Empirical-Bayes Adjustments for Multiple Comparisons Are Sometimes Useful
Author(s): Sander Greenland and James M. Robins
Source: *Epidemiology*, Vol. 2, No. 4 (Jul., 1991), pp. 244-251
Published by: Lippincott Williams & Wilkins
Stable URL: http://www.jstor.org/stable/20065674
Accessed: 14/05/2013 17:58

Empirical-Bayes Adjustments for Multiple Comparisons Are Sometimes Useful

Sander Greenland[1] and James M. Robins[2]

Rothman (Epidemiology 1990;1:43–46) recommends against adjustments for multiple comparisons. Implicit in his recommendation, however, is an assumption that the sole objective of the data analysis is to report and scientifically interpret the data. We concur with his recommendation when this assumption is correct and one is willing to abandon frequentist interpretations of the summary statistics. Nevertheless, there are situations in which an additional or even primary goal of analysis is to reach a set of decisions based on the data. In such situations, Bayes and empirical-Bayes adjustments can provide a better basis for the decisions than conventional procedures. (Epidemiology 1991;2:244–251)

Keywords: Bayesian statistics; epidemiologic methods; risk.

Rothman[1] recently made a general recommendation against adjustments for multiple comparisons. He assumed that the investigator contemplating such adjustments had purely scientific objectives: to summarize the data in a relevant and informative manner, to summarize comparisons of data with hypotheses, and to interpret any data patterns in light of background theory and knowledge. Rothman characterized the latter objective as an informal, narrative, and even creative attempt to separate promising leads for further investigation from other, less promising avenues.

To the extent that objectives conform to those just described, we concur with his recommendation. Nevertheless, an investigator may have objectives beyond purely scientific ones. For example, the investigator may have an administrative objective of providing a rational basis for resource allocation. To meet this objective, certain methods, which may be viewed as multiple-comparisons procedures, may dramatically outperform ordinary single-comparison procedures.

For reasons listed by Rothman,[1] reasonable multiple-comparisons procedures rarely correspond to conventional multiple-comparisons procedures such as Scheffé or Bonferroni adjustments. As Rothman points out, conventional adjustments are generally unacceptable because they (1) assume a "universal null hypothesis" and (2) sacrifice power for the sake of keeping the study's type I error probability at or below a certain fixed significance level (α-level). These two properties of conventional procedures are indeed undesirable in most settings, although exceptions occur. Fortunately, multiple-comparisons procedures need not have these properties. Among those that need not are *empirical-Bayes* methods,[2–7] which may be viewed as generalizations of Stein estimators.[2–4] We will argue that these procedures can aid in using epidemiologic data to make decisions when background information is limited.

Scientific versus Decision Analyses

We first need to describe, from a statistical point of view, what sets scientific analyses apart from administrative analyses. A scientific analysis seeks to describe the data and the degree of conformity or conflict of the data with various statistical models (hypotheses), always within a broader context of other, usually untested assumptions. For example, a logistic-regression analysis of a cohort would provide point estimates of odds ratios (which would serve as summary relative risks when the disease was rare) and provide tests that these ratios are one. All the estimates and tests could be obtained from the likelihood function for the full logistic model. If we thought the binomial-logistic model was an adequate approximation for the analysis, we could view this likelihood function as a sufficient data summary.[8] If the sample size was large enough, we could use the maximum-likelihood estimates, their estimated standard errors, and their correlations as a summary of this likelihood function.

The likelihood function is not sufficient to reach a decision based on the data: Any decision should take account of other available information on the effects under study, as well as the costs and benefits associated

From the [1]Department of Epidemiology, UCLA School of Public Health, Los Angeles, CA 90024-1772 and [2]Occupational Health Program and Department of Biostatistics, Harvard School of Public Health, Boston, MA.

