

THE PROBLEM OF MULTIPLE INFERENCE IN STUDIES DESIGNED TO GENERATE HYPOTHESES

D. C. THOMAS,^{1,2} J. SIEMIATYCKI,³ R. DEWAR,¹ J. ROBINS,⁴ M. GOLDBERG,¹ AND B. G. ARMSTRONG⁵

Thomas, D. C. (Department of Preventive Medicine, U. of Southern California, Los Angeles, CA 90033), J. Siemiatycki, R. Dewar, J. Robins, M. Goldberg, and B. G. Armstrong. The problem of multiple inference in studies designed to generate hypotheses. *Am J Epidemiol* 1985;122:1080-95.

Epidemiologic research often involves the simultaneous assessment of associations between many risk factors and several disease outcomes. In such situations, often designed to generate hypotheses, multiple univariate hypothesis-testing is not an appropriate basis for inference. The number of true positive associations in a collection of many associations can be estimated by comparing the observed distribution of *p* values for the positive associations to a theoretical uniform distribution, or to the observed distribution of negative associations, or to an empiric randomization distribution. None of these approaches, however, will distinguish the true from the false positive associations. Various criteria for selecting a subset of associations to report are considered by the authors, including Bonferroni adjustment of *p* values, splitting the sample for searching and testing, Bayesian inference, and decision theory. The authors prefer an approach in which all associations in the data are reported, whether significant or not, followed by a ranking in order of priority for investigation using empirical Bayes techniques. Methods are illustrated by application to preliminary data from a study aimed at identifying hitherto unsuspected occupational carcinogens.

risk; biostatistics; epidemiologic methods; neoplasms; occupational diseases

Though much of scientific research can be described as hypothesis-testing, the generation of new hypotheses is an essential activity and in fact comprises much of the work of epidemiologists. New hypotheses can arise in a number of ways. Some are suggested by theory, by findings in related fields, or by astute clinical observations.

Some arise serendipitously in the course of studies designed to test other hypotheses. And some are generated by projects that systematically monitor large data bases to consider many possible associations. However generated, a new hypothesis is not usually accepted until the association has been replicated in data other than that

Received for publication January 3, 1984 and in final form March 11, 1985.

Abbreviations: AN, attributable number; EB, empirical Bayes; LR, likelihood ratio; ML, maximum likelihood; RR, relative risk.

¹ Department of Epidemiology and Health, McGill University, Montreal, Quebec, Canada.

² Currently at the Department of Preventive Medicine, University of Southern California, 2025 Zonal Ave., Los Angeles, CA 90033. (Send reprint requests to Dr. Duncan Thomas at this address.)

³ Epidemiology and Preventive Medicine Research

Center, Institut Armand-Frappier, Laval-des-Rapides, Quebec, Canada.

⁴ Occupational Health Program, Harvard School of Public Health, Boston, MA.

⁵ School of Occupational Health and Safety, McGill University, Montreal, Quebec, Canada.

This work was supported by grants from the National Cancer Institute of Canada and Fonds de la recherche en santé du Québec.

The authors thank Dr. David Oakes for helpful comments on an earlier draft of this paper.

which generated it. Any study which collects information on a large number of "stimulus" and "response" variables has a high probability of producing wild goose chases which can consume much valuable research time and resources to refute. This is simply the price that must be paid for the advancement of our knowledge about true associations. (In addition to this statistical limitation, large data bases also tend to have less detailed and less accurate data on any particular association than studies designed specifically around a single association. While important, this issue is not of concern here.)

The Boston Collaborative Drug Surveillance Project (1) provides an example of a large-scale monitoring project. In the course of routine analyses of these data, the association between reserpine and breast cancer was noted, and this finding was published (2). (In this case, publication occurred simultaneously with two other reports of the same association, lending greater credibility to the finding.) Eight further studies followed soon after, most of them negative. Reasons for the discrepancies have been reviewed elsewhere (3) and include various methodological differences that are not of concern here; to some extent, however, the original reports may simply have been a fluke. The high probability of such flukes arising in large exploratory studies is sometimes viewed as cause for concern.

While the problem is clearest in studies designed to monitor many variables, the same issues arise when a study designed to test a particular hypothesis is reanalyzed to consider a variety of alternatives not proposed originally. For example, MacMahon et al. (4) recently reported an association between coffee consumption and pancreatic cancer in an epidemiologic case-control study originally aimed at examining the role of smoking. The results were criticized on various methodologic grounds (5), including the way in which the serendipitous finding was interpreted and reported.

It was argued, for example, that because many hypotheses were considered in the post hoc reanalyses, a more stringent criterion for claiming "statistical significance" should have been used (e.g., multiplying p values by the number of tests carried out). Though widespread, this view is not universally held (6, 7), and we feel it is inappropriate for reasons explored below.

This paper discusses several approaches to the problem of reporting results when a large number of associations have been examined, either in routine analyses of large data bases or post hoc reanalyses of studies designed to test other hypotheses. First, we discuss whether an investigator should report all associations examined, select a subset of "significant" ones, or rank them on some priority scale. Second, we describe several methods for judging whether the distribution of associations is any different from what might be expected merely by chance, considering the number of associations examined.

To simplify presentation, we will use the terms "positive association" and "negative association" to signify, respectively, those associations between disease and exposure with relative risks (RRs) greater than 1.0 (i.e., harmful) and less than 1.0 (i.e., protective). This is in contrast with the usage in some reports where "negative association" signifies "no association" or a true RR of 1.0; the latter we shall term a "null association." Implicit in all of the approaches developed below is the assumption that the distribution of RRs in the population sampled is a mixture of two distributions, a "spike" consisting of those associations for which the true RR exactly equals 1, and a continuous distribution consisting of those for which the true RR \neq 1. Our prior belief is that there is a finite probability (probably large) of RR = 1, and a continuous distribution of non-null RRs larger and perhaps also smaller than 1. We denote by T_0 the number of associations for which the true RR = 1, and by T_+ and T_- the numbers for which the true RR > 1 and

RR < 1, respectively. In a later section of the paper, methods for estimating T_0 , T_+ , and T_- are discussed.

