GMDS Academy 2025

A brief history of meta-analysis

Wolfgang Viechtbauer
Maastricht University
https://www.wvbauer.com

The Beginnings ...

- really have to start in the 1700-1800's with astronomy and geodesy
  - repeated but imprecise measurements of celestial objects, distances, angles, ...
- two crucial developments relevant for the combination of such information:
  - errors compensate, not multiply
  - errors should be balanced

Errors Compensate, not Multiply

- estimates made under different conditions
- "cannot/should not combine them because the error in one estimate may tarnish the rest"
- but law of large numbers:
  - let $\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ and $E(X_i) = \mu$
  - then $\bar{X}_n \xrightarrow{P} \mu$ as $n \to \infty$ $\left(\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \epsilon) = 0\right)$
  - and if $Var[X_i] = \sigma^2$ then $Var[\bar{X}_n] = \frac{\sigma^2}{n}$
- Bernoulli (1713), Poisson (1837), Chebyshev (1846), ...

Errors Should Be Balanced

- let's say we have 3 equations and 3 unknowns
  - $1 = a + 1b + 2c$
  - $4 = a + 2b + 1c$
  - $3 = a + 1b + 1c$
- we all know how to solve this, right?

Errors Should Be Balanced

- let's say we have 3 equations and 3 unknowns
  - $1 = a + 1b + 2c$
  - $4 = a + 2b + 1c$
  - $3 = a + 1b + 1c$
- we all know how to solve this, right?
  - $1 = (4) + 1(1) + 2(-2)$
  - $4 = (4) + 2(1) + 1(-2)$
  - $3 = (4) + 1(1) + 1(-2)$

Errors Should Be Balanced

- but what if we have 4 equations?
  - $1 = a + 1b + 2c$
  - $4 = a + 2b + 1c$
  - $3 = a + 1b + 1c$
  - $5 = a + 3b + 2c$
- how do we solve this?

Errors Should Be Balanced

- can satisfy the first three equations exactly
  - $1 = (4) + 1(1) + 2(-2)$
  - $4 = (4) + 2(1) + 1(-2)$
  - $3 = (4) + 1(1) + 1(-2)$
  - $5 \neq (4) + 3(1) + 2(-2) = 3$
- but "that would force all of the error onto the fourth equation; rather, we need to try to balance the errors in such a way that they are borne nearly equally by all four equations" (Legendre, 1798)

Method of Least Squares

- linear least squares (Gauss, Legendre):
  - $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + e_i$
  - select $\beta_0, \beta_1, \ldots, \beta_p$ such that $\sum_{i=1}^{n} e_i^2$ is minimized
- in example:
  - $1 = (2.2) + (1.8)1 + (-1.4)2 \boxed{+ (-0.2)}$
  - $4 = (2.2) + (1.8)2 + (-1.4)1 \boxed{+ (-0.4)}$
  - $3 = (2.2) + (1.8)1 + (-1.4)1 \boxed{+ (\ \ 0.4)}$
  - $5 = (2.2) + (1.8)3 + (-1.4)2 \boxed{+ (\ \ 0.2)}$
- Gauss also proposed that the normal distribution be used to model measurement errors

Pearson (1904)

- arguably the earliest 'meta-analysis'
- examined effectiveness of a vaccine against typhoid based on studies conducted among soldiers in the British Empire
- outcomes: infection and mortality
- data summarized in terms of 2×2 tables

APPENDIX B.

A.—INCIDENCE EXPERIENCE.

	Inoculated.	Non-inoculated.	Totals.
	1. Hospital Staffs in South Africa.		
Escaped	265	204	469
Cases	32	75	107
Totals	297	279	576
	2. Garrison of Ladysmith in South Africa.		
Escaped	1,670	9,040	10.710
Cases	35	1,489	1 524
Totals	1,705	10,529	12,234
	3. Methuen's Column in South Africa.		
Escaped	2,509	10,724	13,233
Cases	26	257	283
Totals	2,535	10,981	13,516
	4. Single Regiments in South Africa.		
Escaped	1,135	1,203	2,338
Cases	72	82	154
Totals	1,207	1,285	2,492
	5. Army in India.		
Escaped	10,798	109,034	119,832
Cases	84	1,475	1.559
Totals	10,882	110,509	121,391

Pearson (1904)

- calculated the (tetrachoric) correlation between the two variables and averaged the results

The following table gives the results of calculating the correlation coefficients of the tables in Appendix B:

INOCULATION AGAINST ENTERIC FEVER:

Correlation between Immunity and Inoculation.

I. Hospital Staffs	+ 0.373	±	0.021
II. Ladysmith Garrison	+ 0.445	±	0.017
III. Methuen's Column	+ 0.191	±	0 026
IV. Single Regiments	+ 0.021	±	0.033
V. Army in India	+ 0.100	±	0.013
Mean value	+ 0.226		

Maximum Likelihood Estimation

- Fisher formalized the idea in the 1910's
- let $f(x_i|\theta)$ be the pdf of random variable x_i that is a function of parameter (vector) θ
- under independence, the joint density of $(x_1, x_2, \ldots, x_n)$ is $\prod_{i=1}^{n} f(x_i|\theta) \equiv f(x|\theta)$
- given x, find $\hat{\theta}$ such that $f(x|\hat{\theta})$ is maximized
- then $\hat{\theta}$ is the MLE of θ
- fundamental method for fitting all kinds of statistical models (including meta-analysis)

Combining Tests of Significance

- long history of methods for combining the results from independent significance tests (Tippett, 1931; Fisher, 1932; Pearson, 1933; Stouffer et al., 1949; Wilkinson, 1951; Mosteller & Bush, 1954; Good, 1955; Lipták, 1958; Lancaster, 1961; …)

21·1. The Combination of Probabilities from Tests of Significance

When a number of quite independent tests of significance have been made, it sometimes happens that although few or none can be claimed individually as significant, yet the aggregate gives an impression that the probabilities are on the whole lower than would often have been obtained by chance. It is sometimes desired, taking account only of these probabilities, and not of the detailed composition of the data from which they are derived, which may be of very different kinds, to obtain a single test of the significance of the aggregate, based on the product of the probabilities individually observed.

