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 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 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; N = (337, 167, 186, 212)′; M_1 = 451; L = (log 0.80, log 1.16, log 1.57)′ = (−0.223, 0.148, 0.451)′; and 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) 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

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

To do so, we let X be the matrix with the first column equal to $\mathbf{x}$ and the second column equal to the vector with elements that are the square of the corresponding elements of $\mathbf{x}$. A pooled estimate of $\beta = (\beta_1, \beta_2)'$ is $\hat{\beta} = VX'G\mathbf{L}$, with covariance-matrix estimate $V = (X'GX)^{-1}$, and a chi-squared statistic for model fit is $\mathbf{e}'G\mathbf{e}$, where $\mathbf{e}$ is the residual vector $\mathbf{L} - X\hat{\beta}$. The degrees-of-freedom is equal to the length of $\mathbf{e}$ 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

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

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.

STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:954–970
Published online 3 August 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3013

Facilitating meta-analyses by deriving relative effect and precision estimates for alternative comparisons from a set of estimates presented by exposure level or disease category

Jan Hamling[1,*,†], Peter Lee[1], Rolf Weitkunat[2] and Mathias Ambühl[3]

[1]*P.N. Lee Statistics and Computing Ltd, Sutton, Surrey SM2 5DA, U.K.*
[2]*Philip Morris Products SA, Research & Development, Neuchâtel, Switzerland*
[3]*Consult AG, Berne, Switzerland*

SUMMARY

Epidemiological studies relating a particular exposure to a specified disease may present their results in a variety of ways. Often, results are presented as estimated odds ratios (or relative risks) and confidence intervals (CIs) for a number of categories of exposure, for example, by duration or level of exposure, compared with a single reference category, often the unexposed. For systematic literature review, and particularly meta-analysis, estimates for an alternative comparison of the categories, such as any exposure *versus* none, may be required. Obtaining these alternative comparisons is not straightforward, as the initial set of estimates is correlated. This paper describes a method for estimating these alternative comparisons based on the ideas originally put forward by Greenland and Longnecker, and provides implementations of the method, developed using Microsoft Excel and SAS. Examples of the method based on studies of smoking and cancer are given. The method also deals with results given by categories of disease (such as histological types of a cancer). The method allows the use of a more consistent comparison when summarizing published evidence, thus potentially improving the reliability of a meta-analysis. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: systematic review; meta-analysis; contrast; dose response

INTRODUCTION

In a case–control study of breast cancer risk in young women by Smith *et al.* [1], odds ratios, adjusted for age and other covariates, were presented for passive smoking exposure among

*Correspondence to: Jan Hamling, 17 Cedar Road, Sutton, Surrey SM2 5DA, U.K.
†E-mail: JanHamling@pnlee.co.uk

Contract/grant sponsor: Philip Morris International

Received 12 February 2007
Accepted 19 June 2007

lifelong non-smokers. Compared to women with no reported lifetime exposure, odds ratios with 95 per cent (CIs) were given as 2.82 (1.00–7.93) for 1–200 cigarette-years and as 2.24 (0.75–6.68) for >200 cigarette-years. Recently, in their 'Proposed Identification of Environmental Tobacco Smoke as a Toxic Air Contaminant', the California EPA [2] included a long section on passive smoking and breast cancer. A table in that paper (7.4.1B) included summary estimates for overall exposure from 19 studies, one of which was an estimate from the Smith *et al.* [1] study of 2.53 (1.12–5.71). This estimate was stated to be 'calculated by summarizing the adjusted lifetime exposure categories', although no further details were given on how the calculation was done.

We readily found that the combined estimate could be obtained precisely by conducting a simple fixed-effects meta-analysis [3] on the two individual estimates. However, this approach assumes that the estimates for 1–200 and >200 cigarette-years are independent, which is clearly not the situation as both estimates involve the same reference group of zero cigarette-years. Indeed, as the reference group included far fewer cases (10 cases) than the two exposed groups (46 and 38 cases, respectively), the erroneous calculation might have substantially underestimated the width of the CI for the combined estimate.