To illustrate the various approaches, we will use preliminary data from a large data base we are assembling for the purpose of identifying new occupational carcinogens (8–10). In this project, cases with any of 12 cancer sites are being interviewed for information on possible exposure to several hundred chemicals. Each cancer site group is then compared to all of the others combined as referents to obtain RRs for each possible exposure-disease combination, recognizing that in view of the proportional incidence design, any carcinogen which affected a broad spectrum of sites might go undetected or underestimated. The analysis is done in two stages. First, all associations are examined separately using the Mantel-Haenszel procedure (11) to adjust each for a set of a priori confounders, but not for each other. A subset of associations (based on these results) is then subjected to stepwise polytomous logistic regression analysis, forcing confounding variables in first, to adjust each association for all other associations included in the model. As argued below, it would be neither feasible nor desirable to adjust for all associations under consideration. For the purpose of this methodologic discussion, a restricted data set based on all 12 sites of cancer but only 57 of the several hundred exposure factors will be used. Even this restricted data set generates estimates of $T = 12 \times 57 = 684$ RRs. What reporting strategy should be adopted?

GENERAL APPROACHES TO REPORTING ASSOCIATIONS

Three general strategies are possible: 1) to select a number of "significant" or otherwise remarkable associations and report only those; 2) to report the entire matrix of T relative risks and associated confidence intervals or p values; and 3) to rank the T associations on some scale of priority for further investigation. Each of these general

strategies has several variants which are discussed below.

Selective reporting of associations

In part due to the pressure to limit the length of manuscripts for publication and in part due to the predominance of hypothesis testing in statistics, there has been a tendency for investigators to select only a subset of associations they consider worthy of reporting. The five approaches described in this section are all intended to provide some basis for selective reporting of associations using binary-valued ("report"/"no report") decision rules.

The significance level criterion. Statistical significance, central to the frequentist tradition of statistics, is widely used by scientists and journal editors as a basis for deciding which associations to report, the $p < 0.05$ level being conventional. When multiple hypotheses are considered, the question then arises whether the same degree of significance should be applied to each association. A common recommendation is to control the overall "experiment-wise" Type I error rate by adopting a suitably conservative "test-wise" alpha level, e.g., by setting the test-wise level equal to the experiment-wise level divided by the number of tests (commonly known as the Bonferroni adjustment (12)). One would then report only those associations that were "significant" using this stricter test-wise level. The difficulty is that in a large-scale hypothesis-generating study, this criterion becomes so conservative that the probability of detecting any true associations is virtually nil. (The problem is essentially the same as the trade-off between sensitivity and specificity in screening for disease.) In our preliminary data, for example, a 5 per cent experiment-wise level would translate to a test-wise level of $\alpha = 0.0000073$, at which only three associations were significant (all being variants of the already established association between nickel and lung cancer). Taken to its logical extreme, one might argue that an investigator should control his "career-

wise" alpha-level, or even that all investigators should agree to control the "discipline-wise" level. In practice, not even the strictest frequentist behaves this way. Some, however, would argue that the basic distinction is whether or not multiple hypotheses are being tested *in the same data set*. In our opinion, the reuse of the same data indicates the need for multivariate methods to obtain unconfounded estimates of causal parameters but has no implication for the choice of alpha level.

Splitting the sample. Another approach, still grounded in a frequentist philosophy, is to split the sample of subjects into two groups, one of which is used to explore a large number of hypotheses, the other being used to test those that were significant in the first analysis. Only those hypotheses that were significant in both subsamples are then reported. There is some debate over whether such splitting should be done randomly or along natural lines, such as by geography or date of onset of disease. Random splitting of the data has nothing to be said for it, for the following reason—its effect is simply to replace a nominal significance level of α by an overall significance level of α^2 , since the two tests are strictly independent; thus, the same objective could be achieved by analyzing the combined data with the stricter significance level α^2 . The statistical power of the split-sample approach, for the same overall significance level, is, however, consistently much poorer, as shown in table 1, derived by the method given in Appendix 1.

Splitting the sample along natural lines suffers from exactly the same loss of statistical power, but at least mimics the effect of studying two separate populations. For example, an association may have greater credibility if it is apparent in both males and females, whites and nonwhites, two cities or time periods, or using two different methodologies (assuming no true effect modification or differential bias). The sizes of the two subsamples should be based on the relative importance the investigator assigns to searching versus testing activities. In particular, it would probably be unwise to use a data set covering a unique population solely for searching, as that would effectively preclude any independent confirmation except by prospective observations.

Bayesian inference. Bayesian inference uses some estimate of the prior credibilities of each of the T associations to calculate the posterior probabilities that $RR > 1$ for each association, given the study data. One might then choose to report those associations for which this posterior probability (or some other summary of the posterior distribution of RRs) was "sufficiently large," how large being arbitrary like the choice of significance level in frequentist inference. In actual practice, most scientists behave like intuitive Bayesians, combining the evidence from their data with evidence from other studies as well as other disciplines in the discussion sections of their papers, to arrive at a qualitative judgment of the posterior probability that any particular association is true. Application

TABLE 1
Statistical power of a randomly split study* compared with the power of a single study†

Significance level for the split study α	Significance level for the single study α^2	Power of the single study ($1 - \beta$)		
		0.800	0.950	0.990
0.10	0.0100	0.691§	0.878	0.956
0.05	0.0025	0.681	0.872	0.953
0.01	0.0001	0.665	0.862	0.949

* Decision rule for the split study: report association if significant at the α level of significance in both halves.

† Decision rule for the single study: report association if significant at the α^2 level of significance.

§ Power of the split study, $1 - \beta'$, assuming both subsamples are of equal size.

of formal Bayesian methods would, however, require quantitative specification of these prior probabilities. The difficulty of doing this has until recently precluded their practical use in epidemiology, particularly for systematic investigations of many associations and especially in exploratory studies where there is no prior information on most of them.

