Invited Commentary: Benefits of Heterogeneity in Meta-analysis of Data from Epidemiologic Studies

Jesse A. Berlin

In their commentary in this issue of the Journal, Colditz et al. (1) have made a number of interesting points and raised some equally interesting questions related to heterogeneity in meta-analyses of epidemiologic data. In selecting a title for this invited commentary, I have chosen to emphasize explicitly the positive aspects of heterogeneity. I would like to complement (and compliment) the material presented by Colditz et al. by expanding on three major areas: 1) meta-analysis, particularly the exploration of heterogeneity, as original research; 2) the use of meta-analysis in the planning of future research; and 3) some potential pitfalls of which one needs to be aware in the exploration of sources of heterogeneity. The important message of my comments will be that heterogeneity of study results can provide many benefits and should be viewed as a strength of meta-analysis, not a barrier to its use.

META-ANALYSIS AS ORIGINAL RESEARCH

Meta-analysis shares a number of characteristics with so-called primary studies that are traditionally viewed as original research (2). It has the potential to resolve conflicting research results in much the same way as a "definitive" clinical trial or epidemiologic study does. Because it involves comparisons across studies, meta-analysis can lead to insights when study design, exposure assessment or exposure levels, study populations, etc., are found to relate to study outcome. Meta-analysis involves establishing predefined eligibility criteria for data to be included, data collection, and data analytic methods and requires careful interpretation of results with conclusions limited by the available data and study populations. In these regards, it is similar to any other observational study and can provide new information in the same way that a cohort study can. Several examples will serve to illustrate this point. Most of the examples emphasize a source of interstudy variability not directly addressed by Colditz et al. (1), namely variability in study populations with respect to demographic and/or clinical characteristics.

Oral contraceptives and risk of breast cancer: menopausal status as an effect modifier

In a meta-analysis of studies of oral contraceptives and the risk of breast cancer, cited by Colditz et al. (1) in another context, Romieu et al. (3) found an association between increasing odds ratio and increasing duration of oral contraceptive use. This positive association was limited to case-control studies in which the cases were mostly premenopausal, i.e., in which the study or subgroup had an upper age limit of 45 years. Within the premenopausal studies, there was very little remaining heterogeneity, as evidenced both by a non-significant test for heterogeneity and by the fact that fixed- and random-effects models gave nearly identical results. In contrast, studies that included postmenopausal cases showed no evidence of a positive duration-response relation and did show substantial residual heterogeneity.

In a subsequent reanalysis of these data in a methodological paper, Berlin et al. (4) confirmed the above results and found that the magnitude of the odds ratio also depended on the calendar years during which cases were accrued, presumably because of the changing formulations of oral contraceptives.

Calcium intake and bone mass: menopausal status as an effect modifier

In another example related to menopausal status, Cumming (5) reviewed the evidence relating calcium intake to bone mass. He found a positive correlation between calcium intake and bone mass in cross-sectional studies, but this association was greater in studies of premenopausal women. In intervention trials in postmenopausal women, calcium slowed the rate
of bone loss, and this slowing was greatest in studies in which the baseline calcium was low.

**Streptokinase after myocardial infarction: location of infarct as an effect modifier**

In a meta-analysis of randomized trials explicitly addressing subgroup comparisons, Midgette et al. (6) examined the use of streptokinase in patients with suspected acute myocardial infarction. The authors found a summary relative risk of 0.72 (95 percent confidence interval (CI) 0.65–0.79), favoring streptokinase, in patients with suspected acute anterior myocardial infarction and a summary relative risk of 0.87 (95 percent CI 0.76–1.01) in those with suspected inferior infarction. They concluded that there is a protective effect of intravenous streptokinase in anterior infarction, but that the protective effect in inferior infarction is smaller and less certain.