Because researchers often wish to derive alternative comparisons for data presented in categories relative to a common reference group, Easton *et al.* [4] proposed an alternative method for the presentation of results, using 'floating absolute risks' (FARs) and their CIs, which allows such alternative comparisons to be estimated easily and validly. Greenland *et al.* [5] discussed the FAR method, admitting that it 'can supply useful statistics and trend graphs', but arguing that 'it does not yield valid confidence intervals for relative risks'. In reply, Easton and Peto [6] pointed out that the FAR CIs were never intended actually to be CIs for relative risks, but were only intended to facilitate their calculation by adding the floating variances of the log relative risks for the two categories being compared. Easton and Peto [6] also noted that an alternative approach suggested by Greenland *et al.* [5] in fact gave results identical to their approach. Whatever the merits of the FAR method, very few studies have ever reported results in this manner; hence, the problem of obtaining valid estimates from data presented as odds ratios (for case–control studies), or relative risks (for cohort studies), by categories of exposure remains.

In 1992, Greenland and Longnecker [7] described a method to solve a related problem. Given the numbers of cases and controls and covariate-adjusted odds ratios and CIs by the level of exposure, but in the absence of data on the covariances of the adjusted log odds ratios, they wished to estimate the increase in log odds per unit of exposure taking appropriate account of the non-independence of the odds ratios. Their method starts by using the odds ratios and the marginal totals over exposure to derive a corresponding set of pseudo-numbers (or 'effective' numbers) of cases and controls consistent with the input data. These numbers (which have no direct meaning by themselves), together with the CIs of the adjusted odds ratios, could then be used to estimate the required covariances, and hence the unit increase in log odds and its CI. Greenland and Longnecker [7] showed that their approach provided more efficient estimates of the combined odds ratio and CI than other methods previously available and also described how their method could be extended to cohort studies. Practical examples of the method were presented in papers published in 1993 [8], and, much more recently [9], the latter paper also providing a command, glst, written for Stata 9.1, for implementing the method.

Some years ago, one of us (J.H.) developed a program, using Microsoft Excel, to carry out an analogous but somewhat different method based on Greenland and Longnecker's [7]

Statist. Med. 2008; **27**:954–970
DOI: 10.1002/sim

effective numbers approach. In our method, we generate a set of numbers consistent with both the adjusted odds ratio (or relative risk) and its CI, which can then be used to make any comparison required including a dose-related trend. This method has proved invaluable to us when conducting a variety of meta-analyses. In the context of the Smith *et al.* [1] results, our method gives a combined estimate of 2.58 (95 per cent CI 0.96–6.94) rather than the estimate of 2.53 (1.12–5.71) given by the California EPA [2]. Our estimate, which we believe to be more appropriate, shows the observed association to have a p-value of 0.060 rather than 0.025, with the associated meta-analysis weight (inverse-variance of log odds ratio) lower, at 3.93 compared with 5.79.

In the past, we had provided only a brief description of the method, as an appendix to a paper on lung cancer and passive smoking [10]. The objective of this paper is to clarify the details of the method, and to make it readily available to researchers both as an Excel spreadsheet and as a SAS macro.

The method (and software) also takes into account an alternative situation, where individual odds ratios (or relative risks) are presented by diagnostic category (e.g. histological subtype of lung cancer) with a common control group and where estimates are required for combined categories (e.g. all lung cancer). The method is illustrated by worked examples. It should be noted that the accuracy of the combined estimate is limited by the accuracy to which the values are quoted in the study report. While the method works well in practice with results presented to the usual two decimal places, some journals present odds ratios and relative risks to only one decimal place. While the method usually works well here too, we have seen data presented from very large studies where the lower and upper 95 per cent CIs are the same to one decimal place. Here, the method would infer pseudo-numbers that were infinite and hence it would fail. However, provided the source data are presented as non-overlapping and exhaustive categories with a common reference group and are given to sufficient accuracy, we have found the method described below to be widely applicable.

METHOD

The method will be described initially for a case–control study giving results for several categories of exposure. The extension of the method to prospective studies and to studies giving results for categories of disease rather than for categories of exposure will then be described. The method described in this paper has been implemented both in an Excel spreadsheet and in a SAS macro. Both implementations and their accompanying documentation are available for downloading from the web page www.pnlee.co.uk/software.htm. These implementations are summarized in Appendix A (Excel) and Appendix B (SAS).

