METRICS META-RESEARCH INNOVATION CENTER AT STANFORD

The epidemic proliferation of useless systematic reviews

John P.A. Ioannidis, MD, DSc

Professor of Medicine, of Health Research and Policy, of Biomedical Data Science, and of Statistics

Co-Director, METRICS

Stanford University

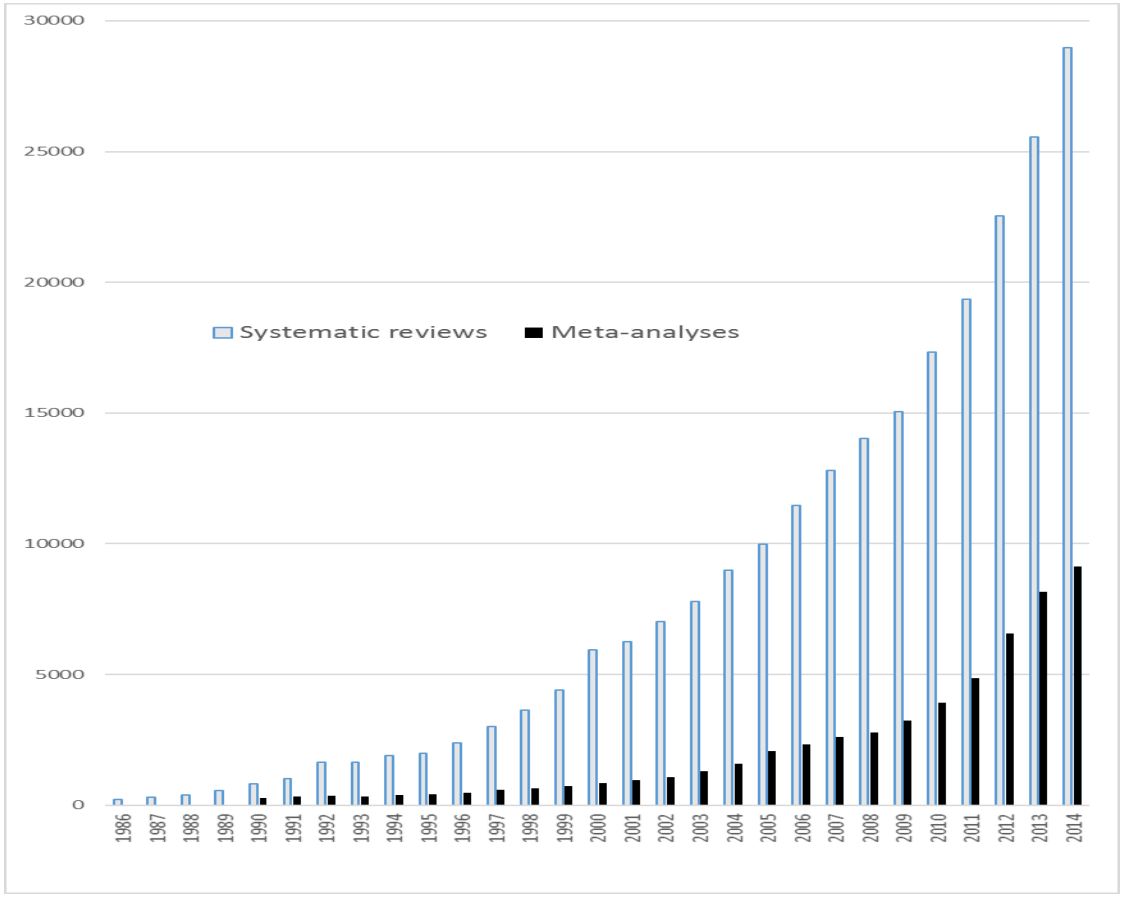
Some potentially heretical statements

- Production of useless systematic reviews and meta-analyses has acquired epidemic proportions
- Systematic reviews have become a prolific business
- Systematic reviews and meta-analyses have become a dangerous, misleading marketing tool

How many systematic reviews and meta-analyses

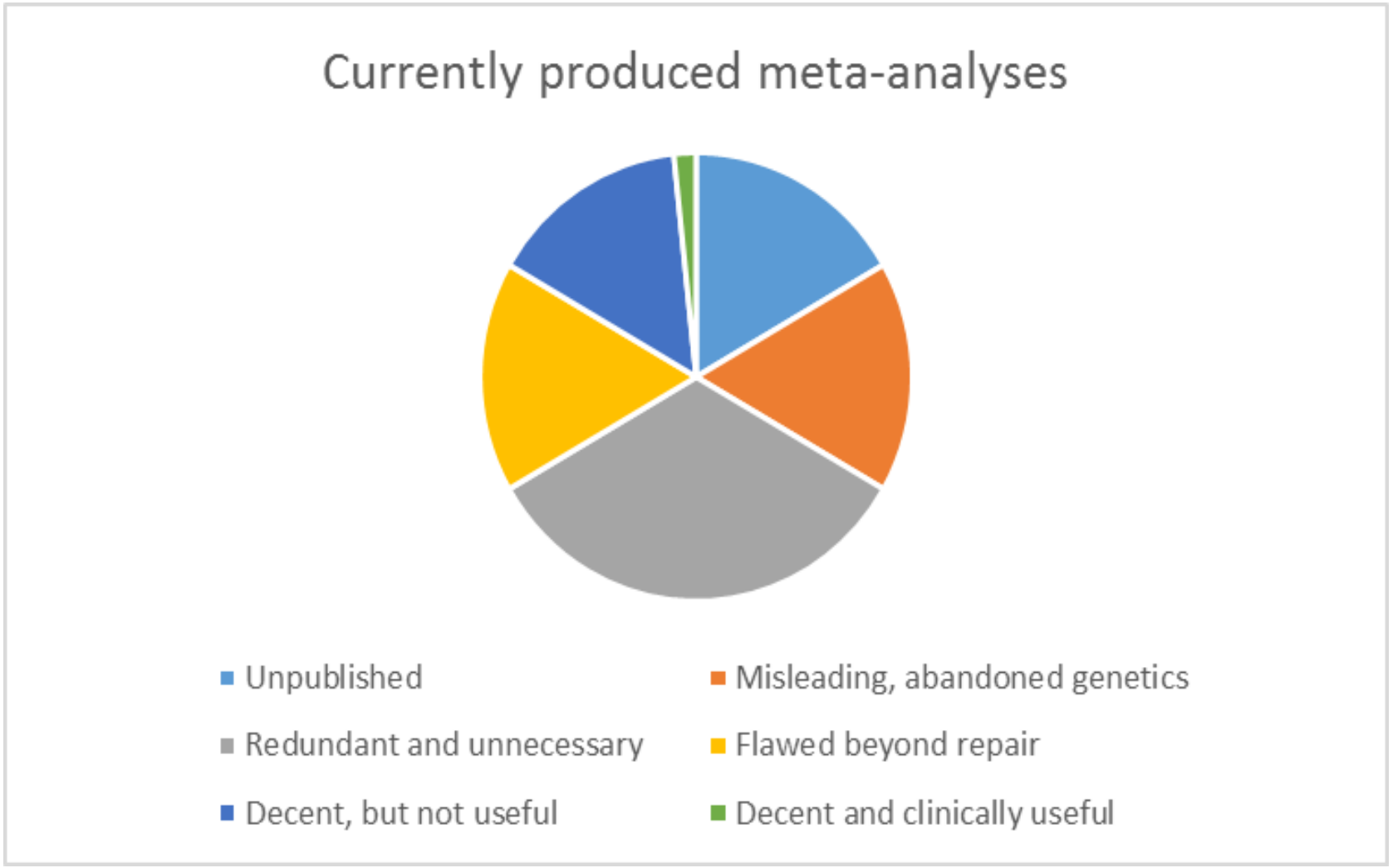
- As of November 2019, there are 106760 published meta-analysis articles indexed in PubMed as publication type "meta-analysis"
- There are over 1000 new ones every month
- There are approximately 250000 published systematic reviews in PubMed, with another 2500 new ones every month
- I estimate that >5000 network meta-analyses have been done, but <40% have been published

The meta-analysis epidemic



The meta-pie

(see Ioannidis, Milbank Quarterly 2016)



Why Most Clinical Research Is Not Useful

John P. A. Ioannidis[1,2]*

1 Stanford Prevention Research Center, Department of Medicine and Department of Health Research and Policy, Stanford University School of Medicine, Palo Alto, California, United States of America, **2** Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Palo Alto, California, United States of America

* jioannid@stanford.edu

Summary Points

- Blue-sky research cannot be easily judged on the basis of practical impact, but clinical research is different and should be useful. It should make a difference for health and disease outcomes or should be undertaken with that as a realistic prospect.
- Many of the features that make clinical research useful can be identified, including those relating to problem base, context placement, information gain, pragmatism, patient centeredness, value for money, feasibility, and transparency.
- Many studies, even in the major general medical journals, do not satisfy these features, and very few studies satisfy most or all of them. Most clinical research therefore fails to be useful not because of its findings but because of its design.
- The forces driving the production and dissemination of nonuseful clinical research are largely identifiable and modifiable.
- Reform is needed. Altering our approach could easily produce more clinical research that is useful, at the same or even at a massively reduced cost.

