Focus on Data: Statistical Significance, Effect Size and the Accumulation of Evidence Achieved by Combining Study Results Through Meta-analysis

Johannes Ledolter[1,3] and Randy H. Kardon[2,3]

[1]Department of Business Analytics, Tippie College of Business, University of Iowa, Iowa City, IA, United States
[2]Department of Ophthalmology and Visual Sciences, University of Iowa, Iowa City, IA, United States
[3]Center for the Prevention and Treatment of Visual Loss, Iowa City VA Health Care System, Iowa City, IA, United States

Correspondence: Johannes Ledolter, Department of Business Analytics, Tippie College of Business, University of Iowa, Iowa City, IA, 52242, USA; johannes-ledolter@uiowa.edu.

Received: May 17, 2020
Accepted: July 7, 2020
Published: August 14, 2020

Citation: Ledolter J, Kardon RH. Focus on data: Statistical significance, effect size and the accumulation of evidence achieved by combining study results through meta-analysis. *Invest Ophthalmol Vis Sci.* 2020;61(10):32. https://doi.org/10.1167/iovs.61.10.32

Purpose. To provide information and perspectives on statistical significance and on meta-analysis, a statistical procedure for combining estimated effects across multiple studies.

Methods. Methods are presented for performing a meta-analysis in which results across multiple studies are combined. An example of a meta-analysis of optical coherence tomography thickness of the retina in patients with multiple sclerosis across multiple studies is provided. We show how to combine individual study results and how to weight the results of each study based on its reliability. The method of a meta-analysis is used to derive from all study results a pooled estimate that is closest to the unknown common effect.

Results. Differences between the two most common methods for meta-analysis, the fixed-effects approach and the random-effects approach, are reviewed. Meta-analysis is applied to the study of the differences in the thickness of the retinal nerve fiber layers of healthy controls and patients with multiple sclerosis, showing why this is a useful procedure for combining estimated effects across multiple studies to derive the magnitude of retinal thinning caused by multiple sclerosis.

Conclusions. This review provides information and perspectives on statistical significance and on meta-analysis, a statistical procedure for combining estimated effects across multiple studies. A discussion is provided to show why statistical significance and low probability values are not all that matter and why investigators should also look at the magnitude of the estimated effects. Combining estimated effects across multiple studies with proper weighting of individual results is the goal of meta-analysis.

Keywords: statistical significance, probability value, effect size, reproducibility, meta-analysis, fixed and random effects

Statistical Significance Is Not All That Matters

In 2005, Ioannidis[1,2] wrote several influential articles that suggest that up to 50% of biomedical studies are not reproducible. Why is this so? Various explanations can be given to support this claim, among them are the following.

- Institutional pressure to be funded and published may lead to either intentional or unintentional "chasing of probability values" and to conclude that results are significant, even though a fresh, unbiased view of the evidence may reveal otherwise. False-positive conclusions may result from the inappropriate treatment of outliers, the use of incorrect statistical analysis methods, ignoring violations of assumptions (e.g., non-normality of data, unequal variances, and missing values) that are critically relevant to the adopted methods, or chance alone. In the end, this can result in a "looking away" from facts that may contradict what one wants to see.
- Publication bias. Only statistically significant results tend to get published. This bias excludes studies that have been underpowered from the start, with little chance of detecting meaningful effects. The result is the publishing of small studies only when the results are statistically significant, which will happen 5% of the time even when there are no differences (assuming an alpha level of 0.05 for statistical testing).
- Statistical significance and low probability values are not all that matters. One must also look at the magnitude of the estimated effects. Cohen's d (Cohen[3]) relates the difference of two group means to the pooled standard deviation; it relates the size of the effect to the standard deviation of individual measurements. General rule of thumb guidelines consider Cohen's d of 0.2 as a small effect, 0.5 as a medium-sized effect, and 0.8 as a large

effect. Cohen's d supplements the results of inferential testing and provides perspective on meaningful effects. Statistical significance does not amount to much if the magnitude of the estimated effect is not scientifically or clinically relevant. One must not confuse statistical significance of estimated effects with the practical significance of estimated effects. Probability values alone do not tell the complete story, as even small and meaningless effects can be identified as significant with a large sample size.

