WINPEPI PROGRAMS

PAIRSetc

MANUAL

Version 3.59

Revised Aug. 21, 2016

PAIRSetc is a WINPEPI program (Abramson 2004, 2011), part of the PEPI suite of computer programs for epidemiologists. ("PEPI" is an acronym for "Programs for EPIdemiologists".)

PAIRSetc provides procedures for use in comparisons of paired and other matched observations, appraising their differences and agreement. The "etc" in its name indicates its ability to deal with matched sets larger than pairs. It may be used for analyses and meta-analyses of cross-sectional, cohort or case-control studies, and trials, and in reliability studies. It can analyse stratified data. There are 34 modules to choose from.

Analysis of paired observations

Analysis of sets of 3 or more matched observations

Analysis of sets of varying numbers of matched observations

Effect of misclassification on paired dichotomous data:

Power

Sample sizes

HOW TO USE WINPEPI: an ABC

A. Obtain the latest version

The latest set of WINPEPI programs and manuals can be downloaded free from www.brixtonhealth.com.

B. Install

Run *winpepisetup.exe*. This will put the programs and manuals in a folder of your choice (replacing any previous versions in that folder) and will place a WINPEPI portal (a "WINPEPI" icon) on your desktop. It may be convenient to pin the Portal to the Start menu or the Taskbar.
If you downloaded *winpepifiles.zip*, you will have to copy its contents to a folder of your choice, and manually put a shortcut to *winpepi.exe* on your desktop.

C. Use the WINPEPI Portal and find the procedure you want

There are seven WINPEPI programs: DESCRIBE (for descriptive epidemiology) COMPARE2 (to compare two independent groups or samples), PAIRSetc (to compare matched observations). LOGISTIC and POISSON (for multiple logistic and Poisson regression), WHATIS (various utilities, including a calculator), and ETCETERA (miscellaneous procedures). Each program has a number of modules (over 120 in all), each of which offers a number of statistical procedures.

Open the WINPEPI Portal, which provides access to all the programs and manuals, and to WINPEPI's Finder, which is an alphabetical index to the statistical procedures. The Portal also provides access to a published overview of the programs and their learning/teaching potential, and to the web-site offering the latest version of WINPEPI. Among other options, it provides a magnifying glass, for users with poor vision or small monitors. The Finder can also be accessed (in any WINPEPI program) by pressing F9 or clicking on "Winpepi".

If you know what program and module are required, open the program by clicking on it in the Portal. Otherwise, search the Finder for the procedure you require. The Finder will tell you what module to use.

THE ESSENTIAL REQUIREMENT IS THAT YOU SHOULD KNOW WHAT YOU WANT.

If you open the Finder and search for "*Multiple linear regression*", for example, you will be directed to ETCETERA J, i.e. to module J of ETCETERA . You would then open ETCETERA and click on J.

You may be offered alternatives. For an *equivalence test for proportions*, for example, the FINDER will say "COMPARE2 A, PAIRSetc A", i.e., either module A of COMPARE2 or module A of PAIRSetc. If the observations are independent, COMPARE2 is appropriate; if they are paired, PAIRSetc is appropriate.

You may have to open the programs to find precisely what each module offers. For example, a search for "*Diagnostic tests, of*", will direct you to "DESCRIBE L1, L4, L5, PAIRSetc D1, D2, D3". When you open DESCRIBE, clicking on "L" will reveal that module L1 refers to "Yes/No" tests, and L4 and L5 to tests with a range of results. In PAIRSetc, modules D1, D2 and D3 (respectively) are appropriate for comparing normally-, log-normally-, or non-normally-distributed results with a gold standard.

It is unwise to use a statistical procedure whose use one does not understand. This manual cannot supply this knowledge, and it is certainly no substitute for the basic understanding of statistics and epidemiological thinking that is essential for the wise choice of methods and the correct interpretation of their results.

D. Open the WINPEPI program and select a module

Open the selected program, via the Portal or by clicking on its icon or name in Explorer.

You will generally be presented with a menu, from which you should make a selection. Some options may be offered in the horizontal menu at the top of the opening screen.

A data-entry screen will then appear. You may be asked to make a further choice before entering the data, and various options may be offered At each stage, simple instructions are provided (in yellow); pop-up hints may be shown. Additional help may be obtained by pressing F1 or clicking on "Help" in the top menu. For further information, the program's manual can be accessed by clicking on "Manual" in the top menu.

E. Enter the data

Two of the programs can read data files. But in most instances, data must be entered at the keyboard or pasted from a text file or spreadsheet. This usually requires prior counting and summarization, either manually or by using statistical software that processes primary data.

Manual entry of data is usually easy. If entries are required in different boxes, pressing *Enter* or *Tab* after entering a number will generally take you to the next box; and pressing *Escape* will clear the entry. If several entries are required in the same box, press *Enter* or *Space* after each entry.

Pasting data: If the data are available in a text file (created, for example, by Notepad or Microsoft Word) or a spreadsheet, they can be copied to the Windows clipboard [usually by pressing *Ctrl-Insert* or *Ctrl-C*], and then pasted into a data-entry box [usually by pressing *Shift-Insert* or *Ctrl-V*]. This can simplify data entry in boxes that require a number of entries (in rows or columns). [Also, data can be copied from a data-entry box and pasted to a text file for future re-use; first, press *Ctrl-A* to mark it for copying.] The following instructions can be accessed by pressing F2 (in any WINPEPI program) or clicking on *"Help – Pasting"*.

Precautions:

- The data must be pasted into the box as a single block, and not piecemeal.
- There must be no missing values (e.g., empty cells in a spreadsheet).
- The data must be in the format required in the box, with spaces between the numbers; exact alignment of the columns is not necessary. For example

```
45 66  1
20  3 132
53 11  44
```

- If a defined number of rows is required, this number must be entered first, e.g. in the "Number of categories" box.
- If a column of row numbers is shown on the left (1, 2, etc.), ensure that the"1" is visible before pasting.
- The cursor must be in the top left corner of the box when the "paste" keys are pressed.

F. Run the program

G. Select the results you need

Do not be confused by the multiplicity of results. You can scroll down until you find the results you need; and ignore everything else. If you want an odds ratio and its confidence intervals, you can ignore all other results.

WINPEPI programs offer more options than most users will ever need, and will usually display more results than are needed. **IGNORE THE OPTIONS AND RESULTS YOU DON'T REQUIRE**.

On the other hand, you may find some of the other results helpful.

Very often, the program will provide alternative tests and measures of effect, often with confidence intervals estimated by alternative methods. If there is disagreement between the results, you may find appropriate advice in the manual, which describes the procedures and their uses and limitations, with literature references..

H. (Maybe) continue the analysis

After getting the first results, it may be decided to continue the analysis. It may, for example, be decided to repeat the analysis (by clicking on "Repeat") and make changes in the data or the options. After performance of a logistic regression analysis, options are offered for the use of the logistic coefficients to compute a probability, risk ratio, etc.

If *stratified dat*a are entered, clicking on "Next stratum" permits entry of another stratum, and clicking on "All strata" provides a combined analysis of all the strata. Similarly, a *meta-analysis* can be performed by entering a table for each study as a separate stratum, and then pressing "All strata". (If summary data (e.g. risk rauos) are available for each study, a series of tables is not needed; module I of COMPARE2 might then be used.)

I. Saving the results

By default, all results (except graphs) are automatically saved in *pepi.txt* in the Winpepi folder, with a warning if its size exceeds 500K. This file can be accessed via the Portal. The default procedure can be viewed or changed by clicking on "Saving" in the top menu; this also provides access to *pepi.txt*. Optionally, graphs can be saved as BMP files.

Results produced during the current session are also saved (temporarily) in *pepi.tm*p, which can be viewed by clicking on "View" in the top menu.

The results of a single analysis can be saved (in a new file) by clicking on "Print or save" or "Print".

J. Adding comments

Click on "Note" (in the top menu) to add a note to the previously-shown results, for saving with the results in *pepi.txt*.

K. Printing the results

The results of an analysis can be printed by clicking on "Print or save" or "Print". Graphs can be printed at low or high resolution. Also, selected results can be printed from *pepi.txt*.

L. Pasting the results to a text file

All results shown on the screen are automatically copied to the Windows clipboard, from which they can be pasted into a Microsoft Word or other text file (preferably for display in a Courier or similar font, to ensure proper alignment of tabulated results). Optionally, graphs can be copied to the clipboard, replacing any results. that are there.

Notes

The programs are 32-bit applications, written with Delphi 5, and will run in any version of Microsoft Windows (including Windows 7), except Windows 3. They can be run from a portable device such as a USB flash drive.)

The manuals that accompany the programs require a PDF reader, such as Adobe Acrobat or Foxit Reader.

The programs and manuals refer to dichotomous variables as "Yes-No" variables, and to interval- or ratio-scale variables as "numerical".

P-values derived from *z* and *t* functions are generally correct to five decimal places, those based on *chi*-square, to four decimal places, and those based on the *F* function to three decimal places.

WINPEPI does not adhere strictly to the conventional definitions of "*risk"* (ratios with count denominators. e.g. prevalence) and *"rate*" (ratios with person-time denominators, e.g incidence density), except when the distinction is important. Risks may be referred to as rates when this is unlikely to cause confusion.

A DO-IT-YOURSELF THREESOME

1. **PLANNING A STUDY**: "Research Methods in Community Medicine: Surveys, Epidemiological Research, Programme Evaluation, Clinical Trials" (J.H. Abramson and Z.H. Abramson), sixth edition, 2008. John Wiley & Sons.

2. **ANALYSING THE FINDINGS**: The WinPepi suite of computer programs for epidemiologists, with their manuals. Can be downloaded free from www.brixtonhealth.co

3. **INTERPRETING THE RESULTS**: "Making Sense of Data: A Self-Instruction Manual on the Interpretation of Epidemiological Data" (J.H. Abramson and Z.H.Abramson), third edition, 2001. Oxford: Oxford University Press.

Acknowledgements

Acknowledgements are due to the late Eric Peritz, who collaborated in developing the original hand-held calculator programs, to Paul Gahlinger, who wrote the early DOS-based programs, to Kevin Sullivan, who suggested the creation of a Windows version, to Garry Anderson, who keeps a friendly but strict eye on quality and accuracy, and to Mark Myatt, Ray Simons, **Bud Gerstman,** and other colleagues and users for their suggestions, criticism, and practical help.

Wilko C Emmens's XYgraph unit (version 2.2) creates the graphs displayed by WINPEPI programs.

PAIRSetc's MODULES : A GUIDE

PAIRSetc provides procedures for use in comparisons of paired and other matched observations, appraising their differences and agreement. The "etc" in its name indicates its ability to deal with matched sets larger than pairs.

Modules A to D compare *paired observations*:

- **Module A** deals with *"Yes-No" (dichotomous) variables*.
- **Modules B and C** deal with variables with *three or more categories* (**B** for nominal categories and **C** for ordered categories.
- **Module D** deals with *numerical (interval-scale or ratio-scale) variables* (**Module D1** for normally-distributed variables, **Module D2** for log-normally distributed variables, and **Module D2** if normality is not assumed).

Modules E to I analyse *larger matched sets* with the same number of observations per set):

- **Modules E and F** for *"Yes-No" variables* (**Module E** for case-control studies using multiple matched controls, and **Module F** for other studies).
- **Module G** for *nominal or ordinal-category variable*s.
- **Modules H and I for** *numerical (interval-scale or ratio-scale) variables* (**Module H** for comparisons of two groups or methods, and **Module I** for comparisons of 3 or more .matched samples or replicates).

Modules J to M analyse matched sets of observations varying in size:

- **Modules J and K for** *"Yes-No" variables* (**Module J** for *case-control studies* with varying numbers of matched controls, and **Module K** to compute *kappa*).
- **Modules L and M** deal with *numerical (interval-scale or ratio-scale) variables* (**Module L** for comparisons of two matched groups[**Module L1**] or two methods of measurement [**Module L2**], and **Module M** for comparisons of different-sized sets of replicate measurements).

Modules Mis1 to Mis3 (accessed by clicking on "Misclass" in the top menu) appraise the possible effect of *misclassification o*n a paired 2 x 2 table.

M**odules P1 to P3** (accessed by clicking on "Power" in the top menu) estimate the *power* of various tests.

Modules S1 to S7(accessed by clicking on "Sample size" in the top menu) estimate the *sample sizes* required for various tests.

Kappa is computed by **Modules A, B, C, E , G, J, and K**.
Replicate numerical observations are compared by **Modules D, I, and K**.
Methods of measuring a numerical variable are compared by Modules **D, H, and L2**.

The options include:

Analysis of incompletely paired data (in Modules A and D1)
Analysis of crossover trial (in Module A)
Measures of predictive accuracy (in Module A)
Reconstruction of paired 2x2 table from odds ratio or P-value (in Module A).
Assessment of intrarater and interrater agreement (in Modules A2 and I2)
Kappa for binocular data (in Module B)
Regression to the mean (assessment and adjustment) (in Module D6).
Analysis of clustered data (in Modules A and D1)
Measures of disagreement (in Module I1).

A. PAIRED OBSERVATIONS: "YES-NO" (DICHOTOMOUS) VARIABLE

This module is appropriate for the analysis of paired observations (in different subjects or the same subject), where the dependent variable is a dichotomy ("yes-no"). It appraises differences and agreement. It can analyse matched-control trials and matched case-control studies, before-after studies, and other comparisons of paired subjects or observations, such as comparisons of husbands and wives, or of diagnoses made by two different observers or procedures. It can handle clustered and stratified data and data collected by *inverse sampling*. An option is offered for the *reconstruction* of the paired 2x2 table (based on a *P* value or odds ratio), for use in incompletely reported studies.

The numbers of pairs with each combination of findings are entered in a 2 x 2 table in which A and B are the paired sets, and "yes" refers to the presence of the characteristic under study; in a case-control-study, "yes" usually refers to exposure to a risk factor or protective factor. Numbers of pairs are entered, not numbers of observations. An option is offered for the entry of supplementary unpaired observations, which are then also used in the analysis.

The controls in a case-control study or trial, and the unexposed in a cohort study should be designated "B". To test for *equivalence*, the bounds of "equivalence" must be defined, by specifying the largest difference that is to be regarded as negligible (e.g. 0.05).

If the data are stratified, enter each stratum in turn. For *meta-analyses*, enter each study as a separate stratum. If there are *clusters* of paired observations that may not be independent (e.g. various pairs of observations of the same person, or by the same observer), enter each cluster as a separate stratum. Click on "All strata" whenever combined results are required.

For each table, the program provides **tests for the difference** between the paired observations, a **test of equivalence** (optional), the **odds ratio** (with a low-bias estimator), the **proportions** (of "yes") **and their difference and ratio,** the **relative difference**, **correlation coefficients** (including ***phi,*** the approximate **tetrachoric correlation coefficient,** and **point-biserial correlation coefficients**), the **number needed to avoid one event**, **attributable or prevented fractions (**for paired case-control studies), ***kappa*****, percentage agreemen**t , and related results, **Gwet's AC1, Brennan and Prediger's G-index, Scott's *Pi* coefficient, Peirce's *I* coefficient, and Martin and Femia's delta-based measures of agreement,** the **intraclass odds ratio,** a measure of the **distinguishability of categories,** and the **probability and odds of replication**. Optionally, **measures of predictive accuracy** (in either direction) are displayed. For *stratified data*, the program provides overall tests for the difference, heterogeneity **tests and measures**, the overall **odds ratio**, and ***kappa* and related results**. Four sets of tests and measures are provided fo **clustered data**. A test and confidence intervals for the odds ratio are provided for studies using **inverse sampling**.

Optionally, this module can analyze a **crossover trial** with a "yes-no" outcome, comparing two treatments, A and B, applied in sequence to the same subjects, after random allocation of the subjects to an AB (Treatment A first) or BA (Treatment B first) group. Each group must be entered as a separate stratum. The program compares the **proportions of "yes"** for the different treatments and sequences, estimates confidence intervals for the difference between "yes" outcomes, and provides **odds ratios**, **tests for a period effec**t, **McNemar, Mainland-Gart, Prescott,** and **Schouten-Kester tests**, and the **number needed to treat.**

Tests for the difference between paired observations

For each table, and (if stratified data are entered) for the combined (pooled) data, the program provides Fisher's and mid-P exact tests (unless the user aborts their computation or numbers are very large), McNemar tests (with and without a continuity correction), and a modified Wald test for differences between A and B. Lui (2001b) recommends use of the McNemar test, uncorrected for continuity, rather than Fisher's exact test, which (like the corrected McNemar test) "can be quite conservative and hence lose much efficiency". The uncorrected McNemar test is more powerful, and performs well even when the number of discordant pairs is as low as 6. On the basis of a study of 9595 scenarios, Fagerland *et al.* (2013) recommend use of the McNemar mid-P test. The modified Wald test (May and Johnson 1997) is said to be valid in most data situations, and to be as powerful or more powerful than the McNemar test in small to moderate samples.

Heterogeneity tests and measures

For stratified data (i.e., a series of tables), the program provides *heterogeneity tests* that compare the odds ratios in the different strata, and the *kappa* values in the different strata. These permit appraisal of the modifying effect of the stratifying variable. The greater the similarity, the higher the P-value. The tests should be interpreted with caution, since their power is low; if the result is significant at the 0.05 level, the hypothesis of homogeneity can be rejected; but "a high p-value ... does not show that the measure is uniform, it only means that heterogeneity ... was not detected by the test" (Rothman and Greenland 1998: 276); the larger the strata, the more valid the test.

The program also provides two *measures of heterogeneity*, *H* and *I-squared* (Higgins and Thompson 2002), with their approximate 95% intervals. An *H* value of less than 1.2 suggests absence of noteworthy heterogeneity, whereas a value exceeding 1.5 suggests its presence, even if the heterogeneity test is not significant. *I-squared* expresses the proportion of variation that can be attributed to heterogeneity (in a meta-analysis, to interstudy variation) rather than to sampling error; a value greater than 50% may be considered substantial heterogeneity (Higgins and Green 2006).

Estimates of the supposed common underlying value of the odds ratio or *kappa* are of questionable value if the findings in the various strata are very disparate. If the results are not uniform, explorations of possible causes - e.g. associations with study design or quality or with the sizes or other characteristics of the samples - may be revealing

Test of equivalence

The program offers an equivalence test for the proportions of "yes" in two matched samples. This test may be appropriate if no statistically significant difference has been found, e.g. in "negative trials" that compare a new treatment with an established standard treatment, where there may be a reason to prefer the new treatment if it is at least as effective as the standard treatment.

To use the test, the bounds of "equivalence" must be defined by specifying the largest difference between proportions (e.g., 0.05) that is to be regarded as negligible.

Two hypotheses are tested: these are the hypotheses that there is more than a specified "negligible" difference in a specific direction – i.e. (a) that the first proportion is (more than negligibly) larger than the second proportion, and (b) that the second proportion is (more than negligibly) larger than the first proportion. If both tests yield significant results, both these hypotheses are rejected, and the results imply that both the one-sided differences are negligible - that is, the proportions are equivalent. If only one test is significant, this indicates that one proportion is at least as high as (i.e., "not inferior to") the other. The larger of the two P values is displayed as the P value for the equivalence test (Liu *et al*. 2002).

Non-significant results may be attributable to small sample size.

Standardized mean difference

The *population standardized mean difference* (Cohen's *d*) is the mean difference between two groups expressed in standard deviation units. Two estimates are provided - a logit-based estimate (1989) and a probit-based estimate (Glass et al. 1981). Computer simulations comparing 13 different methods of computing the standardized mean difference when the variable has been artificially dichotomized (Sanchez-Mecca et al. 2003) showed that these two estimates perform well, although both produce somewhat exaggerated variances.

Standardized mean differences provide a useful basis for meta-analyses if there are studies that present results based on dichotomization of a quantitative (continuous) outcome variable (in a 2x2 table format).

Odds ratio

The odds ratio is computed with its exact Fisher's and mid-P confidence intervals, unless numbers are very large or the computation is interrupted by the user, in which case Poisson-based confidence intervals are substituted. Wilson-score confidence intervals (recommended by Fagerland et al. 2014) are also computed. Jewell's low-bias estimator of the odds ratio (Jewell 1984) is shown. (Alternative confidence intervals are computed for studies using inverse sampling: see below.)

If stratified data are entered, a pooled odds ratio is computed for the combined data, with exact Fisher's and mid-P confidence intervals or Poisson-based confidence intervals.

Proportions, difference between proportions, ratio of proportions

The proportions of "yes" observations in the two samples, their absolute difference, and their ratio are displayed, all with their 90%, 95%, and 99% confidence intervals.

For the difference between proportions, the program displays intervals based on the score method (Newcombe and Altman (2000: 52), on improved Wald intervals (Agresti and Min 2005), and on an improved Wald method with Bonett-Price Laplace adjustment (Bonet and Price 2012; recommended by Fagerland *et al*. 2014).

For the ratio of proportions, Wald, Wilson, and Wilson-cc (continuity-corrected) intervals are provided; these intervals are very similar, unless the sample is very small. The Wilson intervals are as good as or better than the Wald intervals, according to a simulation study by Bonett and Price (2006), and their use is recommended by Fagerland *et al.* (2014). Bonett

and Price regard the continuity-corrected intervals as attractive because, although they tend to be wider than other intervals, their coverage probability cannot drop too far below the specified level of confidence.

The *proportions* of "yes" observations in the two samples, their absolute difference, and their ratio are displayed, all with their 90%, 95%, and 99% confidence intervals.

For the difference between proportions, the program displays intervals based on the score method (Newcombe and Altman (2000: 52), and improved Wald intervals (Agresti and Min 2005).

For the *ratio of proportions*, Wald, Wilson, and Wilson-cc (continuity-corrected) intervals are provided; these intervals are very similar, unless the sample is very small. The Wilson intervals are as good as or better than the Wald intervals, according to a simulation study by Bonett and Price (2006), who regard the continuity-corrected intervals as attractive because, although they tend to be wider than other intervals, their coverage probability cannot drop too far below the specified level of confidence.

Relative difference

The program computes the relative difference between the proportions, with its confidence intervals. This measure (Fleiss *et al.* 2003: 379-380) is defined as the difference between the numbers of "yes" responses in the samples, divided by the number of controls with "no" responses in sample B. It may be useful in the analysis of clinical trials in which a group receiving a new treatment (entered as sample A) is compared with a control group receiving a standard treatment (sample B). If "yes"' indicates a favourable response to treatment, the relative difference is a measure of the relative value of the new treatment, based on the assumption that the new treatment can benefit only those patients who fail to improve under the standard treatment. It is the proportion of subjects who are expected to respond to the new treatment, among those who fail to respond to the standard treatment (Lui 2004: 56).

Incompletely paired data

Optionally, the difference between proportions can be tested or its confidence intervals estimated even if pairing is incomplete , i.e. if some observations are paired and others are unpaired, for example because of refusals, recording errors, or drop-outs.

A *P value* is computed, based on the pooled results of significance tests of the paired data and the unpaired data, along the lines suggested by Kuan and Huang (2013), provided that the direction of the difference between proportions is the same in both sets of data, and provided that its calculation does not involve division by zero. Similarly, an *odds ratio* (with its confidence intervals) is estimated from the total set of observations, using a weighted average of the odds ratio estimators usually used for paired and unpaired observations (Miller and Looney 2012); this procedure requires nonzero entries in all cells. The above results are valid only if "missingness" is random and not influenced by membership of set A or B or by the value of the "yes-no" variable.

The procedure described by Tang et al. (2009) for the estimation of confidence interval is valid if "missingness" is random and not influenced by membership of set A or B or by the value of the "yes-no" variable; computer simulations show that if the sample is small or data

are sparse, the confidence interval may be unduly narrow and the confidence level may be below the specified level.

Other procedures (the Z_1, Z_2, and Z_3 tests) are appropriate if "missingness" is influenced either by membership of set A or B or by the value of the "yes-no" variable. These tests are described by Choi and Stablein (1988), who recommend use of the Z_1 test (based on unpaired observations only) if there are few pairs and many unpaired observations, or of the Z_2 test (the McNemar test, based on paired observations only) if there are many pairs and neither set, or only one set, has many unpaired observations, or of the Z_3 test (based on both paired and unpaired observations) if there are many pairs and both sets have many unpaired observations. The procedure described by Bland and Butland (undated) is also performed, as an alternative to Z_3; this provides confidence intervals for the difference between the proportions, as well as a significance test.

A modified McNemar test, in which fictional pairmates are allotted to unpaired observations in such a way as to reduce the contrast between the proportions, is offered for use if "missingness" is influenced both by membership of set A or B and by the "yes-no" variable. This is the Z_7 test of Choi and Stablein (1988). It is very conservative, and of limited value. It should be used only if the number of unpaired observations is extremely small in comparison with the number of pairs.

Tests are provided for comparing the paired and unpaired observations with respect to their proportions of "yes", and to the magnitude of the differences between sets A and B.

Correlation coefficients

Four correlation coefficients between the variables (A and B) in the 2x2 table are computed: the *phi* coefficient, which is the usual (Pearson) coefficient, applied to binary variables, and is appropriate if both variables are natural dichotomies based on qualitative characteristics (e.g cases and controls, or exposed and nonexposed); the *tetrachoric correlation coefficient* (see below), which is appropriate if both variables are quantitative ones that have been artificially dichotomized; and two *point-biserial correlation coefficients*, appropriate if one variable is (depending on which variable is naturally dichotomous and the other is a dichotomized quantitative variable (and depending on which variable is naturally dichotomous).

Tetrachoric correlation coefficient

An approximate tetrachoric correlation coefficient is computed, providing an estimate of what the correlation would be if the distributions were not dichotomised, assuming an underlying distribution that is continuous and approximately normal. The program computes an approximate coefficient, with its 95% confidence interval. The computation is not performed if there is a zero cell or undue unevenness of the marginal totals (see Methods).

Number needed to avoid one event

The program reports the number of individuals who are needed in the group with a lower rate in order to avoid a single case, with its approximate 95% confidence interval. These results apply to studies that compare the proportions of cases (of disease, etc.) in paired subjects

exposed and not exposed to a risk or protective factor or treatment, and to two-period crossover trials.

In a clinical trial the number needed has been called the "number needed to treat"' or "number needed to treat (benefit)" (Altman 1998), i.e. the number of patients who must be treated in order to prevent one event (Sinclair and Bracken 1994, Feinstein 1995). In an observational study of a supposed cause of disease, it indicates the number of people whose exposure must be prevented in order to prevent one event (assuming that the findings reflect a cause-effect relationship and that the causal factor and its effect are modifiable).

The number is the reciprocal of the risk difference, and the 95% confidence limits for the number needed in a group to avoid one case are the reciprocals of the 95% confidence limits for the risk difference. Since the confidence interval for the rate difference may straddle zero, the confidence interval for the number needed to avoid one case may straddle infinity. A confidence interval of 5.5 to -2.2 is reported as "5.5 to infinity (in the one group), then up to 2.2 in the other group".

Attributable or prevented fractions

Attributable and prevented fractions in the exposed and in the population are computed, with their confidence intervals. These are appropriate for case-control studies where the cases are randomly selected and the disease is rare. Confidence intervals based on large-sample standard errors are provided; they should be used with caution if numbers are small.

The computation of the fractions and their standard errors is based on the methods described by Kuritz and Landis (1987, formulae 4 to 9). If the attributable fraction AF is negative the cases and controls are reversed for the purposes of computation, and the calculated attributable fraction is reported as the prevented fraction PF. If a lower confidence limit for an AF is negative, the equivalent PF is displayed in parentheses.

The confidence intervals of the attributable and prevented fractions in the exposed are computed by Kuritz and Landis's formulae 10 and 11. The confidence intervals of these fractions in the population are generally based on the quadratic-equation method proposed by Lui (2001a, method 5). If the odds ratio is 4 or more or 0.25 or less, however, or the proportion of cases who are exposed is 50% or more, use is instead made of logit-transformed estimators, as recommended by Lui (2001a, method 3).

Kappa, percentage agreement, and related results

Kappa is generally used to measure the agreement between two "yes"-"no" ratings (by different observers or tests, or by the same observer on different occasions) of the same individuals. In addition to this use as a measure of reliability, it may be used to measure concordance in other situations where paired samples are compared (Fleiss *et al*. 2003: 618-619). In a matched case-control study or matched-control trial, *kappa* may serve as an indication of the effectiveness of a matching procedure – it indicates the extent to which the findings in matched pairs are more similar than findings in individuals from different pairs (Fleiss *et al*. 2003: 618).

