How Much Can We Trust Meta-Analyses: The Good, the Bad, and the Ugly (with case examples)

> Tullio Palmerini University of Bologna Italy













The pyramid of evidence based medicine



EBM Pyramid and EBM Page Generator, Copyright 2006-2014 Trustees of Dartmouth College and Yale University. All Rights Reserved. Produced by Jan Glover, David Izzo, Karen Odato and Lei Wang.

Sackett DL, Straus SE, Richardson WS, et al. *Evidence-based medicine: how to practice and teach EBM*. 2nd ed. Edinburgh: Churchill Livingstone, 2000.



		CLASS I Benefit >>> Risk Procedure/Treatment SHOULD be performed/ administered	CLASS IIa Benefit >> Risk Additional studies with locused objectives needed IT IS REASONABLE to per- form procedure/administer treatment	CLASS IIb Benefit ≥ Risk Additional studies with broad objectives needed; additional registry data would be helpful Procedure/Treatment MAY BE CONSIDERED	CLASS III No Benefit Procedum/ Test Treatment COR III: No Senefit Not Helpful No Prives Benefit COR III: Harm Excess Cost w/s Benefit Harmful
SIMALE OF LENIAMIT (FRECISION) OF INCALMENT EFFECT	LEVEL A Multiple populations evaluated* Data derived from multiple randomized clinical triaks or meta-analyses	 Recommendation that procedure or treatment is useful/effective Sufficient evidence from multiple randomized trials or meta-analyses 	 Recommendation in favor of treatment or procedure being useful/effective Some conflicting evidence from multiple randomized trials or meta-analyses 	 Recommendation's usefulness/efficacy less well established Greater conflicting evidence from multiple randomized trials or meta-analyses 	 Recommendation that procedure or treatment is not useful/effective and may be harmful Sufficient evidence from multiple randomized trials or meta-analyses
	LEVEL B Limited populations evaluated* Data derived from a single randomized trial or nonrandomized studies	 Recommendation that procedure or treatment is useful/effective Evidence from single randomized trial or nonrandomized studies 	 Recommendation in favor of treatment or procedure being useful/effective Some conflicting evidence from single randomized trial or nonrandomized studies 	 Recommendation's usefulness/efficacy less well established Greater conflicting evidence from single randomized trial or nonrandomized studies 	 Recommendation that procedure or treatment is not useful/effective and may be harmful Evidence from single randomized trial or nonrandomized studies
	LEVEL C Very limited populations evaluated* Only consensus opinion of experts, case studies, or standard of care	 Recommendation that procedure or treatment is useful/effective Only expert opinion, case studies, or standard of care 	 Recommendation in favor of treatment or procedure being useful/effective Only diverging expert opinion, case studies, or standard of care 	 Recommendation's usefulness/efficacy less well established Only diverging expert opinion, case studies, or standard of care 	 Recommendation that procedure or treatment is not useful/effective and may be harmful Only expert opinion, case studies, or standard of care

SIZE OF TREATMENT EFFECT



TCTAP2019

-CYRI

What is an indirect comparison ?







Problem formulation according to the PICO approach

- <u>Population of interest</u> eg elderly male >2 weeks after myocardial infarction)
- <u>Intervention (or exposure)</u> eg intracoronary infusion of progenitor blood cells
- <u>Comparison</u> eg patients treated with progenitor cells vs standard therapy
- <u>Outcome(s)</u> eg change in echocardiographic left ventricular ejection fraction from discharge to 6-month control





Study selection: Systematic review

 1st - screening of titles and abstracts (BioMedCentral, clinicaltrials.gov, EMBASE, LILACS, and PubMed, Conference proceedings, Website)

 2nd – potentially pertinent citations are then retrieved as full reports and appraised according to prespecified and explicit inclusion/exclusion criteria

 3rd – studies fullfilling both inclusion and exclusion criteria, are then included in the systematic review

Stent Thrombosis With Everolimus-Eluting Stents Meta-Analysis of Comparative Randomized Controlled Trials

Tullio Palmerini, MD; Ajay J. Kirtane, MD, SM; Patrick W. Serruys, MD; Pieter C. Smits, MD; Elvin Kedhi, MD; Dean Kereiakes, MD; Diego Sangiorgi, MStat; Letizia Bacchi Reggiani, MSc; Christoph Kaiser, MD; Hyo-Soo Kim, MD; Antoinette De Waha, MD; Flavio Ribichini, MD; Gregg W. Stone, MD