(Fisher, 1932)

ChatGPT – Please simplify!

ChatGPT ⌄ Share ···

Sure! Here's an even simpler version:

Sometimes we run several separate tests, and none of them are clearly significant by themselves. But together, their results look less likely to be due to chance. In that case, we might want one overall test that combines all the p-values to see if, as a group, they show a real effect.

Fisher's Method

- p = probability of obtaining an outcome as or more extreme than the one we observe assuming the null hypothesis (H_0) is true
- if H_0 is true, then $p \sim \text{Uniform}(0, 1)$
- then we can show that $-2\ln[p] \sim \chi^2_2$
- now assume H_0 is true for $i = 1, \ldots, k$ tests
- then $-2\sum \ln[p_i] \sim \chi^2_{2k}$

Example

- want to know if x and y are correlated
- test $H_0: \rho = 0$ in three different studies
- $n = 50$ in all three ($k = 3$) studies
- find $r_1 = .26$, $r_2 = .22$, and $r_3 = .20$
- then $p_1 = .07$, $p_2 = .12$, and $p_3 = .16$
- so $-2\sum \ln[p_i] = 13.22$
- under a χ^2 distribution with $\text{df} = 6$, this yields a combined p-value of .04
- reject $H_0: \rho_1 = \rho_2 = \rho_3 = 0$

Combining Tests of Significance

- tests a fairly uninteresting null hypothesis
- does not tell us about the size of the effect
- gives the same weight to each study

Test of Significance	=	Size of the Effect or Strength of Association	×	Size of the Study
↓		↓		↓
this is what combined tests of significance are based on		this is what we typically want to know		how much we should weight the evidence

Weighting the Evidence

- estimates (y_i) have different precisions
- pay more attention to more precise estimates by giving them more weight

$$\hat{\mu} = \frac{\sum w_i y_i}{\sum w_i}$$

- if we know the standard errors (SEs) of the estimates, the most efficient weights are:

$$w_i = \frac{1}{\mathrm{SE}_i^2}$$

THEORIA
COMBINATIONIS OBSERVATIONUM
ERRORIBUS MINIMIS OBNOXIAE.

AUCTORE

CAROLO FRIDERICO GAUSS.

PARS PRIOR.

SOCIETATI REGIAE EXHIBITA FEBR. 15. 1821.

GOTTINGAE

APUD HENRICUM DIETERICH.

1823.

Vbi plura obſeruationum genera, ſeu plures determinationes ex obſeruationibus petitae, quibus haud eadem praeciſio concedenda eſt, comparantur, *pondus* earum relatiuum nobis erit quantitas ipſi *mm* reciproce proportionalis, dum *praeciſio* ſimpliciter ipſi *m* reciproce proportionalis habetur. Quo igitur pondus per numerum exprimi poſſit, pondus certi obſeruationum generis pro vnitate acceptum eſſe debet.

"When we compare several classes of observations, or several quantities derived from the observations, not having the same precision, we will take their relative **weights** to be quantities **proportional to the reciprocal of** ***mm*** **[m = SE]**. Likewise their *precisions* will be proportional to the reciprocal of *m*." (Gauss, 1825)

If Pearson (1904) had read Gauss (1825)

The following table gives the results of calculating the correlation coefficients of the tables in Appendix B:

INOCULATION AGAINST ENTERIC FEVER:
Correlation between Immunity and Inoculation.

I. Hospital Staffs ...	... ...	+ 0.373	± 0.021
II. Ladysmith Garrison	... ...	+ 0.445	± 0.017
III. Methuen's Column	... ...	+ 0.191	± 0.026
IV. Single Regiments	... ...	+ 0.021	± 0.033
V. Army in India ...	... ...	+ 0.100	± 0.013
Mean value ...	... ...	+ 0.226	

$$\hat{\mu} = \frac{\frac{1}{0.021^2} \times 0.373 + \cdots + \frac{1}{0.013^2} \times 0.100}{\frac{1}{0.021^2} + \cdots + \frac{1}{0.013^2}} = 0.234$$

Agricultural Research

- "Agricultural **experiments on the same factor** or group of factors are usually **carried out at a number of places** and **repeated over a number of years**. […] The agricultural experimenter is thus frequently confronted with the results of a set of experiments on the same problem, and has **the task of analysing and summarizing these**." (Yates & Cochran, 1938)

Agricultural Research

- laid out methods that are still in use today (Cochran, 1937, 1943, 1954; Cochran & Carroll, 1953; Rao, Kaplan, & Cochran, 1981; Yates & Cochran, 1938)
- **reformulated some core ideas:**
  - estimates are not equally precise (different variances)
  - compute a weighted average of the estimates, with weights inversely proportional to the variances
- **a new idea:**
  - estimates may be more variable than one would expect given their variances (→ heterogeneity)