- Repeated positive results, even though not statistically significant in each individual study, can add up to significant findings if results are combined through meta-analysis, a method that is discussed in the following section, Borrowing Strength: Combining Results From Different Studies Through Meta-Analysis. However, meta-analysis is compromised when nonsignificant studies remain unpublished, because then only the positive studies are combined, biasing the result toward significance.
- Confidence intervals are preferable to probability values. Confidence intervals express both the magnitude of the estimated effect and the uncertainty of the estimate. The uncertainty of the estimate gives perspective on how likely the results are repeatable.
- Probability values provided on a continuous scale are preferable to a binary (no/yes) report of statistical significance at an arbitrarily chosen criteria level (e.g., $P \leq 0.05$). A result with a probability value of 0.105 is not all that different from one with a probability value of 0.095 or even 0.047, especially if the dataset is small and if there is uncertainty whether all assumptions that went into the statistical test that generated the probability value were actually satisfied.

In a 2019 special issue of *The American Statistician* (see Wasserstein et al.[4]), the American Statistical Association recommends against abusive use of probability values; the lead editorial suggests abandoning the use of the term "statistically significant" altogether.

Borrowing Strength: Combining Results From Different Studies Through Meta-Analysis

A meta-analysis is a statistical procedure for combining estimated effects across multiple studies. Individual study results are measured with their errors and calculated confidence intervals. The aim of a meta-analysis is to derive a pooled estimate that is closest to the unknown common effect.

Although there are many different methods for meta-analysis, with each version making slightly different assumptions, all existing methods yield a weighted average of individual study results. The difference is in the way these weights and the uncertainty (confidence interval) of the resulting weighted estimate are calculated.

A meta-analysis assumes that the results of multiple studies are independent. Studies that are combined should include different patients, and must not just reanalyze the same experimental data. Independence implies that results of one study have no bearing on the results of the other studies.

Example

Retinal imaging biomarkers are important for early recognition and monitoring of inflammation and neurodegeneration in multiple sclerosis (MS). With the introduction of spectral domain optical coherence tomography, measurements on the thickness of retinal nerve fiber layers are now readily available, but it is important to know which retinal layers show atrophy associated with neurodegeneration in MS. Petzold et al.[5] conducted a comprehensive review of published studies that addressed this issue and they used meta-analysis methods to combine the results of all relevant studies. Our illustration here compares the thickness of the peripapillary retinal nerve fiber layer in eyes of healthy controls with eyes of patients having MS, but without a history of optic neuritis. The Table summarizes the results of the 18 studies identified in the review by Petzold et al.[5]

The Table reports for each study the number of enrolled participants and the mean and standard deviation of the retinal nerve fiber layer thickness (in microns), for both the MS group and the group of healthy controls. The difference of the two group means, y, is reported in (bold face) column 8 of the Table. A negative difference (effect) indicates that the mean retinal nerve fiber layer thickness of MS patients is smaller than that of healthy controls. The standard error of the estimated effect, $\hat{\sigma} = \sqrt{(s_{MS}^2/n_{MS}) + (s_{HC}^2/n_{HC})}$, is shown in (bold face) column 9. It is derived from the standard deviations and the sample sizes of the two groups. The limits of the 95% confidence interval for a study's mean effect are given in columns 10 and 11. We use percentiles of the standard normal distribution for calculating the confidence intervals. Alternatively, one can use percentiles of the t-distribution, with degrees of freedom given by the Welch approximation; see Ledolter et al.[6] However, this makes little difference as the sample sizes are fairly large. The t-ratio for the mean difference is shown in column 12. The two-sided probability value, testing whether or not the mean difference is significant, is shown in column 13; three of the 18 studies (with P values greater than 0.05 in bold face) were considered not statistically significant.

We use this example to illustrate the basic concepts behind a meta-analysis. The investigator must make certain choices on methods (described elsewhere in this article) when carrying out a meta-analysis, and these choices can affect the results. Also, it is important to establish objective criteria for including studies, because the results of a meta-analysis depend on which studies are included.