Potentially re	levant articles: ,632		
Review of titles		Excluded: Not a comparison between DES: Post-hoc, subgroup, follow up, or	n=1,595 n=1,455
		pooled analyses of included trials:	n=140
Articles requir n=	ing full review: 37		
Full text review		Excluded:	n=26
		Not a RCT: RCT with same polymer and drug:	n=23 n=1
		RCT with no events:	n=2
Articles meeting criteria: n=11			



Palmerini et al. Circ Cv Int 2012



Statistical pooling: fixed versus random effect method

Fixed effect method

Random effect method



One distribution of true effect Each trial is measuring the true effect Natural random sampling variation

Acknokledges the occurrence of variation of true effects among studies Mean effect of studies with their own distribution

Biondi-Zoccai, Network Meta-Analysis: Evidence Synthesis with Mixed Treatment Comparison 2014

Data synthesis: effect size for dichotomus events

- Relative risks (RR) are defined as the ratio of incidence rates, and are thus used for dichotomic variables).
- Odds ratios (OR) are defined as the ratio of the odds (P/[1-P]) and also used for dichotomic variables.
- Hazard ratio (HR) adjust for different duration of follow up and allow for data censoring.
- The absolute risk difference (RD) is the difference between the incidence of events in the experimental vs control groups. Depends on the prevalence of disease and allows calculation of NNT (1/RD).





Meta-analysis



•the good





Advantages of metaanalyses

 Quantitative synthesis with increased statistical power of the available evidence





Meta-analysis in DES trials

- At least 70 RCTs with almost 90.000 randomized patients
- Most studies have an nonferiority design
- Many of them are underpowerd because of an imbalance between observed and expected events
- Most of them have a composite endpoint
- None was powered for stent thrombosis



1-year definite ST

		~
Study		%
ID	OR (95% CI)	Weight
BASKET PROVE	0.20 (0.01, 4.18)	2.15
CIBELES •	0.19 (0.01, 4.02)	2.17
COMPARE	0.22 (0.08, 0.66)	15.25
COMPARE II (NOBORI)	0.61 (0.20, 1.86)	7.49
EXAMINATION	0.28 (0.09, 0.87)	11.93
EXCELLENT •	0.34 (0.07, 1.68)	3.84
ISAR TEST IV	0.50 (0.15, 1.67)	6.82
NEXT (NOBORI)	0.25 (0.03, 2.24)	3.44
PLATINUM	- 1.01 (0.20, 5.01)	2.56
RESET	0.83 (0.25, 2.74)	5.14
RESOLUTE	0.21 (0.06, 0.74)	12.02
SORT OUT IV	0.22 (0.05, 1.03)	7.73
SPIRIT II	0.12 (0.00, 2.87)	1.90
SPIRIT III	5.48 (0.30, 99.39)	0.57
SPIRIT IV	0.30 (0.11, 0.83)	11.42
SPIRIT V DIAB	0.16 (0.01, 4.02)	1.72
TWENTE	0.11 (0.01, 2.08)	3.85
ESSENCE DIABETES	(Excluded)	0.00
SEA-SIDE	(Excluded)	0.00
Overall (I-squared = 0.0%, p = 0.753)	0.36 (0.26, 0.51)	100.00
	Γ	





Network meta-analysis: 49 RCTs and 50,844 pts



2-year definite stent throm	oosis*	Odds Ratio [95%]
CoCr-EES vs BMS	⊢−● −1	0.35 (0.17-0.69)
CoCr-EES vs PES		0.34 (0.19-0.62)
0.01	0.1 Favors Stent 1	10 Favors Stent 2



Palmerini et al. Lancet 2012;379:1393-402



Stroke with PCI versus CABG: 14 RCTs

Study			N. pts wi	th stroke	/total n. pts
			OR (95% CI)	CABG	PCI
ARTS 1 (2001)		•	1.49 (0.42, 5.32)	6/605	4/600
AWESOME (2001)		•	1.44 (0.24, 8.71)	3/232	2/222
BARI (1996)	_		3.52 (0.73, 17.01)	7/914	2/915
Budriot (2011)			5.05 (0.24, 106.53)	2/101	0/100
EAST (1994)		-	3.09 (0.32, 30.01)	3/194	1/198
ERACI 2 (2001)			5.04 (0.24, 105.67)	2/225	0/225
GABI (1994)		*	5.20 (0.25, 109.07)	2/177	0/182
LE MANS (2008)		*	5.10 (0.24, 108.77)	2/53	0/52
Leipzig (2002) -		-	3.03 (0.12, 75.13)	1/110	0/110
MASS II (2004)	_		3.09 (0.62, 15.50)	6/203	2/205
RITA (1993)	_		5.13 (0.60, 44.08)	5/501	1/510
SIMA (2000)	-		0.34 (0.01, 8.63)	0/59	1/62
SYNTAX (2009)			6.11 (1.36, 27.37)	12/897	2/903
Seoul (2005)		*	→ 5.16 (0.21, 128.36)	1/70	0/119
Fixed effects (I-squared=0.0%)		\diamond	2.94 (1.69, 5.09)	52/4341	15/4403
Random effects		\diamond	2.94 (1.69, 5.09)		
		Ĩ			
.01 .05 .1	.5 1	5 10	100		
PCI wors	e	CABG worse			



