Random-Effects Meta-Analyses Are Not Always Conservative

Charles Poole¹ and Sander Greenland²

It is widely held that random-effects summary effect estimates are more conservative than fixed-effects summaries in epidemiologic meta-analysis. This view is based on the fact that random-effects summaries have higher estimated variances and, consequently, wider confidence intervals than fixed-effects summaries when there is evidence of appreciable heterogeneity among the results from the individual studies. In such instances, however, the random-effects point estimates are not invariably closer to the null value nor are their p values invariably larger than those of fixed-effects summaries. Thus, random-effects summaries are not predictably conservative according to either of these two connotations of the term. The authors give an example from a meta-analysis of water chlorination and cancer in which the random-effects summaries are less conservative in both of these alternative senses and possibly more biased than the fixed-effects summaries. The discussion of when to use random effects and when to use fixed effects in computing summary estimates should be replaced by a discussion of whether summary estimates should be computed at all when the studies are not methodologically comparable, when their results are discernibly heterogeneous, or when there is evidence of publication bias.

Meta-analysis remains a controversial topic in epidemiology, with respect not only to how but also to whether and when it should be done (1–15). One point of contention is when to use random effects and when to use fixed effects in computing summary estimates of effect. Some favor the use of random effects and question the basis for using fixed effects (16, 17). Others find the inclusion of random effects "peculiar" (2, p. 130) or oppose it so strongly as to consider it "wholly wrong" (18, p. 242).

We view the opposition of random-effects summaries and fixed-effects summaries as misleading and counterproductive, for the following reason: If the two summaries differ to a meaningful extent, there must be meaningful discrepancies (heterogeneity) among the study-specific effect estimates. In this situation, we contend that any summary will be inadequate. Instead of summarizing, the meta-analyst should report the discrepancies and seek explanations for them (1–11, 19). That such a search may fail to find a convincing explanation does not make a summary estimate any more adequate or any less misleading.

Received for publication June 1, 1998, and accepted for publication January 22, 1999.

¹ Department of Epidemiology, University of North Carolina School of Public Health, Chapel Hill, NC.
² Department of Epidemiology, UCLA School of Public Health, Los Angeles, CA.

Reprint requests to Dr. Charles Poole, Department of Epidemiology (CB 7400), University of North Carolina School of Public Health, Chapel Hill, NC 27599-7400.

Despite these misgivings about meta-analytical summaries of discrepant results, it appears that most meta-analysts are committed to producing a summary of some kind. Meta-analysts rarely entertain the possibility of concluding that study-specific results are too heterogeneous to aggregate. Even rarer are analysts who conclude as Ness and Powles (20, p. 2) did: "The studies included differed in: the type of study, the measurement and reporting of exposure, the period of follow-up and outcome selected. For these reasons, no attempt was made to arrive at a summary statistic."

Given a commitment to aggregate results from different studies, no matter how discrepant their methods and results might be, one rationale given for preferring random-effects summaries is that they are said to be "conservative" (21, pp. 141, 144; 22, p. 973; 23, p. 95). "Conservative" in this context is generally taken to mean that random-effects summaries tend to have higher estimated variances and, therefore, wider confidence intervals than fixed-effects summaries computed from the same study-specific results. We show that, when random-effects summaries are conservative in this sense, they need not be conservative in other, equally important, senses. Specifically, random-effects summaries can be farther from the null value and can have smaller p values, so they can appear more strongly supportive of causation or prevention than fixed-effects summaries. Furthermore, it is possible for a misleading appearance of strongly supportive results to be produced by a plausible form of publication bias.
to which random-effects summaries are more vulnerable than fixed-effects summaries.

METHODS

Suppose we desire a summary of $K$ effect estimates, indexed by $i$, with no consideration for the influence that study characteristics (populations, study designs, exposure contrasts, analytical methods, and so on) might have had on these estimates. In each study population, let $\theta_i$ be the true magnitude of a measure of effect. $\theta_i$ is typically the natural logarithm of a relative risk (i.e., a ratio of incidence rates or of incidence proportions).

