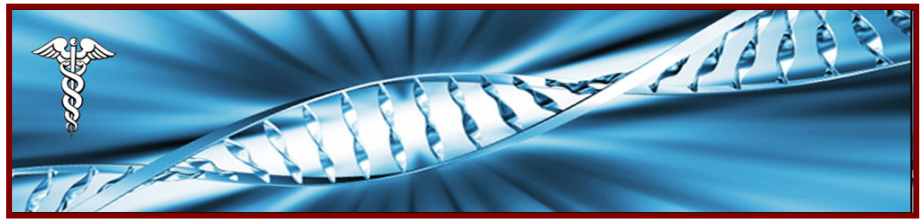


TransMedRi Workshop in Biostatistics: Critical evaluation of statistical analysis in scientific paper

Introduction to Meta-Analysis

Daniele Fanelli

ISSTI
THE INSTITUTE FOR THE STUDY OF SCIENCE TECHNOLOGY AND INNOVATION

THE UNIVERSITY OF EDINBURGH

What is meta-analysis?

- Invented in 1977 by Smith & Glass, to prove the efficacy of psychotherapy
  - (dozens of meta-analyses later, the debate is still not over!)
- Increasing popularity and diffusion

- Combining results of multiple studies around a specified research question
  - To get an overall estimate of the effect
- The most precise form of literature review, one step above a systematic review

- But a meta-analysis is much more….

What is meta-analysis for

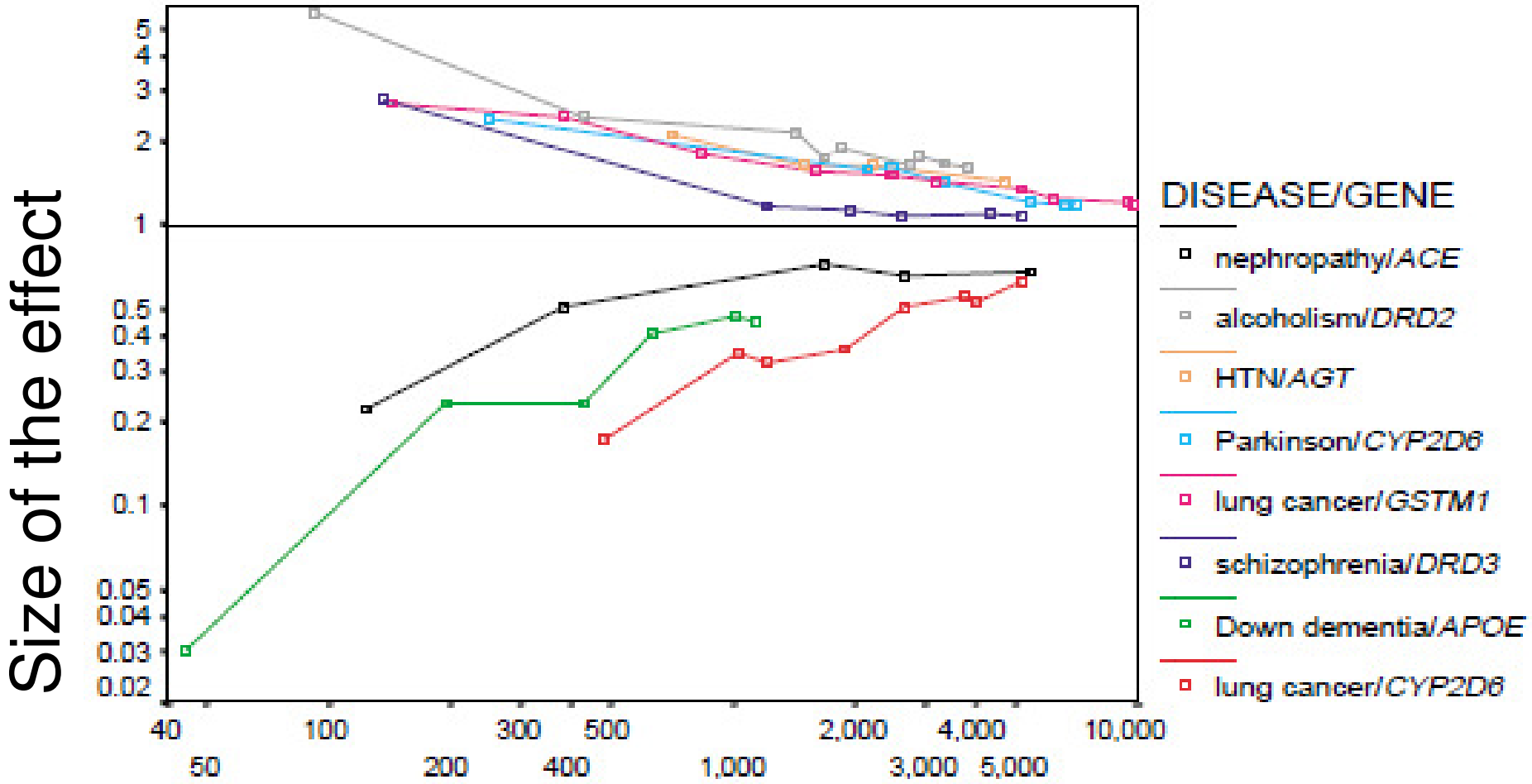
Health effects of soft drinks, juice and milk