Kappa, like other measures of agreement, reflects the agreement concerning specific subjects by specific raters, and can be generalized to a broader group only if the subjects are

representative of the broader group. As a measure of inter-rater reliability, its value depends on the choice of raters. Uses and misuses of *kappa* in epidemiology are discussed by (among others) Sim and Wright (2005), MacLure and Willett (1987), Thompson and Walter (1988a, 1988b), Kraemer and Bloch (1988), Bloch and Kraemer (1989), and Feinstein and Cicchetti (1990). Flight and Julious (2014) emphasize that because of "the disagreeable behaviour of the kappa statistic", it should always be interpreted in conjunction with the percentage agreement, prevalence-adjusted bias-adjusted *kappa*, prevalence index, bias index and maximum attainable *kappa* (see below). Note that *kappa* for binocularratings by two observers is offered by module B of this program.

The probability of chance agreement is taken into account in the calculation of *kappa*. A value of 1 indicates perfect agreement (after allowing for this probability of chance agreement) between ratings; 0 indicates no agreement other than what can be attributed to chance, and a negative value indicates less than chance agreement. Fleiss *et al*. (2003) suggest that a value of 0.75 or more indicates excellent agreement, and 0.40 or less indicates poor agreement. Cicchetti and Sparrow (1981) divide Fleiss's 0.40–0.74 group into 0.60–0.74: good; and 0.40–0.59: fair. Alternative guidelines are: over 0.80, very good agreement; 0.61–0.80, good; 0.41–0.60, moderate; 0.21–0.40, fair; and 0.20 or less, poor agreement (Landis and Koch 1977, Altman 1991). These levels may be taken into account in the appraisal of confidence intervals, e.g. by seeing whether the lower confidence limit lies above 0.40 (Basu and Basu 1995).

A one-tailed test is done, indicating whether *kappa* is significantly higher than zero. If *kappa* is 0.4 or more, a second test is done, indicating whether it is significantly higher than 0.4; and if it is 0.6 or more, a third test is done, indicating whether it is significantly higher than 0.6.

Confidence intervals are estimated both from the standard error and by a goodness-of-fit approach (Donner and Eliasziw 1992). The latter intervals are more accurate than those based on the standard error, especially in small samples; if any of the expected frequencies is <1, the intervals are labelled as approximate.

Paradoxical values of *kappa* – inconsistency with the apparent agreement – may occur because of bias (systematic one-sided variation between two ratings, i.e. "different calibration" of the observers or tests, expressed by a difference between their frequencies of "yes" responses) or because of a skewed "yes"-"no" distribution (inequality between the prevalences of "yes" and "no") (see, *e.g.*, Feinstein and Cicchetti 1989 and Gwet 2010: 30-34). As an indication of bias, the program displays Byrt's *bias index* (Byrt *et al.* 1993*);* the McNemar test appraises the significance of this bias. As indicators of imbalance between prevalences of "yes" and "no", it displays Byrt's *prevalence index* and an *index of asymmetry in agreement* (Lantz and Nebenzahl 1996). It also displays Lantz and Nebenzahl's *index of asymmetry in disagreement.* All four of these indices range from 0 to 100%. A high bias index tends to elevate *kappa*, and a high prevalence index tends to decrease *kappa*.

Two adjusted values of *kappa* – BAK (*bias-adjusted kappa*) and PABAK (*prevalence-adjusted bias-adjusted kappa*) – are computed (Byrt *et al.* 1993) to provide an indication of the above effects on *kappa.* These adjusted values are conditional on the observed percentage agreement. BAK is the value that *kappa* would take if there were no bias; it is equivalent to Scott's *pi* coefficient of agreement (Scott 1955) and to the intraclass *kappa* coefficient. Low *kappa* values are likely to be affected by such bias. PABAK is the value that *kappa* would take if, in addition, the prevalence of each category (as expressed by the mean of the two

raters' totals for the category) was equal. PABAK may be useful in appraising agreement when the percentage agreement is high and *kappa* is paradoxically low; it approximates to the highest possible *kappa* if the percentage agreement is above about 50% (Lantz and Nebenzahl 1996). PABAK is called *kappa-no*r by Lantz and Nebenzahl (1996), and is equivalent to Maxwell's *RE* (random error) coefficient of agreement (Maxwell 1977) and Bennett's S coefficient (Bennett et al. 1954). It should be noted that simulation studies have suggested that PABAK may substantially overestimate agreement (Hoehler 2000).

The program also displays the *maximum attainable kappa* consistent with the marginal totals,

The *percentage agreement* is reported. This is the percentage of individuals who are placed in the same category by both ratings. Unlike *kappa*, it is not corrected for chance agreement. Its significance is tested, using a one-sided test of the null hypothesis that agreement is not more than might be expected by chance. The *percentage of positive agreement* (*Ppos*) and *percentage of negative agreement* (*Pneg*) (Cicchetti and Feinstein 1990) are also shown, with their 95% confidence intervals. The percentage of positive agreement is the percentage of "yes" ratings that are paralleled by a "yes" rating by the other observer or test, among all "yes" ratings; and the percentage of negative agreement is the percentage of "no" ratings that are paralleled by a "no" rating by the other observer or test, among all "no" ratings. An imbalance between these two percentages may be of interest; the program reports the difference between them, with its 95% confidence interval. Cicchetti and Feinstein recommend that, because of its sometimes paradoxical results, *kappa* should always be accompanied by *Ppos* and *Pneg*. Three alternative methods are used to estimate 95% confidence intervals for the indices of positive and negative agreement and the difference between them. Samsa's method (Samsa 1996) is based on the assumption (not always true) that the two observers or tests have a similar tendency to rate subjects as "yes" or "no" (i.e., that they are "similarly calibrated"). According to a simulation study (Graham and Bull 1998), its intervals tend to be too wide. The program also applies two alternative procedures proposed by Graham and Bull: a *delta* method, which performs adequately if the sample size is 200 or more, and a Bayesian method, which is recommended if there are under 200 paired observations.

In clinical practice, the percentage of positive agreement (i.e, concordant positive ratings as a percentage of all positive ratings) represents the probability that, if a subject has been given a positive rating by a typical observer, another typical observer will concur. Similarly, the proportion of negative agreement expresses the probability of concurrence with a negative rating (Samsa 1996). The program displays separate probabilities that a second rating will agree with a first "yes" or "no" rating, depending on whether rating A or B is made first.

The program also displays two indices of agreement suggested by Chamberlain *et al*. (1975): the *proportionate positive agreement* (P_{ppa} or *ppa) index* and the *proportionate negative agreement* (P_{pna} or *pna) index*. The proportionate positive agreement index is the percentage of individuals with concordant "yes" ratings, among all individuals with at least one "yes" rating; and the proportionate negative agreement index is the percentage of individuals with concordant "no" ratings, among all individuals with at least one "no" rating.

Approximate 95% confidence intervals for the measures of positive and negative agreement are estimated by three methods. Samsa's method (Samsa 1996) is based on the assumption (not always true) that the two observers or tests have a similar tendency to rate subjects as "yes" or "no" (i.e., that they are "similarly calibrated"); according to a simulation study

(Graham and Bull 1998), its intervals tend to be too wide. The *delta* method described by Graham and Bull (1998) performs adequately if the sample size is 200 or more. The Bayesian method is recommended if there are under 200 paired observations.

If *stratified data* are entered (e.g. observations of individuals in different age groups), the heterogeneity of the *kappa* values in the different strata is tested, measures of heterogeneity (see above) are provided, three estimates of the overall *kappa* are computed, with their confidence intervals, and overall values of the percentage agreement and of the percentage agreement for each category are reported. The first estimate of the overall *kappa* is precision-based; it is produced by weighting each *kappa* by the inverse of its variance (Fleiss *et al*. 2003: 607). The second uses the methods of Donner and Klar (1996); computation of the overall kappa is based on the common correlation model (in which the expected responses for each pair of observations are based on the overall prevalence of the two possible responses). The associated heterogeneity test (which appraises compatibility of the stratum-specific estimates with the overall *kappa*) and estimation of confidence intervals are based on a goodness-of-fit approach, which has been shown to provide satisfactory confidence intervals for combined samples with as few as 50 subjects (Donner and Eliasziw 1992). The third estimate is obtained by weighting the *kappa* values by the sizes of the samples in the strata. A simulation study suggests that this is preferable to the precision-based method if *kappa* is not zero (Barlow *et al*. 1991).

Gwet's AC_1 statistic

The AC1 statistic is, like *kappa*, a chance-corrected measure of the extent of agreement between raters (Gwet 2002a, 2002b, 2008, 2010). Its main difference from *kappa* is that it bases the probability of agreement-by-chance on only the hard-to-classify subjects, using a model that in effect estimates their number. AC1 has been recommended for use instead of *kappa* on the grounds that its estimate of the probability of chance agreement is more appropriate, and that it is less influenced by differences in the propensity to give positive ratings and by differences in the prevalences of the response categories. It is hence more robust, avoiding paradoxical results. Monte Carlo simulation has demonstrated that it is less biased and has a smaller variance than *kappa*, the G-index, or the *pi* coefficient (Gwet 2008). But along with recommendations that it is preferable to *kappa (e.g.* Lombard *et al.* 2004; Stegmann and Lucking 2005; Haley *et al*. 2008, Wongpakaran *et al.* 2013), Blood and Spratt (2007) warn that "…the AC1 and AC2 statistics … remain infants in the statistical world … as is always the case with new statistics, caution should be exercised in their use and further examination should occur before they are adopted as the standard'".

The program calculates AC1's standard error and 90%, 95%, and 99% confidence intervals on the assumption that the subjects are a random sample of an infinitely large population. If the sample is drawn from a defined target population and the sampling fraction is known, the correct standard error can be computed as $\sqrt{[SE^2 \times (1 - f)]}$, where SE is the reported standard error and *f* is the sampling fraction, e.g. 0.1 (1 in 10).

Other chance-corrected measures of agreement

The *G-index*, or Brennan-Prediger coefficient (Brennan and Prediger 1981, Gwet 2010: 38) is a simple coefficient that bases the chance-probability of agreement only on the number of response categories. The program calculates the G-index's standard error and 90%, 95%, and

99% confidence intervals on the assumption that the subjects are a random sample of an infinitely large population. If the sample is drawn from a defined target population and the sampling fraction is known, the correct standard error can be computed as $\sqrt{[SE^2 \times (1 - f)]}$, where SE is the reported standard error and *f* is the sampling fraction, e.g. 0.1 (1 in 10).

Scott's *pi coefficient* (Scott 1955, Gwet 2010: 21) differs from *kappa* in that it is based on marginal probabilities (the probabilities that each response category will be selected) that are common to both raters, not those that are specific to each rater (Gwet 2010: 38), The program calculates the *pi coefficient*'s standard error and 90%, 95%, and 99% confidence intervals on the assumption that the subjects are a random sample of an infinitely large population. If the sample is drawn from a defined target population and the sampling fraction is known, the correct standard error can be computed as $\sqrt{[SE^2 \times (1 - f)]}$, where SE is the reported standard error and *f* is the sampling fraction, e.g. 0.1 (1 in 10). The *pi* coefficient is equivalent to BAK and to the intraclass *kappa* coefficient.

The *intraclass kappa coefficient* is highly recommended (Kraemer *et al*. 2002, Kraemer 2006) as a measure of reliability. It is appropriate when two sets of ratings of the same "Yes-No" variable are compared. It is identical to BAK and to Scott's *pi* coefficient.

As an alternative to *kappa*, the program also reports a modified version of *Peirce's i coefficien*t (Peirce 1884). This coefficient is based on a "mixture model" that assumes that a proportion of cases are "obvious" and classified correctly (using one of the pairs of marginal values as a "gold standard"), whereas others are "ambiguous" and classified randomly ("guessed"). The modified *i* coefficient suggested by Abar and Loken (2010) for use as a measure of reliability in 2x2 tables (e.g. to compare two raters) is the average of two *i* coefficients, one taking the row margins as fixed, and one taking the column margins as fixed. Computer simulations indicate that the modified *i* coefficient and *kappa* generally provide similar estimates of chance-corrected reliability, but that *kappa* tends to be downwardly biased when "guessing" tendencies are different for the two ratings, e.g. if one rater tends to choose "yes" in ambiguous cases and the other tends to choose "no".

Intraclass odds ratio

The intraclass odds ratio has been recommended as a measure of interrater agreement on binary measurements (Locatelli and Rousson 2016). It is the odds ratio between two exchangeable measurements made on the same subject, and can be interpreted as a ratio of the probabilities of concordance and discordance between the two raters. "A kappa value of 0.75, which is usually taken as the threshold for a good reliability, corresponds to an intraclass odds ratio of at least
49, meaning a probability of concordance at least 49 times higher than the probability of discordance [sic]. This may suggest... considering e.g. an intraclass odds ratio of 25 as being already a good reliability." (Locatelli and Rousson 2014).

Martin and Femia's *delta*-based measures of agreement

These measures of agreement are, like *kappa*, chance-corrected. They have been proposed as alternatives to *kappa* (Martin and Femia 2004, 2008) that are free of *kappa*'s limitations. The program estimates the "overall index", which is the chance-corrected number of "A:yes-B:yes" and "A:no-B:no" pairs, expressed as a percentage of all pairs) - i.e., it is a chance-corrected index analogous to the percentage agreement - and its two component "agreement"

indices, namely the chance-corrected percentage of agreements with respect to "yes" responses (A:"yes"-B:"yes" pairs, as a percentage of all pairs), and the chance-corrected percentage of agreements with respect to "no" responses ('A:"no"-B:"no" pairs, as a percentage of all pairs). It also provides measures of the raters' "consistency" with respect to "yes" responses and "no"responses respectively; these are chance-corrected indices analogous to the percentages of positive and negative agreement (Ppos and Pneg). The measures are asymptotic estimators. Negative indices may be regarded as zero. Approximate standard errors are calculated.

The estimator of total agreement may sometimes be deceptive, providing a non-zero value when there is no agreement (Martin and Femia 2008). This may be suspected if it is similar to either of the agreement indices and "the marginals are unbalanced in the same direction" (e.g., the column 1 total exceeds the column 2 total, and the row 1 total exceeds the row 2 total). A warning message is displayed if the latter condition applies.

Measures of predictive accuracy

As an option, the program provides a number of measures of predictive accuracy in both directions (using variable A as a predictor of variable B, and using B as a predictor of A). Besides the *percentage agreement* (the “proportion correct”) and *kappa* (which in this context may be termed the *Heidke skill score* or the *Doolittle skill score,* referring mainly to skill in weather forecasting), which are described above, these include Goodman and Kruskal’s *lambda, Peirce’s i coefficient* (which is *Youden’s index* and may also be termed *Peirce’s skill score*), the *true skill statistic*,the *Hansen-Kuipers skill score,* the *Hansen-Kuipers discriminant,* or the *Kuipers performance index),* the *critical success index* (also called the *ratio of verification* or the *threat score*), the *Gilbert skill score, Yule’s Q* (the *odds ratio skill score*), and *Shannon's H coefficient (entropy)* The pros and cons of the various measures are discussed in especial detail in publications on weather forecasting, such as Jolliffe and Stephenson (2003).

Kappa is the proportion of correct forecasts, after eliminating those forecasts that would have been correct purely due to chance.

Goodman and Kruskal’s *lambda* (Goodman and Kristal 1954, Siegel and Castellan 1988: 298-303) is a coefficient of forecasting efficiency that expresses the capacity of one variable to "predict" another. It is a "proportional reduction of error" index, i.e., an assessment of the proportion of incorrect predictions that can be prevented if information about the predictor variable is available. Lambda ranges from 0 (if the one variable is of no help in predicting the other) to 1 (if the one variable perfectly specifies the categories of the other). Its value is influenced by the relative sizes of the groups that are compared.

Peirce’s i coefficient (Peirce 1884) can be interpreted as the accuracy for “yes” outcomes plus the accuracy for “no”outcomes. It answers the question: “How well did the forecast separate the ‘yes’ outcomes from the ‘no’ outcomes?

The *critical success index* is the proportion of correct forecasts of a “yes” outcome, when correct forecasts of a “no” outcome are completely ignored.

The *Gilbert skill score* is a modification of the critical success score, expressing the proportion of correct forecasts of a “yes” outcome when allowing for the number of correct

forecasts that would have been obtained by chance (using a formula that does not ignore the correct forecasts of a "no" outcome).

Yule's Q, which is based on the odds ratio and is not influenced by the incidence of "yes" outcomes, has been recommended as a powerful way of testing the association between forecasts and observations (Stephenson 2000).

Shannon's H coefficient (entropy) is a meaure of the unpredictability of a variable. It indicates how much information is required in order to predict the distribution of the variable. It is here reported as the predictive accuracy (0 to 1) of the other variable.

Distinguishability of categories

A measure of the *distinguishability of the categories* (Darroch and McCloud 1986) is computed. This may be useful in a methodological study in which the matched observations represent separate ratings. The measure ranges from 100% if there are no disagreements, to zero if disagreements outnumber agreements.

Inverse sampling

Inverse sampling refers to the addition of pairs to the sample until a prespecified number of pairs with a specific combination of attributes has been found. The computation is based on the assumption that it is the number of pairs with an A: "no", B: "yes" combination that was specified in advance (the two sets of observations should be labelled accordingly when entering the findings). This method of sampling is appropriate only if subjects are accrued sequentially and their attributes can be determined rapidly

The program provides an appropriate test for the difference between the observations, and exact 90%, 95%, and 99% *confidence intervals for the odds ratio*.

Probability and odds of replication

P_{rep}, which predicts the probability that an effect will be replicated in other studies, was proposed by Killeen (2005) as an alternative to significance tests in evaluating research and as an aid in practical decision making (Sanabria and Killeen 2007}. The measure predicts the probability that a replication will find a difference in the same direction (i.e., a "same-sign" result, not necessarily statistically significant) as that found in the original study. Its appropriateness and accuracy have been debated (Iverson *et al.* 2009, Lecoutre and Killeen 2010, Killeen 2010)). Iverson *et al.*. argue that it overestimates the probability of replication. Cumming (2005), who states that "Killeen's P_{rep} is wonderful, but may be difficult to understand", prefers to refer to it as the average probability of replication (*APR*), i.e. the chance of a same-sign result, when averaged over studies in similar populations. As Killeen (2005) points out, a particular value of P_{rep} may be more or less representative of P(rep) values found for other studies carried out under similar conditions.

The program also reports the odds in favour of obtaining a same-sign effect, i.e. P_{rep} / [1 - P_{rep}], as suggested by Baguley (2012).

Clustered data

Some studies are based on clusters of paired “yes”-”no” observations that may not be independent, e.g. pairs of observations of the same person, or by the same observer. The study might, for example, be a clinical trial of the effects of treatment applied to the eyes of patients with early signs of cataract, based on before-after appraisals of visual acuity (“impaired” or “normal”). Since a person's two eyes may be similar, the findings may not be independent, and a simple McNemar test based on the pooled data might yield a spuriously high level of significance. Clustering may similarly occur in a study in which paired observations are made on multiple teeth belonging to the same person, or on multiple blood or tissue samples, or in a study in which different observers of the same subject participate. In such studies the data comprise clusters of related observations, one cluster per subject or per observer. Clusters may contain different numbers (one or more) of pairs of observations.

To analyse clustered data, all that is required is to enter each cluster as a separate stratum. When the combined strata are analysed, the effect of clustering is appraised and allowed for.

Four procedures are provided for this purpose: those described by Eliasziw and Donner (1991), by Obuchowski (1998), and by Durkalski *et al*. (2003), and a modification of the Obuchowski test, proposed by Yang *et al*. (2010).. The Eliasziw-Donner procedure adjusts the McNemar test and estimates adjusted confidence intervals for the odds ratio. The adjusted McNemar chi-square differs from the unadjusted McNemar chi-square only if at least one cluster contains two or more discrepant pairs of observations, the Obuchowski and Durkalski procedures provide significance tests and adjusted confidence intervals for the difference between the proportions of “yes” responses in the sets of paired observations, and the modified Obuchowski procedure provides a significance test.

The relative value of the four tests varies in different circumstances. The Obuchowski test is slightly less powerful than the Eliasziw-Donner test (Obuchowski 1998), and is more powerful than the Durkalski test if cluster size is very variable (Durkalski et al. 2003). On the basis of computer simulations, Yang *et al*. (2010) conclude that their modified Obuchowski test is the most powerful, the original Obuchowski test is the most conservative, and the performance of Durkalski's test varies between the original and modified Obuchowski tests. They recommend use of the modified Obuchowski test if the clusters are of equal size. If the clusters differ in size, they recommend use of Durkalski's test if the number of clusters is small (below 50), and the modified Obuchowski test if there is a large number of dusters.

Crossover trial

The crossover study must have a “yes-no” outcome, where “yes” may, in different studies, indicate "success", e.g. reduction of a symptom, or a patient’s preference for a treatment, or "failure", e.g. occurrence of a symptom. It compares two treatments, A and B (one of which might be a placebo) that are applied in sequence to the same subjects, with (if necessary) an intervening “washout” period sufficiently long to remove the effects of the first treatment. Subjects are randomly allocated to groups that receive Treatment A first (AB sequence) or second (BA sequence).

The program computes the *proportions of "yes" results* for the two treatments, separately for the first and second periods, with significance tests for the differences between the treatments. The proportions among "informative" subjects are computed as well as those among all subjects, "informative" subjects being defined as those who have different outcomes to the two treatments. Significance tests also compare the proportions of " yes" results for the two treatments (among "informative" subjects and among all subjects) among subjects in the AB (Treatment A first) and BA (Treatment B first) groups. Multiple testing is not taken into account.

The program also reports the *overall difference* between the proportions of "yes" results (among all subjects), with its 90%, 95%, and 99% confidence intervals (Schouten and Kester 2010).

Odds ratios comparing the two treatments with respect to their odds in favour of a "yes" result are computed separately for the AB-sequence and BA-sequence groups, as well as for the pooled data.

The comparison of the treatments may be confounded by the effect of their sequence, unless there are adequate washout periods. Clues to the occurrence of a period effect (e.g., a carry-over effect whereby the first treatment affects the outcome in the second period) may be provided by comparisons of the proportions of " yes" results (and their differences) in the two periods (i.e., in the AB and BA sequences), and by tests for an order effect (see below). According to Freeman (1989) and Senn (2002), reliance on tests for an order effect may be misleading.

If a period effect is suspected, the usual recommendation is to base the assessment solely on the first-period findings, i.e. to disregard the second period and treat the trial as a simple parallel-group comparison. Specifically, the results when one treatment is given first are compared with the results when the other treatment is given first. But even if there is a carry-over effect, recourse to the first-period comparison may not always be necessary. It is the less effective treatment that is more likely to be influenced (in the second period) by a carry-over effect of the other treatment. In a trial in which a high proportion of "yes" results points to the success of the treatment, it may therefore be sufficient to concentrate on the treatment with a lower proportion of yes results in the first period, and base the assessment solely on the first-period findings only if this treatment's proportion of "yes" results is substantially higher in the second period than in the first (Cleophas et al. 2009). Calculations suggest that tests for the treatment effect remain powerful even if there is a substantial carryover effect, so that a possible carryover effect can be ignored if the findings point to a significant treatment effect (Cleophas *et al.* 2009).

Two tests that point to a possible period (e.g. carry-over) effect are performed: (a) a test for a discrepancy between the AB and BA groups in their proportions of "yes-yes" and "no-no"results (Armitage and Hill 1982); it has been suggested that a critical level of $P < 0.1$ should be used for this test, rather than $P < 0.05$ (Nagelkerke et al. 1986); and (b) the Armitage-Hills test for treatment-by-period interaction (Armitage and Hills 1982), which is similar to the b est. These tests assume valid randomization of the subjects.

The relative effects of the two treatments are appraised not only by the tests of the differences between proportions in each period, but also by M*cNemar, Mainland-Gart, Prescot*t, and *Schouten-Kester tests.* McNemar tests, which are based on the findings in

"informative" subjects, are performed for the combined data and for each period separately. The Mainland-Gart test (Mainland 1963, p. 237; Gart 1969) is based on the findings in the "informative" subjects in both periods, using a 2 x 2 table formed by removing the noninformative subjects. Prescott's test (Prescott 1981) is a test for linear trend of the A:B relationship in a contingency table that includes the noninformative subjects as a middle group; it may be misleading if numbers are very small. The Schouten-Kester test (Schouten and Kester 2010) is based on the average of the treatment differences found in the two periods. The Prescott and Schouten-Kester tests make allowance for a possible period effect.

The program reports the *number needed to treat* in order to avoid a single "yes" result (if "yes" indicates failure) or to produce a single yes" result (if "yes" indicates success), with its approximate 95% confidence interval, based on the separate data for each sequence, and then on the pooled data. The number needed is the reciprocal of the risk difference, and its 95% confidence limits are the reciprocals of the 95% confidence limits for the risk difference. Since the confidence interval for the rate difference may straddle zero, the confidence interval for the number needed may straddle infinity. A confidence interval of 5.5 to -2.2 is reported as "5.5 to infinity (in the one group), then up to 2.2 in the other group".

Reconstruction of 2x2 table

An option is offered for the reconstruction of the paired 2x2 table, based on an odds ratio or a two-tailed P value (and the proportions of "yes" in the two groups). This may be helpful in meta-analyses of studies with incompletely reported data. The reconstructed table is then analysed in the usual way by this module.

The procedure is described by Hirji and (2011), who deal not only with the reconstruction of the table, but with the calculation of confidence intervals for the risk difference, the risk ratio, and the odds ratio. They point out that since the table can often be reconstructed by using the odds ratio, use of the P value will rarely, if at all, be necessary.

The results cannot be regarded as exact, since they are influenced by the accuracy of the entered odds ratio or P value, by rounding-off, and (if a P-value is used) by which significance test yielded the P value. However, Hirji and Fagerland say that if the P-value is known to two significant digits, the results are sufficiently accurate. They give an example showing that P-values ranging from 0.015 to 0.024 (all of which might be entered as 0.02) can produce 95% C.I.s ranging from 1.04-18.06 to 1.32-18.68 - changes which, they say, are "neither dramatic nor practically meaningful". They recommend use of their methods provided there are more than 50 pairs and the data are not too skewed or sparse.

METHODS

Tests for the difference between paired observations

The Fisher's and mid-P exact tests use an efficient algorithm for calculating the coefficients of the conditional distribution (Martin and Austin 1991, 1996), using code from David O. Martin's public-domain EXACTBB program. The McNemar tests use formulae 4.3 and 4.4 of Siegel and Castellan (1988: 43). The formula for the modified Wald test (May and Johnson 1997) is

$$\text{chi-sq} = (b - c)^2 / [(b + c + 1) - (b - c)^2 \ / \ n]$$

where b and c are the numbers of discrepant pairs
n = total number of pairs

Heterogeneity tests and measures

The test for the heterogeneity of odds ratios in different strata is based on a multiple-sample goodness-of fit test (Sokal and Rohlf 1981: 711-716; Zar 1996: 471-473), using log-likelihood chi-squares (without corrections for continuity) in each stratum; 0.0000001 is added to cells with frequencies of zero. The tests are for goodness of fit with an equal distribution of pairs with discrepancies in different directions.

The test for the heterogeneity of *kappa* values is based on the method of Donner and Klar (1996).

The *measures of heterogeneity* (Higgins and Thompson 2002) are *H* and *I-squared*. *H* is computed by Higgins and Thompson's formula 6, and increased to 1 (indicating absence of heterogeneity) if it less than 1. A test-based interval is computed by Method III. *I-squared* and its 95% interval are computed from *H*, using formula 10.

Test of equivalence

The program uses a test based on restricted maximum likelihood estimation (RMLE), without a continuity correction. This method, described by Nam (1997), has been evaluated and recommended by Liu *et al*. (2002), who explain how to replace the standard errors in the basic formulae (formulae 4 and 5) with RMLE-based values.

Standardized mean difference

The Cox estimate uses formulae 18 and 19 of Sanchez-Mecca et al. (2003), and the Glass et al. estimate uses formulae 20 and 21.

Odds ratio

The odds ratio is b/c or c/b, where b and c are the numbers of discrepant pairs. The low-bias estimator of the odds ratio (Jewell 1984) is $b / (c + 1)$ or $c / (b + 1)$.

