APPENDIX G
Methods for derivation of relative risks

G1 Notation and terminology

A record in the RR database refers to a single comparison between an exposed (smoking) group and an unexposed group (usually a non-smoking group, but sometimes an alternative smoking group). The comparison is either *adjusted* or *unadjusted*, referring to adjustment (or lack of adjustment) for potential confounding factors.

For a case-control study, an unadjusted comparison is based on a 2×2 table of the numbers of cases and controls versus exposed and unexposed subjects, denoted by:

	Cases	Controls
Unexposed	a_0	b_0
Exposed	a_1	b_1

For a prospective study, the b_i represent the at risk population, or the number of man-years at risk.

In many circumstances, the data as originally presented compare a single unexposed level with n exposure levels, giving a $2 \times (n+1)$ table:

	Cases	Controls
Unexposed	a_0	b_0
Exposed 1	a_1	b_1
Exposed 2	a_2	b_2
…	…	…
Exposed n	a_n	b_n

In some studies the cases are sub-divided according to histological type of lung cancer. These may be presented against a single common control group:

	Cases Type 1	Cases Type 2	…	Cases Type l	Controls
Unexposed	$A_{1,0}$	$A_{2,0}$		$A_{l,0}$	b_0
Exposed 1	$A_{1,1}$	$A_{2,1}$		$A_{l,1}$	b_1
Exposed 2	$A_{1,2}$	$A_{2,2}$		$A_{l,2}$	b_2
…	…	…		…	…
Exposed n	$A_{1,n}$	$A_{2,n}$		$A_{l,n}$	b_n

or may each have a separate control group:

	Cases Type 1	Controls for Type 1	…	Cases Type *l*	Controls for Type *l*
Unexposed	$A_{1,0}$	$B_{1,0}$		$A_{l,0}$	$B_{l,0}$
Exposed 1	$A_{1,1}$	$B_{1,1}$		$A_{l,1}$	$B_{l,1}$
Exposed 2	$A_{1,2}$	$B_{1,2}$		$A_{l,2}$	$B_{l,2}$
…	…	…		…	…
Exposed *n*	$A_{1,n}$	$B_{1,n}$		$A_{l,n}$	$B_{l,n}$

Occasionally there may be a single case group but two control groups (e.g. hospital and population controls).

Tables may be presented separately for several strata (e.g. age groups), or for separate levels of a potentially confounding factor (e.g. pollution or occupational exposure), thus forming, e.g. a $2 \times (n + 1) \times m$ table.

The relative risk and its lower and upper 95% confidence limits are denoted by RR, LCL and UCL respectively. ϕ denotes a factor related to the variance of the RR

$$\phi = N_{95} \times \sqrt{\text{var}(RR)} \qquad [1]$$

N_c denotes the inverse standard normal value for confidence level c (e.g. $N_{95} = 1.96$).

G2 Basic method for unadjusted RR

As described in section 3.4.5, an unadjusted RR and its CI are calculated from a 2×2 table by:

$$RR = (a_1\, b_0) / (a_0\, b_1) \qquad [2]$$

$$LCL = RR / \phi \qquad [3]$$

$$UCL = RR \times \phi \qquad [4]$$

where ϕ is given by

$$\ln(\phi) = N_{95}\sqrt{((1/a_0) + (1/a_1) + (1/b_0) + (1/b_1))} \quad \text{for a CC study,} \qquad [5]$$

or $$\ln(\phi) = N_{95}\sqrt{((1/a_0) + (1/a_1) - (1/b_0) - (1/b_1))} \quad \text{for a prospective study.} \qquad [6]$$

Since for a prospective study the b_i are much larger than the a_i, the approximation

$$\ln(\phi) = N_{95}\sqrt{((1/a_0) + (1/a_1))} \qquad [7]$$

may be used to calculate the CI if RR, a_1 and a_0 are known but b_1 and b_0 are unknown.

