A Simple and Robust Way of Concluding Meta-Analysis Results Using *P* values, Standardized Effect Sizes, or Other Statistics

Po-Huang Chyou, PhD

Meta-analysis is a powerful tool to estimate measures of associations or effects based on published or unpublished reports. However, problems exist in many meta-analyses, particularly related to study heterogeneity. This article proposes a way of concluding meta-analysis results using *P* values, taking heterogeneity into account. There is little published research focused on evaluating conclusiveness of summary results of reported meta-analyses. Generally, a *P* value is directly linked to the test statistic $z=b/s_b$ following a standard normal distribution with mean zero and unit variance, where b is an estimator of β and s_b is the estimated standard error of b for any study included in a meta-analysis. This forms the basis of the proposed method for deriving overall test statistics and corresponding *P* values used for comparing results of meta-analyses. Two published meta-analyses were chosen and specific software was applied. Results are consistent with the two published meta-analysis reports in terms of *P* values for significance and direction of summary measure of treatment effect. This proposed method can be utilized to safeguard against improper conclusions of published meta-analyses due to heterogeneity. Exploring more sophisticated statistical methods for situations when the key assumption applied to this proposed method is violated could be pursued and could expand the scope of applications beyond this method.

Keywords: Meta-analysis; Heterogeneity; *P* value

Meta-analysis is a powerful tool to estimate overall measures of associations or effects based on published or unpublished reports.[1] However, many problems exist in most, if not all, meta-analyses. One specific problem that may cause doubt about the result of a meta-analysis is study heterogeneity.[2] Two of the most commonly cited issues relevant to heterogeneity are different treatment effect parameters (generally expressed as β_1, β_2, etc; called regression coefficients in linear regression analysis, odds ratio logistic regression analysis, or relative risk regression analysis) and different variances of these parameter estimates (eg, σ_1^2, σ_2^2). Other important issues may include different study populations (eg, Asian populations, European populations), different study sample sizes (eg, n_1, n_2), different study designs (eg, clinical trials, cohort studies, case-control studies, cross-sectional studies), and the

appropriateness of a fixed effects model versus a random effects model.[2] There has been very little published research that focused particularly on evaluating whether or not the summary result of a reported meta-analysis is accurate.

The present article proposes a simple and robust way of concluding meta-analysis results using *P* values and standardized effect sizes. *P* values are used as the means of converting meta-analysis results to defined/known test statistics which are expressible as a function of the estimates of the βs and σs described above. These test statistics are then applied to deriving the standardized normal z values (ie, standardized effect sizes) based on the sign (positive or negative) of βs. With a given level of significance (ie, type I error), say 0.05, the conclusiveness of meta-analysis results can be determined. Alternatively, the standardized normal z

values of the proposed method can also be derived if estimates of βs, σs, or the confidence limits are reported in the meta-analysis. This method has been tested using published meta-analysis results. Interpretations and limitations of the results from this method are also discussed.

Methods

Possible outcome measures such as odds ratio, relative risk, risk difference, rate ratio, and the number needed to treat may be singularly chosen to obtain a summary measure in a reported meta-analysis. Beyond each of these measures, a corresponding P value which is used to compare with the level of significance is almost always reported. In general, a P value is directly linked to the test statistic $z=b/s_b$, which follows a standard normal distribution with mean 0 and unit variance,[3] where b is an estimator of β and s_b is the estimated standard error of b for any study included in a meta-analysis. Formally, let us assume there are k studies included in a meta-analysis, and let p_i denote the ith P value directly obtained from the ith study, where i = 1, 2,…, k. With the given p_i, the inverse standardized normal test statistic z_i can be derived from the probability equation: $pr(z_i > z_0$ or $z_i < -z_0) = p_i$, for i = 1,2,…, k, where z_0 represents the assumed constant for hypothesis testing. Then, these z_is will be averaged to obtain the sample mean and denoted as: zbar = $\sum z_i/k$, which is N(0, 1/k) since each z_i is N(0, 1) for i = 1, 2,..., k. The confidence interval (CI) for its true mean μ_{zbar} can then be constructed as follows: zbar $\pm$ $z_{\alpha/2}$ × s_{zbar}, where α is the assigned level of significance, $z_{\alpha/2}$ is a standardized normal value, and $s_{zbar} = \{\sum (z_i - zbar)^2/ [k \times (k-1)]\}^{1/2}$. If the derived CI excludes zero, then evidence exists that the meta-analysis of interest has shown a significant treatment effect. In general this treatment effect is presented as a summary measure such

as odds ratio, rate ratio, or relative risk. In addition, the derived results of zbar and S_{zbar} can be used to define the test statistic z_{bar}/s_{zbar} which is N(0, 1), and therefore its corresponding P value can easily be derived. The resulting P value can be used to confirm or negate the reported result of the same meta-analysis.

