lopaciuk 1999	3/101	7/101		8.0 %	0.43 [0.12, 1.54]
Lopez 2001	0/81	3/77		2.5 %	0.13 [0.01, 1.22]
Pini 1994	6/93	4/94		8.0 %	1.54 [0.43, 5.49]
Romera 2009	5/119	7/122		9.6 %	0.72 [0.23, 2.31]
Veiga 2000	2/50	1/50		2.5 %	1.97 [0.20, 19.43]
Total (95% CI)	**1209**	**1332**		**100.0 %**	**0.77 [0.54, 1.10]**

Total events: 55 (LMWH), 74 (VKA)

Heterogeneity: $Chi^2 = 11.68$, df = 10 (P = 0.31); $I^2 = 14\%$

Test for overall effect: Z = 1.44 (P = 0.15)

Test for subgroup differences: Not applicable

Forest plot – b

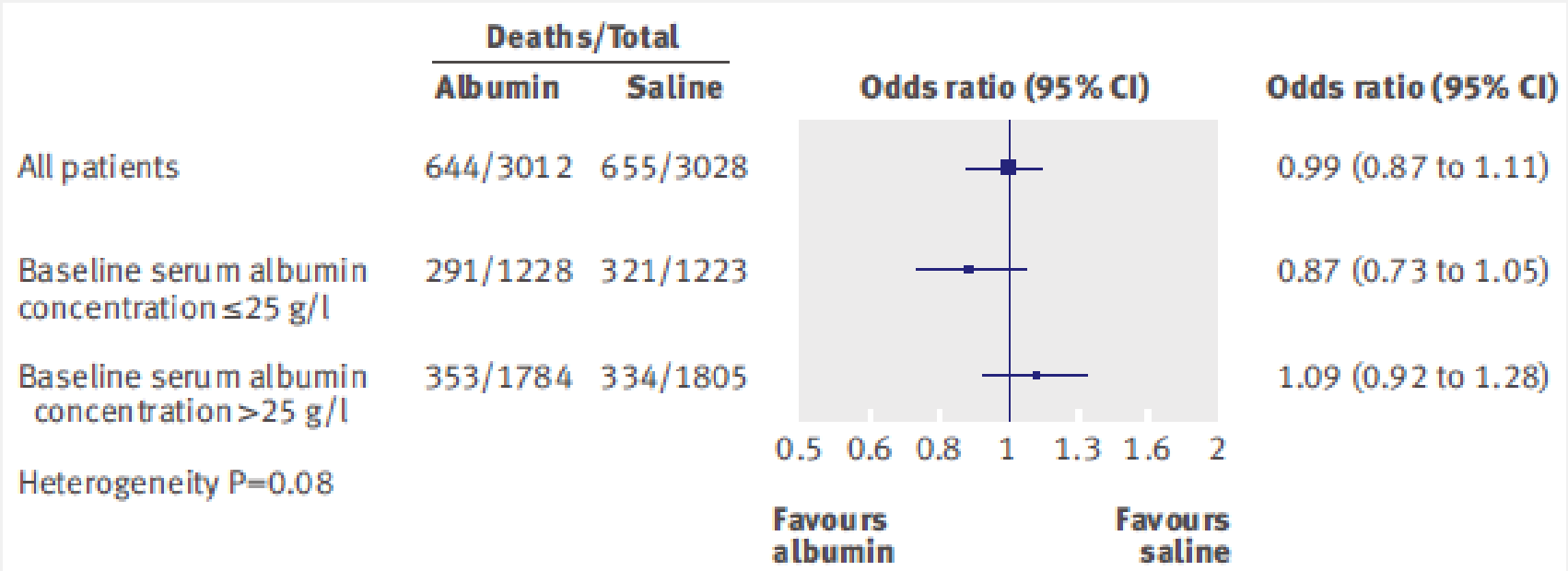
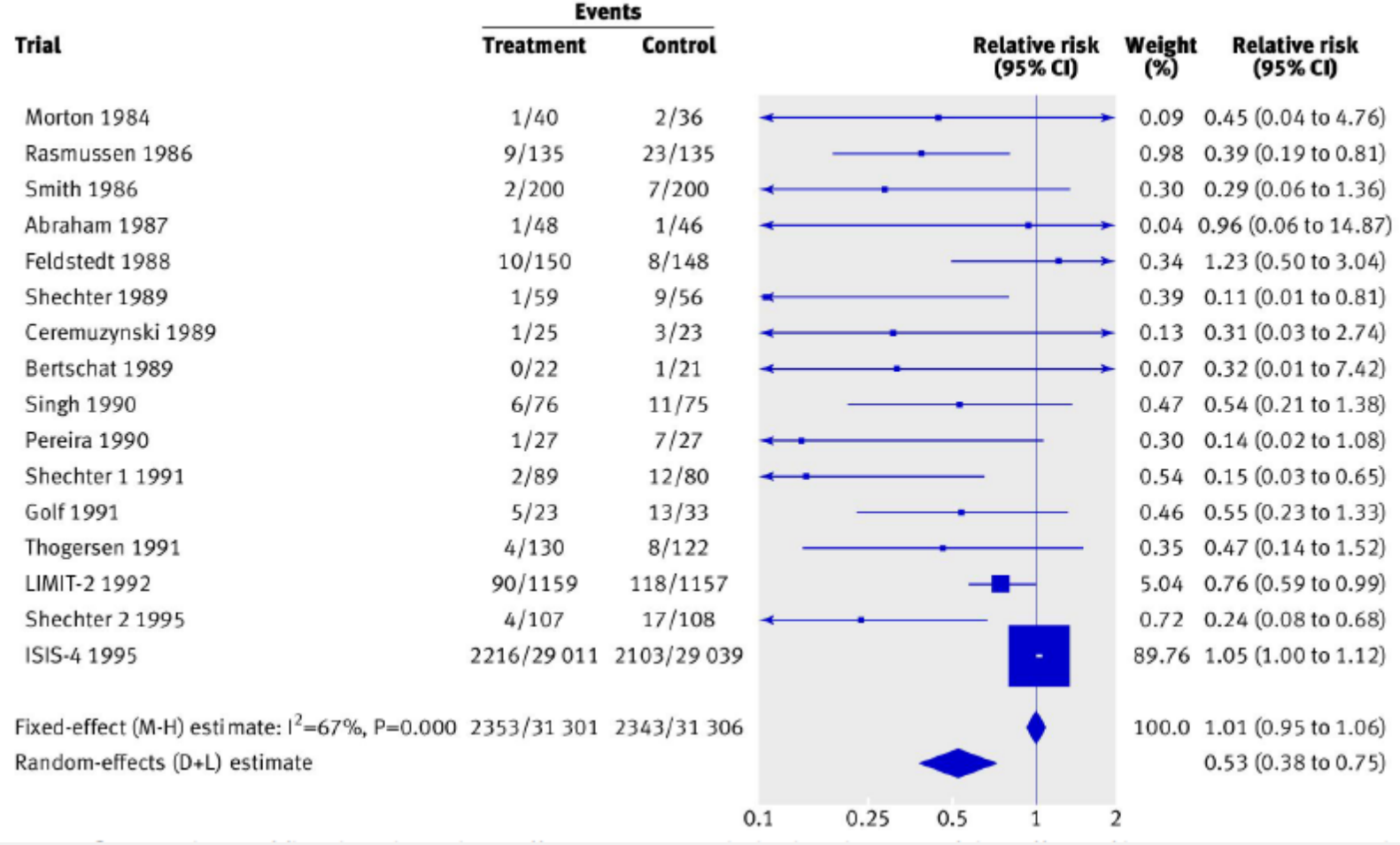