A meta-analytic summary estimate is computed as a weighted average, $\bar{\theta} = \frac{1}{n} \sum_{i=1}^{n} \theta_i$, using weights inversely proportional to $v_i + s_i^2$, where $v_i$ is the estimated variance of $\theta_i$ and $s_i^2$ is an estimate of the variance ($\sigma_i^2$) of a population of effect parameters from which it is assumed that the $\theta_i$ were randomly drawn. The variance of the summary is estimated as follows:

$$V = \frac{1}{\sum_{i=1}^{K} (v_i + s_i^2)^{-1}}$$

A fixed-effects summary (24, 25) is computed under the assumption that the effect measure has the same value in all the study populations ($\theta_1 = \theta_2 = \ldots = \theta_K = \theta$) and, therefore, that the distribution is a constant with $\sigma^2 = 0$. Thus, $s^2$ is set to zero. A random-effects summary (25, 26) is computed under the assumption that the population effect measures have a random (e.g., normal) distribution and, therefore, that $\sigma^2 > 0$. In this case, $s^2$ is estimated from the study-specific results.

The parameter $\sigma^2$ is often called the “between-study component of variance.” It is usually estimated as $s^2 = \frac{\max(Q - df, 0)}{c}$, where $Q$ is a standard homogeneity test statistic, $df = K - 1$ are its degrees of freedom, and $c$ is a function of the $v_i$ (25). Because $s^2$ is constrained to be no smaller than zero, the random-effects estimate of $V$ must be at least as large as the fixed-effects estimate, and the confidence interval for the random-effects summary must be at least as wide as the confidence interval for the fixed-effects summary. The random-effects interval will be even wider if a correction is made for the fact that $\sigma^2$ is estimated (as $s^2$) (27), though in some examples (including the one below) the effect of the correction is negligible.

A rare modification of this approach is to use a formula for $s^2$ that has only $Q - df$ in its numerator (23), thus permitting negative values for $s^2$ (when $Q < df$) and narrower random-effects confidence intervals that are narrower than their fixed-effects counterparts. Of interest here, however, is the increasingly popular approach in which random-effects summaries are computed, and interpretations based on them, when evidence against the homogeneity assumption is present ($Q > df$) or even strongly present ($Q >> df$). The example below illustrates this approach.

EXAMPLE

Morris et al. (28) conducted a meta-analysis of epidemiologic studies on the relation of water chlorination to cancer. The authors based their interpretations on random-effects summaries. They did not assess heterogeneity or compare the random-effects summaries with fixed-effects summaries. The authors described an increase in the random-effects variance estimate as “conservative” because it “minimized the probability of a type I error in the meta-analyses” (28, p. 956). In a previous publication, one of the authors had noted that “[the] combining of heterogeneous material is a commonly accepted threat to the validity of meta-analysis” and had described random-effects summaries as “conservative” when heterogeneity is evident (22, p. 973).

Table 1 shows random-effects summaries, fixed-effects summaries, and homogeneity $p$ values computed by us from the individual study results extracted by Morris et al. (28, table 3) from the original publications. Noting that all of the random-effects summaries are greater than 1.00, Morris et al. suggested that “the association of chlorination by-products with cancer, although modest, is more general than the two statistically significant sites” (bladder cancer and rectal cancer) (28, pp. 957, 959). After cautioning, “Precise cause and effect cannot be determined,” the authors performed a computation suggesting that thousands of “attributable cases” of these two cancers occur in the United States each year, “[if] we posit that the association demonstrated between chlorination by-products and cancer represents a causal relationship in some way” (28, p. 962). They concluded, “The present study has identified a clear and significant association between neoplastic disease and the consumption of water containing chlorination by-products” (28, p. 962).

