American Journal of Epidemiology
Copyright © 1992 by The Johns Hopkins University School of Hygiene and Public Health
All rights reserved

Vol 135, No. 11
Printed in U.S.A

Methods for Trend Estimation from Summarized Dose-Response Data, with Applications to Meta-Analysis

Sander Greenland and Matthew P. Longnecker

Meta-analysis often requires pooling of correlated estimates to compute regression slopes (trends) across different exposure or treatment levels. The authors propose two methods that account for the correlations but require only the summary estimates and marginal data from the studies. These methods provide more efficient estimates of regression slope, more accurate variance estimates, and more valid heterogeneity tests than those previously available. One method also allows estimation of nonlinear trend components, such as quadratic effects. The authors illustrate these methods in a meta-analysis of alcohol use and breast cancer. *Am J Epidemiol* 1992;135:1301–9.

epidemiologic methods; logistic models; meta-analysis; risk assessment

Meta-analytic methods for clinical trial data often assume that sufficient data are available from each study to allow use of ordinary analytic methods. Nevertheless, meta-analyses of observational studies often have to rely on the limited data available from research reports, and they may have to reconstruct the more complete data required for regression analysis (1).

To obtain a regression slope from a research report, one may have to pool estimates for responses at different levels of exposure or treatment. Current methods for pooling estimates assume independence of the estimates, an assumption that is never true because the estimates for separate exposure levels depend on the same reference (unexposed) group. We present two new methods of pooling that account for the correlation between estimates, and we compare the results of applying these methods with the results from methods that assume independence.

TREND ESTIMATION FROM A SINGLE REPORT

As a motivating example, consider the case-control data in table 1 on alcohol and breast cancer, first presented by Rohan and McMichael (2). From these data, we wish to estimate the coefficient β in the logit-linear (linear-logistic) model

$$\lambda(x, z) = \alpha + \beta x + \boldsymbol{\delta}'\mathbf{z},$$

where x is alcohol intake, $\mathbf{z}$ is the vector of covariates, and λ is the log odds of being a case in the study versus being a control. We do not have access to the original data, nor did the published article present enough data to allow us to fit the model to the data. Nevertheless, we can construct an estimate of β by using weighted least squares to regress the adjusted log

Received for publication July 16, 1990, and in final form March 3, 1992.

From the Department of Epidemiology and the Center for Occupational and Environmental Health, School of Public Health, University of California, Los Angeles, CA.

This research was supported by the Center for Occupational and Environmental Health, UCLA School of Public Health. Dr. Sander Greenland was also supported by grant HS 05753–01 from the National Center for Health Services Research, and Dr. Matthew P. Longnecker was also supported by a grant from the International Life Sciences Institute.

TABLE 1. Case-control data on alcohol use and breast cancer, as presented by Rohan and McMichael (2)

Alcohol (g/day)	Assigned dose (g/day)	No. of cases	No. of controls	Total	Crude OR*	Adjusted OR†
0	0	165	172	337	1.0	1.0‡
<2.5	2	74	93	167	0.83	0.80 (0.51–1.27)§
2.5–9.3	6	90	96	186	0.98	1.16 (0.73–1.85)
>9.3	11	122	90	212	1.41	1.57 (0.99–2.51)
Total		451	451	902		

* OR, odds ratio.

† Odds ratio from age-matched conditional logistic regression including variables for history of benign breast disease, bilateral oophorectomy, smoking, education, family history of breast cancer, ages at first and last menstrual period, age at first live birth, ever use of oral contraceptives, ever use of replacement estrogens, and practice of breast self-examination.

‡ Referent.

§ Numbers in parentheses, 95% confidence interval.

odds ratios from table 1 on the exposure doses listed in column 1 of the table (1). Doing so yields an estimated β of $b = 0.0334$, with an estimated variance for b of $v = 0.0003494$.

Given the logistic model, the estimator b of β obtained using the preceding method is consistent for β. Nevertheless, b is inefficient; worse, the variance estimate v obtained from this regression underestimates the true variance of b (see Appendix 1). In effect, the variance estimator for b assumes that the log odds ratios are uncorrelated, an assumption that is never satisfied in practice and is often grossly violated. We have therefore developed a new approach that yields an efficient point estimator and a consistent variance estimator under assumptions more likely to be approximated in practice. Our approach is based on constructing an approximate covariance estimate for the adjusted log odds ratios from a fitted table that conforms to the adjusted log odds ratios.