Bayesian decision theory. Rather than using arbitrary critical values for the posterior probability in deciding which associations to report, a Bayesian decision rule would combine these posterior probabilities with estimates of the costs and benefits of alternative decisions. Thus, the benefit of a true positive report (or the cost of failing to report a true positive) might be taken to be proportional to the true "attributable number" (AN, the number of cases of the disease in the population attributable to the exposure factor); this might be further refined by taking into account the feasibility of reducing the population exposure, the loss of life expectancy caused by the disease, and so on. The costs of a false positive would comprise such things as the expense of further studies to test the association, the concern to exposed individuals, and so on; as a first approximation, these might be taken to be proportional to the apparent attributable number. The decision concerning whether or not to report each association would then be based on whether the expected utility were positive (net benefit) or negative (net cost). Like Bayesian inference, the approach suffers from the practical difficulty of specifying prior probabilities for each association, as well as corresponding costs and benefits.

To overcome this difficulty, one might apply the same logic, not to evaluate the merits of reporting each association individually, but to evaluate the overall performance of alternative decision rules on repeated sampling. Thus, while it might be difficult to specify the costs and benefits of reporting any particular association, it should be much easier to obtain a general

consensus on a reasonable cost/benefit structure *on average*. The expected overall utility of various reporting strategies can then be simulated from an assumed prior distribution, the power function of the reporting strategy, and the relative weights given to the costs and benefits, in the hope of finding a strategy which consistently maximized the utility over a broad range of choices.

Presenting all associations

There is a bias in the epidemiologic and toxicologic literature due to the fact that "statistically significant" associations are more likely to be published than "nonsignificant" ones. Whether this results from editorial policy or self-censorship by investigators is not at issue here. The ability of the community of epidemiologists to evaluate the evidence concerning a suspect risk factor and to pool the results from the world's experience is compromised by this unmeasurable bias. It is therefore as important to publish findings of no effect as findings of statistically significant associations. Furthermore, any method of decision rules or ranking requires a set of values and assumptions to be imposed by the investigator which in some sense distort the simple messages in a data set. In one form or another, estimates of all exposure-disease associations in a data set (and corresponding confidence intervals) should be published. With hundreds or thousands of associations, this could become a very large table indeed, but various compromises are possible. For example, those exposures that were not associated (according to some reasonable criterion) with any disease need not be presented in detail but only listed; or the full table need not be published in a scientific journal but only made available as a technical report.

Ranking associations

An attractive and practical compromise between simply reporting the entire matrix on the one hand or selecting significant associations to report on the other, is to

report all associations together with a ranking according to the priority with which they should be replicated. This could be done in several ways. On the one hand, it could be argued that the most important associations to investigate are those most likely to be causal, and it is sometimes stated (13) that the best single index of causality is the strength of the RR. On the other hand, from a public health point of view it could be argued that the appropriate index for ranking is the AN (or variants of it as described above).

Whatever measure of the magnitude of the association is used, an essential element of the ranking is the strength of the evidence in the data as summarized by the p value. For example, an RR = 10 may be much less important than an RR = 1.5 if the first is based on 0.1 expected case ($p > 0.2$) and the latter on 100 ($p < 0.001$). Conversely, a p value of 0.00001 may be less important than one of 0.001 if the former derives from an RR of 1.5 and the latter an RR of 10. A further difficulty with point estimates is that even if they are individually unbiased (e.g., maximum likelihood (ML) estimates asymptotically), they have the following undesirable ensemble property: given that a particular estimate is among the largest of a set of estimates of related parameters, it is more likely to be an overestimate than an underestimate of its true value, and conversely. (This is analogous to the familiar "regression to the mean" phenomenon.) Furthermore, estimates with large sampling variability are likely to be *more* over- (or under-) estimated than those with smaller variability. A possible combination of magnitude and variability would be to rank the lower confidence limits on the chosen parameter, but the choice of confidence level would be arbitrary, and of course the ranking would only be relevant to those estimates in the top half of the distribution. For example, of the 17 associations selected in our preliminary stepwise analysis, two that shared the same ranking of 80 per cent

lower limits differed by five ranks in their 95 per cent limits. Empirical-Bayes (EB) methods (14) were developed to provide less arbitrary "best estimates" for ensembles of related parameters.

Let ρ_k , $k = 1, \dots, K$, denote a set of true but unknown population parameters (e.g., log odds ratios or attributable numbers) to be estimated, and let $\hat{\rho}_k$ be ML estimates of ρ_k . The standard Bayesian approach is to provide a prior distribution $f(\rho_k | \theta_k)$ for the population parameters, where the parameters θ_k of each prior distribution are assumed to be known. The resulting estimates of $\hat{\rho}_k$ are systematically "pulled back" from their ML values towards their prior expectations by proportions that depend on the relative variances of the observed values and prior distributions. Thus, the more variable estimates are pulled back more than the precise ones and their relative ranking may be altered. In the EB approach, the population parameters ρ_k are assumed to have been sampled from a single distribution parametered by a common value θ (known as a "hyperparameter") which could therefore be estimated from the data rather than having to be specified by prior knowledge.

One might reasonably wonder how one could justify assuming that two associations, concerning different exposures and/or different diseases, share a common distribution. The answer lies in the concept of "exchangeability", which essentially states that in the absence of prior knowledge about either association, one is as likely to be true as another. Thus, for example, if we discover strong evidence that one previously unsuspected chemical is carcinogenic, it becomes likely that there exist other unsuspected carcinogens and the prior distribution for all other associations is shifted upwards. (Extensions of the technique described elsewhere (15) allow prior beliefs about the relative credibilities of the various hypotheses to be incorporated into the parameters of the prior distribution.)

Appendix 2 provides a description of our

implementation of these procedures. We assumed that the log odds ratio estimates \hat{r}_k were independently normally distributed around their true values ρ_k with variances \hat{s}_k^2 (which for simplicity we treat as known) and that the true values ρ_k were distributed as a mixture of two normal distributions with masses α and $1 - \alpha$, means μ_i , and variances σ_i^2 ($i = 1$ for the "true null" associations and $i = 2$ for the true positive associations). Certainly, other prior distributions could be considered. For example, Oakes (unpublished manuscript) assumed that the observed numbers of exposed cases had a Poisson distribution, with expectations having a gamma prior distribution whose parameters can depend on confounding variables (see reference 15 for details). Further work is needed in order to take account of the covariances of the \hat{r}_k .

