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]

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).

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)
$\vdots$	$\vdots$
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
$\vdots$	$\vdots$	$\vdots$
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

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 $(i=0,\dots,n)$, $2(n+1)$ 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

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

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

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

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

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: Cases	Number of subjects: Controls	Odds ratio (95 per cent CI) Adjusted
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: Cases	Number of subjects: Controls	Odds ratio (95 per cent CI) Unadjusted, calculated
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: Cases	Number of subjects: Controls	Odds ratio (95 per cent CI) Unadjusted, calculated
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 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).

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 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

This can be used to remove B_i from the expression for V_i (equation (2)) by using

$$\frac{1}{A_i}+\frac{1}{B_i}=\frac{1}{A_i}\left(1+\frac{A_0}{B_0}R_i\right)$$

and substituting in equation (2) giving

$$A_i=\frac{\left(1+\frac{A_0}{B_0}R_i\right)}{\left(V_i-\frac{1}{A_0}-\frac{1}{B_0}\right)}$$

Estimating the number of controls for level i

A similar approach gives

$$B_i=\frac{\left(1+\frac{B_0}{A_0R_i}\right)}{\left(V_i-\frac{1}{A_0}-\frac{1}{B_0}\right)}$$

Optimizing the estimates of A_i and B_i

The values p and z were calculated using the numbers in the 2×2 table of cases/controls who are exposed/unexposed. Now the estimated values of A_i and B_i can be used to calculate similar values p' and z' using

$$p'=\frac{B_0}{\sum_{i=0}^{n}B_i}$$

$$z'=\frac{\sum_{i=0}^{n}B_i}{\sum_{i=0}^{n}A_i}$$

The sum of squared differences between p' and p and between z' and z is calculated as

$$\left(\frac{p-p'}{p}\right)^2+\left(\frac{z-z'}{z}\right)^2$$

Excel's Solve function is then used to bring this value as close as possible to zero by adjusting the values of A_0 and B_0. The resulting values of A_i and B_i are our estimates of the table of (effective) numbers of cases and controls by exposure level.

As described in the documentation available on the web page www.pnlee.co.uk/software.htm, the Excel spreadsheet also outputs chi-squared and p-values for heterogeneity and trend corresponding to the table of effective numbers of subjects and based on trend coefficients entered by the user. The formulae used for the two chi-squared values are derived by conditioning on the marginal totals, and are based on the K-dimensional hypergeometric distribution (see Breslow and Day [17] formulae 4.38 and 4.39).

Note that, on some occasions, clicking the 'Solve' button may generate a message that a feasible solution could not be found. In these circumstances, a solution may sometimes be obtained either by

DOI: 10.1002/sim

adjusting the precision setting of the iterative process (using a feature available on the spreadsheet) or by starting the iterative process from a different starting point. By default, the starting numbers of exposed and unexposed controls (or at-risk subjects) are set to be the numbers actually in the study. These can be altered by unprotecting the worksheet and entering alternative starting values. These actions are described in more detail in the documentation.

APPENDIX B: THE SAS IMPLEMENTATION

The SAS implementation is given as the macro RREst. The macro, which can be downloaded from www.pnlee.co.uk/software.htm along with documentation, estimates effective table frequencies from published results given as a set of measures of relative risk and confidence intervals. As for the Excel spreadsheet, it may be applied to odds ratios from case–control studies or risk ratios from cohort studies, and it handles results categorized by exposure levels as well as by categories of disease.

As an example, the macro is applied to the data taken from Smith *et al.* considered in the introduction of this paper. The initial 2×2 table of overall numbers of subjects in this study is

Exposure	Cases	Controls
Unexposed	10	22
All exposed	84	77

Two input data sets are required, one containing the odds ratios and confidence intervals and the other corresponding to this 2×2 table.

