RREst_trend – Relative risk estimation with various trend estimates

Author: Jan Hamling Date: 1 February 2019

Contents

1 Introduction

An epidemiological study may report a relative risk and confidence interval for each of several categories, each compared with a single unexposed group. The user may want a different comparison, such as the overall relative risk (RR)[*] and confidence interval (CI) for exposure versus no exposure. Standard meta-analysis techniques cannot be used because the RRs provided are not independent: they share a single comparison group.

RREst is an Excel spreadsheet program to manipulate these non-independent relative risks and confidence intervals to give the RR and CI of a comparison different from the one provided. It can also assess the significance of the heterogeneity between the categories; and estimate the rate of increase in risk per unit dose of exposure to the risk factor considered (the trend over the categories of exposure).

The features of the program are:

- Estimation of effective numbers of participants in each category, both cases and controls/non-cases/at risk.
- The user can specify the contrast they require, by specifying which of the available categories should make up the baseline and which the comparison group, together with the possibility of excluding some categories.
- Estimation of RR (CI) for the specified contrast.
- Estimation of heterogeneity among the selected categories.
- The user can specify trend coefficients (dose values).
- Estimation of significance of trend according to the dose values entered.
- Estimation of trend (risk per unit dose) according to the dose values entered.
- Estimation of trend (risk per unit dose) for ordered non-numeric categories, based only on the distribution (of the population considered in the study) over the categories. Here the dose unit represents the difference between the lowest and highest possible value of the underlying scale. This is referred to as the Uniform scale.
- Case control, prospective and cross-sectional studies are handled.
- Categories of exposure or of disease are handled (although trend estimation is meaningless for results presented as categories of disease).

[*] The term ‘relative risk’ is intended to describe any of the risk ratio measures such as odds ratio.

- Results are displayed to the selected number of decimal places. This can be up to 6 decimal places, but it is not usually appropriate to quote the result values to more than one or two decimal places because the data entered are usually available to no more than two decimal places. Results are rounded values.
- Estimation by iterative process (to generate the effective numbers of participants) uses a user-specified precision. This can be adjusted easily if convergence problems arise.

The method used to estimate the required comparison involves estimating the numbers of participants associated with each of the categories entered. It is then a simple process to calculate the RR and CI of an alternative comparison, simply by combining the estimated numbers of participants into a baseline group and a comparison group. Where the study provides adjusted RRs, the estimated numbers are 'effective' numbers of participants – the number of participants from a population equivalent to a notional 'adjusted' population. These effective numbers are also referred to as pseudo-numbers.

When data are entered by categories of exposure, various trend estimates are provided. Previous versions of this spreadsheet provided an assessment of the significance of the trend over the categories. It now also provides an estimate of the Rate (the rate of increase in risk per unit dose) and CI for the trend, first by a method based on the dose values entered, and secondly by a method (described as the Uniform scale) that can be used when ordered non-numeric categories (such as None, A little, A lot) are provided in the study report. The results of both methods of assessing the Rate are presented. The user must select the more appropriate result.

2 Using the program

2.1 First impressions

The program uses macros to perform the estimation processes. Select 'Enable macros' when the spreadsheet is first opened.

Most of the input cells (the "Data entry area") and the Results cells are visible when the spreadsheet is first opened, and the whole spreadsheet can be printed on three sides of A4 paper.

The spreadsheet has input cells and display-only cells. The input cells are coloured pale turquoise. Note that Excel's usual features are available in the input areas so, for example, preliminary calculations (such as the total number of exposed participants) can be entered as calculations.

The display-only cells are protected: a message will be shown if the user attempts to enter a value. This prevents the user from deleting cell calculations unintentionally. Pressing Tab moves the cursor into the next input cell.

To the right of the input cells (on the second side of the printed spreadsheet) are some instructions on using the spreadsheet together with some of the cells used in the underlying calculations.

To the right of that (on the third side of the printed spreadsheet) are some details of the trend tests together with cells relating to the trend calculations.

If the spreadsheet is to be saved once data have been entered, the user may select a suitable location and filename.

2.2 Entering data, selecting options

The first input cells are for a heading. This is often used for a description of the source of the data.

The Results are shown just below the heading. The number of decimal places shown can be changed by clicking the up and down arrows to the right of the Results cells. Also in this area is the 'Calculate' button. Pressing this button runs the macros that calculate the effective numbers of participants and the Results values shown in this area.

