RREst – Relative risk estimation program

Author: Jan Hamling

Date: 28 August 2007

T:\Jan\RREst\RREst9_Aug07.doc

CONTENTS

1. Introduction

RREst is a spreadsheet program to manipulate non-independent relative risks (RRs) [1] and confidence intervals (CIs) to give the RR and CI of a comparison different from the ones provided. For example, the report of an epidemiological study may give the RR and CI for several categories of exposure, each compared with a single unexposed group. The user may need the overall RR and CI for exposure versus no exposure. Standard meta-analysis techniques cannot be used because the RRs are not independent.

[1] The term 'relative risk' is intended to describe any of the risk ratio measures such as odds ratio.

The features of the program are:

- Specification of the contrast required (i.e. specifying which of the available categories make up the baseline and which the comparison group);
- Estimation of RR (CI) for the specified contrast;
- Estimation of heterogeneity among the selected categories;
- Specification of trend coefficients (dose values);
- Estimation of trend according to the trend coefficients among the selected categories;
- Case control and prospective studies are handled;
- Categories of exposure or of disease are handled;
- Results are displayed to the required number of decimal places (up to 6) (but it is not usually appropriate to quote the resulting RR (CI) to more than one or two decimal places). Results are rounded values;
- Results are calculated to a specified precision. This can be adjusted easily if convergence problems arise.

The method used involves estimating the numbers of subjects associated with each of the available categories. It is then a simple process to calculate the RR and CI of an alternative comparison, simply by combining the estimated numbers of subjects into a baseline group and a comparison group. Where the study provides adjusted RRs, the estimated numbers are ‘effective’ numbers of subjects – the number of subjects from a population equivalent to a notional ‘adjusted’ population.

2. Using the program

First impressions

The program uses macros to perform the estimations needed. Select ‘Enable macros’ when the spreadsheet is first opened.

Most of the input cells and the results cells are visible when the spreadsheet is first opened, and the whole spreadsheet prints on one double-sided A4 page.

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 total the number of exposed subjects) can be entered as calculations.

The display-only cells are protected so that 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 the cells used in the underlying calculations and some instructions on using the spreadsheet.

The spreadsheet is presented as a ‘read-only’ file, which prevents accidental overwriting of the original. If the spreadsheet is to be saved once data have been entered, it should be saved under a new name.

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 will be 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 boxes.

Below the results are two drop-down boxes. The first of these allows the study type (case control or prospective) to be chosen. The second is used to specify how the original 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 subjects whereas a prospective study involves subjects 'at risk'. 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 data***.

The next group of input fields makes up a 2x2 table of initial values/estimates for the numbers of cases and controls (or 'at risk' subjects) according to their being exposed or unexposed to the risk factor.

The RRs and CIs to be analysed are entered below this. Up to 29 comparison levels can be entered. A description can be entered for each one in the column headed 'Category'. The first category represents the baseline of the original data, with the RR set to 1. The lower and upper limits of the CIs are entered in the columns 'Lower' and 'Upper' respectively.

The column headed 'Contrast' is used to specify how the program should combine the data when calculating the Results fields for RR and lower and upper CI, i.e. which categories should be grouped to form the analysis baseline and which the comparison group. It is also used to exclude categories from the calculation of results (RR (CI), Heterogeneity and Trend calculations). Enter the values:

- 0 for categories that will form the RR analysis baseline
- 1 for categories that will form the RR comparison group
- -1 for categories that should be ignored in the calculation of results.

The column headed ‘Dose’ is used to specify trend coefficients for the trend analyses. For each category enter a positive value to be used as a trend coefficient (dose level) for the category.

To the right of these input columns the program will generate two columns of estimated numbers of subjects. These columns will show estimates for each Category entered. To make the program calculate these values, click the ‘Solve’ button. This runs an iterative process which finds a ‘best’ solution based on the values entered in the 2x2 table and in the RR, Lower and Upper columns (e.g. the first row of the 2x2 table will be used as the initial estimates for the first row of estimated numbers) – see Section 3 and Appendices B and C for details.

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

Sequence of events

In Excel click “Open” and select the RREst9 Excel file. Click “Enable macros”. The spreadsheet will then be displayed.

