Critical evaluation of clinical trial data

Guy Goodwin

Warneford Hospital

Disclosures of interest

Grants	Lundbeck, Servier, Wellcome Trust,
Honoraria	AstraZeneca, BMS, Lundbeck, Medscape, Otsuka, Pfizer, Servier,Takeda
Shares	P1vital
Paid positions	University of Oxford
Advisory boards	AstraZeneca, BMS, Cephalon Lundbeck, Merck, Medscape, Otsuka, P1Vital, Servier, Sunovion, Takeda

Objectives

- Placebo
  - The antidepressant controversy
- The key components of successful clinical trials
- What it means for individual patients
- Reasons to be cheerful

The need for a deductive process

Galen's 80 AD said "All who drink of this remedy recover in a short time except those whom it does not help, who all die. Therefore, it is obvious that it fails only in incurable cases." (Galen's view on evidence 180 AD)

Why do we need clinical trials?

- Surgery?
  - Surgeons don’t do clinical trials
- Drug effects
  - Modest ES
  - Underpin practice
- To justify costs
  - Efficacy and safety as commercial barriers
  - Efficiency as reimbursement barrier

Evidence-based guidelines for treating bipolar disorder: Revised third edition recommendations from the British Association for Psychopharmacology

GM Goodwin[1], PM Haddad[2], IN Ferrier[3], JK Aronson[4], TRH Barnes[5], A Cipriani[1], DR Coghill[6], S Fazel[1], JR Geddes[1], H Grunze[7], EA Holmes[8], O Howes[9], S Hudson[10], N Hunt[11], I Jones[12], IC Macmillan[13], H McAllister-Williams[3], DR Miklowitz[14], R Morriss[15], M Munafò[16], C Paton[17], BJ Saharkian[18], KEA Saunders[1], JMA Sinclair[19], D Taylor[20], E Vieta[21] and AH Young[22]

Psychopharm

Journal of Psychopharmacology
1–59

DOI: 10.1177/0269881116636545
jop.sagepub.com
SAGE

Table 1. Traditional evidence categories.

Evidence categories	Treatment studies	Observational studies
I	Meta-analysis of RCTs, at least one large, good-quality, RCT or replicated, smaller RCTs	Large representative population samples
II	Small, non-replicated RCTs, at least one controlled study without randomization or evidence from at least one other type of quasi-experimental study	Small, well designed but not necessarily representative samples
III	Non-experimental descriptive studies, such as uncontrolled, comparative, correlation and case-control studies	Non-representative surveys, case reports
IV	Expert committee reports or opinions and/or clinical experience of BAP expert group	

Randomized Controlled Trials (RCTs) must have an appropriate control treatment arm; for primary efficacy this should include a placebo condition although for psychological treatments this may not be met. BAP: British Association for Psychopharmacology.

Table 2. Grades of recommendation and their relationship with supporting levels of evidence.

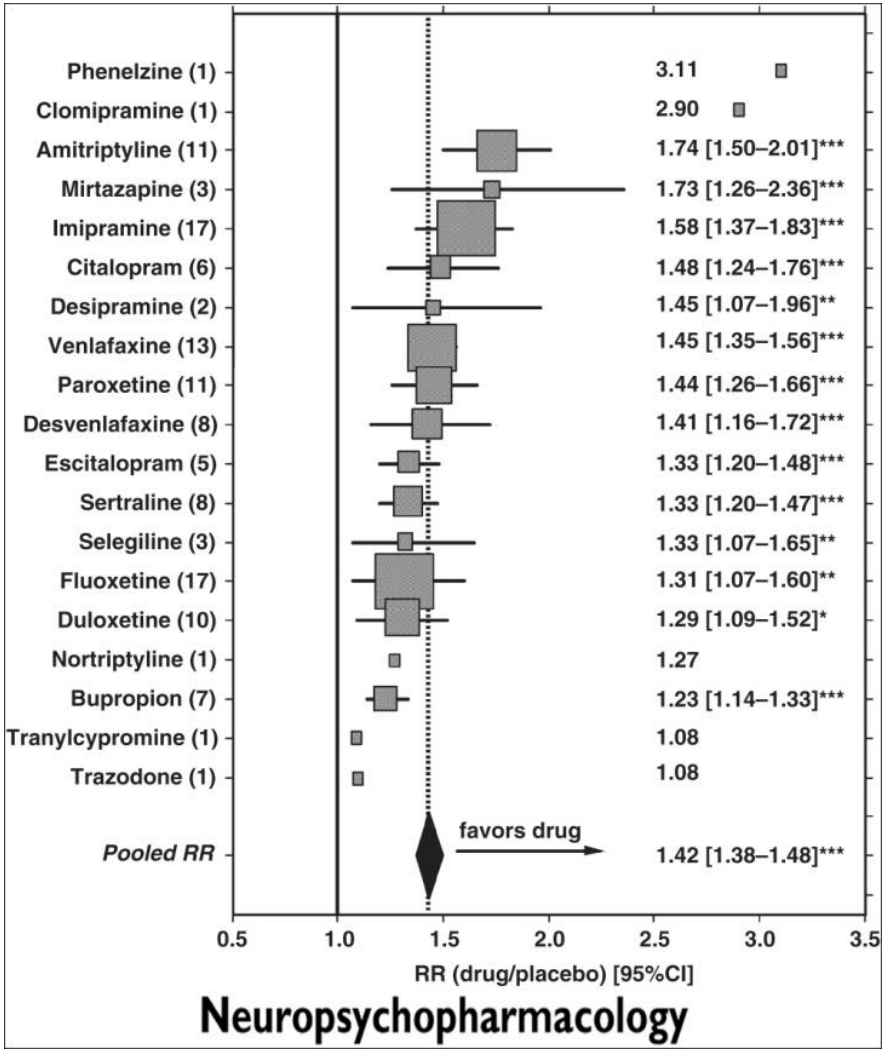
Grade of recommendation	Underlying methodology	Symbol
High	RCTs or double upgraded observational studies	****
Moderate	Downgraded RCTs or upgraded observational studies	***
Low	Double downgraded RCTs or observational studies	**
Very low	Triple downgraded RCTs or downgraded observational studies or case series/reports	*

Drug treatment trials

- High standards of regulation and compliance with procedures
- Company bias?
  - Trial design
  - Publication
- Patient populations
  - Commercial CROs
  - Generalizability
  - Tends to reduce apparent efficacy (placebo)