S

A (2016) Why Most Clinical
l. PLoS Med 13(6): e1002049.
ned.1002049

2016

Table 1. Features to consider in appraising whether clinical research is useful.

Feature	Questions to Ask
Problem base	Is there a health problem that is big/important enough to fix?
Context placement	Has prior evidence been systematically assessed to inform (the need for) new studies?
Information gain	Is the proposed study large and long enough to be sufficiently informative?
Pragmatism	Does the research reflect real life? If it deviates, does this matter?
Patient centeredness	Does the research reflect top patient priorities?
Value for money	Is the research worth the money?
Feasibility	Can this research be done?
Transparency	Are methods, data, and analyses verifiable and unbiased?

doi:10.1371/journal.pmed.1002049.t001

OPEN ACCESS Freely available online

PLOS | ONE

The Geometric Increase in Meta-Analyses from China in the Genomic Era

John P. A. Ioannidis[1,2*], Christine Q. Chang[1], Tram Kim Lam[1], Sheri D. Schully[1], Muin J. Khoury[1,3]

Table 1. Meta-analyses in PubMed According to Publication Year.

Year	All	China	US
1995	429	0	165
1996	482	1	197
1997	596	3	250
1998	639	0	235
1999	741	0	305
2000	849	2	335
2001	948	3	366
2002	1078	11	400
2003	1289	19	401
2004	1594	28	467
2005	2063	33	541
2006	2331	77	681
2007	2594	97	696
2008	2773	179	756
2009	3229	302	774
2010	3904	540	896
2011	4739	828	965
2012 (until search)	2270	464	446

In 2019 (to-date)

“Meta-analysis”: 5837 form China and 2822 from USA

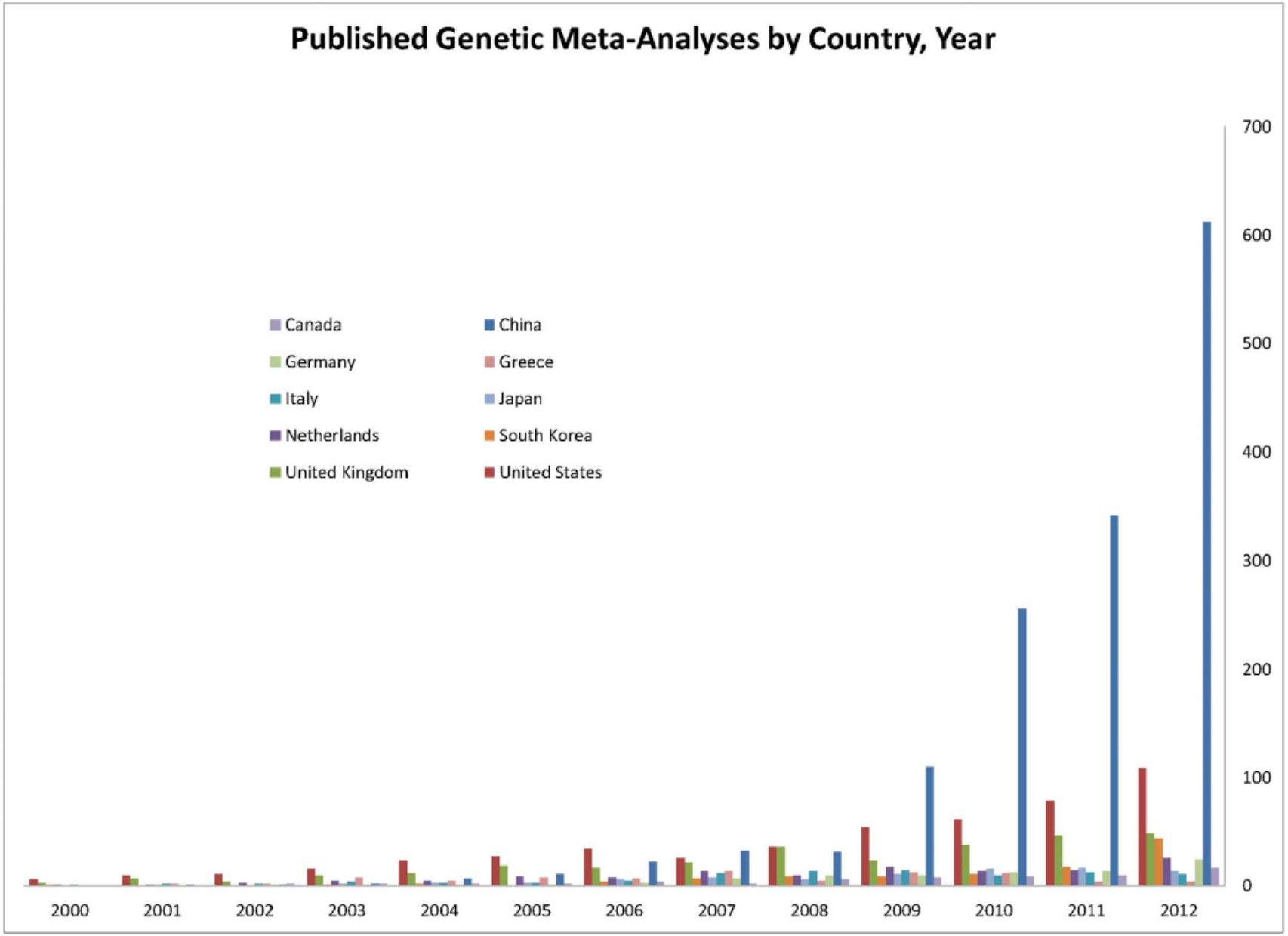
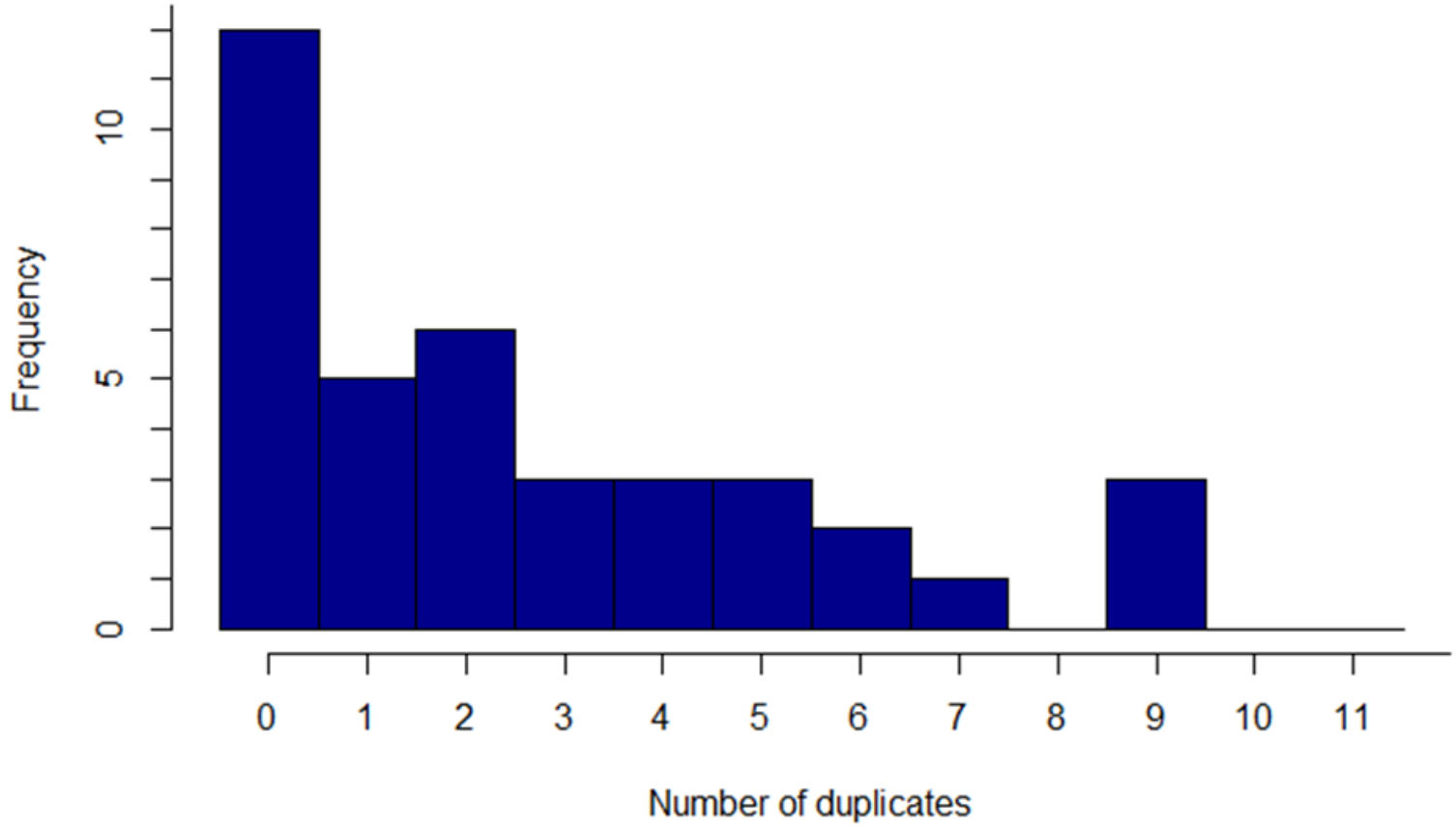


