J Clin Epidemiol Vol. 52, No. 8, pp. 725–730, 1999

0895-4356/99/$–see front matter
PII S0895-4356(99)00066-9

Exploring Heterogeneity in Meta-Analysis: Is the L'Abbé Plot Useful?

Fujian Song
NHS Centre for Reviews and Dissemination, University of York, York, UK

ABSTRACT. By using a published meta-analysis as an example, this paper discusses the use of L'Abbé plot for investigating the potential sources of heterogeneity in meta-analysis. As compared with other graphic procedures, the L'Abbé plot is useful to identify not only the studies having different results from other studies, but also the study arms that are responsible for such differences. This may be important for determining the focus of heterogeneity investigations. Results of stochastic simulation indicate that, purely because of random variation, studies with event rates of around 50% are more likely to be identified as outliers in a L'Abbé plot. This paper also demonstrates that different methods may identify different trials as "outliers" in meta-analysis. J CLIN EPIDEMIOL 52;8:725–730, 1999. © 1999 Elsevier Science Inc.

KEY WORDS. Meta-analysis, heterogeneity, L'Abbé plot, outlier

It has been widely recognized that meta-analysis is not only important for estimating average treatment effect but also crucial for investigating potential sources of heterogeneity [1,2]. Heterogeneity in meta-analysis may be defined as "the variability or differences between studies in the estimates of effects" and can be categorized as statistical heterogeneity, methodological heterogeneity, and clinical heterogeneity [3]. The potential sources of clinical heterogeneity include different study participants, different interventions, and different outcome measures across individual studies [4].

Anello and Fleiss suggested that a meta-analysis may be analytic or exploratory [5]. An analytic meta-analysis aims to estimate overall effect by pooling results from individual studies, while an exploratory meta-analysis aims to explain why the results of individual studies are different. The results from exploratory meta-analysis may not always be reliable enough to guide clinical decision making. However, it is important to explore the available data thoroughly and to examine possible alternative explanations for observed results. Such efforts may provide empirical evidence for identifying future research priorities.

To investigate the sources of heterogeneity in a meta-analysis, the association between the treatment effect and other study characteristics is often examined by using methods such as meta-regression and subgroup analysis [1,6].Since the treatment effect in controlled trials is generally measured by the relative or absolute difference between comparison groups, the original results from individual groups may have been overlooked in many meta-analyses.

L'Abbé and colleagues suggested a method for showing variations in observed results by plotting the event rate in the treatment group on the vertical axis and that in the control group on the horizontal axis [7]. By using an example of a published meta-analysis and by method of computer simulation, this article discusses the use of L'Abbé plot for investigating the potential sources of heterogeneity in meta-analysis.

EXAMPLE

Dolan-Mullen and colleagues conducted a meta-analysis to assess the effect of prenatal smoking intervention [8]. By pooling results from 11 randomized controlled trials, it was found that intervention increased the rate of smoking cessation during pregnancy (overall rate ratio 2.08; 95% confidence interval 1.74, 2.49), although statistical test showed significant heterogeneity in rate ratio across the trials ($\chi^2 = 35.26$, df = 10, $P < 0.001$). Figure 1, known as a Forrest plot [9], shows the results of the primary trials in this meta-analysis.

Figure 2 is a L'Abbé plot for the rate of smoking cessation in both the treatment and the control groups. It can be seen that the rate of smoking cessation varies greatly in both the treatment groups (4.9–31.9%) and in the control groups (1.4–17.2%). Using the likelihood method suggested by Martuzzi and Hills [10], the degree of heterogeneity in the rates of smoking cessation was the same between the treat-

*Address correspondence to: Fujian Song, Ph.D., Senior Research Fellow, NHS Centre for Reviews and Dissemination, University of York, York, YO10 5DD, UK. Tel: (44) 1904 433656, Fax: (44) 1904 433661, E-mail: <fs4@york.ac.uk>.