Neuropsychopharmacology

Figure 1

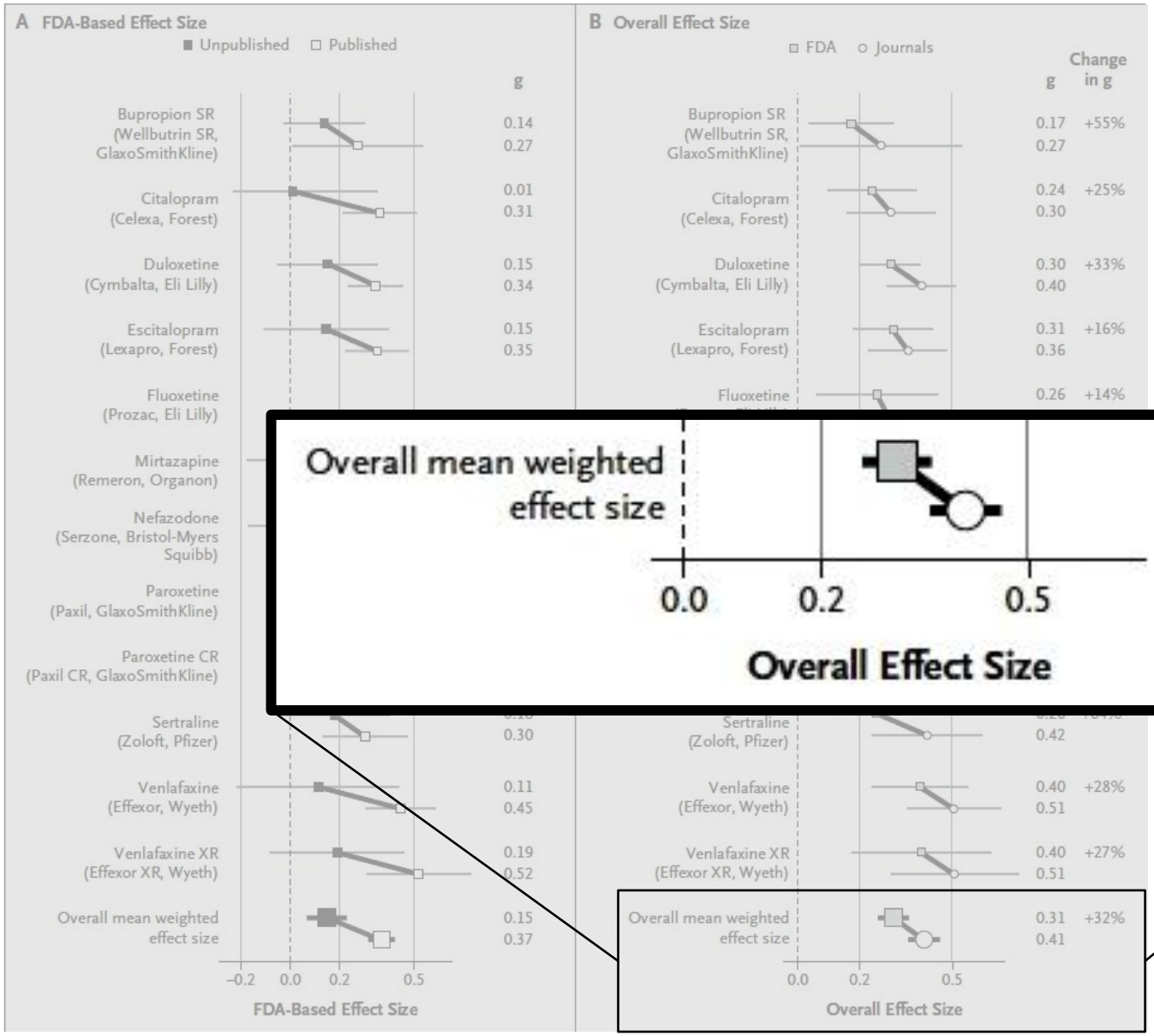


Published meta-analyses

Neuropsychopharmacology (2012) **37,** 851-864;
doi:10.1038/npp.2011.306

The publication problem

Published ± published data



Data from 74 RCTs submitted to FDA comparing published and unpublished data

Norman et al 2008

BIAS

- the key reason for this exaggeration and misrepresentation ……can be summed up in one word: bias. “This can be conscious, subconscious, or unconscious,”
- ‘publication bias,’ gives a falsely exaggerated impression of the science on a subject because not all studies that get conducted get published and the ones that do tend to have extreme results.

ESSAY

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • DOI: 10.1371/journal.pmed.0020124

- Abstract
- Modeling the Framework for False Positive Findings
- Bias
- Testing by Several Independent Teams
- Corollaries
- Most Research Findings Are False for Most Research Designs and for Most Fields
- Claimed Research Findings May Often Be Simply Accurate Measures of the Prevailing Bias
- How Can We Improve the Situation?
- References

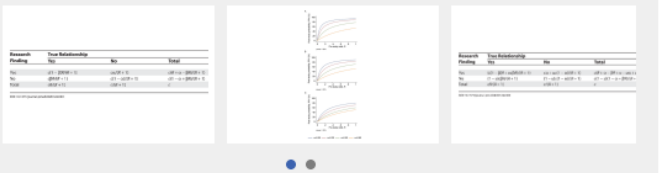
Reader Comments (31)
Media Coverage (77)
Figures

Abstract

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Figures



Citation: Ioannidis JPA (2005) Why Most Published Research Findings Are False. PLoS Med 2(8): e124. doi:10.1371/journal.pmed.0020124