**Corticosteroids and alcoholic hepatitis: study design and patient characteristics as effect modifiers**

Alcoholic hepatitis is a serious acute liver condition with hospital mortality as high as 65 percent (7). Corticosteroids have been used, although not approved, for this indication, and randomized clinical trials have been done to assess the efficacy of those drugs in the prevention of mortality. Imperiale and McCullough (7) performed a meta-analysis to summarize the findings of those studies. The summary relative risk was 0.63 (95 percent CI 0.50–0.80), indicating a 37 percent reduction in the risk of mortality in the corticosteroid-treated patients. The corticosteroid studies exhibited a high degree of heterogeneity, however, so further analyses were done according to subgroup membership. The subgroups, presumably defined in advance of their analysis, were based on the inclusion or exclusion of patients with gastrointestinal bleeding and hepatic encephalopathy. The authors also used a simple measure of study quality to define the most reliable studies. In the best quality studies of patients with hepatic encephalopathy, without gastrointestinal bleeding, and with baseline equivalence of treatment groups, the estimated relative risk was 0.51 (95 percent CI 0.33–0.77). In patients without hepatic encephalopathy, there was no apparent benefit of therapy, with a relative risk of 1.01 (95 percent CI 0.35–2.91), even in the best quality studies. Thus, these authors demonstrated a protective effect of a therapy that had been in use for nearly 20 years. They also identified subgroups for whom treatment appeared to be the most beneficial. Given these results, I invite the reader to consider the ethics of conducting another randomized clinical trial, especially in patients with hepatic encephalopathy and without gastrointestinal bleeding.

**Level of exposure and magnitude of relative risk: cigarettes and stroke**

Information about multiple levels of exposure may also be available either within or across studies and can lead to investigations of dose-response-like relations. For example, evidence of broadly defined dose-response patterns can be provided by separate combinations of relative risks from studies or subgroups with varying levels of exposure. In a meta-analysis of cigarette smoking and the risk of stroke, Shinton and Beevers (8) grouped exposure categories into three categories: low (less than 10 cigarettes/day), intermediate (10–20 cigarettes/day), and high (more than 20 cigarettes/day). The summary relative risks for the three categories were 1.37, 1.45, and 1.82, respectively, consistent with an increasing risk with increasing exposure.

**Level of exposure and magnitude of relative risk: examples from occupational epidemiology**

Occupational epidemiologic studies may not always have ideal measures of exposure, but sometimes may be grouped into broad categories of exposure to examine the relation between exposure level and relative risk. In an analysis of occupational studies of asbestos exposure and gastrointestinal malignancy, Frumkin and Berlin (9) used a surrogate measure of asbestos exposure to group studies into high- or low-exposure categories. The indirect measure of exposure they used was the observed standardized mortality ratio (SMR) for lung cancer in each study, based on the assumption of a monotonic increase in lung cancer SMR and level of asbestos exposure.

Frumkin and Berlin (9) found that the studies involving high-exposure cohorts showed elevated SMRs for gastrointestinal cancer, whereas studies classified as low exposure showed smaller or no increases in mortality. Prior to stratification on these imputed exposure levels, the summary SMRs showed a relatively high level of heterogeneity, which was greatly reduced after stratification. For example, the summary SMR for all 15 studies of colorectal cancer was 111 (95 percent CI 88–141), with highly significant heterogeneity (p < 0.001). The nine studies with low imputed exposure levels had a summary SMR of 86 (95 percent CI 69–106), with substantial (p = 0.06) remaining heterogeneity. The six studies with high imputed exposure levels showed a much stronger association between asbestos exposure and colorectal cancer, with
a summary SMR of 161 (95 percent CI 134–193) and very little evidence ($p = 0.93$) of any substantial residual heterogeneity.

Similarly, a meta-analysis examining petroleum workers and their risk of leukemia (10), followed by an exchange of letters (11, 12), indicated that petroleum workers in certain job categories may be at higher risk of leukemia than others.