Accepted for publication on 30 March 1999.

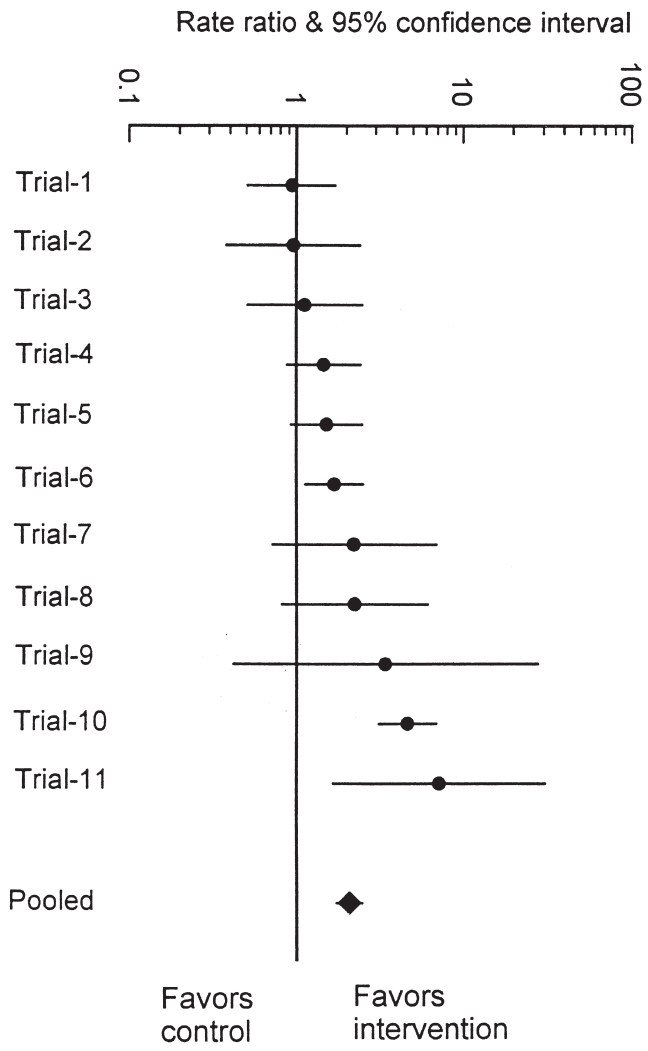

FIGURE 1. Forrest plot for relative effect of prenatal smoking cessation intervention: results of eleven trials in Dolan-Mullen *et al.*'s meta-analysis [8].

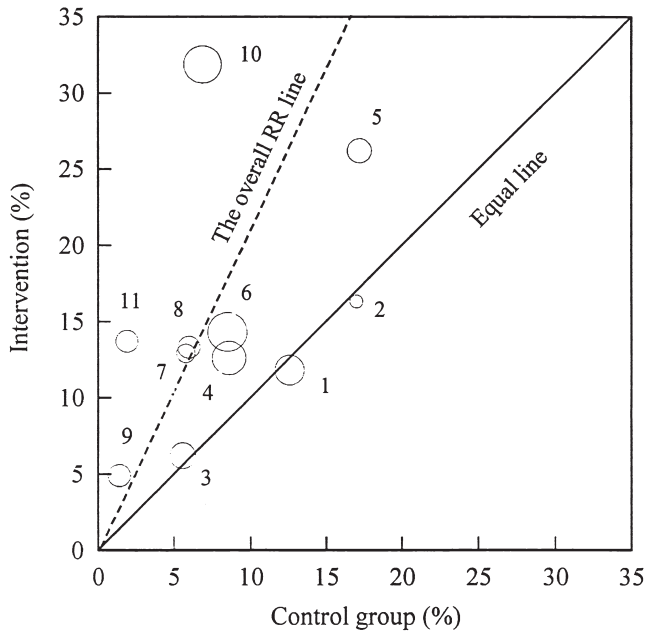

FIGURE 2. L'Abbé plot for rates of smoking cessation in the intervention and control group: results of eleven trials in Dolan-Mullen *et al.*'s meta-analysis [8]. Each circle represents an individual trial and larger circles represent trials with more participants. Solid diagonal line indicates that the rate of smoking cessation is equal in the two arms within trials. The dotted line may be called "the overall RR line," as it represents a rate ratio of 2.08 which is estimated by pooling the results of all eleven studies.

ment arms and between the control arms (variance of rate ratio is about 0.205 for both arms).