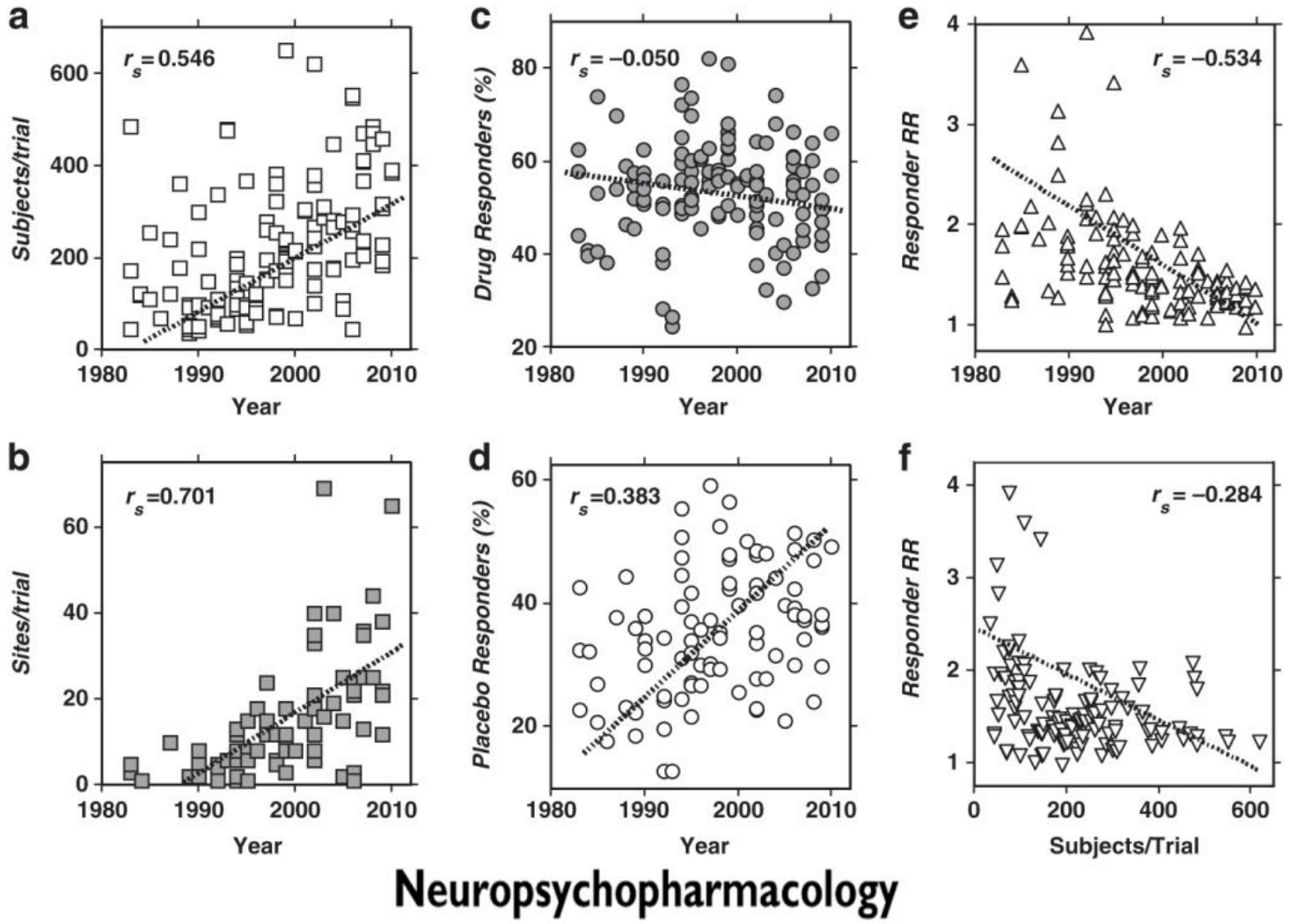
Published: August 30, 2005

Trials and effect sizes have changed over time

The drugs haven’t

Neuropsychopharmacology

Figure 2

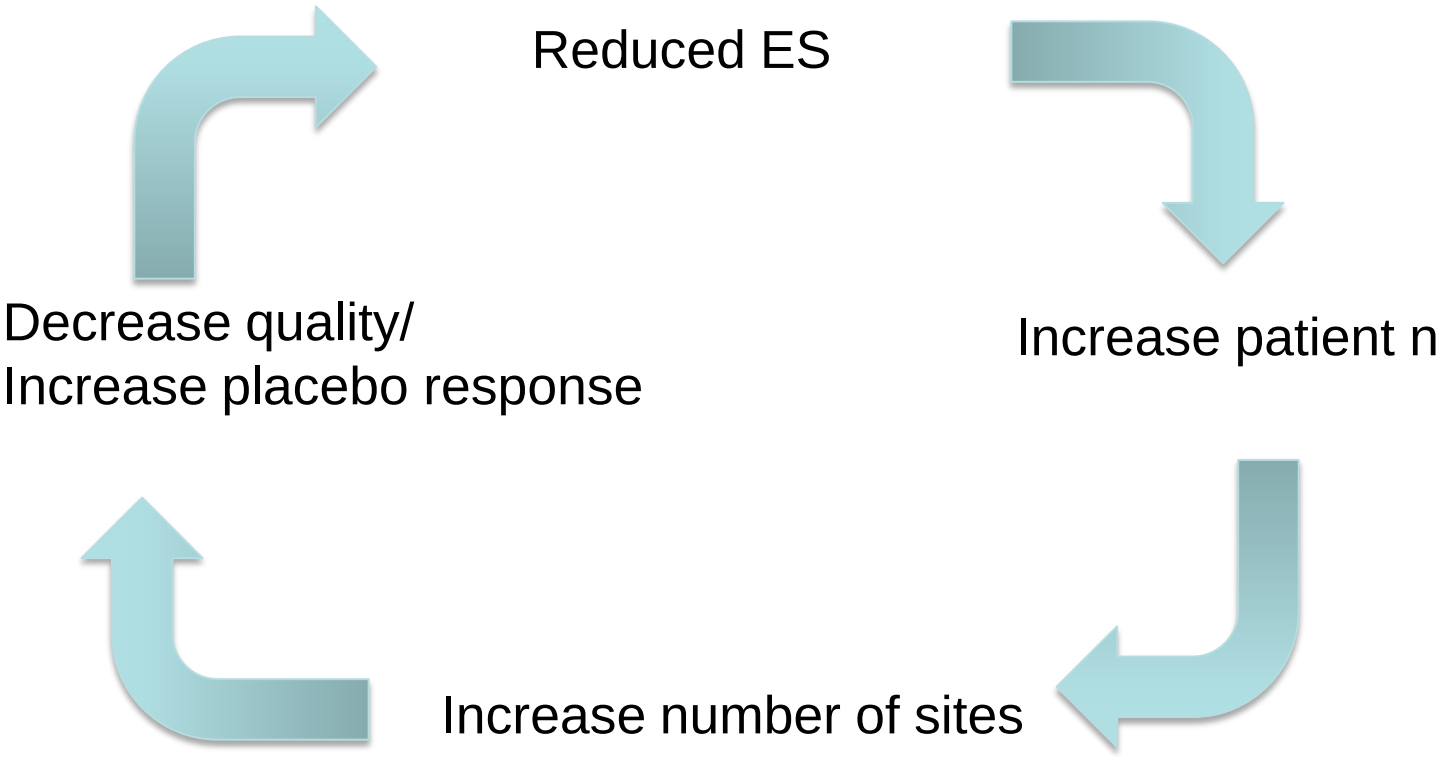


Neuropsychopharmacology (2012) **37,** 851-864;
doi:10.1038/npp.2011.306

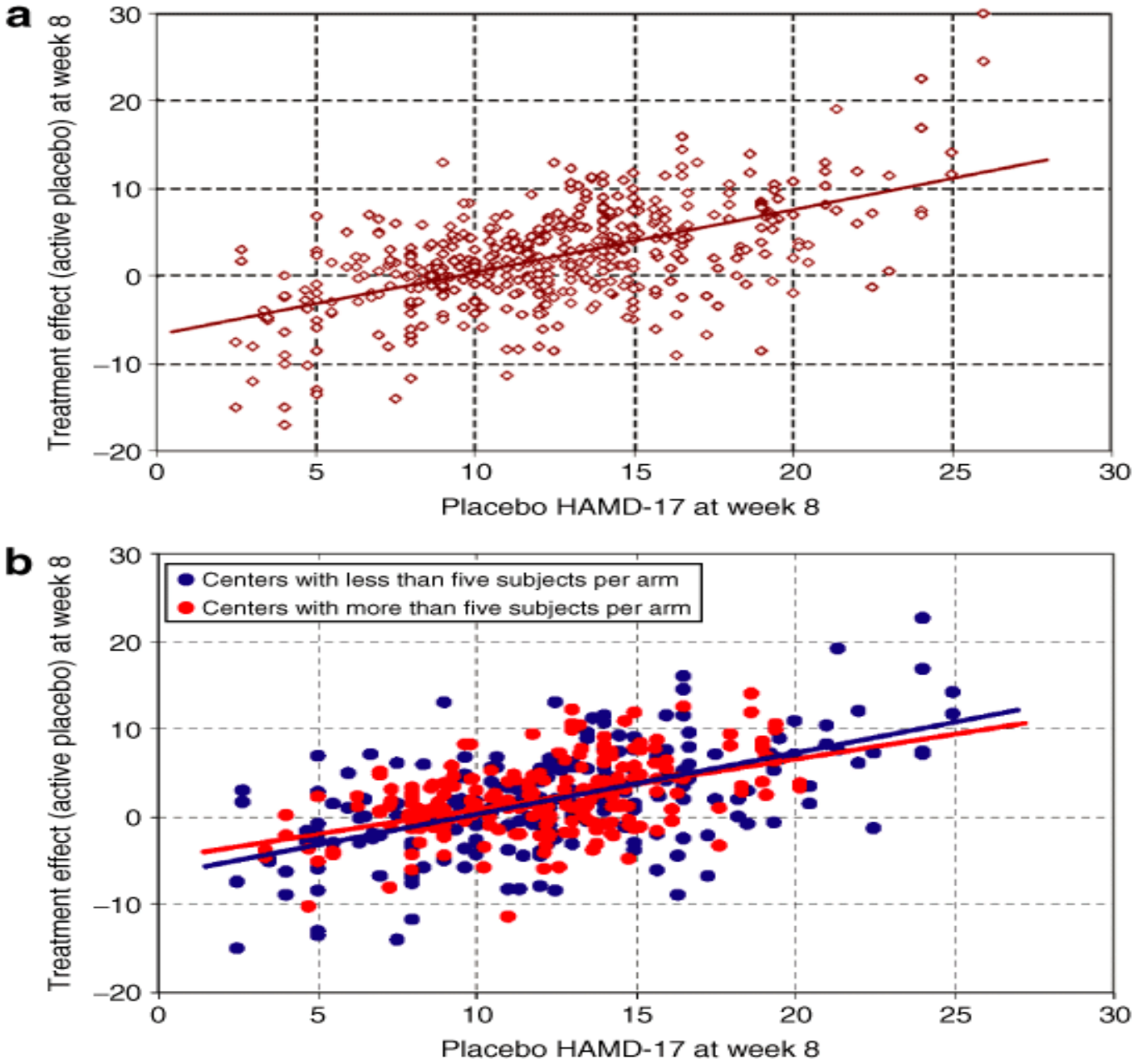
Drivers

- Commercial pressure
  - Do studies to strict time lines
- Outsourcing to ‘CROs’
- Increases in patient numbers
- Increases in trial sites

Maybe….


Reduced ES
Increase patient n
Increase number of sites
Decrease quality/
Increase placebo response

Merlo-Pich, E.
Clinical Pharmacology & Therapeutics (2008); **84**, 3, 378–384 doi:10.1038/clpt.2008.70



Conclusions

- Reduced Drug/placebo difference driven by placebo response
- Increase in placebo response driven by Site number
- Site/patient number is a perverse incentive for CROs

- A quality issue?
  - Fewer sites, ‘better’ patients