Case–control studies giving results by categories of exposure

Suppose, in a case–control study, the subjects are divided into $n+1$ groups—an unexposed group ($i=0$) and n exposed groups ($i=1,\ldots,n$)—and estimates are available (for each exposed group) of the odds ratio compared with the unexposed group (R_i) and its lower and upper 95 per cent confidence limits (L_i to U_i).

Statist. Med. 2008; **27**:954–970
DOI: 10.1002/sim

The published study odds ratios and CIs are, therefore,

Exposure category	Odds ratio (95 per cent CI)
Unexposed: 0	1
1	R_1 (L_1–U_1)
2	R_2 (L_2–U_2)
⋮	⋮
n	R_n (L_n–U_n)

Corresponding to this is an underlying, but unknown, distribution of numbers of subjects:

Exposure category	Cases	Controls
Unexposed: 0	A_0	B_0
1	A_1	B_1
2	A_2	B_2
⋮	⋮	⋮
n	A_n	B_n
Total	A	B

This can be regarded as n 2×2 tables of the form:

	Cases	Controls
Unexposed	A_0	B_0
Exposed	A_i	B_i

For each of these, the odds ratio satisfies the equation:

$$R_i = \frac{A_i B_0}{A_0 B_i} \tag{1}$$

The variance of the log odds ratio $\log_e(R_i)$ is approximated by

$$V_i = 1/A_0 + 1/B_0 + 1/A_i + 1/B_i \tag{2}$$

and the 95 per cent CI of the log odds ratio ($\log_e(U_i)$ to $\log_e(L_i)$) is given by

$$\log_e(R_i) \pm 1.96\sqrt{V_i} \tag{3}$$

The CI for the odds ratio is calculated by exponentiating these values [11]. For alternative CIs, 1.96 can be replaced by the appropriate normal deviate $z_{(1-\alpha/2)}$. For example, 1.645 and 2.58 correspond, respectively, to α levels of 10 and 1 per cent, or 90 and 99 per cent CIs.

For various purposes, it may be necessary to estimate the odds ratios for alternative comparisons, e.g. all the exposed subjects combined *versus* the unexposed subjects, one exposure *versus* another or high exposure *versus* low exposure. The approach used is similar to that of Greenland and

Statist. Med. 2008; **27**:954–970
DOI: 10.1002/sim

Longnecker [7] in that one first reconstructs the underlying table of numbers—in this instance, of cases and controls in each exposure category—and then derives odds ratios and CIs for the required comparison, simply by grouping together the relevant exposure levels into a 2×2 table of cases and controls by exposure and applying equations (1)–(3).

In order to estimate the $2(n+1)$ numbers A_i, B_i $(i = 0, \ldots, n)$, $2(n+1)$ equations are required. The odds ratios and CIs for the exposed categories provide $2n$ equations. For a solution to be obtained, two further pieces of data that are generally reported for epidemiological studies are used—p, the proportion of unexposed subjects among the total number of controls $(B_0/\sum_{i=0}^{n} B_i)$, and z, the relative frequency of controls to cases overall $(\sum_{i=0}^{n} B_i/\sum_{i=0}^{n} A_i)$. (The rationale behind the selection of these two specific data items is discussed later in this paper.)

The $2(n+1)$ equations can now be written down. A preliminary step is to obtain the variance of the $\log_e(R_i)$ estimate for each exposure level. Reorganizing equation (3) gives

$$V_i = \left\{ \frac{\log_e(U_i/L_i)}{3.92} \right\}^2 \tag{4}$$

The $2(n+1)$ equations can then be written as

$$p = B_0/B \tag{5}$$

$$z = B/A \tag{6}$$

$$R_i = A_i B_0 / A_0 B_i \quad (i = 1, \ldots, n) \tag{7}$$

$$V_i = 1/A_0 + 1/B_0 + 1/A_i + 1/B_i \quad (i = 1, \ldots, n) \tag{8}$$

where p, z, and R_i are given, the V_i have been calculated using (4); and

$$B = \sum_{i=0}^{n} B_i$$

$$A = \sum_{i=0}^{n} A_i$$

These can be solved iteratively, as described in the Appendices.