Table 1 shows the study settings and the treatment intensity in individual studies included in Dolan-Mullen *et al.*'s meta-analysis. The different intervention strategies may be associated with heterogeneous results from included studies. For example, the rate of smoking cessation in the treatment group was highest in Trial-10 (31.9%) in which the intervention was composed of at least one personal visit and monthly phone contact plus biweekly mailed materials with homework. On the other hand, the rate of smoking cessation in the treatment group was only 4.9% in Trial-9 in which the intervention was much less intensive, including mainly educational videotape or self-help pamphlet and brief interaction with health educator.

It seems that "usual care" provided in different settings may also be an important source of heterogeneity in this meta-analysis. In two trials that were conducted in Health Maintenance Organizations (HMOs) (Trial-2 and Trial-5), the rates of smoking cessation were high in both the treatment groups (16.3 and 26.2%, respectively) and the control groups (17.0% and 17.2%, respectively) (see Figure 2), although the estimated relative effects in these two trials were either negative or very small (see Figure 1). Conversely, in two trials that were conducted in the urban hospital clinics (Trial-7 and Trial-9), the rates of smoking cessation were very low in the treatment groups (12.9% and 4.9%, respectively) and in the control groups (5.8% and 1.4%, respectively), although the relative effects estimated in these two trials were greater than that in the two trials conducted in HMOs. Therefore, by comparing relative effects in Figure 1, it may be suspected that prenatal smoking cessation interventions in HMOs were less efficacious than those in urban hospital clinics. However, Figure 2 reveals that the rates of smoking cessation in the comparison groups were much higher among the participants in HMOs than that in the urban hospital clinics. The less favourable relative effects in trials conducted in HMOs may be partially explained by the more effective "usual care" provided (see Table 1).

The results of logistic regression analysis [11] confirm that the care intensity was positively associated with a

TABLE 1. Intensity of smoking cessation intervention and setting of health care services in studies included in Dolan-Mullen *et al.*'s meta-analysis[a]

Study	Control events/N	Intervention events/N	Control intensity[b]	Intervention intensity[b]	Setting
Trial-1	16/125	19/127	1	4	Private (U.S.A.)
	10/105	10/98	1	4	Public (U.S.A.)
Trial-2	8/47	7/43	2	4	HMO (U.S.A.)
Trial-3	11/198	12/193	NA	5	Public + private (U.S.A.)
Trial-4	18/209	56/444	1	2	Public (Sweden)
Trial-5	20/116	33/126	3	6	HMO (U.S.A.)
Trial-6	35/414	57/400	3	6	Public (U.S.A.)
Trial-7	4/69	9/70	1	3	Public (U.S.A.)
Trial-8	5/84	12/90	2	4	Public + private (Canada)
Trial-9	1/70	4/71	1	3	Public (U.S.A.)
		2/52		2	
Trial-10	79/395	167/388	1	5	Public + private (U.S.A.)
Trial-11	2/104	14/102	2	4	Public (U.S.A.)
		6/103		4	

Abbreviation: NA = not available; HMO = health maintenance organization.

[a]Data are obtained from ten published studies. Trial-3 is not published and there were no details about the intervention in the control group. Data from Trial-1 could be separated for public and private patients; Trial-9 and Trial-11 had two intervention groups. For reasons unknown, data from published trials are different from that reported in Dolan-Mullen *et al.*'s meta-analysis for Trial-1 and Trial-10.