The EB procedure provides estimates of the hyperparameters α , μ_i , σ_i^2 for the prior distribution, together with estimates of the posterior distribution of ρ_k for each association. For the prior considered here, the posterior distribution is most easily summarized by two quantities described in Appendix 2, $\hat{\pi}_{1k}$, the posterior probability that $\rho_k \neq 0$, and $\hat{\rho}_{1k}$, the posterior expected value of ρ_k given that $\rho_k \neq 0$. In our application, we found that the rankings of $\hat{\pi}_{1k}$ and $\hat{\rho}_{1k}$ were virtually identical, so we report only the latter.

For simplicity, the ML estimates \hat{r}_k and EB estimates $\hat{\rho}_{1k}$ are given in table 2 only for the subset of 17 associations selected by the stepwise logistic analysis. The two sets of estimates are shown on both RR and AN scales, in descending order of their \widehat{AN}_{EB} . Not surprisingly, the two scales provided quite different rankings, reflecting differences in the frequency of exposure and disease. For example, the largest $\widehat{RR}_{ML} = 4.76$ produced one of the smallest $\widehat{AN}_{ML} = 5.2$ because both disease and exposure were rare, whereas the second smallest $\widehat{RR}_{ML} = 1.55$ produced the second largest $\widehat{AN}_{ML} = 24.1$. On both scales, the proportion by which the EB estimate is pulled back to-

ward its prior mean is inversely related to the expected number of exposed cases. As we assumed a common prior distribution for RRs, their EB estimates are all pulled back toward the single value and span a fairly narrow range. The corresponding prior means for ANs (assuming a common distribution of RRs) depend on their sample sizes, so their EB estimates show much greater variation.

In estimating the hyperparameters, we used all 684 Mantel-Haenszel estimates, as the stepwise subset would not represent a random sample of all associations or even of the subpopulation of true positive associations. Despite the fact that the two sets of estimates were obtained in different ways and are therefore not strictly comparable, we prefer to report EB estimates derived from the logistic regression as they are adjusted for each other and redundant associations have been eliminated. Because of computing costs and problems of identifiability, it is not feasible to obtain logistic estimates for all 684 associations, nor would that be desirable because it would lead to overadjustment and consequently inflated variances.

The various hyperparameter estimates are summarized in table 3. The simplest model, assuming all $\rho_k = 0$, produced poor fit (likelihood ratio (LR) $\chi_2^2 = 82.78$, for the improvement compared to the single normal model). The addition of a spike at $\rho = 0$ to the single normal further improved the fit (improvement LR $\chi_1^2 = 4.06$) and produced larger estimates of the mean and variance of the non-null distribution; 95 per cent of the non-null RRs are estimated to lie in the range of 0.9-2.0. The estimate of the proportion of true-null associations $1 - \alpha = 0.52$ is lower than obtained by methods described in the next section, though its variance is large. The low estimate may also reflect departures from the null value caused by uncontrolled confounding or inappropriate choice of reference. Replacing the spike for the null associations by a second normal distribution

TABLE 2
Maximum likelihood (ML) and empirical Bayes (EB) estimates of risk parameters for associations selected by stepwise analysis§

Total no. of cases	Proportion exposed	Expected no. of exposed cases	Relative risk estimates				Attributable number estimates			
			ML*	EB†	Prior†	Pull back %	ML*	EB†	Prior†	Pull back %
246	0.198	48.7	1.55	1.43	1.30	47	24.2	19.3	13.7	46
212	0.108	22.9	1.89	1.52	1.30	62	18.6	11.3	6.6	61
100	0.198	19.8	1.85	1.48	1.30	67	14.4	8.7	5.6	65
246	0.039	9.6	3.88	1.86	1.30	78	24.9	8.0	2.8	77
69	0.306	21.1	1.58	1.39	1.30	67	10.4	7.4	5.7	65
187	0.091	17.0	1.88	1.45	1.30	74	13.9	7.3	4.9	73
255	0.051	13.0	2.22	1.54	1.30	74	15.0	6.9	3.8	73
100	0.167	16.7	1.53	1.37	1.30	69	8.1	5.8	4.7	68
40	0.306	12.2	1.90	1.42	1.30	80	8.6	4.5	3.3	77
187	0.088	7.1	1.97	1.50	1.30	70	6.6	3.5	2.1	69
255	0.013	3.3	4.71	1.96	1.30	81	11.7	3.1	1.0	80
212	0.027	5.7	2.01	1.42	1.30	83	5.6	2.4	1.7	82
36	0.108	3.9	3.12	1.51	1.30	88	6.7	1.9	1.1	86
69	0.043	3.0	2.85	1.47	1.30	89	5.1	1.4	0.9	88
69	0.032	2.2	2.81	1.45	1.30	90	3.8	1.0	0.6	89
40	0.041	1.6	4.76	1.56	1.30	92	5.2	0.9	0.5	91
246	0.005	1.2	3.61	1.41	1.30	95	3.1	0.5	0.4	95

* ML estimate derived from logistic regression coefficient, adjusted for the other associations selected by stepwise analysis and for a priori confounders.

† Based on \hat{p}_i , the expected value of $\ln RR$ given $RR \neq 1$.

‡ Estimated prior mean, assuming common distribution for RRs, fitted to Mantel-Haenszel estimates of RR for all 684 associations.

§ Entry criterion: $p < 0.10$ for the score statistic conditional on previously entered associations, no associations eliminated.

TABLE 3
Hyperparameter estimates from the empirical Bayes analysis of all 684 Mantel-Haenszel odds ratios

Proportion of RRs = 1 α	Distribution of RRs \neq 1		Log likelihood	Likelihood ratio χ^2 (df)*
	Median μ	Variance (1nRR) σ^2		
1.00	-	-	-467.29	82.78 (2)
0.00	1.16	0.031	-425.90	4.06 (1)
0.52	1.36	0.045	-423.87	

* df, degrees of freedom.

to allow for this, however, produced only a trivial improvement in fit.

ESTIMATING THE NUMBER OF ASSOCIATIONS TO BE EXPECTED BY CHANCE

Irrespective of the reporting strategy used, one of the issues in evaluating a set of results is to determine whether the number of observed associations exceeds the number to be expected by chance, and if so by how many. There are at least four methods that can be used to estimate the numbers of true and false positive associations

among those observed. Table 4 presents the observed numbers of associations at several levels of significance, together with estimates of the numbers expected by the various methods.

