



# American Journal of EPIDEMIOLOGY

Volume 140

Number 9

November 1, 1994

Copyright © 1994 by The Johns Hopkins University

School of Hygiene and Public Health

Sponsored by the Society for Epidemiologic Research

---

## POINT/COUNTERPOINT: META-ANALYSIS OF OBSERVATIONAL STUDIES

---

### Meta-analysis/Shmeta-analysis

Samuel Shapiro

---

*Editor's note: This and the following two commentaries (pages 779 and 783) are based on presentations made at the symposium "Meta-analysis of Observational Studies" at the 26th Annual Meeting of the Society for Epidemiologic Research, Keystone, Colorado, June 16–18, 1993. Dr. Shapiro's response to Drs. Petitti and Greenland follows on page 788.*

---

But now it seems that many of the strong relationships (i.e., high relative risks) have been identified, and we are left to struggle in an era of weak associations. More and more we spend our time trying to sort out whether relative risks of 1.3 or 1.6 reflect causality, or whether they simply reflect one or another kind of bias. Should we even care about weak associations? Are they important? They are important and we do need to care: For common illnesses, weak associations can translate into many thousands of cases.

—Lynn Rosenberg, *Presidential Address, 26th Annual Meeting of the Society for Epidemiologic Research, Keystone, Colorado, 1993*

---

Received for publication August 17, 1993, and in final form May 20, 1994.

Abbreviations: CI, confidence interval; IARC, International Agency for Research on Cancer.

From the Slone Epidemiology Unit, 1371 Beacon Street, 3rd Floor, Brookline, MA 02146. (Reprint requests to Dr. Shapiro at this address.)

In this presentation, I will not consider the statistical methods of meta-analysis or its application to randomized controlled trials. Nor will I consider the combined analysis of "raw" data. My focus will be solely on the meta-analysis of published data from nonexperimental (observational) studies.

Let me open by giving a definition of meta-analysis, bowdlerized from a definition of risk analysis that appeared in the *Wall Street Journal*. I think it gets to the heart of the matter: "Meta-analysis begins with scientific studies, usually performed by academics or government agencies, and sometimes incomplete or disputed. The data from the studies are then run through computer models of bewildering complexity, which produce results of implausible precision" (1).

Why has meta-analysis enjoyed such an exponential surge of popularity? The answer, I think, is that meta-analysis offers the Holy Grail of attaining statistically stable estimates for effects of low magnitude. In so doing, it ignores what is an absolute limit to epidemiologic inference. In the nonexperimental domain, epidemiologic

---

This commentary is based on a presentation made at the 26th Annual Meeting of the Society for Epidemiologic Research, Keystone, Colorado, June 16–18, 1993. The wording has been adapted to make it suitable for written presentation. Otherwise, there have been no material changes.

methods can only yield valid documentation of large relative risks. Relative risks of low magnitude (say, less than 2) are virtually beyond the resolving power of the epidemiologic microscope: We can seldom demonstrably eliminate all sources of bias, and we can never exclude the possibility of unidentified and uncontrolled confounding. If many studies—preferably based on different methods—are nevertheless congruent in producing markedly elevated relative risks, we can set our misgivings aside. If, however, many studies produce only modest increases, those increases may well be due to the same biases in all of the studies.

It may or may not be true, as Lynn Rosenberg suggested in her presidential address, that we are running out of large relative risks; and it is certainly true that small ones may be important. However, it does not follow from this that meta-analysis can resolve the dilemma. If the same systematic biases are indeed present across a range of studies, the only effect of meta-analysis is to reinforce them, to produce spurious statistical stability, and thereby to discourage further research. Moreover, it is particularly in the domain of low-level associations that the temptation to perform a meta-analysis is the strongest. No one would suggest a meta-analysis to determine whether cigarettes cause lung cancer, because the answer is obvious.