Random-Effects (RE) Model

- y_i = observed effect in the ith study
- θ_i = true effect in the ith study
- model: $y_i = \theta_i + e_i$
- $\theta_i \sim N(\mu, \tau^2)$ ⟶ heterogeneity in the true effects
- $e_i \sim N(0, v_i)$ ⟶ sampling error ($v_i = SE_i^2$)
- this is still the main 'workhorse' in many meta-analyses today

Yates and Cochran (1938)

Table IX. *Responses to fertilizers in a series of experiments on sugar beet*

Washed roots (tons per acre)

Station	Mean yield	Linear response to N	P	K	Standard error
Allscott	10·97	−0·24	+0·63	+0·57	±0·519
Bardney	11·44	+1·23	+0·35	+0·01	±0·285
Brigg	13·42	+0·11	−0·38	−0·21	±0·603
Bury	13·83	+2·08	−0·05	−0·22	±0·351
Cantley	12·90	+0·20	+0·32	+0·14	±0·453
Colwick	10·12	+1·05	+0·87	−0·07	±0·287
Ely	12·46	−1·14	+0·80	−0·08	±0·886
Felstead	11·28	+3·34	+0·11	+0·23	±0·356
Ipswich	12·45	+1·64	+0·57	+0·34	±0·344
King's Lynn	19·54	+0·52	+0·12	−0·57	±0·481
Newark	14·10	+1·37	+0·54	−0·33	±0·198
Oaklands	12·84	0·00	−0·14	+0·40	±0·622
Peterborough	17·99	−0·14	+1·02	−1·34	±0·618
Poppleton	14·21	+2·72	−0·21	−0·18	±0·357
Wissington	14·55	+3·32	+0·19	+0·38	±0·443

Yates and Cochran (1938)

Table X. *Analysis of variance of response to nitrogen*

	Degrees of freedom	Sum of squares	Mean square
Average response	1	17·1949	17·1949
Response × centres	14	25·5891	1·8278
Pooled estimate of error			0·2345

The variation of the response from centre to centre has therefore an estimated variance of

$$1{\cdot}8278 - 0{\cdot}2345 = 1{\cdot}5933, \longrightarrow \hat{\tau}^2$$

excluding variance due to error. This method of estimation is not fully efficient, but may be used in cases such as the present in which the variation is large in comparison with the experimental errors.

Weights in the RE Model

- in the RE model, we again compute a weighted average of the estimates

$$\hat{\mu} = \frac{\sum w_i y_i}{\sum w_i}$$

- but now use the weights:

$$w_i = \frac{1}{\hat{\tau}^2 + \mathrm{SE}_i^2}$$


```
> # estimates and corresponding standard errors
> yi <- c(-0.24, 1.23, 0.11, 2.08, 0.20,
          1.05, -1.14, 3.34, 1.64, 0.52,
          1.37, 0.00, -0.14, 2.72, 3.32)
> sei <- c(0.519, 0.285, 0.603, 0.351, 0.453,
           0.287, 0.886, 0.356, 0.344, 0.481,
           0.198, 0.622, 0.618, 0.357, 0.443)

> # load the metafor package
> library(metafor)

> # fit the random-effects model
> rma(yi, sei^2, method="HE")
```