These few examples demonstrate that heterogeneity of study results can lead to helpful insights into the problems being addressed. An important general principle is that there were, in all of the cited examples, strong a priori reasons for choosing the particular factors by which the authors stratified their findings. Thus, as Colditz et al. (1) suggest, in meta-analysis, as in primary studies, it is important to invest time and energy in developing a detailed protocol, in which as many contingencies as possible are addressed.

**META-ANALYSIS IN THE DESIGN OF FUTURE STUDIES**

As a consequence of the exploration of sources of heterogeneity, the meta-analyst may identify features of study design that have implications for future research. I shall present several examples of how that process can work. Colditz et al. (1) cited examples of meta-analyses that warrant emphasis in this regard. They pointed out two earlier meta-analyses (13–15) of associations between cancer and alcohol in which the cohort studies showed stronger associations than did the case-control studies. Thus, they argue, an initial stratification of results by study design provides results useful in considering the design of subsequent studies.

Two meta-analyses provide fairly conclusive findings that a particular type of exposure measurement is preferable over others. There are two commonly used methods of assessing type A behavior: the structured interview (16) and the Jenkins Activity Scale (17). Booth-Kewley and Friedman (18), in a quantitative review of psychologic predictors of heart disease, reported that studies using the structured interview found stronger associations between type A behavior and heart disease than those found by studies using the Jenkins Activity Scale. Both assessment tools were presumably measuring the same underlying characteristic. Assuming that the interview studies did not suffer from more serious biases away from the null than did the activity scale studies, the results of the meta-analysis point clearly to the structured interview as the preferred method of assessment to use in future studies.

Needleman and Gatsonis (19) studied the effects of different definitions of exposure on the association between lead exposure and the intelligence quotient (IQ) of children. Studies using blood lead showed a stronger detrimental effect of lead (average partial correlation coefficient, $r = -0.152$) than did studies using tooth lead (average partial correlation coefficient, $r = -0.08$). Both of these summary correlations are fairly small, and there may be other differences among studies than the choice of exposure measure. Nevertheless, the results of stratifying the meta-analysis by type of lead measurement could help inform the choice of exposure measure used in future studies. An important point in this example is that biologic considerations (e.g., the use of a presumably more transient measure, blood lead, versus a measure that presumably reflects more chronic exposure, i.e., tooth lead) would also play an important role in determining the choice of measures to be used.

Another consequence of the exploration of heterogeneity may be to suggest or confirm aspects of study protocol that may be important to finding associations. For example, Gray et al. (20) analyzed studies of the relation between age and testosterone levels in men. They found that studies taking blood samples in the morning tended to show a stronger relation between testosterone and age than did studies using only afternoon samples. This result is perfectly consistent with what endocrinologists already knew (but did not always apply in the design of studies): There is diurnal variation in testosterone levels, with peak levels in the early morning, and diurnal variation is more pronounced in younger than in older men.

**POTENTIAL LIMITATIONS TO THE INVESTIGATION OF SOURCES OF HETEROGENEITY**

Occasionally, a meta-analyst is fortunate enough to be working with individual-level data. In most situations, relative to published data, individual-level data would provide the analyst with great flexibility to address issues of control of confounding and exploration of effect modification or subgroup effects. More often, however, published reports of research are the only source of data available to the meta-analyst. One consequence of analyzing data at the level of the study, rather than at the level of the individual, is that study-level characteristics are assumed to apply to all of the members of a study. This principle is similar to the assignment of group-level values for exposure in ecologic studies or the assignment of exposure levels based on broad job categories in occupational epidemiologic studies. This problem can make interpretation of the individual-level question somewhat difficult when investigating sources of heterogeneity.
For example, consider the analyses of published studies of oral contraceptives and risk of breast cancer already described (3, 4). Separate analyses could be performed for case-control studies with an upper age limit of 45 years or less and for those studies that included postmenopausal women. These “postmenopausal” studies were not, however, limited to postmenopausal women, but also included premenopausal women. The results for the premenopausal women from these postmenopausal studies were unavailable for inclusion in the summary of the studies of premenopausal women. Ideally, we would like to perform completely separate analyses of premenopausal and postmenopausal women and/or fit regression models with appropriate interaction terms to address the presumed underlying source of heterogeneity. Without access to individual-level data, and especially without information on the actual menopausal status of individual women, compromise is necessary, and only an indirect answer to the question of menopausal status can be obtained.