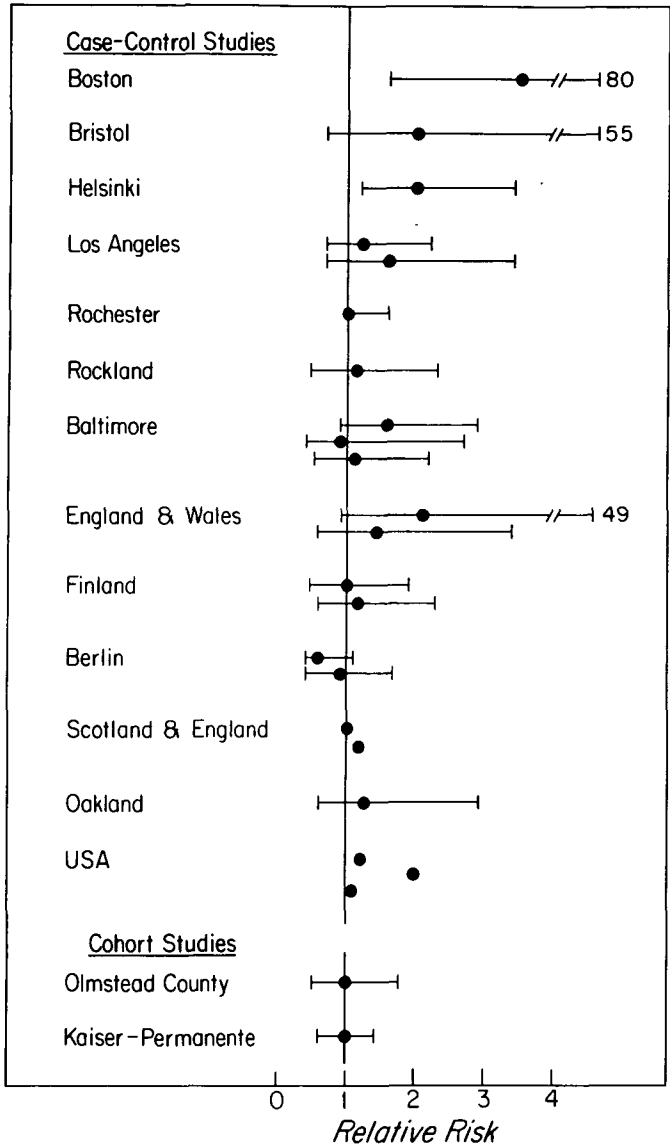
The temptation to undertake a meta-analysis becomes particularly strong when an “eyeball” review of an array of relative risks, mostly nonsignificant, suggests a result that may well turn out to be significant on meta-analysis. For example, consider the risk of breast cancer in relation to the use of reserpine (2–6). An initial study raised the hypothesis that reserpine (then a popular antihypertensive) increases the risk (2). Two further studies were then conducted (3, 4), and all three were published back-to-back in 1974. In 1980, after 12 additional studies had been published, all 15 were reviewed by the World Health Organization International Agency for Research on Cancer (IARC) (5) (figure 1), which concluded that “the studies consid-

ered to be most satisfactory, methodologically, showed little or no evidence of an increased risk” (p. 234). Yet, as is obvious from an “eyeball” inspection of figure 1, a meta-analysis—had the approach been in use at the time—would almost certainly have produced a significantly elevated summary relative risk, perhaps with significant heterogeneity, but that could have been taken care of in a random effects model.

That meta-analysis might also have discouraged the further research that was published after the IARC review appeared (6) (table 1). In a more detailed, more rigorous, and much larger repetition of the original study, although carried out in virtually the same population and with essentially the same methods, it emerged that the initial association that stimulated the entire cascade of studies was almost certainly a statistical fluke. For practical purposes, the final repeat study settled the matter. Yet an earlier meta-analysis, had it been performed, would have conferred false validity to a spurious result.