Fig 5 | Forest plot from study comparing resuscitation with albumin or saline in intensive care showing unadjusted odds ratio of death stratified by baseline albumin concentration[18]

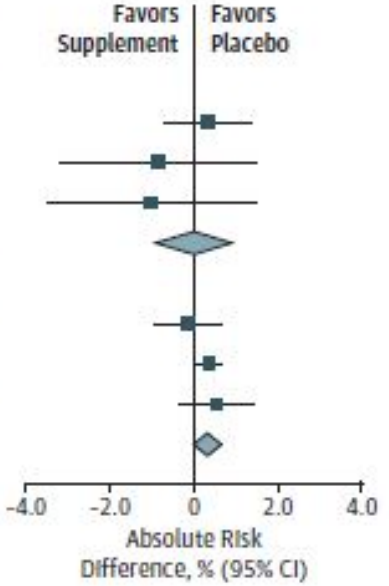
Forest plot – c



Forest plot – d

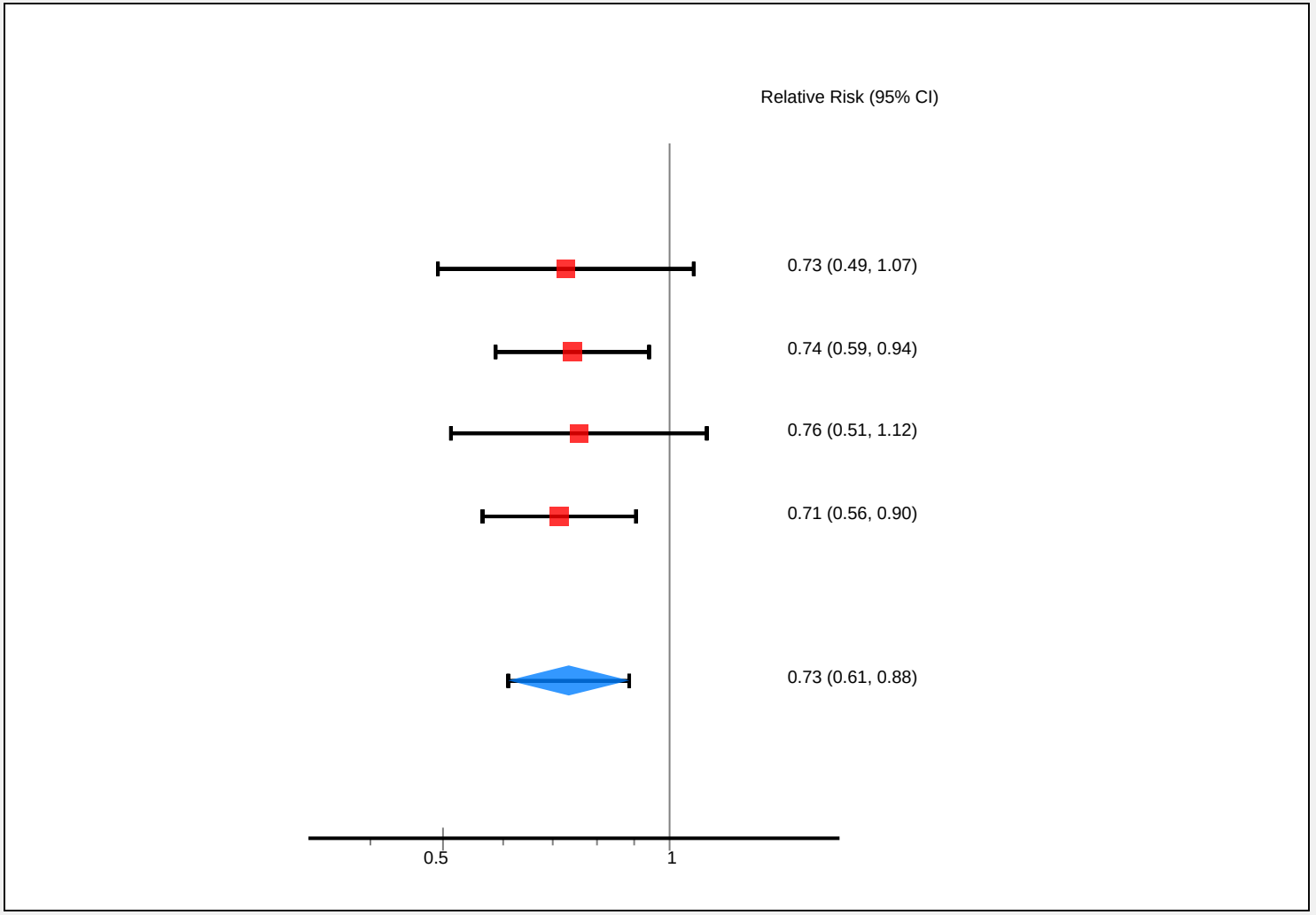
Figure 4. Comparison of Incident Kidney Stones in Randomized Trials Comparing Calcium or Both Vitamin D and Calcium With Placebo