Conclusion	Funding Source[a]			Exact *p* for Trend
	All Industry Benefit[b]	No Industry	All Industry Antagonism[c]	
Favorable	14	24	0	0.037
Neutral	5	8	1	—
Unfavorable	3	20	1	—

No fancy statistics, but accuracy in collecting studies

(Lesser et al. 2007, PLoS Medicine)

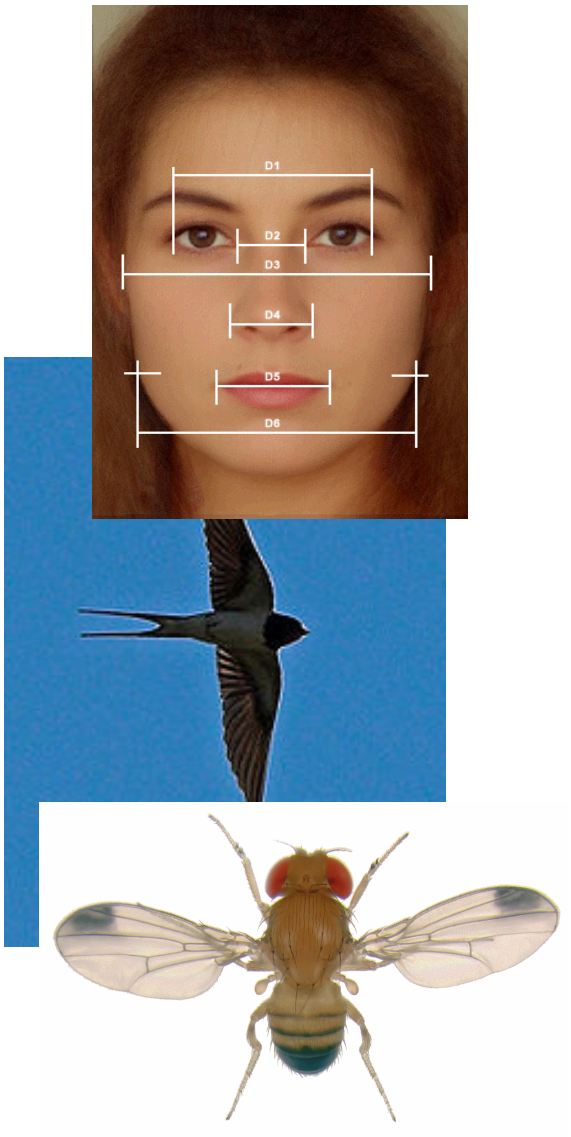
What meta-analysis is for

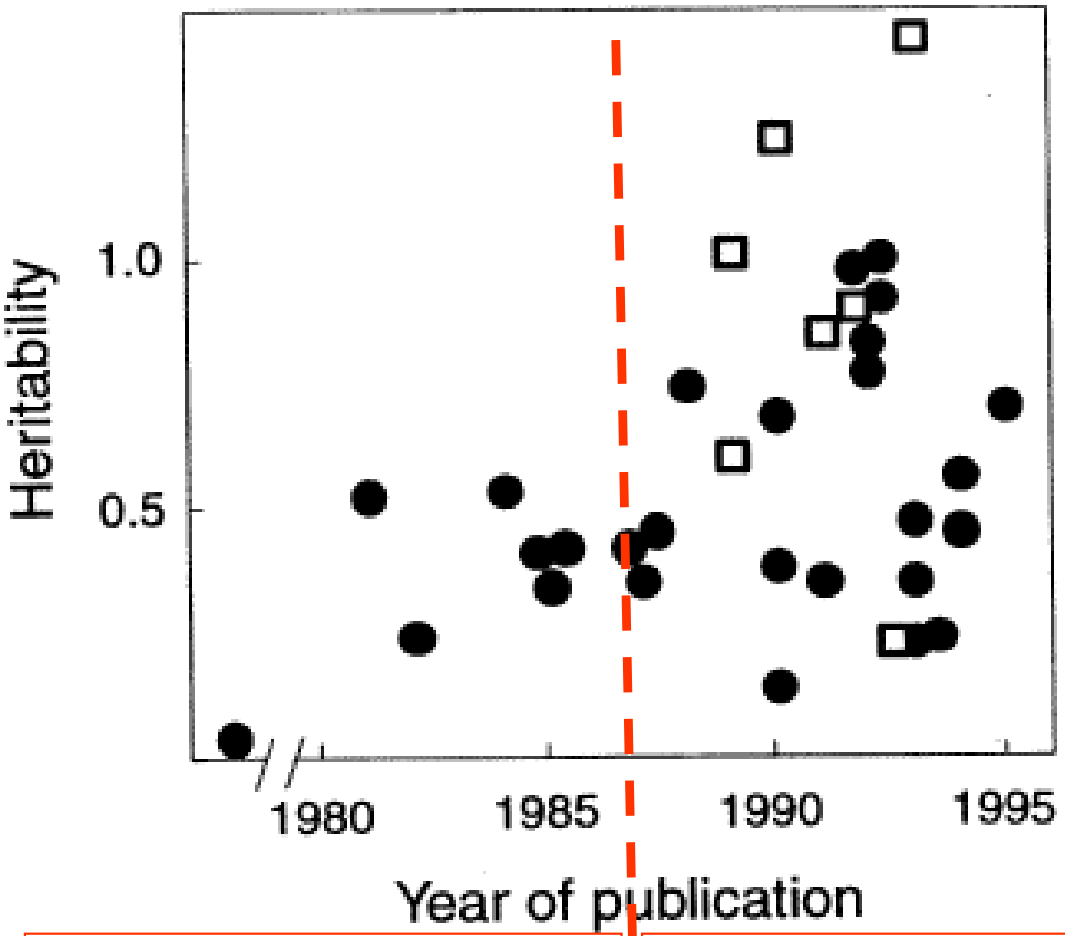


Total number of subjects (studies)

(Ioannidis et al. 2001, Nature Genetics)

What meta-analysis is for





Low H predicted | **High H possible**

(Alatalo et al. 1997, Nature)

Meta-analysis is cool

- Higher-level thinking about evidence
- An accurate way to examine the evidence, and the science that produced it
- And get conclusive answers (hopefully)

- BUT meta-analysis is not magic
  - (remember the psychotherapy debate?)
- Main criticisms
  - Apples and oranges
  - Subjectivity and biases
- GIGO rule still applies
  - Meta-analysis is only as good as what goes in it
  - It has unavoidable limits and problems that must be acknowledged

In this seminar

- Overview of meta-analysis procedures
- Risks/caveats

Phases of a meta-analysis

- Problem specification
- Study search, retrieval and selection
- Effect sizes
- Coding of studies
- Analysis
  - Overall effect
  - Homogeneity analysis
  - Further analyses (regression, sensitivity, splitting, publication bias, etc…)
- Interpretation

Problem specification

- Needs to be precise
  - E.g. Bad = How many scientists cheat?
  - Better = How many scientists admit to fabrication/falsification of data or results, when asked in surveys?
- Will guide specification of eligibility criteria for:
  - Study characteristics
    - Type of study (e.g. anonymous survey, not asking about opinions)
    - Subjects (e.g. respondents involved in publishable research)
    - Research design/specific outcome (e.g. must have “none” category)
    - Range of studies: time, geographical/linguistic/cultural range
    - Other relevant characteristics (e.g. specific funding source)
    - Type of publication
- Ideally, all this should be specified in advance
  - In practice, must adapt to what you find in the literature
    - **Risk of bias**
    - **Ensure transparent documentation of study retrieval, inclusion criteria, selection**