Figure 1. Annual number of meta-analyses of genetic associations for the 10 most-prolific countries in the period 2000–2012; data are derived from HuGE Navigator (last update January 13, 2012).

doi:10.1371/journal.pone.0065602.g001

Strict duplicates in genetic epi meta-analyses (5 year window)



Industry and contractors

Network meta-analyses performed by contracting companies and commissioned by industry

Ewoud Schuit[1,2] and John PA Ioannidis[1,2*]

Systematic reviews as a prolific business

- Over 100 service-offering companies perform systematic reviews
- Dozens of them perform even network meta-analyses
- Their production is >3 times larger than what is published in the literature
- Most of what they do is not published

Overlapping meta-analyses on the same topic: survey of published studies

Konstantinos C Siontis *resident physician*[1], Tina Hernandez-Boussard *assistant professor*[2], John P A Ioannidis *professor*[3 4]

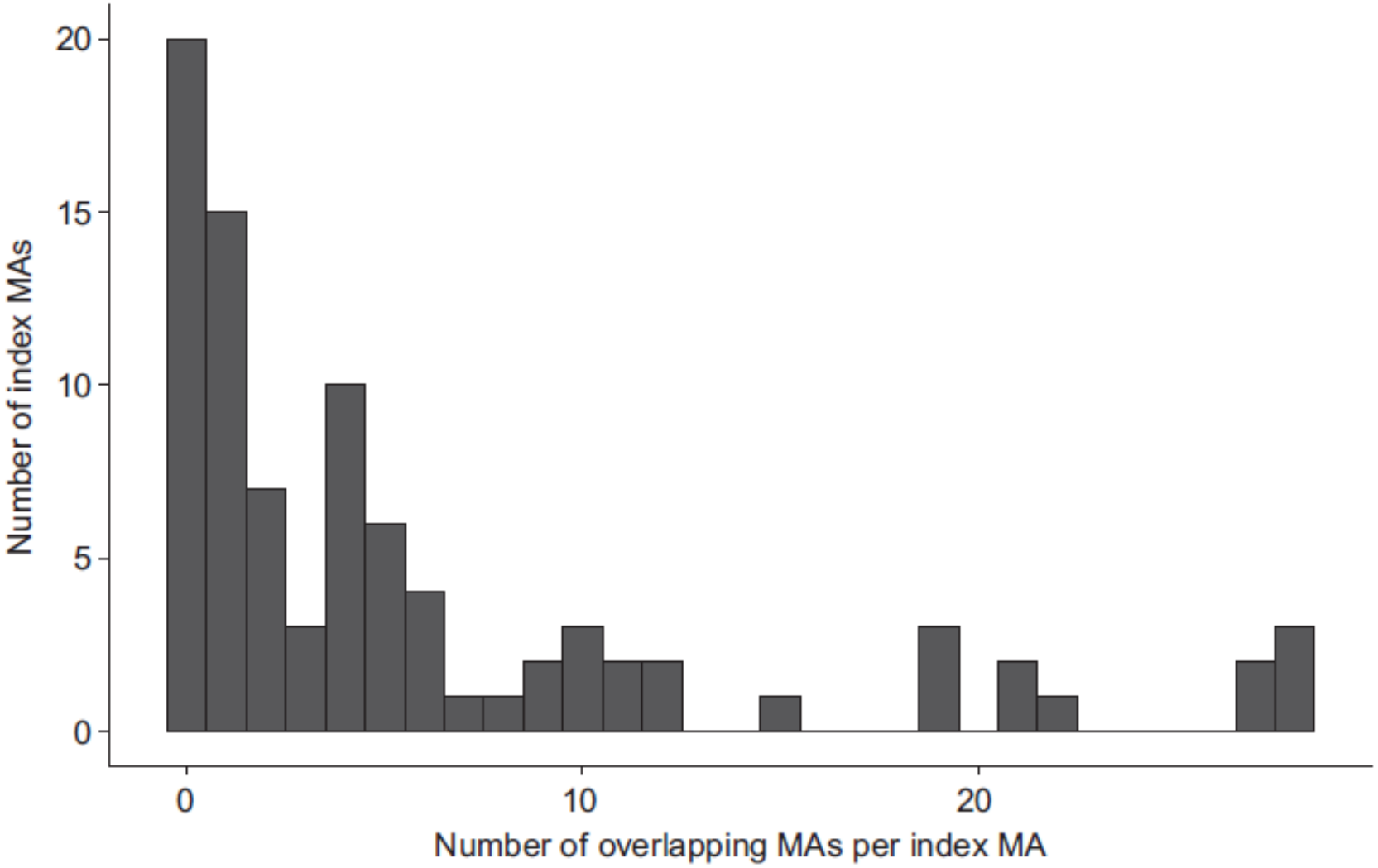
Study selection and methods Meta-analyses published in 2010 were identified, and 5% of them were randomly selected. We further selected those that included randomized trials and examined effectiveness of any medical intervention. For eligible meta-analyses, we searched for other meta-analyses on the same topic (covering the same comparisons, indications/settings, and outcomes or overlapping subsets of them) published until February 2013.

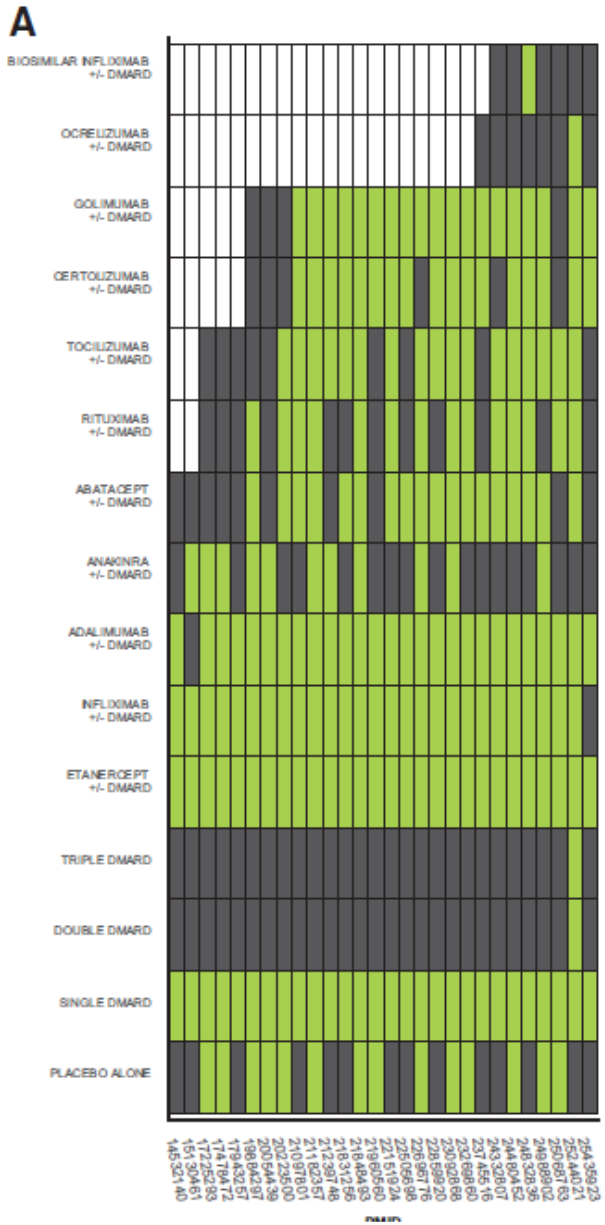
Table 5| **Mapping of potential redundancy in 11 meta-analyses on use of statins for prevention of atrial fibrillation after cardiac surgery**