Source	No. of Patients With Kidney Stones/Total No. (%)		Risk Ratio (95% CI)	Absolute Risk Difference, % (95% CI)
	Supplement	Placebo		
Calcium				
Lappe et al,[21] 2007	3/445 (0.7)	1/288 (0.4)	1.94 (0.20-18.57)	0.33 (-0.69 to 1.35)
Riggs et al,[20] 1998	0/119 (0)	1/117 (0.9)	0.33 (0.01-7.97)	-0.85 (-3.18 to 1.47)
Reid et al,[22] 2008	0/191 (0)	1/99 (1.0)	0.17 (0.01-4.22)	-1.01 (-3.50 to 1.48)
Subtotal: $I^2 = 0.0\%$; $P = .42$			0.68 (0.14-3.36)	0.00 (-0.88 to 0.87)
Vitamin D with calcium				
Lappe et al,[21] 2007	1/446 (0.2)	1/288 (0.4)	0.65 (0.04-10.28)	-0.12 (-0.93 to 0.69)
WHI,[27] 2011	449/18 176 (2.5)	381/18 106 (2.1)	1.17 (1.03-1.34)	0.37 (0.06 to 0.67)
Lappe et al,[25] 2017	16/1102 (1.5)	10/1095 (0.9)	1.59 (0.72-3.49)	0.54 (-0.36 to 1.44)
Subtotal: $I^2 = 0.0\%$; $P = .69$			1.18 (1.04-1.35)	0.88 (0.05 to 0.80)