Below the Results section is the Data entry area. The first values to be entered are provided as two drop-down boxes. The first of these allows the study type (case control, prospective or cross-sectional) to be chosen. The second is used to specify how the data are categorised – by levels of exposure to the risk factor or by categories of disease. The choices made for these two drop-downs affect both the titles shown for the input cells and the calculations performed by the spreadsheet. For example, a case control study involves control participants whereas a prospective study involves participants 'at risk' and these terms are used as headings in various parts of the spreadsheet. The method of calculating variance is different in these two situations. Therefore **be sure to make the right selections using these two drop-downs before entering the rest of the data**.

The next group of input fields, entitled 'No. of participants', is a 2x2 table. For a case-control study, these are the numbers of cases and controls considered in the study report according to their being exposed or unexposed to the risk factor considered. These values relate to the study as it is reported so, for example, the numbers unexposed are those in the study report's reference group (rather than those to be included subsequently in a user-specified contrast). For a cross-sectional study, numbers of cases and controls are replaced by numbers with or without the outcome of interest (cases and non-cases), while for a prospective study, they are replaced by numbers of cases of the outcome and numbers at risk.

The RR and CI values provided by the study report for each of the categories are entered below this. A category description can be entered for each one. Taken together, the categories should provide information on all the participants considered in the study report (all those counted in the 2x2 table). The first category represents the reference group used in the study report, and so has OR/RR set to 1. Up to 29 additional categories can be entered. For studies categorised by level of exposure, the categories should generally be entered in the order of increasing amount of exposure to the risk factor.

The column headed 'Contrast' is used to specify how the user would like the spreadsheet to combine the categories when calculating the Results fields labelled 'Overall risk', i.e. which categories should form the analysis baseline and which the comparison group. It can also be used to exclude categories from these calculations. Enter the values:

0 for categories that will form the analysis baseline

1 for categories that will form the comparison group

-1 for categories that should be ignored in the calculation of Results.

The Contrast settings have no effect on the calculation of effective numbers of participants which always makes use of all categories in the study.

The column headed ‘Dose’ is used to specify coefficients for trend analysis. For each category enter a positive value to be used as a trend coefficient (mean dose). This should be proportional to the amount of exposure to the risk factor. For trend results to be calculated, the categories should be entered in the order of increasing amount of exposure, so the dose values should go from low to high.

To the right of these input columns the program will generate two columns of estimated numbers of participants (the pseudo-numbers). These columns will show estimates for each category entered. To make the program calculate these values, click the button marked ‘Calculate’. This runs an iterative process which finds a ‘best’ solution based on the values entered in the 2x2 table and in the OR/RR, CI Lower and CI Upper columns – see Section 2.4 and Appendices B and C for details. The Calculate process also calculates Results values.

The last input cells (found below the Data entry area) are headed ‘Notes’. These can be used for further details of the study, for calculations (cell formulae) or any other purpose the user wishes.

2.3 Sequence of events

In Excel open the spreadsheet RREst_trend.xlsm. If requested, click “Enable macros”. The spreadsheet will then be displayed.

Optionally, enter text in the Heading.

Appropriate drop-down options should be chosen for ‘Study type’ (case control, prospective or cross-sectional) and ‘Categorised by’ (by levels of exposure or by categories of disease) before the other data are entered.

Values should then be entered in the 2x2 table of ‘No. of participants’ and in the columns Category, OR/RR, CI Lower and CI Upper, Contrast and Dose.

Once changes have been made to the data entered, the Results values are not valid until they have been recalculated. The Results fields are set to zero until the calculations have been completed successfully.

Press the button 'Calculate'.

This will generate the columns of estimated numbers of participants for each category (the effective numbers of participants, the pseudo-numbers). This is done using an iterative process. A results window may pop up with the text:

> Solver could not find a feasible solution.

If this happens, see section 4 below for guidance.

The Calculate process then generates estimates in the Results area: Overall risk, Heterogeneity and the various Trend measures. Information and error messages may be shown in the bottom right of the Results area.

2.4 Calculation details

The calculation of Estimated numbers of cases and controls/non cases/at risk uses the method described by Hamling et al. (2008) [1]. This is described in more detail in Appendices A and B below. Only the settings for 'Study type' and 'Categorised by', the 2x2 table and the OR/RR (CI) values are used for this; Contrast and Dose are ignored.