to use the Hedges or Cochran estimator of τ^2

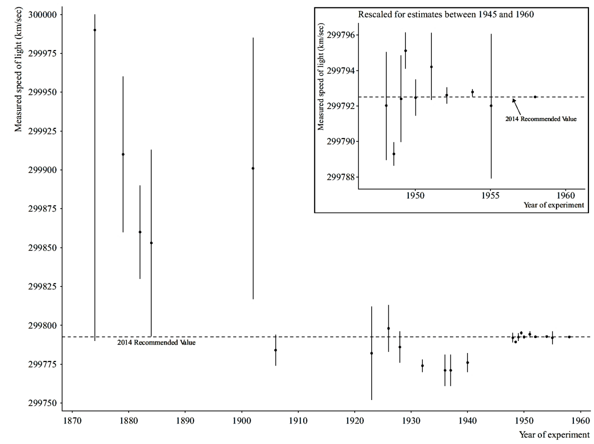

```
Random-Effects Model (k = 15; tau^2 estimator: HE)

tau^2 (estimated amount of total heterogeneity): 1.5933
tau (square root of estimated tau^2 value):      1.2622
I^2 (total heterogeneity / total variability):   91.73%
H^2 (total variability / sampling variability):  12.09

Test for Heterogeneity:
Q(df = 14) = 112.3803, p-val < .0001

Model Results:

estimate      se     zval     pval    ci.lb    ci.ub
  1.1502  0.3476   3.3093   0.0009   0.4690   1.8313  ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$= \hat{\tau}^2$

$\hat{\mu} =$

Physics

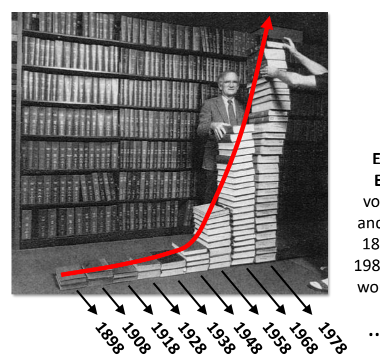
- "Let us suppose that **a given constant has been measured in several different ways**, […] it seems quite legitimate to **combine** by means of [weighted] least squares **the various results**, i.e., to weight them according to their stated probable errors, and to derive the probable error in the **final weighted average** by the usual formulas." (Birge, 1932)



(Hedges, 2019)

Information Explosion

"The individual scientist is being **overloaded with scientific information** […] and can no longer keep up with and assimilate all the information being produced"
Garvey & Griffith (1971)



Evidence of the Information Explosion, the yearly bound volumes of the *Physical Review* and *Physical Review Letters* from 1898, 1908, 1918, ..., 1978. The 1988 stack was left out because it would have been much too high.

(Fitch, 1989)

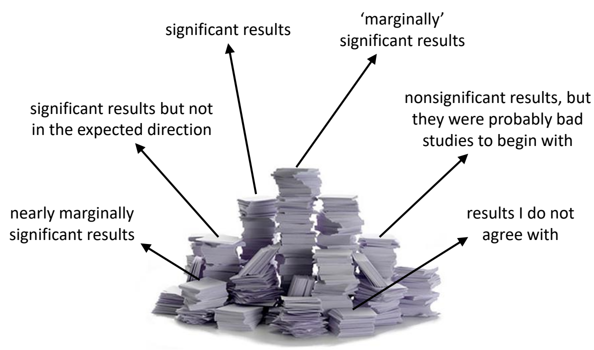
Narrative Literature Review

a description/summary of the current state of knowledge on a particular topic supported by empirical findings as well as the underlying theories and models

Non-Replicable Process

how replicable is the process of a person reading dozens or even hundreds of papers, thinking about them, and then writing down his or her conclusions?

Back to Statistical Significance ...



Vote Counting

- examine all relevant studies conducted
- categorize based on statistical significance
  - statistically significant (with $Trt > Ctrl$ or $r > 0$)
  - not statistically significant
  - statistically significant (with $Trt < Ctrl$ or $r < 0$)
- declare most frequent category the 'winner'
- inconsistent when power of studies is low (Hedges & Olkin, 1985): as $k \to \infty$, method fails to find a true effect or association

Glass (1976)

- based on his presidential address at the 1976 Annual Meeting of the AERA
- "**Meta-analysis** refers to the […] **statistical analysis of** a large collection of analysis results from individual **studies for the purpose of integrating the findings**. It connotes a **rigorous alternative to** the casual, **narrative discussions** of research studies which typify our attempts to make sense of the rapidly expanding research literature." (Glass, 1976)

Key Idea

- quantify the size, direction, and/or strength of the effect or association in each study and use this as 'primary data' in further analyses
- recall Pearson (1904)
- but now using effect size measures, such as odds and risk ratios, risk differences, raw or standardized mean differences, hazard ratios, correlation coefficients, ...
- used inverse-variance weights (but not RE models)

Smith & Glass (1977)

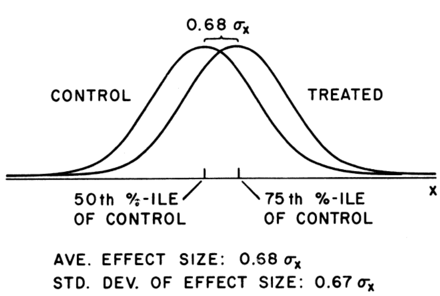
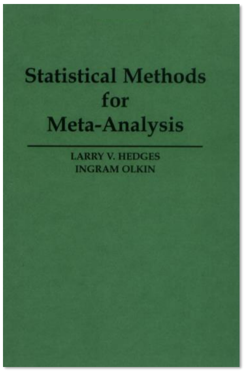


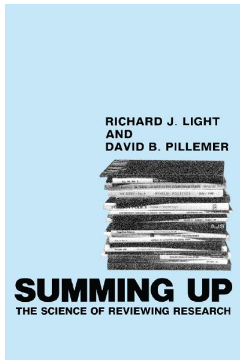
Figure 1. *Normal curves illustrating the aggregate effect of psychotherapy in relation to untreated control groups. (Data based on 833 effect size measures from 375 studies, representing about 40,000 treated and untreated subjects.)*

Early Meta-Analyses

- effectiveness of psychotherapy and counseling (Smith & Glass, 1977)
- effects of interpersonal expectations on behavior (Rosenthal & Rubin, 1978)
- relation between class size and academic achievement (Glass & Smith, 1979)
- differential validity of employment tests for (non-)Caucasian workers (Hunter et al., 1979)

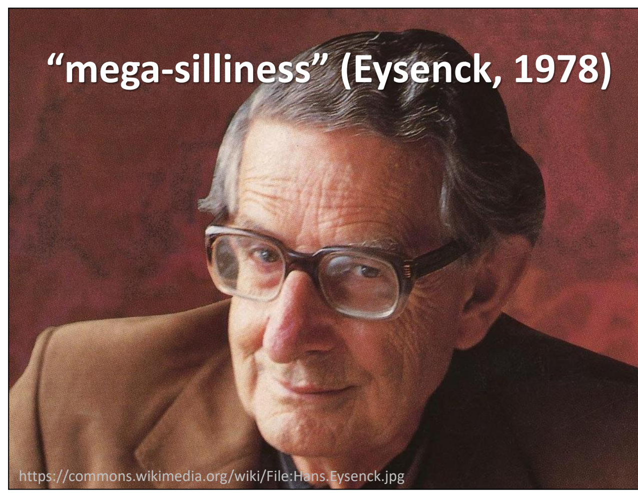
Some Early Books (1985/1984)

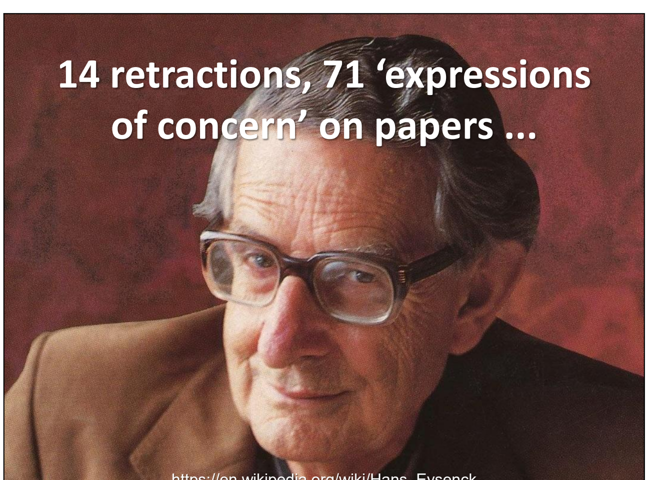