TABLE 1. Analysis of Toxoplasmosis Prevalence in El Salvador (Computed from Table 3 of Reference 3)

City	No. Cases	No. in Sample	Observed SMR	Standard Error (SE)	Z Score*	Empirical-Bayes SMR
1	12	36	129	30	0.96	104
2	221	278	121	4	5.49	119
3	101	119	119	5	3.94	116
4	13	16	115	12	1.32	108
5	23	26	114	8	1.72	109
6	113	166	113	6	2.10	110
7	18	22	111	6	1.85	109
8	53	78	110	9	1.13	106
9	123	154	109	5	1.90	108
10	243	354	108	4	1.93	107
11	62	79	106	7	0.89	105
12	189	254	105	5	1.08	104
13	98	125	104	6	0.63	103
14	277	393	103	4	0.68	102
15	183	245	102	5	0.50	102
16	384	544	102	4	0.62	102
17	206	305	101	4	0.33	101
18	42	58	100	9	0.05	100
19	23	27	98	13	−0.13	99
20	33	45	97	9	−0.31	98
21	67	92	97	7	−0.47	98
22	147	196	96	5	−0.82	97
23	129	192	94	6	−0.95	96
24	57	89	92	7	−1.19	94
25	85	124	90	7	−1.44	93
26	199	323	90	5	−2.04	92
27	93	160	89	6	−1.90	91
28	77	126	86	6	−2.19	89
29	51	84	84	8	−2.03	89
30	57	91	83	7	−2.32	88
31	30	69	76	11	−2.27	87
32	10	35	71	18	−1.64	92
33	60	120	70	6	−4.63	77
34	14	65	68	15	−2.13	89
35	11	57	60	16	−2.51	87
36	2	17	33	22	−3.08	86

*Computed as (SMR − 100)/SE from three-digit values of SMR and SE in Reference 3. Cities 25, 32, 35, and 36 are east of the Rio Lempa.

with each possible decision. A decision analysis requires, in addition to the likelihood function, a loss function, which indicates the cost of each action under the various possible values for the unknown parameter (benefits would be expressed as negative costs). Construction of a loss function requires one to quantify costs in terms of dollars, lives lost, or some other common scale.

There are two major branches of current statistics: frequentist and Bayesian. Oakes[9] provides an excellent nontechnical comparison of these and other schools of thought. The literature on statistical decision analysis is vast, and we cannot even begin to review it here; Raiffa and Schlaiffer[10] and Wald[11] have written classic Bayesian and frequentist texts. Instead, we will provide an illustration of a statistical decision problem.

An Administrative Example

The following example is, as far as we know, the earliest and simplest example of an empirical-Bayes analysis of epidemiologic data (although several others have since appeared, for example, Reference 7). The data in Table 1 concern toxoplasmosis-antibody prevalence in 36 cities in El Salvador, as estimated from a serosurvey.[11] For technical details of the analysis, see Efron and Morris[3]; for a nontechnical overview of the methods, see their subsequent *Scientific American* article,[4] in which they discuss many of the points made here.

A total of 3,506 subjects out of 5,174 was positive, for a crude prevalence of 68%. Efron and Morris[3] gave city-specific results in terms of $(O_i - E_i)/E_i$, where O_i is the observed and E_i is the expected number of cases in city *i*.

Our presentation differs only in that we display the SMRs ($100 \times O_i/E_i$) instead. The expected values are based on the age distribution of the total sample, so that the average value of the SMR_i weighted by E_i is 100. (Sex was not included in the adjustment, presumably because the sex ratio does not vary across cities and therefore cannot account for any of the intercity variations in prevalence.)

The SMR data in Table 1 could serve many different objectives. Suppose that we know little about the determinants of toxoplasmosis prevalence within El Salvador, but we have limited funds for intervention to prevent further cases in each town. Suppose also that we want the per capita allocation for each city to be proportional to its age-adjusted prevalence, so that areas with high prevalence (for reasons other than age) receive more funds per capita than areas with low prevalence. How should we determine the relative per capita allocation?

One naive answer would be to make the per capita allocations proportional to the observed SMRs. The major drawback of this answer is that it takes no account of sampling error. Sampling error *is* present: The data are from a sample survey, and the targets of interest are the true (citywide) SMRs, not the sample estimates. The effect of sampling error may be very large; for example, the city 1 SMR is within two standard errors of most of the other SMRs, so that the true city 1 SMR may be *less* than most of the other SMRs despite the contrary appearance of the estimates.

The sixth column of Table 1 presents the Z scores, $Z_i = (\text{SMR}_i - 100)/\text{SE}_i$, where SE_i is the estimated standard error of SMR_i. One naive way of accounting for sampling error would be to make per capita allocations depend on Z scores. While, for some cities, this might be an improvement over allocations proportional to the SMRs, it would still not properly account for sampling error. Z scores are too heavily dependent on sample size: A city showing only a small departure from its expectation can have a large Z score simply by having a large sample size (see, for example, cities 10 and 26).