Method 1: The Fixed-Effects Model

The fixed-effects model calculates a weighted average of the reported estimated study effects, y_i, in column 8 of the Table. The sample variance of an estimated study effect, denoted by $\hat{\sigma}_i^2$, reflects the reliability of the estimate; it is obtained by squaring the standard error explained above and shown in column 9 of the Table. The reciprocal of this variance, $\hat{\sigma}_i^{-2}$, represents the weight that is attached to the ith study effect, so that reliable effects (less variability in the effect across the patients studied) contribute to the weighted average more than unreliable ones (more variability in the effect across patients). The weights $\hat{\sigma}_i^{-2}$ and the normalized weights $w_i = \frac{\hat{\sigma}_i^{-2}}{\sum \hat{\sigma}_i^{-2}}$ are shown in the last two columns of the Table. The weighted (pooled) average of the n (here $n = 18$) estimated study effects

Table. Results of 18 Studies Assessing the Difference in Mean Thickness (in Microns) of the Peripapillary Retinal Nerve Fiber Layer in Eyes of Healthy Control Patients and in Eyes of Patients With MS Without Optic Neuritis

	MS without Optic Neuritis			Healthy Controls			Difference		95% Confidence Interval					Fixed-Effects
Studies	**Eyes**	**Mean**	**SD**	**Eyes**	**Mean**	**SD**	**Mean = y**	**SE $\hat{\sigma}$**	**Lower Bound**	**Upper Bound**	**t-Ratio**	***P* Value**	$1/\hat{\sigma}^2$	**Weights**
Balk (2014)	279	85.5	10.1	126	91.7	6.8	**−6.2**	**0.86**	−7.88	−4.52	−7.24	0.000	1.365	0.195
Behbehani (2015)	72	102.7	11.5	51	111.3	8.7	**−8.6**	**1.82**	−12.17	−5.03	−4.72	0.000	0.301	0.043
Behbehani (2016)	10	89	16.2	40	94.9	6	**−5.9**	**5.21**	−16.11	4.31	−1.13	**0.257**	0.037	0.005
Esen (2016)	54	89.2	11.2	60	96.7	8.2	**−7.5**	**1.86**	−11.14	−3.86	−4.04	0.000	0.290	0.042
Feng (2013)	12	92	8.5	28	102.1	8.1	**−10.1**	**2.89**	−15.77	−4.43	−3.49	0.000	0.120	0.017
Gelfand (2012)	820	91.5	13.7	106	101.3	10.1	**−9.8**	**1.09**	−11.94	−7.66	−8.98	0.000	0.839	0.120
Gonzalez–Lopez (2014)	104	90.7	12.7	140	99.3	8.7	**−8.6**	**1.45**	−11.43	−5.77	−5.95	0.000	0.478	0.068
Huang–Link (2015)	15	87.7	9.7	68	93.6	8.9	**−5.9**	**2.73**	−11.25	−0.55	−2.16	0.031	0.134	0.019
Khalil (2016)	38	92.4	17.7	23	117.8	26.2	**−25.4**	**6.17**	−37.50	−13.30	−4.12	0.000	0.026	0.004
Klistorner (2014)	53	94	6.5	50	99	9.8	**−5**	**1.65**	−8.23	−1.77	−3.03	0.002	0.368	0.053
Knier (2016)	36	97	10.4	38	100.1	8.6	**−3.1**	**2.23**	−7.46	1.26	−1.39	**0.164**	0.202	0.029
Lange (2013)	37	93.2	14.4	100	98.4	8.8	**−5.2**	**2.53**	−10.15	−0.25	−2.06	0.040	0.157	0.022
Oberwahrenbrock (2012)	571	90.2	12.3	183	100.6	9.4	**−10.4**	**0.86**	−12.09	−8.71	−12.03	0.000	1.337	0.191
Oberwahrenbrock (2013)	66	99.9	11.3	66	100.7	8	**−0.8**	**1.70**	−4.14	2.54	−0.47	**0.639**	0.344	0.049
Petracca (2016)	50	86.9	13.6	40	92.8	12.4	**−5.9**	**2.75**	−11.28	−0.52	−2.15	0.032	0.133	0.019
Soufi (2015)	55	91	11	58	104	8.7	**−13**	**1.87**	−16.67	−9.33	−6.94	0.000	0.285	0.041
Walter (2012)	150	87.6	11.1	61	92.9	9.9	**−5.3**	**1.56**	−8.35	−2.25	−3.40	0.001	0.412	0.059
Xu (2016)	41	89.3	11.5	41	97.1	11.5	**−7.8**	**2.54**	−12.78	−2.82	−3.07	0.002	0.155	0.022