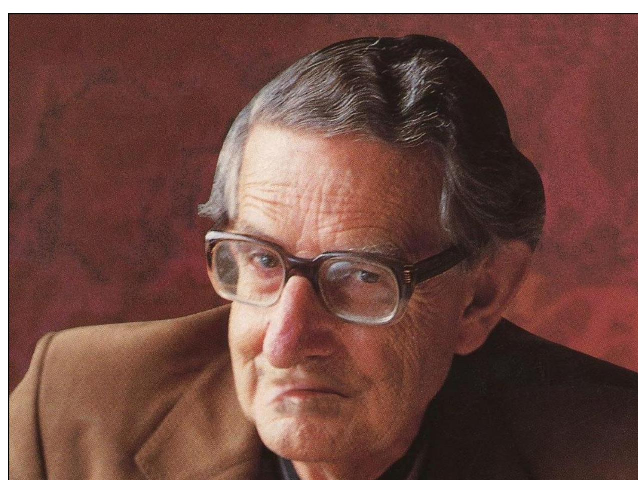


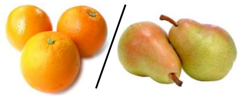
But there were critics …

- “mega-silliness” (Eysenck, 1978)
- “meta-analysis/shmeta-analysis” (Shapiro, 1994)
- “statistical alchemy” (Feinstein, 1995)
- …







 and  Critique

- two responses:
  - want to know the characteristics of fruits in general (fruit salad!)
  - want to examine systematic differences between various fruits
- two options:
  - a priori exclusion
  - a posteriori examination

Smith & Glass (1977)

Type of therapy	Average effect size	No. of effect sizes	Standard error of mean effect size	*Mdn* treated person's percentile status in control group
Psychodynamic	.59	96	.05	72
Adlerian	.71	16	.19	76
Eclectic	.48	70	.07	68
Transactional analysis	.58	25	.19	72
Rational-emotive	.77	35	.13	78
Gestalt	.26	8	.09	60
Client-centered	.63	94	.08	74
Systematic desensitization	.91	223	.05	82
Implosion	.64	45	.09	74
Behavior modification	.76	132	.06	78

Systematic Reviews

- research synthesis as a scientific process
- based on replicable and systematic methods that are meant to "limit bias in the assembly, critical appraisal, and synthesis of all relevant studies on a specific topic" (Last, 2001)
- **methods should be made explicit**
- synthesis part can make use of qualitative or quantitative methods

Cochrane and the EBM Movement

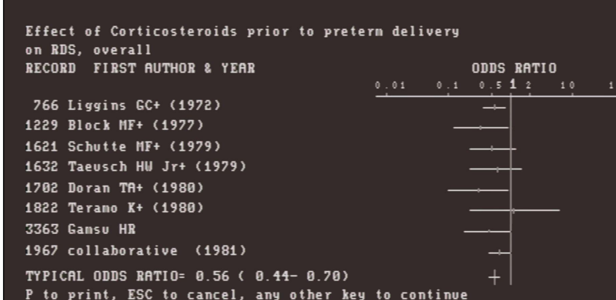
- Cochrane (1972). *Effectiveness and efficiency: Random reflections on health services.*
- advocated the use of RCTs to form the evidence base for clinical decision making
- but often there is not just one RCT ...

Cochrane and the EBM Movement

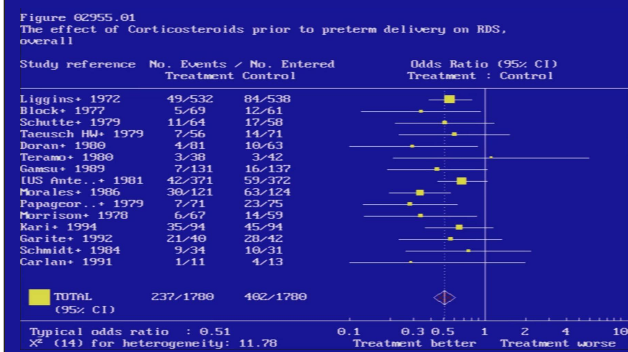
- 1985: Bibliography of Controlled Trials in Perinatal Medicine
- 1989: Effective Care in Pregnancy and Childbirth
- 1988-1993: The Oxford Database of Perinatal Trials
- 1993: Cochrane Collaboration founded
- 1993: Cochrane Pregnancy and Childbirth Database
- 1993: RevMan software released
- 1995: Cochrane Database of Systematic Reviews

THE COCHRANE COLLABORATION®

Oxford Database of Perinatal Trials



Cochrane Pregnancy and Childbirth Database



DerSimonian & Laird (1986)

- re-described the random-effects model
- proposed a more efficient estimator of τ^2 based on the method of moments

Meta-analysis in clinical trials

R DerSimonian, N Laird - Controlled clinical trials, 1986

This paper examines eight published reviews each reporting results from several related trials. Each review pools the results from the relevant trials in order to evaluate the efficacy of …

☆ Cited by 41202 Related articles »

Some More Recent Developments

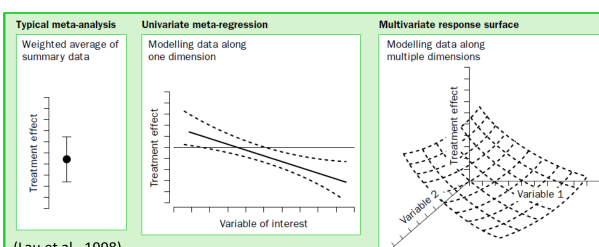
- publication bias
- meta-regression
- models based on exact likelihoods
- Bayesian models/methods
- multilevel/multivariate models
- network meta-analysis

Publication Bias

- affects all review methods (not a problem specific to meta-analysis!)
- in fact, due to meta-analysis:
  - increased awareness of publication bias
  - development of systematic methods to detect and address publication bias
  - see Marks-Anglin & Chen (2020) for an overview
  - increased emphasis on the importance of trial registries and pre-registration

Meta-Regression

- size of the effect or strength of the association may depend on study-level differences (e.g., Hedges, 1982; Lau et al., 1998; Pocock et al., 1981)



(Lau et al., 1998)

Exact Likelihood Methods

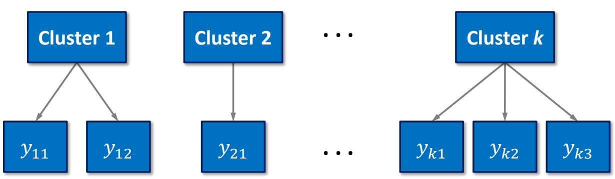
- $y_i|\theta_i \sim N(\theta_i, v_i)$ may not be accurate when studies are small and/or when dealing with rare events
- can instead assume that event counts arose from binomial or Poisson distributions
- leads to the various generalized linear mixed-effects models that can be used for meta-analysis (e.g., van Houwelingen et al., 1993; Turner et al., 2000; Jackson et al., 2018)
- essentially the same methods that are used in the analysis of multicenter trials (Senn, 2000)

Bayesian Methods

- long history of Bayesian thinking about the combination of information (e.g., Carlin, 1992; DuMouchel & Harris, 1983; Lindley, 1971)