As an example, we consider again the study by Smith *et al.* [1], but here we consider active rather than passive smoking (because, for active smoking, the paper reports results both by categories of exposure and overall). Table I of that paper gives results of analyses relating the total amount smoked (cigarette-years) to breast cancer—those results are reproduced here as Table I.

In order to assess the evidence relating smoking to breast cancer, it would be useful to have a single odds ratio and CI for 'Ever smoked' (1 or more cigarette-years) against 'Never smoked'. This involves combining the results given for 1–200 cigarette-years smoking with those for >200 cigarette-years smoking. A meta-analysis of the pair of results presented would be invalid because the results are not independent—they share a common comparison group.

No completely unadjusted analysis is given (that labelled as unadjusted in Table I of Smith *et al.* [1] actually being matched for age and general practitioner), but we can use the numbers of cases and controls to calculate these (see Table II).

DOI: 10.1002/sim

Table I. Odds ratios of breast cancer by amount smoked, taken from Smith *et al.* [1].

Cigarette-years	Number of subjects		Odds ratio (95 per cent CI) Adjusted
	Cases	Controls	
Never smoked	348	355	1
1–200	236	239	1.00 (0.78–1.29)
>200	167	157	1.02 (0.76–1.37)

Table II. Unadjusted odds ratios of breast cancer by amount smoked, calculated from the numbers of subjects given by Smith *et al.* [1].

Cigarette-years	Number of subjects		Odds ratio (95 per cent CI) Unadjusted, calculated
	Cases	Controls	
0	348	355	1
1–200	236	239	1.00731 (0.79808–1.27140)
>200	167	157	1.08509 (0.83389–1.41195)

Table III. Overall unadjusted odds ratio of breast cancer, calculated from the numbers of subjects given by Smith *et al.* [1].

Cigarette-years	Number of subjects		Odds ratio (95 per cent CI) Unadjusted, calculated
	Cases	Controls	
0	348	355	1
>0	403	396	1.03815 (0.84766–1.27145)

Using these values for odds ratios and CIs (to five or more decimal places) and assuming that the numbers of subjects are unknown, the Excel solving method described in Appendix A generated an estimated table of numbers of subjects (not shown) with a maximum absolute inaccuracy of less than 0.02. As the odds ratios and CIs were input to fewer decimal places, the maximum inaccuracy increased to 0.05 for four decimal places, to 0.30 for three, and to 5.54 (an error of 1.6 per cent) for two.

The numbers of cases and controls from Table II can be combined into a single 'Exposed' group and used to calculate an overall unadjusted odds ratio (95 per cent CI) (see Table III).

Using the numbers of subjects estimated by the Excel method gave an estimated odds ratio (95 per cent CI) of 1.03815 (0.84766–1.27144), accurate to four decimal places.

Published study reports do not give results to this level of detail. Reducing the odds ratio (CI) values entered in the Excel spreadsheet to two decimal places resulted in estimated numbers of subjects as shown in Table IV and an estimated odds ratio (95 per cent CI) of 1.04 (0.85–1.27), which, to two decimal places, is the same as the value calculated above (using the actual numbers of cases and controls).

DOI: 10.1002/sim

Table IV. Numbers of subjects estimated using the Excel spreadsheet when the input values are reduced to two decimal places (Smith *et al.* [1] data).

	Number of subjects	
Cigarette-years	Cases	Controls
0	349.221	356.751
1–200	241.540	244.304
>200	163.943	153.650

Table V. Effective numbers of subjects (representing the 'adjusted' population), estimated using the Excel spreadsheet (Smith *et al.* [1] data).

	Actual numbers		Estimated effective numbers from adjusted results	
	Cases	Controls	Cases	Controls
Never smoked	348	355	295.811	296.990
1–200 cigarette-years	236	239	205.264	206.082
200+ cigarette-years	167	157	127.206	125.209

Other examples give odds ratio (CI) values that differ in the second decimal place. This degree of inaccuracy would typically have no appreciable effect on a meta-analysis.