Confidence intervals for odds ratios are estimated by treating the two values as Poisson variates, with their ratio (the odds ratio) distributed binomially (Morris and Gardner 2000: 65). Exact probabilities and confidence intervals are computed with an efficient algorithm for calculating the coefficients of the conditional distribution (Martin and Austin 1991, 1996), using code from David O. Martin's public-domain EXACTBB program. THe Wilson score CIs are calculated by formula 45 of Fagerland et al. 2014) [they are omitted iftheir ca;clation requires a division by jzero].

Proportions, difference between proportions, ratio of proportions

Confidence intervals for the *proportions* (of "yes") are computed by the method described by Newcombe and Altman (2000: 46-47). Confidence intervals for the *difference between proportions* are computed by the score method with a continuity correction, as recommended by Newcombe and Altman (2000: 52), which is method (10) of Newcombe (1998b), and by the improved Wald method ("Wald + 2") recommended by Agresti and Min (2005: formula 2, with N set at 2).

The Wald confidence intervals for the *ratio of proportion*s are based on formulae 16-2 and 16-3 of Rothman and Greenland (1998); also formula 1 of Bonett and Price (2006). The Wilson and Wilson-cc intervals are described by Bonett and Price (2006); PAIRSetc uses an adaptation of the Gauss code provided by these authors.

Incompletely paired data

The *odds ratio* and its confidence intervals are estimated by the procedure described by Miller and Looney (2012) (formulae 1 to 4).

The *P value* for the combined data (Kuan and Huang 2013) is based on pooling of the results of a McNemar test of the paired data and a chi-square test of the unpaired data. The *Z* values provided by the two tests are combined, after weighting them by the square roots of (respectively) twice the number of pairs (as recommended, in a similar context, by Choi and Stablein 1988), and the number of unpaired observations.

For *incompletely paired data*, the confidence intervals of the difference between proportions are computed by the asymptotic method described by Tang *et al.* (2009: formulae 1 and 2), calculating the weights as $n / (n + m_1)$ and $n / (n + m_2)$, as suggested by Choi and Stablein (1982). According to supplementary explanations provided by Ling A (personal communication), if m_1 is zero the term in formula 1 that has m_1 as its denominator is ignored, and b_1 (required in formula 2) is set at zero; if m_2 is zero the term in formula 1 with m_2 as its denominator is ignored, and b_2 is set at zero; if the computed standard error is zero, the whole calculation is repeated after substituting an adjustment constant of 0.5 for a zero number of discrepant pairs (in either direction); in Table V of Tang *et al.* (2009), the correct C.I. by this method is (-1, -0.1590), and not (--/796, -0.071) as misprinted. (Ling A, personal communication)

The Z_1, Z_2, Z_3, and Z_7 tests are described by Choi and Stablein (1988). Z_3 uses a weighted combination of Z_1 (the result for unpaired observations) and Z_2 (the result for paired observations). In computing the weighting factor used in the calculation of Z_3, each pair is counted as two observations.

For Z7, missing observations are replaced by fictional results, in such a way as to reduce the contrast between the proportions. For example, if the number of A: "yes", B: "no" pairs (n_{10}) exceeds the number of A: "no", B: "yes" pairs (n_{01}), missing observations are changed to "yes" if the known result for A is "no", and to "no" if the known result for B is "yes"; whereas if n_{01} exceeds n_{10}, missing observations are changed to "no" if the known result for A is "yes", or to "yes" if the known result for B is "no". The adjusted proportions are reported, and a McNemar test is performed:

$$\text{chi-sq. (1 d.f.)} = (n_{10} - n_{01})^2 / (n_{10} + n_{01})$$

If the adjustments produce a reversal in the direction of the relationship between n_{10} and n_{01}, the adjusted proportions are not reported, and P is reported as 1.

The procedure described by Bland and Butland (undated) is performed only if there is at least one unpaired observation in each set. The procedure uses weighted averages of the differences (between proportions) observed in the paired and unpaired data. There are misprints in the formulae for the variances of these differences: in each formula, the "plus" sign between the two terms is misprinted as a "minus" sign. The comparison of paired and unpaired observations is based on these variance formulae. The test for the difference between the differences in the paired and unpaired dat is omitted if it involves division by zero.

Relative difference

The relative difference is calculated by formula 13.16 of Fleiss *et al.* (2003: 379), and its confidence intervals by the log-transformation method described by Lui (2004: 57: formula 3.22).

Number needed to avoid one event

The number is the reciprocal of the risk difference, and the 95% confidence limits for the number needed in a group to avoid one case are the reciprocals of the 95% confidence limits. The program uses the method described by Walter (2001) for a crossover design with discrete data (formulae 2 and 3).

Correlation coefficients

The *phi* coefficient is computed by formula 16.20 in Sheskin (2007).

The formula used for the *tetrachoric correlation coefficient* (Edwards and Edwards 1984) is

$$(OR^{\pi/4} - 1) / (^{OR\pi/4} + 1)$$

where $OR = ad/bc$

a and d = numbers of concordant pairs

b and c = numbers of discordant pairs

This simple method, which was used by Stata until recently, provides an approximation that is acceptable in many situations (Digby 1983, referring to an almost identical formula [with ¾ instead of pi/4]) but that can be

very inaccurate (Uebersax (2000). V. Wiggins, of the Stata Corporation, in a reply cited by Gunther and Hofler (2006), says that the approximation works well when the marginals in both directions are above 10%. PAIRSetc does not display the coefficient unless this condition is met, and there are no zero cells. An approximate 95% confidence interval is estimated from a large-sample estimate of the standard error (cited by Digby (1983).

The *point-biserial correlation coefficients* are computed by formula 3 of Ulrich and Wirtz (2004).

Attributable and prevented fractions

The computation of the fractions and their standard errors is based on the methods of Kuritz and Landis (1987, formulae 4 to 9). If the attributable fraction AF is negative the cases and controls are reversed for the purposes of computation, and the calculated attributable fraction is reported as the prevented fraction PF. If a lower confidence limit for an AF is negative, the equivalent PF is displayed (and vice versa), using the formulae

$$PF = 1 - 1 / (1 - AF)$$
$$AF = 1 + 1 / (PF - 1)$$

The confidence intervals of the attributable and prevented fractions in the exposed are computed by Kuritz and Landis's formulae 10 and 11. The confidence intervals of these fractions in the population are generally based on the quadratic-equation method proposed by Lui (2001a, method 5); their computation is sometimes prevented by a need to calculate the square root of a negative value. If the odds ratio is 4 or more or 0.25 or less, however, or the proportion of cases who are exposed is 50% or more, use is instead made of logit-transformed estimators, as recommended by Lui (2001a, method 3).

Kappa and related results

The basic formulae are provided by Fleiss *et al*. 2003: chapter 18). *Kappa* is calculated by formula 18.12. For tests of the null hypothesis that *kappa* is zero (formulae 18.14 and 18.35), the standard error (for an underlying zero value of *kappa*) is calculated by formula 18.13. For tests of the hypothesis that *kappa* has an underlying value other than zero, and for confidence intervals, the standard error appropriate for non-zero values is calculated by formula 18.15.

The *maximum attainable value* of *kappa* is computed by calculating *kappa* when taking the marginal totals as fixed but modifying the body of the table so as to represent the maximum possible agreement, by using, for each cell indicating agreement, the smaller of the two relevant marginal frequencies.

Confidence intervals are estimated by two methods: by using the standard error (if the upper confidence limit exceeds 1, it is reduced to 1), and by the goodness-of-fit approach explained by Donner and Eliasziw (1992), which uses a model in which the expected frequencies of "yes"-"yes", "yes"-"no", and "no"-"no" observations are computed from the overall prevalence of "yes" responses.

Bias is appraised by the McNemar chi-square test (see above) .

Indices of asymmetry in agreement and disagreement are calculated (as percentages) by formulae provided by Lantz and Nebenzahl (1996), who refer to them as indices of symmetry. The index of asymmetry in agreement is $|a - d| / (a + d)$ x 100, where *a* and *d* are the numbers of nondiscrepant pairs, and the index of asymmetry in disagreement is $|b - c| / (b + c)$ x 100 where *b* and *c* are the numbers of discrepant pairs. The *bias index* is $|b - c| / N$ x 100, and the *prevalence index* is $|a - d| / N$ x 100, where *N* is the sample size (Byrt *et al.* 1993)

BAK (*bias-adjusted kappa*) and PABAK (*prevalence-adjusted bias-adjusted kappa*) are computed by the methods described by Byrt *et al.* (1993).

In the combined analysis of several samples or strata, the estimate of the supposed *common or overall value* of *kappa* is calculated in three ways: by computing a weighted mean, using the inverse of the variance of each *kappa* as its weight (Fleiss *et al*. 2003: formula 18.21); by the methods of Donner and Klar (1996), which use the common correlation model (the expected responses for each pair of observations are based on the overall prevalence of the two possible responses); and by computing a weighted mean, using the size of the stratum as the weight. The confidence intervals of the common kappa are estimated by formula 18.23 of Fleiss *et al.* (2003) and by the goodness-of-fit approach of Donner and Klar (1996).

The *heterogeneity* tests are based on formula 18.22 of Fleiss *et al*. (2003) and the goodness-of-fit approach of Donner and Klar (1996). The measures of heterogeneity (Higgins and Thompson 2002) are described above.

The *percentage agreement* is $(a + d) / n$,
where a = "yes-yes" pairs
b = "yes-no" pairs
c = "no-yes" pairs
d = "no-no" pairs
$n = a + b + c + d$

Significance is tested by a binomial test comparing the total number of complete agreements with the number expected by chance (Sheskin 2007: 729-730).

If stratified data are entered, the overall values of the percentage agreement are based on the pooled data; this is equivalent to weighting the stratum-specific values by sample sizes.

The formulae for the Chamberlain indices of agreement (Chamberlain *et al*. 1975), *ppa* (the *proportionate positive agreement index*, and *pna* the *proportionate negative agreement index)*, are

$$ppa = a / (a + b + c)$$

and $$pna = d / (d + b + c)$$

and the formulae for the Cicchetti-Feinstein indices (Cicchetti and Feinstein 1990), *Ppos* (the *positive agreement inde*x), and *Pneg* (*the negative agreement index*) are:

$$Ppos = 2a / (2a + b + c))$$

and $$Pneg = 2d / (2d + b + c))$$

(These indices are related: $ppa = Ppos / (2 – Ppos)$, and $pna = Pneg / (2 – Pneg)$.
Three alternative methods are used to estimate 95% confidence intervals for these indices: Samsa's method and *delta* and Bayesian methods. Samsa's procedure (Samsa 1996) is based on a variance calculated as $P(1 – P) / g$, where *P* is the index and *g* is the number of subjects rated in the same way by one of the raters or tests; e.g.., for subjects rated "no", g = (b + d) or (c + d). Since either rater or test may be chosen for this purpose, the method will yield two different confidence intervals if calibration is not identical. PAIRSETC therefore uses the mean of these two alternative numbers (rounded off downwards).
The *delta* and Bayesian methods are described by Graham and Bull (1998).

These three methods are also used to estimate 95% confidence intervals for the difference between the percentages of positive agreement (*Ppos*) and negative agreement (*Pneg*). The estimates are $(Ppos – Pneg) \pm 1.96\sqrt{}[(\text{var}(Ppos) + \text{var}(Pneg)]$ using variance estimates obtained by Samsa's method, and $(Ppos – Pneg) \pm 1.96\sqrt{}\text{var}(Ppos – Pneg)$ by the method of Graham and Bull, who provide a formula for var(Ppos – Pneg). The interval based on the Samsa variances must be regarded as very approximate since, as pointed out by Graham and Bull, it ignores the covariance between *Ppos* and *Pneg.* The Bayesian estimates (see below) are based on a Monte Carlo procedure. In the output, the difference is expressed as *Ppos - Pneg* if *Ppos* is larger, and as *Pneg - Ppos* if *Pneg* is larger.

The *Bayesian intervals* (for *Ppos, Pneg* and their difference) are estimated by Monte Carlo procedures in which 5000 samples are generated, using the algorithm presented in Appendix A of Graham and Bull (1998), with an almost noninformative" prior distribution of 0.25 in each cell. The 95% interval estimates are obtained from the 2.5th and 97.5th percentiles of the simulated distributions. The beta variates are generated by algorithm BB or BC depending on the relative sizes of the adjusted cell values) of Cheng (1978). The random numbers used by these procedures are generated by a pseudo-random number generator described by Wichman and Hill (1985), which derives each number in turn from three seed numbers (in the range 1 – 30,000) which it modifies for subsequent use. The initial seed numbers are generated by Delphi's inbuilt random-number procedures: RANDOMIZE, which derives a preliminary seed from the system clock, and Delphi's RANDOM procedure (which generates three random numbers from which the required seed numbers are computed), supplemented by an additional randomizing shuffle, using the algorithm of Bays and Durham, as described by Press *et al.* (1989: 215-217). The Bayesian procedure may yield slightly different results each time it is repeated.

The *proportionate positive agreement index* and *proportionate negative agreement index* and their confidence intervals are computed by the formula (Graham and Bull 1998)

$$P / (2 - P),$$

where P = (%age of positive or negative agreement or its lower or upper confidence limit) / 100.

Gwet's AC_1 statistic

Gwet's AC1 is calculated by formula 4.1 of Gwet (2010: 61), and its variance by formula 5.7 of Gwet (2010: 94). The program calculates AC1's standard error on the assumption that the subjects are a random sample of an infinitely large population. If the sample is drawn from a defined target population and the sampling fraction is known, the correct standard error can be computed as $\sqrt{t}[SE^2 \times (1 - f)]$, where SE is the reported standard error and *f* is the sampling fraction, e.g. 0.1 (1 in 10).

Other chance-corrected measures of agreement

Brennan and Prediger's *G-index* is calculated by formula 2.18 of Gwet (2010: 38), and its variance by Gwet's formula 5.10). The program calculates the G-index's standard error on the assumption that the subjects are a random sample of an infinitely large population. If the sample is drawn from a defined target population and the sampling fraction is known, the correct standard error can be computed as $\sqrt{t}[SE2 \times (1 - f)]$, where SE is the reported standard error and *f* is the sampling fraction, e.g. 0.1 (1 in 10).

Scott's *pi coefficien*t is calculated by formula 2.6 of Gwet (2010: 21), and its variance by Gwet's formula 5.8 The program calculates the standard error of *pi* on the assumption that the subjects are a random sample of an infinitely large population. If the sample is drawn from a defined target population and the sampling fraction is known, the correct standard error can be computed as $\sqrt{t}[SE2 \times (1 - f)]$, where SE is the reported standard error and *f* is the sampling fraction, e.g. 0.1 (1 in 10).

The formula for the modified *Peirce's i coefficient* (Abar and Loken 2010) is

$$0.5\{(ad - bc) / [(a + c)(b + d)] + (ad - bc) / [(a + b)(c + d)]\}$$

where a b
c d represent the four cells of the 2x2 table.

[An error in the reporting of this coefficient was corrected in version 3563.]

Intraclass odds ratio

Formula 8 of Locatelli and Rousson (2016) is used for the intraclass odds ratio and formula A5 for the variance of its logarithm.

Intraclass *kappa* coefficient

The intraclass *kappa* coefficient is computed by formula 2.40 of Vanbelle (2009):

$$Kappa = (Po - Pe) / (1 - Pe)$$

where $Po = (a + d)/ N$

$$Pe = ((2a + b + c) / 2N)^2 + ((2d + b + c) / 2N)^2$$

$$N = a + b + c + d$$

where a b
c d represent the four cells of the 2x2 table.

Martin and Femia's delta-based measures of agreement

Formulae for the asymptotic estimators for chance-corrected overall agreement (the overall index), for agreement with respect to "yes" and "no" responses, and for consistency, and for their variances, are provided by Martin and Femia (2008: Table 6). Before computation, 1 is added to each of the cells in the 2 x 2 table, to improve the performance of the estimators. It is assumed that the total sample size is fixed in advance, but not the row or column marginal totals.

Measures of predictive accuracy

The formulae for *lambda* (Goodman and Kristal 1954, Siegel and Castellan 1988: formulae 9.37 and 9.39) are:

lambda = (*SumCell1* - *maxR*) / (*n* - *maxR*) if variable A is the predictor;

lambda = (*SumCell 2* - *maxC*) / (*n* - *max*C) if variable B is the predictor,

where *SumCell1* = (the larger of *a* and *c*) + (the larger of *b* and *d*)
SumCell2 = (the larger of *a* and *b*) + (the larger of *c* and *d*)
MaxR = the larger of the two row totals, i.e., (*a* + *b*) or (*c* + *d*)
MaxC = the larger of the two column totals, i.e.. (*a* + *c*) or (*b* + *d*)
a, b, c, and *d* are the frequencies in the 2x2 table:

		A	
		Yes	No
B	Yes	*a*	*b*
	No	*c*	*d*

n = *a* + *b* + *c* + *d*

The standard error of *lambda* (on which its approximate 95% confidence interval is based) is calculated by formulae 9.38 and 9.40 of Siegel and Castellan (1988). If the lower limit is less than 0, it is taken as 0; if the upper limit is above 1, it is taken as 1.

Proportion correct = (a + d) / n
Its confidence interval is estimated by Wilson's method, recommended by Newcombe and Altman (2000: 46-47).

Kappa may be calculated as (*PC* – *E*) / (*1* – *E*)
where *PC* = proportion correct
E = [(*a* + *c*) / *n*] [(*a* + *b*) / *n*] + [(*b* + *d*) / *n*] [[(*c* + *d*) / n]
or as 2(*ad* – *bc*) / [(*ac*+ *c*)(*c* + *d*) + (*a* + *b*)(*b* + *d*)]
The standard error used for estimating 95% confidence intervals for *kappa* is calculated by formulae 18.15 to 18.18 of Fleiss *et al.* (2003).

Peirce's i coefficient = (ad – bc) / [(a + c)(b + d)]
The formulae for its standard error and confidence interval (as Youden's index) are provided by Youden (1950) and cited by Salmi (1986). They is appropriate if the "yes" outcomes and the "no" outcomes are at least 20, and if the index is not very close to zero or one.

The *critical success index* is *a* / (*a* + *b* + *c*)

The *Gilbert skill score* is (*a* – *F*) / (*a* + *b* + *c* – *F*)
where *F* = [(*a* + *b*)(*a* + *c*)] / *n*

Yule's Q is computed, after adding 0.1 to each cell frequency, by the formula
(*ad* – *bc*) / (*ad* + *bc*)
Its 95% confidence interval is estimated, if its value is > -1 and < 1, by formula 11 of Bonett and Price (2007).

The two formulae for *Shannon's H coefficient* are provided in "2-way Contingency Table Analysis", available on the Internet at http://statpages.org/ctab2x2.html):
H(c) = - ((*c1/t*)log2(*c1/t*) + *(c2/t*)log2*(c2/t)*)
H(r) = - (*(r1/t*)log2(*r1/t*) + (*r2/t*)log2(*r2/t*))
where *H(c)* is the coefficient for variable
H(r) is the coefficient for variable A
c1 = *a*+*c*
c2 = *b*+*d*
r1 = *a*+*b*
r2 = *c*+*d*
t = *c1*+*c2*

Probability of replication

P_{rep} is computed from the McNemar chi-square3.

P_{rep} estimates the non-centrality parameter from the maximum of (chi-square - 1) and zero (Saxena and Adam, 1982), and evaluates that non-central chi-square by use of an approximation given by Sankaran (1963) (extracted from a Wikipedia article entitled "Noncentral chi-squared distribution"), modified by Killeen (personal communication) by multiplying the standard deviation by the square root of 2.

Distinguishability of categories

This measure is computed by the method described by Darroch and McLeod (1986).

Inverse sampling

The difference between the observations is tested by the formula (Lui 1996) :

$$\text{chi-square (1 d.f.)} = (b - c)^2 / 2c$$

where b = A: "yes", B: "no"
c = A: "no", B: "yes"

Exact confidence intervals for the odds ratio are computed by formula 5.58 of Lui (2004: 112).

Clustered data

The *Eliasziw-Donner procedure* to adjust for the presence of clusters of non-independent paired observations estimates a weighted average within-cluster intraclass correlation coefficient, *rho,* using information on both concordant and discordant pairs, by the methods described in Section 4 of the paper by Eliasziw and Donner (1991). The program reports the value of r*ho*. (*Rho* cannot be computed if the clusters contain only one discrepant pair of observations, and the procedure is then not performed). Using the methods described in Section 2 of the paper, a correction factor for the McNemar test is then computed (the program divides the McNemar chi-square by this factor). An adjusted variance is computed for the prevalence of discrepancies in one direction, $b / (b+c)$. Confidence intervals are estimated for this prevalence, and converted to adjusted confidence intervals for the pooled odds ratio. 'Not computed" is reported if a computational difficulty is encountered.

The *Obuchowski procedure* for comparing correlated proportions in clustered data uses formula 6 of Obuchowski (1998) to compute a chi-square test statistic; for this purpose the estimator of the variance of the difference between the proportions of "yes" responses in the sets of paired observations is computed by formula 4, after substituting the pooled (mean) proportion for the specific proportions in formula 2, and replacing the covariance estimator computed by formula 3 with that provided by formula 7. A 95% confidence interval for the difference between proportions is based on the variance estimator in formula 2; the square root of this variance is displayed as the standard error of the difference. The formula for the *modified Obuchowski test* is provided by Yang *et al*. (2010), and appears just before their formula 1 .

The *procedure described by Durkalski et al.* (2003) for the analysis of clustered matched-pair data computes chi-square by formula 15. The test is not performed if the clusters contain only one discrepant pair of observations, since it then yields the same result as the unadjusted McNemar test. A 95% confidence interval for the difference between proportions (formula 18) is based on the variance estimator in formula 17; the square root of this variance is displayed as the standard error of the difference.

The Obuchowski and Durkalski procedures are described briefly by McCarthy (2007).

Crossover trial

Differences between proportions of "yes" results are tested by formula 3.15 of Fleiss et al. (2003). The *overall difference* between the proportions is based on the means of the results in the two periods, and its 90%, 95%, and 99% confidence intervals are based on the variance formula provided by Schouten and Kester (2010: p. 194).

The *McNemar tests* use a continuity correction (Siegel and Castellan 1988 , formula 15.2).

The *Mainland-Gart test* uses the formula provided by Senn (2002: p. 130).

For *Prescott’s test,* a Cochran-Armitage trend test (Armitage *et al.* 2002, equation 15.1) is performed, applied to a 2 x 3 contingency table showing, for each of the two sequence groups, the number of subjects with “yes” for A and “no” for B, the number with the same responses for A and B, and the number with “no” for A and “yes” for B (Jones and Kenward 2003: p. 114).

The *Schousten-Kester test* uses the variance formula (assuming a period effect) provided by Schouten and Kester (2010) in their Appendix A.

The *tests for a carry-over effect* are described by Armitage and Hills (1982) – a *chi*-square test for the discrepancy between the AB and BA groups in their proportions of “yes-yes” and “no-no” results (Hills and Armitage 1979), and the Armitage-Hills test (Armitage and Hills 1982), which is a trend test applied to a 2 x 3 contingency table showing, for each of the two sequence groups, the numbers of subjects with “yes-no” and “no-yes” results, with an intermediate category for the subjects with (pooled) “yes-yes” and “no-no” results.

The number needed to avoid/produce a single "yes" result is the reciprocal of the risk difference, and the 95% confidence limits for the number needed in a group to avoid one case are the reciprocals of the 95% confidence limits for the risk difference. The program uses the method described by Walter (2001) for a crossover design with discrete data (formulae 2 and 3).

Reconstruction of 2x2 table

Optionally, use can be made of an odds ratio or a two-tailed P-value. If an odds ratio is entered, the table is constructed by employment of the formulae in row 3 of Table 4 of Hirji and Fagerland (2011). If a P-value is entered, the formulae in row 1 are used.

A2. CONCURRENT ASSESSMENT OF INTERRATER AND INTRARATER RELIABILITY ("YES-NO" VARIABLE)

This module assesses **interrater and intrarater reliability** in a study that compares "yes-no" ratings of the same subjects made by two raters, each of whom rate each subject twice. The "raters" may be different observers, different measuring instruments, or different methods or conditions of measurement.

The four ratings of each subject are required.

The program provides three *measures of agreement* (equivalent to *kappa*): one inter-rater reliability and (for each rater) a measure of intrarater reliability, with its standard error.

Measures of reliability

The *measures of reliability* (which are equivalent to *kappa*) are computed by the method described by Shoukri and Donner (2001), who conclude that the use of two ratings of each subject (instead of one) may allow fewer subjects to be included in studies of interrater reliability, with no net loss in efficiency.

This procedure may also be appropriate in studies where there have to be two ratings by each rater, as in a study of the presence of some lesion in the eyes, or in studies of twins.

METHOD

The computation is based on a nested beta-binomial model. The interrater reliability is computed by formula 8 of Shoukri and Donner (2001), the intrarater reliabilities by formulae 10 and 11, and their variances by formula 12.

B. PAIRED OBSERVATIONS: THREE OR MORE CATEGORIES, NOT ORDERED

This module is appropriate for the analysis of paired observations (in different subjects or the same subject) where the dependent variable is a nominal-scale one (i.e., with categories that are not ordered). It appraises differences and agreement between the two sets of observations. It can be used to analyse matched-control trials and case-control studies, before-after studies, and other comparisons of paired subjects or observations, such as comparisons of husbands and wives, and diagnoses of the same individuals by two different observers or diagnostic techniques.

The number of categories must be entered, and then the numbers of pairs with each combination of findings are entered in a *k* x *k* table in which the paired sets of observations are arbitrarily designated A and B. The numbering and sequence of the categories is arbitrary, except that if there is a reference category it should be given the highest number. Numbers of pairs are entered, not numbers of observations.

If the data are stratified, enter each stratum in turn; for *meta-analyses*, enter each study as a separate stratum. Click on "All strata" whenever combined results are required.

For each table, the program provides **tests for the difference** between the two sets of observation **(**extended McNemar test, and Stuart-Maxwell and Bhapkar tests for **marginal homogeneity)** , showing the sources of disagreement (if there are up to seven categories), and computes **odds ratios and related tests**, ***kappa* and related results**, and a measure of the **distinguishability of categories.**

For *stratified data*, the program provides overall **tests for the difference** (based on the pooled data) and ***kappa* and related results.**

The module also provides an option for the **comparison of binocular data** (i.e. findings concerning the presence of an abnormality or other attribute in both eyes) reported by two raters. Agreement between the raters is expressed by *kappa* coefficients, and McNemar tests assess the difference between the raters, the difference between the eyes, and rater-eye interaction.

Tests for the difference between paired observations

For each table, the program provides extended McNemar tests for off-diagonal symmetry and the Stuart-Maxwell and Bhapkar testS for marginal heterogeneity. If stratified data are entered, extended McNemar tests are done on the combined (pooled) data.

The *extended McNemar ("symmetry") test* (Bowker's test for off-diagonal symmetry) tests the symmetry of the findings; e.g. for categories 1 and 2 (and similarly for each other pair of categories) it tests whether the probability that the observation will be in category 1 in one set

of observations and in category 2 in the second is the same as the probability of the reverse combination, namely category 2 in the first set and category 1 in the second. Ordinary (Pearson's) and log-likelihood chi-squares are computed. Comparisons of zero cells do not contribute to the chi-square. If there are comparisons of zero cells, alternative P values are shown, based on different degrees of freedom, namely the total number of pairs compared (Bowker 1948) and this total number reduced by the number of zero-cell comparisons (Hoenig *et al*. 1995, Evans and Hoenig 1998).

As a guide to the sources of disagreement (Maxwell 1970), the contribution that each pair of categories makes to a significant McNemar chi-square ($P < 0.05$) is reported (if there are up to seven categories).

The *Stuart-Maxwell and Bhapkar tests* for marginal heterogeneity (Stuart 1955, Maxwell 1970, Bhapkar 1966) test the hypothesis that the probabilities of the various categories are the same in the two sets of observations (are the totals of the columns the same as the totals of the rows?) The Bhapkar test is more powerful than the Stuart-Maxwell test if the sample is small; for larger samples the two tests are essentially equivalent (Dunnigan 2013). For certain sets of data, these tests are impractical (Dunnigan 2013), and are omitted. The specific categories that manifest significant differences can be pin-pointed (see "Odds ratios and related tests", below).

The results of the extended McNemar and Stuart-Maxwell or Bhapkar tests cannot be expected to be the same, except that symmetry implies marginal homogeneity (but not vice versa).