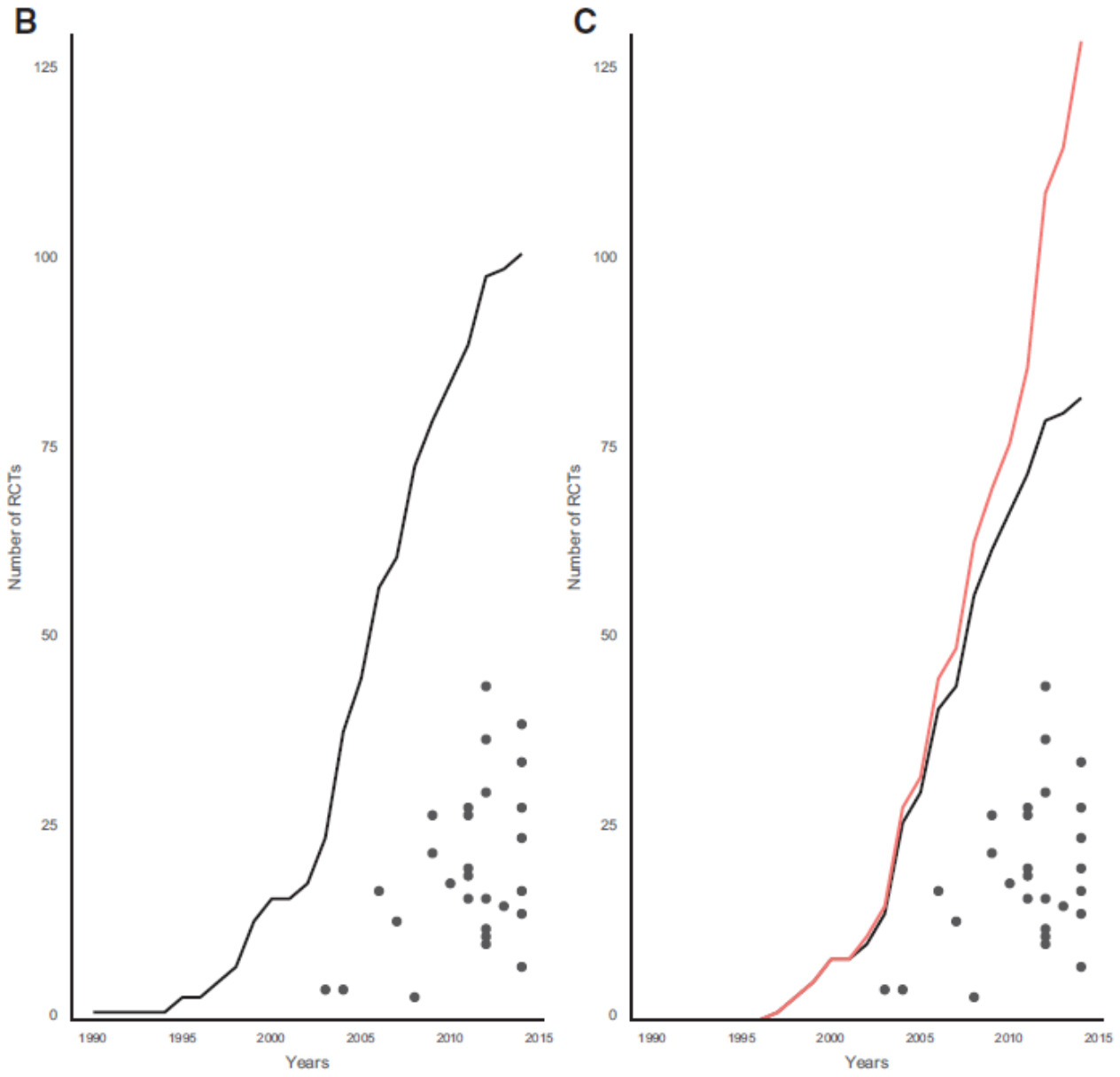
	1	2	3	4	5	6	7	8	9	10	11
Month/year published	2/2008	6/2008	7/2009	9/2009	12/2009*	8/2010	9/2010	2/2011	2/2012	4/2012	11/2012
Month/year of last search	6/2007	2/2008	5/2008	7/2008	12/2008*	8/2009	2/2010	8/2010	4/2011	5/2010	12/2010
Eligible studies	RCT†	RCT+O	RCT	RCT+O	RCT+O	RCT	RCT	RCT	RCT	RCT	RCT
Effect size (95% CI)	0.60‡ (0.27 to 1.37)	0.67‡ (0.51 to 0.88)	0.57‡ (0.42 to 0.78)	0.78‡ (0.67 to 0.90)	0.68‡ (0.59 to 0.79)	0.57§ (0.45 to 0.72)	0.54§ (0.43 to 0.68)	0.40‡ (0.29 to 0.55)	0.56§ (0.45 to 0.69)	0.40‡ (0.29 to 0.55)	0.40‡ (0.29 to 0.55)

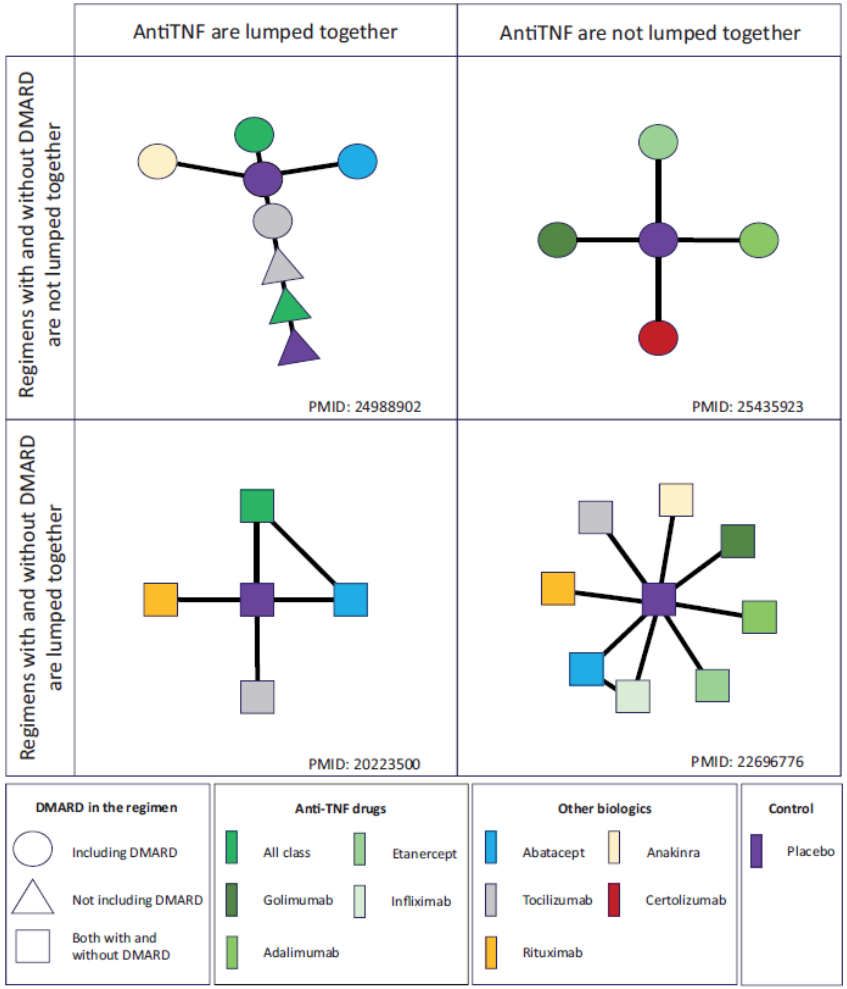
Overlapping network meta-analyses on the same topic: survey of published studies

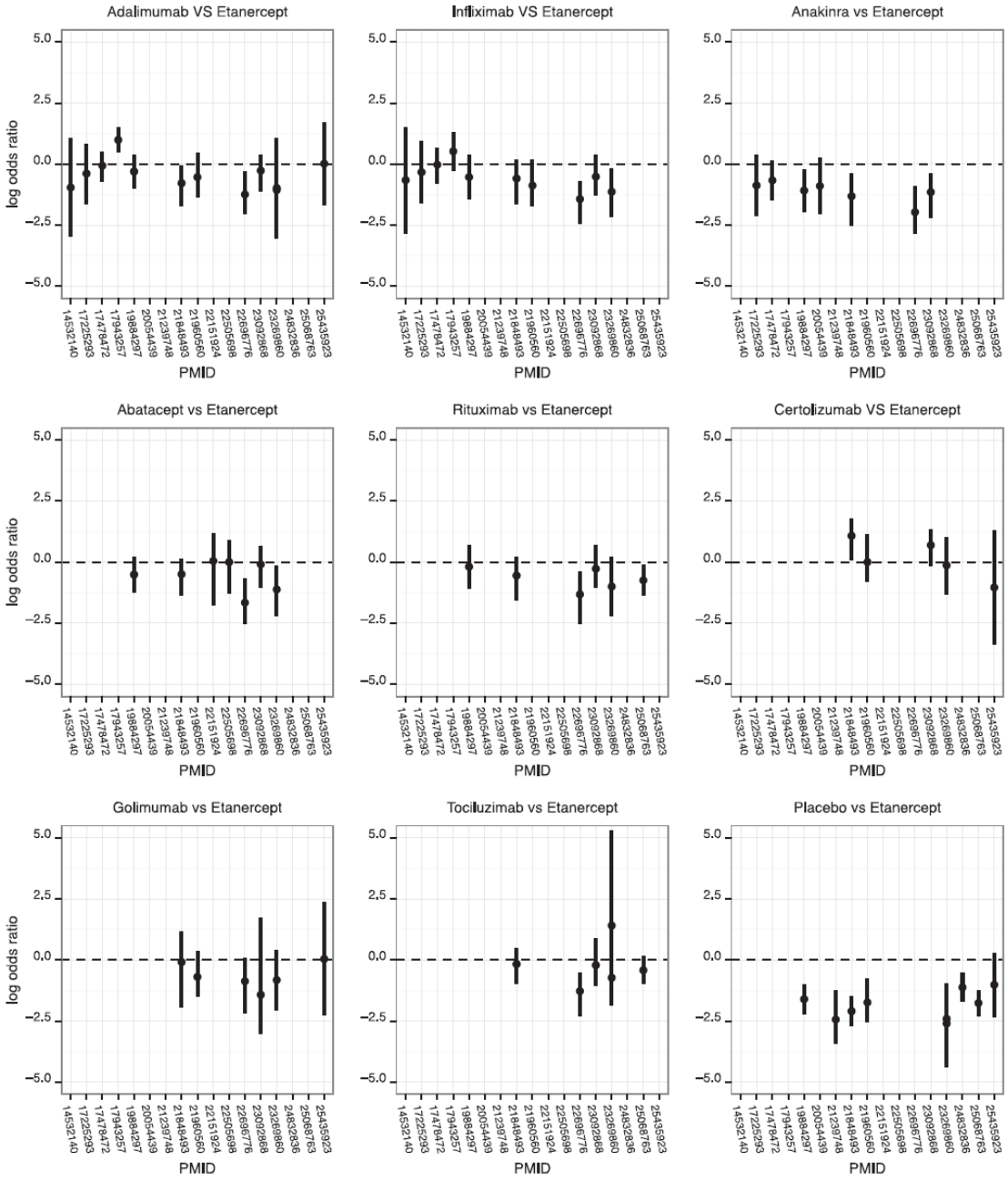
F Naudet,[1] E Schuit[1,2] and J P A Ioannidis[1,3,*]