Palmerini et al; JACC 2012



DAPT trial

Outcome	Continued Thienopyridine (N=5020)	Placebo (N=4941)	Hazard Ratio, Thienopyridine vs. Placebo (95% CI)†	P Value†
	no. of patients (%)		
Stent thrombosis‡	19 (0.4)	65 (1.4)	0.29 (0.17-0.48)	< 0.001
Definite	15 (0.3)	58 (1.2)	0.26 (0.14-0.45)	< 0.001
Probable	5 (0.1)	7 (0.1)	0.71 (0.22-2.23)	0.55
Major adverse cardiovascular and cerebrovascular events	211 (4.3)	285 (5.9)	0.71 (0.59–0.85)	<0.001
Death	98 (2.0)	74 (1.5)	1.36 (1.00-1.85)	0.05
Cardiac	45 (0.9)	47 (1.0)	1.00 (0.66-1.52)	0.98
Vascular	5 (0.1)	5 (0.1)	0.98 (0.28-3.39)	0.98
Noncardiovascular	48 (1.0)	22 (0.5)	2.23 (1.32-3.78)	0.002
Myocardial infarction	99 (2.1)	198 (4.1)	0.47 (0.37-0.61)	<0.001
Stroke	37 (0.8)	43 (0.9)	0.80 (0.51-1.25)	0.32
Ischemic	24 (0.5)	34 (0.7)	0.68 (0.40-1.17)	0.16
Hemorrhagic	13 (0.3)	9 (0.2)	1.20 (0.50-2.91)	0.68
Type uncertain	0	1 (<0.1)	_	0.32



TCTAP 2019

Mortality with Extended Duration DAPT After DES: A Pairwise and Bayesian Network Meta-Analysis of 10 RCTs and 31,666 Pts



ES=effect size

Palmerini T, Stone GW. Lancet 2015:on-line



Advantages of meta-analyses

Application to any clinical research question

 Thorough appraisal of the internal validity of primary studies

Explore clinical and statistical heterogeneity





Meta-analysis

•the bad







Interpreting a meta-analysis

 The strength of a metaanalysis is the strength of individual trials and of methods used





Potential problem

Individual Small Trial

Meta-analysis of several such trials

If all the trials have the same systemic bias, all we have done is to tighten the confidence intervals!





Mortality with DES vs BMS

Observational studies

RCT





Kirtane et al Circulation 2009



Type of bias	Description	Relevant domains in the Collaboration's 'Risk of bias' tool
Selection bias.	Systematic differences between baseline characteristics of the groups that are compared.	 Sequence generation; Allocation concealment.
Performance bias.	Systematic differences between groups in the care that is provided or in exposure to factors other than the interventions of interest.	 Blinding of participants, personnel and outcome assessors; Other potential threats to validity.
Attrition bias.	Systematic differences between groups in withdrawals from a study.	 Incomplete outcome data; Blinding of participants, personnel and outcome assessors.
Detection bias.	Systematic differences between groups in how outcomes are determined.	 Blinding of participants, personnel and outcome assessors; Other potential threats to validity.
Reporting bias.	Systematic differences between reported and unreported findings.	 Selective outcome reporting; (see also Chapter <u>10</u>).



Meta-analysis: misleading results

•and the ugly

Mixing apples with pears!

Dfferent type of study (obs. vs RCT) Different clinical setting Difference in endpoint definition Different follow up duration Different drug formulation







Exploring statistical heterogeneity

 Statistical heterogeneity is the variation in true treatment effect among patients within a single study or between different studies







Statistical inconsistency

- Statistical inconsistency (I²) has been recently introduced to overcome the risk of alpha and beta error of standard tests for statistical heterogeneity
- It is computed as [(Q df)/Q] x 100%, where Q is the chi-squared statistic and df is its degrees of freedom
 l²< 25%: low inconsistency
 l² between 25% and 50%: moderate inconsistency
 - I² > 50%: severe inconsistency





Mortality with Omega 3 fats Pts with or without CAD Purifed supplement/fish oil capsules