Figure 1 shows a histogram of the distribution of observed SMRs. The effect of sampling error on this distribution depends on the underlying distribution of true SMRs for the cities. How we account for the error should depend on what we are willing to assume about the distribution of true SMRs. We will assume here that the histogram of the true distribution is similar in shape to the observed distribution. More specifically, for the moment, we will make two assumptions:

1. Geographic determinants of age-specific toxoplasmosis-antibody prevalence are randomly distributed across cities, so that the values of the true SMRs are randomly distributed across cities. (This means, for example, that the true SMRs of two neighboring cities are no more likely to be close in value than the true SMRs of two distant cities.)
2. The resulting distribution of true SMRs is approximately normal (or at least roughly "bell"-shaped), with a variance V.

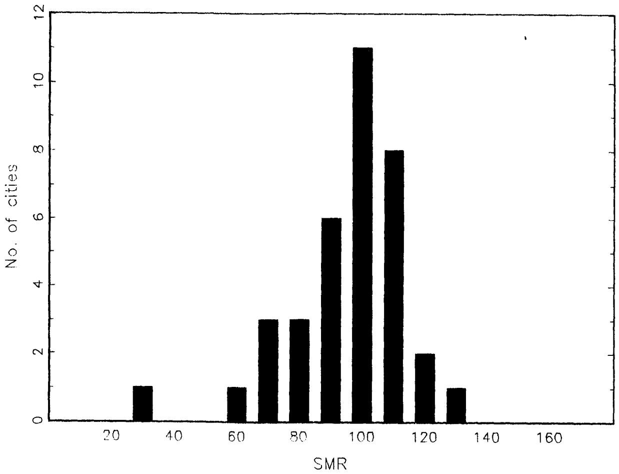

FIGURE 1. Toxoplasmosis SMRs.

Thus, the distribution of the observed SMRs is the sum of two distributions: the distribution of true SMRs (described by assumptions 1 and 2) and the distribution of sampling errors. This model is often described as a "two-stage sampling model": First, nature "samples" the true SMRs from a true-SMR distribution; second, the investigator takes samples to estimate these true SMRs. Under this model, the variance V of the true SMR distribution must be less than that of the observed SMR distribution, since the variance of the observed distribution is the sum of the variance of the true distribution and the average variance of the sampling errors. We will discuss these assumptions later.

Under the above assumptions, extreme (small or large) SMR estimates are more likely to represent unstable observations than are central estimates; indeed, in Table 1, the six largest standard errors occur among the seven SMRs more than 25% above or 20% below the mean. This result is analogous to the well-known "regression to the mean" phenomenon observed in taking measurements on people: When subjects are sampled from a population in which true values have a bell-shaped distribution, but the values are measured with error, extreme measured values are likely to be the product of extreme errors. Consequently, subjects that are extreme on the first measurement will probably regress toward the population mean upon a second measurement. The methods we will describe attempt to account for such effects; the "subjects" being measured are, however, towns rather than people, and the object of a measurement is an SMR rather than a blood pressure.

An intuition for the relation between the true and observed SMRs may come from viewing the true and observed SMR distributions as representing a distribution before and after misclassification due to sampling error. As an example, suppose the true distribution is as in Figure 2. For simplicity, suppose the effect of sampling error is to produce the following probabilities of classification: 20% for getting an observed (sample) SMR two categories above or below the correct category (10% each), 40% for one category above or below (20% each), and 40% for the correct category. Application of this error law to the distribution in Figure 2 yields Figure 3 for the misclassified SMR distribution. As can be seen, the smallest observed values are likely to be smaller than the true SMR in the population from which they arose. Likewise, the largest observed values are likely to be larger than the true SMR from which they arose.

Under assumptions 1 and 2, a "good" set of estimates of the true SMRs will take account of the effect of sampling error by regressing (shrinking) the observed SMRs toward the estimated mean of the true SMRs. Column 7 of Table 1 presents the results of regressing the observed SMRs in column 4 toward their mean (of 100). Each SMR is moved toward the mean according to the generalized Stein rule discussed by Efron and Morris.[2] The amount of movement for each SMR varies directly with the standard error of the SMR; to a good approximation, the estimates in column 7 equal 100 + B_i (SMR_i − 100), where B_i is the empirical-Bayes factor $\hat{V}/(\hat{V} + SE_i^2)$, $\hat{V}$ is the estimated variance of the true SMRs distribution. Thus, the SMRs for cities 1 and 36 (which have large standard errors) are moved toward the mean by a large proportion of their original distance from the mean, while the SMR for city 2 (which has a small standard error) is moved very little.