[b]Intensity of interventions for smoking cessation was determined arbitrarily (mainly for the purpose of illustration) according to the following method: Counselling: 1 = once, <10 min; 2 = twice or >10 min; 3 = more than twice. Information package: 1 = booklets, once; 2 = booklets, mailing frequently; 2 = manual for guiding behaviour changes; 1 = videotapes; 1 = audiocassette; 2 = personal letters.

higher rate of smoking cessation and participants in the public clinics were less likely to stop smoking as compared to those in the private clinics or HMOs (Table 2).

STOCHASTIC SIMULATION

Stochastic simulation technique was used to explore factors related to the random discrepancies between study points and the overall RR line in the L'Abbé plot. The simulation results reveal that random variation in the distance between a study point and the overall RR line is negatively related to the sample size (Figure 3). It is also observed that the random variation in the distance is greatest when the event rates in the control and the treatment group are 50%.

Therefore, the absolute distances between individual studies and the overall RR line may not be comparable unless the impact of random variation has been adjusted for according to the number of participants and the event rates. For instance, Figure 3 could be used to estimate the standard deviation for a given control rate and the number of subjects in individual studies in Dolan-Mullen *et al.*'s meta-analysis. Then the estimated standard deviations (SDs) can be used to standardize the absolute distances between individual studies and the overall RR line. Taking Trial-1 as an example (Table 3), its distance form the overall RR line is 0.0623, the control rate is 0.126 and the average number of subjects is 157. It can be estimated from Figure 3 that the standard deviation of distance is about 0.032. Then the standardized distance between Trial-1 and the overall RR line is 1.947 (i.e., 0.0623/0.032).

Table 3 presents the standardized distance between individual studies and the overall RR line in Dolan-Mullen *et al.*'s meta-analysis. In four of the eleven individual studies, the distance from the average line may be considered as statistically significant at the 10% level because these studies are more than 1.64 standard deviations away from the average.

IDENTIFYING OUTLIERS

In Dolan-Mullen *et al.*'s meta-analysis [8], L'Abbé plot was used to identify "outliers" according to the distance be-

TABLE 2. Results of logistic regression analysis for exploring heterogeneity in Dolan-Mullen *et al.*'s meta-analysis

Model	Coefficients (standard error)	Odds ratio (95% CI)
Intensity	0.182 (0.0251)	1.20 (1.14, 1.26)
Setting, public[a]	−0.598 (0.130)	0.55 (0.43, 0.71)
Setting, public + private[a]	0.296 (0.129)	1.34 (1.05, 1.73)

[a]HMO/private as the baseline.