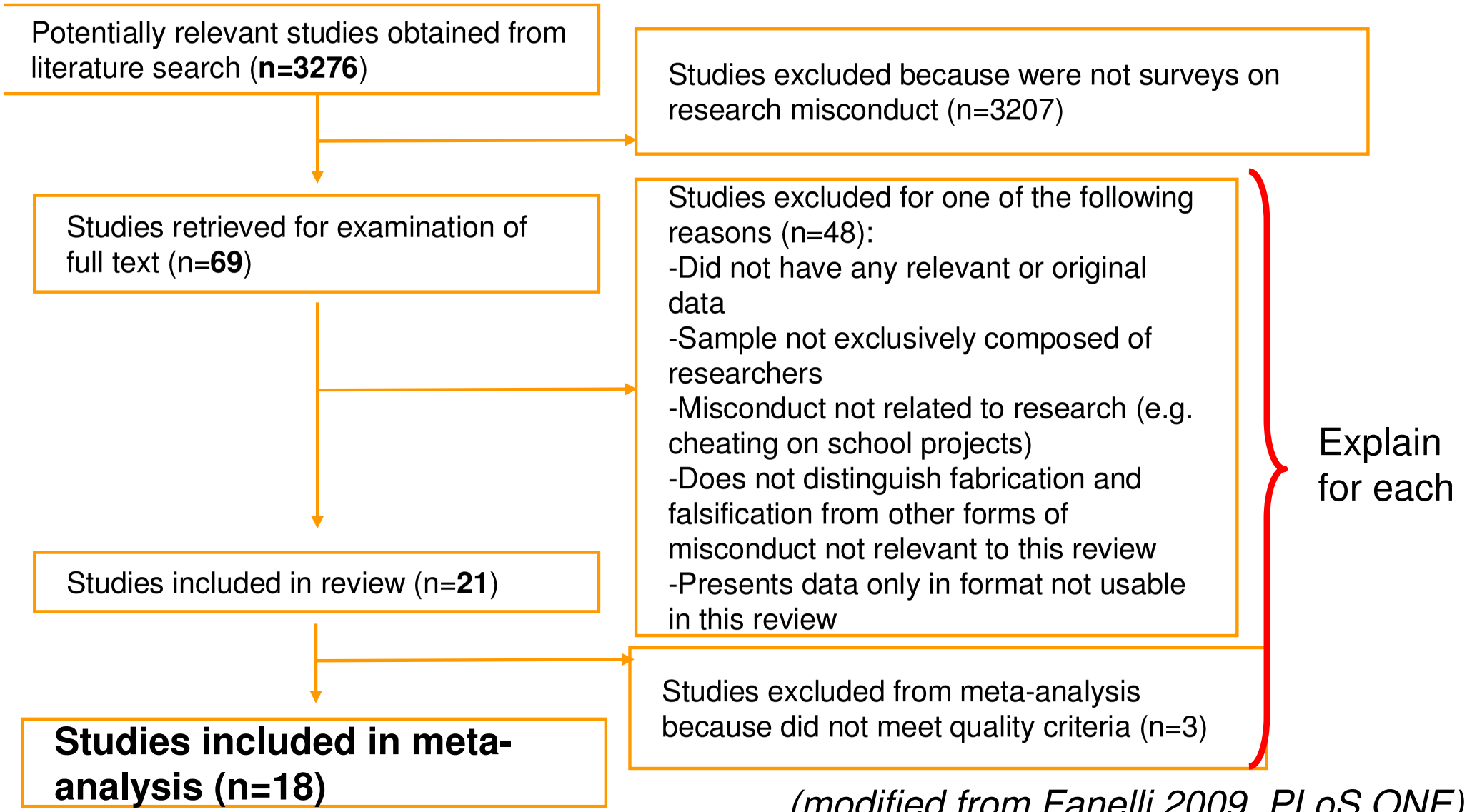
Study search and retrieval

- Ideally, you should have all studies ever made
- In practice, do as thorough a search as possible, compromising between breadth and time
  - Electronic databases (keywords compromising between generic and specific terms)
  - Examine key reviews and journals
  - Conferences for papers/authors of interest
  - Grey literature databases (Government, PhD theses)
- Come up with a complete list of potentially interesting papers
  - Title+abstract -> Select potentially interesting studies (again, compromise) to retrieve
  - Text ->include/exclude papers based on specified criteria

Flow diagram

"research misconduct" OR "research integrity" OR "research malpractice" OR "scientific fraud" OR "fabrication, falsification" OR "falsification, fabrication"

42 literature databases, 14 journals, 8 grey literature databases, 2 internet scientific search engines, and references lists



(modified from Fanelli 2009, PLoS ONE)

Effect size

- The fundamental difference between M-A & Sys. Rev
  - Many stop at the level of systematic review!
- A standardized measure of outcome
  - Could be yes/no, favourable/unfavourable, P-values
  - Best when encodes direction+magnitude
- Weighted by a measure of precision
  - Could be anything, but statisticians have concluded that…
  - Inverse variance weight: $1/SE^2$
  - SE calculated by statistical theory
  - SE formula not available for all statistical outcomes (e.g. multivariable)
    - Proportion, mean, pre-/post-, independent groups, correlation
    - Extract/convertible from common stats (t, chi^2, F, regression…)
- The “art” of meta-analysis is identifying the proper research question and strategy to get a reliable, standardized effect size across studies.
  - E.g. the number of scientists that do not reply 0, “never”, “none” etc… in surveys on misconduct

Effect sizes calculations

Table 3.2

Effect Size, Standard Error and Inverse Variance Weight Formulas for each Effect Size Type