The preceding analysis is unrealistic, in that the unadjusted results are generally of little interest. For many associations, there are a number of established confounding factors that should be taken into account in the risk estimates quoted in a systematic review. In order to handle adjusted results, we follow Greenland and Longnecker [7] in supposing that a table of pseudo-numbers of subjects can be estimated that represents an 'adjusted' population—the numbers of subjects effectively used when an adjusted analysis is carried out. The process described above can then be carried out in exactly the same way as before, but using the adjusted odds ratios and CIs provided in the report of the study.

As an example, the adjusted odds ratios and CIs for total amount smoked (never smoked, 1–200 cigarette-years, 200+ cigarette-years, to two decimal places) from Table I of the paper by Smith *et al.* [1] were entered in the Excel version of the method. The estimated effective numbers of subjects (see Table V) were rather lower than the actual numbers of subjects, as would be expected since adjustment usually increases the variance of an estimate [12].

The adjusted odds ratio (CI) for ever smoked *versus* never smoked, which we estimated as 1.0076 (0.8074–1.2574) from this table, can be compared with the adjusted result actually given in Table I of the paper of 1.01 (0.81–1.26), which is the same to two decimal places.

In the example above, there is very little variation in risk by level of exposure, and it is unsurprising that the method comes up with an apparently appropriate answer. As an example with more variation in risk, we consider data from the lung cancer case–control study of Matos *et al.* [13]. Odds ratios for current smoking *versus* lifelong non-smoking were reported (in Table III of that paper) overall and by various aspects of the smoking habit, all adjusted for the same list of covariates. The relevant data by age at the start of smoking are given in Table VI.

DOI: 10.1002/sim

Table VI. Odds ratios of lung cancer by age at starting to smoke, taken from Matos *et al.* [13].

	Number of subjects		Odds ratio (95 per cent CI)
Age at start	Cases	Controls	Adjusted
Non-smoker	11	110	1
<15	45	41	11.3 (5.3–24.3)
15–19	49	58	8.6 (4.1–18.2)
20+	18	33	5.3 (2.3–12.5)

Table VII. Odds ratios of lung cancer by the number of cigarettes smoked per day, taken from Matos *et al.* [13].

	Number of subjects		Odds ratio (95 per cent CI)
Cigarettes/day	Cases	Controls	Adjusted
Non-smoker	11	110	1
1–14	5	32	1.6 (0.5–5.0)
15–24	42	54	8.0 (3.4–16.8)
25+	65	46	15.0 (7.1–31.9)

From these data, we estimated the combined odds ratio for current smoking as 8.54 (4.32–16.87) using the Excel method. This compares seemingly well with the values published by Matos *et al.* [13] of 8.5 (4.3–16.7), given that the odds ratios and CIs were shown to only one decimal place.

Interestingly, basing the calculation on the data by number of cigarettes/day suggested a possible error in the source paper (see Table VII). Here, the Excel method gave a combined estimate of 9.06 (4.48–18.34), which is not so close to the 8.5 (4.3–16.7) given by Matos *et al.* [13]. This may be because the odds ratio for 15–24 cigarettes/day is some distance away from the centre of the 95 per cent CI on a log scale (the square root of 3.4×16.8 being 7.56 and not 8.0) and suggests a possible typographical error.

Case–control studies giving results by categories of disease

The odds ratio and variance definitions given as (1) and (2) above can also be used for the 2×2 table below:

	Exposed	Unexposed
Controls	E_0	U_0
Cases	E_i	U_i

which allows for a number of distinct categories of disease, such as different histological types of a cancer, rather than categories of exposure, and uses a common control group. The method described above for results by exposure is equally applicable to results by disease, with E_0, U_0, E_i, U_i corresponding, respectively, to A_0, B_0, A_i, B_i. Here, p is the proportion of controls among

DOI: 10.1002/sim

Table VIII. Odds ratios of environmental tobacco smoke by the type of lung cancer, taken from Fontham *et al.* [14].

Subjects	Number of subjects		Odds ratio (95 per cent CI) Adjusted
	Exposed	Unexposed	
Controls	158	1095	1
Adenocarcinoma	62	426	1.04 (0.75–1.46)
Other histological types	24	128	1.79 (1.08–2.95)

the unexposed, and z is the ratio of unexposed to exposed, overall. Both p and z can be calculated for any study that reports the numbers of subjects studied.