The key components of a clinical trial

- Who are the patients?
- What are the treatment options?
  - How many
  - How ethical
- Is the study randomized and the randomization concealed?
- Is the study blind?
- Are the end points sensitive

Who are the right patients?

- Age
- Severity
- Primary or secondary care
- What site?
  - CROs
- What country?

Who are the right patients?

It wont be those with the mildest illness?

From: **Antidepressant Drug Effects and Depression Severity: A Patient-Level Meta-analysis**

JAMA. 2010;303(1):47-53. doi:10.1001/jama.2009.1943

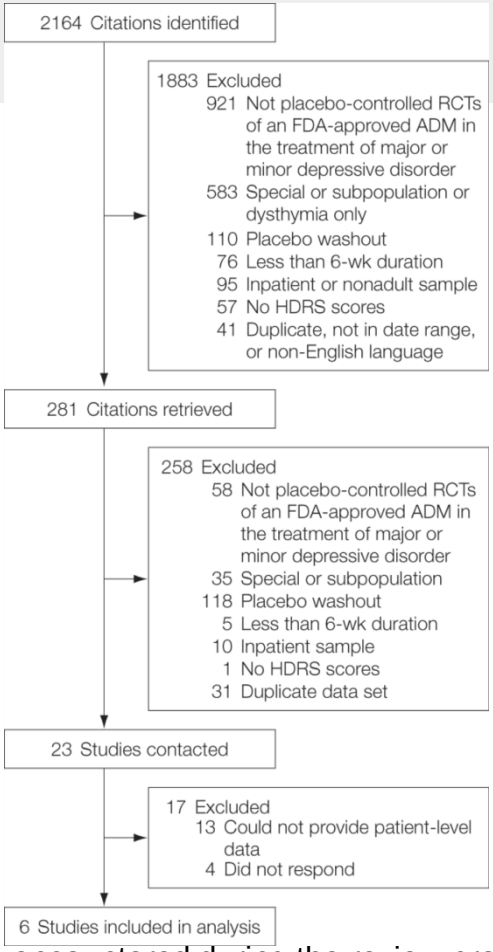


Figure Legend:

Reasons for exclusion describe the first reason for exclusion that was encountered during the review process. Several articles had multiple reasons for exclusion. RCTs indicates randomized controlled trials; FDA, US Food and Drug Administration; ADM, antidepressant medication; HDRS, Hamilton Depression Rating Scale.