A second concern in exploring sources of heterogeneity is that other differences may exist among studies apart from the study characteristic we are examining. For example, in the meta-analysis of studies of type A behavior discussed above (18), the authors showed that the studies using interviews found stronger associations between type A behavior and heart disease than did the studies using the activity scale. Suppose that case-control studies using interviews, perhaps even because of the interview process, were far more subject to recall bias than the studies using the self-reported activity scales. In that case, the stronger association in the interview studies would not necessarily argue in favor of using interviews in subsequent studies. The point is that comparisons between types of studies are just as subject to potential bias by unmeasured confounders as are comparisons between exposed and unexposed individuals in any observational study.

The idea that comparisons may be confounded has led several authors to advocate the use of regression models to explore heterogeneity (2, 3, 21–24). Regression models could be used to examine the effects of a given study-level variable while controlling for the effects of other variables. Unfortunately, study characteristics are sometimes so highly correlated with each other that it can be difficult to separate the effects of one from another. For example, Antman and Berlin (25) examined data showing the decline in the risk of ventricular fibrillation over the previous 20 years using data from all of the randomized trials of lidocaine prophylaxis in the prevention of primary ventricular fibrillation after acute myocardial infarction. Their primary interest was in examining the relation between year of publication and risk of ventricular fibrillation. An interesting association described in a subsequent paper (23) was that route of administration of lidocaine (intramuscular vs. intravenous) was a strong predictor of the risk of ventricular fibrillation in the treated groups from the randomized trials. Curiously, route of administration was also a strong predictor of risk in the control groups, where, presumably, route of administration of placebo (or of doing nothing) should not be related to risk. This and other associations between pairs of predictor variables made interpretation of regression models of risk quite challenging (23).

Thus, what I believe meta-analysts are beginning to learn and need to have reinforced is that working with study-level data presents some challenges that are not faced when analyzing individual-level data. At the same time, exploration of heterogeneity also faces the same challenges as the analysis of any observational epidemiologic data, in that issues of confounding, interaction, and model-building strategies require strict attention.

CONCLUSIONS

Colditz et al. (1) have presented a wide-ranging discussion of many aspects and implications of heterogeneity in meta-analyses of epidemiologic studies. In this paper, I have supplemented their discussion with a few related observations. In general, the two papers point to several main conclusions:

1) Heterogeneity is common in meta-analyses of epidemiologic data and probably should be viewed as the expectation, rather than the exception. Despite that fact, many authors fail to assess, report, and explore the sources of heterogeneity;
2) Analyses of heterogeneity should be undertaken, but should be pursued and interpreted cautiously in the spirit of an exploratory data analysis;
3) The exploration of heterogeneity can lead to insights about modification of apparent associations by various aspects of study design, exposure measurements, and study populations;
4) Relations discovered in the process of exploring heterogeneity may be useful in the planning and execution of subsequent studies;
5) The exclusion of outlying results solely on the basis of their disagreement with other studies can lead to seriously biased summary estimates and should be avoided; and
6) Sources of heterogeneity to be explored are most easily interpreted when they are identified in advance of the analysis, not when they are suggested only by the data.

Am J Epidemiol Vol. 142, No. 4, 1995
Meta-analysis is like any other form of data analysis in that it requires strict adherence to methodological guidelines, careful planning, the use of a priori definitions and analytic strategies, and extremely careful interpretation that does not go beyond the limits of the data.

REFERENCES