If required, enter text in the ***Heading*** area.

As was mentioned above, appropriate drop-down options should be chosen for ***Study type*** (case control or prospective) and ***Categorised by*** (by levels of exposure or by categories of disease) before the data are entered.

Values should then be entered in the 2x2 table of ***No. of subjects*** and in the columns ***Category, RR, Lower*** and ***Upper*** (these last two representing the CI) and ***Dose***. ***Contrast*** values can also be entered.

Next, click the ***Solve*** button to generate the columns of estimates of the numbers of subjects for each category (the effective numbers of subjects). These will be used in all subsequent calculations. A results window will pop up with the text:

Solver found a solution or

Solver could not find a feasible solution.

Click OK. (If the second of these appears, see section 4 below for guidance.)

The user can then enter one or more ***Contrast*** specifications in turn, because the related calculations use only the values already generated by the Solve process. Each time a Contrast specification is entered, the ***Results: RR (CI), Heterogeneity and Trend*** are recalculated immediately.

Note that, because the Contrast, Heterogeneity and Trend calculations make use of the columns of estimated numbers of subjects, the Results values are not valid until the Solve button has been used.

Note also that a prospective study giving RRs by disease level will have the ‘at risk’ category as the first level in the Category/RR(CI) columns. This category includes all the subjects in the study so it makes no sense to specify additional baseline levels. The program will always use 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.

3. Troubleshooting: Excel can't find the Solver

The Solve button makes use of an Excel feature called Solver. If this feature was not included when Excel was installed then clicking the Solve button will generate a message such as:

"Compile Error: Sub function not defined"

If this happens it is necessary to install the Solver feature.

To install the Solver:

1. Start Excel.
2. Click Tools: Add-Ins
3. Tick the box next to "Solver Add-in" and click OK

The system will then probably ask you to insert the MS Office CD so that the relevant code can be installed. Follow instructions as they appear on the screen.

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

1. Start Excel
2. Open RREst
3. Click Tools: Protection: Unprotect sheet
4. Click Tools: Macros: Visual Basic Editor
   A new window will appear showing Visual Basic code in sub-windows.
5. Click Tools: References
6. Tick the box next to SOLVER.
7. If SOLVER is listed as MISSING, then:
   Click Browse
   Find and add the file Solver32.dll or SOLVER.xla, 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.xla

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

Note that Excel may ask for a .dll file or for a .xla file. Note which of these Excel is asking for. We have found that, even when a .xla file is needed, the Browse system may show only files of type:

Type Libraries (*.olb, *.tlb, *.dll)

If this happens, click the drop-down against ‘File of type:’ and choose ‘All Files (*.*)’. It will then be possible to see and select files of type .xla.

8. Check that SOLVER is now listed and ticked. Click OK.
9. Click File: Save.
10. Click File: Close and return to Microsoft Excel.
11. In Excel click Tools: Protection: Protect sheet: OK.

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

Occasionally clicking the Solve button will give the message:

"Solver could not find a feasible solution"

There are two reasons why this can happen:

- 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 more common than the second. To allow the user to investigate whether this is the situation, the spreadsheet provides a drop-down box showing the ***Solver Precision*** setting used and allowing the user to change the setting. Whenever the Precision setting is changed the Solve process is performed automatically, so the new estimates are displayed and the Solver result window is shown.

If adjusting the Precision setting doesn't give a feasible solution, then the second reason may apply. It is worth checking the data values entered. It is easy to make a data entry error. It is also not uncommon for there to be typographical errors in published study reports. There are a number of simple checks that can be used to identify possible errors in reported odds ratios, relative risks and confidence intervals [1]. Similarly, the numbers of subjects included in the published analyses are sometimes poorly or misleadingly reported.

If this does not solve the 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 in the extreme bottom right of the spreadsheet),
- Calculate the 'Start values' (shown just above the columns of Estimated numbers of subjects).

For a case control study with RRs given by levels of exposure, P represents the proportion of unexposed subjects 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 in the extreme bottom right of the spreadsheet. These values give a measure of the accuracy of the iterative process performed when the Solve button is pressed. They are described in more detail below.