If more conservative results are results that would encourage less definitive conclusions, it is clear that random-effects summaries are less conservative than fixed-effects summaries in this example. The fixed-effects summary estimates of relative risk are almost equally divided between those above and below 1.00 (table 1). None is as high as 1.20. For three cancer sites (brain, rectum, and stomach), the $p$ value for the fixed-
Random-Effects Meta-Analyses Not Always Conservative

TABLE 1. Summary of relative risk estimates, 95% confidence intervals, two-sided \( p \) values, and two-sided homogeneity \( p \) values derived from study-specific results on water chlorination and cancer presented by R. D. Morris et al.*

<table>
<thead>
<tr>
<th>Cancer site</th>
<th>No. of results</th>
<th>Random-effects summaries†</th>
<th>Fixed-effects summaries</th>
<th>Homogeneity ( p ) value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Relative risk</td>
<td>Two-sided ( p ) value</td>
<td>Relative risk</td>
</tr>
<tr>
<td>Bladder</td>
<td>7</td>
<td>1.20 (1.08, 1.33)‡</td>
<td>0.001</td>
<td>1.19 (1.10, 1.30)</td>
</tr>
<tr>
<td>Brain</td>
<td>2</td>
<td>1.28 (0.52, 3.14)</td>
<td>0.6</td>
<td>0.96 (0.76, 1.20)</td>
</tr>
<tr>
<td>Breast</td>
<td>4</td>
<td>1.17 (0.90, 1.53)</td>
<td>0.2</td>
<td>0.92 (0.88, 0.95)</td>
</tr>
<tr>
<td>Colon</td>
<td>7</td>
<td>1.11 (0.91, 1.35)</td>
<td>0.3</td>
<td>0.92 (0.90, 0.95)</td>
</tr>
<tr>
<td>Colon and rectum</td>
<td>8</td>
<td>1.15 (0.97, 1.36)</td>
<td>0.1</td>
<td>1.05 (0.99, 1.11)</td>
</tr>
<tr>
<td>Esophagus</td>
<td>5</td>
<td>1.16 (0.89, 1.51)</td>
<td>0.3</td>
<td>1.13 (0.91, 1.40)</td>
</tr>
<tr>
<td>Kidney</td>
<td>4</td>
<td>1.16 (0.89, 1.51)</td>
<td>0.3</td>
<td>1.05 (0.95, 1.15)</td>
</tr>
<tr>
<td>Liver</td>
<td>4</td>
<td>1.15 (0.95, 1.40)</td>
<td>0.2</td>
<td>1.15 (0.96, 1.39)</td>
</tr>
<tr>
<td>Lung</td>
<td>5</td>
<td>1.01 (0.86, 1.19)</td>
<td>0.9</td>
<td>0.95 (0.92, 0.97)</td>
</tr>
<tr>
<td>Pancreas</td>
<td>6</td>
<td>1.05 (0.91, 1.22)</td>
<td>0.5</td>
<td>0.96 (0.92, 1.01)</td>
</tr>
<tr>
<td>Rectum</td>
<td>6</td>
<td>1.38 (1.01, 1.67)</td>
<td>0.04</td>
<td>1.06 (0.99, 1.14)</td>
</tr>
<tr>
<td>Stomach</td>
<td>6</td>
<td>1.14 (0.94, 1.37)</td>
<td>0.2</td>
<td>1.01 (0.95, 1.06)</td>
</tr>
</tbody>
</table>

† Random-effects summaries, confidence intervals, and \( p \) values computed from the study-specific results presented by Morris et al. (table 3). These values differ slightly in some cases from those reported by Morris et al. (table 4).
‡ Numbers in parentheses, 95% confidence interval.

As a specific example, consider cancer of the rectum, one of the sites for which the random-effects summary "achieved statistical significance" (29) and thus became the subject of a computation of potentially attributable cases. The random-effects summary for this cancer site, like the fixed-effects summary, is dubious because the individual results of which it is composed strongly conflict with one another. The conflict, reflected in the very low homogeneity \( p \) value (table 1), is seen in greater detail by the large gaps among the study-specific confidence intervals, some of which do not even overlap (table 2). A single summary, whether computed with fixed effects or random effects, masks these important facts and provides an unsuitable basis for computing numbers of potentially attributable cases.