Effect Size Type	*Effect Size Statistic*	*Standard Error*	*Inverse Variance*
	One Variable Relationships–Central Tendency Description		
Proportion—direct method	$ES_p = p = \frac{k}{n}$	$SE_p = \sqrt{\frac{p(1-p)}{n}}$	$w_p = \frac{n}{p(1-p)}$
Proportion—logit method	$ES_l = \log_e\left[\frac{p}{1-p}\right]$	$SE_l = \sqrt{\frac{1}{np} + \frac{1}{n(1-p)}}$	$w_l = np(1-p)$
Arithmetic mean	$ES_m = \overline{X} = \frac{\Sigma x_i}{n}$	$SE_m = \frac{s}{\sqrt{n}}$	$w_m = \frac{n}{s^2}$
	Two Variable Relationships—Pre–Post Contrasts		
Mean gain—unstandardized	$ES_{ug} = \overline{X}_{T2} - \overline{X}_{T1} = \overline{G}$	$SE_{ug} = \sqrt{\frac{2s_p^2}{n}(1-r)} = \sqrt{\frac{s_g^2}{n}}$	$w_{ug} = \frac{n}{2s_p^2(1-r)}$
Mean gain—standardized	$ES_{sg} = \frac{\overline{X}_{T2} - \overline{X}_{T1}}{s_p} = \frac{\overline{G}}{s_g/\sqrt{2(1-r)}}$	$SE_{sg} = \sqrt{\frac{2(1-r)}{n} + \frac{ES_{sg}^2}{2n}}$	$w_{sg} = \frac{2n}{4(1-r) + ES_{sg}^2}$
	Two Variable Relationships—Group Contrasts		
Mean difference—unstandardized	$ES_{um} = \overline{X}_{G1} - \overline{X}_{G2}$	$SE_{um} = s_p\sqrt{\frac{1}{n_{G1}} + \frac{1}{n_{G2}}}$	$w_{um} = \frac{n_{G1}n_{G2}}{s_p^2(n_{G1} + n_{G2})}$
Mean difference—standardized	$ES_{sm} = \frac{\overline{X}_{G1} - \overline{X}_{G2}}{s_p}$ $ES'_{sm} = \left[1 - \frac{3}{4N-9}\right]ES_{sm}$	$SE_{sm} = \sqrt{\frac{n_{G1}+n_{G2}}{n_{G1}n_{G2}} + \frac{(ES'_{sm})^2}{2(n_{G1}+n_{G2})}}$	$w_{sm} = \frac{2n_{G1}n_{G2}(n_{G1}+n_{G2})}{2(n_{G1}+n_{G2})^2 + n_{G1}n_{G2}ES_{sm}^2}$
Proportion difference	$ES_{pd} = p_{G1} - p_{G2}$	$SE_{pd} = \sqrt{p(1-p)\left(\frac{1}{n_{G1}} + \frac{1}{n_{G2}}\right)}$	$w_{pd} = \frac{n_{G1}n_{G2}}{p(1-p)(n_{G1}+n_{G2})}$
Logged odds-ratio	$ES_{LOR} = \log_e\left(\frac{ad}{bc}\right)$	$SE_{LOR} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$	$w_{LOR} = \frac{abcd}{ab(c+d) + cd(a+b)}$
	Two Variable Relationships—Association between Variables		
Product-moment r	$ES_r = r$ $ES_{Zr} = .5\log_e\left[\frac{1+ES_r}{1-ES_r}\right]$	$SE_{Zr} = \frac{1}{\sqrt{n-3}}$	$w_{Zr} = n - 3$

Note. See text for definition of terms.

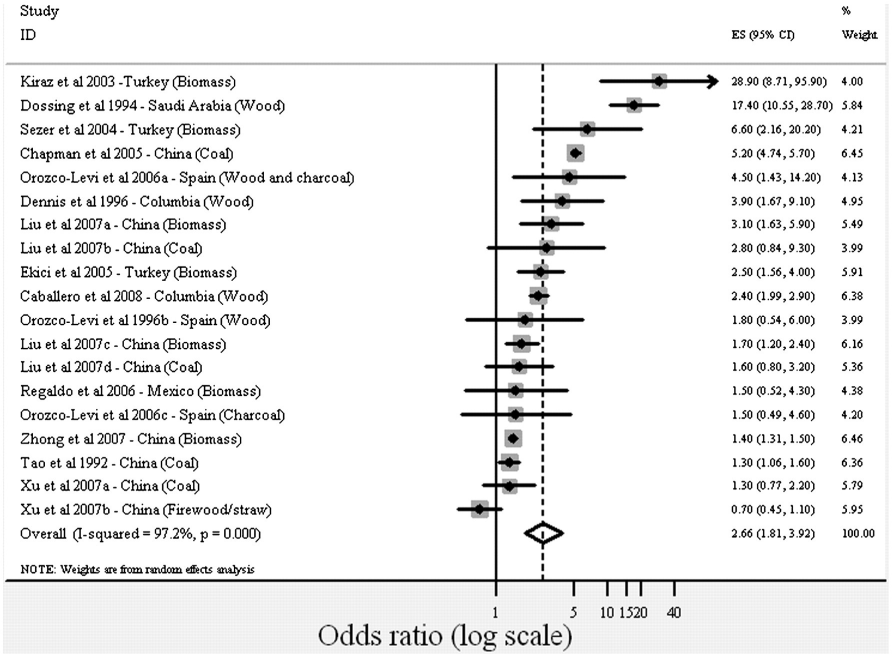
(Lipsey & Wilson, 2001, Practical Meta-Analysis)