It is not uncommon that actual P values may not be reported in meta-analysis, but instead, other relevant statistics such as rate ratio and corresponding (eg, 95%) confidence limits (CL) are included. The current proposed method is still applicable in this situation. Take odds ratio (OR) as an example: the standardized z (= b/s_b) statistic of this proposed method can be obtained using methods based on the CI/CL for each study, where b = LOR (that is, the log OR), and s_b = ln(OR$_u$/OR)/1.96, or s_b = ln(OR/OR$_l$)/1.96, the standard deviation of LOR. OR$_l$ and OR$_u$ represent the lower and upper CL for the OR, respectively.

Data Analysis and Statistical Software

Data from two published meta-analyses were chosen for comparisons. The first data set was published by Burr et al,[4] who considered 12 epidemiological studies that investigated the link between presence/absence of the platelet PlA polymorphism and presence/absence of coronary heart disease. Reported statistics included LORs, mainly from a logistic regression model, and the corresponding P values. The second data set used was published by Gareen et al.[5] They considered 32 studies (readers can refer to tables 2-3 in Gareen's paper) used to assess the association between intrauterine devices and pelvic inflammatory disease. Reported statistics mainly included rate ratios and 95% CL.

Table 1. Comparisons between reported meta-analysis in log-odds ratios and P values and the proposed method.

Burr et al (2003)[4]	Reported		Proposed	
	LOR	*P*	Converted OR	z
	1.055	0.005	2.87	2.8070
	-0.100	0.388	0.90	-0.8633
	0.626	0.006	1.87	2.7478
	0.017	0.886	1.02	0.1434
	1.068	0.023	2.91	2.2734
	-0.025	0.835	0.98	-0.2083
	-0.117	0.596	0.89	-0.5302
	-0.381	0.111	0.68	-1.5937
	0.507	0.006	1.66	2.7478
	0.000	0.999	1.00	0.0013
	0.385	0.061	1.47	1.8735
	0.405	0.111	1.50	1.5937

LOR, log-odds ratio; OR, odds ratio

The commercially available statistical software package PASS[6] was used to obtain the inverse normal z statistic values, and a SAS[7] program was written and carried out for obtaining zbar and the corresponding CI based on the results of each of the published meta-analyses.

Results

Table 1 compares the reported results of the meta-analysis conducted by Burr et al[4] and the results obtained using the current proposed method. Burr et al[4] reported LOR and *P* values. For our purposes, LOR was converted to OR (= exp[LOR]) and used to calculate z values. Based on the calculated z values, the current proposed method led us to an estimate of effect (b) of 0.9160 with a standard error (s_b) of 0.4586 with a 95% confidence interval of (0.0172, 1.8149), which does not include 0. The final test statistic z = 0.9160/0.4586 = 1.9976 was obtained, yielding a two-sided *P* value of 0.0457. This is consistent with the findings of Burr et al.[4] In Burr's study,[4] it was concluded that the PlA polymorphism is a risk factor for coronary heart disease, with the mean of the distribution of the LOR estimated to be 0.2 (ie, OR = exp[0.2] = 1.2), and the null hypothesis that this mean is 0 is rejected, with a two-sided *P* value of 0.049. In the same study, both LOR and standard error LOR (SELOR) were available. Therefore, the alternative approach proposed here may also be applied. After directly calculating the ratio LOR/SELOR for each of the 12 studies, the standardized z = 0.9123 with a standard error of 0.4554, which corresponds to a final z statistic of 0.9123/0.4554 = 2.003, yielding a *P* value of 0.0452. The difference (0.0457 vs. 0.0452) must be due to rounding.

Table 2 shows the comparisons between the values reported for the meta-analysis conducted by Gareen et al[5] and the results obtained using the method currently proposed in this paper. In Gareen's study[5] rate ratio (RR) and corresponding 95% CL were reported. In the current study, RR was further converted to LRR (= log[RR]) and its standard error (SERLRR) was derived using the lower and upper CL of Gareen's reported study. Results were very similar using either the reported lower or upper CL. Therefore, only the SERLRR values obtained using the upper CL are included in table 2. Also in table 2, each of the 32 z values was calculated by LRR/SERLRR. Based on the calculated z values, the current proposed method resulted in a test statistic of 3.9147 with a standard error of 0.4999. Furthermore, the final test statistic z = 3.9147/0.4999 = 7.8309 was obtained, yielding a two-sided *P* value of < 0.0001. This is consistent with Gareen's report[5] in which it was concluded that there was a consistent positive association of intrauterine devices with both symptomatic and asymptomatic pelvic inflammatory disease (RR = 4.1, 95% CL = 2.9-5.8, and *P* value < 0.001).