The 'Overall risk' OR/RR and CI values are calculated using these effective numbers of participants together with the values in the Contrast column. The categories with 0 in Contrast are grouped together as the baseline. The categories with 1 in Contrast together form the comparison group. Categories with -1 in Contrast are excluded.

A prospective study giving results categorised by disease will have the 'at risk' as the first Category. This category includes all the participants in the study so it makes no sense to specify additional baseline levels. The calculations will always use only the 'at risk' level as the baseline. Any request (using the Contrast column) for the baseline to include a disease category, or for the baseline to exclude the 'at risk' category will be ignored.

'Heterogeneity' and 'Trend (Breslow 1980)' results are calculated using Breslow and Day's separately published methods for case control or cross-sectional [2] and prospective [3] studies. See Appendix C. These calculations also take account of the Contrast column by excluding

categories with -1 in Contrast. The calculation of this trend result makes use of the Dose values for the categories included.

'Trend: rate of increase in risk per unit dose' results use the method given by Berlin et al. (1993) [4] together with the correction for the non-independence of results by exposure level described by Greenland and Longnecker (1992) [5]. These sources describe the method for case-control and cross-sectional studies. Orsini et al. (2012) [6] provide the modifications to be used for prospective studies. Categories with -1 in Contrast are excluded. See Appendix D for details. Note that the method requires inclusion of the Unexposed category. Entering -1 in Contrast against the Unexposed category will therefore not exclude that category from this trend calculation and a warning message will be shown in the Results section.

The first of these trend estimates, marked 'Dose as entered', makes use of the values in the Dose column. However, some studies report ordered non-numeric categories such as 'None', 'A little' and 'A lot'. For these no meaningful dose values can be derived. To provide an indication of trend for this type of data, the 'Uniform scale' results are given. For these, the dose values entered are ignored. The scale is based on the distribution of the participants across the exposure levels. Exposure is assumed to range from 0 (least possible) to 1 (most possible), and the N participants in the study are considered to have equally spaced exposures ranging from 1/(2N) to 1−1/(2N). This "uniform scale" allows the combination of effect estimates using different measures of an underlying common exposure. See Appendix E for details.

3 Troubleshooting: Excel can't find the Solver

The 'Calculate' button makes use of the Excel add-in Solver. If this feature was not included when Excel was installed then clicking that button will generate a message such as:

"Compile Error: Sub function not defined"

If this happens it is necessary to install the Solver add-in.

To install the Solver:

1. Within Excel find 'Options'. This may be under the File tab.
2. Click 'Add-ins' in the left-hand panel.
3. Select 'Manage' 'Excel Add-ins' 'Go' at the bottom of the window.
4. Tick the box against Solver add-in and click OK.

Follow the instructions that may appear on the screen.

If this does not solve the problem it may also be necessary to associate the Solver with the spreadsheet's macro code:

1. Open the spreadsheet.
2. Right-click one of the macro buttons and select 'Assign macro' then 'Edit'.
3. Within the code window select the Add-Ins tab, Add-In Manager.
4. Tick the box against Solver add-in and click OK.
5. Close the code window.

If necessary, find and add the file Solver32.dll or SOLVER.xlam, depending on what Excel is requesting. This will probably be found in:

C:\Program Files\Microsoft Office\Office*nn*\Library\Solver\Solver32.dll

or in

C:\Program Files\Microsoft Office\Office*nn*\Library\Solver\Solver.xlam

where *nn* is a number such as 10, 11 or 12, depending on the version of Excel installed.

4 Troubleshooting: The Solve process doesn't find a feasible solution

Occasionally pressing the 'Calculate' button will give the message:

> "Solver could not find a feasible solution"

There are three reasons why this can happen:

- There is a mistake in the data entered;
- The solve process came quite close to a solution but couldn't manage to satisfy the solver's Precision specification;
- The solve process could not step from its starting values to any reasonable solution.

The first of these is the most common reason. It is easy to leave out a decimal point from a numeric value or to specify study type 'Prospective' for a case control study. It is also not uncommon for there to be errors in the values given in a study report. See Lee (1999) [7] for simple methods of checking the feasibility of the values reported.

To check whether the second of these is the problem, the user should change the setting in the Solver Precision drop-down within the Results section. Whenever the Solver Precision setting is changed the 'Calculate' process is performed automatically, so the new estimates are displayed.