Coding

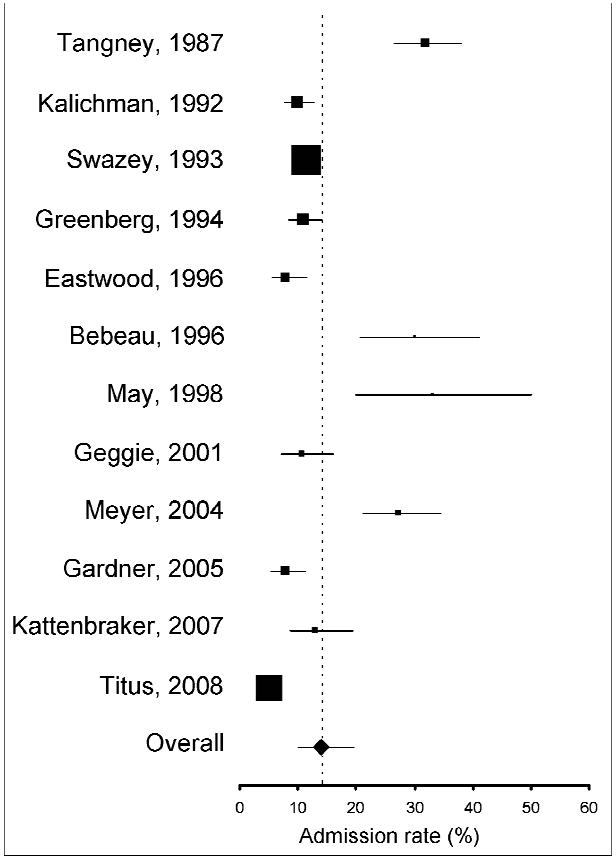
- Like a survey, with papers in place of respondents
  - Key characteristics (e.g. effect size, N, res. Design…)
  - As many other possible factors
    - Which might (but should not) affect ES (e.g. year, type of test, type of publication, statistical significance, missing values…)
- Usually more than one ES per study
  - Clarify, record characteristics for each
- Establishing of coding protocol
  - As objective as possible
    - Define each category, revise if necessary
  - Recommended when doubtful:
    - Two coders, and test for sensitivity/discuss divergence (expensive)
    - Optimize inter-coder reliability and proceed
- Controversial practices
  - Guesstimate missing/uncertain data
  - Confidence ratings on each item
  - Quality scores
    - I would avoid or at least run sensitivity analysis (W-Wo)

Analysis: Overall Effect & forest plot

$$\overline{ES} = \frac{\sum(w_i ES_i)}{\sum w_i} \qquad SE = \sqrt{\frac{1}{\sum w_i}}$$



(Kurmi et al. 2010, Thorax)



(Fanelli 2009, PLoS ONE)

Homogeneity test

$$Q = \sum w_i \left(ES_i - \overline{ES} \right)^2$$

(chi-square distribution)

Statistically significant? Then…

- -Between study variation is due to more than sampling error
- -You are not looking at a single population mean
- -Other study factors at play
- -Do you know the factors?

Fixed effects $d_j = \delta + e_j$

Random effects $d_j = \overline{\delta} + u_j + e_j$ Adds an extra variance comp. (Q, k, w)

Larger Confidence Intervals

But with few studies power is low: Q will be n.s. anyway!

Fixed effects assumptions (no study differences, all possible studies included) might be considered unrealistic

Modelling with ANOVA/Regr. analogs

- ANOVA-analog Qbetween – Qwithin
  - Qb sig. -> the factor affects ES
- Inv. Var. Regression Qr (var expl) – Qe (non)
  - Qr (variance explained) sig –> at least 1 term sig
  - Qw sig -> still study-level variance
  - Qe sig-> study-level variab.
- If Qw or Qe sig-> mixed effect model.
  - Mixed effects adds random component after initial model
- **But P-values are unreliable**
  - **And choice of model changes power/error rates**
  - **Sensitivity analysis**

The power of regression

Table 3. Inverse variance-weighted regression on admission rates.

	Variable	B±SE	P	Stand. Coeff.	Model R^2
Base Model	Constant	−4.53±0.81	<0.0001	0	0.82
	Self-/Non-self	−3.02±0.38	<0.0001	−1.04	
	Mailed/Handed	−1.17±0.4	0.0032	−0.33	
	"Fabricated, Falsified"/"Modified"	−1.02±0.39	0.0086	−0.34	
Candidate co-variables	Year	−0.03±0.03	0.3	−0.14	0.83
	USA/other	−0.71±0.4	0.08	−0.2	0.85
	Researcher/other	−0.33±0.33	0.32	−0.11	0.83
	Biomedical/other	0.17±0.39	0.66	0.06	0.82
	Medical/other	0.85±0.28	0.0022	0.29	0.89
	Social Sc./other	−0.03±0.37	0.94	−0.01	0.82

The table shows model parameters of an initial model including three methodological factors (top four rows) and the parameter values for each sample characteristic, entered one at a time in the basic model. All variables are binary. Regression slopes measure the change in admission rates when respondents fall in the first category.
doi:10.1371/journal.pone.0005738.t003

(Fanelli 2009, PLoS ONE)

Sensitivity analyses

- Jackknifing (e.g. leave one out)
- Different coding criteria
- Different coders
- + excluded studies
- Particular publications
- Different transforms