For case-control and cumulative cohort data, our estimates are computed as follows:

1) Let the reference exposure level be coded zero;
   N_x = the total number of subjects at exposure level x;
   $\mathbf{N}$ = the vector of N_x;
   M_1 = the total number of cases;
   L_x = the adjusted log odds ratio estimate for exposure level x ($x \neq 0$) versus the reference level ($x = 0$);
   $\mathbf{L}$ = the vector of L_x ($x \neq 0$);
   v_x = the estimated variance for L_x (see Greenland (1) for methods of computing v_x from published reports);
   $\mathbf{v}$ = the vector of v_x ($x \neq 0$).
2) Fit cell counts to the interior of the total data table (which has margins N_x and M_1) such that $A_xB_0/(A_0B_x) = \exp(L_x)$, where A_x and $B_x = N_x - A_x$ are the fitted numbers of cases and noncases at exposure level x. (See Appendix 2 for a simple fitting algorithm.)
3) For $x \neq z$, estimate the asymptotic correlation of L_x and L_z by

$$r_{xz} = (1/A_0 + 1/B_0)/s_x s_z,$$

   where s_x^2 = crude variance estimate = $1/A_x + 1/B_x + 1/A_0 + 1/B_0$.
4) Estimate the asymptotic covariance of L_x, L_z by

$$c_{xz} = r_{xz}(v_x v_z)^{1/2}.$$

5) Estimate β by weighted least squares for correlated outcomes:

$$b^* = v_b^* \mathbf{x}' C^{-1} \mathbf{L},$$

$$v_b^* = \widetilde{\text{var}}(b^*) = (\mathbf{x}' C^{-1} \mathbf{x})^{-1},$$

where $\mathbf{x}$ is the vector of observed nonzero exposure levels and $C = \widetilde{\text{cov}}(\mathbf{L})$ has diagonal elements v_x and off-diagonal elements c_{xz}.

Step 5 is easily carried out using a matrix programming language such as GAUSS, SC, APL, S-PLUS, or SAS IML.

Consistency of b^* under the logit model follows immediately from consistency of $\mathbf{L}$. As Appendix 3 shows, b^* is more efficient than b, and v_b^* is consistent for var(b^*) under the assumptions that

1) the crude odds ratio parameters approximately equal the adjusted odds ratio parameters, i.e., the sampling distribution is strictly collapsible (3);
2) the correlation matrices of the crude and adjusted odds ratios are approximately equal;
3) the variances of the crude odds ratios can be approximated by the usual formulas based on the multinomial or Poisson distributions.

Assumption 3 is a standard assumption for unmatched studies. When assumption 3 is violated, it is usually because matching has been employed; nevertheless, numerous studies indicate that the impact of matching on variances is usually small (e.g., see reference 4). Assumptions 1 and 2 will be satisfied when the adjustment factors are only weakly related to the exposure and outcome. Assumption 1 can be checked by comparing the crude odds ratios with the adjusted odds ratios. In any case, some set of externally specified constraints is necessary in order to allow estimation to proceed when the covariate-specific data are unreported, and assumptions 1–3 are far more reasonable than assuming that the L_x's are uncorrelated (which has, up until now, been the only recourse in dose-response meta-analyses). We also note that assumptions 1–3 are sufficient but not necessary for b^* and v^* to outperform b and v.

For the Rohan and McMichael (2) data, we applied the above steps as follows:

1) The exposure categories were assigned levels of 0, 2, 6, and 11 g/day; $\mathbf{N} = (337, 167, 186, 212)'$; $M_1 = 451$; $\mathbf{L} = (\log 0.80, \log 1.16, \log 1.57)' = (-0.223, 0.148, 0.451)'$; and $\mathbf{v} = (0.0542, 0.0563, 0.0563)'$.
2) The fitted cell values were 160.5, 70.3, 95.5, and 124.7 for cases and 176.5, 96.7, 90.5, and 87.3 for controls at exposure levels 0, 2, 6, and 11. As a numerical check on the computations, note that these reproduce the adjusted odds ratios, e.g., 70.3(176.5)/160.5(96.7) = 0.80.
3) $s_2 = (1/70.3 + 1/96.7 + 1/160.5 + 1/176.5)^{1/2} = 0.19095$; similarly, $s_6 = 0.18280$ and $s_{11} = 0.17711$. Thus, $r_{2,6} = (1/160.5 + 1/176.5)/0.19095(0.18280) = 0.3408$; similarly, $r_{2,11} = 0.3518$ and $r_{6,11} = 0.3674$.
4) $c_{2,6} = 0.3408[0.0542(0.0563)]^{1/2} = 0.0188$; similarly, $c_{2,11} = 0.0194$ and $c_{6,11} = 0.0207$.
5) $\mathbf{x} = (2, 6, 11)'$,