If adjusting the Solver Precision setting doesn't give a feasible solution, then the final reason may apply. For this problem it may be useful to try adjusting the spreadsheet's Start values. The rest of this section discusses how to do this.

The spreadsheet uses the 2x2 table to:

- Calculate the values *P* and *Z* (shown at the bottom of the second printed side),
- Calculate the 'Start values' (shown just above the columns of Estimated numbers of participants in columns G and H).

For a case control study categorised by levels of exposure, *P* represents the proportion of unexposed participants among the controls and *Z* represents the relative frequency of controls to cases overall. (The meanings of *P* and *Z* are different for other combinations of Study type and Categorisation but operate identically in the solve process.)

The values *P'*, *Z'* and Sum of Squares are also shown near the *P* and *Z* values. These give a measure of the accuracy of the iterative process performed when the 'Calculate' button is pressed. They are described in more detail below.

The 'Calculate' process works by:

- Copying the pair of 'Start values' into the first row of the columns of Estimated numbers.
- Calculating the other values in the columns of Estimated numbers using these first row values and the OR/RR (CI) for each category (details of the calculations are given in Appendix B below).
- Using the table of Estimated numbers to calculate *P'* and *Z'*.
- Calculating the Sum of Squares value using the formula:

$$\left(\frac{P - P'}{P}\right)^2 + \left(\frac{Z - Z'}{Z}\right)^2$$

  This is a measure of the extent to which the estimates (*P'* and *Z'*) differ from the study-defined values (*P* and *Z*).

- Repeatedly adjusting the values in the first line of the table of Estimated numbers (and recalculating) until the Sum of Squares value is small enough to satisfy the specified precision.

For some studies the 'Calculate' process cannot find an acceptable solution – the solution generated gives an unacceptably large Sum of Squares value. In these circumstances it may be worth trying different Start values. Originally these values are copied (automatically) from the first row of the 2x2 table.

To overwrite the Start value cells with different values it is necessary to turn off worksheet protection: click Review: Unprotect sheet. The spreadsheet will now allow any cell to be overwritten, including those containing formulae, so care is needed when using this feature.

Now enter new values in the Start values cells (cells G20 and H20) and press the 'Calculate' button again. This can be done as many times as necessary to find suitable Start values.

Once worksheet protection has been turned off it is also possible to change other Solve parameters, such as the maximum number of iterations performed. The Solve process can also be set to show the values generated at each iteration. To do any of these, press:

Data: Solver: Options

This shows a number of boxes which allow the user to adjust the solve parameters. There is also a check box for the option ‘Show Iteration Results’.

5 Appendix A: The equations to be solved and the process involved

5.1 The notation

A0, B0 The pair of numbers in the first row of the table of Estimated numbers of participants. For a study giving RRs by exposure levels these represent the number of unexposed cases and the number of unexposed controls/non-cases/at risk participants respectively.

A1 to *An* The number of participants in the first column of the table of Estimated numbers for category 1 onwards. For a study giving OR/RRs by categories of exposure these represent the estimated number of cases in each exposure category (1 to n).

B1 to *Bn* As A_1 to *An* but for the second column of the table of Estimated numbers of participants. For a study giving OR/RRs by categories of exposure these represent the estimated number of controls/non-cases/at risk participants in each exposure category (1 to n).

P Value calculated from the second column of the 2x2 table (first value in second column over total for the second column). For a study giving OR/RRs by categories of exposure this represents the proportion of unexposed participants among the controls/non-cases/at risk participants.

Z Value calculated from the totals of the 2x2 table (second column total over first column total). For a study giving OR/RRs by categories of exposure this represents the relative frequency of controls/non-cases/at risk participants to cases.

P' Estimated value of *P* derived from the table of Estimated numbers of participants.

Z' Estimated value of *Z* derived from the table of Estimated numbers of participants.

R_i Relative risk (OR or RR) for category *i* (entered by the user).

L_i Lower value of confidence interval for category *i* (entered by the user).

U_i Upper value of confidence interval for category *i* (entered by the user).

V_i Variance of $\log_e (R_i)$ for category *i*.

A_b Total estimated number of cases/exposed participants in the categories chosen to be the baseline group (as defined by the entries in the Contrast column).

A_c Total estimated number of cases/exposed participants in the categories chosen to be the comparison group (as defined by the entries in the Contrast column).

B_b Total estimated number of controls/non-cases/at risk/unexposed participants in the categories chosen to be the baseline group (as defined by the entries in the Contrast column).