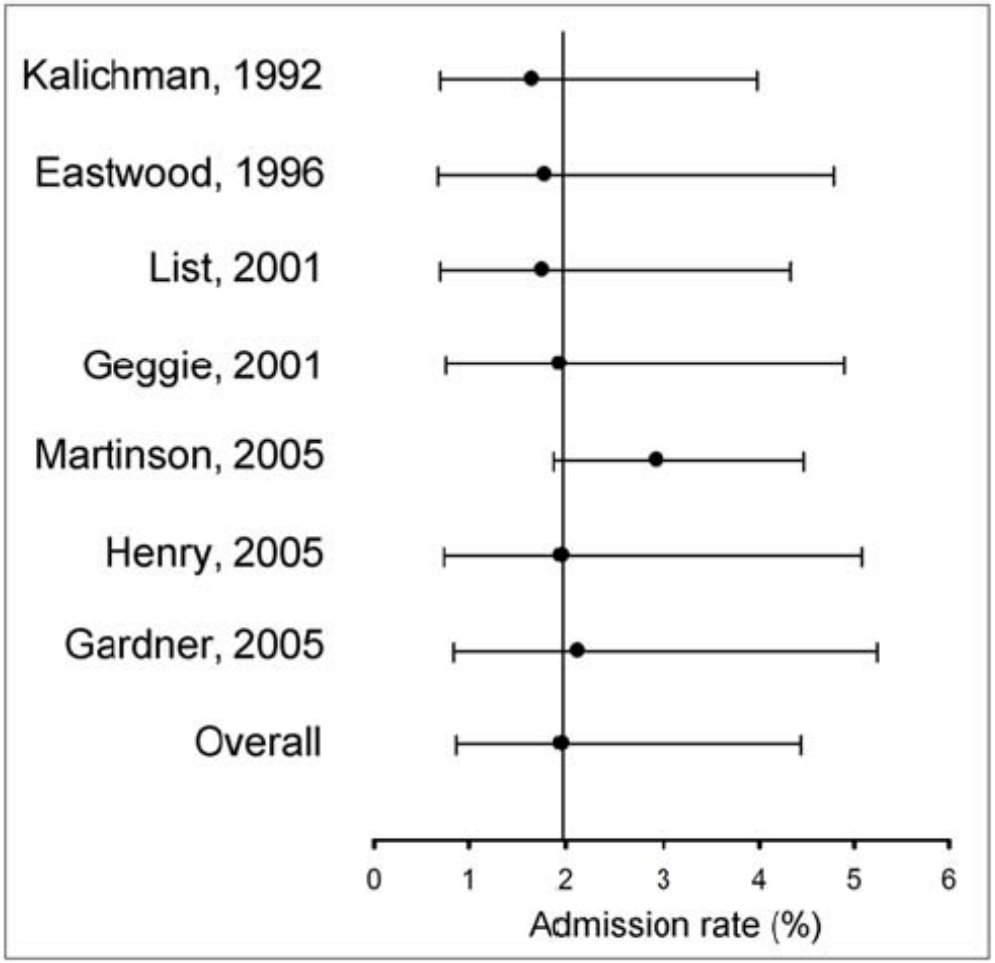
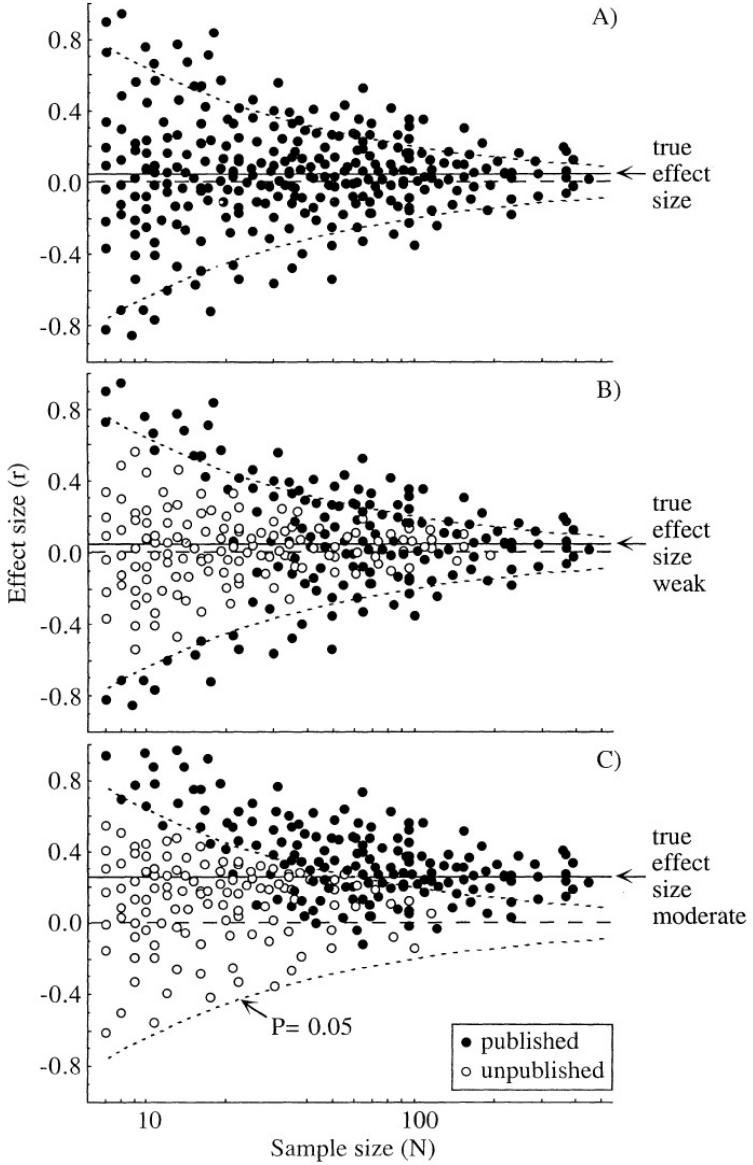