$$\mathbf{C} = \begin{bmatrix} 0.0542 & 0.0188 & 0.0194 \\ 0.0188 & 0.0563 & 0.0207 \\ 0.0194 & 0.0207 & 0.0563 \end{bmatrix},$$

$v_b^* = 0.0004270$, and $b^* = 0.0454$.

The last two numbers should be contrasted with the uncorrected results, $b = 0.0334$ and $v_b = 0.0003494$. The regression-fitted odds ratio for the highest alcohol level (11 g/day) versus no alcohol is exp[11(0.0454)] = 1.65 for the corrected results but exp[11(0.0334)] = 1.44 for the uncorrected results. The inverse-variance weight assigned to this study in a meta-analysis of the type discussed below would be 1/0.0004270 = 2,342 using the covariance-corrected variance but 1/0.0003494 = 2,862 using the uncorrected variance.

Because Rohan and McMichael (2) reported the crude data, we may check assumption 1 by comparing the crude odds ratios with the adjusted odds ratios. All of the crude odds ratios are within 20 percent of the adjusted odds ratios, which indicates that there is no major violation of assumption 1.

The above method extends to analyses of person-time rate ratios, upon appropriate redefinition of terms. Beta becomes the coefficient in a log-linear (exponential) Poisson regression; N_x becomes the total person-time observed at exposure level x; the L_x's become adjusted log rate ratios; cell counts are fitted such that $A_xN_0/(A_0N_x) = \exp(L_x)$; and r_{xz} becomes $1/(A_0s_xs_z)$, where $s_x^2 = M_1/A_xA_0$. For the analysis of risk ratios (as in a cohort study with N_x persons, rather than person-time), these formulas may be applied with $s_x^2 = M_1/A_xA_0 - 1/N_0 - 1/N_x$ and $r_{xz} = (1/A_0 - 1/N_0)/s_xs_z$.

EMPIRICAL COMPARISONS OF THE ESTIMATORS

The objective of the above method is to approximate the logistic coefficient that would have been obtained had either more complete study data or the estimated logistic coefficient been reported, and to provide a less biased variance estimate than was previously available. To compare and evaluate the uncorrected and corrected estimators, we analyzed 10 published data sets (5–14) for which there were enough data reported to compute the maximum likelihood estimate of the logistic coefficient, $\hat{\beta}$.

The results are summarized in table 2. As expected, both b and b^* are fairly close to the logistic coefficient from the full data. Also as expected, the variance estimator v for b appears to underestimate the true variance of b, for it provides values below the estimated variance for $\hat{\beta}$ in 9 out of 10 of the data sets.

The variance estimates for b^* tend to equal or exceed the variance estimates for $\hat{\beta}$; this is somewhat reassuring, given that $\hat{\beta}$ is fully efficient and b^* is generally not unless assumptions 1–3 hold. One large discrepancy occurs for the alcohol-esophageal cancer study (10). This study shows considerable heterogeneity of the alcohol slope across age categories; in such cases, the ordinary (inverse-information) variance estimate for the maximum likelihood estimate is suspect, and some authors recommend refitting the model with a dispersion parameter or with random effects to account for the apparent overdispersion (15). With a random-effect term added to the full-data model, the variance estimate for $\hat{\beta}$ is much closer to that for b^*. We also applied b^* to data sets in which there was statistically significant heterogeneity of the slope across strata (not shown), and found its variance estimate to be much larger than the variance estimate for $\hat{\beta}$ in those cases; this result is again reassuring, since the conventional variance estimate for $\hat{\beta}$ would be an underestimate in such cases (15).