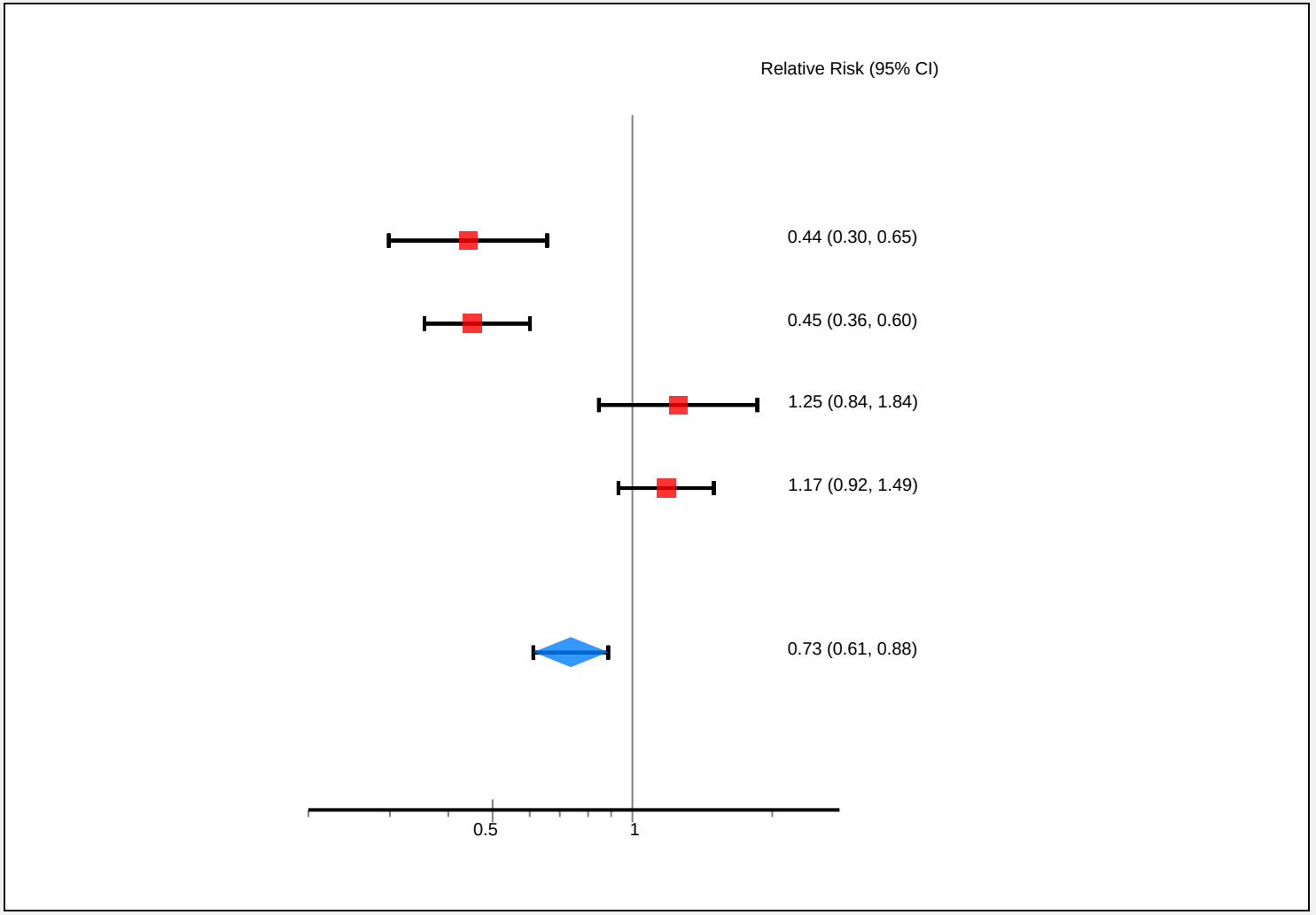


WHI indicates Women's Health Initiative.

Are you happy pooling?



Are you happy pooling?



What criteria were you using?

- similarity of point estimates
  - less similar, less happy
- overlap of confidence intervals
  - less overlap, less happy

Heterogeneity

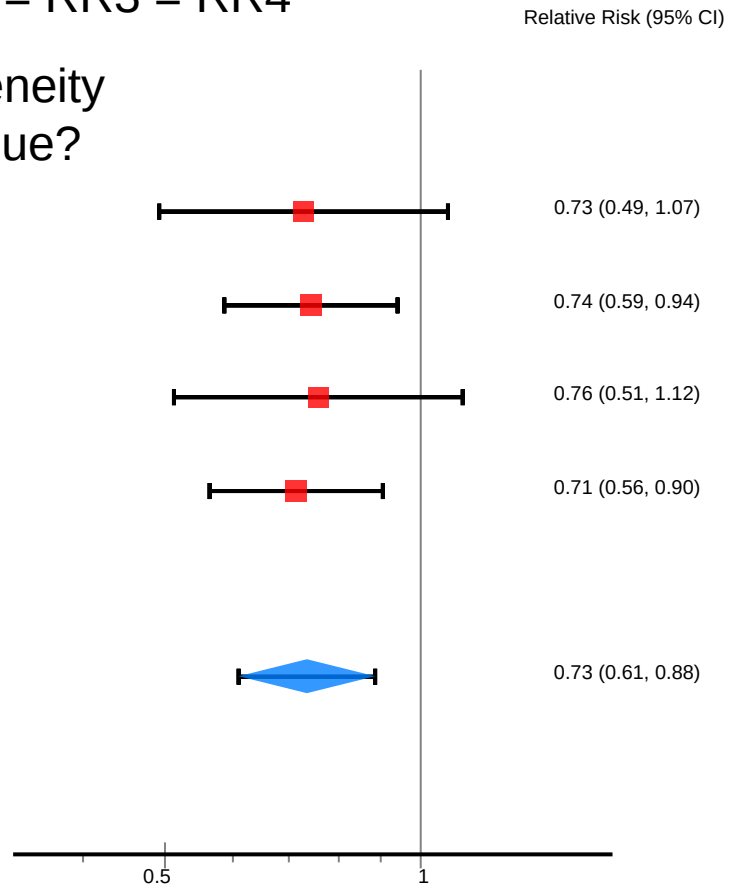
- Humans vary, e.g. in risk of poor outcomes from disease, in response to therapy, and in vulnerability to adverse effects
- Heterogeneity represents this variation in results
- Affects certainty about estimates of effect

- Identified by:
  - Visual inspection
  - Chi^2: “yes” or “no”
  - I^2: 0 to 100%
- Explored by:
  - Patients
  - Interventions
  - Comparisons
  - Outcomes
  - Methods, Systems, +