- evidence accumulates naturally in the Bayesian framework (the evidence from study 1 becomes the prior for study 2, the posterior from this becomes the prior for study 3, and so on ...)
- formalizes the idea of 'borrowing strength'

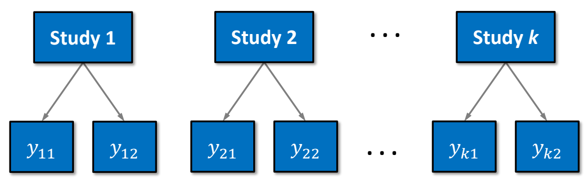
Multilevel Meta-Analytic Data

- multilevel structures can arise when we have multiple estimates for some higher clustering variable (country, author, lab, …)



Multiple (Correlated) Outcomes

- multivariate data arise when multiple outcomes are measured within the studies



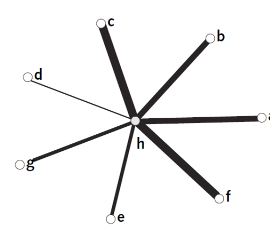
note: *not all studies have to measure all outcomes*

Network Meta-Analysis

- often there are multiple treatments available for the same condition/disease
- studies comparing the effectiveness of these treatments form a network of comparisons

Star-Shaped Networks

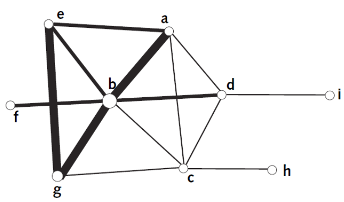
Second-generation antiepileptic drugs in partial epilepsy



a: levetiracetam, b: gabapentin, c: lamotrigine, d: oxcarbazepine, e: tiagabine, f: topiramate, g: zonisamide, h: placebo

Complex Networks

Chemotherapy regimens for ovarian cancer



a: platinum monotherapy, b: platinum-based combination,
c: taxane monotherapy, d: platinum + taxane-based combination,
e: nonplatinum/nontaxane monotherapy,
f: platinum-based combination (ip), g: nonplatinum/nontaxane combination, h: taxane-based combination,
i: platinum/taxane-based combination (ip)

Goals of a Network Meta-Analysis

- synthesize evidence provided by all studies and treatment comparisons in one model
- obtain indirect evidence about comparisons that have not been examined head-to-head
- establish hierarchy of treatment effectiveness
- identify research opportunities

Model Fitting

- analysis conducted with more complex mixed-effects models (e.g., Berkey et al., 1998; Konstantopoulos, 2011; Lu & Ades, 2004; Lumley, 2002; Raudenbush et al., 1988; Salanti et al., 2008; Senn et al., 2013; van Houwelingen et al., 2002)
- estimation via ML/REML estimation or MCMC

ML/REML Estimation

- under normality assumptions:
  - $\boldsymbol{y} \sim MVN(\boldsymbol{X\beta}, \boldsymbol{M})$
  - $\boldsymbol{X}$ model/design matrix ($k \times p$)
  - $\boldsymbol{\beta}$ vector with the fixed effects ($p \times 1$)
  - $\boldsymbol{M}$ marginal var-cov matrix ($k \times k$)
- form of $\boldsymbol{M}$ depends on model, but generally:

$$\boldsymbol{M} = \boldsymbol{D} + \boldsymbol{V}$$

structured matrix that depends on a number of parameters (e.g., $\tau^2, \sigma_1^2, \sigma_2^2, \rho$, …)

matrix with the (approximately) known sampling (co)variances

ML/REML Estimation

- log likelihood:

$$ll_{ML} = -\frac{k}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{M}| - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\beta})'\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{X\beta})$$

- where $\boldsymbol{W} = \boldsymbol{M}^{-1}$
- for given $\boldsymbol{M}$, the MLE of $\boldsymbol{\beta}$ is:

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}$$

- can construct the profiled log likelihood:

$$ll_{ML} = -\frac{k}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{M}| - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{Xb})'\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{Xb})$$

- depends only on the parameters in $\boldsymbol{M}$

ML/REML Estimation

- can also write profiled log likelihood as:

$$ll_{ML} = -\frac{k}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{M}| - \frac{1}{2}\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}$$

- where $\boldsymbol{P} = \boldsymbol{W} - \boldsymbol{W}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}$
- MLEs of the parameters in $\boldsymbol{M}$ can be found using various optimizations methods: Fisher scoring, Newton-Raphson, Nelder-Mead, quasi-Newton, …
- once parameters in $\boldsymbol{M}$ have been estimated, we can obtain the MLE of $\boldsymbol{\beta}$ as described above

ML/REML Estimation

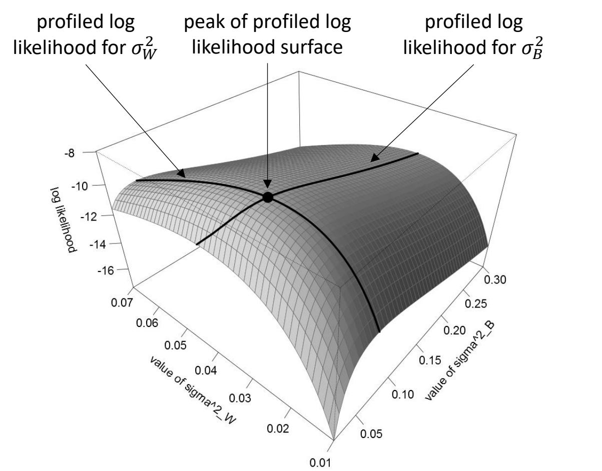
- MLEs of variance components tend to be negatively biased (esp. in small samples)