APPLICATION TO META-ANALYSIS

The coefficient and variance estimates obtained from research reports often form the primary data for meta-analysis. Differences among the coefficients may be analyzed using techniques analogous to the standard inverse-variance weighting techniques used in contingency table analysis (1); if there is no evidence of important differences among the coefficients, one may conveniently summarize the meta-analytic results by computing a pooled (overall) coefficient estimate. The primary impact of our correction method on such meta-analyses will be to alter the relative weighting of the study-specific coefficients and to produce a more accurate variance estimate for the pooled coefficient estimate.

We recomputed the meta-analysis of alcohol use and breast cancer by Longnecker et al. (16) using both our covariance-corrected method and the uncorrected method (1). The results are given in table 3. The change in weight produced by the correction ranged from −30 percent to 10 percent. Letting k index the listed studies ($k = 1, \ldots, 16$), the fixed-effects corrected pooled

TABLE 2. Estimated regression coefficients and weights from full-data maximum likelihood estimation ($\hat{\beta}$) and from weighted least squares regression on adjusted log relative risks, with (b^*) and without (b) correction for covariance of log relative risks, for 10 data sets*

Description of study (ref.)	Method	Estimate	SE†	Weight ($1/SE^2$)	% weight is above or below MLE† weight
Arsenic exposure and lung cancer in men (5)	Full data ($\hat{\beta}$)	0.336	0.0524	364	
	Corrected (b^*)	0.311	0.0510	384	5.5
	Uncorrected (b)	0 322	0.0480	434	19.2
Alcohol consumption and colorectal cancer in men (6)	Full data ($\hat{\beta}$)	0.102	0.0373	719	
	Corrected (b^*)	0.101	0.0400	625	−13.1
	Uncorrected (b)	0.091	0.0316	1,000	39.0
Alcohol consumption and breast cancer in women (7)	Full data ($\hat{\beta}$)	0.116	0.0279	1,280	
	Corrected (b^*)	0.115	0 0275	1,320	3.1
	Uncorrected (b)	0.090	0.0222	2,030	58.6
Coffee consumption and myocardial infarction in women (8)‡	Full data ($\hat{\beta}$)	0.123	0.0814	151	
	Corrected (b^*)	0.131	0.0846	140	−7.3
	Uncorrected (b)	0.088	0.0734	186	23.2
Cigarette smoking and myocardial infarction in women (9)	Full data ($\hat{\beta}$)	1 08	0.100	100	
	Corrected (b^*)	1.06	0.103	94.3	−5.7
	Uncorrected (b)	1.09	0.098	104	4.0
Alcohol consumption and esophageal cancer in men (10)	Full data ($\hat{\beta}$)	1 09	0.103	94.3	
	Full data with random effects	1 10	0.117	73.1	
	Corrected (b^*)	1.03	0.122	67	−28.7; −8 1§
	Uncorrected (b)	1.13	0 097	106	12.4; 45.0
Cigarette smoking and lung cancer in men (11)	Full data ($\hat{\beta}$)	0.740	0.0257	1,510	
	Corrected (b^*)	0.707	0.0292	1,170	−22.5
	Uncorrected (b)	0.902	0 0246	1,650	9.3
Cigarette smoking and lung cancer in men (12)	Full data ($\hat{\beta}$)	0.472	0.0499	402	
	Corrected (b^*)	0.454	0.0598	280	−30.3
	Uncorrected (b)	0.668	0.0634	249	−38.1
Passive smoking and lung cancer in women (13)	Full data ($\hat{\beta}$)	0.311	0.109	84.2	
	Corrected (b^*)	0.309	0 109	84.2	0
	Uncorrected (b)	0.326	0.0987	103	22.3
Sunlight exposure and basal cell skin cancer (14)	Full data ($\hat{\beta}$)	0.479	0.127	62.0	
	Corrected (b^*)	0.478	0.125	64.0	3.2
	Uncorrected (b)	0.480	0.119	70.6	13.9

* All full-data regressions included age; weighted least squares regressions were on log relative risks adjusted for age, with age treated categorically in both types of analyses. Exposure levels were coded as 0, 1, 2, . . ., etc., in all analyses.
† SE, standard error, MLE, maximum likelihood estimate.
‡ In this data set, the covariate was smoking (treated categorically), not age.
§ Second set of numbers is for random-effects estimate

coefficient estimate for these data is $b_p^* = (\Sigma_k b_k^*/v_k^*)/(\Sigma_k 1/v_k^*) = 0.00823$, with estimated standard error $s_p^* = (\Sigma_k 1/v_k^*)^{-1/2} = 0.00132$; for comparison, the uncorrected pooled estimate is $b_p = (\Sigma_k b_k/v_k)/(\Sigma_k 1/v_k) = 0.00789$, with estimated standard error $s_p = (\Sigma_k 1/v_k)^{-1/2} = 0.00121$. The small difference in point estimates is unsurprising, given the high precision of the results and the fact that both estimators are consistent, but the uncorrected summary somewhat overstates the precision of the pooled results.