Homogenous

Ho: RR1 = RR2 = RR3 = RR4

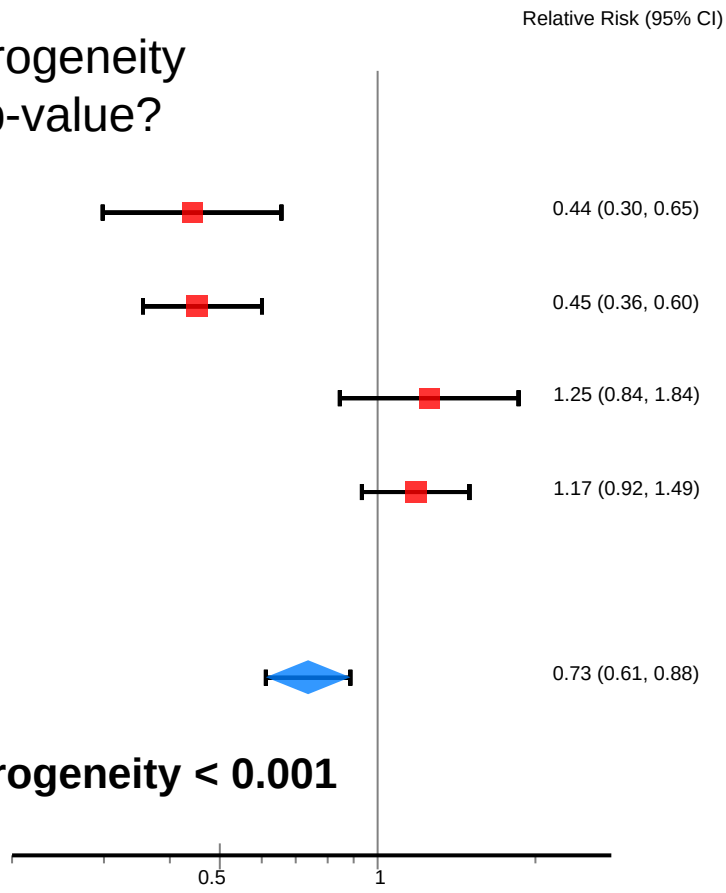
test for heterogeneity
what is the p-value?



p=0.99 for heterogeneity

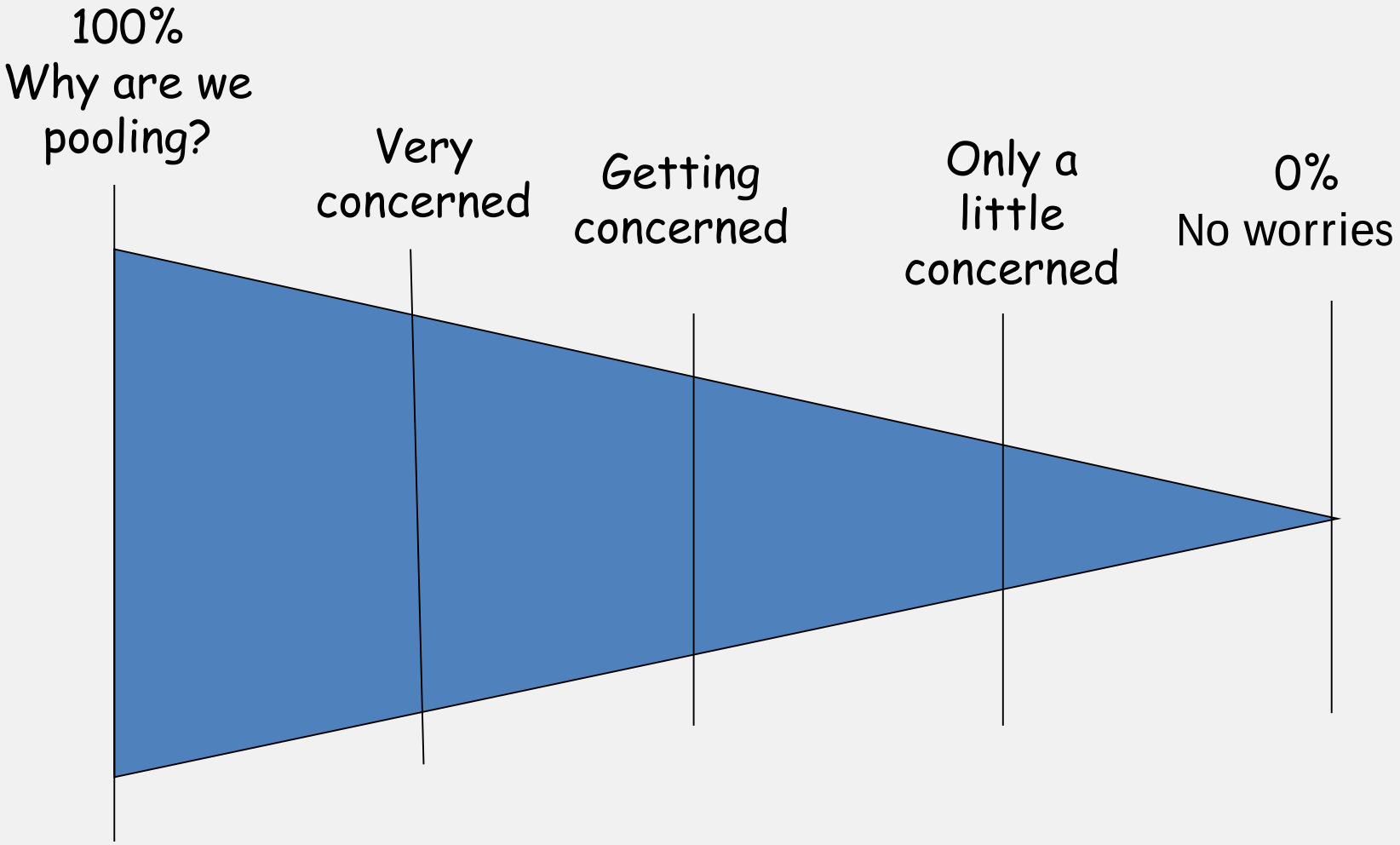
Heterogeneous

test for heterogeneity
what is the p-value?