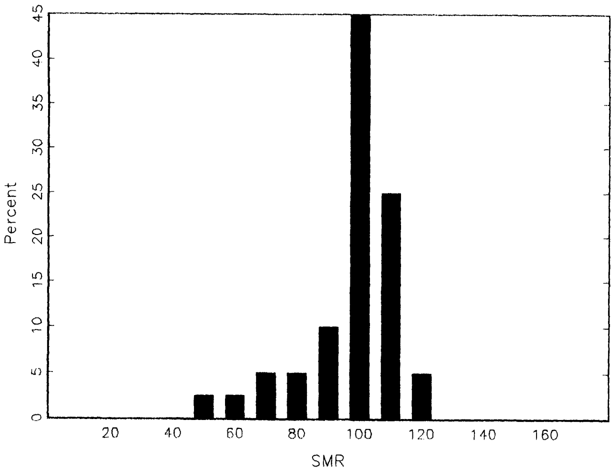

FIGURE 2. Example of true SMRs.

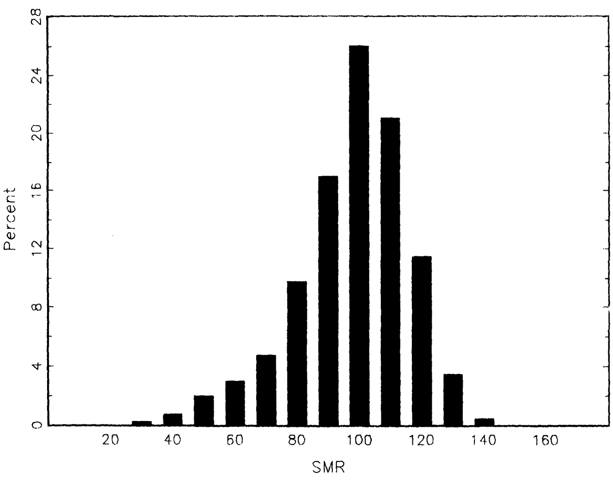

FIGURE 3. Misclassified SMRs.

Should we prefer the empirical-Bayes estimates in column 7 of Table 1 over the ordinary estimates in column 4? Under the above assumptions, there are frequentist and Bayesian arguments for doing so. Both arguments assume that: (1) our objective is to minimize the *total* costs arising from errors in the observed SMRs, and (2) the cost of each error is a strictly increasing function of the absolute distance between the estimated SMR and the true SMR. In mathematical terms, these arguments assume that, if the absolute error in the estimate for city *i* is denoted by A_i and the cost of an error of size A is C(A), then (1) our objective is to minimize total cost $\Sigma_i C(A_i)$, and (2) $C(A_i) > C(A_j)$ if and only if $A_i > A_j$.

Given the preceding assumptions, the frequentist reason for favoring the empirical-Bayes adjustment is that, over repeated sampling (that is, upon multiple resurveys), our expected total cost incurred when using the empirical-Bayes estimates will be smaller than the expected total cost when using the ordinary estimates. One should note, however, that the errors (and hence costs) for some cities may be higher under the empirical-Bayes approach if those cities differ from the majority on important and uncontrolled predictors of the disease; this problem is analogous to confounding in etiologic studies.

The Bayesian reason for favoring the empirical-Bayes adjustment sounds identical to the frequentist reason:

Given the assumptions, the expected total cost is smaller for the empirical-Bayes estimates than for the ordinary estimates. Nevertheless, the "expected" in the Bayesian argument does not refer to an average over repeated sampling; instead, it refers to the average over a subjective (personal) probability distribution for the 36 true SMRs—a distribution with mean 100, variance *V*, and (in our example) assumed to have an approximate "bell" shape. Furthermore, there is an even stronger Bayesian reason for favoring the empirical-Bayes adjustment: Under the assumptions, for any given city we should (subjectively) expect the cost to be smaller if we use the adjustment.