With any pooling technique, it is important to check for between-study heterogeneity of the estimated parameters (1). Given K studies to be pooled, the corrected heterogeneity test statistic is

$$X_{h^*}^2 = \sum_k (b_k^* - b_p^*)^2/v_k^*,$$

which has an approximate $K - 1$ df chi-squared distribution if the study-specific slopes are homogeneous and the v_k^*'s are consistent for the variances of the b_k^*'s. If

TABLE 3. Estimated regression coefficients, standard errors, and weights, corrected and uncorrected for covariance of log relative risks, for 16 studies of alcohol use and breast cancer reviewed by Longnecker et al. (16)*

Article in Longnecker et al. (16)	Corrected			Uncorrected		
	b^*	SE†	Weight ($1/SE^2$)	b	SE	Weight ($1/SE^2$)
Hiatt and Bawol, 1984 (1)‡	0.00434	0.00247	164,000	0.00385	0.00230	207,000
Hiatt et al., 1988 (2)	0.0109	0.00410	59,600	0.0122	0.00379	65,600
Willett et al., 1987 (3)	0.0284	0.00564	31,400	0.0248	0.00537	34,700
Schatzkin et al., 1987 (4)	0.118	0.0476	441	0.129	0.0457	478
Harvey et al., 1987 (5)	0.0121	0.00429	54,200	0.0137	0.00408	60,000
Rosenberg et al., 1982 (6)	0.0870	0.0232	1,860	0.0902	0.0202	2,440
Webster et al., 1983 (7)	0.00311	0.00373	71,800	0.000625	0.00333	90,000
Paganini-Hill and Ross, 1983 (8)	0.00000	0.00940	11,300	0.000000	0.00965	10,700
Byers and Funch, 1982 (9)	0.00597	0.00658	23,100	0.00810	0.00687	21,030
Rohan and McMichael, 1988 (10)	0.0479	0.0205	2,378	0.0367	0.0188	2,837
Talamini et al., 1984 (11)	0.0389	0.00768	16,900	0.0394	0.00725	19,000
O'Connell et al., 1987 (12)	0.203	0.0946	112	0.203	0.0946	112
Harris and Wynder, 1988 (13)	−0.00673	0.00419	56,900	−0.00674	0.00403	61,500
Le et al., 1984 (14)	0.0111	0.00481	43,300	0.0107	0.00418	57,300
La Vecchia et al., 1985 (15)	0.0148	0.00635	24,800	0.0146	0.00530	35,600
Begg et al., 1983 (16)	−0.000787	0.00867	13,300	0.000128	0.00794	15,900
Pooled estimate	0.00823	0.00132		0.00789	0.00121	

* Coefficients are the increase in log relative risk of breast cancer associated with average daily alcohol consumption of 1 g. O'Connell et al. (12) reported only two categories of alcohol intake; thus, the correction had no effect.
† SE, standard error.
‡ Numbers in parentheses, Longnecker et al.'s (16) reference no.

the full-data coefficient $\hat{\beta}_k$ and its variance estimate $\hat{v}_k$ are available for study k, these may be substituted for b_k^* and v_k^* in the formulas for b_p^*, v_p^*, and $X_{h^*}^2$.

Because the uncorrected variances tend to underestimate the variances of the uncorrected estimators, the uncorrected heterogeneity statistic

$$X_h^2 = \sum_k (b_k - b_p)^2/v_k$$

will tend to be inflated above its nominal $K - 1$ df chi-squared distribution, and so it will produce an invalid (supranominal) heterogeneity test. For the data in table 3, however, both statistics are so large ($X_{h^*}^2 = 75.3$ and $X_h^2 = 87.2$ on $16 - 1 = 15$ df) that the homogeneity hypothesis is untenable. Thus, in this example, the pooled slope estimates are inappropriate summaries of the studies, and further heterogeneity analysis (such as "meta-regression" (1)) is needed.