Despite the recognition that the possibility of bias can seldom be eliminated and that possible confounding can never be, some meta-analysts have proceeded on the assumption that exactly those objectives can be achieved by assigning appropriate quality weights to each of a series of individual studies. An example with which I am closely familiar is alcohol and breast cancer, because our group was responsible for publishing both the association that stimulated several further studies (7) and a subsequent null result (8). In the interim, a quality-weighted meta-analysis of the first 16 published studies (9) yielded summary relative risk estimates for the daily intake of 24 g of alcohol of 1.7 (95 percent confidence interval (CI) 1.4–2.2), derived from 12 case-control studies, and 1.4 (95 percent CI 1.0–1.8), derived from four cohort studies. Those estimates were interpreted as strongly supportive evidence for an association.

Fifteen quality criteria were applied. Each study was scored according to how



**FIGURE 1.** Risk of breast cancer in relation to reserpine use: relative risk estimates and 95% confidence intervals in 13 case-control and two cohort studies. Confidence intervals are not provided if they were not published or could not be estimated from the published data. Some studies estimated more than one relative risk.

well it fulfilled the relevant criterion, and the scores were used as weights. For illustrative purposes, attention will be confined to the case-control studies, and the quality criteria are given in table 2. Each criterion is debatable, as is the scoring system itself, and the question is whether the weighting enhanced the validity of the meta-analysis.

Table 3 gives the ranking of the case-control studies according to quality score,

together with the highest level of alcohol intake analyzed in each study, and its associated relative risk. There was no correlation between the ranks and the relative risk estimates. The scores made no difference. This is hardly surprising, since the criteria that were applied were inadequate, largely subjective, and incomplete. For example, in the study with the highest score (10), the cases and controls were drawn from a

TABLE 1. Data from two case-control studies on reserpine use and breast cancer

Study*	No of subjects	Reserpine use		Relative risk	95% confidence interval
		No.	%		
Boston Collaborative Drug Surveillance Program (2), 1974					
Cases	150	11	7	3.5	1.6-8.0
Controls	600	13	2		
Shapiro et al. (6), 1976-1982					
Cases	1,881	65	3	0.8	0.5-1.1
Controls	1,523	64	4		

\* The reference number is given in parentheses.

screening program: the possible selection biases thereby introduced were not taken into account in assigning the scores. In addition, the information was obtained by mailed questionnaire after an interval of at least 5 years after the diagnosis of breast cancer, and at a time when the hypothesis was known and publicized. For an exposure as difficult to record as alcohol intake, information bias was likely. In a narrative review, the relative risk estimate of 1.7 would be considered as weak support, at most, for a causal inference.

Table 3 also illustrates two further points:

1. The highest dose levels varied considerably among the studies and were estimated differently. In addition, dissimilar categories of alcohol intake were treated in the meta-analysis as if they were the same (e.g., current drinking; recent drinking; habitual drinking of wine; drinking during the previous 5 years; weighted average of three periods during life; intake 3 years previously); and there was misclassification, inasmuch as several assumptions had to be made to convert those categories into grams per day of alcohol consumed.
2. There was no correlation between the relative risk estimates among the studies and the amount of alcohol consumed. For example, the estimate was 1.1 (95 percent CI 0.3-1.8) for an intake of  $\geq 43$  g/day in the data of Webster et al. (11), but 16.7 (95 percent CI 3.1-89.7) for an intake of  $\geq 8$  g/day in the data of Talamini et al. (12). Despite the lack of correlation, the data were deemed to be combinable.

One question that arises is why the meta-analysis of a series of studies that were so different in their particulars nevertheless

produced a statistically significant summary effect. That is not a difficult question to answer: It is likely that various sources of bias and confounding had effects that were generally in the same direction among the studies; the net effect could readily have been a summary relative risk estimate of 1.7. Alcohol intake is uniquely susceptible to misclassification and biased reporting. Information bias was possible and even likely, and in the same direction, in most of the case-control studies. Plausible sources of selection bias, also operating in the same direction (for example, detection bias due to screening), could also have been present in more than one study.