From: **Antidepressant Drug Effects and Depression Severity: A Patient-Level Meta-analysis**

JAMA. 2010;303(1):47-53. doi:10.1001/jama.2009.1943

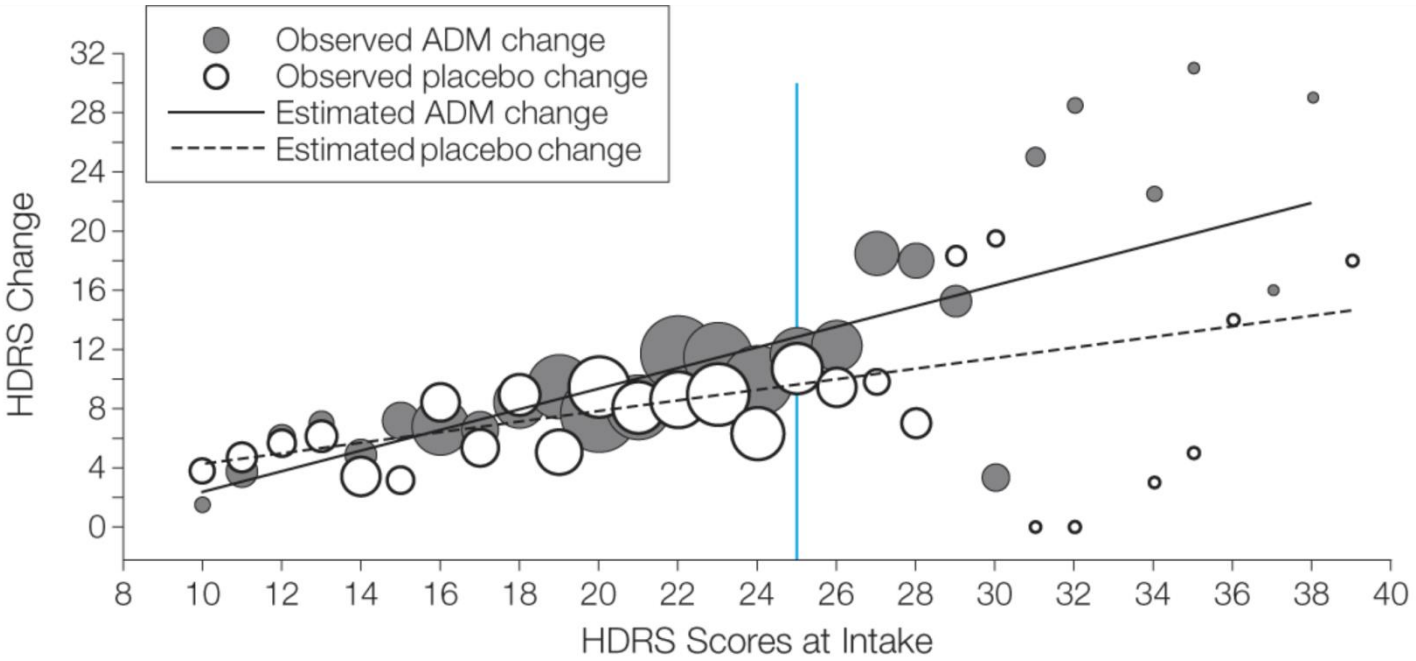


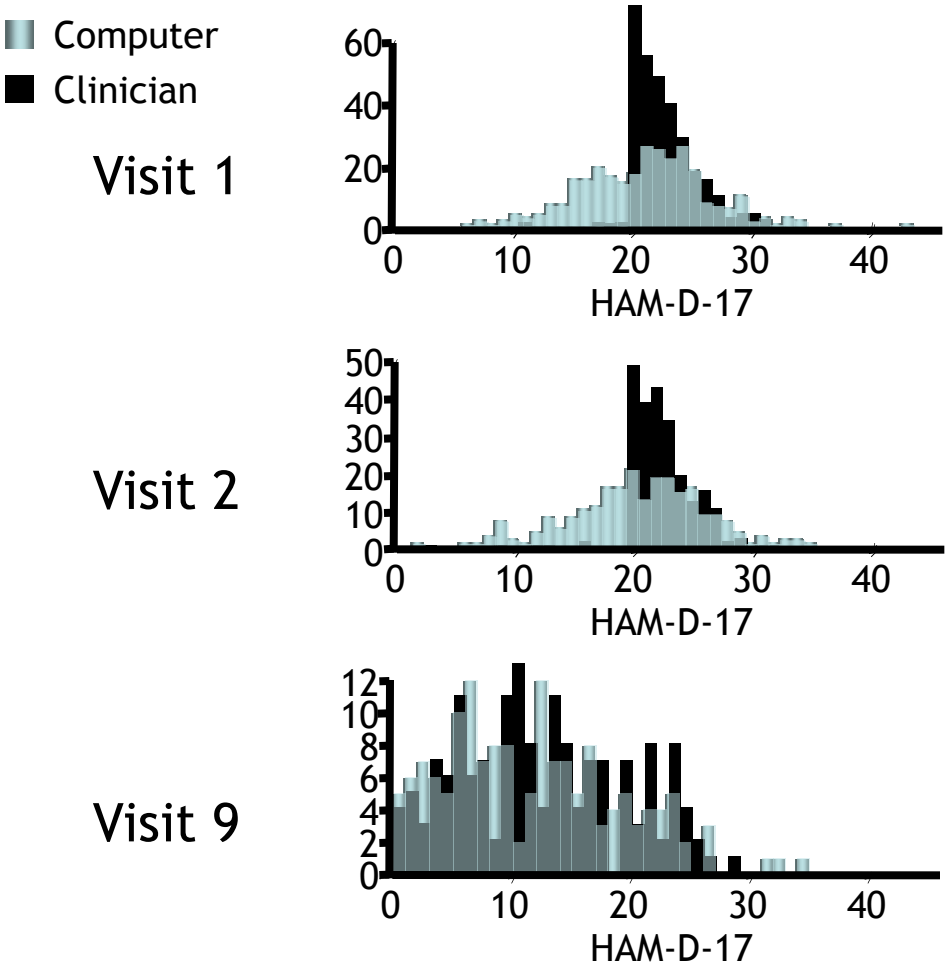
Figure Legend:

Circles represent observed (raw) mean change in depressive symptoms from intake to the end of treatment at each initial Hamilton Depression Rating Scale (HDRS) score for both the antidepressant medication (ADM) and placebo conditions. The size (area) of the circles is proportional to the number of data points that contributed to each mean. Regression lines represent estimates of change in depression symptoms from intake to end of treatment for ADM and placebo conditions as a function of baseline symptom severity. These regression lines were estimated from a model of the baseline severity × treatment interaction, controlling for the effects of the study from which the data originated. The National Institute for Clinical Excellence threshold for clinical significance (an HDRS point difference ≥3) was met for intake HDRS scores of 25 or greater, indicated by the blue line.

Practical choice of patients

- Equipoise
  - You are uncertain which treatment is best
  - The patient will not be at greater risk from one treatment versus the other
- In a placebo controlled trial
  - Milder symptoms
  - No suicidal risk
  - No children and young people
- But you still want patients to have symptoms as severe as possible?

Objective vs. Subjective HAM-D Ratings



DeBrota D, et al. Poster presented at NCDEU, 1999.