SD, standard deviation.

$$effect_{pooled} = \sum\nolimits_{i=1}^{n} w_i y_i$$

is the estimate of the unknown common treatment effect. The standard error of the pooled estimate is given by

$$se(effect_{pooled}) = \sqrt{\frac{1}{\sum \hat{\sigma}_i^{-2}}}.$$

These are the generalized least squares estimates of a population mean and its standard error when observations have unequal variances $\hat{\sigma}_i^2$; see Abraham and Ledolter[7] (page 128).

For the example in the Table, $effect_{pooled} = -7.70$ microns and $se(effect_{pooled}) = 0.378$ microns. The 95% confidence interval of $-7.70 \pm (1.96)(0.378)$ extends from -8.44 to -6.96. The probability value for testing the hypothesis whether or not the common effect is zero is less than 0.0001.

Method 2: The Random-Effects Model

The fixed-effects model assumes that all included studies are drawn from the same population. This assumption is unrealistic because the studies are heterogeneous and treatment or disease effects differ owing to diverse measurement devices and algorithms, and local study conditions including genetic and environmental influences on the populations being studied. The random-effects model relaxes this assumption, which makes it a more realistic model in most situations.

The random-effects model assumes that the treatment effect from the ith study, y_i, is distributed as $y_i|\mu_i \sim N(\mu_i, \sigma_i^2)$, where μ_i is the true underlying treatment effect of the ith study and σ_i^2 is the corresponding within-study variance. This variance is estimated by $\hat{\sigma}_i^2$, the squared entry of column 9 of the Table. The random-effects model further assumes that $\mu_i \sim N(\mu, \tau^2)$, where μ and τ^2 denote the overall treatment effect and the between-study variance, respectively. These two distributions imply the marginal distribution $y_i \sim N(\mu, \sigma_i^2 + \tau^2)$.

Random-effects procedures for a meta-analysis differ by how the between-study variance τ^2 gets estimated. The DerSimonian and Laird[8] estimate, $\hat{\tau}_{DL}^2$, is commonly used. DerSimonian and Laird use the Q statistic, $Q = \sum\nolimits_{i=1}^{n} \sigma_i^{-2}(y_i - \bar{y})^2$, where $\bar{y} = \sum\nolimits_{i=1}^{n} \sigma_i^{-2} y_i \Big/ \sum\nolimits_{i=1}^{n} \sigma_i^{-2} = \sum\nolimits_{i=1}^{n} w_i y_i$ is the pooled estimate under the fixed-effects model. Under the assumptions of the random-effects model, the expectation of Q is $E(Q) = (n-1) + (S_1 - \frac{S_2}{S_1})\tau^2$, where $S_1 = \sum\nolimits_{i=1}^{n} \sigma_i^{-2}$and $S_2 = \sum\nolimits_{i=1}^{n} \sigma_i^{-4}$. Replacing all unknown variances σ_i^2 with their estimates $\hat{\sigma}_i^2$ and solving the equation $Q = E(Q)$ for τ^2 leads to the DerSimonian and Laird estimate $\hat{\tau}_{DL}^2 = \max(0, \frac{Q-(n-1)}{S_1 - \frac{S_2}{S_1}})$. With this estimate, the generalized least squares estimate of the common treatment effect μ is given by the weighted average

$$effect_{DL} = \frac{\sum\nolimits_{i=1}^{n} (\hat{\sigma}_i^2 + \hat{\tau}_{DL}^2)^{-1} y_i}{\sum\nolimits_{i=1}^{n} (\hat{\sigma}_i^2 + \hat{\tau}_{DL}^2)^{-1}} = \sum\nolimits_{i=n}^{n} \tilde{w}_i y_i,$$

$$\text{with weights } \tilde{w}_i = \frac{(\hat{\sigma}_i^2 + \hat{\tau}_{DL}^2)^{-1}}{\sum\nolimits_{i=1}^{n} (\hat{\sigma}_i^2 + \hat{\tau}_{DL}^2)^{-1}}.$$