ANALYSIS OF NONLINEAR TRENDS IN POOLED DATA

The methods discussed so far are useful when one's goal is to pool slope estimates from several reports (1). A more flexible method for meta-analysis of trend involves pooling of study data *before* trend analysis. We will refer to this as the "pool-first" method. Let $\mathbf{x}_k$ and $\mathbf{L}_k$ be the vectors of nonzero exposure levels and log odds ratios or log rate ratios observed in study k; let C_k be the estimated covariance matrix for $\mathbf{L}_k$; let $\mathbf{x} = (\mathbf{x}_1', \ldots, \mathbf{x}_k')'$ and $\mathbf{L} = (\mathbf{L}_1', \ldots, \mathbf{L}_k')'$; and let G be the block-diagonal matrix with k'th diagonal block C_k^{-1}. A pooled estimate $\tilde{\beta}$ of the common slope β is given by $\tilde{v}\mathbf{x}'G\mathbf{L}$, with variance estimate $\tilde{v} = (\mathbf{x}'G\mathbf{x})^{-1}$; assuming each C_k is a consistent estimator of $\text{cov}(\mathbf{L}_k)$, and the slope is in fact constant across studies, $\tilde{v}$ will be consistent for $\text{var}(\tilde{\beta})$.

For linear-logistic estimation, the "pool-first" method is algebraically equivalent to the method of pooling the corrected coefficient estimates from each study. The advantage of the "pool-first" method is that it is easily extended to fitting and testing nonlinear logistic models. For example, suppose we wish to estimate β_1 and β_2 in the quadratic logit model

$$\lambda(x, z) = \alpha_k + \beta_1 x + \beta_2 x^2 + \boldsymbol{\delta}_k' \mathbf{z}_k.$$

To do so, we let X be the matrix with the first column equal to **x** and the second column equal to the vector with elements that are the square of the corresponding elements of **x**. A pooled estimate of $\boldsymbol{\beta} = (\beta_1, \beta_2)'$ is $\tilde{\boldsymbol{\beta}} = VX'G\mathbf{L}$, with covariance-matrix estimate $V = (X'GX)^{-1}$, and a chi-squared statistic for model fit is $\mathbf{e}'G\mathbf{e}$, where **e** is the residual vector $\mathbf{L} - X\tilde{\boldsymbol{\beta}}$. The degrees-of-freedom is equal to the length of **e** minus 2. The chief limitation of this method is that it cannot incorporate studies that report only a slope estimate: A study must report dose-specific odds ratios or rate ratios to be included; fortunately, such reporting is standard practice.

For illustration, we applied the preceding method to the studies reported in table 3 and obtained $\tilde{\beta}_1 = 0.00934$ for the linear term and $\tilde{\beta}_2 = -0.0000258$ for the quadratic term, with standard errors of 0.00229 and 0.0000429, respectively. The goodness-of-fit statistic is 99.9 on $49 - 2 = 47$ df, very significant. The results thus indicate that the pooled quadratic effect is small compared with the pooled linear effect (at least within the range of alcohol use reported by most women in these studies), and that a quadratic term explains little of the heterogeneity of trend across studies. As was demonstrated by the large value of X_{h*}^2 given above, nonsignificance of the quadratic term does *not* imply that the homogeneous linear model is adequate.

DISCUSSION

The methods given here are readily modified to allow more general model forms than logistic or exponential. We have not pursued this generalization, however, because empirical studies indicate that the asymptotic theory used here (17) may be unreliable as a practical guide for models with parameters that are not linear in the logit or log scales; see the paper by Moolgavkar and Venzon (18) for some striking examples and further references.

Because the corrected estimates involve somewhat more computation than the uncorrected estimates, it seems natural to ask under what conditions the correction will be worth the effort. From the structure of the correlation formulas, it appears that the impact of the correction on individual study weights depends in part on the percentage of subjects who are in the reference category of exposure. Nevertheless, knowledge of the proportion of subjects in the reference group does not reliably identify studies for which the correction will make an important difference.

Because the relative weighting of the studies will not change as dramatically as the absolute weighting, we would not expect a large impact of the correction on overall pooled estimates of effect. Nevertheless, the correction could have substantial impact on heterogeneity analyses, especially when apparent "outlier" studies are based on limited numbers in the reference category of exposure.