Study or subcategory	High omega 3 fat	Low omega 3/contr	ol Relative risk	Relative risk
RCT data	(11/14)	(11/14)	(ranuoni) (95% CI)	(141100111) (95% 61
Borchgrevink 1966	10/100	14/100		0.71 (0.33 to 1.53
Natvig 1968	43/6716	40/6690	-	1.07 (0.70 to 1.64
Burr (DART) 1989	93/1015	131/1018	-8-	0.71 (0.55 to 0.92
Kaul 1992	0/58	1/49	< ∎	0.28 (0.01 to 6.78
Leaf 1994	0/275	2/276	< ∎	0.20 (0.01 to 4.16
Sacks (HARP) 1995	0/41	1/39	← =	0.32 (0.01 to 7.57
Eritsland 1996	8/317	6/293		1.23 (0.43 to 3.51
Singh 1997	30/242	26/118		0.56 (0.35 to 0.91
GISSI-P 1999	477/5665	554/5658		0.86 (0.77 to 0.97
Johansen 1999A	1/250	3/250	<	0.33 (0.03 to 3.18
von Schacky 1999	1/112	2/111	← ・ · · · · · · · · · · · · · · · · · ·	0.50 (0.05 to 5.39
Brox 2001	0/80	1/40	← :	0.17 (0.01 to 4.05
Nilsen 2001	11/150	11/150	+	1.00 (0.45 to 2.24
Bemelmans 2002	3/109	1/157		- 4.32 (0.46 to 41.00
Burr 2003	283/1571	242/1543	-	1.15 (0.98 to 1.34
Subtotal (95% CI)	16 701	16 492	•	0.87 (0.73 to 1.03
Total events: 960 (high omega 3	fats), 1035 (low omega 3/cont	trol)		
Test for heterogeneity: $\chi^2=24.12$,	df=14, P=0.04, /2=42.0%			
Test for overall effect: z=1.58, P=	0.11			

TCTAP 2019

Hooper et al; BMJ 2006



Mortality in pts with STEMI: Different type of studies pooled

4	Culprit or	nly PCI	Multivesse	I PCI		Odds Ratio	Odds Ratio
Study or Subgroup	Events	Total	Events	Total	Weight	IV, Random, 95% CI	IV, Random, 95% CI
Prospective studies							
Di Mario 2004	0	17	1	52	0.9%	0.98 [0.04, 25.20]	
Khattab 2008	2	45	1	28	1.5%	1.26 [0.11, 14.53]	
Politi 2010	7	84	2	65	3.1%	2.86 [0.57, 14.27]	
Subtotal (95% CI)		146		145	5.4%	1.98 [0.57, 6.85]	
Total events	9		4				
Heterogeneity: Tau ² = 0	0.00; Chi ² =	0.52, df =	= 2 (P = 0.77); I ² = 09	%		
Test for overall effect: 2	Z = 1.08 (P =	= 0.28)					
Retrospective studies							
Cavender 2010	1321	25802	246	3134	20.9%	0.63 [0.55, 0.73]	•
Corpus 2004	20	354	5	26	5.9%	0.25 [0.09, 0.74]	
Dziewierz 2010	42	707	9	70	9.2%	0.43 [0.20, 0.92]	
Hannan 2010	10	503	17	503	8.9%	0.58 [0.26, 1.28]	
Kong 2006	31	1350	5	632	7.0%	2.95 [1.14, 7.62]	
Poyen 2003	2	81	1	86	1.5%	2.15 [0.19, 24.20]	
Qarawani 2008	1	25	4	95	1.7%	0.95 [0.10, 8.88]	
Roe 2001	10	79	17	79	8.1%	0.53 [0.23, 1.24]	
Schaaf 2010	60	124	19	37	9.7%	0.89 [0.43, 1.85]	
Toma 2010	94	1983	25	217	14.5%	0.38 [0.24, 0.61]	-
Varani 2008	8	156	12	147	7.3%	0.61 [0.24, 1.53]	
Subtotal (95% CI)		31164		5026	94.6%	0.62 [0.45, 0.84]	•
Total events	1599		360				
Heterogeneity: Tau ² = 0	0.11; Chi ² =	20.44, df	= 10 (P = 0.	03); l² =	51%		
Test for overall effect: 2	Z = 3.04 (P =	= 0.002)					
Total (95% CI)		31310		5171	100.0%	0.66 [0.48, 0.89]	•
Total events	1608		364				
Heterogeneity: Tau ² = 0	0.11; Chi ² =	24.31, df	= 13 (P = 0	03); l ² =	47%		
Test for overall effect: 2	Z = 2.69 (P =	= 0.007)					
N							
Network meta-anal	ysis						
All studies (n=17)						0.70 [0.46, 1.14]	● ●
							+
							0.01 0.1 1 10 10