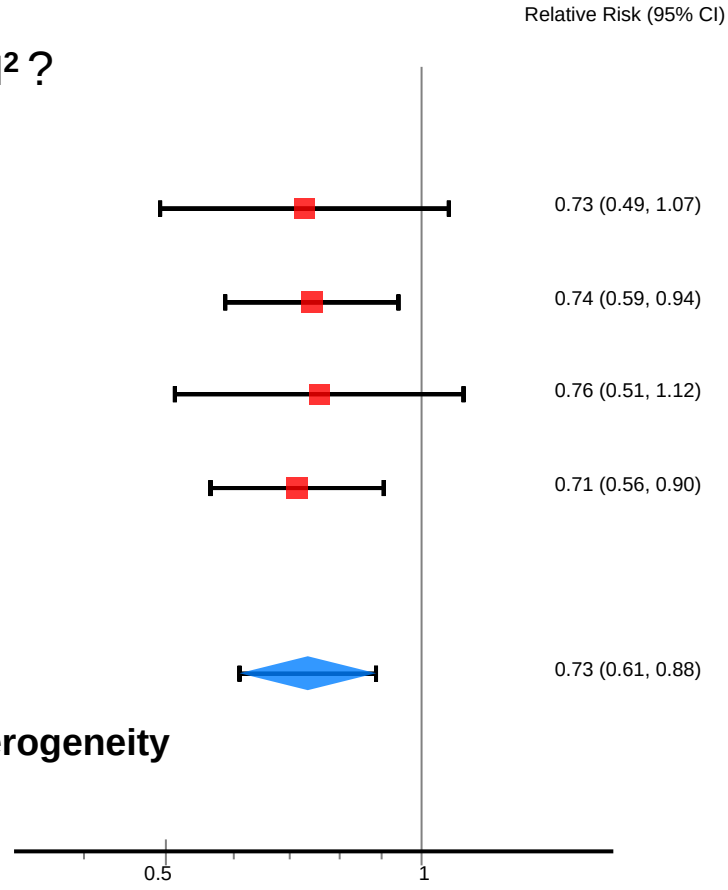
p-value for heterogeneity < 0.001

I^2 Interpretation



Homogenous

What is the I^2?

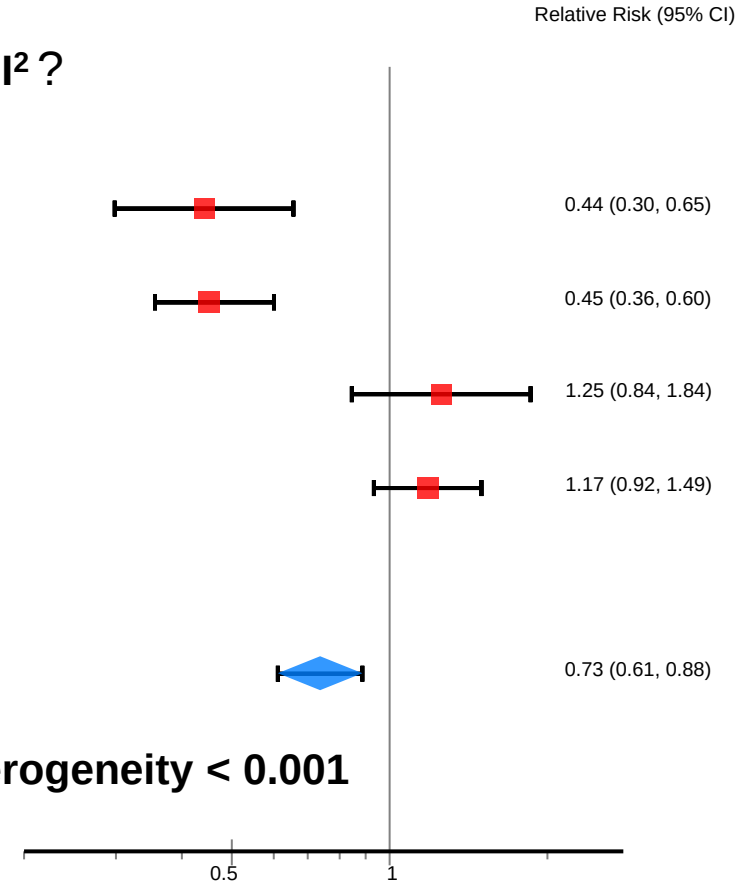


p=0.99 for heterogeneity

I^2=0%

Heterogeneous

What is the I^2?