B_c Total estimated number of controls/non-cases/at risk/unexposed participants in the categories chosen to be the comparison group (as defined by the entries in the Contrast column).

R Relative risk for the comparison group compared with the baseline group (as defined by the entries in the Contrast column).

L Lower value of confidence interval for *R*.

U Upper value of confidence interval for *R*.

V Variance of $log_e (R)$.

5.2 The equations

0) $$V_i = \left\{ \frac{log_e \left({}^{U_i}/_{L_i} \right)}{3.92} \right\}^2 \qquad (i=1, \ldots n)$$

This is the variance of the log of *Ri*. These variance values are fundamentally what determine the estimated numbers of participants in each category, in that the width of a confidence interval reduces as the number of participants increases. Note that U_i and L_i are entered by the user.

1) $$P' = {B_0}\Big/{\sum_{i=0}^{n} B_i} \qquad \text{(i=1, ... n)}$$

The estimate of *P* calculated using the table of estimated numbers of participants.

2) $$Z' = {\sum_{i=0}^{n} B_i}\Big/{\sum_{i=0}^{n} A_i} \qquad \text{(i = 1, ... n)}$$

The estimate of Z calculated using the table of estimated numbers of participants.

3) $$R_i = \frac{A_i B_0}{B_i A_0} \qquad \text{(i = 1, ... n)}$$

The relationship between the relative risk for level *i* and the Estimated numbers of participants (A_i and B_i). The value of R_i is known because it is entered by the user. This equation is used in calculating A_i and B_i (see Appendix C).

4) The variance of log_e (R_i) in terms of the Estimated numbers of participants (A_i and B_i).

For case control and cross-sectional studies:

$$V_i = \frac{1}{A_0} + \frac{1}{B_0} + \frac{1}{A_i} + \frac{1}{B_i} \qquad \text{(i = 1, ... n)}$$

For prospective studies giving RRs by exposure level:

$$V_i = \frac{1}{A_0} - \frac{1}{B_0} + \frac{1}{A_i} - \frac{1}{B_i} \qquad \text{(i = 1, ... n)}$$

And for prospective studies giving RRs by disease category:

$$V_i = -\frac{1}{A_0} - \frac{1}{B_0} + \frac{1}{A_i} + \frac{1}{B_i} \qquad \text{(i = 1, ... n)}$$

These equations are used in calculating *Ai* and *Bi* (see Appendix C).

5) Generally the Overall risk RR (the contrast's point estimate) is given by

$$R = \frac{A_c B_b}{B_c A_b}$$

but, for a prospective study giving RRs by disease category, the baseline group is always the 'at risk' category:

$$R = \frac{A_c B_0}{B_c A_0}$$

(As equation (3) but for the requested contrast.)

6) To calculate the variance of $log_e(R)$ for the Overall risk (as equation (4) but for the requested contrast):

For a case control or cross-sectional study:

$$V = \frac{1}{A_b} + \frac{1}{B_b} + \frac{1}{A_c} + \frac{1}{B_c}$$

For a prospective study giving RRs by exposure level:

$$V = \frac{1}{A_b} - \frac{1}{B_b} + \frac{1}{A_c} - \frac{1}{B_c}$$

And for a prospective study giving RRs by disease category (which always uses the 'at risk' participants as the baseline):

$$V = -\frac{1}{A_0} - \frac{1}{B_0} + \frac{1}{A_c} + \frac{1}{B_c}$$

7) $log_e(U) = log_e(R) + 1.96\sqrt{V}$

8) $log_e(L) = log_e(R) - 1.96\sqrt{V}$

5.3 The steps involved in the solution

As soon as the user has entered the RRs and CIs for the levels of exposure/disease, the *Vi* (the variance of the log of each *Ri*) are calculated using formula (0). This is possible because the values depend on the entered CI values only.

When the user presses Calculate:

- The values of A_0 and B_0 are copied from the Start values into the first line of the table of Estimated numbers of participants. These values will be the starting point for the

iterative process. For RRs given by exposure level these are the numbers of unexposed (cases and controls/non-cases/at risk participants).

- The spreadsheet calculates:

a) V_{extra} values using the V_i and the estimates of A_0 and B_0:

For a case control or cross-sectional study:

$$V_{extra} = V_i - \frac{1}{A_0} - \frac{1}{B_0}$$

For a prospective study giving RRs by exposure level:

$$V_{extra} = V_i - \frac{1}{A_0} + \frac{1}{B_0}$$

And for a prospective study giving RRs by disease category:

$$V_{extra} = V_i + \frac{1}{A_0} + \frac{1}{B_0}$$

See Appendix C for the derivation of these formulae.