- restricted maximum likelihood estimation tends to provide approximately unbiased estimates
- restricted log likelihood:

$$ll_{REML} = -\frac{(k-p)}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{X}'\boldsymbol{X}| \\ -\frac{1}{2}\ln|\boldsymbol{M}| - \frac{1}{2}\ln|\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}| - \frac{1}{2}\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}$$

- (fixed effects are automatically profiled out)
- estimate parameters in $\boldsymbol{M}$, then obtain MLE of $\boldsymbol{\beta}$

Example: Multilevel Model

- consider the multilevel random-effects model
  - $y_{ij} = \mu + w_i + u_{ij} + e_{ij}$
  - $w_i \sim N(0, \sigma_B^2)$
  - $u_{ij} \sim N(0, \sigma_W^2)$
  - $e_{ij} \sim N(0, v_{ij})$
- ll_{ML} contains three parameters: μ, σ_B^2, and σ_W^2
- profiled log likelihood depends on σ_B^2 and σ_W^2



Is Meta-Analysis More 'Objective'?

- yes and no
- \+ we have systematic methods
- \+ methods are reported / replicable
- – many (subjective) decisions need to be made (researcher degrees of freedom, the 'garden of forking paths'; Gelman & Loken, 2014)
- maybe the wrong question ...
- are MAs more transparent? YES!

References

Berkey, C. S., Hoaglin, D. C., Antczak-Bouckoms, A., Mosteller, F., & Colditz, G. A. (1998). Meta-analysis of multiple outcomes by regression with random effects. Statistics in Medicine, 17(22), 2537-2550.

Birge, R. T. (1932). The calculation of errors by the method of least squares. Physical Review, 40(2), 207-227.

Carlin, J. B. (1992). Meta-analysis for 2x2 tables: A Bayesian approach. Statistics in Medicine, 11(2), 141-158.

Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments. Supplement to the Journal of the Royal Statistical Society, 4(1), 102-118.

Cochran, W. G. (1943). The comparison of different scales of measurement for experimental results. Annals of Mathematical Statistics, 14(3), 205-216.

Cochran, W. G. (1954). The combination of estimates from different experiments. Biometrics, 10(1), 101-129.

Cochran, W. G., & Carroll, S. P. (1953). A sampling investigation of the efficiency of weighting inversely as the estimated variance. Biometrics, 9, 447-459.

Cochrane, A. L. (1972). Effectiveness and efficacy: Random reflections on health services. Nuffield Trust.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. Controlled Clinical Trials, 7(3), 177-188.

DuMouchel, W., & Harris, J. E. (1983). Bayes methods for combining the results of cancer studies in humans and other species. Journal of the American Statistical Association, 78(382), 293-308.

Eysenck, H. J. (1978). An exercise in mega-silliness. American Psychologist, 33, 517.

Feinstein, A. R. (1995). Meta-analysis: Statistical alchemy for the 21st century. Journal of Clinical Epidemiology, 48(1), 71-79.

Fisher, R. A. (1932). Statistical methods for research workers (4th ed.). Oliver and Boyd.

Fitch, V. L. (1989). An APS president reflects on his two-year term. Physics Today, 42(12), 49-53.

Garvey, W. D., & Griffith, B. C. (1971). Scientific communication: Its role in the conduct of research and creation of knowledge. American Psychologist, 26(4), 349-362.

Gauss, C. F. (1825). Theoria combinationis observationum erroribus minimis obnoxiae. Göttingen, Germany: Dieterich.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. American Scientist, 102(6), 460.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5(10), 3-8.

Glass, G. V., & Smith, M. L. (1979). Meta-analysis of research on class size and achievement. Educational Evaluation and Policy Analysis, 1(1), 2-16.

Good, I. J. (1955). On the weighted combination of significance tests. Journal of the Royal Statistical Society, Series B, 17(2), 264-265.

Hedges, L. V. (1982). Fitting continuous models to effect size data. Journal of Educational Statistics, 7(4), 245-270.

Hedges, L. V. (2019). The statistics of replication. Methodology, 15(Supplement 1), 3-14.

Hedges, L. V., & Olkin, I. (1985). Statistical methods for meta-analysis. Academic Press.

Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. Psychological Bulletin, 86(4), 721-735.