The Solve process works by:

- Copying the pair of 'Start values' into the first line of the columns of Estimated numbers.
- Calculating the other values in the columns of Estimated numbers using these first line values and the 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:

$$\text{Sum of Squares} = \left(\frac{P - P'}{P}\right)^2 + \left(\frac{Z - Z'}{Z}\right)^2$$

  This is a measure of the extent to which the estimates (P' and Z') differ from the specified values (P and Z).
- Repeatedly adjusting the values in the first line of the table of Estimated numbers (and recalculating) until a small enough value of Sum of Squares is given.

For some studies the Solve 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 line of the 2x2 table.

To overwrite the Start value cells with different values it is necessary to turn off worksheet protection. To do this, click:

Tools : Protection : 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 and click the Solve 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:

Tools : 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’.

APPENDIX B: The equations to be solved and the process involved

The notation

A_0, B_0	The pair of numbers in the first line of the table of Estimated numbers of subjects. For a study giving RRs by exposure levels these represent the number of unexposed cases and the number of unexposed controls/'at risk' subjects respectively.
A_1 to A_n	The number of subjects in the first column of the table of Estimated numbers for level 1 onwards. For a study giving RRs by exposure levels these represent the estimated number of exposed cases in each level of exposure (1 to n).
B_1 to B_n	As A_1 to A_n but for the second column of the table of Estimated numbers of subjects. For a study giving RRs by exposure levels these represent the estimated number of exposed controls/ 'at risk' subjects in each level of exposure (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 RRs by exposure levels this represents the proportion of unexposed subjects among the controls/'at risk' subjects.
Z	Value calculated from the totals of the 2x2 table (second column total over first column total). For a study giving RRs by exposure levels this represents the relative frequency of controls/'at risk' subjects to cases.
P'	Estimated value of P derived from the table of Estimated numbers of subjects.
Z'	Estimated value of Z derived from the table of Estimated numbers of subjects.
R_i	Relative risk for level i (entered by the user).
L_i	Lower value of confidence interval for level i (entered by the user).
U_i	Upper value of confidence interval for level i (entered by the user).
V_i	Variance of $\log_e (R_i)$ for level i.

A_b	Total estimated number of cases/exposed subjects in the levels chosen to be the baseline group (as defined by the entries in the contrast column).
A_c	Total estimated number of cases/exposed subjects in the levels chosen to be the comparison group (as defined by the entries in the contrast column).
B_b	Total estimated number of controls('at risk')/unexposed subjects in the levels chosen to be the baseline group (as defined by the entries in the contrast column).
B_c	Total estimated number of controls('at risk')/unexposed subjects in the levels 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).

The equations

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

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

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

The estimate of P calculated using the table of estimated numbers of subjects.

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

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

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

The relationship between the relative risk for level i and the Estimated numbers of subjects (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 subjects (A_i and B_i).

For case control 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 A_i and B_i (see Appendix C).

5) Generally 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 required contrast (as equation (4) but for the requested comparison):

For a case control 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' subjects 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) $\log_e(L) = \log_e(R) - 1.96\sqrt{V}$

<u>The steps involved in the solution</u>

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

When the user clicks Solve:

- 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 subjects. 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/'at risk' subjects).

- The spreadsheet calculates:

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

  For a case control 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 subjects 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 subjects for the requested comparison (A_b, A_c, B_b and B_c) using the contrast definition and the estimated numbers of subjects 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(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: the cell containing the Sum of Squares value calculated at (j) above

  Target value: 0

Variable cells: the first two cells in the table of Estimated numbers of subjects (which contain the values of A_0 and B_0)

This routine 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 the requested comparison.

Notice that the calculation of the contrast ((c)-(g) above) does not affect the results of the Solver routine. However, the results of the comparison 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 the Solver has produced a solution, a range of comparisons can be generated simply by changing the values in the 'Contrast' column.

APPENDIX C: Calculating values for estimated numbers of subjects

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

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

(3) $$R_i = \frac{A_i B_0}{B_i A_0}$$

(4) For case control 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 which give 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 which give RRs by disease category:

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

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 subjects we need A_i independently of B_i and B_i independently of A_i.

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 studies, for prospective studies which give RRs by exposure level and for prospective studies which give RRs by disease category.