Assuming a fixed alpha level

The most straightforward approach to estimating the expected numbers uses the fact that, under the global null hypothesis that $RR = 1$ for all associations, the distribution of p values is uniform. (The uniformity of p values under the null hypothesis remains true even if the associations are not independent; however, their tendency to cluster will then inflate the variance of estimated expected numbers.) The expected number of "significant" associations can therefore be obtained by multiplying the alpha level by the total number of tests. Column A in table 4 shows expectations based on this method. For example, at the 5 per cent level of significance, $0.05 \times 684 = 34.2$ would be expected to be significant by chance, of which half (17.1) would be positive; in contrast, the observed number of significant positive associations was 36. Of course, the choice of α level is arbitrary. One way to mitigate this arbitrariness is to summarize across α levels by integration, as described below.

Whatever summary is used, however, this approach still relies on the assumption of the uniformity of the p value distribution, an assumption which can fail either because of the presence of a substantial number of true-positive associations or because of departures from the assumptions of the test on which the p values are based.

TABLE 4
Frequencies of significant positive associations at various α levels

Significance level (one sided) α	No. of significant positive associations			
	Observed	Expected*		
		A	B	C
<i>Based on all T = 684 associations</i>				
0.25	106	85.5	78	85.5
0.10	60	34.2	19	42.5
0.05	36	17.1	9	24.8
0.01	22	3.4	0	8.3
0.001	9	0.3	0	1.8
<i>Based on subset of T = 218 associations with at least 5 expected exposed cases</i>				
0.25	47	27.3	31	25.6
0.10	30	10.9	14	8.8
0.05	16	5.5	7	3.2
0.01	10	1.1	0	0.5
0.001	6	0.1	0	0.2

* A) Based on the global null hypothesis, assuming a uniform distribution of p values; B) based on observed number of negative associations, assuming symmetry around the null; C) Based on four random permutations of the disease classification of the study subjects.

For example, if the test is based on asymptotic theory, p value distributions obtained from small samples may not be uniform even if the global null hypothesis is true. This situation can be improved by restricting the p value comparisons to those based on sufficiently many expected cases. In the bottom part of table 4, the associations are restricted to those based on five or more expected exposed cases. This restriction has the undesirable effect of precluding detection of associations where less than five were expected. The association between vinyl chloride and angiosarcoma, for example, which was discovered by a cluster of three cases with about 0.0004 expected, would not have been counted under this restriction. Furthermore, the restriction does not eliminate any non-uniformity under the null hypothesis caused by uncontrolled confounding or selection bias.

Assuming all negative associations are noncausal

A variant of the above approach is to derive the expected number of positive associations from the observed number of negative associations, i.e., to assume on prior grounds that all observed significant associations with relative risks below 1.0 (protective effects) are false. For example, it seems reasonable to assume that no occupational exposure decreases the risk of cancer. On the global null hypothesis, the distributions of p values should be symmetric around 0.5 (i.e., $RR = 1$). Thus, the number of such false significantly low risks is an estimate (whose validity is discussed below) of the expected number of false significantly high risks. Subtracting this expected number from the observed number of significantly high risks gives an estimate of the number of true associations. This is the basis for the expectations given in column B of table 4. For example, at $\alpha = 0.05$, 36 positive associations were observed compared to nine negative associations, so this estimate of the number of true associations is 27. Again, this is dependent on the alpha

level chosen, but the estimates can be presented for a range of α levels.

This approach is dependent on the assumption of symmetry of p values around the null hypothesis, which can fail because of small numbers or confounding. The former is the most likely reason for the shortfall of significant negative associations in the top part of table 4, since many of the associations are based on very small expected numbers of exposed cases where it becomes virtually impossible to obtain a significant negative association. Therefore, the method should only be applied (if at all) to those associations based on large numbers, as in the bottom part of table 4. Aside from the problem of small numbers, uncontrolled confounding is likely to induce asymmetry in the distribution of p values, probably inducing more spurious positive than spurious negative associations. Furthermore, if the study associations are not adjusted for each other, some negative associations are bound to arise not because the exposure is truly positive, but because it is negatively associated with other hazardous exposures.

p value plotting

A third method, proposed by Schweder and Spjotvoll (16), is essentially a generalization of the first. Under the global null hypothesis, p values are uniformly distributed and hence their cumulative distribution is linear. Any departure from a straight line at the small- p end would indicate the presence of some true positive associations. By fitting a line to the majority of the p values, the point on the vertical axis where it intersects the origin gives an estimate of the number T_0 of true null hypotheses in the data. This technique essentially provides an integral over the entire range of p values of the difference between the observed p value density function and a uniform density. However, unlike the above methods where the expected numbers are derived by assuming the global null hypothesis, this method uses the best estimate of the number of true null hypotheses derived by visual or least-squares fitting to a subset

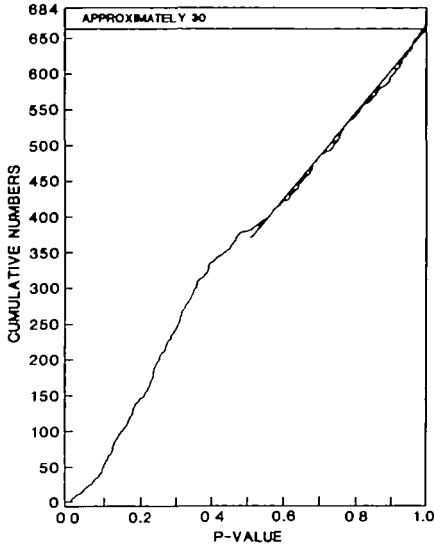


FIGURE 1. Cumulative distribution of p values (null value 0.5) for all 684 associations.

of the data. Figure 1 shows the p value plot of all $T = 684$ tests. The distribution is clearly non-uniform and no straight line can be confidently drawn through all of the data. If fitting is restricted to the range of p values from about 0.55 to 0.95, an estimate of T_+ of about 30 is obtained. For the subset of 218 associations based on at least five expected exposed cases (figure 2), the distribution is much more uniform, leading to estimates of T_+ of 19 or 24, depending on whether the negative associations are included in the fitting.