These values are used in the next calculation.

b) Estimates of the numbers of participants in each level (A_1 to A_n and B_1 to B_n) using the estimates of A_0, B_0 and the V_{extra} values, making use of a combination of equations (3) and (4) - see Appendix C for details of the calculations.

c) Totals of the estimated numbers of participants for the requested comparison (A_b, A_c, B_b and B_c) using the contrast definition and the estimated numbers of participants A_0 to A_n and B_0 to B_n.

d) R (the RR of the required comparison) using equation (5)

e) V (the variance of the $\log_e$ (RR) of the comparison) using equation (6)

f) Log (CI) of the required comparison, using equations (7) and (8)

g) The CI of the required comparison, by exponentiating the values given in (f)

h) P' using equation (1)

i) Z' using equation (2)

j) The sum of squares value:

$$\left(\frac{P - P'}{P}\right)^2 + \left(\frac{Z - Z'}{Z}\right)^2$$

- The spreadsheet's Solver routine then runs an iterative process with the following definitions:

  Solution cell: cell W56, the cell containing the Sum of Squares value calculated at (j)

  Target value: 0

  Variable cells (the values modified by Solver's iterative process): the first two cells in the table of Estimated numbers of participants (which contain the values A_0 and B_0)

  Solver modifies the values in the variable cells (A_0 and B_0) in an attempt to reach 0 in the Sum of Squares cell. Each time a new value is tried, all the calculations (a)-(j) above are reworked and so a new Sum of Squares value results. The final values of A_0 and B_0 generate our best estimates for A_0-A_n and B_0-B_n and hence the RR (CI) of Overall risk, the requested contrast.

Notice that the calculation of the contrast ((c)-(g) above) does not affect the results of the Solver routine. However, the values in Overall risk do depend on the results of the Solver (the final values for A_0 to A_n and B_0 to B_n). This means that, once Solver has produced a solution, a range of comparisons can be generated simply by changing the values in the 'Contrast' column and pressing 'Calculate'.

6 Appendix B: Calculating values for estimated numbers of participants

(given the entered RR (CI) values and the estimated values for A_0 and B_0)

Paragraphs (a) and (b) of Appendix A (section 5.3) describe the calculation of these values as using a combination of equations (3) and (4), i.e. the equations:

3) $$R_i = \frac{A_i B_0}{B_i A_0} \qquad (i = 1, \ldots n)$$

4) For case control and cross-sectional studies:

$$V_i = \frac{1}{A_0} + \frac{1}{B_0} + \frac{1}{A_i} + \frac{1}{B_i} \qquad (i = 1, \ldots n)$$

For prospective studies giving RRs by exposure level:

$$V_i = \frac{1}{A_0} - \frac{1}{B_0} + \frac{1}{A_i} - \frac{1}{B_i} \qquad (i = 1, \ldots n)$$

And for prospective studies giving RRs by disease category:

$$V_i = -\frac{1}{A_0} - \frac{1}{B_0} + \frac{1}{A_i} + \frac{1}{B_i} \qquad (i = 1, \ldots n)$$

Note that A_i and B_i appear in both (3) and (4). In order to calculate values for the table of Estimated numbers of participants we need A_i independently of B_i and *vice versa*.

Note that the values of R_i are known (their values were entered by the user), while V_i, A_0 and B_0 have already been estimated in previous calculations.

The following lines manipulate (3) and (4) to give equations for A_i and B_i respectively. This is shown separately for case control/cross-sectional studies, for prospective studies which give RRs by exposure level and for prospective studies which give RRs by disease category.

6.1 Case control/cross-sectional studies

From (4):

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

Putting this in (3):

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

giving

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

Similarly, from (4)

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

Putting this in (3):

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

giving

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

Notice that the calculations for both A_i and B_i involve dividing by:

$$\left(V_i - \frac{1}{A_0} - \frac{1}{B_0}\right)$$

For a case control study, this set of values is calculated in the spreadsheet under the title V_{extra}.