The 2 × 2 table may be as given originally, calculated from a matched-pairs table, or estimated from a percentage distribution, in which case it may be subject to rounding error. It may also be calculated by summing over exposure levels, over disease levels, over control groups, and/or over strata/confounder levels.

G3 Correction for a zero cell

If one cell in a 2 × 2 table is equal to zero, then the basic formulae are adjusted by adding 0.5 to each cell:

$$RR = ((a_1+0.5)\,(b_0+0.5)) \,/\, ((a_0+0.5)\,(b_1+0.5)) \qquad [8]$$

and

$$\ln(\phi) = N_{95}\sqrt{((1/(a_0+0.5))+(1/(a_1+0.5))+(1/(b_0+0.5))+(1/(b_1+0.5)))} \qquad [9]$$

for a CC study, (and similarly as above for a prospective study).

G4 Adjusting for a potential confounder and combining independent RRs

If results are given separately for different *m* levels of a potentially confounding factor, either as a 2 × 2 × *m* table, or as *m* RRs and CIs, then the overall RR and CI, adjusting for the factor, is calculated by the method of Fleiss and Gross[1]. If the original *m* RRs were adjusted, then the new estimate is adjusted for both the original and the new factors.

This method is also used when RRs and CIs were given originally for specific types of lung cancer, each with their separate control group, to combine over the disease groups.

G5 Converting CI from different confidence level

If a RR and CI were originally presented with a different confidence level *c* (e.g. *c* = 90%) then the 95% CI is calculated using formulae [3] and [4] with:

$$\ln(\phi) = (\ln(UCL_c) - \ln(LCL_c)) \,/\, (2 \times N_c) \qquad [10]$$

G6 Inverting from a different denominator

If a RR and CI were originally presented with the exposed and unexposed groups reversed from those required (e.g. plain vs filter rather than filter vs plain), then the required values are calculated as:

$$RR = 1 / RR_O \quad [11]$$

$$LCL = UCL_O \quad [12]$$

$$UCL = LCL_O \quad [13]$$

where the subscript *O* indicates the values as originally presented.

G7 Ratio of rates

Prospective studies often present mortality rates rather than RRs. If they are presented separately for the exposed and unexposed groups (R_1 and R_0), then the RR is calculated by:

$$RR = R_1 / R_0 \quad [14]$$

If CIs for the mortality rates are also available (L_1, U_1 and L_0, U_0), then the CI for the RR can be calculated by using:

$$\ln(\phi) = \sqrt{((((\ln(U_0) - \ln(L_0))/(2 \times N_{95}))^2 + ((\ln(U_0) - \ln(L_0))/(2 \times N_{95}))^2)} \quad [15]$$

G8 Using symmetry of the CI

When only two of the RR, LCL and UCL are given, then the third is calculated to give a CI symmetrical about the RR. For instance if UCL is missing, then formula [4] is used with:

$$\phi = RR / LCL \quad [16]$$

G9 Combining non-independent RRs

The method of Fry and Lee[2] is used to combine non-independent RRs. This is most commonly applied to adjusted results presented for *n* exposure (smoking) levels relative to a single unexposed (non-smoking) level. The method defines the parameters P to be the proportion of unexposed subjects in the control group/at risk population, and Z the relative frequency of the control group/at risk population to the case group. The hypothetical underlying $2 \times (n + 1)$ table of 'adjusted' cases/controls $\times$ exposure level is then estimated to give the same RRs and CIs as original, and to give P and Z as close as possible to their original values. These numbers can be

summed to combine smoking groups as required, and the resulting 2×2 table is then used to calculate the adjusted RR and CI for the new combination using formulae [2] – [5].[a]

A number of points can be noted:

- both the numerator and the denominator may be either a single original level or a combination of the original levels.
- if the numerator is the original base group and the denominator is one of the original levels, then this would give the same result as inverting (section G6)
- all the groups from the original table are included in the estimation process even if not all are required for the combinations of interest.
- the parameters P and Z required for the estimation process are generally available. Any specific problems with these values are noted in the RR database as comments. The numbers of unexposed cases and controls/at risk are also generally available and are used as the starting values for the iterative solution (although no comment is entered if some other values are used).
- although the base groups is generally 'unexposed' (i.e. non-smoking), the method is equally applicable when the base group in the original table is a smoking level, for instance a base group of current smokers, with several levels of ex-smokers by duration of ex-smoking and a never smoker level; or a base group of always plain cigarette smokers, with several levels of filter smokers by time of switching.
- the method is also applicable when the original results are given by smoking and another factor. For instance, if a table gives results by both smoking level and pollution level, relative to a non-smoking non-polluted base group, then the method is used to obtain estimates for smoking relative to non-smoking regardless of pollution by summing both smoking groups and non-smoking groups over all the pollution levels. Equally, the RR and CI for smoking within each level of pollution can be obtained (by choosing the appropriate numerator and denominator groups in turn), and then combined by the method of Fleiss and Gross[1] to obtain an estimate for smoking adjusted for pollution (in addition to the original adjusting factors).

A variant of the method is also used to combine over disease groups. In this case the source table gives RRs and CIs for a single smoking comparison (unexposed vs exposed), but for l disease groups, generally different histologic types of lung cancer, each compared to a single shared control group. The parameters are redefined, with P now the proportion of controls among the unexposeds, and Z the relative frequency of unexposed to exposed. The hypothetical underlying $(l+1) \times 2$

[a] Note that due to an error, the case-control formula [5] for ϕ was used throughout when formula [6] should have been used for prospective studies. However the method of estimating the relative risk was unaffected and, since for a prospective study the b_i are large compared to the a_i, this would have little effect on the estimated confidence limits (at most one digit in the second decimal place in samples tested).

table of cases/controls × exposure can then be estimated in the same way, the counts summed to combine disease groups, and the resulting 2 × 2 table used to calculate the adjusted RR and CI for the new combination.

Note that

- in this case, the original base (control) group is not combined with other groups, the method being used to combine lung cancer groups when the source paper only presented results for individual histologic types.
- the method may also be used to obtain results for lung cancer versus a combined control group when the source paper presented results versus two control groups separately

G10 Using SMRs, or expected values

When the observed numbers of cases are given together with SMRs or expected values relative to a standard (e.g. national) population, then the 'ratio of two standardised ratios' is calculated as described by Altman *et al*[3] using the program CIA.

G11 CI estimated from crude numbers

When an adjusted RR is presented without any CI, but the corresponding 2 × 2 table (or at least the numbers of cases for a prospective study) is available, then the original RR is used and a CI is estimated for it by assuming its width is the same as for the equivalent unadjusted RR, i.e. by using formulae [3] – [7] but not formula [2]. The RRs can then be further combined if necessary using the Fleiss and Gross, or Fry and Lee methods as appropriate (section G4 and G9).

If adjusted mortality rates are presented without CIs, but with the corresponding 2 × 2 table (or at least the numbers of cases for a prospective study), then the RR can be calculated as the ratio of the rates (as section G7), and the CI estimated from the crude numbers. However if these are then required to be combined over smoking levels, then, instead of using the method of Fry and Lee, the hypothetical numbers of 'adjusted controls/at risk' are estimated by dividing the numbers of cases by the rates, these numbers are summed to form the required combined smoking groups, and the resulting 2 × 2 table used to estimate the RR and CI for the combination by the usual formulae.

References

1. 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. *J Clin Epidemiol* 1991;**44**:127-39.

2. 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 Built Environ* 2000;**9**:303-16.

3. Altman DG, Machin D, Bryant TN, Gardner MJ, editors. *Statistics with confidence,* 2nd edition. London: BMJ Books; 2000.