We wish to emphasize that the correction we have discussed here is concerned only with improving the statistical properties of the slope estimators. It cannot compensate for biases in the pooled studies, publication bias in identification of studies, noncomparability of exposure or outcome measurements across studies, or any of the other problems that should be addressed in a careful meta-analysis.

REFERENCES

1. Greenland S. Quantitative methods in the review of epidemiologic literature. Epidemiol Rev 1987;9:1–30.
2. Rohan TE, McMichael JA. Alcohol consumption and risk of breast cancer. Int J Cancer 1988;41:695–9.
3. Whittemore AS. Collapsibility of multidimensional contingency tables. J R Stat Soc [B] 1978;40:328–40.
4. Thomas DC, Greenland S. The relative efficiencies of matched and independent sample designs for case-control studies. J Chronic Dis 1983;36:685–97.
5. Breslow NE, Day NE, eds. Statistical methods in cancer research. Vol 2. The design and analysis of cohort studies. Appendix V. Lyon, France: Inter-

national Agency for Research on Cancer, 1987. (IARC scientific publication no. 82).
6. Longnecker MP. A case-control study of alcohol consumption in relation to risk of cancer of the right colon and rectum in men. Cancer Causes and Control 1990;1:5–14.
7. Willett WC, Stampfer MJ, Colditz GA, et al. Moderate alcohol consumption and the risk of breast cancer. N Engl J Med 1987;316:1174–80.
8. Rosenberg L, Slone D, Shapiro S, et al. Coffee drinking and myocardial infarction in young women. Am J Epidemiol 1980;111:675–81.
9. Shapiro S, Slone D, Rosenberg L, et al. Oral-contraceptive use in relation to myocardial infarction. Lancet 1979;1:743–7.
10. Breslow NE, Day NE, eds. Statistical methods in cancer research. Vol 1. The analysis of case-control studies. Lyon, France: International Agency for Research on Cancer, 1980:151. (IARC scientific publication no. 32).
11. Kahn HA. The Dorn study of smoking and mortality among U.S. veterans: report on eight and one-half years of observation. In: Haenszel W, ed. Epidemiological approaches to the study of cancer and other chronic diseases. Bethesda, MD: National Cancer Institute, 1966. (NCI monograph no. 19).
12. Frome EL. The analysis of rates using Poisson-regression models. Biometrics 1983;39:665–74.
13. Hirayama T. Non-smoking wives of heavy smokers have a higher risk of lung cancer: a study from Japan. BMJ 1981;282:183–5.
14. Vitaliano PP. The use of logistic regression for modelling risk factors: applications to non-melanoma skin cancer. Am J Epidemiol 1978; 108:402–14.
15. McCullagh P, Nelder JA. Generalized linear models. 2nd ed. New York: Chapman and Hall Ltd, 1989.
16. Longnecker MP, Berlin JA, Orza MJ, et al. A meta-analysis of alcohol consumption in relation to risk of breast cancer. JAMA 1988;260:652–6.
17. Bishop YMM, Feinberg SE, Holland PW. Discrete multivariate analysis: theory and practice. Cambridge, MA: The MIT Press, 1975.
18. Moolgavkar SH, Venzon DJ. General relative risk regression models for epidemiologic studies. Am J Epidemiol 1987;126:949–61.
19. Seber GAF, Wild CJ. Nonlinear regression. New York: John Wiley and Sons, Inc, Publishers, 1989.

APPENDIX 1

Inefficiency of the Uncorrected Point Estimator and Inconsistency of the Uncorrected Variance Estimator

Let n be the total sample size. The uncorrected estimator b may be written

$$b = (\mathbf{x}'W^*\mathbf{x})^{-1}\mathbf{x}'W^*\mathbf{L}$$
$$= \sum_x w_x x L_x / s,$$

where W is the diagonal matrix with diagonal elements $w_x = 1/v_x$ and $s = \Sigma w_x x^2$; the uncorrected variance estimator for b obtained from a weighted least squares regression program (after division by the computed residual mean square) will be $1/s$. The asymptotic variance of $\sqrt{n}(b - \beta)$ is, however, consistently estimated by

$$n\mathbf{x}'WC_aW\mathbf{x}/s^2 = n/s + n\mathbf{x}'WC_0W\mathbf{x}/s^2,$$
$$= n/s + n\sum_{j\neq k} x_j w_j c_{ajk} w_k x_k / s^2, \quad \text{(A1)}$$

where $C_a = [c_{ajk}]$ is the covariance-matrix estimator for $\mathbf{L}$ from the complete data and $C_0 = C_a - W^{-1}$. Since the second term of expression A1 is positive, n/s must underestimate the asymptotic variance of $\sqrt{n}(b - \beta)$ by an amount proportional to the covariances of the L_x's.