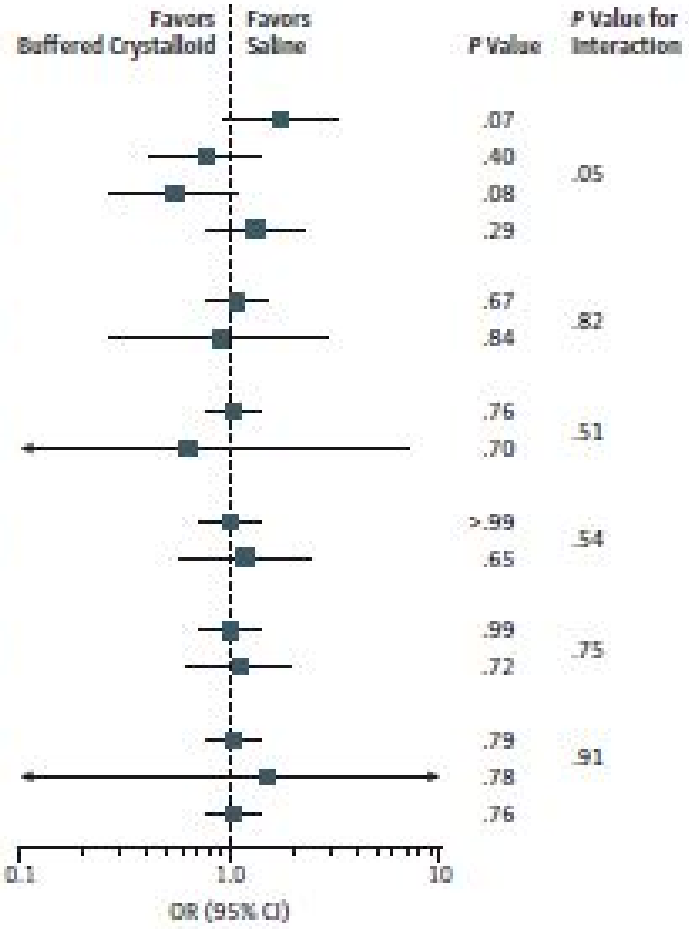
p-value for heterogeneity < 0.001

I^2=89%

Why Not Use Subgroups?

Figure 3. Risk of Acute Kidney Injury by Subgroup for Patients Admitted to the Intensive Care Unit Receiving Buffered Crystalloid vs Saline Fluid Therapy

Subgroup	No. of Individuals		No. of Events (%)		OR (95% CI)	P Value	P Value for Interaction
	Buffered Crystalloid	Saline	Buffered Crystalloid	Saline			
Location							
ICU 1	190	204	29 (15.3)	19 (9.3)	1.75 (0.95-3.25)	.07	.05
ICU 2	387	378	20 (5.2)	25 (6.6)	0.77 (0.42-1.41)	.40	
ICU 3	200	169	15 (7.5)	22 (13.0)	0.54 (0.27-1.08)	.08	
ICU 4	290	274	38 (13.1)	28 (10.2)	1.32 (0.79-2.23)	.29	
Sepsis							
No	1032	983	95 (9.2)	85 (8.6)	1.07 (0.79-1.46)	.67	.82
Yes	35	42	7 (20.0)	9 (21.4)	0.89 (0.27-2.86)	.84	
Trauma							
No	1034	977	101 (9.8)	91 (9.3)	1.05 (0.78-1.41)	.76	.51
Yes	33	48	1 (3.0)	3 (6.3)	0.63 (0.06-7.07)	.70	
APACHE II							
<25	985	946	77 (7.8)	74 (7.8)	1.00 (0.72-1.39)	>.99	.54
≥25	82	79	25 (30.5)	20 (25.3)	1.18 (0.57-2.41)	.65	
Cardiac surgery							
No	597	547	76 (12.7)	70 (12.8)	1.00 (0.71-1.42)	.99	.75
Yes	470	478	26 (5.5)	24 (5.0)	1.11 (0.63-1.96)	.72	
TBI							
No	1045	1000	101 (9.7)	93 (9.3)	1.04 (0.77-1.40)	.79	.91
Yes	22	25	1 (4.5)	1 (4.0)	1.50 (0.09-26.00)	.78	
Overall	1067	1025	102 (9.6)	94 (9.2)	1.06 (0.78-1.41)	.76	



Sources of error …

- Apophenia: tendency to see patterns in ‘noise’ or randomness
- While adaptive in some situations, can lead us astray when analyzing study data
- Play of chance vs. distorted signal vs. true signal

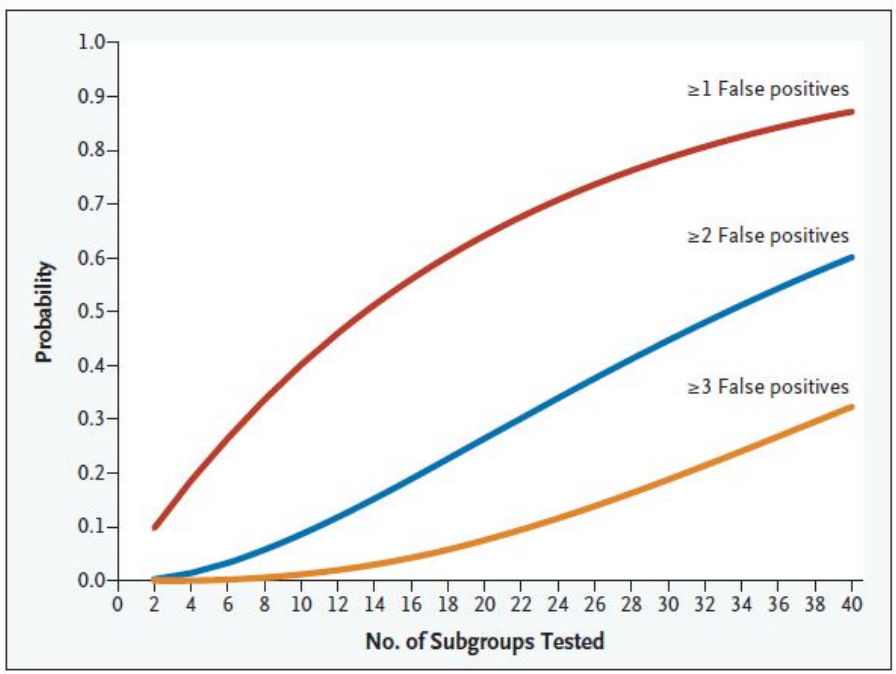
Subgroups: Inform? Mislead?