Following guidance we consider sound (5, 11, 36–38), we should seek explanations for the conflict among the study-specific results. One a priori hypothesis, posed as far back as the earliest study in this analysis (30), stems from a recognition that most chlorinated water tends to be surface water (e.g., from rivers) and most unchlorinated water tends to be ground water (e.g., from wells). Thus, the distinction between surface water and ground water is sometimes used as a proxy for the presence and absence of chlorination (table 2). Surface water and ground water differ in many other ways, however. For example, surface water is often more highly polluted with a wide range of synthetic chemicals other than by-products of chlorination, and ground water tends to have higher con-
table 2. Exposure contrasts and results of studies of rectal cancer and chlorinated drinking water derived from results extracted by R. D. Morris et al. for a meta-analysis of water chlorination and cancer

<table>
<thead>
<tr>
<th>Author(s) and year</th>
<th>Reference no.</th>
<th>Index water</th>
<th>Reference water</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atawanga et al., 1978</td>
<td>30</td>
<td>Chlorinated</td>
<td>Uncalorinated ground</td>
</tr>
<tr>
<td>Brenniman et al., 1980</td>
<td>31</td>
<td>Chlorinated ground</td>
<td>Unchlorinated ground</td>
</tr>
<tr>
<td>Wilkins and Comstock, 1981</td>
<td>32</td>
<td>Chlorinated surface</td>
<td>Unchlorinated surface</td>
</tr>
<tr>
<td>Young et al., 1981</td>
<td>33</td>
<td>Chlorinated</td>
<td>Unchlorinated ground</td>
</tr>
<tr>
<td>Gettablet et al., 1982</td>
<td>34</td>
<td>Chlorinated surface</td>
<td>Unchlorinated ground</td>
</tr>
<tr>
<td>Zeller et al., 1985</td>
<td>35</td>
<td>Chlorinated surface</td>
<td>Unchlorinated ground</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Study name</th>
<th>Relative risk (95% CI)</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brenniman et al. (31)</td>
<td>1.98 (1.95, 1.98)</td>
<td>0.055</td>
</tr>
<tr>
<td>Zierler et al. (35)</td>
<td>2.2 (2.0, 2.5)</td>
<td>0.139</td>
</tr>
<tr>
<td>Other studies</td>
<td>1.6 (1.2, 2.2)</td>
<td>0.039</td>
</tr>
</tbody>
</table>

Informal inspection of the study-specific results (table 2) suggests that the two studies in which the water source (surface or ground) was held constant by design found weaker associations than the studies in which the comparisons were essentially between chlorinated surface water and unchlorinated ground water. More formally, a weighted random-effects meta-regression suggests that the four studies in which water source was not held constant were estimating a relative risk that was about 60 percent higher than the relative risk being estimated by the studies in which the water source was held constant (estimated ratio of relative risks = 1.6; 95 percent confidence interval: 1.1, 2.2). In addition to the intrinsic differences between surface water and ground water previously described, there may be a greater difference of concentrations of chlorination by-products between chlorinated surface water and unchlorinated ground water than between chlorinated and unchlorinated ground water, or between chlorinated and chloraminated surface water.

Another possible explanation for discrepant results is publication bias. The reluctance of investigators to publish results close to the null value and their extreme reluctance to publish implausible results are well documented (39-42). As one student of the problem concluded, “In summary, publication bias is a serious problem for the meta-analyst” (43, p. 408). Some authorities explicitly advocate withholding implausible results from the published record (44). It has been noted in some situations (2, 5, 45) that the reluctance to publish near-null and implausible results may be greater if those results are imprecise (i.e., if the relative risk estimates have wide confidence intervals). In this case, a random-effects summary would be biased to a greater degree than a fixed-effects summary because the random-effects summary gives greater relative weight to imprecise results, especially when heterogeneity is evident. The stronger the evidence of het-

centrations of dissolved minerals. Without taking special precautions in the design of a study, it would be difficult or impossible to distinguish a chlorination effect from a surface-water effect.