Find an assay sensitive population

Not too severe, not too mild, properly measured, real impairment

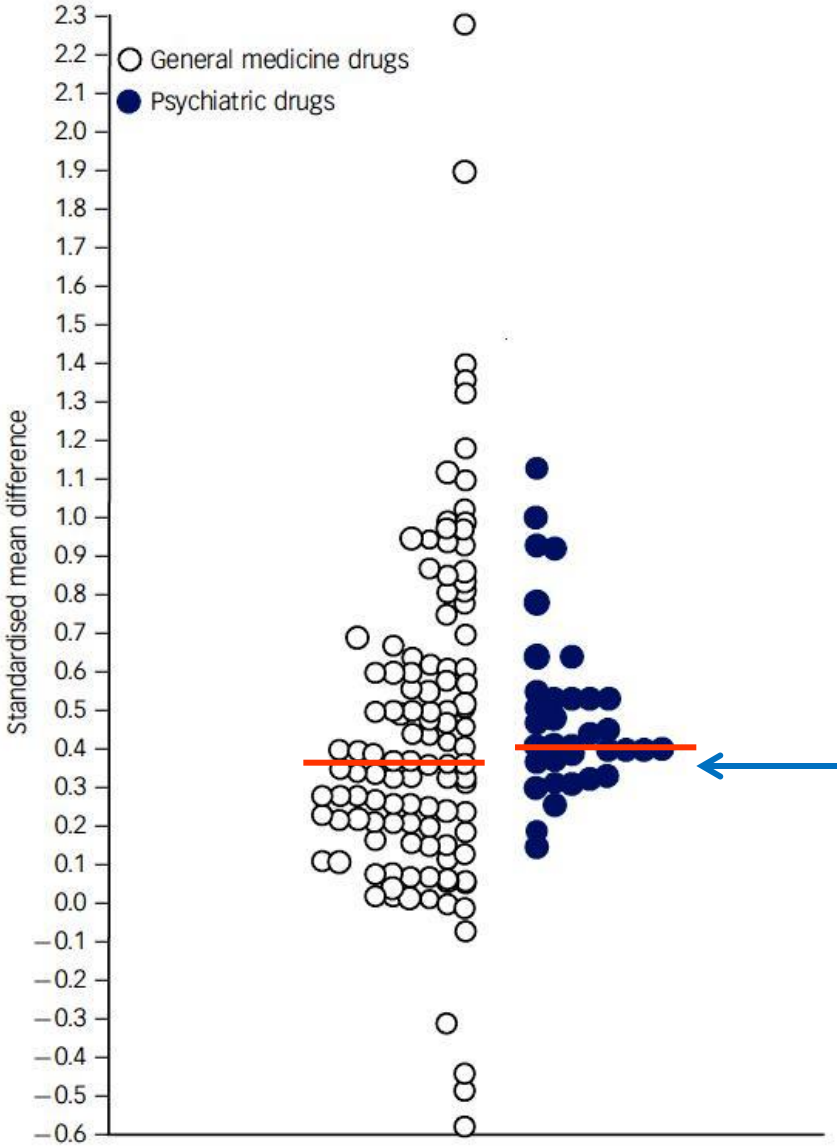
Is the study randomized and the randomization concealed?

Is the study blind?

Should we be embarrassed by all this?

Small effect sizes, placebo effects etc

Effect sizes in psychiatry and general medicine



Review of 94 meta-analyses
- 48 drugs in 20 medical diseases
-16 drugs in 8 psychiatric disorders

General equivalence of effect sizes

Leucht et al 2012 Br J Psychiat 200:97-106

Antidepressants

The status of clinical trials

- They are experiments
- The more controlled, the more artifical
- They support clinical practice, but do not define it
- Evidence based medicine is a little over reliant on the ‘double blind RCT’ as the definitive proof of everything

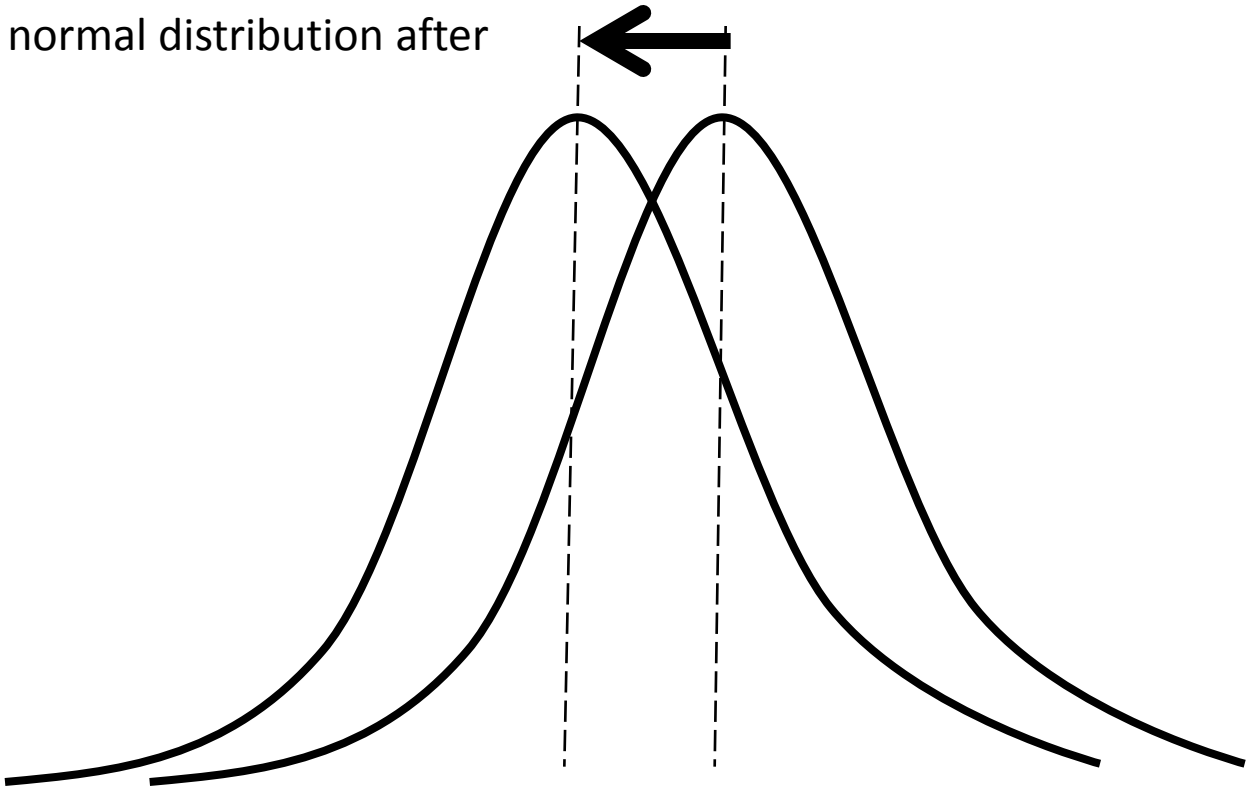
What are the implications for individual patients?

What do you say to patients about whether drugs work?

Distribution of scores and interpretation of outcome

Difference between means – how do we interpret this?

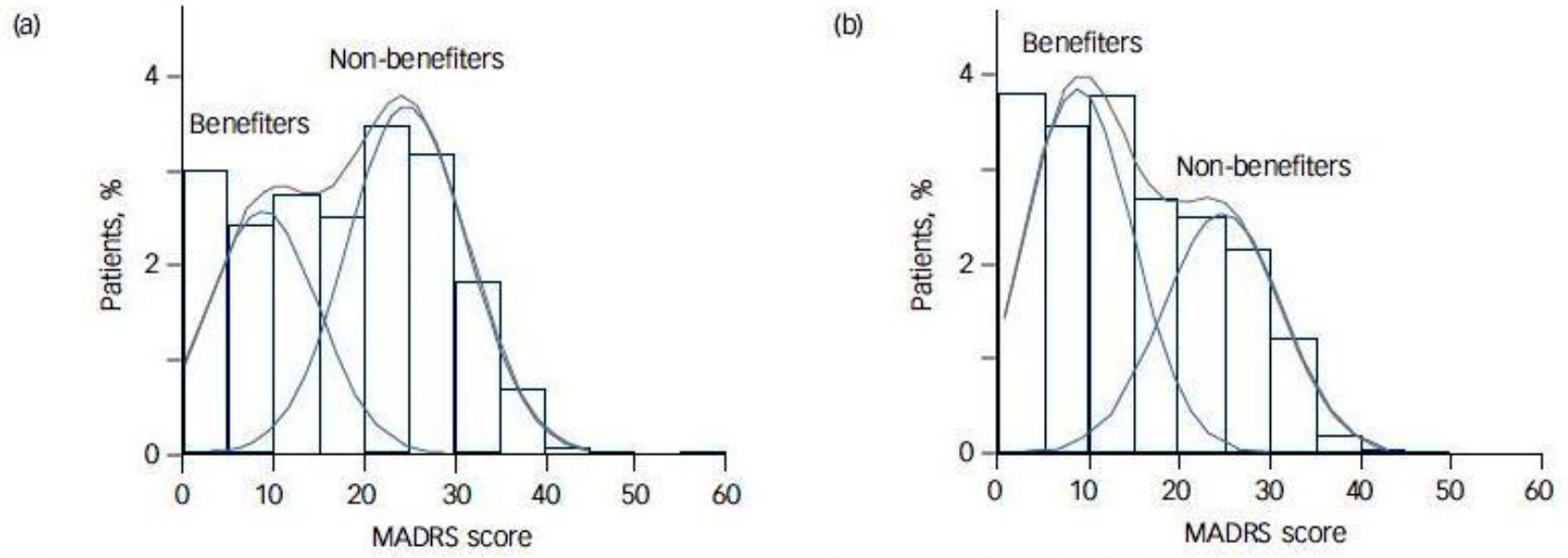
Assumption of normal distribution after treatment