A
BIOSIMILAR INFLIXIMAB +/- DMARD
OCRELIZUMAB +/- DMARD
GOLIMUMAB +/- DMARD
CERTOLIZUMAB +/- DMARD
TOCILIZUMAB +/- DMARD
RITUXIMAB +/- DMARD
ABATACEPT +/- DMARD
ANAKINRA +/- DMARD
ADALIMUMAB +/- DMARD
INFLIXIMAB +/- DMARD
ETANERCEPT +/- DMARD
TRIPLE DMARD
DOUBLE DMARD
SINGLE DMARD
PLACEBO ALONE
145321.40
151304.61
172252.83
174784.72
179432.57
198842.97
200654.39
202235.00
210978.01
211823.57
212397.48
218312.56
218484.93
219980.60
221519.24
225056.98
226967.76
228589.20
230926.88
232698.60
237455.16
243328.07
244804.52
248328.36
249889.02
250887.63
252440.21
254359.23


B
C
125
100
75
50
25
0
Number of RCTs
1990
1995
2000
2005
2010
2015
Years


AntiTNF are lumped together
AntiTNF are not lumped together
Regimens with and without DMARD are not lumped together
Regimens with and without DMARD are lumped together
PMID: 24988902
PMID: 25435923
PMID: 20223500
PMID: 22696776
DMARD in the regimen
Including DMARD
Not including DMARD
Both with and without DMARD
Anti-TNF drugs
All class
Etanercept
Golimumab
Infliximab
Adalimumab
Other biologics
Abatacept
Anakinra
Tocilizumab
Certolizumab
Rituximab
Control
Placebo


Adalimumab VS Etanercept
Infliximab VS Etanercept
Anakinra vs Etanercept
Abatacept vs Etanercept
Rituximab vs Etanercept
Certolizumab VS Etanercept
Golimumab vs Etanercept
Tociluzimab vs Etanercept
Placebo vs Etanercept
log odds ratio
5.0
2.5
0.0
−2.5
−5.0
14532140
17225293
17478472
17943257
19884297
20054439
21239748
21848493
21960560
22151924
22505698
22696776
23092868
23269860
24832836
25068763
25435923
PMID

Palpacuer *et al. BMC Medicine* (2019) 17:174
https://doi.org/10.1186/s12916-019-1409-3

BMC Medicine

RESEARCH ARTICLE Open Access

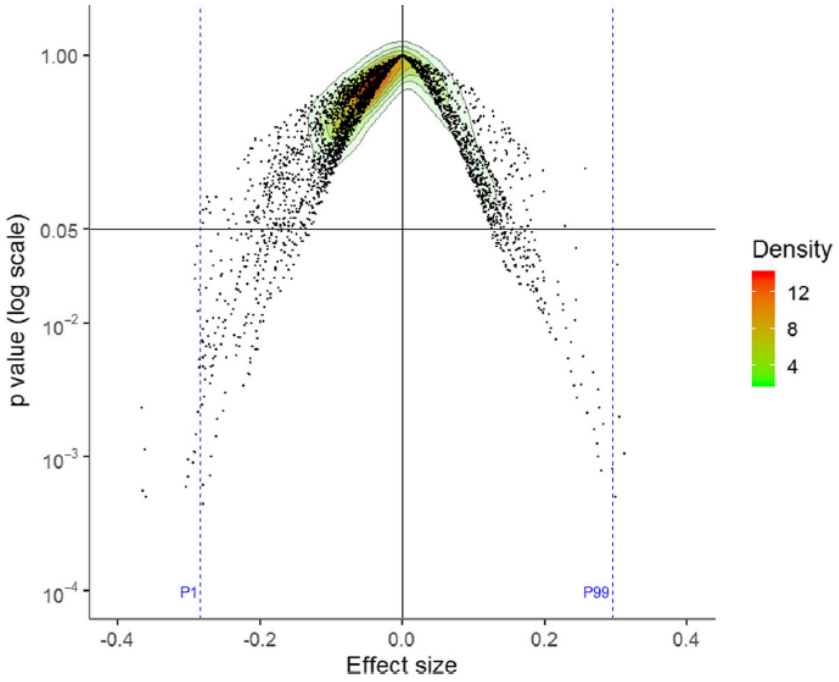
Vibration of effects from diverse inclusion/exclusion criteria and analytical choices: 9216 different ways to perform an indirect comparison meta-analysis

Clément Palpacuer[1,2*], Karima Hammas[3,4], Renan Duprez[5], Bruno Laviolle[1,6,7], John P. A. Ioannidis[8,9] and Florian Naudet[1,6,7,8]

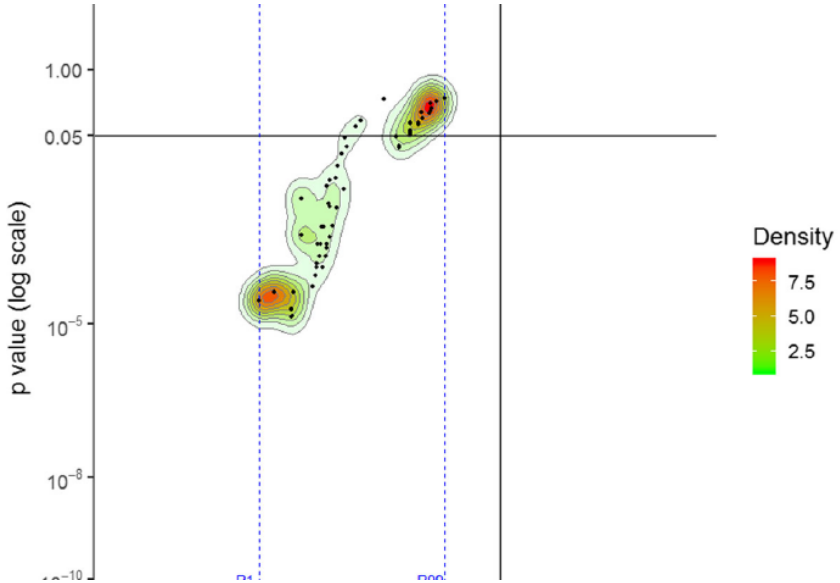
Table 1 Definition of the different methodological choices and number of possible analytical scenarios

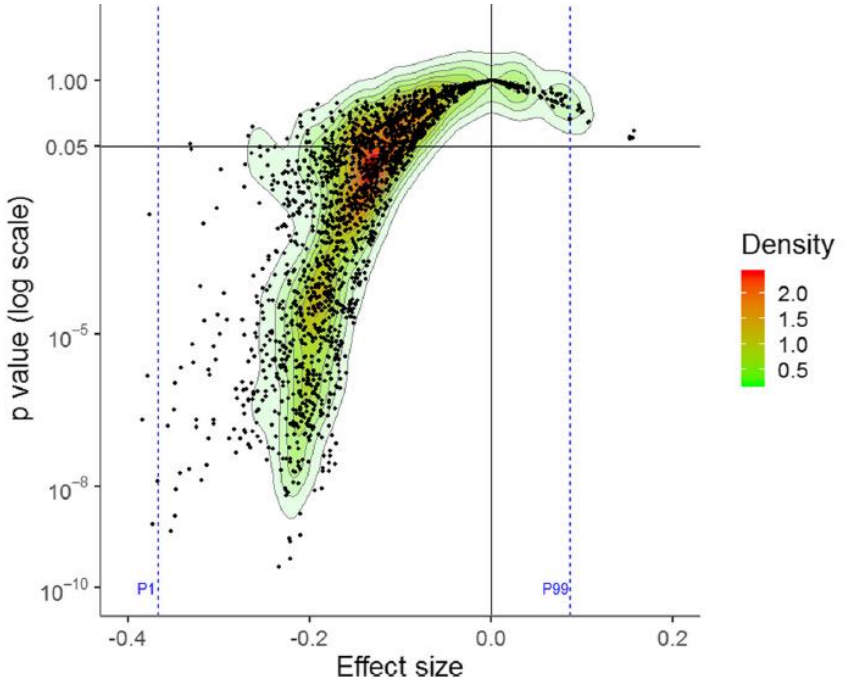
Category	Criteria	Number of possibilities
Medical condition	Inclusion of all studies (AUDs and/or AD) Exclusion of studies including patients with AUDs	2
Abstinence[†]	Inclusion of all studies (abstinent or non-abstinent patients) Exclusion of studies requiring a minimum period of abstinence of 5 days or more before the beginning of the study	2
Gender	Inclusion of all studies (mixed gender, males only or females only) Exclusion of studies with males or females only	2
Somatic comorbidity	Inclusion of all studies (patients with or without systematic somatic comorbidities) Exclusion of studies on patients with systematic somatic comorbidities (e.g. studies on patients with HIV)	2
Psychiatric comorbidity	Inclusion of all studies (patients with or without systematic psychiatric comorbidities) Exclusion of studies on patients with systematic psychiatric comorbidities (e.g. studies on depressed patients)	2
Psychological support	Inclusion of all studies (with or without psychological intervention) Exclusion of studies with no psychological intervention	2
Treatment and dose	Only approved dose and route of administration Approved dose and route of administration OR closest dose to the approved dose Maximum dose tested	3
Treatment duration	Inclusion of all studies, regardless of treatment duration Exclusion of studies with a treatment duration of less than 12 weeks	2
Outcome[‡]	Quantity of alcohol consumed* Frequency of drinking** Abstinence***	3
Publication	Published and unpublished studies (e.g. study reports, ClinicalTrials.gov) Exclusion of unpublished studies	2
Risk of bias	Inclusion of all studies, regardless of the risk of selective outcome reporting Exclusion of studies with a high risk of selective outcome reporting	2
Analysis	Fixed effect model Random effect model	2
Total of possible combinations		9216