Discussion

Using the proposed method, we were able to consistently reproduce the results of the well-conducted, published meta-analysis reports that were used for comparisons. It has been said that meta-analysis is the analysis of the analyses. Here, an analysis of the analysis of the analyses was conducted. In this study, we propose a simple and robust way of concluding the results of published meta-analyses. It is simple because no sample size-adjustment and variance-adjustment or weighting is necessary. It is robust because the test statistic used consistently follows a standard normal distribution which takes into account the possible differences in effect size (the numerator of b/s_b) and its corresponding variance or standard deviation (the denominator of b/s_b) among studies included in the meta-analysis. The proposed method has value in that it will boost confidence in the results of published meta-analyses if agreement exists. If there is a disagreement between the findings using the method proposed in this paper and the reported meta-analysis results, the author would strongly recommend to readers that they question the conclusiveness of the reported findings. It is important for the conclusion of any meta-analysis results to be further confirmed. For example, the suggested increase in risk of cardiovascular events observed in a meta-analysis was later refuted by a well-conducted, placebo-controlled, randomized clinical trial.[8]

Limitations

The key assumption made for the method proposed herein is that the test statistic utilized follows a standard, normal, independent identical distribution (IID). This is justifiable only when the study sample size is sufficiently large (eg, 50 or more) in each of the studies included in a meta-analysis. Even though an unweighted average of our proposed method can be visualized to be unbiased,[9] it may induce a larger mean square error than is expected when there exist outliers. Also, our proposed method is exactly equivalent to the fixed effect pooled estimate on the standardized effect sizes. Although it remains to be determined whether examples (real data) exist where a conclusive ($P < 0.05$) meta-analysis run using the random effect model is proven inconclusive ($P > 0.05$) by the proposed methods or vice versa (inconclusiveness vs. conclusiveness); a hypothetical example is given in table 3 as an illustration. Table 3 shows that all LORs are equal (OR = 1.5), and the *P* value is < 0.001 (conclusive). However, the proposed method results in a *P* value of 0.1314 (inconclusive), presumably due to the extreme observed z value of 113.2342.

Another limitation is that the proposed approach relies on *P* values and some other parameter estimates (b) presented in a meta-analysis, which determine the applicability of this current approach. Another workable scenario of this proposed approach is when both b and s_b are reported or derivable from a meta-analysis. Finally, only two published data sets were chosen and applied to the current proposed method. However, the proposed method is simple enough to allow readers to adopt and conduct an investigation of their own to any meta-analysis data they have collected or identified.

Table 2. Comparisons between reported meta-analysis in rate ratios and confidence limits and the proposed method.

Gareen et al (2000)[5]	Reported			Proposed		
	RR	RR$_l$	RR$_u$	LRR	SERLRR$_u$	z
	1.9	0.38	9.1	0.6418	0.7992	0.8031
	3.3	2.2	4.9	1.1939	0.2017	5.9196
	2.6	2.1	3.1	0.9556	0.0897	10.6475
	4.4	2.2	9.2	1.4816	0.3763	3.9370
	4.9	2.7	9.0	1.5892	0.3102	5.1233
	2.8	1.5	5.4	1.0296	0.3351	3.0727
	5.8	0.011	3029.0	1.7579	3.1929	0.5505
	2.1	1.4	3.2	0.7419	0.2149	3.4524
	8.6	5.3	14.0	2.1518	0.2486	8.6548
	1.9	0.38	9.3	0.6419	0.8103	0.7921
	6.4	1.5	27.0	1.8563	0.7345	2.5274
	1.9	0.48	7.8	0.6419	0.7206	0.8908
	2.2	1.7	2.7	0.7885	0.1045	7.5460
	2.1	1.6	2.8	0.7419	0.1468	5.0549
	2.9	1.8	4.7	1.0647	0.2464	4.3219
	1.1	0.26	4.6	0.0953	0.7300	0.1306
	2.3	1.2	4.7	0.8329	0.3646	2.2843
	9.3	3.9	23.0	2.2300	0.4620	4.8271
	2.3	0.91	6.0	0.8329	0.4892	1.7026
	2.9	2.1	4.0	1.0647	0.1641	6.4892
	4.1	1.1	15.0	1.4110	0.6618	2.1322
	2.3	1.4	3.8	0.8329	0.2562	3.2514
	17.0	7.8	37.0	2.8332	0.3968	7.1404
	12.0	0.98	135.0	2.4849	1.2349	2.0123
	11.0	3.4	35.0	2.3979	0.5905	4.0605
	1.5	0.77	2.8	0.4055	0.3185	1.2733
	6.7	1.1	40.0	1.9021	0.9116	2.0865
	10.0	3.1	33.0	2.3026	0.6091	3.7800
	132.0	57.0	304.0	4.8828	0.4256	11.4721
	9.0	1.5	54.0	2.1972	0.9142	2.4035
	1.9	1.3	2.8	0.6419	0.1978	3.2443
	12.0	3.4	45.0	2.4849	0.6744	3.6848