Bimodality of post-treatment scores

Bimodal distribution explained 60% of variance v 6% with unimodal model

Placebo **Escitalopram**



Benefiters scores went from mean of 30 to 10

Non-benefiters scores went from mean of 30 to 25

Proportion benefiting from escitalopram and not from placebo = 19%

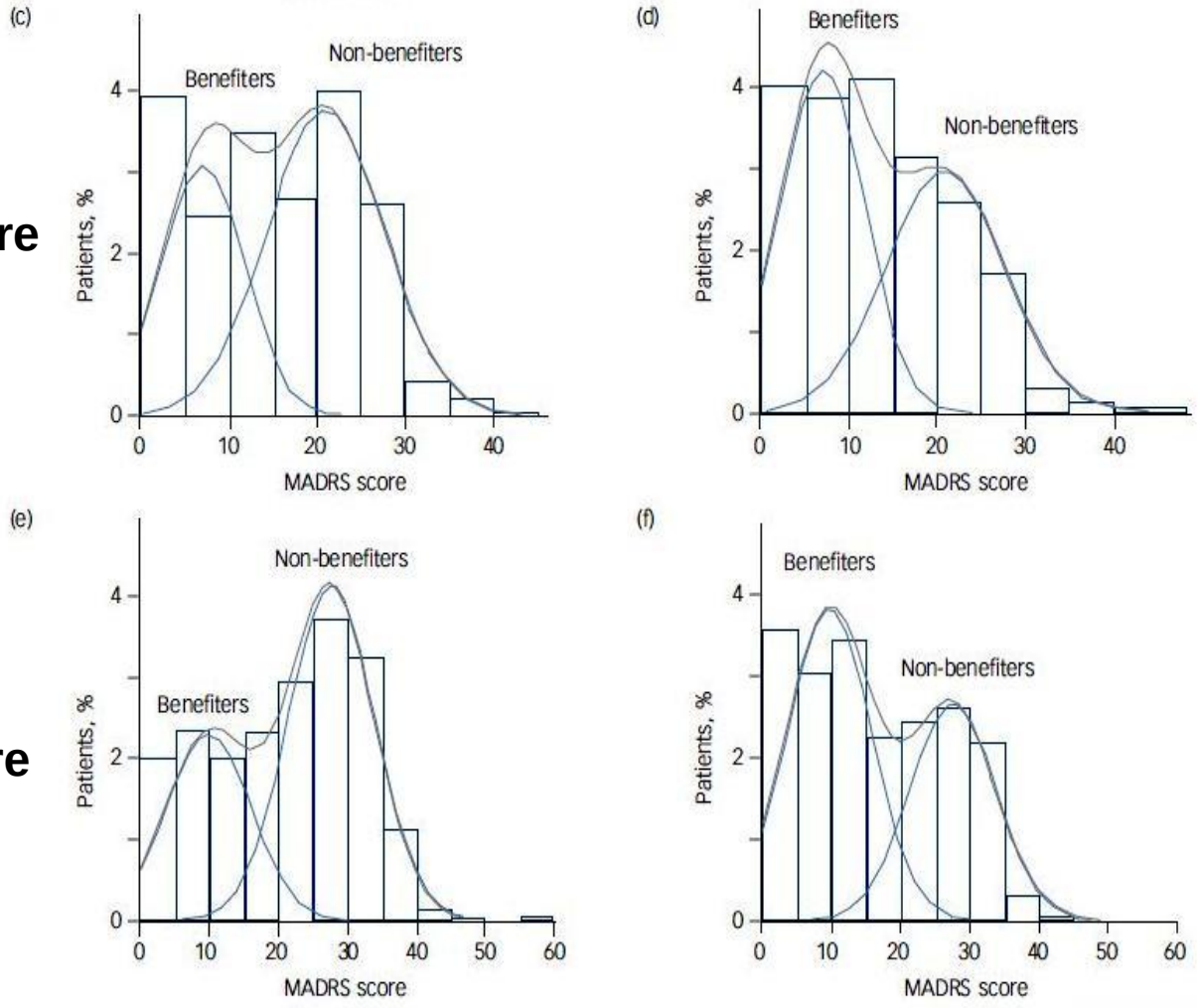
Thase et al 2011

Bimodality of post-treatment scores

Placebo **Escitalopram**

Less severe

More severe

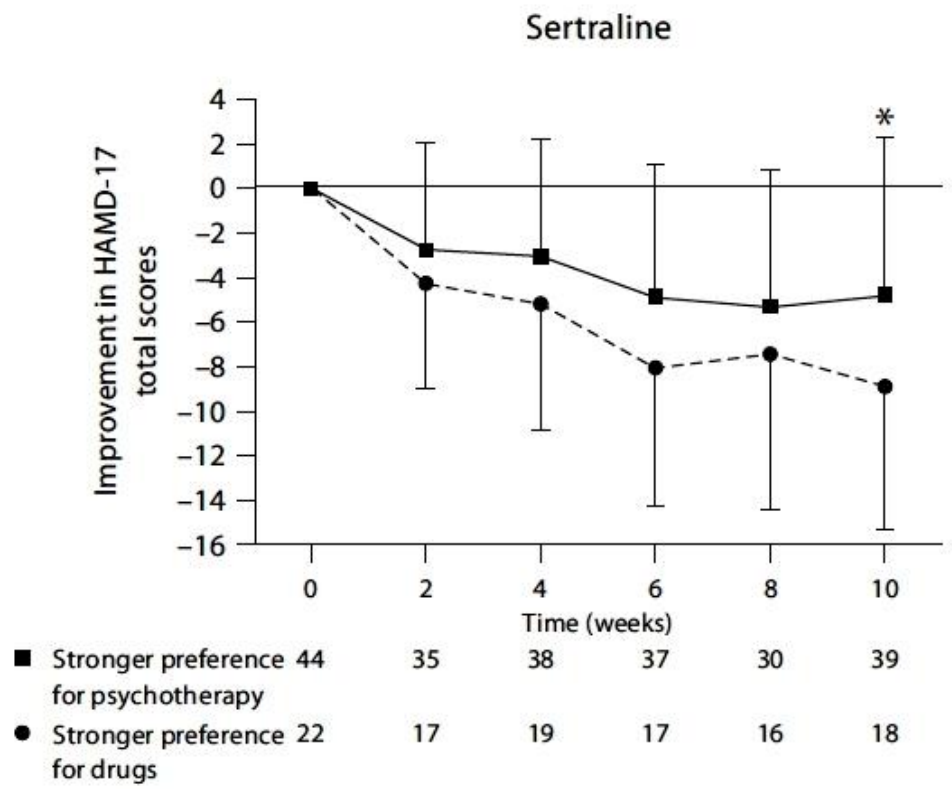


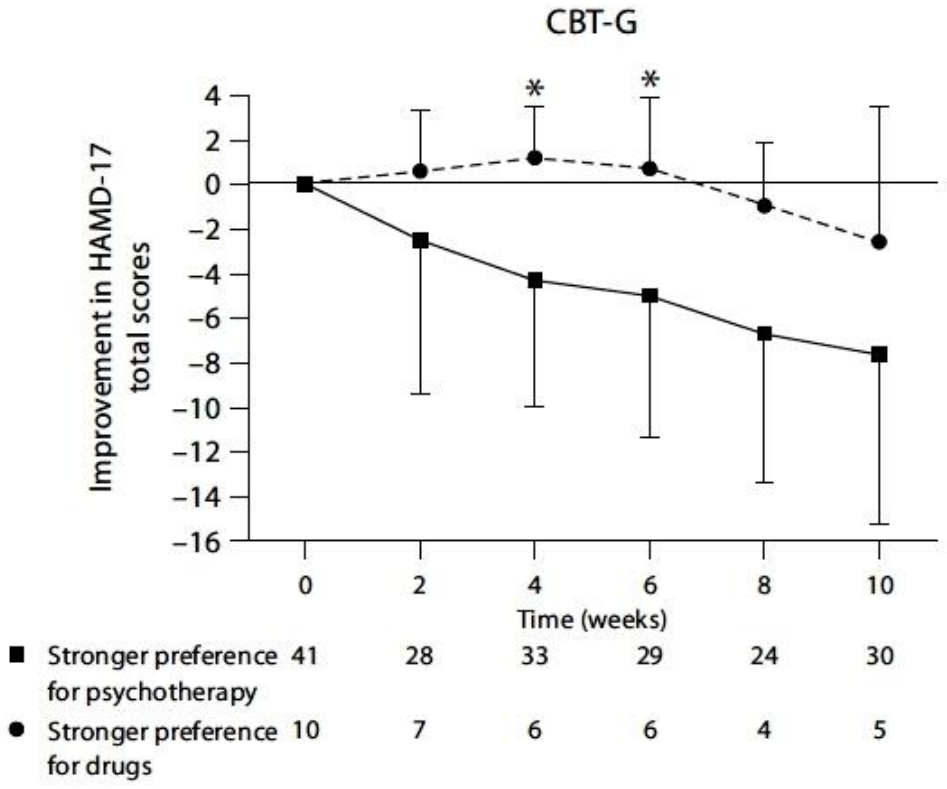
Thase et al 2011

Effect of preference on RCT outcome

Mergl et al 2011 Psychother Psychosom 80:39-47

- Primary care patients with preference determined before treatment.
- Randomised to sertraline or group CBT or patient choice





What is the ‘placebo response’

- Baseline rating inflation
- Regression to the mean
- Spontaneous recovery
- Intensity of follow-up and assessment
- The attributional properties of an inert pill
  - Patient expectations
  - Doctor expectations

Implications for individual patients

- Drug response on average about 20% greater than placebo
- Drug response is more often a complete response/remission
- How do you enhance the placebo response?
  - Clarity about the drug and its effects/side effects
  - Enhance expectations of treatment
  - Listen to the patient

Reasons to be cheerful

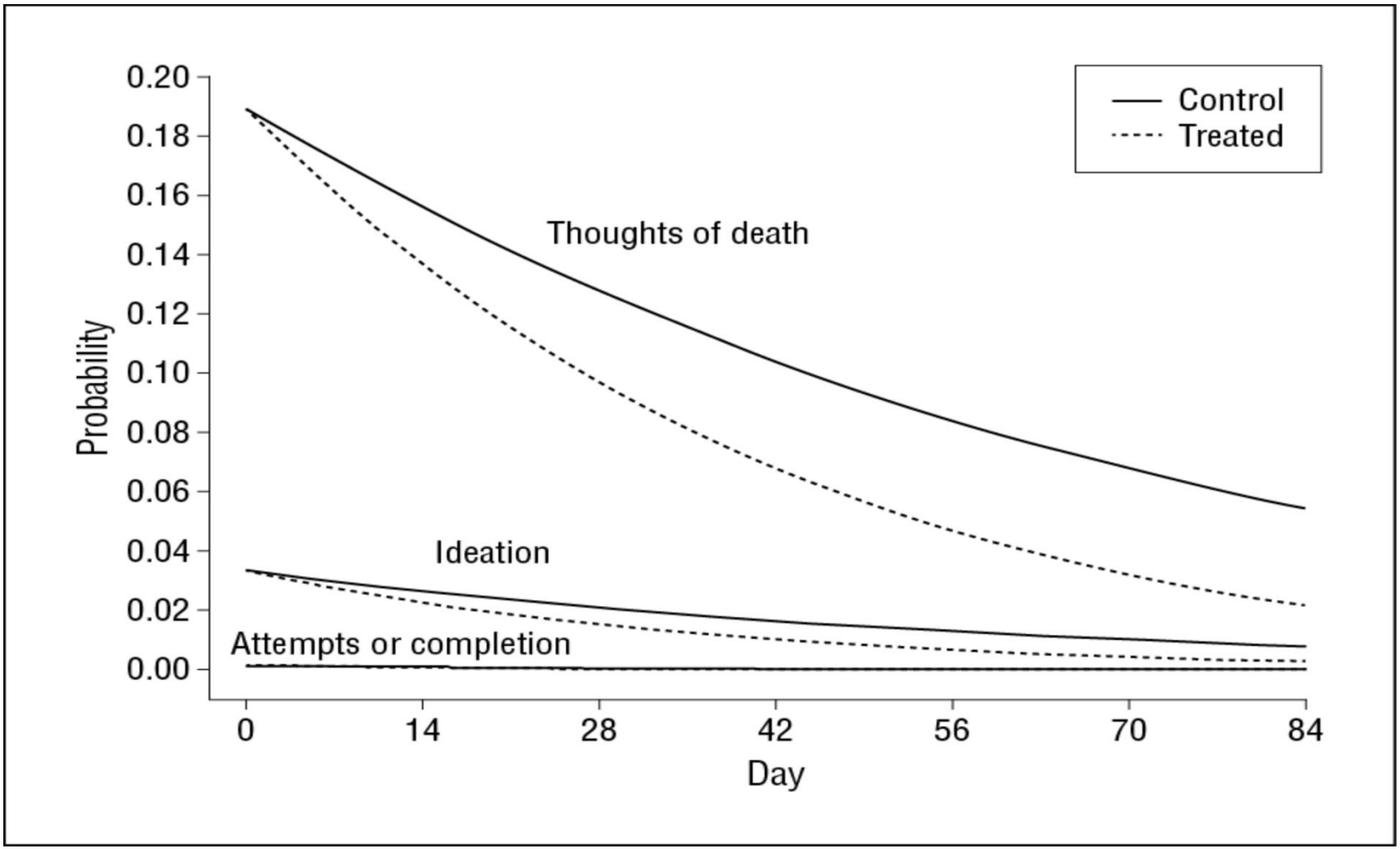
New findings on safety of antidepressants