Jackson, D., Law, M., Stijnen, T., Viechtbauer, W., & White, I. R. (2018). A comparison of seven random-effects models for meta-analyses that estimate the summary odds ratio. Statistics in Medicine, 37(7), 1059-1085.

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis. Research Synthesis Methods, 2(1), 61-76.

Lancaster, H. O. (1961). The combination of probabilities: An application of orthonormal functions. Australian Journal of Statistics, 3(1), 20-33.

Last, J. M. (2001). A dictionary of epidemiology. Oxford University Press.

Lau, J., Ioannidis, J. P. A., & Schmid, C. H. (1998). Summing up evidence: One answer is not always enough. Lancet, 351(9096), 123-127.

Light, R. J., & Pillemer, D. B. (1984). Summing up: The science of reviewing research. Harvard University Press.

Lindley, D. V. (1971). The estimation of many parameters. In V. P. Godambe & D. A. Sprott (Eds.), Foundations of statistical inference (pp. 435–455). Holt, Rinehart, and Winston.

Lipták, T. (1958). On the combination of independent tests. A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei, 3, 171-197.

Lu, G., & Ades, A. E. (2004). Combination of direct and indirect evidence in mixed treatment comparisons. Statistics in Medicine, 23(20), 3105-3124.

Lumley, T. (2002). Network meta-analysis for indirect treatment comparisons. Statistics in Medicine, 21(16), 2313-2324.

Marks-Anglin, A., & Chen, Y. (2020). A historical review of publication bias. Research Synthesis Methods, 11(6), 725-742.

Mosteller, F., & Bush, R. R. (1954). Selected quantitative techniques. 1, 289-334. Addison-Wesley.

Pearson, K. (1904). Report on certain enteric fever inoculation statistics. British Medical Journal, 2(2288), 1243-1246.

Pearson, K. (1933). On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. Biometrika, 25(3/4), 379-410.

Pocock, S. J., Cook, D. G., & Beresford, S. A. A. (1981). Regression of area mortality rates on explanatory variables: What weighting is appropriate?. Applied Statistics, 30(3), 286-295.

Rao, P. S. R. S., Kaplan, J., & Cochran, W. G. (1981). Estimators for the one-way random effects model with unequal error variances. Journal of the American Statistical Association, 76(373), 89-97.

Raudenbush, S. W., Becker, B. J., & Kalaian, H. (1988). Modeling multivariate effect sizes. Psychological Bulletin, 103, 111-120.

Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. Behavioral and Brain Sciences, 3, 377-415.

Salanti, G., Higgins, J. P. T., Ades, A. E., & Ioannidis, J. P. A. (2008). Evaluation of networks of randomized trials. Statistical Methods in Medical Research, 17(3), 279-301.

Senn, S. (2000). The many modes of meta. Drug Information Journal, 34, 535-549.

Senn, S., Gavini, F., Magrez, D., & Scheen, A. (2013). Issues in performing a network meta-analysis. Statistical Methods in Medical Research, 22(2), 169-189.

Shapiro, S. (1994). Meta-analysis/Shmeta-analysis. American Journal of Epidemiology, 140(9), 771-778.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. American Psychologist, 32(9), 752-760.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., & Williams Jr, R. M. (1949). The American soldier: Adjustment during army life (Vol. 1). Princeton University Press.

Tippett, L. H. C. (1931). The methods of statistics. Williams & Norgate.

Turner, R. M., Omar, R. Z., Yang, M., Goldstein, H., & Thompson, S. G. (2000). A multilevel model framework for meta-analysis of clinical trials with binary outcomes. Statistics in Medicine, 19(24), 3417-3432.

van Houwelingen, H. C., Arends, L. R., & Stijnen, T. (2002). Advanced methods in meta-analysis: Multivariate approach and meta-regression. Statistics in Medicine, 21(4), 589-624.

van Houwelingen, H. C., Zwinderman, K. H., & Stijnen, T. (1993). A bivariate approach to meta-analysis. Statistics in Medicine, 12(24), 2273-2284.

Wilkinson, B. (1951). A statistical consideration in psychological research. Psychological Bulletin, 48(2), 156-158.

Yates, F., & Cochran, W. G. (1938). The analysis of groups of experiments. The Journal of Agricultural Science, 28(4), 556-580.

Thank You for Your Attention!

Questions, Comments?