Odds ratios and related tests

The program provides odds ratios based on the contrast between each pair of categories (if there are up to 10 categories). If the odds ratio based on the contrast between two categories, e.g. 1 and 2 (displayed as "1:2") is above 1, this means that the odds in favour of 1 rather than 2 are higher in sample A than in sample B.

The consistency of these odds ratios based on pairs of categories is tested. For example, if the odds ratio for category 1 versus category 2 is 3.0 and the odds ratio for category 2 versus category 3 is 4.0, the odds ratio for category 1 versus category 3 would be expected to be 12.0. Inconsistency with such expectations suggests that the odds ratios may be modified by the matching variables (Pike, Casagrande, and Smith 1975). A low P value is indicative of inconsistency.

Maximum-likelihood estimates of mutually consistent odds ratios based on the contrast between each pair of categories are computed; these estimates are not very meaningful if the test points to mutual inconsistency.

The program also computes odds ratios based on a comparison of each category with all other categories combined, and does McNemar tests to appraise their significance; alternative P-values are provided for tests of hypotheses formulated before and after seeing the results. If the Stuart-Maxwell test shows significant disagreement, these odds ratios and tests pinpoint the specific disagreements (i.e., about specific categories) that are responsible.

Confidence intervals are displayed for odds ratios contrasting each category with the reference category (the category with the highest category number), assuming mutual consistency.

Kappa and related results

The program computes an overall *kappa* value (for the complete set of categories), and a separate *kappa* value for each category. In each instance, a one-tailed test is done, indicating whether *kappa* is significantly higher than zero. If *kappa* is 0.4 or more, a second test is done, indicating whether it is significantly higher than 0.4; and if it is 0.6 or more, a third test is done, indicating whether it is significantly higher than 0.6. Confidence intervals for *kappa* are estimated from its standard error. Flight and Julious (2014) emphasize that because of "the disagreeable behaviour of the kappa statistic", it should always be interpreted in conjunction with the percentage agreement, prevalence-adjusted bias-adjusted kappa, prevalence index, bias index and maximum attainable kappa (see below).

Paradoxical values of *kappa* may occur because of bias (systematic one-sided variation between two ratings) – indicated by the extended McNemar test (see above) – or a skewed distribution (inequality between the prevalences of the categories in the two samples). Two adjusted values of the overall kappa – BAK (*bias-adjusted kappa*) and PABAK (*prevalence-adjusted bias-adjusted kappa*) – are therefore computed (Byrt *et al.* 1993*)*. These adjusted values are conditional on the observed percentage agreement. BAK is the value that *kappa* would take if there were no systematic one-sided variation between the ratings; it is equivalent to Scott's *pi* coefficient of agreement (Scott 1955). Low *kappa* values are likely to be affected by such bias. PABAK is the value that *kappa* would take if, in addition, the prevalence of each category (as expressed by the mean of the two raters' totals for the category) was equal. PABAK may be useful in appraising agreement when the percentage agreement is high and *kappa* is paradoxically low; it approximates to the highest possible *kappa* if the percentage agreement is above about 50% (Lantz and Nebenzahl 1996). PABAK is called *kappa-no*r by Lantz and Nebenzahl (1996), and is equivalent to Maxwell's *RE* (random error) coefficient of agreement (Maxwell 1977) and Bennett's S coefficient (Bennett et al. 1954). It should be noted that simulation studies have suggested that PABAK may substantially overestimate agreement (Hoehler 2000).

The program also displays the *maximum attainable overall kappa* consistent with the marginal totals, i.e. consistent with the observed level of bias.

Kappa is generally used to measure the agreement between two ratings (by different observers or tests, or by the same observer on different occasions) of the same individuals. In addition to this use as a measure of reliability, it may be used to measure concordance in other situations where paired samples are compared (Fleiss *et al*. 2003: 617-618). In a matched case-control study or matched-control trial, *kappa* may serve as an indication of the effectiveness of a matching procedure – it indicates the extent to which the findings in matched pairs are more similar than findings in individuals from different pairs (Fleiss *et al*. 2003: 618). *Kappa*, like other measures of agreement, reflects the agreement concerning specific subjects by specific raters, and can be generalized to a broader group only if the subjects are representative of the broader group. As a measure of inter-rater reliability, its value is determined by the selection of raters. Uses and misuses of *kappa* in epidemiology are discussed by (among others) Sim and Wright (2005), MacLure and Willett (1987), Thompson and Walter (1988a, 1988b) Kraemer and Bloch (1988), and Gwet (2010: 30-34)

The probability of chance agreement is taken into account in the calculation of *kappa*. A value of 1 indicates perfect agreement (after allowing for this probability of chance agreement) between ratings; 0 indicates no agreement other than what can be attributed to chance, and a negative value indicates less than chance agreement. Fleiss *et al*. (2003) suggest that a value of 0.75 or more indicates excellent agreement, and 0.40 or less indicates poor agreement. Cicchetti and Sparrow (1981) divide Fleiss's 0.40–0.74 group into 0.60–0.74: good; and 0.40–0.59: fair. Alternative guidelines are: over 0.80, very good agreement; 0.61–0.80, good; 0.41–0.60, moderate; 0.21–0.40, fair; and 0.20 or less, poor agreement (Landis and Koch 1977, Altman 1991).

The *percentage agreement* is also shown. This is the percentage of individuals who are placed in the same category by both ratings, and (unlike *kappa*) it is not corrected for chance agreement. In a study in which the same individuals are rated by two observers, this is the percentage of subjects who are placed in the same category by both raters). Its significance is tested, using a one-sided test of the null hypothesis that agreement is not more than might be expected by chance. The percentage agreement is also shown separately for each category (if there are up to six categories), together with the *percentage of positive agreement* and the *percentage of negative agreement*. The percentage of positive agreement is the percentage of "yes" ratings (for a specific category) that are paralleled by a "yes" rating by the other observer or test, among all "yes" ratings for that category; and the percentage of negative agreement is the percentage of "no" ratings (for a specific category) that are paralleled by a "no" rating by the other observer or test, among all "no" ratings for that category. In clinical practice, the percentage of agreement for a specific rating represents the probability that, if a subject has been given that rating by a typical observer, another typical observer will concur.

If *stratified data* are entered (e.g. observations of individuals in different age groups), the heterogeneity of the overall kappa values in the different strata is tested, measures of heterogeneity (see above) are provided, two estimates of the overall kappa are computed, with their confidence intervals. The first estimate of the overall kappa is precision-based; it is produced by weighting each kappa by the inverse of its variance (Fleiss *et al*. 2003: 602). The second estimate is obtained by weighting the kappa values by the sizes of the samples in the strata. A simulation study suggests that this is preferable to the precision-based method if kappa is not zero (Barlow et al. 1991). A heterogeneity test is done, and supplemented by two measures of heterogeneity, H and I-squared (Higgins and Thompson 2002), with their approximate 95% intervals. An H value of less than 1.2 suggests absence of noteworthy heterogeneity, whereas a value exceeding 1.5 suggests its presence, even if the heterogeneity test is not significant. I-squared expresses the proportion of variation that can be attributed to heterogeneity (in a meta-analysis, to interstudy variation) rather than to sampling error; a value greater than 50% may be considered substantial heterogeneity (Higgins and Green 2006). Overall values of the percentage agreement are reported. These are based on the pooled data; this is equivalent to weighting the stratum-specific values by sample sizes.

Other measures of chance agreement

The A*C1 statistic* is, like *kappa*, a chance-corrected measure of the extent of agreement between raters (Gwet 2002a, 2002b, 2008, 2010). Its main difference from *kappa* is that it bases the probability of agreement-by-chance on only the hard-to-classify subjects, using a model that in effect estimates their number. AC1 has been recommended for use instead of kappa on the grounds that its estimate of the probability of chance agreement is more

appropriate, and that it is less influenced by differences in the propensity to give positive ratings and differences in the prevalences of the response categories. It is hence more robust, avoiding paradoxical results. Monte Carlo simulation has demonstrated that it is less biased and has a smaller variance than *kappa*, the G-index, or the *pi* coefficient (Gwet 2008). But along with recommendations that it is preferable to *kappa (e.g.* Lombard *et al.* 2004; Stegmann and Lucking 2005; Haley *et al*. 2008, Wongpakaran *et al.* 2013), Blood and Spratt (2007) warn that "…the AC1 and AC2 statistics … remain infants in the statistical world … as is always the case with new statistics, caution should be exercised in their use and further examination should occur before they are adopted as the standard'".

The program calculates AC1's standard error and 90%, 95%, and 99% confidence intervals on the assumption that the subjects are a random sample of an infinitely large population. If the sample is drawn from a defined target population and the sampling fraction is known, the correct standard error can be computed as $\sqrt{[SE^2 \times (1 - f)]}$, where SE is the reported standard error and *f* is the sampling fraction, e.g. 0.1 (1 in 10).

The G-*index*, or Brennan-Prediger coefficient (Brennan and Prediger 1981, Gwet 2010: 38) is a simple coefficient that bases the chance-probability of agreement only on the number of response categories. The program calculates the G-index's standard error and 90%, 95%, and 99% confidence intervals on the assumption that the subjects are a random sample of an infinitely large population. If the sample is drawn from a defined target population and the sampling fraction is known, the correct standard error can be computed as $\sqrt{[SE^2 \times (1 - f)]}$, where SE is the reported standard error and *f* is the sampling fraction, e.g. 0.1 (1 in 10).

Scott's *pi coefficient* (Scott 1955, Gwet 2010: 21) differs from *kappa* in that it is based on marginal probabilities (the probabilities that each response category will be selected) that are common to both raters, not those that are specific to each rater (Gwet 2010: 38) The program calculates the *pi* coefficient's standard error and 90%, 95%, and 99% confidence intervals on the assumption that the subjects are a random sample of an infinitely large population. If the sample is drawn from a defined target population and the sampling fraction is known, the correct standard error can be computed as $\sqrt{[SE^2 \times (1 - f)]}$, where SE is the reported standard error and *f* is the sampling fraction, e.g. 0.1 (1 in 10).

Comparison of binocular data

This option compares two raters' reports of binocular findings - that is, their findings concerning the presence of an abnormality or other "yes-no" attribute in both eyes of the same subjects. Account is taken of the positive correlation generally present between observations made in fellow eyes (Oden 1991). The procedure may also be used (with appropriate construal of the terms "right eye" and "left eye") in comparisons of two raters' observations of other paired bodily structures, or in studies where a first-eye/second-eye grouping is more relevant than a right/left grouping.

The option requires the entry of the raters' findings in the two eyes of the same subjects, in a 4x4 cross-tabulation showing (for each rater) the numbers of subjects with a positive finding in both eyes, in the right eye only, in the left eye only, and in neither eye.

Kappa is computed for each eye separately, and for the pooled data on both eyes. Confidence intervals (90%, 95%, and 99%) are estimated for the kappa values, and (where appropriate)

significance tests are performed, comparing kappa with prespecified values of 0.4 and 0.6. The confidence intervals are appropriate if the samples are large.

Modified McNemar tests (Schouten 1993) assess the difference between the raters, the difference between the eyes, and rater-eye interaction.

Distinguishability of categories

A measure of the distinguishability of pairs of categories is computed. This may be useful in a methodological study in which the matched observations represent separate ratings. The value is 100% if there are no disagreements, and zero if disagreements outnumber agreements. An average distinguishability index is reported, as well as the distinguishability of each pair of categories.

METHODS

Tests for the difference between paired observations

The *extended McNemar ('symmetry') test* is described by Bowker (1948), Everitt (1977: 114-115) and Zar (1998: formula 9.22). There are $k(k\text{-}1)/2$ degrees of freedom (where k = number of categories). Corresponding cells that both have zero values are omitted from the calculation of this chi-square, and if there are such comparisons an alternative P is computed, after reducing the degrees of freedom by the number of zero-cell comparisons (Hoenig *et al.* 1995, Evans and Hoenig 1998).

The contributions that a specific pair of categories (*i* and *j*) makes to a significant chi-square ($P < 0.05$) are computed by formula 6 of Maxwell (1970):

$$\text{chi-square} = (n_{ij} - n_{ji})^2 \ / \ (n_{ij} + n_{ji})$$

where n_{ij} = the number in the cell in column *i* of row *j*
n_{ji} = the number in the cell in column *j* of row *I*

In McNemar tests for single categories, the degrees of freedom are defined as $k\text{-}1$ (where k = number of categories) for testing *a posteriori* hypotheses (Fleiss *et al.* 2003: 382).

The *Stuart-Maxwell chi-square test* (Stuart 1955, Maxwell 1970) is performed if there are up to 20 categories. It is computed by a matrix operation (Fleiss *et al.* 2003: 381-383; Everitt 1977: 115-116. The test is not done if any cell is the only non-zero cell both in its column and in its row, or in 3x3 tables that have more than 3 zero cells. To avoid computational problems in extreme situations, some zero divisors are replaced by 0.000001 (the "perturbation method" used by Dunnigan 2013).

The *Bhapkar test* (Bhapkar 1966) is performed if there are up to 20 categories, and is also computed by a matrix operation.

Odds ratios and related tests

If there is a zero observed frequency of pairs in any cell, adjusted odds ratios are computed, by adding 0.5 in each cell.

The test for the consistency of odds ratios between pairs of categories, the maximum-likelihood estimation of mutually consistent odds ratios, and the estimation of confidence intervals are described by Pike, Casagrande and Smith (1975).

Kappa and related results

The basic formulae are provided by Fleiss *et al*. (2003: chapter 18). Kappa for single categories and for the total distribution (*overall kappa*) are calculated by formulae 18.10 to 18.12. For tests of the null hypothesis that

kappa is zero (formulae 18.14 and 18.35), the standard error (for an underlying zero value of *kappa*) is calculated by formula 18.13. For tests of the hypothesis that *kappa* has an underlying value other than zero, and for confidence intervals, the standard error appropriate for non-zero values is calculated by formulae 18.15 to 18.18. Confidence intervals are estimated from the standard error (if the upper confidence limit exceeds 1, it is reduced to 1).

The *maximum attainable value* of *kappa* is computed by calculating *kappa* when taking the marginal totals as fixed but modifying the body of the table so as to represent the maximum possible agreement, by using, for each cell indicating agreement, the smaller of the two relevant marginal frequencies.
Bias is appraised by the extended McNemar (symmetry) test (see above). BAK (*bias-adjusted kappa*) and PABAK (*prevalence-adjusted bias-adjusted kappa*) use the methods described by Byrt *et al.* (1993).

In the combined analysis of several samples or strata, the estimate of the supposed *common or overall value* of *kappa* is calculated in two ways: by computing a weighted mean, using the inverse of the variance of each *kappa* as its weight (Fleiss *et al*. 2003: formula 1.21); and by computing a weighted mean, using the size of the stratum as the weight. The confidence intervals of the common kappa are estimated by formula 18.23 of Fleiss *et al*. (2003).

The *heterogeneity test* is based on formula 18.22 of Fleiss *et al*. (2003). The *measures of heterogeneity* (Higgins and Thompson 2002) are *H* and *I-squared*. *H* is computed by Higgins and Thompson's formula 6, and increased to 1 (indicating absence of heterogeneity) if it less than 1. A test-based interval is computed by Method III. *I-squared* and its 95% interval are computed from *H*, using formula 10.

The significance of the *percentage agreement* is tested by a binomial test comparing the total number of complete agreements with the number expected by chance (Sheskin 2007: 729-730).

Some computations are omitted if division by zero or other problems are encountered. In some instances, zero values are changed to 0.00001 to permit computation.

Other chance-corrected measures of association

Gwet's AC1 is calculated by formula 4.1 of Gwet (2010: 61), and its variance by formula 5.7 of Gwet (2010: 94). The program calculates AC1's standard error on the assumption that the subjects are a random sample of an infinitely large population. If the sample is drawn from a defined target population and the sampling fraction is known, the correct standard error can be computed as $\sqrt{[SE^2 \times (1 - f)]}$, where SE is the reported standard error and *f* is the sampling fraction, e.g. 0.1 (1 in 10).

Brennan and Prediger's *G-index* is calculated by formula 2.18 of Gwet (2010: 38). and its variance by formula 5.10 of Gwet (2010: 95). The program calculates the G-index's standard error on the assumption that the subjects are a random sample of an infinitely large population. If the sample is drawn from a defined target population and the sampling fraction is known, the correct standard error can be computed as $\sqrt{[SE^2 \times (1 - f)]}$, where SE is the reported standard error and f is the sampling fraction, e.g. 0.1 (1 in 10).

Scott's *pi coefficient* is calculated by formula 2.6 of Gwet (2010: 21).

Comparison of binocular data

Single-eye kappa estimates and their standard errors are computed by the methods described by Fleiss *et al.* (2003: chapter 18). *Kappa* is calculated by formula 18.12. For tests of the study hypotheses that *kappa* exceeds 0.4 or 0.6 (formula 13.19), and for confidence intervals (formula 18.20), the standard error appropriate for non-zero values is calculated by formula 18.15 .

The *pooled kappa* estimate is computed from a 2x2 table containing the sums of the single-eye observed frequencies, and the sums of the single-eye expected frequencies (Schouten 1993, Table IIIc), using formula 18.12 of Fleiss (2003). As pointed out by Schouten, this *kappa* is identical to the weighted *kappa* computed from the original 4x4 cross-tabulation (Table I), using weights of 1 for complete agreement on both eyes, 0.5 for agreement on only one eye, and 0 for disagreement on both eyes. The large-sample nonzero standard error is computed from this 4x4 table, using the method described by Fleiss *et al.* (1969) for weighted *kappa* (formulae 15 to 21), with the above weights. This standard error is used for tests of the study hypotheses that *kappa* exceeds 0.4 or 0.6 (Fleiss *et al.* 2003, formula 18.19), and for estimating confidence intervals (formula 18.20).

The modified McNemar tests use the formula

$$\text{chi-square} = (|M| - 0.25)^2 / B$$

where *M* and *B* are computed from the original 4x4 table by the method described by Schouten (1993: 2212 and 2213). For the comparison of raters, *M* is derived from the weights in Schouten's Table IVc, and *B* from the squares of these weights. For the comparison of eyes, *M* is derived from the weights in Table V, and *B* from the squares of these weights. For the test of rater-eye interaction, *M* is derived from the weights in Table VI, and *B* from the squares of these weights.

Distinguishability of categories

This measure is computed by the method described by Darroch and McLeod (1986), using maximum-likelihood estimates (without assuming quasi-symmetry) in formula 21b (Shoukri 2000).

C. PAIRED OBSERVATIONS: THREE OR MORE ORDERED CATEGORIES

This module is appropriate for the analysis of paired observations (in different subjects or the same subject) where the dependent variable has three or more categories that fall into a sequence. It appraises differences and agreement between the two sets of observations. It can be used to analyse matched-control trials and case-control studies, before-after studies, and other comparisons of paired subjects or observations, such as comparisons of husbands and wives, and diagnoses of the same individuals by two different observers or diagnostic techniques.

The number of categories must be entered, and then the numbers of pairs with each combination of findings are entered in a *k* x *k* table in which the paired sets of observations are arbitrarily designated A and B. The categories must be entered in the correct sequence; if there is a reference category it should be given the highest number. Numbers of pairs are entered, not numbers of observations. S*cores* of 1, 2, 3, etc. are allotted to the categories (for use in computing a weighted *kappa*), but these default scores can optionally be changed to numbers that are believed to better express the relative closeness of the categories.

If the data are stratified, enter each stratum in turn; for *meta-analyses*, enter each study as a separate stratum. Click on "All strata" whenever combined results are required.

For each table, the program provides **tests for the difference** between the two sets of observation , including tests appropriate for ordered categories (Mann-Whitney test, Fleiss-Everitt test, Wilcoxon signed-rank test, permutation test, and a McNemar test of overall bias or directional change) and tests that ignore the sequence of the categories (extended McNemar test, and Bhapkar tests), and computes **odds ratios and related tests**, ***kappa* and related results**, Gwet's AC2 coefficient, other measures of agreement (Brennan and Prediger's G-index and Scott's *Pi* coefficient).a measure of the **distinguishability of categories,** and **rank correlation coefficients and other measures of ordinal association.**

For *stratified data*, the program provides overall **tests for the difference**, a **heterogeneity test**, a **generalized odds ratio**, and ***kappa* and related results.**

Tests for the difference between paired observations

The Mann-Whitney test, Fleiss-Everitt test for three ordered categories, Wilcoxon signed-rank test, and permutation test take account of the sequence of the categories, whereas the extended McNemar test and the Stuart-Maxwell and Bhapkar tests ignore their sequence.

The *Mann-Whitney test for paired data* (Agresti 1984: 208-209) is a large-sample test that compares the frequencies in two sets of paired observations; if the data are arranged in the format of a square contingency table these are the marginal distributions. A two-tailed P-value is shown, labelled as "approximate" if there are under 50 pairs of observations. If stratified data are entered, the test is also done on the combined data, after weighting the test

statistics in the strata in three different ways – equally, by the sample sizes in the strata, and by the square roots of the sample sizes – as well as a simple test on the pooled data.

The *Fleiss-Everitt test*, which is done if there are three categories, tests whether in one set of observations there tend to be more values at one end of the scale and fewer at the other, compared with the other set of observations (Fleiss *et al*. 2003: 382-384). If stratified data are entered, the test is also done on the combined (pooled) data.

The *Wilcoxon signed-ranks test* (Siegel and Castellan 1988: 87-95) tests whether the median discrepancy between paired observations is zero. It is done only if the differences between each pair of adjacent categories can be assumed to be equal, that is, if the scores allotted to adjacent categories are equally spaced. The test is appropriate if the differences between paired observations are an acceptable basis for ranking the differences in the characteristic that is measured.

The *permutation test for matched pairs* (Siegel and Castellan 1988: 95-100 is appropriate for interval-scale variables, and is therefore done only if the scores allotted to adjacent categories are equally spaced. It is not done if there are more than 20 pairs. The test provides exact P-values; one-tailed P-values are displayed if $P < .05$; the one-tailed value is doubled to provide a two-tailed value.

The *McNemar test of overall bias or directional change* examines the paired observations, to test whether the values in one set of observations tend to be significantly higher than those in the other set. This may be useful when (for example) comparing two methods of study, or determining whether there was a change in values after some treatment.

The *extended McNemar ("symmetry") test* (Bowker's test for off-diagonal symmetry) tests the symmetry of the findings; e.g. for categories 1 and 2 (and similarly for each other pair of categories) it tests whether the probability that the observation will be in category 1 in one set of observations and in category 2 in the second is the same as the probability of the reverse combination, namely category 2 in the first set and category 1 in the second. Ordinary (Pearson's) and log-likelihood chi-squares are computed. The test is equivalent to the test for goodness of fit with a symmetry model described by Agresti (1984: 202). If stratified data are entered, the test is also done on the combined (pooled) data. Comparisons of zero cells do not contribute to the chi-square. If there are comparisons of zero cells, alternative P values are shown, based on different degrees of freedom, namely the total number of pairs compared (Bowker 1948) and this total number reduced by the number of zero-cell comparisons (Hoenig *et al*. 1995, Evans and Hoenig 1998).

As a guide to the sources of disagreement (Maxwell 1970), the contribution that each pair of categories makes to a significant McNemar chi-square ($P < 0.05$) is reported (if there are up to seven categories).

The *Stuart-Maxwell and Bhapkar tests* for marginal heterogeneity (Stuart 1955, Maxwell 1970, Bhapkar 1966) test the hypothesis that the probabilities of the various categories are the same in the two sets of observations (are the totals of the columns the same as the totals of the rows?) The Bhapkar test is more powerful than the Stuart-Maxwell test if the sample is small; for larger samples the two tests are essentially equivalent (Dunnigan 2013). For certain sets of data, these tests are impractical (Dunnigan 2013), and are omitted. The

specific categories that manifest significant differences can be pin-pointed (see "Odds ratios and related tests", below).

The results of the extended McNemar and Stuart-Maxwell or Bhapkar tests cannot be expected to be the same, except that symmetry implies marginal homogeneity (but not vice versa).

Heterogeneity test

If stratified data are entered, goodness of fit with a symmetry model is tested twice, once using the pooled data, and once using the sum of the goodness-of-fit chi-squares in the separate strata. The difference between the two goodness-of-fit chi-squares is an indication of the effect of the stratifier variable(s), and is displayed as a heterogeneity test. The result should be interpreted with caution, since test has a low power. The symmetry model is based on the assumption that the probability of discrepant pairs in which the case is in category 1 and the control in category 2 is the same as the probability of pairs in which the case is in category 2 and the control in category 1 (and similarly for other pairs of categories); i.e. the odds ratio (as generally computed for paired data) is 1 (Agresti 1984: 202).

The heterogeneity of *kappa* values is also tested (see below).

Odds ratios and related tests

The *generalized odds ratio* or GOR (the odds ratio for ordinal data) is displayed. This is the odds in favour of a higher score in one sample than in the other, i.e. the ratio of pairs with a higher score in one sample to pairs with a higher score in the other sample. It is assumed that this odds ratio is the same for each pair of categories (Agresti 1984: 203). The ratios in both directions are displayed, with their 90%, 95% and 99% confidence intervals. If stratified data are entered, the assumed common values of the GOR are displayed (with their 96% confidence intervals); these are weighted averages of the stratum-specific GOR values, and are of questionable value is there is marked heterogeneity.

The program provides odds ratios based on the contrast between each pair of categories (if there are up to 10 categories). If the odds ratio based on the contrast between two categories, e.g. 1 and 2 (displayed as "1:2") is above 1, this means that the odds in favour of 1 rather than 2 are higher in sample A than in sample B.

The consistency of these odds ratios based on pairs of categories is tested. For example, if the odds ratio for category 1 versus category 2 is 3.0 and the odds ratio for category 2 versus category 3 is 4.0, the odds ratio for category 1 versus category 3 would be expected to be 12.0. Inconsistency with such expectations suggests that the odds ratios may be modified by the matching variables (Pike, Casagrande, and Smith 1975). A low P value is indicative of inconsistency.

Maximum-likelihood estimates of mutually consistent odds ratios based on the contrast between each pair of categories are computed; these estimates are not very meaningful if the test points to mutual inconsistency.

The program also computes odds ratios based on a comparison of each category with all other categories combined, and does McNemar tests to appraise their significance; alternative P-values are provided for tests of hypotheses formulated before and after seeing the results.

Confidence intervals are displayed for odds ratios contrasting each category with the reference category (the category with the highest category number), assuming mutual consistency.

Kappa and related results

As measures of the agreement between the matched observations, the program provides two *weighted kappa* estimates (which take account of the sequence of the categories), an ordinary *overall kappa* (for the complete set of categories, but ignoring the sequence), and (if there are up to six categories) separate kappa values for each category. In each instance, a one-tailed test is done, indicating whether kappa is significantly higher than zero. If kappa is 0.4 or more, a second test is done, indicating whether it is significantly higher than 0.4; and if it is 0.6 or more, a third test is done, indicating whether it is significantly higher than 0.6. Confidence intervals are estimated from the standard error. Flight and Julious (2014) emphasize that because of "the disagreeable behaviour of the kappa statistic", it should always be interpreted in conjunction with the percentage agreement, prevalence-adjusted bias-adjusted kappa, prevalence index, bias index and maximum attainable kappa (see below).

In the computation of weighted kappa, the weight given each pair of observations depends on the size of the difference between the categories in which the pair-mates fall. Default scores of 1, 2, 3, etc. are allotted to the categories for this purpose; but these scores can optionally be changed to numbers that are believed to better express the relative closeness of the categories. Two weighting schemes are used – linear and quadratic (Sim and Wright 2005). Either may be chosen; but since the results differ, study reports should specify the method used. The linear weights are proportional to the size of the difference between scores, whereas the quadratic weights are proportional to the square of the difference. If there are 4 categories, the linear weight is 0.67 if the difference between scores is 1, 0.33 if it is 2, and 0 if it is 3. Quadratically-weighted *kappa* values tend to increase with the number of categories, whereas linearly-weighted values are less sensitive (Brenner and Kliebsch 1996).