Vibration of effects in meta-analysis



s for the indirect comparison of nalmefene to naltrexone. A negative effect size favours nalmefene,





for the direct comparison of naltrexone to placebo. A negative effect size favours naltrexone,

Meta-analyses on antidepressants for depression

- We identified 185 eligible meta-analyses published over these 7 years.
- 54 (29%) had authors who were employees of the assessed drug manufacturer and 147 (79%) had some industry link (sponsorship or authors who were industry employees and/or had conflicts of interest).

Meta-analyses as a marketing tool

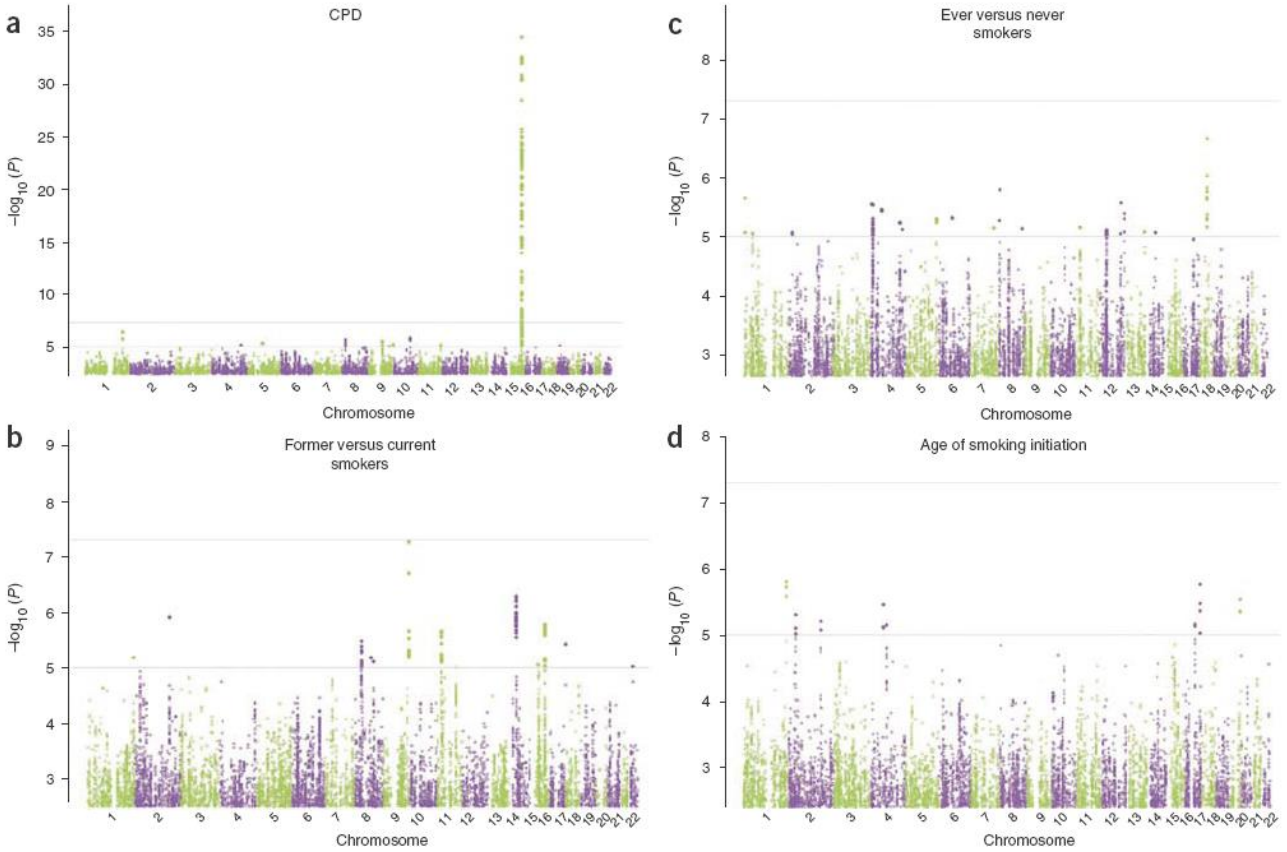
- Only 58 of the 185 meta-analyses on antidepressants for depression (31%) had any negative statements in the concluding statement of the abstract.
- Meta-analyses including an author who were employees of the manufacturer of the assessed drug were 22-times less likely to have negative statements about the drug than other meta-analyses (1/54 [2%] vs. 57/131 [44%], $p<0.001$).

Meta-analyses>>>>Trials

- 67 meta-analyses for 14 trials on direct oral anticoagulants as stroke prevention in atrial fibrillation
- For all drugs only one very large trial dominates the evidence or is the only evidence

Doundoulakis et al. and Siontis and Ioannidis, Circ Cardiovasc Qual Out 2018

Large-scale collaboration, reproducible research and prospective meta-analysis



Raw: can we even trust the data?

Restoring Study 329: efficacy and harms of paroxetine and imipramine in treatment of major depression in adolescence

Joanna Le Noury,[1] John M Nardo,[2] David Healy,[1] Jon Jureidini,[3] Melissa Raven,[3] Catalin Tufanaru,[4] Elia Abi-Jaoude[5]

ABSTRACT

OBJECTIVES

To reanalyse SmithKline Beecham's Study 329 (published by Keller and colleagues in 2001), the primary objective of which was to compare the efficacy and safety of paroxetine and imipramine with placebo in the treatment of adolescents with unipolar major depression. The reanalysis under the restoring invisible and abandoned trials (RIAT) initiative was done to see whether access to and reanalysis of a full dataset from a randomised controlled trial would have clinically relevant implications for evidence based medicine.

DESIGN

Double blind randomised placebo controlled trial.

SETTING

12 North American academic psychiatry centres, from 20 April 1994 to 15 February 1998.

PARTICIPANTS

275 adolescents with major depression of at least eight weeks in duration. Exclusion criteria included a range of comorbid psychiatric and medical disorders and suicidality.

INTERVENTIONS

Participants were randomised to eight weeks double blind treatment with paroxetine (20-40 mg), imipramine (200-300 mg), or placebo.

MAIN OUTCOME MEASURES

The prespecified primary efficacy variables were change from baseline to the end of the eight week acute treatment phase in total Hamilton depression scale (HAM-D) score and the proportion of responders (HAM-D score ≤8 or ≥50% reduction in baseline HAM-D) at acute endpoint. Prespecified secondary outcomes were changes from baseline to endpoint in depression items in K-SADS-L, clinical global impression, autonomous functioning checklist, self-perception profile, and sickness impact scale; predictors of response; and number of patients who relapse during the maintenance phase. Adverse experiences were to be compared primarily by using descriptive statistics. No coding dictionary was prespecified.