We forego the mathematical and logical details of these arguments, which can be found elsewhere.[2–6] We note that the proper form of the adjustment will vary according to (1) the form of the distributional assumptions (that is, the form of the sampling-error distribution and the form of the distribution of true values), (2) the form of the cost function (for example, intervention costs might vary from city to city, and one would certainly want to take this into account), and (3) the manner of adjustment for covariates. For example, in the toxoplasmosis data, there is some skewness of the SMR distribution to the left, which can be accounted for by a geographic variable (see below). But these details are not at the heart of the multiple-comparisons controversy.

Instead, we wish to emphasize the general implications of the technical results: If we want to minimize costs or allocate fixed resources on the basis of some true population values, but our estimates of these values are contaminated by "random error," then a Bayes or empirical-Bayes adjustment can be preferable to no adjustment. This is so, regardless of whether we view the "random error" as variability that we cannot yet predict, as Rothman and Bayesians do, or as an objective (chance) phenomenon, as frequentists do.

Empirical-Bayes Assumptions

The preferability of the empirical-Bayes adjustment depends, of course, on the correctness of the assumptions it uses (beyond those that underlie both the ordinary estimates and the empirical-Bayes estimates). This is why, in employing such adjustments, the assumptions used should be grounded in the best available information and should be critically examined as part of the analysis. In the above example, both assumptions appear to be false. Assumption 1 (randomness of SMRs) appears false because the observed SMRs in the eastern third of the country (cities 25, 32, 35, and 36) form a cluster of low values. Assumption 2 (approximate normality) appears false insofar as the lower tail of Figure 1 looks too heavy.

Both these observations can be "explained" (in the statistical sense) by a geographic covariate that indicates whether a city is east or west of the Rio Lempa (the country's major river)[4]: When this covariate is "regressed out" of the observed SMRs (by weighted linear regression of the SMRs on the covariate), the residuals conform well to approximate normality. Thus, by including this covariate as a regressor in the empirical-Bayes analysis, we can replace the untenable random-determinants and approximate-normality assumptions (1 and 2) with more tenable assumptions of random determinants and approximate normality within the regions east or west of the river. Under these modified assumptions, the empirical-Bayes adjustment proceeds by regressing the residuals (from the regression of the SMRs on the geographic covariate) to their mean of zero. The resulting empirical-Bayes SMR estimate for a city east (west) of the Rio Lempa is the empirical-Bayes adjusted residual for the city *plus* the weighted mean SMR for all cities east (west) of the Rio Lempa. (The remaining available covariates [city size, rainfall, and elevation] did not appear to be associated with the SMRs.[12])

Although empirical-Bayes adjustments require assumptions in addition to those used for ordinary maximum-likelihood estimation, they need not depend on specific distributional assumptions, such as assumption 2. In fact, in their original formulation, they involve no assumptions about the distribution of the parameters being estimated.[6] Rather, the fundamental assumption underlying empirical-Bayes methodology is that information about certain subsets of parameters in the study can provide information about other parameters in the study.

Consider, for example, the prevalence of toxoplasmosis in city 1: Viewed in isolation, without any knowledge of the typical prevalence of toxoplasmosis in El Salvador, we would have no idea that the indirectly adjusted prevalence estimate for city 1, (129/100) (3,504/5,174) = 87%, is suspiciously high. But, upon seeing the other, lower, estimates and the large standard error for the city 1 estimate, city 1 becomes a prime suspect as an estimate inflated by sampling error. The lower estimates from the other cities lead us to expect a lower prevalence for city 1, and the large standard error for the city 1 estimate indicates that the true SMR could easily be much lower than the estimate. In the absence of any other information to allay our suspicion (such as city 1 having an unusually high value for a known determinant of toxoplasmosis), the empirical-Bayes method regresses our city 1 estimate to the mean in direct proportion to its sampling variance.

Of course, the fundamental assumption can fail. To take an extreme example, suppose our study objective was to estimate simultaneously toxoplasmosis prevalence in city 1, dental caries prevalence in city 2, hypertension prevalence in city 3, and myopia prevalence in city 4. We doubt that information about any subset of these prevalences would tell us anything about the remaining prevalences, and so we would not expect the adjusted estimates to be closer to the true values than the unadjusted estimates. Thus, from a Bayesian perspective, we would not recommend applying an empirical-Bayes adjustment to this study.