Paradoxical values of *kappa* may occur because of bias (systematic one-sided variation between two ratings) – indicated by the extended McNemar test (see above) – or a skewed distribution (inequality between the prevalences of the categories in the two samples). An adjusted value of the overall kappa –PABAK (*prevalence-adjusted bias-adjusted kappa*) – is therefore computed (Byrt *et al.* 1993*).* This adjusted value is conditional on the observed percentage agreement. BAK is the value that *kappa* would take if there were no systematic one-sided variation between the ratings; it is equivalent to Scott's *pi* coefficient of agreement (Scott 1955). Low *kappa* values are likely to be affected by such bias. PABAK is the value that *kappa* would take if there were no systematic one-sided variation between the ratings and, in addition, the prevalence of each category (as expressed by the mean of the two raters' totals for the category) was equal. PABAK may be useful in appraising agreement when the percentage agreement is high and *kappa* is paradoxically low; it approximates to the highest possible *kappa* if the percentage agreement is above about 50% (Lantz and Nebenzahl 1996). PABAK is called *kappa-no*r by Lantz and Nebenzahl (1996), and is equivalent to Maxwell's *RE* (random error) coefficient of agreement (Maxwell 1977) and Bennett's *S* coefficient

(Bennett *et al*. 1954). It should be noted that simulation studies have suggested that PABAK may substantially overestimate agreement (Hoehler 2000).

The program also displays the *maximum attainable overall kappa* consistent with the marginal totals, i.e. consistent with the observed level of bias.

Kappa is generally used to measure the agreement between two ratings (by different observers or tests, or by the same observer on different occasions) of the same individuals. In addition to this use as a measure of reliability, it may be used to measure concordance in other situations where paired samples are compared (Fleiss *et al*. 2003: 617-618). In a matched case-control study or matched-control trial, *kappa* may serve as an indication of the effectiveness of a matching procedure – it indicates the extent to which the findings in matched pairs are more similar than findings in individuals from different pairs (Fleiss *et al*. 2003: 617-618).

Kappa, like other measures of agreement, reflects the agreement concerning specific subjects by specific raters, and can be generalized to a broader group only if the subjects are representative of the broader group. As a measure of inter-rater reliability, its value depends on the choice of raters. Uses and misuses of *kappa* in epidemiology are discussed by (among others) Sim and Wright (2005), MacLure and Willett (1987), Thompson and Walter (1988a, 1988b), Kraemer and Bloch (1988), Bloch and Kraemer (1989) and Gwet (2010).

The probability of chance agreement is taken into account in the calculation of *kappa*. A value of 1 indicates perfect agreement (after allowing for this probability of chance agreement) between ratings; 0 indicates no agreement other than what can be attributed to chance, and a negative value indicates less than chance agreement. Fleiss *et al*. (2003) suggest that a value of 0.75 or more indicates excellent agreement, and 0.40 or less indicates poor agreement. Cicchetti and Sparrow (1981) divide Fleiss's 0.40–0.74 group into 0.60–0.74: good; and 0.40–0.59: fair. Alternative guidelines are: over 0.80, very good agreement; 0.61–0.80, good; 0.41–0.60, moderate; 0.21–0.40, fair; and 0.20 or less, poor agreement (Landis and Koch 1977, Altman 1991). These levels may be taken into account in the appraisal of confidence intervals, e.g. by seeing whether the lower confidence limit lies above 0.40 (Basu and Basu 1995).

The *percentage agreement* is also shown. This is the percentage of individuals who are placed in the same category by both ratings, and (unlike *kappa*) it is not corrected for chance agreement. In a study in which the same individuals are rated by two observers, this is the percentage of subjects who are placed in the same category by both raters). Its significance is tested, using a one-sided test of the null hypothesis that agreement is not more than might be expected by chance. The percentage agreement is also shown separately for each category (if there are up to six categories), together with the *percentage of positive agreement* and the *percentage of negative agreement*. The percentage of positive agreement is the percentage of "yes" ratings (for a specific category) that are paralleled by a "yes" rating by the other observer or test, among all "yes" ratings for that category; and the percentage of negative agreement is the percentage of "no" ratings (for a specific category) that are paralleled by a "no" rating by the other observer or test, among all "no" ratings for that category. In clinical practice, the percentage of agreement for a specific rating represents the probability that, if a subject has been given that rating by a typical observer, another typical observer will concur.

If *stratified data* are entered (e.g. observations of individuals in different age groups), the heterogeneity of the overall kappa values in the different strata is tested, measures of heterogeneity (see above) are provided, and two estimates of the overall kappa are computed, with their confidence intervals. The first estimate of the overall kappa is precision-based; it is produced by weighting each kappa by the inverse of its variance (Fleiss *et al*. 2003: 602). The second estimate is obtained by weighting the kappa values by the sizes of the samples in the strata. A simulation study suggests that this is preferable to the precision-based method if kappa is not zero (Barlow *et al*. 1991). A heterogeneity test is done, and supplemented by two measures of heterogeneity, *H* and *I-squared* (Higgins and Thompson 2002), with their approximate 95% intervals. An *H* value of less than 1.2 suggests absence of noteworthy heterogeneity, whereas a value exceeding 1.5 suggests its presence, even if the heterogeneity test is not significant. *I-squared* expresses the proportion of variation that can be attributed to heterogeneity (in a meta-analysis, to interstudy variation) rather than to sampling error; a value greater than 50% may be considered substantial heterogeneity (Higgins and Green 2006). Overall values of the percentage agreement are reported. These are based on the pooled data; this is equivalent to weighting the stratum-specific values by sample sizes.

Gwet's AC2

The AC2 statistic is, like weighted *kappa*, a chance-corrected measure of the extent of agreement between raters concerning an ordered set of response categories (Gwet 2010: 76-78, 80-81). It is a weighted version of the AC1 statistic, treating various kinds of disagreement as partial agreements. The program assumes that the successive categories are equally spaced, with scores of 1, 2,3 etc. A quadratic weight is assigned to each pair of scores, reflecting the degree of agreement. Its main difference from *kappa* is that it bases the probability of agreement-by-chance on only the (estimated) hard-to-classify subjects using a model that estimates their number.

Distinguishability of categories

A measure of the distinguishability of pairs of categories is computed. This may be useful in a methodological study in which the matched observations represent separate ratings. The value is 100% if there are no disagreements, and zero if disagreements outnumber agreements. An average distinguishability index is reported, as well as the distinguishability of each pair of categories.

Rank correlation coefficients and other measures of ordinal association

Kendall's and Spearman's rank correlation coefficients (*tau b* and *rho*, respectively) are computed (with their standard errors and 95% confidence intervals). These have different numerical values but are similar in their ability to detect associations (Siegel and Castellan 1988: 251). The other measures of ordinal association that are provided are Goodman and Kruskal's *gamma* and Somers' asymmetric *D*, which may be regarded as measures of how effectively the order of a pair of observations with respect to one observation can be predicted from their order with respect to the other observation (see Hildebrand, Laing, and Rosenthal 1977). The Somers' *D* statistics are appropriate when one of the observations is clearly the dependent one, e.g. one that comes later in time; Somers' *Dxy* is appropriate when A is dependent, and *Dyx* when B is dependent.

Tau, Kruskal's *gamma*, and Somers' *D* depend on a comparison of the ranks of the paired observations. All possible pairs are taken into account in the computation of *tau*, whereas pairs that tie are disregarded in the calculation of *gamma*, and pairs that tie with respect to one (the independent) observation are omitted from the computation of Somers' *D*. *Tau* is the geometric average of *Dxy* and *Dyx*.

METHODS

Maximum categories = 60 [50 for kappa].

Tests for the difference between paired observations

The *Mann-Whitney test for paired data* is described by Agresti (1984: 208-209). If stratified data are entered, the results of the tests in the strata are combined by Stouffer's method (Stouffer *et al.* 1949: 5; DeMets 1987), based on weighted averages of the test results in the strata, using three different sets of weights for the *Z* values –weighting them equally, by the sample sizes in the strata, and by the square roots of the sample sizes. A simple test is also done on the combined (pooled) data.

The *Fleiss-Everitt test* for ordered categories is described by Fleiss *et al.* (2003: 382-384).

The *Wilcoxon signed-ranks test* uses the formula provided by Siegel and Castellan (1988: 92, formula 5.5), but allowing for the effect of ties on the variance by replacing the denominator (as suggested by Sprent 1993: 53 and Mehta and Patel 1991: 7-10) by $\sqrt{\sum(R_i / 4)}$, where R_i = the rank of the difference between paired observations. Nondiscrepant pairs are ignored. If there are fewer than 20 pairs, significance is appraised by using critical levels for one-tailed P = .05, .025, .01, .005, .0025, and .0005 (derived from Siegel and Castellan 1988: Table H; and Zar 1998: Table B.12).

The *permutation test* is explained by Siegel and Castellan (1988: 95-100).

The *McNemar test of overall bias or directional change* compares *a* and *b*, where *a* is the total number of pairs on one side of the main diagonal of the cross-tabulation [i.e. the line connecting cells with equal values for both tests), and b is the total number on the other side of the diagonal. Chi-square (with one degree of freedom) is then calculated as $(a - b)^2 / (a + b)$ or (with a continuity correction) as $(|a - b| - 1)^2 / (a + b)$. One-tailed and two-tailed P values are reported.

The *extended McNemar test* is described by Bowker (1948), Everitt (1977: 114-115) and Zar (1998: formula 9.22). There are $k(k-1)/2$ degrees of freedom (where *k* = number of categories). Corresponding cells that both have zero values are omitted from the calculation of this chi-square, and if there are such comparisons an alternative P is computed, after reducing the degrees of freedom by the number of zero-cell comparisons (Hoenig *et al.* 1995, Evans and Hoenig 1998).
The contributions that specific pairs of categories make to a significant chi-square (P < 0.05) are computed by formula 6 of Maxwell (1970):

$$\text{chi-square} = (n_{ij} - n_{ji})^2 / (n_{ij} + n_{ji})$$

where n_{ij} = the number in the cell in column *i* of row *j*
n_{ji} = the number in the cell in column *j* of row *i*

In McNemar tests for single categories, the degrees of freedom are defined as $k-1$ (where *k* = number of categories) for testing *a posteriori* hypotheses (Fleiss *et al.* 2003: 382).

The *Stuart -Maxwell chi-square test* (Stuart 1955, Maxwell 1970) is performed if there are up to 20 categories. It is computed by a matrix operation (Fleiss *et al.* 2003: 381-383; Everitt 1977: 115-116. The test is not done if any cell is the only non-zero cell both in its column and in its row, , or in 3x3 tables that have more than 3 zero cells unless there are only 3 categories, in which case the category with perfect agreement is omitted from the calculation of chi-square and, as suggested by Uebersax (2006), P is based both on 1 degree of freedom and (for

a more conservative test) on 2 degrees of freedom. To avoid computational problems in extreme situations, some zero divisors are replaced by 0.000001.

The *Bhapkar* test (BhapkaR 1966) is performed if there are up to 20 categories, and is also computed by a matrix operation.

Heterogeneity test

The tests for goodness of fit with a symmetry model, on which the heterogeneity test is based, are described by Agresti (1984: 202).

Odds ratios and related tests

The *generalized odds ratio*, which is Agresti's α′ (Agresti 1980), is computed by the formula provided by Lui (2004: 126), and its 95% confidence interval by the logarithmic-transformation method of formula 6.14 (Lui 2004: 127). For stratified data, the assumed common value of the GOR is the exponent of a weighted average of the logs of the GOR values in the strata, and its 95% confidence interval is computed from the estimated variance of this weighted average (Agresti 1980: 63).

If there is a zero observed frequency of pairs in any cell, 0.5 is added in each cell.

The test for the consistency of odds ratios between pairs of categories, the maximum-likelihood estimation of mutually consistent odds ratios, and the estimation of confidence intervals are described by Pike, Casagrande and Smith (1975).

Kappa and related results

The basic formulae are provided by Fleiss *et al*. (2003: chapter 18). Kappa for single categories and for the total distribution (*overall kappa*) are calculated by formulae 18.10 to 18.12, and *weighted kappa* by formulae 18.27 to 18.29, using linear or quadratic weights. Linear weights are calculated by a formula (18.31) suggested by Cicchetti and Allison (1971), namely (for each cell),

$$1 - |i - j| / (k - 1)$$

where i and j are the scores of the row and column categories

k is the number of categories.

The formula for quadratic weights (18.30) is

$$1 - |i - j|^2 / (k - 1)^2$$

For tests of the null hypothesis that kappa is zero (formulae 18.14 and 18.35), the standard error (for an underlying zero value of kappa) is calculated by formula 18.13. For tests of the hypothesis that kappa has an underlying value other than zero, and for confidence intervals, the standard error appropriate for non-zero values is calculated by formulae 18.15 to 18.18. Confidence intervals are estimated from the standard error (if the upper confidence limit exceeds 1, it is reduced to 1).

The *maximum attainable value* of *kappa* is computed by calculating *kappa* when taking the marginal totals as fixed but modifying the body of the table so as to represent the maximum possible agreement, by using, for each cell indicating agreement, the smaller of the two relevant marginal frequencies.

Bias is appraised by the extended McNemar (symmetry) test (see above), and BAK (*bias-adjusted kappa*) and PABAK (*prevalence-adjusted bias-adjusted kappa*) by the methods described by Byrt *et al*. (1993).

In the combined analysis of several samples or strata, the estimate of the supposed *common or overall value* of *kappa* is calculated in two ways: by computing a weighted mean, using the inverse of the variance of each *kappa* as its weight (Fleiss *et al*. 2003: formula 18.21); and by computing a weighted mean, using the size of the stratum as the weight. The confidence intervals of the common kappa are estimated by formula 18.23 of Fleiss *et al*. (2003)

The *heterogeneity test* is based on formula 18.22 of Fleiss *et al*. (2003). The *measures of heterogeneity* (Higgins and Thompson 2002) are *H* and *I-squared*. *H* is computed by Higgins and Thompson's formula 6, and increased to 1 (indicating absence of heterogeneity) if it less than 1. A test-based interval is computed by Method III. *I-squared* and its 95% interval are computed from *H*, using formula 10.

The significance of the *percentage agreement* is tested by a binomial test comparing the total number of complete agreements with the number expected by chance (Sheskin 2007: 729-730).

Some computations are omitted if division by zero or other problems are encountered. In some instances, zero values are changed to 0.00001 to permit computation.

Gwet's AC2

Gwet's AC2 is calculated by formula 4.17 of Gwet (2010: 77).

Distinguishability of categories

This measure is computed by the method described by Darroch and McLeod (1986), using maximum-likelihood estimates (without assuming quasi-symmetry) in formula 21b (Shoukri 2000).

Rank correlation coefficients and other measures of ordinal association

The computation of *tau*, *gamma*, and Somers' *D* is based on *S*, the difference between the numbers of concordant and discordant pairs, as explained by Kendall (1970: 45-46) and Agresti (1984: 157-159).

The formula for *tau* makes allowance for tied observations (Siegel and Castellan 1988: 249, formula 9.10). If the number of pairs N > 30, the significance of *S* is tested by a large-sample method whose use Agresti (1984: 180) suggests if the numbers of concordant and discordant pairs both exceed 100. If this condition is not met the program reports P as approximate. The formula is

$$Z = (S - CC) / \sqrt{V}$$

where V = variance of S, making allowance for tied ranks (Kendall 1970: formula 4.3)

As recommended by Kendall (1970:54-58), $CC = 1$ unless one variable has only two values and the other has tied ranks, in which case

$$CC = [(2N - T_F - T_L) / \text{Intervals}] / 2$$

where Intervals = the number of different ranks for the non-dichotomous variable, minus one

T_F and T_L = ties involving the first and last ranks (respectively) of the non-dichotomous variable

Gamma is calculated by a formula provided by Siegel and Castellan (1988: 292, formula 9.32). If N > 30, the significance test for *S* (see above) is used as a test for *gamma*.

Somers' *Dxy* and *Dyx* are calculated by Siegel and Castellan's formulas 9.41 and 9.42 (1988: 304-305). Significance is tested by a Z test (Siegel and Castellan 1988: 309, formula 9.47), based on the variance computed by Siegel and Castellan's formula 9.45.

Spearman's *rho* is computed by a formula that takes account of tied ranks (Siegel and Castellan 1988: 241, formula 9.7). It is not calculated if numbers are too large for the program to handle. A large-sample approximation is displayed as the S.E. of *rho*, namely $\sqrt{[1 / (N - 1)]}$ (Hollander and Wolfe 1999, formula 8.72). The t-test for the significance of *rho* (Siegel and Castellan 1988: 243, footnote), used if N > 30, is based on the null variance. An approximate 95% confidence interval (Zar 1996: 398) is estimated if N is 10 or more and *rho* is 0.9 or less, based on the Fisher *z* transformation

$$z = 0.5\ln[(1 + rho) / (1 - rho)]$$

The confidence limits for *rho* {Fieller, Hartley and Pearson (1957, 1961) are

$$\exp[2(z \pm 1.96SEz) - 1] / \exp[2(z - 1.96SEz) + 1]$$

where $SEz = \sqrt{[1.06 / (N - 3)]}$.

D1. PAIRED NUMERICAL OBSERVATIONS (NORMAL DISTRIBUTION)

This module is appropriate for the analysis of paired numerical observations (in different subjects or the same subject) where a normal distribution is assumed. It appraises differences and agreement between the two sets of observations It can be used to analyse matched-control trials and case-control studies, before-after studies, reliability studies, comparisons of measurement methods, and other comparisons of paired subjects or observations. An option is offered for deriving confidence intervals for the difference between means from the P-value, for use in meta-analyses of incompletely reported studies.

The observations entered may be measurements in paired subjects, e.g. matched cases and controls, or repeated measurements in the same subjects. Each pair of matched observations (labelled "A" and "B") can be entered in a separate line, or pairs with the same values can be entered together, with their frequency; up to 500 lines may be entered. Replicated measurements can be entered in any order, unless “A” and “B” represent defined instruments, observers, times, conditions, etc. An option is offered for the entry of supplementary unpaired observations.

If the data are stratified, enter each stratum in turn. Click on “All strata” whenever combined results are required. In a study of several *clusters*, with paired observations in each cluster, enter each cluster as a separate stratum, and then click on "All strata" for a combined analysis.

In a clinical trial or cohort study that uses paired baseline and follow-up measurements to compare the changes in two groups, enter each group as a separate stratum and then click on “All strata” for a comparison using **analysis of covariance** and for estimates of the **number needed to treat**.

The program provides a **comparison of the paired observations (**including tests for differences, namely the Bradley-Blackwood test, Student's paired *t*-test, and Pitman's test), **measures of agreement** (correlation coefficient and population correlation coefficient, six intraclass correlation coefficients, Lin’s concordance correlation coefficient (with Lin’s accuracy coefficient), repeatability coefficients, the standard error of measurement, the within-subject coefficient of variation, the confidence interval for the “true value” corresponding to an observed measurement, Spearman-Brown coefficients of reliability, St Laurent's correlation coefficient, 95% limits of agreement , and the association between the absolute difference and the mean value), a **measure of disagreement, partial *omega*-squared**, **predictors and odds of replication**, and **ANOVA tables**. Measures of the similarity or dissimilarity of the distributions (**PSR** and **ABC**) are provided. Optionally, **equivalence tests** can be performed.

If *stratified data* are entered, the paired one-tailed *t* tests in the separate strata are combined, and the *heterogeneity* of the P-values in the strata is tested.

Comparison of the paired observations

The program displays means, standard deviations and standard errors for the two sets of observations, and the mean difference between the observations, with its standard deviation, standard error and 90%, 95% and 99% confidence intervals. It also provides linear regression coefficients, with their standard errors.

The tests for differences are the Bradley-Blackwood test, which simultaneously tests the means and variances (Bradley and Blackwood 1989; Bartko 1994), Student's paired *t*-test, which compares the means, and Pitman's test (Pitman 1939) for the equality of variances. Two-tailed P-values are displayed.

Since the paired *t* test is based on the assumption that the differences are normally distributed (Zar 1998: 1634, Armitage et al . 2002: 103) four tests for normality are performed - the Lilliefors test, the D'Agostino-Pearson test, the Shapiro-Wilk W test (Shapiro and Wilk 1965, 1968), and the Shapiro-Francia W' test (Shapiro and Francia 1972)
.
The Lilliefors test (Lilliefors 1967) examines the deviation of the cumulative frequency from the standard normal cumulative distribution; the result is reported as P <0.01 or P < 0.05 or P<0.10 or P<0.15, >.0.1 or P<0.2, >0.15; or P>0.2. The D'Agostino-Pearson test (D'Agostino and Pearson 1973, which is based on tests for skewness and kurtosis, is not performed if fewer than 50 pairs are entered. It. is a method of choice if there are 50 or more observations (D'Agostino *et al.*1990). The Shapiro-Wilk W test is based on the correlation between the ordered values and some constants that would be closely correlated in a sample from a normal population. It is "arguably the best omnibus test" (Royston 1993), although it is affected by tied data. It is performed if there are between 7 and 50 observations. The Shapiro-Francia W' is based on the correlation between the ordered observations and the expected standard normal order statistics. It has about the same overall power as the Shapiro-Wilks W test. It is affected by tied data.

If *stratified data* are entered, the paired one-tailed *t* tests in the separate strata are combined by Stouffer's method (Stouffer et al. 1949, p. 45; DeMets 1987) to produce overall one-tailed tests that control for the stratifying variables. Three different sets of weights are used for this purpose – weighting the test results equally, by the sample sizes in the strata, and by the square roots of the sample sizes. In addition, the *heterogeneity* of the P-values in the strata is tested.

Measures of agreement

The measures of agreement have special relevance to studies of reliability, comparisons of measurement methods, and the clinical application of measurements.

A simple correlation coefficient between the variables, intraclass correlation coefficients, and Lin's concordance correlation coefficient are computed in all instances. Four correlation coefficients between the variables (A and B) in the 2x2 table are computed: the *phi* coefficient, which is the usual (Pearson) coefficient, applied to binary variables, and is appropriate if both variables are natural dichotomies based on qualitative characteristics (e.g cases and controls, or exposed and nonexposed); the *tetrachoric correlation coefficient* (see below), which is appropriate if both variables are quantitative ones that have been artificially dichotomized; and two *point-biserial correlation coefficients*, appropriate if one variable is

naturally dichotomous and the other is a dichotomized quantitative variable (and depending on which variable is naturally dichotomous).

If the paired observations are positively correlated the program also provides measures that may be useful if A and B are replicate measurements, or if they denote two different methods of measurement. The measures for use in studies of replicate measurements are repeatability coefficients, the standard error of measurement, and the confidence interval for the "true value" corresponding to an observed measurement. The measures that are appropriate in comparisons of measurement methods are St Laurent's correlation coefficient (for use if one of the methods is regarded as a "gold standard"), and 95% limits of agreement. The program displays the correlation coefficient between the absolute difference and the mean of A and B, and the linear regression of the difference on the mean.

The simple *correlation coefficient* is seldom helpful in comparisons of methods of measurement (Bland and Altman 1995a; Altman 1991: 401-402), since at best it points to an association between the measurements, and does not tell how closely they agree; moreover, its value tends to be high if the subjects are very different, and low if they are similar. The program also reports the *population correlation coefficient*, the *coefficient of determination* (the square of the simple correlation coefficient), and the *adjusted coefficient of determination* (the square of the population correlation coefficient).

The *"true" correlation coefficient.* Since correlation coefficients are attenuated if the measurements do not have complete reliability (Trafimow 2015), an option is offered for the computation of a "true" correlation coefficient that compensates for imperfect reliability. In effect, this is an estimate of what the coefficient would be if the measurements were completely reliable. As Trafimow states, this "corrected" coefficient will be further from zero than the obtained one, "suggesting that it might actually be more important than it otherwise would seem to be". He stresses the importance of a strong appreciation of the effects of the reliability of the measures on correlation sizes. The reliability measures required for the computation are the correlation coefficients between replicate measurements of each of the two variables; the computation is possible only if replicate measurements are available. *The program must be run three times* - the first two times to compute correlation coefficients for replicate measurements of variable A and replicate measurements of variable B, and the third time to measure the correlation between variables A and B and "de-attenuate" it. The "true" coefficient is squared to provide a "true" coefficient of determination.

Intraclass correlation coefficients, which are appropriate for interval-scale data with an assumed normal distribution, are measures of agreement that express the correlation (in terms of absolute agreement) between measurements within individuals or sets of matched individuals. Six intraclass correlation coefficient (ICC) values are computed (Shrout and Fleiss 1979), with their 95% confidence intervals.

Each ICC is appropriate in a different situation. (a) The values with the rubric "two-way model with fixed raters" are appropriate in studies where the matched observations in each set represent various "unique" raters, and no inferences are made about other raters; "raters" denote the various observers, treatments, methods or conditions of observation, matched individuals, or (in a reliability study of a questionnaire or other scale) questions or other scale items, that were studied. Two such ICCs are provided. The first, which Shrout and Fleiss refer to as model 3.1, uses a single measurement as the unit of analysis, and the second (model 3,k) uses an average measurement. (b) The two ICC values reported as "two-way

model with random raters" are appropriate if the raters were randomly selected from a larger population of raters and it is proposed to generalize the findings to this larger population. If analysis is based on a single measurement, this is model 2,1; if it based on an average measurement, it is model 2,k. (c) The third pair of ICC values, entitled "one-way random model", is appropriate in methodological or other studies where the measurements are replications by the same observer or using the same instrument, and the order in which they are entered does not matter (this does not apply to the other ICC values).. They apply to the use of a single measurement (model 1,1) – e.g. in studies to determine the reliability of a single measurement – or to an average measurement (model 1,k) – e.g. in studies to determine the reliability of an average measurement.

The maximum value of an ICC is 1; the lower limit is an indeterminate negative value. As a rule of thumb, it has been suggested that ICC values above 0.75 should be regarded as evidence of excellent, and values above 0.4 as evidence of good, reliability (Shoukri and Pause 1999: 27).

In the appraisal of replicated measurements a low ICC may express variability of the characteristic measured, as well as low reliability of measurement; this is especially important if measurements were conducted at different times. The usefulness of the ICC in comparisons of two methods of measurement (Bartko 1994; Lee 1992) is constrained by these and other limitations (Muller and Buttner 1994; Bland and Altman 1995a).

The *concordance correlation coefficient* is computed with its 95% confidence interval. Suggested by Lin (1989) as an improved measure of the reproducibility of measurements, its use is appropriate in comparisons where the two observers (or measurement methods) are selected "at random" to represent all observers (or measurement methods) to whom the assessed consistency relates; whereas if they are "fixed"–- e.g. in a comparison of two kinds of measuring instrument – it is more appropriate to use the intraclass correlation coefficient (Mueller and Buettner 1994). It has been tentatively suggested that a Lin coefficient of >0.99 indicates almost perfect agreement, 0.95-0.99 substantial agreement, 0.90-0.95 moderate agreement, and <0.90 poor agreement (NIWA 2009). The Fisher *z* transformation of the coefficient is displayed, with its standard error, for use if the findings are to be compared with those in a different set of paired observations; (for this purpose, the standard error of the difference between two *z* transformations is the square root of the sum of their variances). The value of any correlation coefficient, including Lin's concordance coefficient, is affected by the *range of values* included in the analysis (Lin and Chinchilli 1997) - the wider the range, the stronger the correlation - and this should be taken into account when coefficients are appraised or coefficients based on different samples are compared. The program therefore reports this range (the range of the means of paired values).

Lin's *accuracy coefficient Xa* (Lin *et al* 2012), also referred to as the *bias correction factor Cb* (Lin 1989) is also reported. Lin's CCC has two components - the correlation coefficient, which is a measure of precision that evaluates deviations from the best-fit line, and the accuracy coefficient, which measures how far the best-fit line deviates from a 45-degree line through the origin. The coefficient varies from 0 to 1; the further it is from 1, the greater the deviation. Confidence intervals are reported.

The *coefficients of repeatability* express the expectation (with 95% confidence) of the maximum size of the absolute difference between paired observations. Two coefficients are provided, with their approximate confidence intervals. The first (Bland and Altman 1986;

Chinn 1990) is valid if there is no bias (no systematic difference between the observations), i.e. if the mean difference between observations is zero; this may not be so if the measurement process alters the quantity or if knowledge of the first measurement affects the second. The second coefficient controls for any effect of bias; it is based on the residual within-subjects sum-of-squares, after removal of the between-ratings component.

The *standard error of measurement* (Fleiss 1986: 11) –.also called the "technical error" (Kahn and Sempos 1989: 239-242) or "the SE of an obtained score" (Guilford and Fruchter 1986: 413) – is an index of reliability that expresses variation between observers and other causes of differences between repeated observations. To aid in its interpretation, its ratios to the standard deviation among persons and to the mean value are displayed.