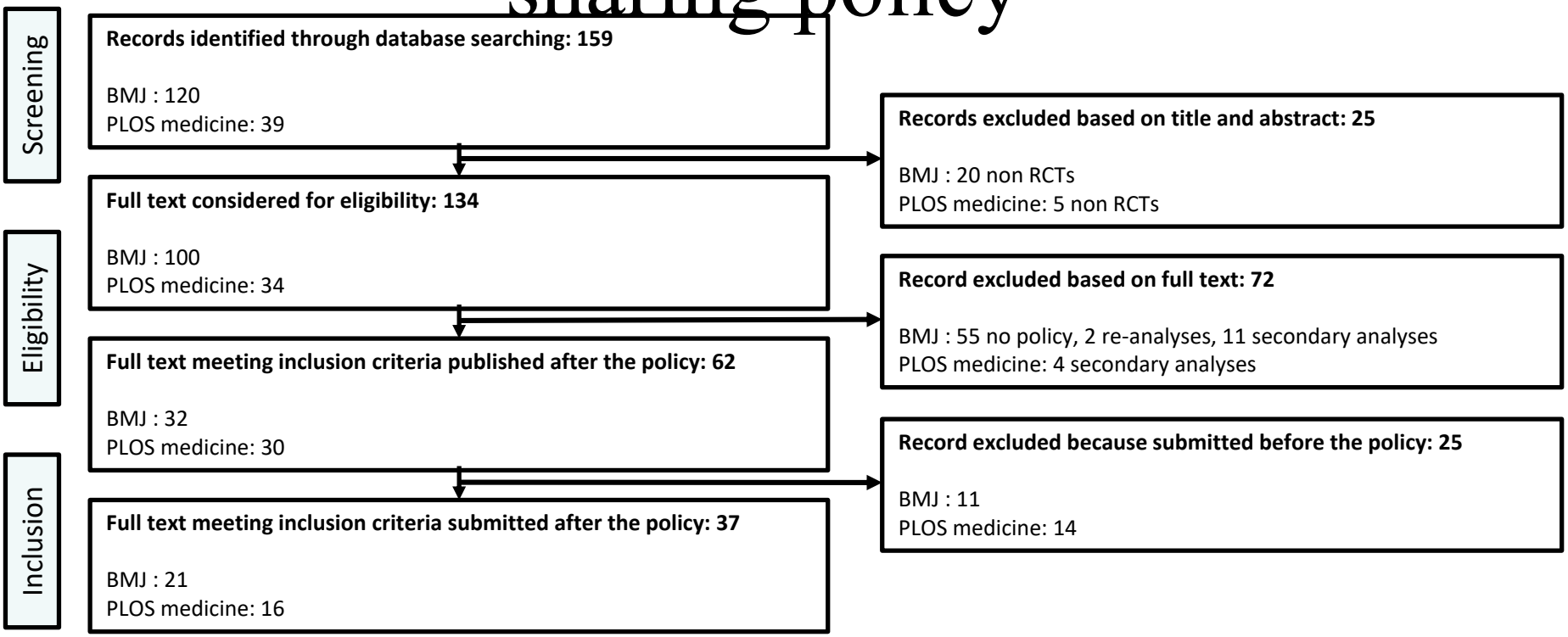
RESULTS

The efficacy of paroxetine and imipramine was not statistically or clinically significantly different from placebo for any prespecified primary or secondary efficacy outcome. HAM-D scores decreased by 10.7 (least squares mean) (95% confidence interval 9.1 to 12.3), 9.0 (7.4 to 10.5), and 9.1 (7.5 to 10.7) points, respectively, for the paroxetine, imipramine and placebo groups (P=0.20). There were clinically significant increases in harms, including suicidal ideation and behaviour and other serious adverse events in the paroxetine group and cardiovascular problems in the imipramine group.

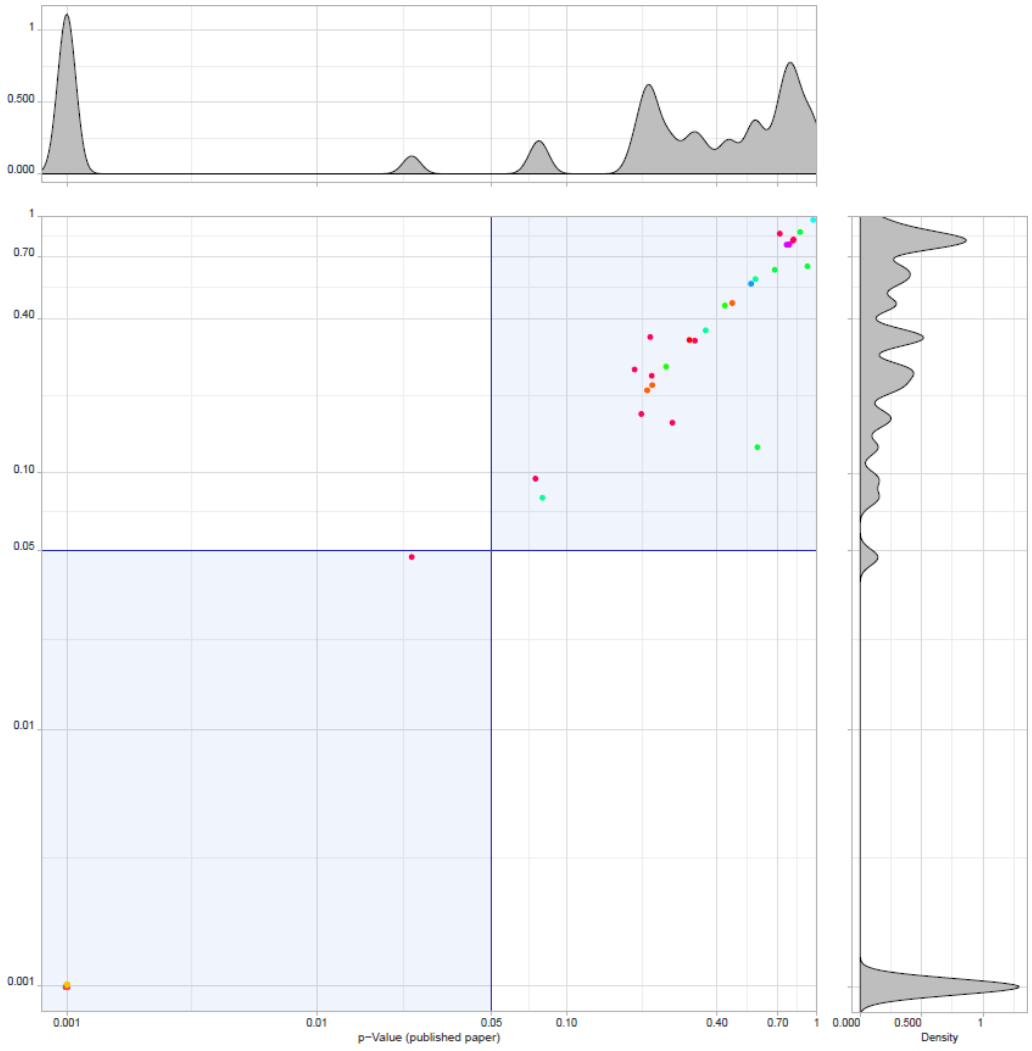
CONCLUSIONS

Neither paroxetine nor high dose imipramine showed efficacy for major depression in adolescents, and there was an increase in harms with both drugs. Access to primary data from trials has important implications for both clinical practice and research, including that published conclusions about efficacy and safety should not be read as authoritative. The reanalysis of Study 329 illustrates the necessity of making primary trial data and protocols available to increase the rigour of the evidence base.

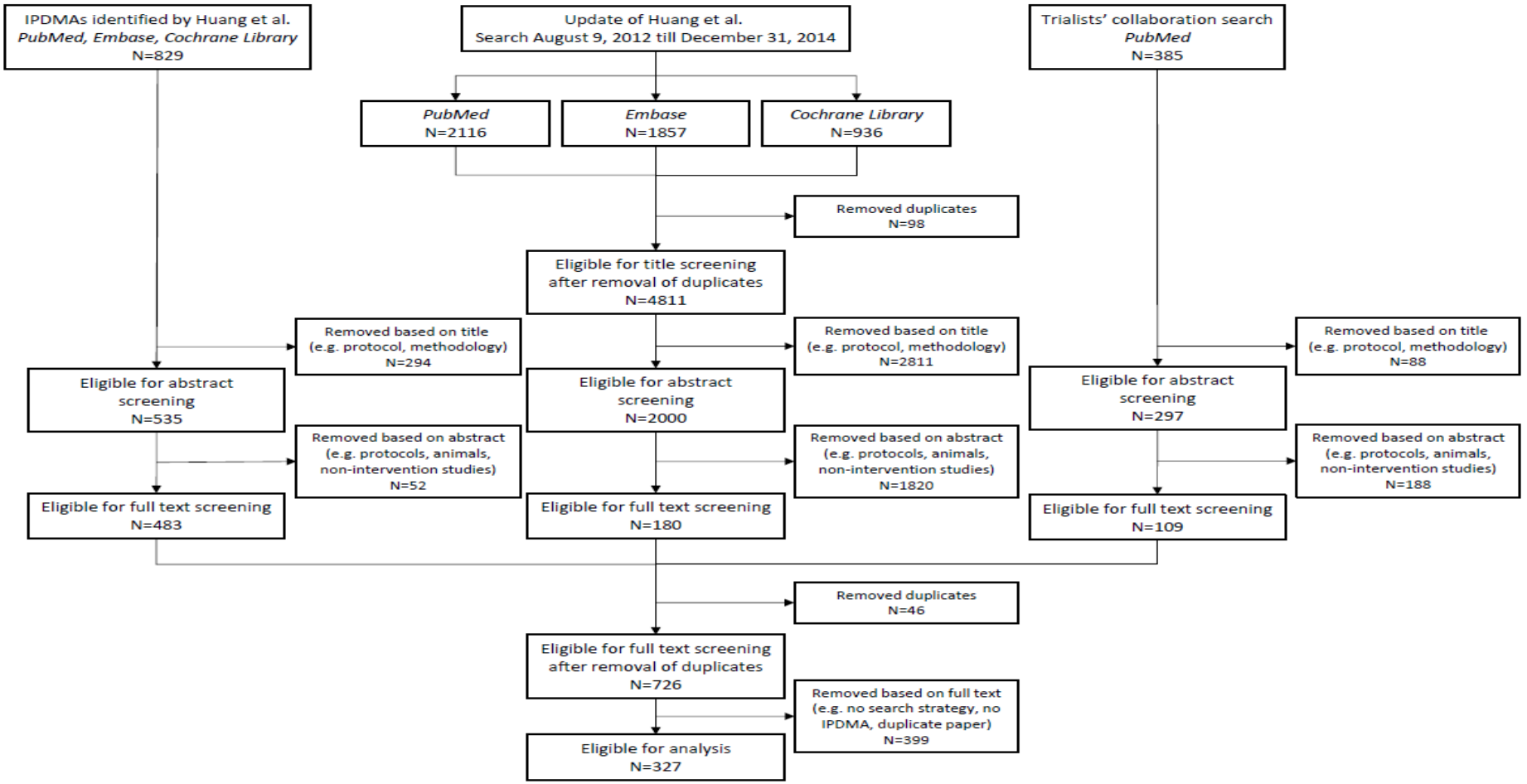
46% retrieval rate for raw data of randomized trials under full data sharing policy