More generally, we cannot be certain that our empirical-Bayes adjusted estimates are closer to their corresponding true values than are the original, unadjusted estimates. Just as with model-based adjustments for age, sex, and other potential confounders, we carry out the empirical-Bayes adjustments because we reason that, if our model assumptions are correct, the adjusted estimates will be closer on average to the true values than the unadjusted estimates. Thus, from the empirical-Bayes or Bayesian perspective, multiple comparisons are not really a "problem" (except in a computational sense). Rather, the multiplicity of comparisons provides an opportunity to improve our estimates through judicious use of any prior information (in the form of model assumptions) about the ensemble of parameters being estimated. Unlike conventional multiple comparisons, empirical-Bayes and Bayes approaches will alter and can improve point estimates and can provide more powerful tests and more precise (narrower) interval estimators (see References 5 and 6 for descriptions and discussions of empirical-Bayes and Bayesian tests and interval estimates). But, like all model-based adjustments, these benefits are purchased at the cost of dependence on model assumptions.[13]

Discussion

We have argued that, under certain precise and often practical formulations of decision problems, empirical-Bayes adjustments for multiple comparisons are useful; so are some closely related Bayes procedures.[6] Why then did Rothman recommend against multiple-comparison adjustment in general? And what characterizes the conditions under which such adjustments are useful? Unsurprisingly, the two questions are closely related.

Rothman[1] argued that the conventional statistical doctrine underlying corrections for multiple comparisons was based on two presumptions:

1. Chance not only can cause the unusual finding in principle, but it does cause many or most such findings.
2. No one would want to earmark for further investigation something caused by chance.

Rothman then argued that both presumptions are wrong. In essence, what we are arguing is that, sometimes, both presumptions are right. In particular, the assumptions underlying our example and others like it imply both these presumptions.

In our toxoplasmosis example, there was a known physical sampling process that contributed random errors to the estimates. For some cities, these sampling errors were of such large magnitude that they were not only capable of causing unusual findings (that is, unusually high or low SMRs), but were also *likely* to cause some unusual findings. Therefore, presumption 1 was satisfied. In addition, presumption 2 was satisfied because no one would want to investigate a city or assign a high (or low) allocation to a city if the city had a high (or low) SMR estimate solely because of survey sampling error.

Rothman did not discuss situations satisfying presumptions 1 and 2, perhaps because he regards such situations as falling in the domain of policy, not science. We maintain, however, that a large portion of epidemiologic research takes place to serve policy needs, and that there is no sharp boundary between scientific and policy-oriented studies: A study may be of only scientific interest to some parties, but of great policy interest to others. Furthermore, many scientific studies satisfy the above presumptions: Consider that presumption 1 will be satisfied in any study that involves randomization (as in clinical trials) or random sampling (as in case-control studies with randomly selected controls). In such studies, no one would want to earmark for further investigation any finding produced by sampling error or by random variation in treatment assignment, and so presumption 2 is also satisfied.

Some of the preceding points are illustrated by a recent multiple-outcome case-control study of chemical exposures and cancer mortality at a large industrial facility.[14] A key question among many was "which, if any, of the 100 or so exposure–outcome combinations should be selected for further study by the company, given that resources are very limited and only a few can be pursued?" Combinations representing well-accepted association (for example, asbestos–lung cancer) could be set aside because much better, earlier data had already been used to establish policy. But, for most of the combinations, there was little or no background information. Also, most combinations (including those showing the largest associations) involved small numbers of exposed cancer cases, and the target cohort was represented only by selected cancer deaths (cases) and selected deaths from other causes (controls). Consequently, regardless of whether one regards "random variation" as truly random or as the product of multiple independent extraneous causes, random variation was as potent a candidate explanation for the largest observed association as was the hypothesis that the exposure had an effect; thus, Rothman's presumption 1 was fulfilled. Furthermore, no one had any desire to further investigate a hypothesized effect of an exposure if the exposure's association with disease were due only to random error; thus, Rothman's presumption 2 was fulfilled.

The form of our problem and our objectives were similar to those in the toxoplasmosis example: We had a multiple-decision problem (to pursue or not pursue each of several associations), and, as discussed below, we wished to minimize our expected total cost of error (as measured by the total distance between the estimates and the true associations). We thus chose to supplement conventional analyses with certain multiple-comparisons procedures—such as empirical-Bayes adjustments[15]—to see which, if any, of the observed associations could not be explained by random variation.