The program computes an approximate 95% *confidence interval for the "true value"* corresponding to an observed measurement or the mean of two or three measurements. These should be used with caution, since they assume that the width of the confidence interval is independent of the magnitude of the value (Guilford and Fruchter 1986: 413).

The *within-subject coefficient of variation* is an indication of the extent to which the measurement error varies according to the magnitude of the measurement (Bland and Altman 1996b). Using this coefficient, the program provides formulae for the approximate 95% confidence interval for the "true value" corresponding to an observed measurement.

St Laurent's gold-standard correlation coefficient is a measure of criterion validity – it is a measure of the agreement between a measurement and a "gold standard" (St Laurent 1998). Two values are displayed, with their approximate 95% confidence intervals, taking A or B in turn as the "gold standard". The procedure assumes that the "gold- standard" measurements and the differences between the two sets of measurements are normally distributed.

The *95% limits of agreemen*t (Bland and Altman 1995a, 1995b, 1999; Altman 1991: 397-400) answer the question, "given a measurement by one method, how far might this be from a measurement by the other method?" These demarcate the bounds of the range that, with a 95% probability, includes the difference between single measurements of the same subject by the two methods. The 95% confidence intervals of the 95% limits of agreement are estimated (the limits of agreement may be very imprecise if the sample is small). The confidence intervals are computed by two methods - those of Bland and Altman (1986, 1999) and those of Donner and Zou (2010). Simulation results suggest that the latter method is preferable Donner and Zou 2010).

Use of the 95% limits of agreement assumes that the differences are reasonably constant throughout the range of measurement. To check this assumption, the program displays the *coefficient of correlation between the absolute difference and the mean* of the two values, and the *regression of the difference on the mean*. The correlation and regression coefficients may be expected to be zero if the mean difference and the scatter of differences do not change with increasing values. If the difference and the mean are correlated, it may be appropriate to repeat the computation after log-transformation of the measurements, since the difference between log-transformed values may not change with increasing values. (To do this, click on "Repeat", then on "Lognormal distribution assumed", and then on "Run".)

Considerable inconsistencies may occur between the limits of agreement and the ICC in the interpretation of agreement, and Costa-Santos *et al.* (2011) suggest that these methods should be used in tandem.

Even when one of the methods of measurement is a new one and the other is an accepted standard, it is preferable to examine the relationship between the difference and the mean value rather than the relationship between the difference and the standard measurement, which (as shown by Bland and Altman 1995b) is likely to be misleading.

Spearman-Brown coefficients of reliability provide estimates of the effect of using the means of replicated observations. They predict what the reliability would be if two, three, four, or five replications were averaged.

Clustered data

In order to effectively remove the correlation associated with data clustering (which may appreciably affect the test results) the program uses a Wilcoxon signed-ranks test, applied to the cluster means. The limitations of this simple method (Galbraith et al. 2010) are that the same weight is given to large and small clusters, and that the non-use of individual observations may reduce power; computer simulations confirm this slight loss of power compared with other, more elaborate, tests that take clustering into account. The procedure may not be appropriate if there are very few clusters.

Test for correlation when data are missing

This optional procedure (Parzen *et al.* 2010) tests the null hypothesis that there is no correlation between two numerical variables, while adjusting for missing data. It uses whatever unpaired values (i.e., values with missing pairmates) have been entered (missing values being indicated by an “x”), as well as the paired values. The test makes no assumptions about the distribution of the values. It is said to be appropriate if (a) the probability that a value is missing ("missingness") is completely random, or (b) if "missingness" depends on the observed data but not on the missing values. In the latter instance the test is stated to be unbiased, unlike tests based solely on the complete pairs, which "can potentially yield misleading inferences". The test is said to have high power to detect a linear correlation or a nonlinear monotonic trend.

The test statistic (*Qa*) is displayed, with the corresponding P value. For comparison, the result of a parallel test based solely on complete pairs - the correlation statistic proposed by Mantel (1963) (which can be correct only if "missingness" is completely random) - is displayed. A difference between the test statistics suggests that "missingness" is not completely random. The program also displays the mean values of the variables in the complete and incomplete pairs, to permit an appraisal of possible bias and a decision on whether to use this procedure incorporating the incomplete pairs.

Measure of disagreement

The *measure of disagreement* between two sets of matched numerical observations proposed by Costa-Santos *et al.* (2010) is based on the differences between the paired observations, in relation to the magnitude of the larger value in the pair. It is applicable to ratio-scale

variables (i.e., those where a zero value indicates absence of the attribute) with positive values. The measure ranges from 0 (no disagreement) to 1 (strong disagreement).

Optionally, a 95% confidence interval is estimated for the measure of disagreement, using a bootstrap procedure. This procedure can produce a long delay.

Partial o*mega*-squared

Partial *omega*-squared (ω^2) is an effect-size index that expresses the proportion of the variability of the dependent variable that is associated with variability on the levels of the independent variable, without taking between-subject variability into account(Sheskin 2007: 762).

By Cohen's criteria, an omega-squared of 0.1379 or more indicates a large effect size, 0.0588 or more (but less than 0.1379) indicates a medium effect size, and 0.0099 or more (but less than 0.0578) indicates a small effect size (Sheskin 2007: 763). Cohen (1988) warns that these criteria should be used only when there is no better basis for evaluation.

Equivalence tests

Optionally, the equivalence of the paired measurements is tested, using the procedure described by Yi et al. (2007). This requires entry of the bounds of "equivalence", i.e., the largest difference between measurements that is to be regarded as negligible or 'acceptable'. The tests are based on a comparison of the within-subject variance with this specified difference (and also with this difference multiplied by 0.5, 0.75, 1.5, or 2). A P value under 0.05 implies good agreement (negligible variation, i.e. equivalence) at a 5% significance level.

Comparison of distributions

The *proportion of similar responses* (*PSR*, also called the *OC* or *overlap coefficient*) and the *area between curves* (*ABC*, also called the *dissimilarity index*) are measures of the similarity or dissimilarity (respectively) of two distributions (Giacoletti and Heyse 2011, Mizuno et al. 2005; Rom and Hwang 1996). Differences between frequency curves reflect differences both in location (means) and in scale (variances).

The PSR measures the degree of overlap of two probability distributions. It ranges from 0%, indicating completely disjoint distributions, to 100%, indicating a complete overlap. It has been suggested that a PSR around 70% is a reasonable criterion for equivalence in clinical studies (Rom and Hwang 1996).

The ABC is a measure of the degree of separation between two distributions. Differences between frequency curves reflect differences in scale (variance) as well as in location (mean). The PSR and ABC are related (PSR = 1 - ABC/2).

The estimators are applicable to normal distributions with similar or different means and variances, although computer simulations have shown that the validity of the procedures is highest if the distributions are normal and variances are equal (Mizuno et al. 2005).

These measures have been suggested as aids in comparisons of the results of two treatments. including crossover studies (Rom and Hwang 1996), and in examining the discriminatory capacity of tests (Giacoletti and Heyse 2011.)

The PSR and ABC values are not reported if either exceeds 100%, which indicates that the procedures are inappropriate for this comparison, probably because the two distributions are almost or completely discrepant - i.e. with very little or no overlap..

ANOVA tables

If the paired observations are positively correlated, an analysis of variance (ANOVA) table for the linear regression between the difference between the two ratings and the mean of the two ratings is displayed.

In all instances, a two-way mixed model ANOVA table is displayed, showing between-subjects, within-subjects and between-ratings sums of squares. (The P-values based on *F* tests in the ANOVA tables are one-tailed.)

Analysis of covariance

In studies that use paired baseline and follow-up measurements ("before" and "after" data) to compare the changes in two groups, as in clinical trials and cohort studies, differences between the initial findings in the two groups may complicate interpretation of the findings. Analysis of covariance (which treats the follow-up value as the dependent variable and the baseline value as a covariate - in effect adjusting each subject's follow-up measurement for his or her baseline measurement) is recommended in such studies, although a simple comparison of the changes in the two groups is a reasonable alternative if there is no baseline imbalance and there is a high correlation (say $r > 0.8$) between baseline and follow-up measurements (Vickers and Altman 2001). The use of analysis of covariance avoids the effects of regression to the mean (the tendency of subjects with initially low values to show a rise, and those with initially high values to show a drop).

The procedure assumes that the slopes in the two groups (expressing the regressions of "after" values on "before" values) are parallel. These slopes are therefore compared, and if the slope coefficients differ significantly ($P < 0.05$) analysis of covariance is deemed inappropriate, and is not performed . Heterogeneity with respect to deviations from the regression lines in the two groups is also tested. A single adjusted (pooled) slope coefficient is computed for the analysis of covariance. The program reports the difference between the "after" values in the two groups, for any given "before" value, i.e. controlling for the "before" value. It tests the significance of this difference, and provides its standard error and 90%, 95%, and 99% confidence intervals. In addition, adjusted mean "after" values are computed for both groups, based on the arbitrary assumption that the overall mean of "before" values is the mean "before" value in each group.

Number needed to treat

If the results of a randomized clinical trial based on before-after measurements are entered (with the results in the treatment group in Stratum 1 and those of the control group in Stratum

2), together with the magnitude of the change (in units or as a percentage) that is defined as indicating successful treatment (the MID, or minimal important change), the number needed to treat is computed, with its approximate 95% confidence interval. Depending on the purpose of the trial, this is the number of individuals who are needed in the treatment group in order to avoid a single case or other harmful event, or to produce a beneficial result. If the results of a cohort study are entered, the number is the number needed (in the group entered as Stratum 1) to avoid or produce one minimal important change.

Three methods are used. The first method dichotomizes the results as "successful" or not successful", and calculates the proportions of successes in the treatment and control groups. The second method uses a "better", "worse" or "neither better nor worse" trichotomy in order to estimate the proportions in each group who are more successful. In each instance the difference between the two proportions is reported, and its reciprocal is the number needed to treat. Approximate 95% confidence intervals for the number needed to treat are computed from the confidence intervals of the proportions, unless the latter straddle zero, which would mean a confidence interval straddling infinity for the number needed to treat. The third method is based on the differences in a continuous scale. This is a sensitivity analysis, making a series of calculations of the number needed to treat, using different values for the assumed correlation (in a crossover study) between a subject's results when in the treatment and control groups.

Probability and odds of replication

P_{rep}, which predicts the probability that an effect will be replicated in other studies, was proposed by Killeen (2005) as an alternative to significance tests in evaluating research and aiding practical decision making (Sanabria and Killeen 2007}. The measure predicts the probability that a replication will find a difference in the same direction (i.e., a "same-sign" result, not necessarily significant) as that found in the original study. Its appropriateness and accuracy have been debated (Iverson *et al*. 2009, Lecoutre and Killeen 2010, Killeen 2010). Iverson *et al*. argue that it overestimates the probability of replication. Cumming (2005), who states that "Killeen's P_{rep} is wonderful, but may be difficult to understand", prefers to refer to it as the average probability of replication (*APR*), i.e. the chance of a same-sign result, when averaged over studies in similar populations. As Killeen(2005) points out, a particular value of P_{rep} may be more or less representative of P_{rep} values found for other studies carried out under similar conditions.

The program also reports the *odds* of obtaining a same-sign effect, i.e. *Prep* / (1 - P_{rep}), as suggested by Baguley (2012), and the probability that (on average) replicated studies will find a difference that lies within a confidence interval found in this study(Cumming *et al.* 2004); if the present study's sample size is 30 or more, this probability is 75.5% for a 90% confidence interval, 83.4% for a 95% interval, and 93.1% for a 99% interval.

Confidence intervals derived from *P*- value

An option is offered for deriving confidence intervals for the difference between means from the P-value, for use in meta-analyses of incompletely reported studies., using the procedure described by Hirji and Fagerland (2011). If the *P*-value was based on a paired t test, the number of pairs must be entered.

METHODS

To avoid computational problems in extreme situations, zero divisors are replaced by 0.000001.
At least three pairs of observations must be entered.

Comparison of the paired observations

Formulae for the Bradley-Blackwood test, Student's paired *t*-test, and Pitman's test are provided by Bartko (1994). Linear regression methods are explained in all basic statistics textbooks.

If *stratified data* are entered, the results of the one-tailed *t* tests in the strata are combined by Stouffer's method (Stouffer *et al*. 1949: 5; DeMets 1987), based on weighted averages of the *z* values computed for each test by transforming its one-tailed P-value to the corresponding normal score (Hedges and Olkin 1985: 39). Three different sets of weights are used – weighting the *z* values equally, by the sample sizes in the strata, and by the square roots of the sample sizes. In addition, a heterogeneity test is performed, comparing the P-values in the strata (Wolf 1986: 45). The heterogeneity test uses Wolf's formula:

$$\text{chi-square } (k - 1 \text{ d.f.}) = \sum(z_i - \text{MeanZ})^2$$

where k = number of strata,
z_i = z value in stratum $_i$
MeanZ = mean z value.

Tests for normality

The Lilliefors test for normality (Lilliefors 1967) is explained by Sprent (1993: 77-78); it uses the critical values provided in Table IV. The D'Agostino-Pearson test (D'Agostino 1986, D'Agostino and Pearson 1973) uses formula 6.19 of Zar (1998). The Shapiro-Wilk W test uses the formulae provided by Conover (1999: p. 450) to compute the test statistic, employing the coefficients in Conover's Table A16, and then uses Table A18 to convert the test statistic to an approximately normal random variable, from which an approximate P value is obtained. The Shapiro-Francia W' test uses the method described by Royston (1989), employing the inverse standard normal distribution function formula described by Hamaker (1978). Tied data are treated as sequential.

Measures of agreement

The *correlation coefficient* is computed by formula 19.1 of Zar (1998).

The *phi* coefficient is computed by formula 16.20 in Sheskin (2007).

The formula used for *the tetrachoric correlation coefficient* (Edwards and Edwards 1984) is

$$(OR\pi/4 - 1) / (OR\pi/4 - 1)$$

where OR = ad/bc
a and d = numbers of concordant pairsπ
b and c = numbers of discordant pairs
This simple method, which was used by Stata until recently, provides an approximation that is acceptable in many situations (Digby 198
3, referring to an almost identical formula [with ¾ instead of pi/4]) but that can be very inaccurate (Uebersax (2000). V. Wiggins, of the Stata Corporation, in a reply cited by Gunther and Hofler (2006), says that the approximation works well when the marginals in both directions are above 10%. PAIRSetc does not display the coefficient unless this condition is met, and there are no zero cells. An approximate 95% confidence interval is estimated from a large-sample estimate of the standard error (cited by Digby (1983).

The point-biserial correlation coefficients are computed by formula 3 of Ulrich and Wirtz (2004).

The formula for the *population correlation coefficient* (Abdi and Williams 2010) is

$$\sqrt{\{1 - [(1 - R^2) * (N - 1) / (N - 2)]\}}$$

where R = correlation coefficient
N – number of paired observations

The *coefficient of determination* is R^2, and the *adjusted coefficient of determination* is the square of the population correlation coefficient.

The "true" correlation coefficient (*RAB*) is computed by the formula

$RAB = rAB / \sqrt{(rA * rB)}$ (Trafimow 2015),

Where *rAB* = computed correlation coefficient between variables A and B
rA = correlation coefficient between replicate measurements of A
rB = correlation coefficient between replicate measurements of B

The result is not shown if RAB is less than -1 or more than 1.

The following formulae (Shrout and Fleiss 1979) are used for the six intraclass correlation coefficients. Shrout-Fleiss ICC models 1,1 and 1,k are computed from a one-way random effects model ANOVA, models 2,1 and 2,k from a two-way random effects model ANOVA, and models 3,1 and 3,k from a two-way mixed effects model ANOVA.

$$\text{ICC model 1,1} = (MSB - MSW) / [MSB + (k - 1)MSW]$$
$$\text{ICC model 1,k} = (MSB - MSW) / MSB$$
$$\text{ICC model 2,1} = (MSB - MSE) / [MSB + (k - 1)MSE + k(MSJ - MSE) / N]$$
$$\text{ICC model 2,k} = (MSB - MSE) / [MSB + (MSJ - MSE) / N]$$
$$\text{ICC model 3,1} = (MSB - MSE) / [MSB + (k - 1)MSE]$$
$$\text{ICC model 3,k} = (MSB - MSE) / MSB$$

where MSB = between-subjects mean square
MSE = residual within-subjects mean square
MSW = within-subjects mean square
N = number of subjects
k = number of observations in matched set

Formulae for confidence intervals for the six ICC models are provided by McGraw and Wong (1996a and 1996b) in their Table 7, where they are referred to as ICC(1) and ICC(k) for Case 1, and ICC(A,1) and ICC(A,k) for Cases 2 and 3. The formulae (except those for models 2,1 and 2,k) are set out in a convenient code by Steinley and Wood (2000). Linear interpolation is used to estimate F values that are based on non-integer degrees of freedom (and 1 d.f. is substituted for <1 d.f.) in the computation of confidence intervals for models 2,1 and 2,k; the latter results may differ slightly from those provided by SPSS, which handles non-integer degrees of freedom differently.

Intraclass correlation coefficients are not computed if the correlation coefficient is 1 or –1. ICCs indicative of the reliability of the mean of two ratings are not shown if they fall outside the (–1,+1) range..

The *Spearman-Brown prediction formula* (Fleiss 1986: 14-15: formula 1.30) for reliability (*R)* is

$$R = Nr / [1 + (N - 1)r]$$

where N = number of replicates that are averaged
r = intraclass correlation coefficient (model 1,1)

Fleiss's formula 1.31 is used to estimate the number of replicates required to obtain a reliability of 0.75 or 0.8:

$$N = P(1 - r) / [r(1 - P)]$$

where P = 0.75 or 0.8

The *concordance correlation coefficient* is computed by formula 19.76 of Zar (1998: 409), with *n* substituted for (*n* – 1) in the denominator, and its 95% confidence interval is based on variance formula 2 of Lin (1989), as corrected by Lin (2000). [Version 1.14 and earlier versions of PAIRSetc used Zar's formulae, which yield slightly different results.] The confidence interval is not computed if the correlation coefficient is 1 or –1 or if its estimation requires division by zero.

Lin's *accuracy coefficient* is calculated by formula 2.27 of Lin *et al* (2012*).* Its 90, 95, and 99% confidence limits are based on his equation 2.31, using a logit transformation. If the sample is small, dividing the CCC by the reported coefficient does not exactly coincide with the reported correlation coefficient, as it should (Lin *et al*. 2012), because of the use of *n* or *n*-1 in different formulae.

The formulae for the two *repeatability coefficients* (Bland and Altman 1986; Chinn 1990) are

$$1.96\sqrt{(\sum D^2 / N)} \text{ or}$$
$$1.96\sqrt{(2.SSW / N)}$$

and (controlling for any effect of bias)

$$1.96\sqrt{[2.SSE / (N - 1)]}$$

where D = difference between paired observations
N = number of pairs
SSW = within-subjects sum-of-squares
SSE = residual within-subjects sum-of-squares (excluding the between-ratings component).

Approximate 95% confidence intervals are obtained by substituting confidence limits for SSW and SSE, estimated by the method described by Zar (1998: formula 7.16), in the above formulae.

The formula for the *standard error of measurement* SEm is provided by Kahn and Sempos (1989: 240). SEm is also the square root of the within-subjects mean square shown in the ANOVA table (Fleiss 1986: 11). The formula for the SD among persons is also provided by Kahn and Sempos (1989: 241).

The 95% *confidence intervals for the "true value"* are estimated from the SD of the differences between values, by the method described by Peat *et al.* (1994); the *t*-distribution is used in the computation.

The *within-subject coefficient of variation* (*WSCV*) is computed by the root mean square procedure described by Bland (2006). This yields a result that is similar to but not identical with the logarithmic method described by Bland and Altman (1996b). The approximate 95% confidence interval for the "true value" corresponding to a measurement of X is from X divided by (1.96*GSD*) to X multiplied by (1.96*GSD*),
where *GSD* = geometric standard deviation = $(WSCV + 1)^2$

St Laurent's gold-standard correlation coefficient (St Laurent 1998) is computed by the formula

$$R_g = \sqrt{\{1 / [2B(1 / R_c) - 1] + 1\}}$$

where B = regression coefficient (slope) of the approximate measurement on the gold-standard measurement
R_c = concordance correlation coefficient.

An approximate 95% confidence limit is computed in accordance with St Laurent's Proposition 1.

The *95% limits of agreement* (Chinn 1991, Bland and Altman 1999) are

$$D - 1.96(SD) \text{ and}$$
$$D + 1.96(SD).$$

The 95% confidence limits for the limits of agreement are estimated by the method described by Bland and Altman (1986, 1999) and by the MOVER (Method of Variance Estimates Recovery) method described by Donner and Zou (2010).

The *Spearman-Brown prediction formula* (Wikipedia) is

$$Nr / [1 + (N - 1)r]$$

where N = number of replicates that are averaged
r = intraclass correlation coefficient.

This application of the Spearman-Brown formula was suggested by its use by Solomon (2004).

Comparison of distributions

If the two variance are not equal, PAS is computed by formula 2 of Rom and Hwang (1996)

If they are equal, PSR is computed by formula 2 of Giacoletti and Heyse (2011).

ABC is derived from PSR, using Giacoletti and Heyse's fromula 4.

Clustered data

If clustered data are entered, a Wilcoxon signed-ranks test based on the cluster means is employed. This uses the formula provided by Siegel and Castellan (1988: 92, formula 5.5), but allowing for the effect of ties on the variance by replacing the denominator (as suggested by Sprent 1993: 53 and Mehta and Patel 1991: 7-10) by $\sqrt{\sum[S_i] / 4)}$, where S_i = the square of the rank of the difference between paired observations. Nondiscrepant pairs are ignored. If there are fewer than 20 pairs, significance is appraised by using critical levels for one-tailed P = .05, .025, .01, .005, .0025, and .0005 (derived from Siegel and Castellan 1988: Table H; and Zar 1998: Table B.12). If the sample is larger a normal approximation is used, with allowance made for ties .

Test for correlation when data are missing

The test statistic (*Qa*), which is regarded as a chi-square statistic with one degree of freedom, is calculated by formula 7 of Parzen *et al*. (2010) The variance (the denominator in the formula) is estimated by the bootstrap procedure described by Parzen *et al.* The correlation statistic proposed by Mantel (1963) is calculated from Pearson's correlation coefficient by formula 4 of Parzen *et al.*

The bootstrap procedure uses 2000 random samples with the same number of pairs (complete and incomplete) as the original sample, each sample drawn (with replacement) from the values in the original sample. The variance required for the test is derived from the estimates of covariance (under the null hypothesis) in the 2000 bootstrap samples (formula 9).

The random sampling in this bootstrap procedure uses a pseudo-random number generator described by Wichman and Hill (1985), which derives each number in turn from three seed numbers that it modifies for subsequent use. Initial values for the seed numbers are generated by Delphi's inbuilt random-number procedures, namely RANDOMIZE, using the system clock, and RANDOM, which generates three random numbers from which the required seed numbers are computed. Delphi's RANDOM procedure is augmented by an additional randomizing shuffle, using the algorithm of Bays and Durham, as described by Press *et al.* (1989: 215-217). The formula for each selection is $\mathrm{trunc}(RM) + 1$

where R is a random number in the range $0 < R < 1$
M = the number of candidates.

Measure of disagreement

The formula for this measure (Costa-Santos *et al.* 2010) is

$$\sum L_i / n$$

where $L_i = \log\{[a_i - b_i| / \max(a_i,b_i)] + 1\}.\log(2)$
a_i and b_i are the observations in pair i
n = the number of pairs of observations

If a_i and b_i are equal, L_i is taken as 0.

The measure is not computed if any a_i or b_i is negative. The maximum number of sets of matched observations is 500.

The confidence interval is obtained by a bootstrap procedure, using the basic percentile method (Efron 1981, Efron and Gong 1983) as described by Sheskin (2007: 532-536). The approximate 95% limits are the (2.5)th and (97.5)th percentiles of the distribution of the measures of disagreement (computed by the above method) in 1000 random samples of the same size as the original sample, each drawn (with replacement) from the values in the original sample. Because of resampling, repetitions of the procedure may yield slightly different results.

The random sampling in this bootstrap procedure uses a pseudo-random number generator described by Wichman and Hill (1985), which derives each number in turn from three seed numbers that it modifies for subsequent use. Initial values for the seed numbers are generated by Delphi's inbuilt random-number procedures, namely RANDOMIZE, using the system clock, and RANDOM, which generates three random numbers from which the required seed numbers are computed. Delphi's RANDOM procedure is augmented by an additional randomizing shuffle, using the algorithm of Bays and Durham, as described by Press *et al.* (1989: 215-217). The formula for each selection is

$$\mathrm{trunc}(RM) + 1$$

where R is a random number in the range $0 < R < 1$
M = the number of candidates.

Tests of equivalence

The method is described by Yi *et al*. (2008).

$$\text{Chi-square} = SSW / (D^2 \times 1.96 \times 1.96 \times 2)$$

where SSW = within-subject variance (based on ANOVA)
D = maximum acceptable difference

The P value for the test is 1 minus the P value associated with this chi-square, with $n(k\text{-}1)$ degrees of freedom,

where n = no. of sets of paired measurements
k = no. of repeated measurements (i.e., 2)

Comparison of distributions

If the two variance are not equal, PAS is computed by formula 2 of Rom and Hwang (1996)
If they are equal, PSR is computed by formula 2 of Giacoletti and Heyse (2011).
ABC is derived from PSR, using Giacoletti and Heyse's formula 4.

Partial omega-squared

This is calculated by equation 17.10 of Sheskin (2007).

Probability and odds of replication

Based on Lecoutre, Lecoutre and Poitevineau (2009, formula 13),

$$P_{rep} = 1 - P(|t| / \sqrt{(2)}),$$

where $P(|t| / \sqrt{(2)})$ is the one-tailed probability associated with a t value of $t / \sqrt{(2\}}$, with df degrees of freedom

t *i*s the t value (with df degrees of freedom) obtained by a paired t test

The probability that a replicated study will find a difference that lies within a 100C% confidence interval for the difference (Cumming et al. 2004) is 1 minus double the P value corresponding to a standard normal deviate of $C / \sqrt{(2)}$.

ANOVA tables

The ANOVA tables are explained by Bartko (1994).

Analysis of covariance

The method of calculation is explained in detail by Armitage *et al*. 2002: 332-335) and by Ferguson (1966: 332-339). A t test (Armitage *et al*. 2002: formula 11.20) is used to compare the two slope coefficients, and the pooled slope coefficient is computed by formula 11.23.

Heterogeneity with respect to deviations from the regression lines in the two groups is tested (Snedecor and Cochran 1980: 386) by applying a two-tailed F test to the ratio of the residual mean squares; the residual sums of squares are computed by formula 7.6 of Armitage *et al.* (2002: 292). The standard deviation about regression (the square root of the residual mean square) is reported for each group.

The difference between the "after" values at a given "before" value is computed by formula 11.32 of Armitage *et al*. (2002); its variance is calculated by formula 11.33 and used in a t test (formula 11.35) and for estimating confidence intervals. Adjusted mean "after" values are computed for both groups, based on the assumption that the observed overall "before" mean applies to both groups (formula 11.36).

Analysis of covariance is not done if the slope coefficients in the two groups differ significantly, or if the "before" or "after" values are invariant in either group.

Number needed to treat

The method using a dichotomy is described by Walter (2001: section 3.2). The method using a trichotomy is described by Guyatt et al. (1998); approximate confidence intervals are estimated by a method analogous to that of Walter(2001: formula 3). The method using a continuous scale is described by Walter (2001: section 4.2). If this method yields a number needed to treat exceeding 100, it is reported as ">100".

Confidence intervals derived from P- value

If the number of pairs is entered, the program assumes that the P-value was based on a paired t test; otherwise, a

z test is assumed. Formulae for deriving confidence intervals for the difference between means both for t tests and for z tests. are provided by Hirji and Fagerland (2011).

D2. PAIRED NUMERICAL OBSERVATIONS (LOGNORMAL DISTRIBUTION)

This module is appropriate for the analysis of paired numerical observations (in different subjects or the same subject) that have a lognormal distribution (such as, for example, bronchial responsiveness, recovery times after drug administrations, or the domestic house-dust allergen level). It appraises differences and agreement between the two sets of observations It can be used to analyse matched-control trials and case-control studies, before-after studies, reliability studies, comparisons of measurement methods, and other comparisons of paired subjects or observations.