As an example of this situation, data were taken from the Fontham *et al.* [14] study of environmental tobacco smoke exposure and lung cancer in non-smoking women. Data from Table II of the source paper relating to pipe smoking by the spouse are reproduced in Table VIII.

Here, the odds ratio for all lung cancer types estimated by the method is 1.178 (0.872–1.590), quite similar to the value of 1.19 (0.88–1.60) given in the source paper.

Prospective (cohort) studies giving results by categories of exposure

Consider a prospective study with B_0 unexposed subjects and B_i subjects exposed at level i $(i = 1, \ldots, n)$, of whom A_0 and A_i subjects, respectively, develop the disease being studied. This gives the 2×2 table:

	Diseased	At risk
Unexposed	A_0	B_0
Exposed	A_i	B_i

for a study in which subjects are analysed by categories of exposure. The unexposed population is common to all comparisons. Each individual comparison represents a study of a specific exposure.

Katz *et al.* [15] recommend a method of obtaining a CI for cohort study data (their Method C) in which the log relative risk $\log_e(R_i)$ is taken to be approximately normally distributed with approximate mean:

$$\log_e(R_i) = \log_e\left(\frac{A_i B_0}{A_0 B_i}\right) \tag{9}$$

The variance is estimated as

$$V_i = 1/A_0 - 1/B_0 + 1/A_i - 1/B_i \tag{10}$$

and approximate 95 per cent CIs for the log relative risk are

$$\log_e(R_i) \pm 1.96\sqrt{V_i} \tag{11}$$

Note that these are identical to equations (1)–(3), except for the negative signs in the expression for variance. Therefore, the method presented above for case–control studies is applicable to

DOI: 10.1002/sim

prospective studies as long as relative risks (not odds ratios) are available, the 2×2 table is appropriately defined (as shown above) and the calculation of the variance is modified. The value p is now defined as the proportion of unexposed subjects among those at risk, and z is the ratio of number at risk to the total number of diseased subjects.

Note that, in the above, the relative risk ($A_i B_0 / A_0 B_i$) is estimated by the risk ratio. The method also provides a good approximate solution for rate ratios where 'at risk' is replaced by 'person-years at risk', as the terms in $1/B_0$ and $1/B_i$ generally contribute virtually nothing towards the estimated variance.

Prospective (cohort) studies giving results by categories of disease

Consider a prospective study with U_0 unexposed subjects and E_0 exposed subjects, of whom U_i and E_i subjects, respectively, develop disease category i ($i = 1, \ldots, n$). The 2×2 table becomes

	Exposed	Unexposed
At risk	E_0	U_0
Diseased	E_i	U_i

The at-risk population is common to all comparisons, while each comparison represents an analysis using a distinct definition for the disease of interest. The Katz *et al.* [15] method is therefore applicable to each comparison. Here, p is the proportion of unexposed at-risk subjects among the sum of the unexposed at-risk and the unexposed diseased subjects, and z is the ratio of the sum of the unexposed at-risk and unexposed diseased subjects to the sum of the exposed at risk and exposed diseased subjects.

Use of p and z with adjusted data

Here, we return, for simplicity, to the situation of case–control studies with results given by categories of exposure. When the method is applied to data that are unadjusted for covariates, it seems that, provided the odds ratios and CIs are given to sufficient accuracy, and provided two additional independent pieces of data are available to allow the $2(n + 1)$ equations to be solved, the actual table of numbers of cases and controls by exposure can be estimated correctly. There is no specific reason to select p (the proportion of unexposed subjects among the total number of controls) and z (the relative frequency of controls to cases overall) as the additional data items. One could equally well, for example, derive the correct table of numbers from the odds ratios, the CIs, the total number of cases, and the total number of controls.

When the method is applied to adjusted data, the situation is rather different. One does not actually have any further precise information about the pseudo-numbers other than the odds ratios and CIs. One may know p or z for the unadjusted numbers, but one cannot infer that these values apply to the pseudo-numbers corresponding to the adjusted odds ratios and CIs. One could, for example, imagine a situation where the disease only occurs in subjects with level A of a confounding variable, and that level A is very common in those exposed to the agent of interest. In that situation, p based on the unadjusted data may substantially exceed the appropriate p for the adjusted analysis (with those with levels of the confounding variable other than A not contributing to the adjusted analysis at all).