Figure 5. Sensitivity analysis of admission rates of data fabrication, falsification and alteration in self reports. Plots show the weighted pooled estimate and 95% confidence interval obtained when the corresponding study was left out of the analysis.
doi:10.1371/journal.pone.0005738.g005

(Fanelli 2009, PLoS ONE)

Assessing publication bias

- Funnel plot
  - subjective
- Modeling
  - Regression (on w, 1/SE…)
  - Fail safe
    - N of null studies that would cancel the effect
  - Trim and Fill
    - Recalculate ES adding virtual studies
  - All have limitations, naïve assumptions
- Ensure thorough search
- Combine tests/sensitivity analyses

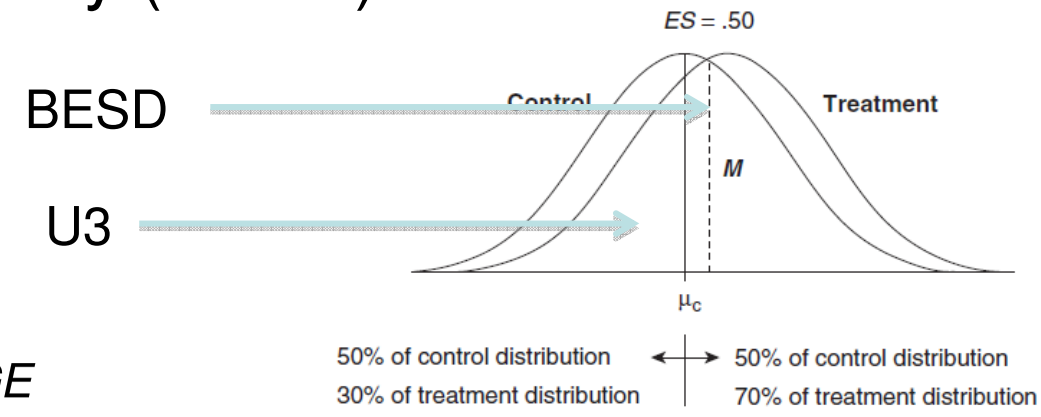


(Palmer, 1999, American Naturalist)

How important is your effect?

- YES, SIZE MATTERS!
- Cohen's 1988 convention:
  - d(st. diff. means) 0.2-small 0.5-med 0.8-large
  - r(correlation) 0.1 0.25 0.4
  - Odds ratios 1.5 2.5 4.5
  - Just a convention!
- More meaningful if expressed in a known metric
  - Will depend on study question etc…
- Binomial Effect Size Display (BESD) and U3
  - Calculated on r
  - Splitting sample
    - T-C, High-Low
    - Success-failure

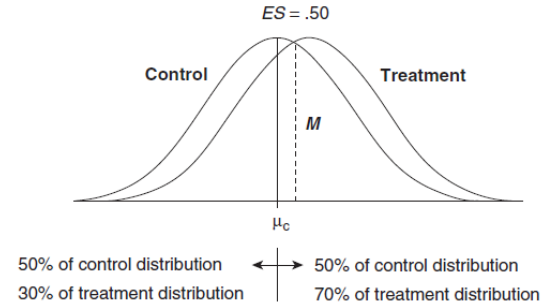
(from Lipsey & Hurley 2009, in The SAGE Handbook of Applied Social Research Methods)



Depiction of the Percentage of the Treatment Distribution Above the Success Threshold Set at the Mean of the Control Distribution

Effect size Equivalents

$$r = \frac{ES}{\sqrt{ES^2 + 4}}.$$



Depiction of the Percentage of the Treatment Distribution Above the Success Threshold Set at the Mean of the Control Distribution

ES	*r*	*U3: % of T Above X_c*	*BESD C Versus T Success Rates*		*BESD C Versus T Differential*
.10	.05	54	.47	.52	.05
.20	.10	58	.45	.55	.10
.30	.15	62	.42	.57	.15
.40	.20	66	.40	.60	.20
.50	.24	69	.38	.62	.24
.60	.29	73	.35	.64	.29
.70	.33	76	.33	.66	.33
.80	.37	79	.31	.68	.37
.90	.41	82	.29	.70	.41
1.00	.45	84	.27	.72	.45

(from Lipsey & Hurley 2009, in The SAGE Handbook of Applied Social Research Methods)