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

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

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

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

Similarly, from (4) $$A_i = \frac{1}{\left(V_i - \frac{1}{A_0} - \frac{1}{B_0} - \frac{1}{B_i}\right)}$$

In (3) $$R_i = \frac{B_0}{A_0 B_i \left(V_i - \frac{1}{A_0} - \frac{1}{B_0} - \frac{1}{B_i}\right)}$$

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

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

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

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}.

Prospective studies giving RRs by exposure level

From (4):

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

Putting this in (3) $$R_i=\frac{A_iB_0\left(-V_i+\frac{1}{A_0}-\frac{1}{B_0}+\frac{1}{A_i}\right)}{A_0}$$

$$R_i=\frac{A_iB_0\left(-V_i+\frac{1}{A_0}-\frac{1}{B_0}\right)+\frac{A_iB_0}{A_i}}{A_0}$$

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

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

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

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

Similarly, from (4) $$A_i = \frac{1}{\left(V_i - \frac{1}{A_0} + \frac{1}{B_0} + \frac{1}{B_i} \right)}$$

In (3) $$R_i = \frac{B_0}{A_0 B_i \left(V_i - \frac{1}{A_0} + \frac{1}{B_0} + \frac{1}{B_i} \right)}$$

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

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

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

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.

Prospective studies giving RRs by disease category

From (4):

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

Putting this in (3) $$R_i = \frac{A_i B_0 \left(V_i + \frac{1}{A_0} + \frac{1}{B_0} - \frac{1}{A_i} \right)}{A_0}$$

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

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

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

Similarly, from (4) $$A_i = \frac{1}{\left(V_i + \frac{1}{A_0} + \frac{1}{B_0} - \frac{1}{B_i} \right)}$$

In (3) $$R_i = \frac{B_0}{A_0 B_i \left(V_i + \frac{1}{A_0} + \frac{1}{B_0} - \frac{1}{B_i} \right)}$$

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

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

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

Notice that, for a prospective study which gives RRs by disease category, 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}. The V_{extra} values are different for the different types of study.

APPENDIX D: Formulae used in the calculation of Heterogeneity and Trend

Case control studies

The data are in the format:

	Exposure level				
	1	2	…	K	Totals
Cases	a_1	a_2	…	a_K	n_1
Controls	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 [3].

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

where the expected value $e_k = E(a_k) = \dfrac{m_k n_1}{N}$

and χ^2_i = Chi-squared statistic on *i* degrees of freedom.

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

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

with e_k as above.

Prospective studies

The data are in the format:

	Exposure level				
	1	2	…	K	Totals
Deaths	d_1	d_2	…	d_K	D
Alive at follow-up	$n_1 - d_1$	$n_2 - d_2$	…	$n_K - d_K$	$N - D$
At Risk	n_1	n_2	…	n_K	N
Dose levels	x_1	x_2	…	x_K	

Section 3.6 of Breslow and Day Volume 2 [4], 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 stratum for stratified analyses.)

Thus, Heterogeneity is assessed using a modified form of (4.38) of Breslow and Day Volume 1 [3].

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

where the expected value $E_k = \dfrac{n_k D}{N}$

Similarly Trend is assessed using a modified form of (4.39) of the same volume:

$$\chi^2_1 = \frac{N^2(N-1)\left\{\sum_{k=1}^{K} x_k(d_k - E_k)\right\}^2}{D(N-D)\left\{N\sum_{k=1}^{K} x_k^2 n_k - \left(\sum_{k=1}^{K} x_k n_k\right)^2\right\}}$$

REFERENCES

1. Lee PN. Simple methods for checking for possible errors in reported odds ratios, relative risks and confidence intervals. *Stat Med* 1999;**18**:1973-81.

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. Breslow NE, Day NE. *The analysis of case-control studies*, Volume 1. Lyon: IARC; 1980. (Davis W, editor. Statistical methods in cancer research.) IARC Scientific Publication No. 32.

4. Breslow NE, Day NE. *The design and analysis of cohort studies*, Volume 2. Lyon: International Agency for Research on Cancer; 1987. (Statistical methods in cancer research.) IARC Scientific Publication No. 82.