It may be useful in reliability studies of a normally-distributed variable, if the simple difference between the observations under comparison is found to increase with the level of the measurement. In such instances, the difference between the logs of the observations (i.e., the ratio of the measurements) may be found to be reasonably constant throughout the range of measurement, facilitating estimation of the agreement between measurements.

Computations are based on the logarithms of the values that are entered, which may be measurements in paired subjects or repeated measurements in the same subjects. Each pair of matched observations (labelled "A" and "B") can be entered in a separate line, or pairs with the same values can be entered together, with their frequency; up to 500 lines may be entered.

If the data are stratified, enter each stratum in turn. Click on “All strata” whenever combined results are required.

In a clinical trial or cohort study that uses paired baseline and follow-up measurements to compare the changes in two groups, enter each group as a separate stratum and then click on “All strata” for a comparison using **analysis of covariance**.

The program provides a **comparison of the paired log-transformed observations**, including the Bradley-Blackwood test, Student's paired *t*-test, and Pitman's test, and **measures of agreement between the log-transformed observations** (correlation coefficient, intraclass correlation coefficient, Lin’s concordance correlation coefficient (with Lin’s accuracy coefficient), repeatability coefficients, the standard error of measurement, the confidence interval for the “true value” corresponding to an observed measurement, St Laurent's correlation coefficient, 95% limits of agreement, and the association between the difference and the mean value**).**

If *stratified data* are entered, the paired one-tailed *t* tests in the separate strata are combined, and the *heterogeneity* of the P-values in the strata is tested.

Comparison of the paired log-transformed observations

The program displays the ratio of the paired values, with its confidence intervals.

The tests for differences between the log-transformed observations are the Bradley-Blackwood test, which simultaneously tests the means and variances (Bradley and Blackwood 1989; Bartko 1994), Student's paired t-test, which compares the means, and Pitman's test for the equality of variances. Two-tailed P values are displayed.

Since the paired t test is based on the assumption that the differences are normally distributed (Zar 1998: 1634, Armitage et al . 2002: 103) four tests for the normality o the log-transformed observationsare performed – the Lilliefors test, the D'Agostino-Pearson test, the Shapiro-Wilk W test (Shapiro and Wilk 1965, 1968), and the Shapiro-Francia W' test (Shapiro and Francia 1972)
.
The Lilliefors test (Lilliefors 1967) examines the deviation of the cumulative frequency from the standard normal cumulative distribution; the result is reported as P <0.01 or P < 0.05 or P<0.10 or P<0.15, >.0.1 or P<0.2, >0.15; or P>0.2. The D'Agostino-Pearson test (D'Agostino and Pearson 1973, which is based on tests for skewness and kurtosis, is not performed if fewer than 50 pairs are entered. It. is a method of choice if there are 50 or more observations (D'Agostino et al.1990). The Shapiro-Wilk W test is based on the correlation between the ordered values and some constants that would be closely correlated in a sample from a normal population. It is "arguably the best omnibus test" (Royston 1993), although it is affected by tied data. It is performed if there are between 7 and 50 observations. The Shapiro-Francia W' is based on the correlation between the ordered observations and the expected standard normal order statistics. It has about the same overall power as the Shapiro-Wilks W test. It is affected by tied data.

If *stratified data* are entered, the paired one-tailed *t* tests in the separate strata are combined by Stouffer's method (Stouffer et al. 1949, p. 45; DeMets 1987) to produce overall one-tailed tests that control for the stratifying variables. Three different sets of weights are used for this purpose – weighting the test results equally, by the sample sizes in the strata, and by the square roots of the sample sizes. In addition, the *heterogeneity* of the P-values in the strata is tested.

Measures of agreement between the log-transformed observations

The measures of agreement have special relevance to studies of reliability, comparisons of measurement methods, and the clinical application of measurements.

A simple correlation coefficient, the intraclass correlation coefficient, and Lin's concordance correlation coefficient are computed in all instances.

If the paired observations are positively correlated the program also provides measures that may be useful if A and B are replicate measurements, or if they denote two different methods of measurement. The measures for use in studies of replicate measurements are repeatability coefficients, the standard error of measurement, and the confidence interval for the "true value" corresponding to an observed measurement. The measures that are appropriate in comparisons of measurement methods are St Laurent's correlation coefficient (for use if one of the methods is regarded as a "gold standard"), and 95% limits of agreement. The program displays the correlation coefficient between the ratio and mean of A and B.

The simple *correlation coefficient* is seldom helpful in comparisons of methods of measurement (Bland and Altman 1995a; Altman 1991: 401-402), since at best it points to an

association between the (log-transformed) measurements, and does not tell how closely they agree.

The *intraclass correlation coefficient* (ICC) is a measure of agreement that expresses the correlation between measurements within individuals or pairs of matched individuals. It ranges from -1 to +1, zero indicating no agreement. The ICC is displayed with its significance level and 95% confidence interval. The ICC is affected by the degree of variation among the subjects, and may be misleadingly low if the subjects are very similar, or if differences between paired observations are large relative to the differences between subjects (Bartko 1994). In the appraisal of replicated measurements a low coefficient may express variability of the characteristic measured, as well as low reliability of measurement; this is especially important if measurements were conducted at different times. The usefulness of the ICC in comparisons of two methods of measurement (Bartko 1994; Lee 1992) is constrained by these and other limitations (Muller and Buttner 1994; Bland and Altman 1995a).

The *concordance correlation coefficient* is computed with its 95% confidence interval. Suggested by Lin (1989) as an improved measure of the reproducibility of measurements, its use is appropriate in comparisons where the two observers (or measurement methods) are selected "at random" to represent all observers (or measurement methods) to whom the assessed consistency relates; whereas if they are "fixed"–- e.g. in a comparison of two kinds of measuring instrument – it is more appropriate to use the intraclass correlation coefficient (Mueller and Buettner 1994). It has been tentatively suggested that a Lin coefficient of >0.99 indicates almost perfect agreement, 0.95-0.99 substantial agreement, 0.90-0.95 moderate agreement, and <0.90 poor agreement (NIWA 2009). The Fisher *z* transformation of the coefficient is displayed, with its standard error, for use if the findings are to be compared with those in a different set of paired observations; (for this purpose, the standard error of the difference between two *z* transformations is the square root of the sum of their variances).

Lin's *accuracy coefficient Xa* (Lin *et al.* 2012), also referred to as the *bias factor Cb* (Lin 1989) is also reported. Lin's CCC has two components - the correlation coefficient, which is a measure of precision that evaluates deviations from the best-fit line, and the accuracy coefficient, which measures how far the best-fit line deviates from a 45-degree line through the origin. The coefficient varies from 0 to 1; the further it is from 1, the greater the deviation. Confidence intervals are reported.

The *coefficients of repeatability* express the expectation (with 95% confidence) of the maximum size of the absolute difference between paired log-transformed measurements (i.e., for the ratio of paired measurements). Two coefficients are provided, with their approximate confidence intervals. The first (Bland and Altman 1986; Chinn 1990) is valid if there is no bias (no systematic difference between the observations), i.e. if the mean difference is zero; this may not be so if the measurement process alters the quantity or if knowledge of the first measurement affects the second. The second coefficient controls for any effect of bias; it is based on the residual within-subjects sum-of-squares, after removal of the between-ratings component.

The *standard error of measurement* (Fleiss 1986: 11) –.also called the "technical error" (Kahn and Sempos 1989: 239-242) or "the SE of an obtained score" (Guilford and Fruchter 1986: 413) – is an index of reliability that expresses variation between observers and other

causes of differences between repeated observations. The standard error of measurement is expressed in logarithmic units.

The program computes an approximate 95% *confidence interval for the "true value"* corresponding to an observed measurement. This should be used with caution, since it assumes that the width of the confidence interval is independent of the magnitude of the value (Guilford and Fruchter 1986: 413).

St Laurent's gold-standard correlation coefficient is a measure of criterion validity – it is a measure of the agreement between a measurement and a "gold standard" (St Laurent 1998). Two values are displayed, with their approximate 95% confidence intervals, taking A or B in turn as the "gold standard". The procedure assumes that the "gold- standard" measurements and the differences between the two sets of (log-transformed) measurements are normally distributed.

The *95% limits of agreement* (Bland and Altman 1995a, 1995b; Altman 1991: 397-400) answer the question, "given a measurement by one method, how far might this be from a measurement by the other method?" These demarcate the bounds of the range that, with a 95% probability, includes the difference between log-transformed measurements of the same subject by the two methods (i.e., the ratio of the measurements). The 95% confidence intervals of the limits of agreement are estimated (the limits of agreement may be very imprecise if the sample is small).

Use of the 95% limits of agreement assumes that the differences are reasonably constant throughout the range of measurement. To check this assumption, the program displays the *coefficient of correlation between the difference and the mean* of the two log-transformed values, and the *regression of the difference on the mean*. The correlation and regression coefficients may be expected to be zero if the mean difference and the scatter of differences do not change with increasing values.

Considerable inconsistencies may occur between the limits of agreement and the ICC in the interpretation of agreement, and Costa-Santos *et al.* (2011) suggest that these methods should be used in tandem.

Even when one of the methods of measurement is a new one and the other is an accepted standard, it is preferable to examine the relationship between the difference and the mean value rather than the relationship between the difference and the standard measurement, which (as shown by Bland and Altman 1995b) is likely to be misleading.

Analysis of covariance

In studies that use paired baseline and follow-up measurements ("before" and "after" data) to compare the changes in two groups, as in clinical trials and cohort studies, differences between the initial findings in the two groups may complicate interpretation of the findings. Analysis of covariance (which treats the follow-up value as the dependent variable and the baseline value as a covariate - in effect adjusting each subject's follow-up measurement for his or her baseline measurement) is recommended in such studies, although a simple comparison of the changes in the two groups is a reasonable alternative if there is no baseline imbalance and there is a high correlation (say $r > 0.8$) between baseline and follow-up measurements (Vickers and Altman 2001). The use of analysis of covariance avoids the

effects of regression to the mean (the tendency of subjects with initially low values to show a rise, and those with initially high values to show a drop).

The procedure assumes that the slopes in the two groups (expressing the regressions of "after" values on "before" values) are parallel. These slopes are therefore compared, and if the slope coefficients differ significantly ($P < 0.05$) analysis of covariance is deemed inappropriate, and is not performed . Heterogeneity with respect to deviations from the regression lines in the two groups is also tested. A single adjusted (pooled) slope coefficient is computed for the analysis of covariance. The program reports the difference between the log "after" values in the two groups, for any given "before" value, i.e. controlling for the "before" value. It tests the significance of this difference, and provides its standard error. Since this is a difference between logs of two values, its antilog is the ratio of the two values. The program therefore reports the ratio of the "after" values in the two groups, for any given "before" value, with 90%, 95%, and 99% confidence intervals for the ratio.

METHODS

If zero values are encountered, 1 is added to all values before log-transforming them. Logs to base 10 are used. To avoid computational problems in extreme situations, zeroes are sometimes changed to 0.0000001 or 0.000001.
At least three pairs of observations must be entered.

Comparison of the paired log-transformed observations

Formulae for the Bradley-Blackwood test, Student's paired *t*-test, and Pitman's test are provided by Bartko (1994). Linear regression methods are explained in all basic statistics textbooks.

If *stratified data* are entered, the results of the one-tailed *t* tests in the strata are combined by Stouffer's method (Stouffer *et al*. 1949: 5; DeMets 1987), based on weighted averages of the *z* values computed for each test by transforming its one-tailed P-value to the corresponding normal score (Hedges and Olkin 1985: 39). Three different sets of weights are used – weighting the *z* values equally, by the sample sizes in the strata, and by the square roots of the sample sizes. In addition, a heterogeneity test is performed, comparing the P-values in the strata (Wolf 1986: 45). The heterogeneity test uses Wolf's formula:

$$\text{chi-square } (k - 1 \text{ d.f.}) = \sum (z_i - \text{MeanZ})^2$$

where k = number of strata,
z_i = z value in stratum $_i$
MeanZ = mean z value.

Tests of lognormality

The *Lilliefors test* (Lilliefors 1967) is explained by Sprent (1993: 77-78); it uses the critical values provided in Table IV.

The *D'Agostino-Pearson test* (D'Agostino 1986, D'Agostino and Pearson 1973) uses formula 6.19 of Zar (1998). The test is not performed if there are under 50 values.

The *Shapiro-Wilk W* uses the formulae provided by Conover (1999: p. 450) to compute the test statistic, employing the coefficients in Conover's Table A16, and then uses Table A18 to convert the test statistic to an approximately normal random variable, from which an approximate P value is obtained.

The *Shapiro-Francia W'* uses the method described by Royston (1989), employing the inverse standard normal distribution function formula described by Hamaker (1978). Tied data are treated as sequential.

Measures of agreement between the log-transformed observations

The *correlation coefficient* is computed by formula 19.1 of Zar (1998).The significance test for the coefficient uses Hotelling's modified *z* transformation (Sokal and Rohlf 1981: 583-587) if $N < 30$.

The following coefficients are computed only if the correlation coefficient is positive.

The *intraclass correlation coefficient* that is computed is a mixed model ICC for two fixed ratings, assuming a two-way mixed analysis of variance model (Bartko 1994).

The *concordance correlation coefficient* is computed by formula 19.76 of Zar (1998: 409), and its confidence interval s are based on variance formula 2 of Lin (1989), as corrected by Lin (2000). [Version 1.14 and earlier versions of PAIRSetc used Zar's formulae, which yield slightly different results.] Confidence intervals are not computed if the correlation coefficient is 1 or –1, or if its estimation requires division by zero.

Lin's *accuracy coefficient* is calculated by formula 2.27 of Lin *et al.* (2012). Its 90, 95, and 99% confidence limits are based on his equation 2.31, using a logit transformation. If the sample is small, dividing the CCC by the reported coefficient does not exactly coincide with the reported correlation coefficient, as it should (Lin *et al.* 2012), because of the use of *n* or *n*-1 in different formulae.

The formulae for the two *repeatability coefficients* (Bland and Altman 1986; Chinn 1990) are

$$1.96\sqrt{(\sum D^2 / N)} \text{ or}$$

$$1.96\sqrt{(2.SSW / N)}$$

and (controlling for any effect of bias)

$$1.96\sqrt{[2.SSE / (N - 1)]}$$

where D = difference between paired observations
N = number of pairs
SSW = within-subjects sum-of-squares
SSE = residual within-subjects sum-of-squares (excluding the between-ratings component).

Approximate confidence intervals are obtained by substituting confidence limits for SSW and SSE, estimated by the method described by Zar (1998: formula 7.16), in the above formulae.

The formula for the *standard error of measurement* SEm is provided by Kahn and Sempos (1989: 240). SEm is also the square root of the within-subjects mean square shown in the ANOVA table (Fleiss 1986: 11). The formula for the SD among persons is also provided by Kahn and Sempos (1989: 241).

The 95% *confidence interval for the "true value"* is estimated from the SD of the differences between (log-transformed) values, by the method described by Peat *et al.* (1994); the *t*-distribution is used in the computation.

St Laurent's gold-standard correlation coefficient (St Laurent 1998) is computed by the formula

$$R_g = \sqrt{\{1 / [2B(1 / R_c) - 1] + 1\}}$$

where B = regression coefficient (slope) of the approximate measurement on the gold-standard measurement
R_c = concordance correlation coefficient.

An approximate 95% confidence limit is computed in accordance with St Laurent's Proposition 1.

The *95% limits of agreement* (Chinn 1991) are

$$D - 1.96(SD) \text{ and}$$

$$D + 1.96(SD).$$

The 95% confidence limits for the limits of agreement (Bland and Altman 1986; Altman 1991: 422-423) are estimated by subtracting and adding *t*.SE. In these formulae,
D = mean of the differences (Value 1 minus Value 2)
SD = standard deviation of the differences
SE = $\sqrt{[SD^2 / N) + (t^2.SD^2 / 2N)]}$, which reduces to SD $\sqrt{[2 + t^2]} / \sqrt{(2N)}$
t = the value in the *t* distribution corresponding to a two-tailed P of 0.05 with (N - 1) degrees of freedom
N = number of pairs

Analysis of covariance

The method of calculation is explained in detail by Armitage *et al.* 2002: 332-335) and by Ferguson (1966: 332-339). A *t* test (Armitage *et al.* 2002: formula 11.20) is used to compare the two slope coefficients, and the pooled slope coefficient is computed by formula 11.23.

Heterogeneity with respect to deviations from the regression lines in the two groups is tested (Snedecor and Cochrane 1980: 386) by applying a two-tailed *F* test to the ratio of the residual mean squares; the residual sums of squares are computed by formula 7.6 of Armitage *et al.* (2002: 292). The standard deviation about regression (the square root of the residual mean square) is reported for each group.

The difference between the "after" values at a given "before" value is computed by formula 11.32 of Armitage *et al*. (2002); its variance is calculated by formula 1.33 and used in a t test (formula 11.35) and for estimating confidence intervals. Adjusted mean "after" values are computed for both groups, based on the assumption that the observed overall "before" mean applies to both groups (formula 11.36).

Analysis of covariance is not done if the slope coefficients in the two groups differ significantly, or if the "before" or "after" values are invariant in either group.

D3. PAIRED NUMERICAL OBSERVATIONS (NORMALITY NOT ASSUMED)

This module is appropriate for the analysis of paired numerical observations (in different individuals or the same individual), where a normal or lognormal distribution is not assumed. It appraises differences and agreement between the two sets of observations It can be used to analyse matched-control trials and case-control studies, before-after studies, reliability studies, comparisons of measurement methods, and other comparisons of paired subjects or observations.

The observations entered may be measurements in paired subjects, e.g. matched cases and controls, or replicated measurements in the same subjects. Each pair of matched observations (labelled "A" and "B") can be entered in a separate line, or pairs with the same values can be entered together, with their frequency; up to 500 lines may be entered.

If the data are stratified, enter each stratum in turn. Click on “All strata” whenever combined results are required.

The program provides a **comparison of the paired observations**, including nonparametric tests (permutation test, sign test, Wilcoxon signed-ranks tests, Hollander's test for bivariate symmetry), the median difference between the two values (and Hodges-Ledhmann estimates), the median ratio of the two values in the population, and the proportion with higher values in one set of observations (with their 95% confidence intervals), **measures of agreement** (95% limits of agreement for untransformed and log-transformed data, and an accuracy estimator for screening/diagnostic tests), a **measure of disagreement**, **nonparametric regression analysis** (including **monotonic regression)**, and **rank correlation coefficients and other measures of association**.

If *stratified data* are entered, the paired one-tailed Wilcoxon signed-ranks tests in the separate strata are combined, and the *heterogeneity* of the P-values in the strata is tested. In a study of several *clusters*, with paired observations in each cluster, enter each cluster as a separate stratum, and then click on "All strata" for a combined analysis.

Comparison of the paired observations

The program displays the median and mean values in the two sets of observations, and the *median difference* between the two values in the population, with its approximate 95% confidence intervals. In a matched-control trial or before-after study, the median difference is an estimator of the treatment effect. On the assumption that the data come from distributions that are identical except in the magnitude of the values, these results express the difference between the population means, as well as the difference between the population medians.

The *median ratio* of the two values in the population, with its approximate confidence 95% confidence interval, is estimated in the same way, after log-transforming the observations. The results that are displayed are the exponents of the median difference computed from log-transformed values and its confidence limits.

The *proportion with a higher value in one set* of values is displayed, with its approximate 95% confidence interval.

The *permutation (randomization) test* for paired replicates is performed only if the number of pairs (N) is 25 or less. It may be slow, since it requires processing of 2^N possibilities, i.e. 33,554,432 if N = 25; optionally, the procedure can be aborted. The test is appropriate for interval-scale variables. It assumes that the difference between paired observations is a measure of the difference in the characteristic that is measured; no assumptions are made about normality or other characteristics of the distribution. Exact one-tailed P-values are displayed if $P < 0.05$; the one-tailed value is doubled and shown as a two-tailed value.

The *Wilcoxon signed-ranks test* (Siegel and Castellan 1988: 87-95) tests whether the median discrepancy between paired observations is zero. It is appropriate if the differences between paired observations are an acceptable basis for ranking the differences in the characteristic that is measured. The test is based on the assumption that the distribution of intra-pair differences is symmetric around their median; if this condition is not met some statisticians suggest transformation of the data in order to enhance symmetry (Altman 1991: 204). The program provides a skewness index (0% = complete symmetry, 100% = extreme asymmetry in either direction) and (if there are no zero or negative values) repeats the test, using log-transformed values, which may reduce asymmetry.

If *stratified data* are entered, the one-tailed Wilcoxon signed-ranks tests in the separate strata are combined by Stouffer's method (Stouffer *et al*. 1949, p. 45; DeMets 1987) to produce overall one-tailed tests that control for the stratifying variables. Three different sets of weights are used for this purpose – weighting the test results equally, by the sample sizes in the strata, and by the square roots of the sample sizes. In addition, the *heterogeneity* of the P-values in the strata is tested.

Hollander's test for bivariate symmetry (exchangeability) tests the null hypothesis that paired numerical observations are interchangeable; for example, in a before-after trial using the same subjects, that there is no treatment effect (Hollander and Wolfe 1999: 94-104). It is sensitive to differences between the paired observations and in their dispersion. The program uses a large-sample approximation (Hollander and Wolfe 1999: 96-97) to determine the P value; the results should be used with caution if the sample is small. A low P indicates that the paired observations are not interchangeable.

The *sign test* is based on the direction, not the magnitude, of the differences between the paired observations. [If the numbers of pairs with differences in each direction are known, they can be entered in module A of the DESCRIBE program.]

The *Hodges-Lehmann procedure* (Sprent 1993:89-90) determines the median of the differences between two matched sets, e.g matched cases and controls (with 90%, 95%, and 99% confidence intervals). [This is not necessarily the same as the deference between the medians, or the median of the differences observed in each matched set.]

A large-sample method of analysis is used if there are over 50 matched sets.

The analysis takes account of tied differences (if the large-sample method is used), but not of variation within matched sets.

Measures of agreement

Lin's *concordance correlation coefficient* (Lin 1989 and 2000), which is based on an assumption of normality, is not computed, although computer simulations have shown that the coefficient is robust and can cope with samples from non-normal distributions (Lin 1989). If it is required, module D1 can be used; the results should then be regarded as approximate.

The value of any correlation coefficient is affected by the *range of values* included in the analysis (Lin and Chinchilli 1997) – the wider the range, the stronger the correlation – and this should be taken into account when coefficients are appraised or coefficients based on different samples are compared. The program therefore reports this range (the range of the means of paired values).

The *95% limits of agreement* (Bland and Altman 1995a, 1995b; Altman 1991: 397-400) are appropriate when paired measurements of the same subjects have been entered in order to compare two observers or methods of measurement. The limits of agreement express the range that, with approximately 95% probability, includes the difference between single measurements of the same subject by the two observers or methods, answering the question, "given a measurement by one observer or method, how far might this be from a measurement by the other observer or method?" The limits are unreliable if the sample is small, and are not displayed if under 20 pairs of observations are entered. Their confidence intervals should be regarded as rough approximations. The limits of agreement method assumes that the differences are reasonably constant throughout the range of measurement. As a check on this assumption, the program displays Kendall's rank correlation coefficient (*tau b*; see below) for the difference and the mean of the two values. This may be expected to be zero if the mean difference and the scatter of differences do not change with increasing values. Even when one of the methods of measurement is a new one and the other is an accepted standard, it is preferable to examine the relationship between the difference and the mean value rather than the relationship between the difference and the standard measurement, which (as shown by Bland and Altman 1995b) is likely to be misleading.

Approximate 95% limits of agreement are also computed for the ratio of the two measurements, together with Kendall's rank correlation coefficient for the relationship between the ratio and the geometric mean of the two values. This computation is based on log- transformed data, and is not done if there are zero or negative observations.

A nonparametric *accuracy estimator* is computed, for use in comparisons of screening/diagnostic test results with ordered "gold-standard" ratings. It estimates the probability that the test will correctly rank the members of a random pair of subjects (chance expectation = 50%).

Measure of disagreement

The measure of disagreement between two sets of matched numerical observations proposed by Costa-Santos *et al*. (2010) is based on the differences between the paired observations, in relation to the magnitude of the larger value in the pair. It is applicable to ratio-scale variables (i.e., those where a zero value indicates absence of the attribute) that have positive

values. The measure is applied to the untransformed values. It ranges from 0 (no disagreement) to 1 (strong disagreement).

Optionally, a 95% confidence interval is estimated for the measure of disagreement, using a bootstrap procedure. This procedure can produce a long delay.

Clustered data

In order to effectively remove the correlation associated with data clustering (which may appreciably affect the test results) the program uses a Wilcoxon signed-ranks test, applied to the cluster medians. The limitations of this simple method, like those of a test using the cluster means (Galbraith *et al*. 2010), are that the same weight is given to large and small clusters, and that the non-use of individual observations may reduce power; computer simulations confirm this slight loss of power compared with other, more elaborate, tests that take clustering into account. The procedure may not be appropriate if there are very few clusters.

In addition, a Wilcoxon signed-ranks test is performed in each cluster, the results are combined (using alternative sets of weights), and the heterogeneity of the P-values in the various strata is tested.

Nonparametric regression analysis

The nonparametric regression analysis procedure (which assumes interval-scale measurements) has the advantage of robustness – i.e., discrepant "outlier" observations have a reduced effect. Estimators of the intercept (*alpha*) and slope (*beta*) coefficients in the population are computed, with 90, 95, and 99% confidence intervals for the latter coefficients. Computation may be slow for large samples and can be aborted by the user. Computation is aborted if the samples are too large for the program to handle.

Three alternative ways of estimating *beta* are used, depending on the total number of pairs and the number of discrepant values. Two estimators of *alpha* are computed, and both are shown if they differ. The first estimator is recommended if it cannot be assumed that deviations from the regression line are symmetrical, and the second is recommended if the symmetry assumption is tenable.

Monotonic regression analysis is a form of nonparametric regression analysis. It expresses the linear relationship between the ranks of the A and B variables. Normality is not assumed. The regression equation has the form

Rank of B = *alpha* + *beta*(rank of A).

The closer the relationship is to monotonicity, the closer the absolute value of *beta* is to 1.

Rank correlation coefficients and other measures of association

Kendall's and Spearman's rank correlation coefficients (*tau b* and *rho* [with its standard error], respectively) are computed. These have different numerical values but are similar in their ability to detect associations (Siegel and Castellan 1988: 251).

Goodman and Kruskal's *gamma* and Somers' asymmetric D may be regarded as measures of how effectively the rank of a pair of observations with respect to one observation can be

predicted from their rank with respect to the other observation (see Hildebrand, Laing, and Rosenthal 1977). The *D* statistics are appropriate when one of the observations is clearly the dependent one, e.g. one that comes later in time; *Dxy* is appropriate when A is dependent, and *Dyx* when B is dependent.

Tau, Kruskal's *gamma*, and Somers' *D* depend on a comparison of the ranks of the paired observations. All possible pairs are taken into account in the computation of *tau*, whereas pairs that tie are disregarded in the calculation of *gamma*, and pairs that tie with respect to one (the independent) observation are omitted from the computation of Somers' *D*. *Tau* is the geometric average of *Dxy* and *Dyx*.

A conservative ("outside") 95% confidence interval is estimated for *tau*. For small samples this estimate may be inordinately wide

METHODS

If zero values are encountered, 1 is added to all values before log-transforming them. At least three pairs of observations must be entered.

Comparison of the paired observations

The estimation of the *median difference* and its confidence intervals is described by Campbell and Gardner (2000). For 25 or fewer pairs, the program uses critical values provided in Table 18.6 of Altman *et al*. (2000); for larger samples, it uses the formula provided by Campbell and Gardner (2000: 42).

The median ratio of the two values in the population is estimated in the same way, after log-transforming the observations. The results that are displayed are the exponents of the median difference computed from log-transformed values) and its confidence limits.

A confidence interval for the *proportion that has higher values* in one set is based on the binomial sign test for two dependent samples (Sheskin 2007: 813); it is appropriate if the sample is not small (since the procedure uses a normal approximation).

The permutation test assumes that under the null hypothesis the differences between paired observations are equally likely to be positive or negative. Taking each of these possibilities for each pair, the sum of the differences is computed for each possible combination of findings. The P-value is the proportion of these outcomes that are as extreme as, or more extreme than, the outcome in the actual observations. The procedure is explained by Siegel and Castellan (1988: 95-100).