The standard error of this estimate is

$$se(effect_{DL}) = \sqrt{\frac{1}{\sum\nolimits_{i=1}^{n} (\hat{\sigma}_i^2 + \hat{\tau}_{DL}^2)^{-1}}}.$$

For the example in the Table, $\hat{\tau}_{DL}^2 = 7.21$. The standard deviation $\hat{\tau}_{DL} = \sqrt{7.21} = 2.69$ microns reflects the between-study variability. The estimate of the overall treatment effect is $effect_{DL} = -7.41$ microns, with a standard error $se(effect_{DL}) = 0.805$ microns. The approximate 95% confidence interval $-7.41\pm$ (1.96)(0.805) extends from -8.98 to -5.83, and the probability value for testing the hypothesis whether or not the common effect is zero is less than 0.0001.

The random-effects model allows for variability among the study effects, whereas in the fixed-effects model all study effects are assumed to originate from a single common mean. In our example, the two pooled estimates, $effect_{pooled} = -7.70$ and $effect_{DL} = -7.41$, are about the same. However, because of the large between-study variability, the standard error from the random-effects model, $se(effect_{DL}) = 0.805$, is twice as large as the standard error from the fixed-effects model $se(effect_{pooled}) = 0.378$. However, the conclusions are unchanged; either method of meta-analysis confirms highly significant differences in the mean retinal thickness of healthy controls and patients with MS.

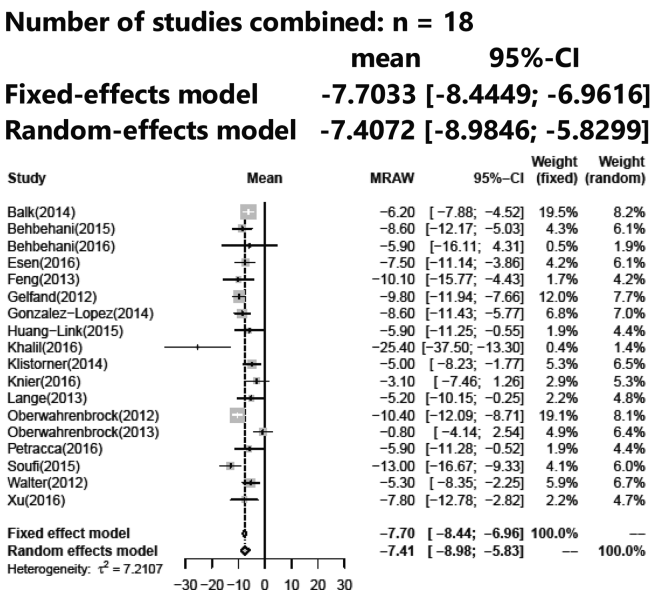
Furthermore, the magnitudes of the estimated effects are large. For each study we pool the standard deviations of the MS and healthy control groups in columns 4 and 7 of the Table, $s_{pooled} = \sqrt{[(n_{MS}-1)s_{MS}^2 + (n_{HC}-1)s_{HC}^2]/(n_{MS}+n_{HC}-2)}$. We then average the standard deviation estimates across the 18 studies, using the fixed effects weights in column 15 of the Table as an indication of their reliability. The resulting estimate, $s = 10.63$ microns, is insensitive to the weights being used (the equally weighted average is 10.87; the weighted average with weights from the random-effects is 10.53). The standard deviation $s = 10.63$ leads to a Cohen's d in the medium to large category (7.70/10.63 = 0.72 and 7.41/10.63 = 0.70).

Software

The R library meta[9] can be used to carry out the analysis and visualize its results. The forest plot in the Figure displays the estimated study effects by squares and their confidence intervals by horizontal lines. The area of the square reflects the precision of the treatment estimate. The vertical line through zero represents the no effect hypothesis. Confidence intervals for individual studies that overlap with this line demonstrate that their effects do not differ from zero. The weights for fixed and random-effects meta-analyses and their resulting pooled estimates, visualized by vertical lines and diamonds as labels, are shown. In this example, the differences between the two estimates and the two vertical lines are minor. The name forest plot refers to the forest of lines it produces. Graphpad Prism 8 also can be used to produce forest plots.