- Suicide
- Foetal malformations

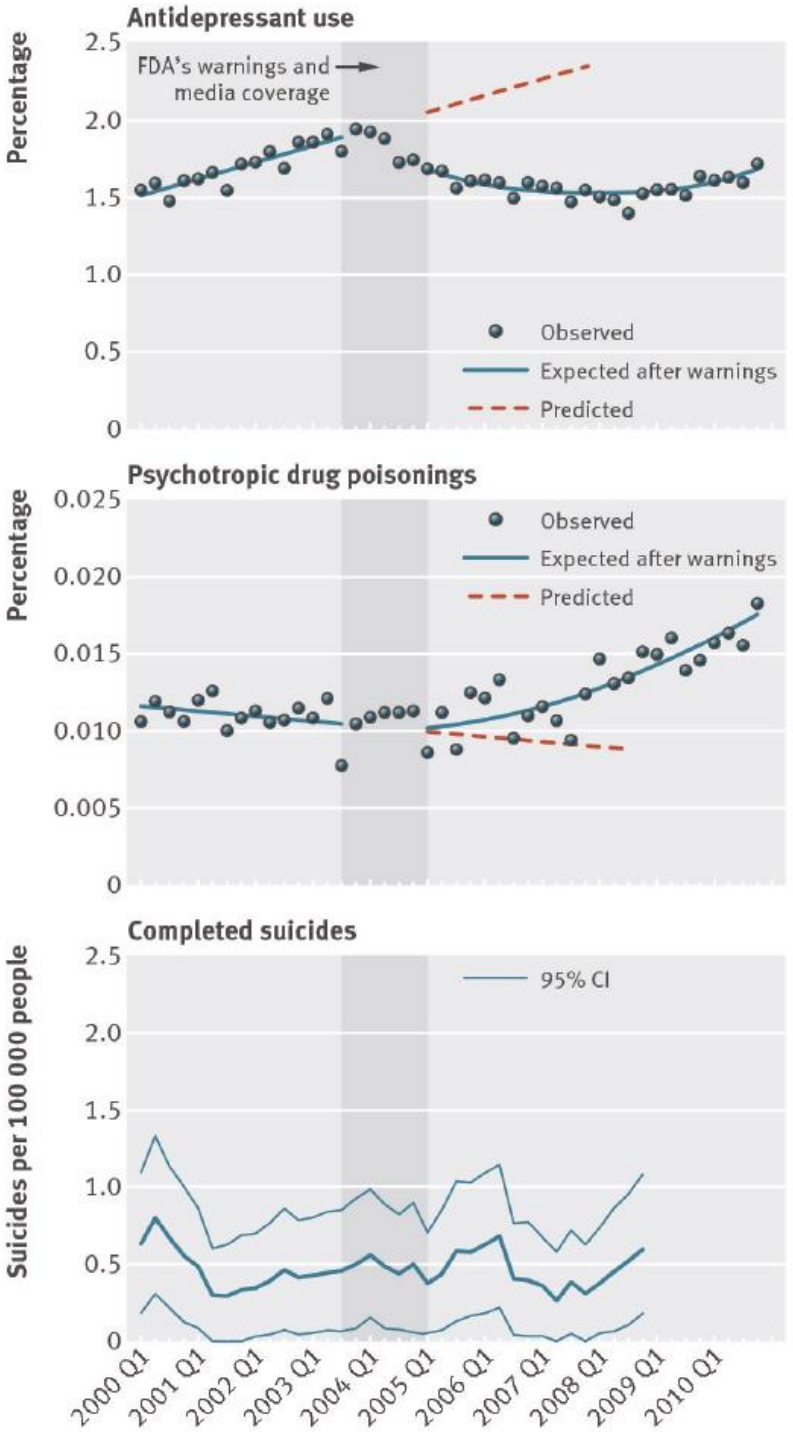
Suicidal Thoughts and Behavior With Antidepressant Treatment: Reanalysis of the Randomized Placebo-Controlled Studies of Fluoxetine and Venlafaxine

Gibbons et al *Arch Gen Psychiatry.* 2012;69(6):580-587. doi:10.1001/

9185 patients (fluoxetine: 2635 adults, 960 geriatric patients, 708 youths; venlafaxine: 2421 adults with IR venla and 2461 adults with extended-release venlafaxine) for a total of 53 260 person-week observations.

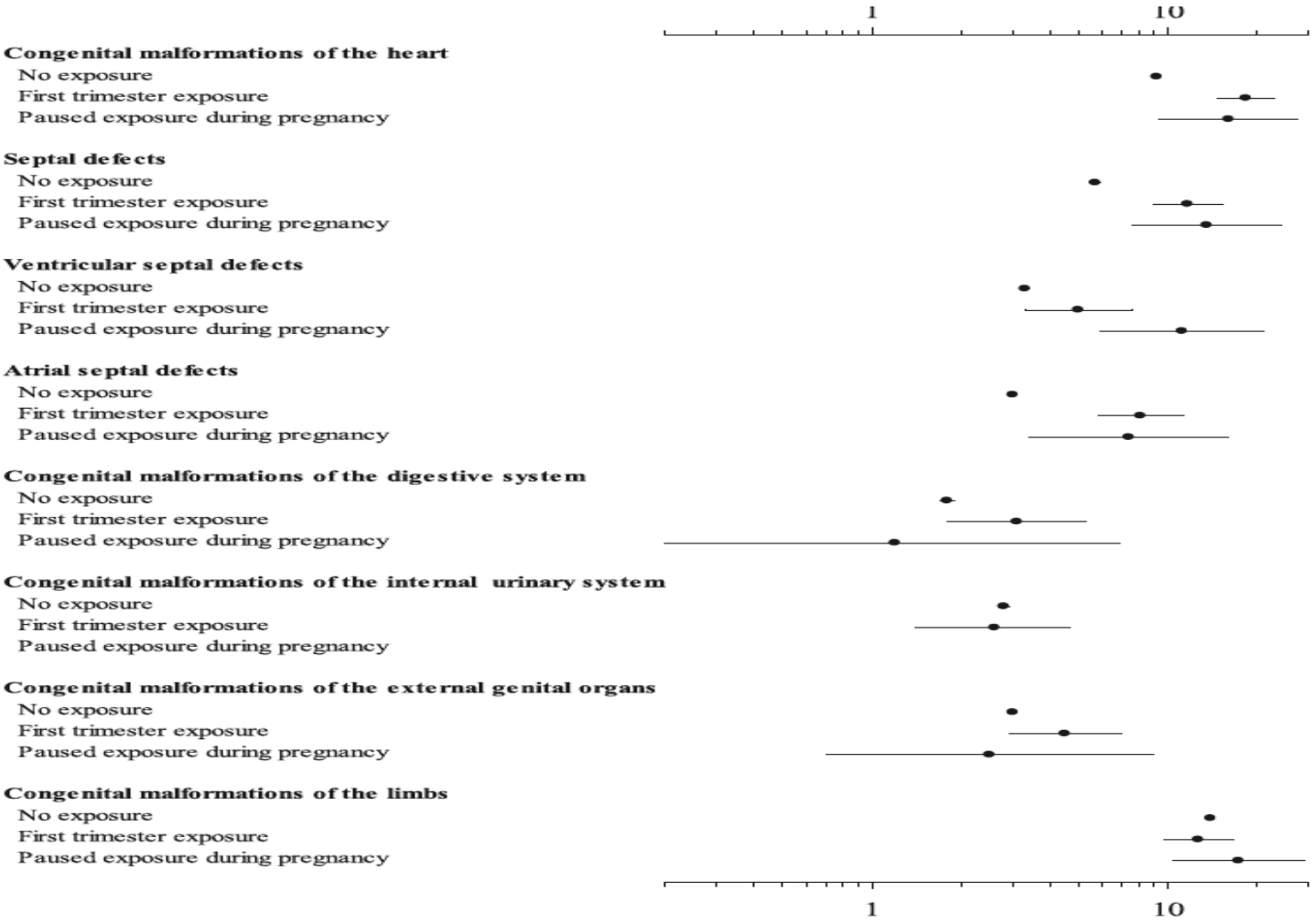


Lu et al BMJ 2014, 348, on line
1million adolescents



Foetal malformation and SSRIs

Jimenez-Solem et al. BMJ Open 2012;
2:e001148. doi:10. 1136/bmjopen-2012-001148



OPEN

Molecular Psychiatry (2015), 1–8

www.nature.com/mp

ORIGINAL ARTICLE

Consistent superiority of selective serotonin reuptake inhibitors over placebo in reducing depressed mood in patients with major depression

F Hieronymus[1], JF Emilsson[1], S Nilsson[2] and E Eriksson[1]

18 out of 32 trials ‘failed’ on HAM-D comaprison

3 out of 32 trials ‘failed’ on depressed mood item

Table 3. Effect sizes and *P*-values for different measures of efficacy in the pooled population.

Measure of efficacy (scoring range)	*Baseline mean (s.d.)*	*Pooled effect size*	*Pooled analysis* P-*value*[a]	*Pooled analysis* P-*value*[b]
HDRS-17-sum (0–52)	23.1 (3.7)	0.27 (0.32[c])	< **0.001**	
Individual items				
Depressed mood (0–4)	2.8 (0.6)	0.40 (0.44[c])	< **0.001**	< **0.001**
Feelings of guilt (0–4)	1.7 (0.7)	0.26	< **0.001**	< **0.001**
Suicide (0–4)	1.1 (0.9)	0.22	< **0.001**	< **0.001**
Insomnia, early (0–2)	1.2 (0.8)	0.08	**0.005**	**0.002**
Insomnia, middle (0–2)	1.3 (0.8)	0.07	**0.009**	**0.005**
Insomnia, late (0–2)	1.2 (0.8)	0.13	< **0.001**	< **0.001**

HDRS-17-sum (0–52) 23.1 (3.7) 0.27 (0.32[c])

Individual items

Depressed mood (0–4) 2.8 (0.6) 0.40 (0.44[c])
Feelings of guilt (0–4) 1.7 (0.7) 0.26
Suicide (0–4) 1.1 (0.9) 0.22
Insomnia, early (0–2) 1.2 (0.8) 0.08
Insomnia, middle (0–2) 1.3 (0.8) 0.07
Insomnia, late (0–2) 1.2 (0.8) 0.13

Conclusions

- Placebo
  - The antidepressant controversy
- The key components of successful clinical trials
- What it means for individual patients
- Reasons to be cheerful