6.2 Prospective studies giving RRs by exposure level

From (4):

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

Putting this in (3):

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

giving

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

Similarly, from (4)

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

Putting this in (3):

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

giving

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

Notice that, for a prospective study which gives RRs by exposure level, the calculations for both A_i and B_i involve dividing by:

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

This set of values is calculated under the title V_{extra}. Notice that these V_{extra} values are different from those for a case control study.

6.3 Prospective studies giving RRs by disease category

From (4):

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

Putting this in (3):

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

giving

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

Similarly, from (4)

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

Putting this in (3):

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

giving

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

Notice that, for a prospective study which gives RRs by exposure level, the calculations for both A_i and B_i involve dividing by:

$$\left(V_i + \frac{1}{A_0} + \frac{1}{B_0}\right)$$

This set of values is calculated under the title V_{extra}. Notice that these V_{extra} values are different for the different types of study.

7 Appendix C: Formulae used in the calculation of Heterogeneity and Trend (Breslow 1980)

7.1 Case control and cross-sectional studies

The data are in the format:

	Exposure level				
	1	2	…	K	Totals
Cases	a_1	a_2	…	a_k	n_1
Controls/Non-cases	c_1	c_2	…	c_k	n_0
Totals	m_1	m_2	…	m_k	N
Dose levels	x_1	x_2	…	x_k	

Heterogeneity is assessed using formula (4.38) from §4.5 of Breslow and Day (Volume 1) [2].

$$\chi^2_{K-1} = (N-1)\left(\frac{1}{n_1} + \frac{1}{n_0}\right)\sum_{k=1}^{K}\frac{(a_k - e_k)^2}{m_k}$$

where the expected value is

$$e_k = E(a_k) = \frac{m_k n_1}{N}$$

and χ^2_{K-1} is a Chi-squared statistic on K-1 degrees of freedom.

Trend is assessed using formula (4.39) of the same volume:

$$\chi^2_1 = \frac{N^2(N-1)\{\sum_{k=1}^{K} x_k(a_k - e_k)\}^2}{n_1 n_0\{N\sum_{k=1}^{K} x_k^2 m_k - (\sum_{k=1}^{K} x_k m_k)^2\}}$$

with e_k as above.

7.2 Prospective studies

The data are in the format:

	Exposure level				
	1	2	…	K	Totals
Deaths	d_1	d_2	…	a_k	D
Alive at follow-up	$n_1 - d_1$	$n_2 - n_2$	…	$n_k - d_k$	$N - D$
Totals	n_1	n_2	…	n_k	N
Dose levels	x_1	x_2	…	x_k	

Section 3.6 of Breslow and Day (Volume 2) [3], page 107 states:

"The cohort statistics are simpler because one does not need to consider the marginal totals $d_{jk} + n_{jk}$ at all. By substituting n_{jk} for both c_{ki} and m_i, N_j for both n_{0i} and N_i and d_{jk} for a_{ki}, many of the statistics developed in §4.5 of Volume 1 are converted into precisely the form needed for cohort analyses." (Note that the *i* and *j* subscripts represent the strata for stratified analyses.)

Thus, Heterogeneity is assessed using a modified form of (4.38) of Breslow and Day (Volume 1) [2].

$$\chi^2_{K-1} = (N-1)\left(\frac{1}{D} + \frac{1}{N-D}\right)\sum_{k=1}^{K}\frac{(d_k - E_k)^2}{n_k}$$

where the expected value is

$$E_k = \frac{n_k D}{N}$$

Similarly, trend is assessed using formula (4.39) of the same volume:

$$\chi_1^2 = \frac{N^2(N-1)\{\sum_{k=1}^{K} x_k(d_k - E_k)\}^2}{D(N-D)\{N\sum_{k=1}^{K} x_k^2 n_k - (\sum_{k=1}^{K} x_k n_k)^2\}}$$

8 Appendix D: Trend (rate of increase per unit dose) using the dose levels entered

As stated in section 2.4, this uses the method described by Berlin et al. (1993) [4] together with the correction for the non-independence of results by exposure level described by Greenland and Longnecker (1992) [5]. These sources describe the method for case-control and cross-sectional studies. Orsini et al. (2012) [6] provides the modifications to be used for prospective studies.