Two of the six studies supplying rectal cancer results for this meta-analysis did take special steps to confront this potential problem. Brenniman et al. (31) confined their study to chlorinated and unchlorinated ground water. Zierler et al. (35) confined their study to surface water treated by chlorination and by chloramination, a disinfection process in which ammonia is added after the chlorination step, in part to reduce the concentration of chlorination by-products. In the four remaining studies, the exposure contrasts were essentially between chlorinated surface water and unchlorinated ground water (table 2).
heterogeneity (i.e., the greater the difference between $Q$ and $df$), the greater the value of $\hat{\omega}$, and the more closely the random-effects summary, $\hat{\theta}_r$, will approximate a simple arithmetic mean of the study-specific estimates, with no extra weighting given to the more precise results.

In the present example, where implausible results would be relative risk estimates less than 1.00, publication bias is a reasonable hypothesis. To determine whether study subjects live in communities using surface water or ground water is a simple matter, and information on the use of chlorination is only slightly more difficult to obtain. Some investigators may have obtained such information, conducted a preliminary analysis, and then decided, based on the results, whether or not to prepare a full report for publication. A result close to the null or, especially, a relative risk estimate below 1.0 may have made the requisite effort seem unattractive.

Unfortunately, there are too few results in the present example (eight at most, table 1) to expect strong evidence of publication bias for any single cancer site, even if such bias were present. A test for publication bias developed by Begg and Mazumdar has “only moderate” power when the meta-analysis consists of approximately 25 studies and does not become “fairly powerful” until there are about 75 studies (46, p. 1088). Begg and Mazumdar developed a stratified version of the test to find evidence of publication bias across results for several diseases in a single meta-analysis and applied it to the study-specific results extracted by Morris et al. (28). Begg and Mazumdar concluded that there was “strong evidence of publication bias” (46, p. 1097).

A similar impression is lent by the shape of a funnel graph (3, 4), in which the effect estimate from each study, $\hat{\theta}_i$, is plotted against a measure of the estimate’s precision, $1 / n_i$. In this example (figure 1), the upward “spout” of the funnel, composed of the more precise results, lies very close to the null value. The “cone” of the funnel, consisting of the less precise results at the bottom of the graph, is shifted toward estimates on the plausible (i.e., high) side of the null value. The image is what would be expected if a sizable number of imprecise and implausible results, though extant, were missing from the published record.

The result responsible for more of the heterogeneity than any other is from the study by Zierler et al. (35). It may be tempting to view this result as an “outlier” and to look for reasons to exclude it. Such an approach would be inappropriate because the basic structure of the study by Zierler et al. was no different from most of the others (30, 31, 33, 34): a case-control study of cancer deaths with controls consisting of deaths from other causes and exposure classification based on residence information from death certificates. (The study by Wilkins and Comstock (32) was a cohort study.) Moreover, the four results other than the result of Zierler et al. are still heterogeneous (two-sided $p = 0.05$).

Even if the result of Zierler et al. were to be improperly excluded, we would consider it inappropriate to combine the remaining five, quantitatively heterogeneous, results into a single quantitative summary, with or without random effects. Public-health interest in this analysis arises not from the qualitative homogeneity of the study-specific results, but from the use of the quantitative, random-effects summary to suggest that thousands of cancer cases are potentially attributable to ingestion of chlorination by-products in the United States each year. (The figure for rectal cancer was originally computed as 6,500 (28), then corrected to 7,500 (28), and then rounded to 8,000 (29).) The quantitative heterogeneity among the study-specific results, their strong association with a key feature of study design, and the evidence of publication bias should call into question any speculative computations of potentially attributable cases.
We do not wish to be misconstrued as claiming that water chlorination has no carcinogenic effect. Causality is a tenable explanation for the patterns of results seen in this literature for rectal cancer, bladder cancer, and perhaps other cancers as well. Publication bias, confounding by characteristics other than chlorination that distinguish surface water from ground water, and other methodological features of the studies are also tenable. We suggest that the estimation of thousands of potentially attributable cancer cases from summary estimates computed from disparate studies without mention of heterogeneity or publication bias (28) sets too low a standard for the use of meta-analysis to guide public-health decision-making.