The SAS Code below reads the first input data set into SAS. In the example, the first data column (`level`) is the label for the category, while `Est`, `lower` and `upper` are the odds ratio and confidence limits. The last column, the variable named `Overall`, defines the contrast that is to be estimated, in this instance comparing the unexposed with the pooled group of all exposed.

```
data Smith_OR;
  input level $ 1-14 Est lower upper Overall;
cards;
Never Smoked     .    .    .    0
1-200          2.82 1.00 7.93 1
More than 200  2.24 0.75 6.68 1
;
run;
```

The second input data set is created in the data step below.

```
data Smith_IniFreq;
  input Case Control;
cards;
10 22
84 77
;
run;
```

The following call requests the macro to calculate the effective frequencies, the resulting estimation of the contrast called 'Overall' and a related confidence interval. In this call, `type=CC` (rather than `type=prospective`) indicates that the study type is case–control, and `levels=exposure` (rather than `levels=disease`) specifies the categorization of the results in the Smith study.

```
%RREst(Smith_OR, Smith_IniFreq,
        type=CC, levels=exposure);
```

The resulting table of effective frequencies and the estimate and confidence interval for the contrast 'Overall' are printed to the SAS output window. The results closely agree with those given above as found using the Excel spreadsheet.

An outline of how the SAS macro proceeds in finding a solution is now given, as in Appendix A, focusing on the situation of a case–control study with results given by level of exposure. The macro solves the system of equations by converting it into an optimization task on the unit square by reparameterization. After defining $\beta_1 = (A_0^{-1} + B_0^{-1})/V_{\min}$ and $\beta_2 = A_0/(A_0 + B_0)$ with $V_{\min} = \min\{V_1, \ldots, V_n\}$, each point (β_1, β_2) on the unit square represents a frequency table satisfying the conditions $A_0>0$, $B_0>0$, and $A_0^{-1} + B_0^{-1} < V_{\min}$ (the latter follows from formula (2)). Conversely, each such table is represented by a point (β_1, β_2). For given values $\beta_1, \beta_2 \in (0, 1)$, A_0 and B_0 are calculated as $A_0 = B_0\beta_2/(1-\beta_2)$ and $B_0 = (\beta_1\beta_2 V_{\min})^{-1}$, and the table entries A_i, B_i, and the quantities $\hat{p}$ and $\hat{z}$ are found using the formulae given in Appendix A. The solution (β_1, β_2) yields $\hat{p}$ and $\hat{z}$ identical to the p and z from the initial 2×2 table. As in the Excel spreadsheet, the search for such a solution is performed with an iterative method, using the SAS procedure PROC NLIN.

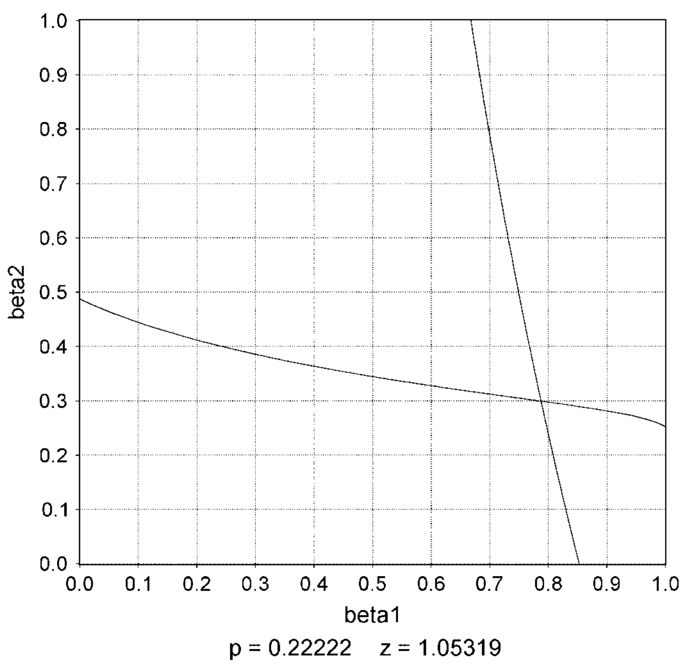

Figure B1. Solution of the optimization problem on the unit square. The solution is located at the intersection point of the two lines.