Vlaar et al; JACC 2011

Different population and study design



Elmariah et al; Lancet 2014

Potential problems

- Meta-analysis is as strong as the studies that are included
- A good meta-analysis of badly designed studies will still result in bad statistics
 - The sources of bias are not controlled by the method





What publication bias is: the file drawer problem

 The publication of studies depends on the nature and direction of results

 Positive studies are more likely to be published than negative studies

 The results is that published studies are systematically different from unpublished studies





Detection of publication bias: smaller studies have larger random error



Effect size estimate (mean difference, log risk ratio, etc)



Publication bias

 However, when small studies are predominately in one direction (usually the direction of larger effect sizes), asymmetry will ensue and this may be indicative of publication bias







Exploring publication bias

Begg's test

Egger's test

Meta-regression on study size





Effect of cilastazol on restenosis

Study	Cilostazol	Control	RR (random)	RR (random)	
or sub-category	n/N	n/N	95% CI	95% CI	Year
DICA					
Toko	7/42	16/40		0.42 10 19 0.901	1997
Teuchikana 1999	22/123	51/129	-	0.45 (0.29 0.70)	1999
Nagaoka	4/23	3/24	-	1 29 (0 25 5 55)	2001
Subtotal (05% CI)	199	193		0.49 (0.31 0.28)	2002
Total quarter 22 (Ciladata) 7	Control)	195	~	0.45 (0.31, 0.76)	
Total events 33 (Cilustazol), 70	147 df = 2(P = 20) R = 10	1.000			
Test for oursell offect: 7 = 2.04	(P = 0.02)	0.2 /0			
Test for overall effect. 2 = 3.04	(7 = .002)				
DCA					
Tsuchikana 1998	0/20	5/17		0.08 (0.00 1.31)	1998
Subtotal (95% CI)	20	17		0.08 (0.00 1.31)	2000
Total events 0 (Cilostazol) 5 (Control)	÷.		0.00 (0.00, 1.01)	
Test for beterogeneity not ann	licable				
Test for overall effect: 7 = 1 77	(P = 0.8)				
Teat for overall ender. 2 = 1.11	(7 = .00)				
Stanling					
Kunichima	3/25	11 /41	_	0.22 (0.10.1.05)	1997
Sakiya	9/82	20/83	_	0.46 (0.22 0.94)	1998
Vanacaki	0/18	3/17		0 14 (0 01 2 44)	1998
Ochisi	0/23	5/21		0.08 (0.00 1.42)	1999
Pork	58/254	65/240	_	0.84 (0.62 1.15)	1999
Kozupa	10/61	19/58	-	0.50 (0.25 0.98)	2001
Also	0/22	15/41		0.04 10.00 0.591	2002
Kamishirada	7/54	17/57	-	0.43 (0.20, 0.97)	2002
Miznauchi	4/65	13/65		0.31 (0.11 0.89)	2002
Incue	4/34	10/32		0.38 (0.13 1.08)	2004
Sekieuchi	24/89	20/86	-	1 16 10 69 1 941	2004
CREST	57/259	92/267	•	0.64 (0.48 0.85)	2005
Hap	5/34	10/37		0.54 (0.21 1.43)	2005
RACTS	47/202	60/194		0.75 (0.54) 04)	2005
Tokeyosu	119/427	128/436	1	0.95 (0.77 1.17)	2005
Chen	7/52	17/54		0.43 (0.19. 0.95)	2006
DECLARE-Long	14/250	23/250		0.61 (0.32, 1.16)	2007
Subtotal (95% CI)	1976	1979	6	0.63 (0.51 0.77)	
Total events: 368 (Cilostazol) 5	528 (Control)		*	,	
Test for beterogeneity Chi2 = 3	$244 \text{ df} = 16 (P = 0.09) I^2$	= 50 7%			
Test for overall effect: 7 = 4.34	(P < .0001)				
Total (95% CI)	2184	2189	•	0.60 [0.49, 0.73]	
Total events 401 (Cilostazol) 6	603 (Control)				
Test for heterogeneity, Chi ² = 4	2.21, df = 20 (P = .003), I2	= 52.6%			
Test for overall effect: Z = 4.99	(P < .00001)				
			0.001 0.01 0.1 1 10 100	1000	
			Favours clostarol Eavours cont	rol	









Conclusions

- The validity of a meta-analysis refers to the soundness of the original studies and the procedures used to combine them.
- Although several potential limitation have been identified in these procedures, systematic reviews and meta-analyses succeed when researchers implement the correct methodology and enforce sound validity checklists.