Briefly, the method is as follows:

- A 2xK table of cases and controls/non-cases/at risk is required. Our table of effective numbers is appropriate for this.
- The variance of the result for each exposure level is estimated using the width of its confidence interval, as in equation 0 of Appendix A.
- The table of effective numbers is used to estimate the correlation of pairs of results. These values are used, together with the variance values, to estimate the covariance of each pair of results.
- The variance-covariance matrix is inverted and used, together with the $\log_e$ RR and dose values, to estimate beta, the change in log relative risk per unit increase in dose (the slope of the relationship between dose and $\log_e$ RR) and its variance.
- The value of beta and its variance are exponentiated to give the rate of increase in risk per unit increase in dose. These values are presented on the spreadsheet.

The method assumes that the unexposed group has a dose value of zero. The spreadsheet allows the user to enter a non-zero value for unexposed dose. When this has been done, the calculations subtract the unexposed dose value from each of the dose values entered and shows these adjusted dose values in column AA of the spreadsheet. These adjusted values are used to estimate beta. Note that subtracting the same value from each dose value does not change the slope of the relationship, so the estimate of beta is unaffected.

These calculations are coded using VBA in the form of macros associated with the spreadsheet. This is necessary because of the need to invert a matrix of variable size, dependent on the number of exposure levels included. Using VBA has the benefit of allowing more complex data checking and error reporting than is possible within Excel itself.

9 Appendix E: Trend (rate of increase per unit dose) using the 'Uniform scale'

Some studies report ordered non-numeric categories such as 'None', 'A little' and 'A lot'. For these no meaningful dose values can be derived.

To provide an indication of trend for this type of data in a manner that allows comparability between studies, 'Uniform scale' results are given. For this, the dose values entered are ignored. Instead, scale values are derived and used in place of dose values.

Scale values are based on the best available estimate of the distribution of the population studied. This is represented by the control participants for a case-control study, the overall participants (cases plus non-cases) for a cross-sectional study, and the at risk population for a prospective study. Values in the table of effective numbers of participants are used in these calculations.

We consider an underlying scale, where 0 indicates the least possible exposure and 1 the most possible exposure, and assume that the participants are ordered by level of exposure. If there are N participants from the population of interest (controls, overall population or at risk population depending on the study design), the individual with the lowest exposure is assumed to come from the lowest Nth of the population with doses ranging from 0 to 1/N (mean 0.5/N), the individual with the next lowest exposure to come from the next lowest Nth with doses ranging from 1/N to 2/N (mean 1.5/N) and so on. Thus if there are, say, three exposure groups in a case-control study, with effective numbers of N_1, N_2 and N_3 (summing to N), the corresponding doses assigned for the three groups will be 1/N multiplied by, respectively, $N_1/2$, $N_1+N_2/2$ and $N_1+N_2+N_3/2$. Thus for example if we have $N_1=50$, $N_2=30$ and $N_3=20$, the mean doses in the three groups will be 0.25, 0.65 and 0.90.

This process necessarily produces a non-zero "dose" for the unexposed category. The "dose" values are therefore adjusted, as before (see Appendix D), by subtracting the value for the unexposed category from each of the values.

Both the original 'uniform scale' values and the values adjusted by subtraction are shown on the spreadsheet.

The rate and its confidence interval are then calculated in the same way as before (see Appendix D) except that the 'Uniform scale' values are used instead of the entered dose values.

10 References

1. Hamling J, Lee P, Weitkunat R, et al. 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. *Stat Med* 2008; 27: 954-970. DOI: 10.1002/sim.3013.

2. Breslow NE and Day NE. *The analysis of case-control studies*. Statistical methods in cancer research, Vol 1, Davis W (ed). Lyon: IARC, 1980. IARC Scientific Publication No. 32, p.338.

3. Breslow NE and Day NE. *The design and analysis of cohort studies*. Statistical methods in cancer research, Vol 2. Lyon: International Agency for Research on Cancer, 1987. IARC Scientific Publication No. 82, p.406.

4. Berlin JA, Longnecker MP and Greenland S. Meta-analysis of epidemiologic dose-response data. *Epidemiology* 1993; 4: 218-228. DOI: 10.1097/00001648-199305000-00005.

5. Greenland S and Longnecker MP. Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. *Am J Epidemiol* 1992; 135: 1301-1309.

6. Orsini N, Li R, Wolk A, et al. Meta-analysis for linear and nonlinear dose-response relations: examples, an evaluation of approximations, and software. *Am J Epidemiol* 2012; 175, 1: 66-73.

7. Lee PN. Simple methods for checking for possible errors in reported odds ratios, relative risks and confidence intervals. *Stat Med* 1999; 18: 1973-1981.