DOI: 10.1002/sim

To visualize the solution, the SAS macro draws a plot of two contour lines on the unit square, one contour joining the points (β_1, β_2) where the condition $\hat{p} = p$ is met, the second analogously corresponding to the condition $\hat{z} = z$. A solution of our problem must be located at an intersection point of the two lines. In the Smith example, p and z are calculated from the initial frequencies of cases and controls as $p = 0.2222$ and $z = 1.0532$. Figure B1 shows the resulting plot with the two contours on the unit square and the solution lying at the intersection of the two lines. The contour plot helps assess the existence and uniqueness of a solution, and may be useful if the solving process fails to find a satisfactory solution.

ACKNOWLEDGEMENTS

We thank Mrs P. Wassell and Mrs D. Morris for their patient and accurate typing of the various drafts of this paper. We are also grateful to Philip Morris International for financial support.

REFERENCES

1. Smith SJ, Deacon JM, Chilvers CED. Alcohol, smoking, passive smoking and caffeine in relation to breast cancer risk in young women. *British Journal of Cancer* 1994; **70**:112–119.
2. California Environmental Protection Agency. *Proposed Identification of Environmental Tobacco Smoke as a Toxic Air Contaminant*, *SRP Version*, 2005. www.arb.ca.gov/toxics/ets/finalreport/finalreport.htm.
3. Fleiss JL, Gross AJ. Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: a critique. *Journal of Clinical Epidemiology* 1991; **44**:127–139.
4. Easton DF, Peto J, Babiker AGAG. Floating absolute risk: an alternative to relative risk in survival and case–control analysis avoiding an arbitrary reference group. *Statistics in Medicine* 1991; **10**:1025–1035.
5. Greenland S, Michels KB, Robins JM, Poole C, Willett WC. Presenting statistical uncertainty in trends and dose–response relations. *American Journal of Epidemiology* 1999; **149**:1077–1086.
6. Easton D, Peto J. Re: ‘Presenting statistical uncertainty in trends and dose–response relations’ [Letter]. *American Journal of Epidemiology* 2000; **152**:393–394.
7. Greenland S, Longnecker MP. Methods for trend estimation from summarized dose–response data, with applications to meta-analysis. *American Journal of Epidemiology* 1992; **135**:1301–1309.
8. Berlin JA, Longnecker MP, Greenland S. Meta-analysis of epidemiologic dose–response data. *Epidemiology* 1993; **4**:218–228.
9. Orsini N, Bellocco R, Greenland S. Generalized least squares for trend estimation of summarized dose–response data. *Stata Journal* 2006; **6**:40–57.
10. Fry JS, Lee PN. Revisiting the association between environmental tobacco smoke exposure and lung cancer risk. I. The dose–response relationship with amount and duration of smoking by the husband. *Indoor and Built Environment* 2000; **9**:303–316.
11. Woolf B. On estimating the relationship between blood group and disease. *Annals of Human Genetics* 1955; **19**:251–253.
12. Lee PN. Simple methods for checking for possible errors in reported odds ratios, relative risks and confidence intervals. *Statistics in Medicine* 1999; **18**:1973–1981.
13. Matos E, Vilensky M, Boffetta P, Kogevinas M. Lung cancer and smoking: a case–control study in Buenos Aires, Argentina. *Lung Cancer* 1998; **21**:155–163.
14. Fontham ETH, Correa P, Reynolds P, Wu-Williams A, Buffler PA, Greenberg RS, Chen VW, Alterman T, Boyd P, Austin DF, Liff J. Environmental tobacco smoke and lung cancer in nonsmoking women. A multicenter study. *Journal of the American Medical Association* 1994; **271**:1752–1759.
15. Katz D, Baptista J, Azen SP, Pike MC. Obtaining confidence intervals for the risk ratio in cohort studies. *Biometrics* 1978; **34**:469–474.
16. Greenland S. Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society*, *Series A* (*General*) 2005; **168**:267–306.
17. Breslow NE, Day NE. The analysis of case–control studies, vol. 1. In *Statistical Methods in Cancer Research*, Davis W (ed.), vol. 32. IARC: Lyon, 1980.

DOI: 10.1002/sim