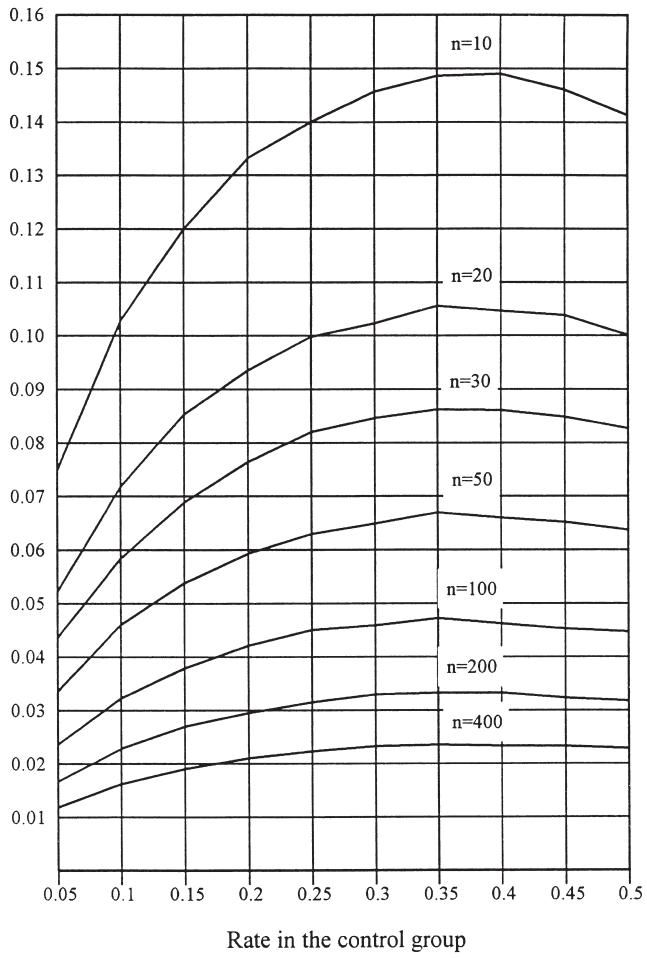

FIGURE 3. Results of stochastic simulation: standard deviation of distance between study points and the overall RR line under different baseline rates in the control group and number of participants in each group (RR = 2.08; simulation times = 10,000).

tween a trial and the overall RR line.Then the "outliers" were excluded one by one until the statistical test of heterogeneity across trials was no longer significant. Trial-2 should be excluded from the analysis because it is farthest from the overall RR line (Figure 2). However, the standardized distance is greatest for Trial-10 (Table 3). On the other hand, according to the Forrest plot (Figure 1), the difference in the relative effect is greatest between Trial-11 and the overall estimate (Table 3).

By reviewing meta-analyses published in *Journal of American Medical Association* (JAMA) and *British Medical Journal* (*BMJ*) in 1995–1996, it was found that 11 of the 31 meta-analyses reported enough data for a L'Abbé plot, although this plot had not been used in any of these meta-analyses. Among these eleven meta-analyses, the outlying trials identified by the Forrest plot were different from that identified by the adjusted and unadjusted L'Abbé method in 6 and 7 meta-analyses respectively. Furthermore, identified "outliers" were different between the adjusted and unadjusted L'Abbé method in 6 meta-analyses.

Three of these 11 meta-analyses showed statistically significant heterogeneity across included studies [12–14]. Table 4 shows the results after excluding "outliers" in the three meta-analyses and Dolan-Mullen *et al.*'s meta-analysis according to different methods. In one meta-analysis [13], the treatment effect after excluding outliers according to the standardized distance in the L'Abbé plot becomes much smaller and no longer statistically significant, as compared to the significant results after excluding outliers identified by other methods (Table 4). Clearly, it is not unusual that different methods may identify different trials as outliers, and the conclusions of a meta-analysis may change by excluding different studies.

DISCUSSION

By examining the rates of smoking cessation in the L'Abbé plot, it is suggested that variations in the intensity of interventions, "usual care" controls, and study settings are potential sources of heterogeneity in Dolan-Mullen *et al.*'s meta-analysis. However, the usefulness of the L'Abbé plot in this example does not necessarily mean that it can reveal important sources of heterogeneity in all meta-analyses, because the potential causes of heterogeneity may be different and data may not be available in other cases.

The results from stochastic simulation demonstrate that random variation in the distance between a study point and the overall RR line in a L'Abbé plot is associated with the number of subjects and the event rates. The distances between studies and the overall RR line should be standardized before making comparisons for investigating clinical or methodological heterogeneity. For example, this article has used standard deviations estimated from computer simulations to adjust distances between studies and the overall RR line (Figure 3). This method is only an approximation but useful until an analytic method is developed.

As other methods that are currently available to investigate heterogeneity in a meta-analysis, the usefulness of L'Abbé plot is limited by the available data reported in the primary studies. Incorrect "outliers" may be identified if random variation in the distance between studies and the overall RR line has not been standardized. In addition, simple regression analysis of the treatment rate on the control rate will yield misleading result [15]. To have a more appropriate graphic illustration of heterogeneity in a meta-analysis, the range of scales used on the vertical axis and on the horizontal axis in a L'Abbé plot should be the same, and the area of point for each individual study should correspond to its sample size.