An efficient estimator for β is the complete-data estimator

$$(\mathbf{x}'C_a^{-1}\mathbf{x})^{-1}\mathbf{x}'C_a^{-1}\mathbf{L} = \sum u_x L_x.$$

The weights $w_x x/s$ used for b are generally not proportional to the optimal weights u_x unless the covariances are zero; hence, b is inefficient.

APPENDIX 2

Iterative Fitting Algorithm for the Crude Table

The algorithm is based on Newton's method (19) for solving the following system for **A**, the vector of fitted numbers of cases at each nonzero exposure level. We have an equation for each observed exposure level,

$$L_x + \log(M_1 - A_+) + \log(N_x - A_x) - \log A_x - \log(N_0 - M_1 + A_+) = 0,$$

where A_+ is the sum of the elements of **A** (note that A_0 is not in **A**, since $A_0 = M_1 - A_+$). An initial value $\mathbf{A}^{(0)}$ may be the crude observed totals, if available, or the null expected value $M_1\mathbf{N}/n$, where **N** is the vector of N_x for $x \neq 0$ and n is the total number of subjects in the data. The algorithm may diverge from poor starting values; in our experience, convergence was always achieved by starting with the crude observed totals rather than the null expected values.

At iteration i, define

$A_0^{(i)} = M_1 - A_+^{(i)}$,

$c_x^{(i)} = 1/A_x^{(i)} + 1/(N - A_x^{(i)})$ for all x (including $x = 0$),

$e_x^{(i)} = L_x + \log A_0^{(i)} + \log(N_x - A_x^{(i)}) - \log A_x^{(i)} - \log(N_0 - A_0^{(i)})$ for $x \neq 0$,

$\mathbf{e}^{(i)}$ = the vector of $e_x^{(i)}$,

$H^{(i)}$ = the matrix with $c_x^{(i)} + c_0^{(i)}$ for on-diagonal elements and $c_0^{(i)}$ for all off-diagonal elements, and

$\mathbf{A}^{(i+1)} = \mathbf{A}^{(i)} + (H^{(i)})^{-1}\mathbf{e}^{(i)}$.

Convergence is achieved when the increments become negligible relative to the $A_x^{(i)}$ and $N - A_x^{(i)}$ for all x. For person-time data, the equations become

$$L_x + \log(M_1 - A_+) + \log N_x - \log A_x - \log N_0 = 0;$$

the expression for $e_x^{(i)}$ is similarly modified; and $c_x^{(i)}$ becomes $1/A_x^{(i)}$.

APPENDIX 3

Properties of the Corrected Variance Estimator

The asymptotic variance of $\sqrt{n}(b^* - \beta)$ is consistently estimated by

$$n\mathbf{x}'C^{-1}C_aC^{-1}\mathbf{x}(v_b^*)^2, \qquad \text{(A2)}$$

where $C = \widetilde{\text{cov}}(\mathbf{L})$ is as defined in step 5 in the text, and C_a is the (unobserved) covariance-matrix estimator for **L** from the complete data. Note that C for a single study may be written $C = W^{-1}RW^{-1}$, where W^{-1} is the diagonal matrix with the variance estimators of the adjusted log odds ratios on the diagonal, and R is the estimated correlation matrix of the crude log odds ratios derived under assumptions 1–3 given in the text using the delta method (17) applied to the crude cross-classification of exposure and outcome. Under assumptions 1–3 in the text, nC converges to nC_a, and so nv_b^* converges to expression A2; hence, nv_b^* is consistent for $\text{var}^A[\sqrt{n}(b^* - \beta)]$ under assumptions 1–3. The assumptions also imply that nv_b^* converges to

$$n(\mathbf{x}'C_a^{-1}\mathbf{x})^{-1},$$

which in turn converges to the asymptotic variance of the maximum likelihood estimator based on the full data; hence, under assumptions 1–3, b^* will be more efficient than the uncorrected estimator b.