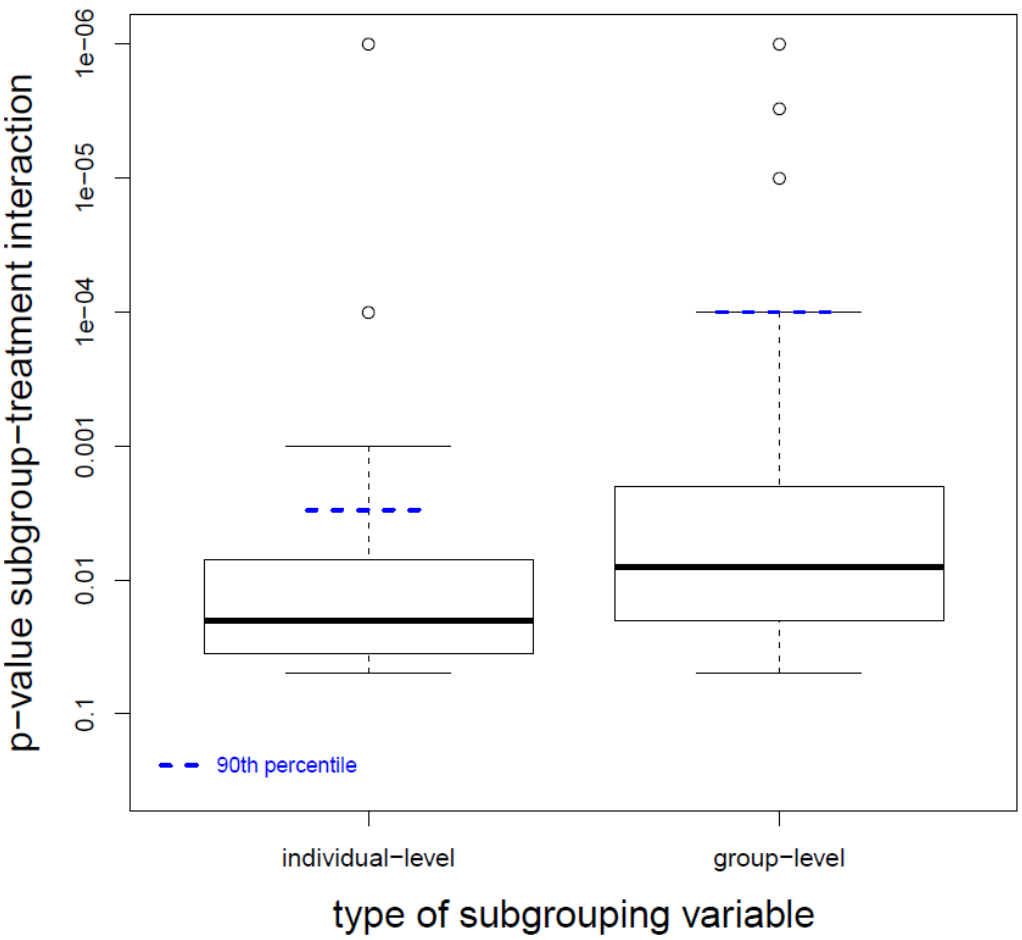


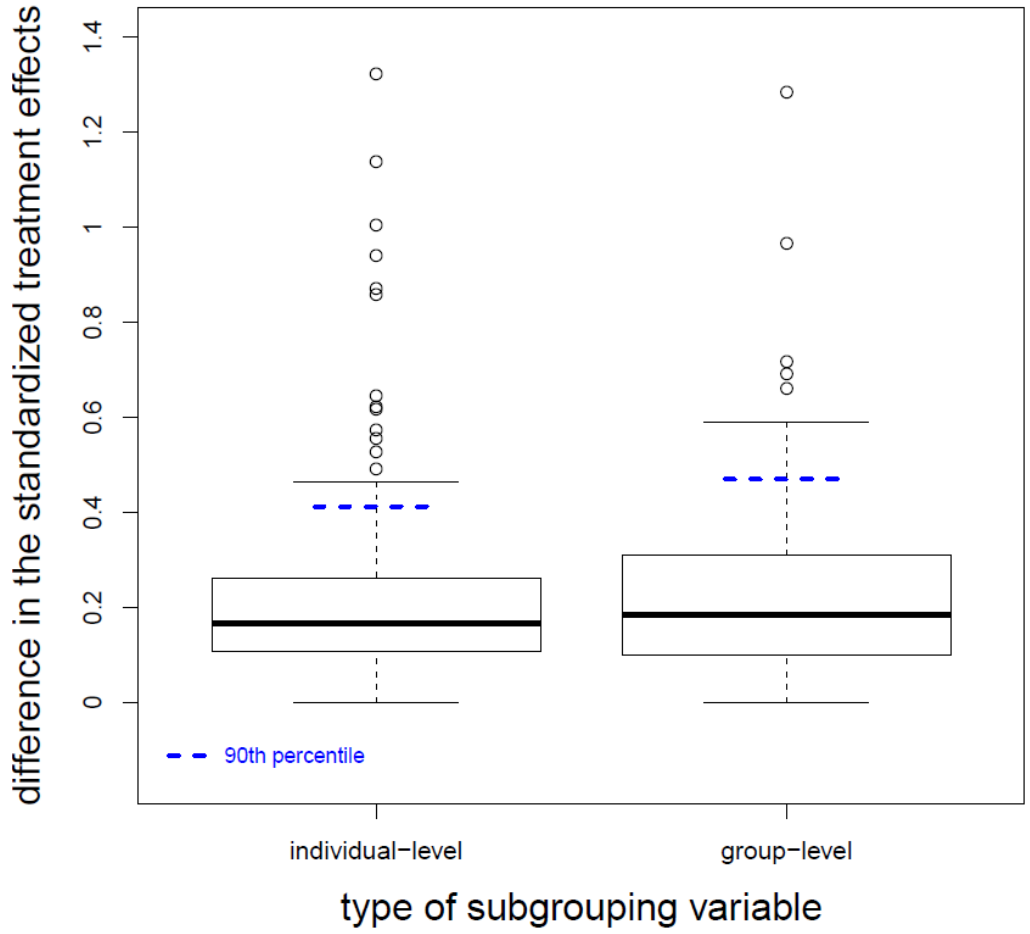
Naudet et al, BMJ 2018


1
0.500
0.000
1
0.70
0.40
0.10
0.05
0.01
0.001
0.001
0.01
0.05
0.10
0.40
0.70
1
p−Value (published paper)
0.000
0.500
1
Density

What have we learned from ~1000 MIPDs




p−value subgroup−treatment interaction
1e−06
1e−05
1e−04
0.001
0.01
0.1
90th percentile
individual−level
group−level
type of subgrouping variable


difference in the standardized treatment effects
0
0.2
0.4
0.6
0.8
1
1.2
1.4
90th percentile
individual−level
group−level
type of subgrouping variable

A prospective view

- Design systematic reviews and meta-analyses as prospective network designs
  - All teams join forces worldwide
  - Data are incorporated prospectively
  - Geometry of comparisons is pre-designed
  - Next study is designed based on enhancing, improving geometry of the network, and maximizing the informativity given the network

Meta-analysis=main type of primary, prospective research

Concluding comments

- The main utility of systematic reviews and meta-analyses has been to reveal how unreliable biomedical evidence is
- This pervasive message should mostly sensitize people to do something about improving the evidence
- Instead, (poor and useless) meta-analyses of (poor and useless) evidence have been entrenched as sort of gold standard and they have become an evolving epidemic
- Individual level data enhance our capabilities, but they require more wide-spread data-sharing and they may not always reveal reliable effect modifications for individualizing treatment
- Retrospective systematic reviews and meta-analyses should be gradually abandoned and primary research should be gradually converted to prospective meta-analysis