As to confounding, the determinants both of alcohol use and of breast cancer are largely unknown, but obviously changing. For alcohol intake, the determinants vary according to age, calendar time, place, and culture, and many of them are unmeasurable or incompletely measurable. If unidentified confounding was present, it is possible that the same confounders could have beset the majority of the studies and have biased the relative risk estimates in the same direction. For example, in many societies, high socioeconomic status is related both to alcohol intake and to breast cancer risk, and it cannot be fully measured or controlled.

If the foregoing considerations were still insufficient to cast doubt on the validity of the meta-analysis, the aggregated data had the following additional faults: both in the meta-analysis and in the individual studies,

**TABLE 2. Criteria used to evaluate the design and data analysis of 12 case-control studies of alcohol consumption and breast cancer and the percentage given favorable quality scores\***

Criteria	% with favorable score†
Applied to case-control studies with hospital controls ( <i>n</i> = 7)	
1. Did at least 90% of hospital controls have conditions other than any of the following: cancer of the liver, oral cavity, larynx, esophagus, or large bowel; myocardial infarction; cerebrovascular disease; trauma; fractures; peptic ulcer disease; or other conditions that involved upper gastrointestinal tract blood loss, cirrhosis, pancreatitis, alcoholism, or alcoholic hepatitis?	57
2. Is it likely that the referral pattern for the control diseases was similar to the referral pattern for the case diseases?	57
Applied only to case-control studies with community controls ( <i>n</i> = 5)	
3. Was the response rate among the controls at least 70%?	60
4. Were the controls people who, had they developed the disease under study, would have been cases?	100
Applied to all case-control studies ( <i>n</i> = 12)	
5. Were the data collected in a similar manner for cases and controls?	75
6. Were all cases interviewed within 6 months of diagnosis?	33
7. Was the same interview schedule used for cases and controls?	75
8. Was the interviewer blinded with respect to the case-control status of the person interviewed?	8
9. Was the time period within which cases and controls were interviewed the same?	67
10. Were the same exclusion criteria applied to cases and controls?	25
11. If the study was a matched case-control study, did the authors either conduct a matched analysis, show that an unmatched analysis was equivalent to a matched analysis and present an unmatched analysis, or adequately account for the matching factors in an unmatched analysis?	56
Applied to all studies ( <i>n</i> = 16)	
12. Was the diagnosis of cancer histologically confirmed in at least 90% of the cases?	75
13. In the analyses, did the authors control for potential confounding by classic breast cancer risk factors in addition to age?	88

\* The data are from Longnecker et al. (9).

† The highest scores ("favorable") were given when the original investigators explicitly supplied information indicating that the answer to the criterion question was affirmative. For most criteria, up to 5 points were awarded; for question 6, the maximum was 3 points. When the answer to a question was not clearly yes or no, then an intermediate score was given. For an answer of unclear/probably yes, 2 points were given, and for unclear/probably no, 1 point was given; zero points were given when the answer was clearly no. The total number of points awarded to a given study was a function of which criteria were applicable and how many points were awarded for each of the applicable criteria. For example, a case-control study with hospital controls would have a total possible methods score of 48 points (questions 1, 2, 5, 7-10, 14, and 15, 5 points each; question 6, 3 points) and a total possible data analysis score of 5 or 10, depending on whether a matched study design was used (questions 11 and 15, 5 points each).

consistency (for example, within strata of age or menopausal status) was incompletely evaluated or not evaluated at all; the categories used to evaluate dose-response effects were variously defined and variously assessed in the different studies; and possible duration effects were hardly exam-

ined. In short, it boggles the mind to suggest that data concerning an issue as complex as alcohol and breast cancer can meaningfully be combined or interpreted as evidence for a causal association.