Statist. Med. 2008; **27**:954–970
DOI: 10.1002/sim

Table IX. Sensitivity analysis showing the effect of varying the values of p and z on the estimated odds ratios and CIs for current smoking (based on data shown in Table VI—Matos *et al.* [13]).

p	z	Odds ratio (95 per cent CI)
0.3	10	8.696 (4.202–17.998)
	1.967	8.739 (4.266–17.901)
	0.5	8.760 (4.390–17.480)
0.4545	10	8.463 (4.194–17.077)
	1.967	8.542 (4.324–16.875)
	0.5	8.621 (4.549–16.337)
0.6	10	8.292 (4.265–16.119)
	1.967	8.432 (4.481–15.867)
	0.5	8.560 (4.774–15.349)

In practice, p and z were selected as the additional items of information to be assumed known for a number of reasons. First, the values were usually readily available from papers presenting results from epidemiological studies. Second, it was clear that constraining total numbers of cases or controls or exposed or unexposed subjects to be the same for the adjusted data as for the unadjusted data is inappropriate, as adjustment tends to increase the width of CI, so that pseudo-numbers based on adjusted data are smaller than the actual numbers used in the unadjusted analysis [12]. Third, it seemed reasonable to suppose that in most circumstances adjustment would not have a large effect on p and z. Finally, odds ratios and CIs for comparisons estimated by the method seem in many situations to be little affected by the precise choice of p and z.

To illustrate the final point, Table IX shows the effect of varying p and z for data from the lung cancer case–control study of Matos *et al.* [13]. The data by age at start of smoking are used for estimating the overall covariate-adjusted odds ratios and CIs for current smoking *versus* lifelong non-smoking. The values of p of 0.4545 and z of 1.967 shown in the table are those derived from the unadjusted data, which led to our estimate of 8.54 (4.32–16.87). Although variation in the estimated odds ratio is evident as p and z change, this is not large, given the substantial variation in p and z allowed in this sensitivity analysis. It is of interest to compare these estimates with the lower odds ratio and narrower CI of 8.25 (5.26–12.95) when the three estimates by age at start are combined by fixed-effects meta-analysis [3], incorrectly assuming that they are independent.

DISCUSSION

A standard method of presenting results from epidemiological studies by level of exposure involves presenting the numbers of cases and controls (or at risk) for each level, together with covariate-adjusted effect estimates (odds ratios or relative risks) and their CIs for all but one level relative to the other (baseline) level. Often researchers are interested in alternative comparisons, for example, combining present with past exposure, in order to be able to compare ever with never exposure. However, the standard data presentation does not in general allow the calculation of valid alternative effect estimates (although it can do so in the simple situation of a pairwise comparison of two of the original exposure levels), and never allows the calculation of their valid CIs. This is because the effect estimates at the different exposure levels are non-independent and the standard data presentation does not give information on covariances between the estimates.

DOI: 10.1002/sim

The idea of generating a table of effective numbers of subjects by exposure level corresponding to all the adjusted effect estimates was put forward by Greenland and Longnecker [7] and applied by Berlin *et al.* [8] as a method of obtaining trend estimates from summarized dose–response data. (By 'corresponding' we mean that applying standard formulae for 2×2 tables to the effective numbers will generate the required effect estimates.) The method presented in this paper is a modification of this, in which the generated table of effective numbers corresponds both to the adjusted effect estimates and to their CIs. The table allows adjusted effect estimates and CIs to be calculated for any alternative comparison of levels (including a dose-related trend statistic), and can help a dose–response meta-analysis and/or support a sensitivity analysis for methodological bias [16]. Our method has also been extended to the situation where the original data are for two exposure levels and multiple disease categories, rather than two disease categories and multiple exposure levels.