To justify the simplest empirical-Bayes analysis (in which a given rate-ratio estimate is regressed to the grand mean of all the estimates based on sampling error in the given estimate), one would have to assume, similar to assumption 1, that the true rate ratios were assigned at random to the different exposure–disease combinations. This random-assignment assumption would not hold if, based on subject matter knowledge (for example, animal experiments or previous epidemiologic studies), one believed that the association of trichloroethylene with kidney cancer was more likely to be causal than the association of asbestos exposure with leukemia. Analogous to the use of the geographic covariate in the toxoplasmosis example, a more sophisticated empirical-Bayes analysis would regress the estimate of a given rate ratio to the estimated mean of those rate ratios thought to be of the same order of magnitude as (that is, exchangeable with) the given rate ratio.[15]

Of course, which rate ratios are thought to be of the same order of magnitude is a subjective judgment that might well provoke disagreement among various observers. This disagreement is somewhat analogous to disagreements about choice of model form, such as the controversy over the appropriate model to use when predicting the effects of low-dose radiation from high-dose radiation studies.

Conclusions

As the previous discussion may show, we are not challenging Rothman's arguments insofar as they pertain to settings in which presumptions 1 and 2 do not hold. But different analysts, or the same analyst at different times, may have different analytic objectives for the same data. When one's analytic objectives are to reach a decision based on the data (such as whether to pursue further study of an association), certain adjustments for multiple comparisons are useful. When one's analytic objectives are to summarize compactly a complex data set, such adjustments may be of more doubtful utility.

In considering a problem that involves multiple outcomes based on multiple estimates, the total costs stemming from errors in the estimates may be highly interdependent, and these interdependencies should influence our decisions. Such interdependencies occur throughout fields with technologic or policy inputs. Epidemiology is no exception: when resources are limited, $100,000 spent to study trichlorethylene and brain tumors is $100,000 not spent to study polychlorinated biphenyls and esophageal cancer. Intelligent multiple-comparisons procedures are intended to aid in making such decisions; they are not designed to replace any component of good scientific analysis, such as accurate description and summarization of data.

Rothman's essay is a valuable warning against thoughtless or indiscriminate use of multiple-comparisons procedures. It would, however, be unfortunate if his essay led some to believe that the more thoughtful of these procedures are never useful.

Acknowledgments

We are most grateful to Charles Poole for his thorough critique of this manuscript.

References

1. Rothman KJ. No adjustments are needed for multiple comparisons. Epidemiology 1990;1:43–46.
2. Efron B, Morris CN. Stein's estimation rule and its competitors: an empirical Bayes approach. J Am Stat Assoc 1973;68:117–130.
3. Efron B, Morris CN. Data analysis using Stein's estimator and its generalizations. J Am Stat Assoc 1975;70:311–319.
4. Efron B, Morris CN. Stein's paradox in statistics. Scientific American 1977;236:119–127.
5. Morris CN. Parametric empirical Bayes inference: Theory and applications (with discussion). J Am Stat Assoc 1983;78:47–65.
6. Maritz JS, Lwin T. Empirical Bayes Methods (2nd ed.). New York: Chapman and Hall, 1989.
7. Thomas DC, Siemiatycki J, Dewar R, et al. The problem of multiple inference in studies designed to generate hypotheses. Am J Epidemiol 1985;122:1080–1095.
8. Leamer EE. Specification Searches. New York: Wiley, 1978.

9. Oakes M. Statistical Inference. Chestnut Hill, MA: Epidemiology Resources Inc., 1990.
10. Raiffa H, Schlaiffer R. Applied Statistical Decision Theory. Cambridge, MA: Harvard University Press, 1961.
11. Wald A. Statistical Decision Functions. New York: Wiley, 1950.
12. Remington JS, Efron B, Cavanaugh E, et al. Studies on toxoplasmosis in El Salvador: Prevalence and incidence of toxoplasmosis as measured by the Sabin-Feldman dye test. Trans Roy Soc Trop Med Hyg 1970;64:252–267.
13. Robins JM, Greenland S. The role of model selection in causal inference from nonexperimental data. Am J Epidemiol 1986;123: 392–402.
14. Salvan AS, Greenland S, Wegman DH, et al. Case-control analyses of cancer mortality in a large-transformer manufacturing plant. Abstracts of the Twelfth Scientific Meeting of the International Epidemiological Association: 45.
15. Greenland S. A semi-Bayes approach to the estimation of correlated multiple associations, with an application to a large occupational cancer-mortality study. Statistics in Medicine 1991 (in press).