A second meta-analysis spanning 38 studies has now been published (13). There

**TABLE 3. Quality ranking of 12 case-control studies of alcohol consumption and breast cancer and relative risk estimates for the highest level of alcohol intake analyzed in each study\***

Source†	Dosage (g/day)	RR
Harvey et al., 1987	≥26	1.7
Rosenberg et al., 1982	≥7	2.0
Webster et al., 1983	≥43	1.1
Paganini-Hill and Ross, 1983	≥26	1.0
Byers and Funch, 1982	>11	1.1
Rohan and McMichael, 1988	>9	1.6
Talamini et al., 1984	>7	16.7
O'Connell et al., 1987	≥2	1.5
Harris and Wynder, 1988	>15	0.9
Le et al., 1984	>34	1.2
La Vecchia et al., 1985	>39	2.1
Begg et al., 1983	>13	1.4

\* The data are from Longnecker et al. (9).

† Listed in quality score rank, highest to lowest.

was significant heterogeneity. With a random effects model, the summary relative risk estimate for the consumption of 24 ounces of alcohol was 1.24 (95 percent CI 1.15–1.34). The second meta-analysis differed from the first in certain particulars, but not in concept. However, perhaps because the data were heterogeneous, or because the summary relative risk point estimate was closer to unity than it had been previously, it is now acknowledged that “the relationship is, on average so modest that whether alcohol consumption causes breast cancer will be difficult to determine with epidemiologic data of the type now available” (14, p. 693).

Serious advocates of meta-analysis have been at pains to point out that the quality of the information yielded by that approach cannot transcend the quality of the individual studies (15). They argue that the technique can nevertheless provide useful quantitative information if the limits are explicitly acknowledged: a bad meta-analysis, the reasoning goes, does not denote a bad concept. I disagree. I believe the concept is bad. But even if, for the sake of the argument, one concedes the point, in practice, the need to acknowledge the limits has been ignored. Meta-analysis is not being used in the way that its cautious advocates recommend.

The next example is taken from a review article on breast cancer (16), and from subsequent correspondence (17). In the review, it was stated that the “widespread use of estrogen-replacement therapy has almost certainly contributed to the higher incidence [of breast cancer] among postmenopausal women” (16, p. 323). In the correspondence that followed (17), the authors explained that “in summarizing a large literature on postmenopausal estrogen-replacement therapy, a full discussion of all publications is not possible. We believe it was appropriate to cite a meta-analysis of 23 primarily case-control studies [18].”

The findings from that meta-analysis (18) are presented in table 4. Among all 23 studies, the summary relative risk estimate for estrogen use was 1.01, but the data were

**TABLE 4. Findings from a meta-analysis of breast cancer and estrogen use\***

	No. of studies	Relative risk	95% confidence interval	p for heterogeneity†
Ever-use: all studies	23	1.01	0.95–1.08	0.006
Ever-use: all studies with adjusted results‡	12	1.05	0.97–1.14	0.006
Ever-use: postmenopausal women	12	0.96	0.89–1.05	0.07
Ever-use: postmenopausal women, adjusted for type of menopause	7	0.99	0.90–1.08	0.04
Highest dose category	?	1.04	0.88–1.24	0.00003
Highest dose category excluding one study	?–1	1.28	1.06–1.54	0.09
Ever-use: positive family history	?	1.25	0.83–1.88	0.006

\* The data are from Armstrong (18).

† The p value for heterogeneity among the individual relative risks that contribute to the summary relative risk for breast cancer.

‡ Results adjusted for type of menopause, whether natural or artificial, or age at menopause, or both.

markedly heterogeneous ( $p = 0.006$ ). Heterogeneity was not reduced if the meta-analysis was confined to the 12 studies with results adjusted for menopausal status, and the relative risk was 1.05. However, heterogeneity was reduced ( $p = 0.07$ ) when the meta-analysis was confined to 12 studies of postmenopausal women, and the relative risk estimate was 0.96. Heterogeneity was further reduced ( $p = 0.04$ ) when the meta-analysis was confined to seven studies of postmenopausal women with adjusted results, and the relative risk estimate was 0.99.