When using this method, some care should be taken to ensure that the categories to be combined are non-overlapping and, together, are equivalent to the stated summary category. For example, smokers categorized as smoking '<20 cigarettes per day' and '20+ cigarettes per day' could reasonably be combined to represent 'All cigarette smokers', provided data were available on cigarette consumption for all (or most) of the sample. However, 'cigarette smokers', 'pipe smokers', and 'cigar smokers' could not be combined into 'smokers' if some subjects appeared in more than one of the three categories. Similarly, the lung cancer categories 'squamous cell carcinoma' and 'adenocarcinoma' could be combined to represent 'Lung cancer: squamous + adeno', whereas including an extra category 'other lung cancer' would justify the title 'All lung cancer'.

The method has the advantage of being widely applicable as it makes use of data values that are generally available in a published study. We put forward no proof that the method always gives a unique solution, and indeed in extreme situations the method can fail to converge (although, as noted in Appendix A, this can often be resolved by using different starting points for the iteration process). However, we have found that the method gave seemingly appropriate estimates in practical applications on many hundreds of sets of study results.

There is certainly scope for further work to gain greater insight into possible circumstances when the method may give unsatisfactory results. However, we feel that the method is a useful one, especially when one is trying to conduct a meta-analysis, as it assists in allowing risk estimates to be presented in a consistent way. While some studies publish estimates for overall exposure and some publish only estimates by level of exposure, and one wishes to incorporate an estimate from every study into the meta-analysis to gain additional power, it is clearly helpful to obtain an estimate for overall exposure from those studies that only give results by exposure level. One can of course try to obtain an estimate from the author using the source data, but that is not always feasible, especially if the study was conducted many years ago. In this circumstance, our method can help to obtain reasonable estimates—certainly better than estimates obtained using methods that ignore the interdependence of the estimates by level. We hope that making available the Excel spreadsheet and the SAS macro on the website www.pnlee.co.uk/software.htm will help to facilitate future meta-analyses.

APPENDIX A: THE EXCEL IMPLEMENTATION

The Excel spreadsheet, which can be downloaded from www.pnlee.co.uk/software.htm, uses the approach described below for solving the equations. The method varies only slightly between

DOI: 10.1002/sim

case–control and prospective studies and between studies giving categories of disease rather than levels of exposure, as described above. The spreadsheet provides drop-boxes for selecting study type (case–control or prospective) and categorization (by exposure levels or by categories of disease), and the details of the spreadsheet's formulae depend on the values selected. The description below is based on a case–control study giving odds ratios and CIs by levels of exposure merely in order to simplify the terminology. Some details of the calculations are different for prospective (cohort) studies, as described above.

The user takes the following actions:

1. Selects study type (case–control) and categorization (by exposure levels) using the drop boxes.
2. Enters a 2×2 table of the overall numbers of subjects in the study—the numbers of cases and controls according to whether exposed or unexposed—as given in the study report.
3. Enters, for each level of exposure, the odds ratio and CI as given in the study report.
4. Specifies how the exposure levels will be grouped for the required estimated odds ratio and CI (note that the user can also specify that individual exposure levels are to be excluded from this estimate).
5. Clicks the 'Solve' button to generate optimized estimates of the effective numbers of cases and controls (A_0 to A_n and B_0 to B_n) and hence the required estimated odds ratio and CI.

The spreadsheet is set up to make the following calculations.

Using the 2×2 table of overall numbers of subjects to estimate p, z, A_0, and B_0

The proportion of unexposed in the population (p) is estimated from the 2×2 table as

$$\frac{\text{Number of unexposed controls}}{\text{Total number of controls}}$$

and the ratio of controls to cases (z) is calculated from the 2×2 table as

$$\frac{\text{Total number of controls}}{\text{Total number of cases}}$$

The 2×2 table of overall numbers of subjects is also used to give initial values for A_0 and B_0 using the numbers of unexposed cases and controls, respectively. These values will not necessarily be used in the final table of estimated numbers of cases and controls because for adjusted odds ratios the numbers in that table will be effective numbers of cases and controls rather than the actual numbers.

Estimating the number of cases for level i

The variance of the estimated log relative risk (V_i) is calculated for each exposure level using equation (4). From these, together with the odds ratio for each exposure level and the initial values for A_0 and B_0, initial estimates are calculated for the number of cases for each exposure level:

From equation (1),

$$B_i = \frac{A_i B_0}{A_0 R_i}$$

DOI: 10.1002/sim