Randomization

The most cumbersome but most reliable approach is based on computer simulation. The real subjects are randomly permuted among the disease groups and the matrix of relative risks and corresponding p values computed. By repeated permutation, one can obtain an expected distribution of p values for comparison against the observed number. Table 4 (column C) and figure 3 give the results of four random permutations of the disease variable (keeping the total numbers of subjects in the 12 groups and the subjects' exposure and confounder values fixed). Thus at $\alpha = 0.05$, 25 positive and six negative associations would have

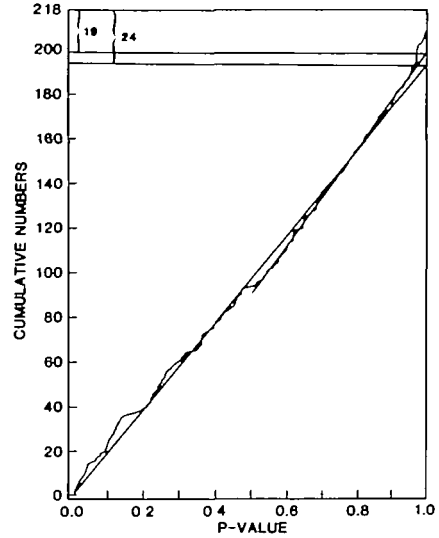


FIGURE 2. Cumulative distribution of p values (null value 0.5) for subset of 218 associations based on at least five expected exposed cases.

been expected by chance compared to the 36 and nine observed. An estimate of the number of true associations, integrated over α levels, can be obtained as described in Appendix 3. There is, however, a degree of arbitrariness in this procedure resulting from the choice of weight function or the range of p values used in the integration: our estimates ranged from 23 to 37 true

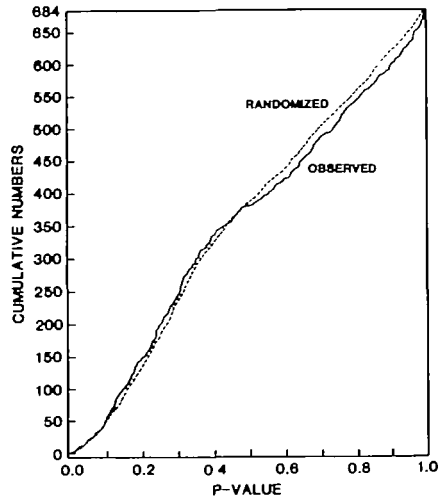


FIGURE 3. Comparison of cumulative distributions of observed and randomized p values (null value 0.5) for all 684 associations.

TABLE 5

Integrated estimates of the number of true positive and true negative associations based on the randomized data

Assumed limit of p values for true associations		No. of associations outside range				Estimated nos. of true associations	
Positive	Negative	Observed		Randomized		\hat{T}_-	\hat{T}_+
		Positive	Negative	Positive	Negative		
0.25*	0.90*	194	96	700	283	37	32
0.20	0.80	146	60	515	170	26	20
0.20	0.85	146	120	515	413	27	25
0.20	0.95	146	146	515	528	27	24
0.30	0.80	238	60	917	170	23	20
0.30	0.85	238	120	917	413	25	24
0.30	0.95	238	146	917	528	25	23
Average of above six estimates:		-	-	-	-	26	23

* Choices producing the largest estimates of T_- and T_+ .

negative and 20 to 32 true positive associations (table 5). The most important advantage of this approach is that it requires no assumptions about the theoretic distribution of the test statistics, so there is no need to eliminate the associations based on small numbers.

DISCUSSION

When many statistical tests are carried out, there will inevitably be some false positives, however the associations are selected. Though it is sometimes assumed that these false positive associations are merely statistical artifacts, they do in fact represent coincidences that exist in the real world, irrespective of whether epidemiologists and statisticians are there to observe them, and irrespective of whether the association is observed as part of a multivariable monitoring system, a single factor case-referent study, or a case report by an astute clinician of a cluster of patients having some common characteristic. Apparent associations that come from a monitoring study, if they have never been observed and reported previously, deserve the same dissemination and attention as the well-documented report of a cluster by a clinician. One of the benefits of a systematic data collection system over the case-report "sys-

tem" is precisely the ability to place observed associations into an overall probabilistic context. As shown above, there are many ways to present results of such analyses.

We recommend the reporting of estimates of RRs and their standard errors for all associations under study. Further, the number of observed positive associations should be described and compared with estimates of the number expected by chance using one or more of the methods presented. Because many of the associations will be interrelated, the use of multivariate methods will be necessary to sort out which associations make independent contributions and which are merely reflections of other associations. These methods impose practical restrictions on the number of associations that can be considered, which can be dealt with by means of a stepwise selection procedure, forcing in a priori confounders first. Other than the need for such selection in using multivariate methods, we do not advocate reporting of only a subset of associations. The results of these univariate and multivariate analyses can then be ranked using the empirical Bayes approach. Finally, each of the high ranking associations should be rigorously analyzed using classic case-control meth-

ods to control confounding factors and to study the variation in risks with duration and intensity of exposure and other factors. Such detailed analyses would allow the investigator greater confidence in pinpointing those hypotheses most worthy of further research.

Multiple inference in epidemiologic research is a long-standing and controversial problem. Our purpose here is to stimulate discussion of the various options.