Summary

Rigor and reproducibility are now emphasized in the design and analysis of experiments to help ensure that results hold up to the test of time and can be replicated in other studies.



```
Number of studies combined: n = 18
                           mean        95%-CI
Fixed-effects model     -7.7033 [-8.4449; -6.9616]
Random-effects model    -7.4072 [-8.9846; -5.8299]
```

Study	MRAW	95%–CI	Weight (fixed)	Weight (random)
Balk(2014)	−6.20	[−7.88; −4.52]	19.5%	8.2%
Behbehani(2015)	−8.60	[−12.17; −5.03]	4.3%	6.1%
Behbehani(2016)	−5.90	[−16.11; 4.31]	0.5%	1.9%
Esen(2016)	−7.50	[−11.14; −3.86]	4.2%	6.1%
Feng(2013)	−10.10	[−15.77; −4.43]	1.7%	4.2%
Gelfand(2012)	−9.80	[−11.94; −7.66]	12.0%	7.7%
Gonzalez−Lopez(2014)	−8.60	[−11.43; −5.77]	6.8%	7.0%
Huang−Link(2015)	−5.90	[−11.25; −0.55]	1.9%	4.4%
Khalil(2016)	−25.40	[−37.50; −13.30]	0.4%	1.4%
Klistorner(2014)	−5.00	[−8.23; −1.77]	5.3%	6.5%
Knier(2016)	−3.10	[−7.46; 1.26]	2.9%	5.3%
Lange(2013)	−5.20	[−10.15; −0.25]	2.2%	4.8%
Oberwahrenbrock(2012)	−10.40	[−12.09; −8.71]	19.1%	8.1%
Oberwahrenbrock(2013)	−0.80	[−4.14; 2.54]	4.9%	6.4%
Petracca(2016)	−5.90	[−11.28; −0.52]	1.9%	4.4%
Soufi(2015)	−13.00	[−16.67; −9.33]	4.1%	6.0%
Walter(2012)	−5.30	[−8.35; −2.25]	5.9%	6.7%
Xu(2016)	−7.80	[−12.78; −2.82]	2.2%	4.7%
Fixed effect model	**−7.70**	**[−8.44; −6.96]**	**100.0%**	**--**
Random effects model	**−7.41**	**[−8.98; −5.83]**	**--**	**100.0%**

Heterogeneity: $\tau^2 = 7.2107$

FIGURE. Results of the meta-analysis of the data from the Table. CI, confidence interval.

We have outlined possible factors that can compromise the reproducibility of published results that have been interpreted as being significant. We have also emphasized the importance of characterizing the effect size, Cohen's d, based on the difference of means relative to the standard deviation of the results. The major emphasis in this tutorial is on how to best combine the results of various studies, through a meta-analysis approach to determine if an overall effect is likely to be significant and consistent across similar studies, which helps to bolster the rigor and repeatability of conclusions.

Acknowledgments

Supported by the Center for the Prevention and Treatment of Visual Loss, Iowa City VA Health Care Center (RR&D C9251-C; RX003002), and an endowment from the Pomerantz Family Chair in Ophthalmology (RK).

Disclosure: **J. Ledolter,** None; **R.H. Kardon,** None

References

1. Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research. *JAMA*. 2005;294:218–228.
2. Ioannidis JPA. Why most published research findings are false. *PLoS Med*. 2005;2:e124.
3. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. New York: Routledge; 1988.
4. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "p < 0.05." Editorial. *Am Stat*. 2019;73: 1–19.
5. Petzold A, Balcer LJ, Calabresi PA, et al. Retinal layer segmentation in multiple sclerosis: a systematic review and meta-analysis. *Lancet Neurol*. 2017;16:797–812.
6. Ledolter J, Gramlich OW, Kardon RH. Focus on data: parametric statistical inference. *Invest Ophthalmol Vis Sci*. 2020;61:11.
7. Abraham B, Ledolter J. *Introduction to Regression Modeling*. Belmont, CA: Duxbury Press; 2006:128.
8. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trial*. 1986;7:177–188.
9. R Software for Statistical Computing. *Library meta*. Vienna, Austria: R Foundation for Statistical Computing. Available at: www.R-project.org/.