- Subgroups may be informative for clinical decisions (in present) and raise hypothesis for further research (in the future)
- Subgroups may also mislead, due to several possible explanations for differences found

Possible explanations of difference in subgroups:

- Hypothesized difference
- Chance
- Other patient difference
- Different co-interventions
- Different outcome measures
- Different risk of bias

Multiple looks; imbalance



Probability That Multiple Subgroup Analyses Will Yield at Least One (Red), Two (Blue), or Three (Yellow) False Positive Results.

- If no difference exists, multiple comparisons risks finding 'false positive' results
- "The more you look, the more you find."
- Using subgroups un-does the prognostic balance from random allocation

Credibility of subgroup analyses

Criteria to assess the credibility of subgroup analyses

Design

- Is the subgroup variable a characteristic measured at baseline or after randomisation?*
- Is the effect suggested by comparisons within rather than between studies?
- Was the hypothesis specified a priori?
- Was the direction of the subgroup effect specified a priori*
- Was the subgroup effect one of a small number of hypothesised effects tested?

Analysis

- Does the interaction test suggest a low likelihood that chance explains the apparent subgroup effect?
- Is the significant subgroup effect independent?*

Context

- Is the size of the subgroup effect large?
- Is the interaction consistent across studies?
- Is the interaction consistent across closely related outcomes within the study?*
- Is there indirect evidence that supports the hypothesised interaction (biological rationale)?

*New criteria.

- Ten criteria in 3 main areas
  - Study design
  - Data analysis
  - Study context
- Greater confidence if most or all are met
- Lower confidence if few or none are met
- Work through with teams during appraisal

SR's of Other Study Types

Annals of Internal Medicine

REVIEW

Screening for Occult Cancer in Patients With Unprovoked Venous Thromboembolism

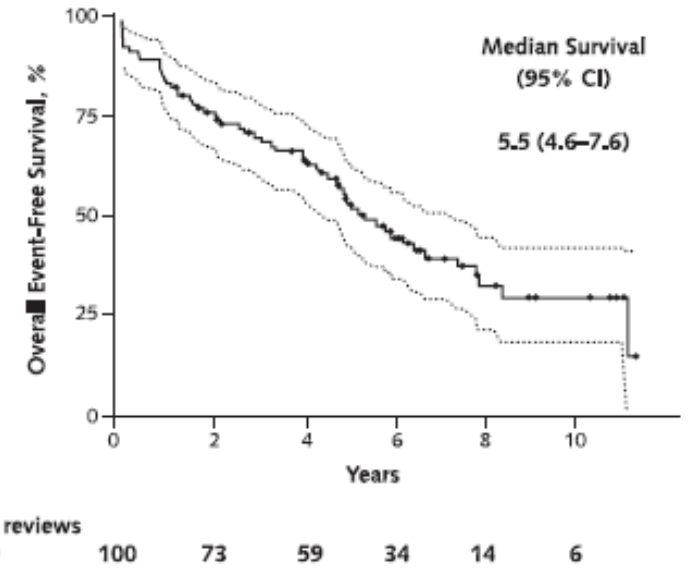
A Systematic Review and Meta-analysis of Individual Patient Data

Nick van Es, MD; Grégoire Le Gal, MD, PhD; Hans-Martin Otten, MD, PhD; Philippe Robin, MD, PhD; Andrea Piccioli, MD, PhD; Ramón Lecumberri, MD, PhD; Luis Jara-Palomares, MD; Piotr Religa, MD, PhD; Virginie Rieu, MD; Matthew Rondina, MD; Mariëlle M. Beckers, MD, PhD; Paolo Prandoni, MD, PhD; Pierre-Yves Salaun, MD, PhD; Marcello Di Nisio, MD, PhD; Patrick M. Bossuyt, PhD; Harry R. Büller, MD, PhD; and Marc Carrier, MD

- Diagnostic test accuracy studies
- Cohort studies of prognosis
- Disease probability for differential diagnosis*
- Other observational studies

How quickly do systematic reviews go out of date?

Figure 2. **Overall survival time (95% CI) free of signals for updating.**



The immediate decrease in survival at time zero reflects the 7 systematic reviews for which signals for updating had already occurred at the time of publication. The low number of reviews at risk after 10 years reflects the fact that the sample spanned 1995 to 2005 and censoring occurred on 1 September 2006. Thus, only reviews published before September 1996 and having no signals for updating could have more than 10 years of observation.

- Survival analysis
- 100 systematic reviews, 1995 – 2005
- Searched for ‘update signals’ (i.e. new trial evidence)

- *Ann Intern Med 2007*

Taking SRs home …

- When well-made and current, SRs synthesize the body of research evidence that can guide important decisions
- SRs have limits, yet we should start with them: ‘how well does this work?’
- We can (and must!) appraise SRs for risk of bias, estimates of effect, and confidence in these estimates

Questions?