REFERENCES

- Jick H. The discovery of drug-induced illness. *JAMA* 1977;296:481-5.
- Boston Collaborative Drug Surveillance Program. Reserpine and breast cancer. *Lancet* 1974;2:669-71.
- Labarthe DR. Methodologic variation in case-control studies of reserpine and breast cancer. *J Chronic Dis* 1979;32:95-104.
- MacMahon B, Yen S, Trichopoulos D, et al. Coffee and cancer of the pancreas. *N Engl J Med* 1981;304:630-3.
- Feinstein AR, Horwitz RI, Spitzer WO, et al. Coffee and pancreatic cancer: the problems of etiologic science and epidemiologic case-control research. *JAMA* 1981;246:957-61.
- Cole P. The evolving case-control study (with discussion). *J Chronic Dis* 1979;32:15-34.
- Miettinen OS. Contribution to the discussion of a paper by Labarthe. *J Chronic Dis* 1979;32:111.
- Siemiatycki J, Day NE, Fabry J, et al. Discovering carcinogens in the environment: a novel epidemiological approach. *JNCI* 1981;66:217-25.
- Siemiatycki J, Gerin M, Hubert J. Feasibility of an exposure-based case-control approach to discovering occupational carcinogens. In: Peto R, Schneiderman M, eds. Quantification of occupational cancer. Banbury Report 1981;9:471-83.
- Siemiatycki J, Gerin M, Richardson L, et al. Preliminary report of an exposure-based, case-control monitoring system for discovering occupational carcinogens. *Terat Carcin Mutag* 1982;2:169-77.
- Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *JNCI* 1959;22:719-48.
- Jones DR, Rushton L. Simultaneous inference in epidemiologic studies. *Int J Epidemiol* 1982;11:276-82.
- Breslow NE, Day NE. *Statistical methods in cancer research. I. The analysis of case-control studies.* IARC scientific publication no. 32. Lyon: IARC, 1980;58-69.
- Morris C. Parametric empirical Bayes inference: theory and applications (with discussion). *J Am Statist Assoc* 1983;78:47-65.
- Thomas DC. The problem of multiple inference in identifying point source environmental hazards. *Environ Health Perspect* 1985;62:411-18.
- Schweder T, Spjøtvoll E. Plots of P -values to evaluate many tests simultaneously. *Biometrika* 1982;69:493-502.
- Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B* 1977;39:1-38.

APPENDIX 1

Calculation of the statistical power of the split sample approach

Let z_1' and z_2' be normal random deviates for any particular association tested in subsamples 1 and 2, respectively. We wish to calculate the power $1 - \beta'$ of the decision rule "report if both tests are significant at a nominal level α ". This is

$$\begin{aligned} 1 - \beta' &= \Pr\{z_1' > Z_\alpha \text{ and } z_2' > Z_\alpha \mid H_1\} \\ &= \Pr\{z_1' > Z_\alpha \mid H_1\} \Pr\{z_2' > Z_\alpha \mid H_1\} \\ &= \{\Pr\{z' > Z_\alpha - E(z' \mid H_1) \mid H_0\}\}^2. \end{aligned} \quad (1.1)$$

We wish to compare this with the power of a pooled analysis of the single sample. This comparison is only meaningful if made at the same overall significance level, which is α^2 because the two tests are independent. For the single sample, we therefore have a power

$$\begin{aligned} 1 - \beta &= \Pr\{z > Z_{\alpha^2} \mid H_1\} \\ &= \Pr\{z > Z_{\alpha^2} - E(z \mid H_1) \mid H_0\}. \end{aligned} \quad (1.2)$$

By definition, we have

$$1 - \beta = \Pr\{z > Z_{1-\beta} \mid H_0\} \quad (1.3)$$

and

$$\sqrt{(1 - \beta')} = \Pr\{z > Z_{\sqrt{1-\beta'}} \mid H_0\}. \quad (1.4)$$

Combining powers 1.1 and 1.4, we obtain

$$\Pr\{z' > Z_{\sqrt{1-\beta'}} | H_0\} = \Pr\{z' > Z_\alpha - E(z' | H_1) | H_0\},$$

and combining powers 1.2 and 1.3, we get

$$\Pr\{z > Z_{1-\beta} | H_0\} = \Pr\{z > Z_\alpha - E(z | H_1) | H_0\},$$

so

$$Z_{\sqrt{1-\beta'}} = Z_\alpha - E\{z' | H_1\} \tag{1.5}$$

and

$$Z_{1-\beta} = Z_\alpha - E\{z | H_1\} \tag{1.6}$$

Since the pooled sample test z is based on twice the sample size of each of the split sample tests z' , we have

$$E\{z' | H_1\} = E\{z | H_1\} / \sqrt{2}. \tag{1.7}$$

This allows us to combine equations 1.5 and 1.6 to produce

$$\begin{aligned} Z_{1-\beta'} &= Z_\alpha - E\{z | H_1\} / \sqrt{2} \\ &= Z_\alpha - (Z_\alpha - Z_{1-\beta}) / \sqrt{2}. \end{aligned} \tag{1.8}$$

Equation 1.8 is solved for $1 - \beta'$ and tabulated for a variety of values of α and $1 - \beta$ in table 1.

APPENDIX 2

Implementation of empirical Bayes estimation techniques

We take the parameter of interest to be the log of the relative risk, partly because its sampling distribution is approximately normal and partly because we find it easier to specify its prior distribution than for other parameters. Once its empirical Bayes estimates have been obtained, estimates of the attributable number (or other parameters) can be obtained by applying the usual conversion formula to the empirical Bayes estimate of the relative risk. Let r_k denote the observed logRR for the k th association and let ρ_k denote its (unknown) true value. We postulate that the distribution of ρ_k is a mixture of two normals,

$$f(\rho_k) = \alpha N(\mu_1, \sigma_1^2) + (1 - \alpha) N(\mu_2, \sigma_2^2), \tag{2.1}$$

where α represents the probability of relative risk originating from population indexed 1. This model is motivated by the recognition that ρ_k may (with probability α) reflect only nonsampling error due to uncontrolled confounding, nonideal choice of reference group, etc., rather than (with probability $(1 - \alpha)$) a true biologic causal relationship. In practice, available data may not always provide enough power to adequately identify all five parameters in equation 2.1. In these situations, simpler special cases of equation 2.1 may be useful. For example, we may assume that $\alpha = 0$ so that $f(\rho_k)$ becomes a simple normal distribution, or $\mu_1 = \sigma_1^2 = 0$ so that $f(\rho_k)$ becomes a mixture of a normal distribution and a spike at $RR = 1$.

The r_k are assumed to be independently normally distributed with mean ρ_k and known variance s_k^2 (not necessarily all equal),

$$g(r_k | \rho_k) = N(\rho_k, s_k^2). \tag{2.2}$$

(Further work is needed to take into account the covariances of the estimates r_k .)