DISCUSSION

In some situations, a random-effects summary effect estimate can be less conservative than a fixed-effects summary by lying farther from the null or by having a lower null $p$ value. Indeed, though Berlin et al. described the random-effects approach as "a more conservative testing procedure" (21, p. 144), they found that it gave results similar to ours: a higher $z$ statistic, and thus a lower $p$ value, in eight of the 22 meta-analyses in which the authors compared random-effects and fixed-effects approaches. Random-effects summaries also may be more biased in some circumstances.

In either situation, analysis of heterogeneity, especially analysis of study characteristics that might explain differences among the results, provides far more important information for interpretation and decision-making than can be provided by any single summary. Thus, we do not assert that fixed-effects summaries are preferable to random-effects summaries. We see either type of summary as having only a minor role in a well-done meta-analysis, unless all the studies are very similar in their methods, in their populations at risk, in their exposure contrasts, and in their results.

Unfortunately, random-effects summarization seems to have become one of the statistical methods, like significance testing, that tend to be applied to epidemiologic data ritualistically and without much thought for important features of the data they might conceal. This unfortunate situation is no doubt due to the false hope that, because the random-effects approach "incorporates the heterogeneity" (47, p. 177) or "take[s] heterogeneity into account" (22, p. 976) in computing the variance estimate and confidence interval for the summary, it somehow accounts for that heterogeneity, explains it, or makes it go away.

Random-effects summaries incorporate heterogeneity under assumptions that, if properly understood, would be immediately rejected in many situations, including our example. One assumption underlying the random-effects summary for rectal cancer, for example, is that the true study-specific associations of chlorination and rectal cancer are "exchangeable." The assumption of exchangeability means, among other things, that one would have no reason to expect the true value of the relative risk in any particular study population to differ systematically from the value in any of the others. In short, exchangeability means that the characteristics of the studies that generated the results make those results potentially combinable. In our view, unless the study-specific results are a priori exchangeable, no attempt should be made to assess their overall heterogeneity or to combine them into a summary.

The assumptions needed for a fixed-effects summary are even less tenable, for they include the assumption that the true relative risks in the different study populations are not merely exchangeable, but identical. The greater robustness of the fixed-effects summary against some forms of publication bias is somewhat accidental and provides insufficient compensation for its deceptively narrow confidence intervals.

Fortunately, the analytical options for conducting an informative meta-analysis are far broader than the choice between random-effects summarization and fixed-effects summarization. When exchangeability is not present, one may profitably adopt a stratification or regression approach to meta-analysis, analogous to common practice in ordinary within-study analysis (36–38, 48–51). In such a comparative, explanatory, or analytical approach, exchangeability assumptions need be made only within levels of covariates that distinguish studies.

The hypotheses that lead one to choose particular study characteristics as stratification or regressor variables often exist long before the meta-analyses are planned, or even before the original studies are conducted, like the longstanding concern about the potential for confounding by surface-water effects in studies of chlorinated water. Nevertheless, to deal with "data-dredging" issues that can arise occasionally in the exploration of heterogeneity (11), random-coefficient (hierarchical regression) models may be used (15, 37, 52, 53). When only small numbers of studies have been published, as in the case of chlorinated water and specific cancers, the ability to use even the simplest method of comparative meta-analysis—stratification—is highly limited. Small numbers of studies, however, should not be considered a license to aggregate, just as random effects should not be considered a talisman that averts the conflict of heterogeneity. For readers who remain distrustful of any statistical methods of meta-analysis, aggregative or explanatory (6, 9, 14), the safest alternative may be narrative description and informal comparisons of individual study results without quantitative summarization.
ACKNOWLEDGMENTS

This work was partially supported by US Environmental Protection Agency contract 6C0126NASX to Dr. Charles Poole.

The authors thank Jesse A. Berlin, Matthew P. Longnecker, and Patricia A. Murphy for valuable comments.

REFERENCES