The L'Abbé plot may have been under-used for exploring causes of heterogeneity in meta-analyses. For example, it was not used in 31 meta-analyses that were published in *JAMA* and *BMJ* in 1995–1996. On the other hand, the

TABLE 3. Standardized distance and difference in relative effect for each study in a meta-analysis[a]

Study	Control event/N	Intervention event/N	Distance[b]	SD[c]	Standardized distance	Difference in relative effect[d]
Trial-1	22/175	16/136	0.0623	0.032	1.947	−0.794
Trial-2	8/47	7/43	0.0829	0.060	1.382	−0.773
Trial-3	11/198	12/193	0.0231	0.018	1.283	−0.619
Trial-4	18/209	56/444	0.0230	0.014	1.643	−0.354
Trial-5	20/116	33/126	0.0419	0.037	1.132	−0.314
Trial-6	35/414	57/400	0.0144	0.015	0.960	−0.208
Trial-7	4/69	9/70	0.0035	0.030	0.117	0.065
Trial-8	5/84	12/90	0.0041	0.027	0.152	0.074
Trial-9	1/70	6/123	0.0083	0.022	0.377	0.494
Trial-10	27/392	124/389	0.0760	0.014	5.429	0.800
Trial-11	2/104	14/102	0.0421	0.022	1.914	1.233

Abbreviation: SD = standard deviation.
[a]Number of smoking cessation events and participants are from Dolan-Mullen *et al.*'s meta-analysis.
[b]Distance between study point and the overall RR line (RR = 2.08) in the L'Abbé plot (Figure 2).
[c]Estimated from Figure 3, according to the control rate and average number of subjects in each group.
[d]Log rate ratio of individual study minus overall log rate ratio (Figure 1).

L'Abbé plot was inappropriately used to solve statistical heterogeneity by identifying and excluding "outliers" in some meta-analyses [8,16–18]. The credibility of a meta-analyses cannot be improved simply by excluding "outliers." This is because, firstly, eligible trials are excluded according to their results, not their designs. Secondly, excluding trials will result in a lower power statistical test of heterogeneity. Thirdly, research evidence is wasted or under-used if the chances to explore the clinical and/or methodological heterogeneity are missed. Finally, there is the practical problem that different methods may identify different trials as outliers, and the conclusions of a meta-analysis may change by excluding different studies.

In summary, the L'Abbé plot as a graphical method is helpful for identifying not only the studies that are of different results from other studies, but also the study arms that are responsible for such differences. This may be important for determining the focus of heterogeneity investigations.

TABLE 4. Results of different methods used for excluding outliers from meta-analyses

Meta-analysis and method	No. of trials excluded	Heterogeneity test	Rate ratio (95% CI)
Dolan-Mullen *et al.* [8]			
All trials (n = 11)	0	P < 0.001	2.08 (1.72–2.50)
Absolute distance	2	P = 0.316	1.59 (1.28–1.96)
Standardized distance	1	P = 0.327	1.56 (1.27–1.92)
Forrest plot	2	P = 0.680	1.47 (1.19–1.82)
McQuay [12]			
All trials (n = 3)	0	P < 0.001	2.91 (2.20–3.86)
Absolute distance	2	Only 1 trial left	2.92 (2.12–4.01)
Standardized distance	2	Only 1 trial left	2.92 (2.12–4.01)
Forrest plot	2	Only 1 trial left	2.92 (2.12–4.01)
Barza *et al.* [13]			
All trials (n = 17)	0	P = 0.012	0.76 (0.61–0.95)
Absolute distance	2	P = 0.113	0.74 (0.58–0.93)
Standardized distance	2	P = 0.167	0.98 (0.74–1.29)
Forrest plot	3	P = 0.115	0.75 (0.60–0.95)
Linde *et al.* [14]			
All trials (n = 12)	0	P < 0.001	2.47 (2.03–2.99)
Absolute distance	2	P = 0.124	2.87 (2.26–3.64)
Standardized distance	2	P = 0.124	2.87 (2.26–3.64)
Forrest plot	4	P = 0.377	2.52 (1.98–3.22)

"Outliers" are excluded until the heterogeneity test is no longer statistically significant at the 10% level.