We use a combination of Newton-Raphson iteration and the E-M algorithm (17) to find the maximum likelihood estimates of the underlying parameters α , μ_i , and σ_i^2 ($i = 1, 2$). The parameter α is most easily estimated from the marginal distribution of r_k .

$$\begin{aligned} M(r_k) &= \int f(\rho_k) g(r_k | \rho_k) d\rho_k \\ &= \alpha N(\mu_1, \sigma_1^2 + s_k^2) + (1 - \alpha) N(\mu_2, \sigma_2^2 + s_k^2), \end{aligned} \tag{2.3}$$

by maximizing the log likelihood $L = \sum_k \ln M(r_k)$ using trial values of the remaining parameters μ_i and σ_i^2 . These parameters are estimated simultaneously using the E-M algorithm. In the E-step, the posterior probability $\hat{\pi}_{ik}$ that r_k arises from underlying distribution i , and the posterior expectations $\hat{\rho}_{ik}$ of the true risks ρ_k given that it comes from underlying distribution i are calculated using the current estimates of all the underlying parameters as follows:

$$\hat{\pi}_{ik} = \alpha_i N(r_k | \mu_i, \sigma_i^2 + s_k^2) / M(r_k) \quad (\alpha_1 = \alpha, \alpha_2 = 1 - \alpha) \quad (2.4a)$$

$$\hat{\rho}_{ik} = E(\rho_{ik}) = (s_k^2 \mu_i + \sigma_i^2 r_k) / (s_k^2 + \sigma_i^2). \quad (2.4b)$$

In the M-step, these estimates are then used to estimate the underlying parameters μ_i and σ_i^2 , as follows:

$$\hat{\mu}_i = \sum_k E(\rho_{ik}) \hat{\pi}_{ik} / \sum_k \hat{\pi}_{ik} \quad (2.5a)$$

$$\hat{\sigma}_i^2 = \sum_k E(\rho_{ik}^2) \hat{\pi}_{ik} / (\sum_k \hat{\pi}_{ik} - 1) - \hat{\mu}_i^2 \quad (2.5b)$$

where

$$\begin{aligned} E(\rho_{ik}^2) &= \hat{\rho}_{ik}^2 + \text{var}(\hat{\rho}_{ik}) \\ &= \hat{\rho}_{ik}^2 + \sigma_i^2 s_k^2 / (\sigma_i^2 + s_k^2). \end{aligned}$$

One iteration of the Newton step to estimate α and of the E-M algorithm to estimate the remaining parameters can be carried out simultaneously. Upon convergence, the estimates of the underlying parameters may give insight into the number of true positive associations and the distribution of errors and true relative risks, whereas the estimates of the posterior probabilities $\hat{\pi}_{ik}$ and posterior expectations $\hat{\rho}_{ik}$ can be used for ranking associations.

APPENDIX 3

Estimation of the number of true null associations using the randomized p values

Let $\hat{f}(p)$ denote the probability density function (pdf) for the observed p values and $\hat{g}(p)$ the pdf for the randomized p values, obtained as described in "Randomization". (The p values are scaled so that 0.5 denotes no association, ~ 0 a significant negative and ~ 1 a significant positive association.)

In line with the general prior discussed in the text, we postulate that

$$f(p) = [T_0 g(p) + T_- h_-(p) + T_+ h_+(p)] / T \quad (3.1)$$

where $h_-(p)$ and $h_+(p)$ are the unknown pdf's for the true negative and true positive associations, respectively, and $T_0 + T_- + T_+ = T$ are the true numbers of each type of association. Note that no assumption is required that $g(p)$ be uniform.

We can therefore estimate T_0 by minimizing the difference between $T\hat{f}(p)$ and $\hat{T}_0\hat{g}(p)$ over some range of p values which, hopefully, excludes most of $h_-(p)$ and $h_+(p)$. More generally, we might adopt a weighted least-squares criterion of the form

$$\text{WSS}(\hat{T}_0) = \int_0^1 w(p) [T\hat{f}(p) - \hat{T}_0\hat{g}(p)]^2 dp \quad (3.2)$$

where $w(p)$ is an arbitrary weight function to be specified. Equation 3.2 is minimized by

$$\hat{T}_0 = T \int w(p) \hat{f}(p) \hat{g}(p) dp / \int w(p) \hat{g}^2(p) dp. \quad (3.3)$$

However, one should avoid using too fine an interval in evaluating these integrals, as the numerator becomes infinitesimal while the denominator remains finite as the interval gets very small. In our data, this phenomenon became apparent as the average number of observations in any interval got less than about 10. Using 20 intervals, the estimates of T_0 for a variety of weight functions ranged from 667 to 678, depending mainly on the weight assigned to the tail areas.

An even simpler criterion is based on the cumulative distributions,

$$F(p) = [T_0 G(p) + T_- H_-(p) + T_+ H_+(p)] / T. \quad (3.4)$$

Thus, if we could choose p_- and p_+ such that $H_-(p_-) \approx 1$ and $H_+(p_+) \approx 0$, then

$$T\hat{F}(p_-) = \hat{T}_0\hat{G}(p_-) + \hat{T}_-$$

$$T\hat{F}(p_+) = \hat{T}_0\hat{G}(p_+) + \hat{T}_+$$

and solving to \hat{T}_0 , we would obtain

$$\hat{T}_0 = T[\hat{F}(p_+) - \hat{F}(p_-)] / [\hat{G}(p_+) - \hat{G}(p_-)] \quad (3.5a)$$

$$\hat{T}_- = T\hat{F}(p_-) - \hat{T}_0\hat{G}(p_-) \quad (3.5b)$$

$$\hat{T}_+ = T - \hat{T}_0 - \hat{T}_-. \quad (3.5c)$$

In practice, we cannot know where to choose p_- and p_+ , so any choice is arbitrary and subject to random variation. We recommend first searching for the values which minimize \hat{T}_0 to find a lower bound (recognizing that as the minimum of a random variable, it is bound to be a biased estimate) and then using an average of \hat{T}_0 values over a variety of conventional p_- and p_+ values in the general neighborhood of the ones which produced this minimum. In our data, a minimum of $\hat{T}_0 = 615$ was obtained at $p_- = 0.25$ and $p_+ = 0.90$; we therefore adopted the values $p_- = 0.20$, and 0.30 and $p_+ = 0.80, 0.85, 0.95$ to describe the range and mean of T_0 , T_- , and T_+ estimates given in table 5.