RR, rate ratio; RR$_l$, 95% lower confidence limit; RR$_u$, 95% upper confidence limit; LRR, log-rate ratio; SERLRR$_u$, standard error of LRR using RR$_u$.

In summary, the current proposed method can be utilized to safeguard against improper conclusions reported by certain published meta-analyses, due primarily to the issue of heterogeneity among studies included in the meta-analysis. Exploration of more sophisticated statistical methods for situations in which the key assumption applied to the method proposed herein is violated is worthy of pursuing and may benefit from simulations with different levels of heterogeneity. This will certainly enhance and expand the scope of applications beyond the method proposed here.

Table 3. A hypothetical example.

Study	LOR	SELOR	z
S1	0.405465	1.880704	0.215592
S2	0.405465	2.572847	0.157594
S3	0.405465	0.849294	0.477414
S4	0.405465	2.951985	0.137353
S5	0.405465	0.261408	1.551082
S6	0.405465	1.209421	0.335256
S7	0.405465	2.207609	0.183667
S8	0.405465	1.236299	0.327967
S9	0.405465	0.789342	0.513675
S10	0.405465	0.003581	113.2342
S11	0.405465	1.850707	0.219087
S12	0.405465	0.083472	4.857507
S13	0.405465	2.237144	0.181242
S14	0.405465	1.945276	0.208436
S15	0.405465	2.759249	0.146948
S16	0.405465	1.695643	0.239122
S17	0.405465	2.049513	0.197835
S18	0.405465	2.529477	0.160296
S19	0.405465	1.779514	0.227852
S20	0.405465	1.17509	0.34505
S21	0.405465	1.973475	0.205457
S22	0.405465	2.165817	0.187211
S23	0.405465	2.966013	0.136704
S24	0.405465	2.137419	0.189698
S25	0.405465	1.552926	0.261097
S26	0.405465	1.86624	0.217263
S27	0.405465	2.702653	0.150025
S28	0.405465	1.846319	0.219607
S29	0.405465	2.54761	0.159155
S30	0.405465	1.004315	0.403723
S31	0.405465	1.780947	0.227668
S32	0.405465	1.177176	0.344439

LOR, log-odds ratio; SELOR, standard error LOR

References
1. Petitti DB. Meta-Analysis, Decision Analysis and Cost-Effectiveness Analysis: Method for Quantitative Synthesis in Medicine. 2nd ed. New York, NY: Oxford University Press; 1994.
2. Elwood JM. Critical Appraisal of Epidemiological Studies and Clinical Trials. 2nd ed. New York, NY: Oxford University Press; 1998.
3. Hogg RV, Craig AT. Introduction to Mathematical Statistics. 4th ed. New York, NY: Macmillan Publishing Company; 1978.
4. Burr D, Doss H, Cooke GE, Goldschmidt-Clermont PJ. A meta-analysis of studies on the association of the platelet PlA polymorphism of glycoprotein IIIa and risk of coronary heart disease. Stat Med 2003;22:1741–1760.
5. Gareen HF, Greenland S, Morgenstern H. Intrauterine devices and pelvic inflammatory disease: Meta-analyses of published studies, 1974-1990. Epidemiology 2000;11:589–597.
6. Hintz J. NCSS and PASS. Statistical & Power Analysis Software. Kaysville, Utah; 2001. Available at: http://www.ncss.com.
7. SAS, version 9.2, Cary, NC: SAS Institute; 2005.
8. Michele TM, Pinheiro S, Iyasu S. The safety of tiotropium--the FDA's conclusions. N Engl J Med 2010;363:1097–1099.
9. Shuster JJ. Empirical vs natural weighting in random effects meta-analysis. Stat Med 2010;29:1259-1265.

Author Affiliation
Po-Huang Chyou, PhD*

*Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, Marshfield, Wisconsin, USA