However, the distances between individual studies and the average in a L'Abbé plot should be standardized before making comparisons for exploring clinical or methodological heterogeneity. Because appropriate use of the available methods for exploring heterogeneity in meta-analysis may provide important results relevant to clinical practice or future research, statistical heterogeneity in meta-analysis should not be simply solved by excluding outlying trials.

References

1. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. **BMJ** 1994; 309: 1351–1355.
2. Lau J, Ioannidis JP, Schmid CH. Summing up evidence: one answer is not always enough. **Lancet** 1998; 351: 123–127.
3. Mulrow CD, Oxman AD. **Cochrane Collaboration Handbook [updated 9 December 1996].** Available in The Cochrane Library [database on disk and CDROM]. The Cochrane Collaboration (Issue 1); 1997.
4. Bailey KR. Inter-study differences: how should they influence the interpretation and analysis of results? **Stat Med** 1987; 6: 351–358.
5. Anello C, Fleiss JL. Exploratory or analytic meta-analysis: Should we distinguish between them? **J Clin Epidemiol** 1995; 48: 109–116.
6. Berlin JA, Antman EM. Advantages and limitations of metaanalytic regressions of clinical trials data. **Online J Curr Clin Trials** 1994; Doc No 134.
7. L'Abbé KA, Detsky AS, O'Rourke K. Meta-analysis in clinical research. **Ann Intern Med** 1987; 107: 224–233.
8. Dolan-Mullen P, Ramirez G, Groff J. A meta-analysis of randomized trials of prenatal smoking cessation interventions. **Am J Obstet Gynecol** 1994; 171: 1328–1334.
9. Moher D, Jadad AR, Klassen TP. Guides for reading and interpreting systematic reviews. III. How did the authors synthe size the data and make their conclusions? **Arch Pediatr Adolesc Med** 1998; 152: 915–920.
10. Martuzzi M, Hills M. Estimating the degree of heterogeneity between event rates using likelihood. **Am J Epidemiol** 1995; 141: 369–374.
11. Thompson SG. Controversies in meta-analysis: the case of the trials of serum cholesterol reduction. **Stat Methods Med Res** 1993; 2: 173–192.
12. McQuay H, Carroll D, Jadad AR, Wiffen P, Moore A. Anticonvulsant drugs for management of pain: A systematic review. **BMJ** 1995; 311: 1047–1052.
13. Barza M, Ioannidis JP, Cappelleri JC, Lau J. Single or multiple daily doses of aminoglycosides: a meta-analysis. **BMJ** 1996; 312: 338–345.
14. Linde K, Ramirez G, Mulrow CD, Pauls A, Weidenhammer W, Melchart D. St John's wort for depression—An overview and meta-analysis of randomised clinical trials. **BMJ** 1996; 313: 253–258.
15. Sharp SJ, Thompson SG, Altman DG. The relation between treatment benefit and underlying risk in meta-analysis. **BMJ** 1996; 313: 735–738.
16. Fiore MC, Smith SS, Jorenby DE, Baker TB. The effectiveness of the nicotine patch for smoking cessation: A meta-analysis. **JAMA** 1994; 271: 1940–1947.
17. Munnangi S, Sonnenberg A. Colorectal cancer after gastric surgery: A meta-analysis. **Am J Gastroenterol** 1997; 92: 109–113.
18. Lachner G, Engel R. Differentiation of dementia and depression by memory test: A meta-analysis. **J Nerv Ment Dis** 1994; 182: 34–39.