When the highest dose category in each study was meta-analyzed (number of studies unspecified), the relative risk estimate was 1.04, but the data were exceedingly heterogeneous ( $p = 0.00003$ ). One study accounted for most of the heterogeneity; when it was excluded, heterogeneity was reduced ( $p = 0.09$ ), and the relative risk estimate was 1.28 (95 percent CI 1.06–1.54), the only “statistically significant” finding in the entire meta-analysis. Finally, when estrogen use was analyzed among women with a family history of breast cancer (number of studies unspecified), the relative risk estimate was 1.25 (95 percent CI 0.83–1.88), and there was marked heterogeneity ( $p = 0.006$ ).

These data are hardly persuasive evidence of causality, and I cannot think of a better empirical demonstration of the futility of meta-analysis than this example. To this must be added that the author of the meta-analysis differed from the reviewers. He judged that the positive results did not support causality, and were probably due to publication bias. His conclusion was that the data were reassuring.

The meta-analysis of breast cancer risk in relation to estrogen use continues. There are now at least seven published studies, some with contradictory results (18–24). There will undoubtedly be a lot more. In time, they will more fully reveal the absurdity of the approach and help to bring about its demise. Epidemiology can only benefit.

Finally, in closing, it is necessary to note that it is already late in the day. Meta-

analysis is now firmly entrenched in the medical and epidemiologic establishments, sometimes under the label “risk analysis.” Contracts and grants have been and are being awarded, and relative risks of low magnitude are being adjudged as causal. Recently, for example, there was a heated exchange of views (25–27) concerning the carcinogenic hazards of water chlorination: a quality-weighted meta-analysis of 10 studies (28) yielded a summary relative risk estimate for all cancer sites of 1.15 (95 percent CI 1.09–1.20). I propose that the meta-analysis of published nonexperimental data should be abandoned.

---

## REFERENCES

1. Davis B. What price safety? Risk analysis measures need for regulation, but it's no science. *Wall Street Journal*, August 6, 1992, p. 1.
2. Boston Collaborative Drug Surveillance Program. Reserpine and breast cancer. *Lancet* 1974; 2:669–71.
3. Armstrong B, Stevens N, Doll R. Retrospective study of the association between use of rauwolfia derivatives and breast cancer in English women. *Lancet* 1974;2:672–5.
4. Heinonen OP, Shapiro S, Tuominen L, et al. Reserpine use in relation to breast cancer. *Lancet* 1974;2:675–7.
5. World Health Organization. Some pharmaceutical drugs. (IARC monographs on the evaluation of the carcinogenic risk of chemicals to man, vol. 24). Lyon, France: International Agency for Research on Cancer, 1980:211–41. (Distributed by WHO Publications Centre USA, Albany, NY).
6. Shapiro S, Parsells JL, Rosenberg L, et al. Risk of breast cancer in relation to the use of rauwolfia alkaloids. *Eur J Clin Pharmacol* 1984;26: 143–6.
7. Rosenberg L, Slone D, Shapiro S, et al. Breast cancer and alcoholic-beverage consumption. *Lancet* 1982;1:267–70.
8. Rosenberg L, Palmer JR, Miller DR, et al. A case-control study of alcoholic beverage consumption and breast cancer. *Am J Epidemiol* 1990;131:6–14.
9. Longnecker MP, Berlin JA, Orza MJ, et al. A meta-analysis of alcohol consumption in relation to risk of breast cancer. *JAMA* 1988;260:652–6.
10. Harvey EB, Schairer MS, Brinton LA, et al. Alcohol consumption and breast cancer. *J Natl Cancer Inst* 1987;78:657–61.
11. Webster LA, Layde PM, Wingo PA, et al. Alcohol consumption and risk of breast cancer. *Lancet* 1983;2:724–6.
12. Talamini R, LaVecchia C, Decarli A, et al. Social factors, diet and breast cancer in a Northern Italian population. *Br J Cancer* 1984;49:723–9.
13. Longnecker MP. Alcoholic beverage consump-