The *Wilcoxon signed-ranks test* uses the formula provided by Siegel and Castellan (1988: 92, formula 5.5), but allowing for the effect of ties on the variance by replacing the denominator (as suggested by Sprent 1993: 53 and Mehta and Patel 1991: 7-10) by $\sqrt{\sum[Si] / 4)}$, where Si = the square of the rank of the difference between paired observations. Nondiscrepant pairs are ignored. If there are fewer than 20 pairs, significance is appraised by using critical levels for one-tailed P = .05, .025, .01, .005, .0025, and .0005 (derived from Siegel and Castellan 1988: Table H; and Zar 1998: Table B.12). If the sample is larger a normal approximation is used, with allowance made for ties

The formula for the *skewness index* is

$$abs[(H - M) - (M - H)] / (H - L)$$

where M is the median of the observed intra-pair differences
H is their top decile
L is their lowest decile.

The deciles are determined by the methods explained by Zar (1998: 26-27). Pairs with no discrepancies are taken into account in the computation of the median difference and the skewness index, but not in the significance test.

If *stratified data* are entered, the one-tailed Wilcoxon signed-ranks tests in the separate strata are combined by averaging their signed *z* values (Stouffer *et al*. 1949, p. 45; DeMets 1987). Three different sets of weights are used for this purpose – weighting the test results equally, by the sample sizes in the strata, and by the square roots of the sample sizes. In addition, a heterogeneity test is performed, comparing the P-values in the strata, using the formula (Wolf 1986: 45):

$$\text{chi-square } (k - 1 \text{ d.f.}) = \sum(Zi - \text{MeanZ})^2$$

where k = number of strata,
Z_i = z value in stratum i
MeanZ = mean z value.

Hollander's *test for bivariate symmetry* is described by Hollander and Wolfe (1999: 94-104).

The *sign test* is an exact binomial test with a binomial probability of 0.5 (Siegel and Castellan 1988: 80-83; formula 4.2).

Hodges-Lehmann procedure:

The differences between the values of cases and controls are calculated in the n matched pairs. As described by Han (2008), each difference is then compared with each other difference, and for each of these $m = n(n-1)/2$ comparisons of two values, the mean of the pair of differences (Walsh value) is computed.
The m means are then ranked in ascending order, and their median is determined. This is the point estimate of the Hodges-Lehmann median difference between cases and controls.

If $n <= 50$, a value R corresponding to the value of n is obtained from Table A12 of Conove r (1999: p. 545), using the $W_{0.005}$, $W_{0.025}$, and $W_{0.05}$ column for the 90%, 95%, and 99% confidence intervals respectively. The lower confidence limit is the Walsh value whose rank is R in the series, and the upper confidence limit is the Walsh value whose rank is R from the upper end of the series,

If $n > 50$, confidence intervals are estimated by a large-sample approximation (Hollander and Wolfe 1999: 132-133, using the formulae provided by Han (2008), but with a correction for tied ranks (Unistat Statistics Software). The lower confidence limit is the Walsh value whose rank is R in the series, where $R = .za./.b$ rounded up to the nearest integer), and the upper confidence limit is the Walsh value whose rank is R from the upper end of the series,

where z = -1.645, -1.96, or -2.5767 (for 90%, 95%, or 99% limits respectively)
$a = \sqrt{[n(n + 1)(2n + 1) / 24 - Tee / 48]}$
$b = n(n + 1) / 4$
n = number of matched sets
Tee = the sum of $(t_i^3 - t_i)$
t_i = the number of ties in each set of tied ranks

The correction for ties (Tee / 48) in calculating a is omitted if it reduces a to zero or a negative value.

Measures of agreement

The significance test for the correlation coefficient uses Hotelling's modified z transformation (Sokal and Rohlf 1981: 583-587) if N < 30.

The *concordance correlation coefficient* is computed by formula 19.76 of Zar (1998: 409), and its confidence intervals are based on variance formula 2 of Lin (1989), as corrected by Lin (2000). [Version 1.14 and earlier versions of PAIRSetc used Zar's formulae, which yield slightly different results.] Confidence intervals are not computed if the correlation coefficient is 1 or –1, or if its estimation requires division by zero.

Approximate *95% limits of agreement* are computed by a nonparametric procedure described by Bland and Altman 1999. They are determined by excluding the lower and upper 2.5% of the observed distribution of differences. Their confidence intervals are based on confidence intervals for the relevant quantiles (Campbell and Gardner 2000: 39). They should be regarded as rough approximations, both because the method of computing confidence intervals for the quantiles assumes a normal distribution, and because when one confidence limit (lower or upper) falls outside the observed range of differences, it is arbitrarily placed at the

same distance from the point estimate as the other confidence limit. The 95% limits of agreement for the ratio of the two measurements are computed in the same way, using log-transformed data.

The nonparametric *accuracy estimator* is based on formula 1 of Obuchowski *et al.* (2004), but with a modification to allow for tied observations. The denominator in the formula, $n(n-1)$, is reduced by T, where $T = \sum[\, t_i(t_i - 1)]$

t_i = number of subjects with a specific constellation i of findings

The estimator is not computed if both sets of observations show no variation.

Measure of disagreement

The formula for this measure (Costa-Santos et al. 2010) is

$$\sum L_i \,/\, n$$

where $L_i = \log\{[a_i - b_i| \,/\, \max(a_i, b_i)] + 1\}.\log(2)$

a_i and b_i are the observations in pair i

n = the number of pairs of observations

If a_i and b_i are equal, L_i is taken as 0.

The measure is not computed if any a_i or bi is negative, or if there are over 500 sets of matched observations.

The confidence interval is obtained by a bootstrap procedure, using the basic percentile method (Efron 1981, Efron and Gong 1983) as described by Sheskin (2007: 532-536). The approximate 95% limits are the (2.5)th and (97.5)th percentiles of the distribution of the measures of disagreement (computed by the above method) in 1000 random samples of the same size as the original sample, each drawn (with replacement) from the values in the original sample. Because of resampling, repetitions of the procedure may yield slightly different results.

The random sampling in this bootstrap procedure uses a pseudo-random number generator described by Wichman and Hill (1985), which derives each number in turn from three seed numbers that it modifies for subsequent use. Initial values for the seed numbers are generated by Delphi's inbuilt random-number procedures, namely RANDOMIZE, using the system clock, and RANDOM, which generates three random numbers from which the required seed numbers are computed. Delphi's RANDOM procedure is augmented by an additional randomizing shuffle, using the algorithm of Bays and Durham, as described by Press et al. (1989: 215-217). The formula for each selection is

trunc(RM) + 1

where R is a random number in the range $0 < R < 1$

M = the number of candidates.

Clustered data

If clustered data are entered, a Wilcoxon signed-ranks test based on the cluster medians is employed. This uses the formula provided by Siegel and Castellan (1988: 92, formula 5.5), but allowing for the effect of ties on the variance by replacing the denominator (as suggested by Sprent 1993: 53 and Mehta and Patel 1991: 7-10) by $\sqrt{\sum[S_i] \,/\, 4)}$, where S_i = the square of the rank of the difference between paired observations. Nondiscrepant pairs are ignored. If there are fewer than 20 pairs, significance is appraised by using critical levels for one-tailed P = .05, .025, .01, .005, .0025, and .0005 (derived from Siegel and Castellan 1988: Table H; and Zar 1998: Table B.12). If the sample is larger a normal approximation is used, with allowance made for ties

The one-tailed Wilcoxon signed-ranks tests in the separate clusters are combined by Stouffer's method (Stouffer *et al*. 1949, p. 45; DeMets 1987) to produce overall one-tailed tests. Three alternative sets of weights are used for this purpose – weighting the test results equally, by the cluster sizes, and by the square roots of the cluster sizes.

The heterogeneity test comparing the P-values in the strata uses the formula (Wolf 1986: 45):

$$\text{chi-square } (k - 1 \text{ d.f.}) = \sum(z_i - \text{MeanZ})^2$$

where k = number of strata,

z_i = z value in stratum $_i$

MeanZ = mean z value.

Nonparametric regression analysis

The nonparametric regression analysis procedures are described by Daniel (1995: 622-625), Sprent (1993: 195-202) and Sen (1968).

Three alternative ways of estimating *beta* (the slope coefficient) are used, depending on the total number of pairs and the number of discrepant values. If up to 30 pairs of observations are entered, Theil's estimator (Theil 1950) is computed by a method described by Sprent (1993: 195-198). If more than 30 pairs are entered, Sen's method (Sen 1968) is used when possible. The program cannot cope with Sen's method if there are more than 146 pairs of observations with different values of the independent variable, and it then employs the abbreviated Theil method (Sprent 1993: 198-202), which uses a systematic sample of the data. For the Sprent and abbreviated Theil methods, which (unlike Sen's method) assume distinct values of the independent variable, the program treats tied observations as if they were not identical by imputing differences of (alternately) 0.000001 or -0.000001.

The point estimate of *beta* (*b*) is the median value of b_{ij}, where $b_{ij} = (y_j - y_i) / (x_j - x_i)$ for each pair of values of the independent variable *x* (x_i and x_j) and the corresponding values of the dependent variable *y* (y_i and y_j). Using Sprent's method, b_{ij} is calculated for all of the N(N-1)/2 possible pairs of values; zero values of (x-x_i) are changed to 0.000001 or -0.000001 (alternately). In Sen's procedure b_{ij} is calculated only if (x_j-x_i) is not zero. In the abbreviated Theil procedure the N pairs of observations are arranged with the values of the independent variable in a monotonically rising sequence, and each of the first N/2 pairs is then linked with the pair situated N/2 positions further along the array; b_{ij} is computed only for these linked observations; zero values of (x_j-x_i) are changed to 0.000001 or -0.000001.

Confidence intervals for *beta* are obtained from an array of values of b_{ij} in order of increasing magnitude. Sen's method (Sen 1968) uses critical values provided by a large-sample formula based on a variance estimate corrected for ties, and Sprent's method (Sprent 1993: 199-202) uses critical values based on the critical value for Kendall's tau for significance at nominal 10%, 5%, and 1% levels in two-tailed tests, obtained from Siegel and Castellan (1988: 363, TableRII) and Sprent (1993: Table IX). Approximate confidence intervals are estimated in a similar way in the abbreviated Theil procedure, using critical values based on formula 2.3 in Sprent (1993: 34).

Two estimators of the *alpha* coefficient are computed (Dietz 1989; Daniel 1995: 623-624). The first is the median of the (y_i - $b.x_i$]) terms for the N pairs of observations, and the second (Daniel 1995: 623-624) is the median of the averages of the (y_i - $b.x_i$) terms calculated for each of the pairwise combinations of observations. The second estimator of *alpha* is not calculated if the abbreviated Theil procedure is used.

The *monotonic regression* analysis uses formulae 1-3 of Conover (1999: 244).

Rank correlation coefficients and other measures of association

The computation of *tau*, *gamma*, and Somers' *D* is based on *S*, the difference between the numbers of concordant and discordant pairs, as explained by Kendall (1970: 45-46) and Agresti (1984: 157-159).

The formula for *tau* makes allowance for tied observations (Siegel and Castellan 1988: 249, formula 9.10). If the number of pairs N > 30, the significance of *S* is tested by a large-sample method whose use Agresti (1984: 180) suggests if the numbers of concordant and discordant pairs both exceed 100. If this condition is not met the program reports P as approximate. The formula is

$$Z = (S - CC) / \sqrt{V}$$

where V = variance of S, making allowance for tied ranks (Kendall 1970: formula 4.3)

As recommended by Kendall (1970:54-58), $CC = 1$ unless one variable has only two values and the other has tied ranks, in which case

$$CC = [(2N - T_F - T_L) \ / \ \text{Intervals}] \ / 2$$

where Intervals = the number of different ranks for the non-dichotomous variable, minus one

T_F and T_L = ties involving the first and last ranks (respectively) of the non-dichotomous variable

A conservative ("outside") 95% confidence interval is estimated for t*au*, using formula 4.12 of Kendall (1970: 64).

Gamma is calculated by a formula provided by Siegel and Castellan (1988: 292, formula 9.32). If N > 30, the significance test for *S* (see above) is used as a test for *gamma*.

Somers' *Dxy* and *Dyx* are calculated by Siegel and Castellan's formulas 9.41 and 9.42 (1988: 304-305). Significance is tested by a Z test (Siegel and Castellan 1988: 309, formula 9.47), based on the variance computed by Siegel and Castellan's formula 9.45.

Spearman's *rho* is computed by a formula that takes account of tied ranks (Siegel and Castellan 1988: 241, formula 9.7). It is not calculated if numbers are too large for the program to handle. A large-sample approximation is displayed as the S.E. of *rho*, namely $\sqrt{[1 / (N - 1)]}$ (Hollander and Wolfe 1999, formula 8.72). The *t*-test for the significance of *rho* (Siegel and Castellan 1988: 243, footnote), used if N > 30, is based on the null variance. An approximate 95% confidence interval (Zar 1998: 392) is estimated if N is 10 or more and *rho* is 0.9 or less, based on the Fisher *z* transformation

$$z = 0.5\ln[(1 + rho) / (1 - rho)]$$

The confidence limits for *rho* {Fieller, Hartley and Pearson (1957, 1961) are

$$\exp[2(z \pm 1.96SEz) - 1] / \exp[2(z - 1.96SEz) + 1]$$

where $SEz = \sqrt{[1.06 / (N - 3)]}$.

If there are 30 or fewer pairs, the significance of t*au* is appraised by using critical levels for one-tailed P = 0.05, 0.025, 0.01, and 0.005 (Siegel and Castellan 1988: Tables RI and RII), and the significance of *rho* by using critical levels for one-tailed P = 0.05, 0.025, 0.01, 0.005, and 0.001 (Siegel and Castellan 1988: Table Q). If N > 30, a Z test is used for *tau* and *gamma*, and a *t*-test for *rho*. The Z test is appropriate for large samples, and P is reported as "approximate" if criteria suggested by Agresti (1984: 180) are not met.

D4. ANALYSIS OF PAIRED SURVIVAL DATA

This module is appropriate for the analysis of trials and follow-up surveys that study paired survival data, e.g. in paired individuals or in the two eyes of the same subjects.

A survival time ("time to event") is the number of time units (usually days or months) from the start of observation until the occurrence of a specified end-point event (such as death, the onset of a disease or complication, recovery from a disease, or return to work) or (if the event has not occurred) until withdrawal from observation. The main reasons for withdrawal, or *censoring*, are loss of contact, circumstances that dictate removal from the study, and conclusion of the study.

Each pair of survival times (A and B) may be entered separately, or (if specific paired values occur more than once) the paired values can be entered with their frequency.
Censored survival times are entered by appending "+", e.g .by entering "37+". Up to 500 pairs of survival times may be entered

To obtain results that are relevant to specific periods that are of interest, these periods can be entered (e.g., 24 months, to obtain information about 2-year survival).

The program provides a Kaplan-Meier life-table analysis for each group of observations (**cumulative survival proportions** with their 95% confidence intervals, **median and mean survival times**, and the **incidence rate** of the event), **comparisons of survival proportions,** the **hazard ratio** (with 95% confidence intervals), the **trends in the early and later periods of follow-up**, and **tests comparing the survival distributions (**Prentice-Wilcoxon and Gehan tests**)**.

Cumulative survival proportions

For each group of observations, the cumulative survival proportions (expressed as percentages) at each survival time entered are estimated by the Kaplan-Meier procedure. Cumulative survival proportions are also computed for any survival times that have been specified as of special interest, with their approximate 95% confidence intervals; these are large-sample limits, and Rothman and Greenland (1998: 289-90) recommend their use only if at least five events were observed and there are at least five survivors under observation at the time of the calculation; a warning is displayed if these conditions are not met.

The step-by-step survival proportions that are reported provide raw data for the construction of survival curves, consisting of horizontal lines with vertical steps whenever the survival proportion changes.

Median and mean survival times

Where possible, median and mean survival times are reported for each group of observations.

Whether survival times are censored or not, the median survival time is defined as the time at which the cumulative survival probability drops to 50% or below. An approximate standard error and 95% confidence interval are reported; these values may be inaccurate if the sample is small (Machin and Gardner 2000: 97)..

If the survival probability is not precisely 50% at the reported median survival time, an alternative median is also reported, based on linear interpolation between the times straddling the 50% mark.

The program also computes the median survival time expected if the distribution is exponential; if this is very different from the observed median, the assumption of exponentiality can be rejected..

The mean survival time is displayed, with its 95% confidence interval. If there are censored survival times, these values are estimates.

Incidence rate of the event

The average rate of events and its confidence intervals are estimated from the mean survival time and its confidence limits. If any survival times are censored, the rate is an estimate.

Comparisons of survival proportions

For specific survival times that have been specified as being of special interest, the program displays the difference between the survival proportions in the two groups of observations, and the ratio of these proportions, with their approximate 95% confidence intervals. The confidence intervals should be used with caution if the survival times were selected *a posteriori*, after examination of the data (Altman 1991: 376).

Hazard ratio

The hazard ratio, which is similar to a relative risk, expresses the relative survival experience of the two groups. The program also displays the values (in each group of observations) on which the hazard ratio is based – the number of observed events and the "extent of exposure" or "expected events", and their ratio.

Trends in the early and later periods of follow-up

As a simple indication of possible time-related differences between the survival distributions, the program summarizes the change in the cumulative survival proportion in each group of observations, in the early and later segments of the follow-up period (usually using the median survival period for Group A as the cutting-point). The change is expressed as the drop in the survival percentage.

Comparison of the changes may point to trends that are different in the two groups or time periods. Differences in trend in the two periods may be obscured in the overall results.

Number needed to avoid one event

For use in studies in which the events are avoidable, the program reports the number of individuals who are needed in the group with a longer survival time, in order to avoid a single case.

Tests comparing the survival distributions

Two tests are performed: the Prentice-Wilcoxon test (Prentice 1978) and the Gehan test (Gehan 1965). Both tests allow for censored observations. One-tailed and two-tailed P values are shown.

When data are heavily censored, great differences can exist between the results of the two tests (O'Brien and Fleming 1987).

Computer simulations indicate that the Prentice-Wilcoxon test is more powerful in most situations, but the Gehan test may be more powerful if the survival times follow an exponential distribution (Woolson and O'Gorman 1992).

METHODS

Cumulative survival proportions

Cumulative survival proportions are estimated by the Kaplan-Meier technique (Kaplan and Meier 1958; Armitage *et al*. 2002: 575-576; Machin and Gardner 2000: 94-96).

95% confidence intervals for survival proportions at specific selected times are computed from the estimated variance of the logit of the proportion, using Greenwood's formula (Rothman and Greenland 1998: 289-90).

Median and mean survival times

The *median survival time* is defined as the time at which the cumulative survival probability drops to 50% or below. Its approximate standard error and 95% confidence interval are computed by the formulae provided by Machin and Gardner (2000: 97-98), based on the survival times at which the survival probabilities reach or cross the 45% and 55% levels, or if these probabilities are equal, the 40% and 60% levels. The effective sample size required for the calculation is the total sample size minus the number censored before the median survival time (Machin and Gardner (2000: 94). If the sample is small, the results are unreliable.

If the survival probability is not precisely 50% at the reported median survival time, an alternative median is also reported, based on linear interpolation between the times straddling the 50% mark (Selvin 1996: 374).

The median survival time expected if the distribution is exponential is the sum of the survival times (whether censored or not) divided by the number of events (Altman 1991: 385).

The **mean survival time** and its confidence intervals are computed in the usual way if no survival times are censored. Otherwise, a nonparametric estimate of the mean is computed, based on formula 11.29 of Selvin (1996: 371); its standard error is computed by formula 11.31 and used for interval estimation; for this purpose, the longest survival time is treated as uncensored, even if it is censored.

A mean/median survival time is also computed, based on the assumption that the distribution is exponential (Selvin 1996, formula 11.19; Altman 1991: 385). Its standard error is computed by Selvin's formula 11.20.

Incidence rate of the event

Since (in a closed population) an incidence rate is the reciprocal of the average time until occurrence of the event (Rothman 1986: 29; Morrison 1979), the reciprocals of the mean survival time (or the estimate of the mean survival time) and its confidence limits are used as estimates of the average rate of events and its confidence limits.

Comparisons of survival proportions

For comparisons of survival proportions, the estimation of the variances and confidence intervals of the differences and ratios is described by Rothman and Greenland (1998, 291-292). Formulae 16-15 and 16-16 are used, based on. the estimated variances of the logits of the proportions (Rothman and Greenland 1998, pp 289-90). The proportions are treated as independent.

Ratio of median survival times

The computation of a confidence interval for the ratio of the median survival times in the two groups (Simon 1986), on the assumption that the survival times have an exponential distribution, is described by Altman (1991: 384-385). The median survival times used for this purpose are those at which the cumulative survival probability drops to 50% or below.

Hazard ratio

The program computes the Pike hazard ratio estimator (Pike 1972).

Trends in the early and later periods of follow-up

Changes in the survival percentage in each group of observations are reported, in the early and later periods of follow-up. The cutting-point used for this purpose is based on the median survival period for Group A (or, if this median is not reached, on the point at which the cumulative survival proportion drops to 60%). The longest survival time entered determines the end of the later period. Where possible, the interval defined for Group A is applied to Group B also. Linear interpolation is used where necessary.

Number needed to avoid one event

The number of individuals who are needed in the group with a longer survival time in order to avoid a single case is computed from the difference between survival proportions and its estimated variance (Altman and Anderson 1999),

Tests comparing survival distributions

The Prentice-Wilcoxon and Gehan tests are done in accordance with the detailed procedures set out by Woolson and O'Gorman (1992). For the Prentice-Wilcoxon test, the *delta* value for each pair of survival times is multiplied by the frequency of the combination, if it is more than 1.

D5. ASSESSMENT OF REGRESSION TO THE MEAN

This module assesses the effect of regression to the mean (RTM), for use in studies in which subjects selected because of their extreme values (high or low) are measured again later to appraise change Once known, the expected change due to RTM can be compared with the change actually observed, to see to what extent RTM can explain the observed difference.

The procedure assumes a normal or lognormal distribution.

The cut-point used for selecting subjects for inclusion in the sample must be entered, together with the mean value and S.D. in the population from which they were drawn. Optionally (to allow for aging or a secular change), the population mean and S.D. at the time of the second measurement can be added. If the distribution is lognormal, the required entries are the log of the cut-point, and means and S.D.s of log-transformed values.

The program computes the **regression-to-the-mean effect**, i.e. (if the distribution is normal) the expected difference between the mean baseline and mean second value or (if the distribution is lognormal) the expected ratio of the mean second value to the mean baseline value. The computed baseline mean is also displayed, for comparison with the observed baseline mean (a discrepancy suggests a skewed distribution).

Regression-to-the-mean effect

If there is of random variation, the second measurement in a follow-up study of subjects selected because of their extreme values will always tend to be less extreme than the first. In trials, RTM effects may be confused with treatment or placebo effects (Barnett *et al.* 2005, Morton and Torgerson 2005).

Regression towards the mean depends not only on the cutpoint used to determine inclusion in the sample and on the distribution (mean and S.D.) in the population, but also on the correlation (in the population) between repeated measurements of the same individuals. The expected changes due to RTM are therefore displayed in a table that lists alternative values, depending on the correlation coefficient (ranging from 0.025 to 0.975). Choice of an appropriate coefficient requires external data; the correlation usually becomes attenuated as the interval between measurements increases. For cholesterol values, correlation coefficients of 0.7 or higher have been reported for measurements taken a year apart (Yudkin and Stratton 1996).

A method sometimes used to reduce the effect of regression towards the mean is to determine the subject's inclusion in the study sample not by using a single baseline measurement, but by using the mean of two or more baseline measurements. The table therefore displays alternative RTM values, for 1, 2, 3, or 4 baseline measurements. Use of more than 4 measurements brings little benefit (Yudkin and Stratton 1996).

If a second population mean and S.D. are not entered, the computation assumes that there is no change in the population distribution.

METHOD

The program estimates the effect of regression to the mean by the formula provided by Davis (1976: formula 3); also Yudkin and Stratton (1996) and Barnett *et al*. (2005). It uses a modification that allows for a change in the population values between the two measurements, as described by Chinn and Heller (1981). It also uses modifications that are appropriate if the baseline value is the mean of 2, 3, or 4 measurements, as described by Davis (1976) and Yudkin and Stratton (1996).

If a lognormal distribution is assumed, the computation is based on the logs that are entered. The RTM effect that is displayed, namely the ratio of the later mean to the baseline mean, is the antilog of the RTM effect computed from the logs (Bland and Altman 1996a).

Some values may not be calculated, or the whole analysis may be skipped, if the cut-point is excessively extreme.

The computed baseline mean that is reported is based on single baseline measurements (Davis 1976: formula 2).

D6. ADJUSTMENT FOR REGRESSION TO THE MEAN

This module provides **significance tests that adjust for the effect of regression to the mean** (Mee-Chua tests), for use in studies in which subjects selected because of their extreme values (high or low) are measured again later to appraise change.

Either individual data or summary data (sample size, means and standard deviations) may be entered. If the mean in the population from which the sample was drawn is not known, or if the correlation between the two sets of measurements is not known, the program provides a sensitivity analysis, computing P values for a wide variety of scenarios.

If the population mean is not known, the program also reports the **lowest possible adjusted P value** and the population mean to which it applies, and the **range of population means** for which the test would be significant, using the extended Mee-Chua procedures proposed by Ostermann *et al*. (2008).

If individual data are entered, a **paired t test** (not controlling for the regression-to-the-mean effect) is also performed, for comparison with the above test.

Regression-to-the-mean effect

If there is random variation, the second measurement in a follow-up study of subjects selected because of their extreme values will always tend to be less extreme than the first. In uncontrolled or inadequately controlled trials, this regression-to-the-mean (RTM) effect may be confused with a treatment or placebo effect (Barnett *et al*. 2005, Morton and Torgerson 2005). The degree of regression towards the mean depends on the criterion used to determine inclusion in the sample and on the distribution (mean and S.D.) in the population from which the sample was drawn, and finds expression in the correlation between repeated measurements of the same individuals.

A test proposed by Mee and Chua (1991) is used. This test, which is based on a linear regression model, requires information on the mean value in the population and the correlation between the two sets of values. The test may be regarded as a replacement, removing the RTM effect, for a paired *t*-test. It assumes that distributions are normal, that there was no change in the population mean and S.D. between the times of the two measurements, that the correlation is constant over the whole range of values, and that effects are additive.

This test has been formulated more simply by Ostermann *et al.* (2008), who extend it to a situation where the population mean is unknown, by suggesting that it be used systematically over a range of reasonable means, and by providing formulae to determine the *lowest possible adjusted P value* and the population mean that would give rise to this lowest P value, and the *range of population means* for which the test would be significant.

If the population mean is not available, the program reports the lowest possible P value and the population mean associated with this P value, and the range of population means that

would be associated with a significant test result (i.e., with a P value less than 0.025 in a one-sided Mee-Chua test of the null hypothesis versus an alternative in the direction of the observed difference between the means of the two sets of values). The program also applies the test to 11 hypothetical situations, using alternative evenly-spaced population means ranging from half to double the mean of the first set of measurements.

If the correlation coefficient is not available, but the population mean is, the test is performed nine times, using alternative correlation coefficients of 0.9, 0.8, … 0.1.

If neither the population mean nor the correlation coefficient is available the test is performed 99 times (nine postulated correlation coefficients, with eleven evenly-spaced population means). For each value of the correlation coefficient, the lowest possible adjusted P value and the associated population mean are reported, as well as the range of population means for which the test would be significant.

When appraising the findings, postulated population means that are not plausible should be ignored. Also, it should be kept in mind that regression to the mean can contribute to the difference between the means only if the population mean is below the first mean (if the second mean is lower than the first), or above the first mean (if the second mean is higher than the first).

Scrutiny of the results to determine the plausibility of scenarios that show significant results may permit judgments on the possibility of a true change, and facilitate "separating the wheat from the chaff in situations when one has to interpret the results of uncontrolled studies" (Ostermann *et al*., 2008).

METHOD

The formula used for the *Mee-Chua test* is equation 4 of Osterman *et al* . (2008), with the covariance (s_{Y1Y2}) replaced by rs_1s_2, where r is the correlation coefficient and s_1 and s_2 are the standard deviations of the two samples. The *lowest possible adjusted P value* is derived from the t value computed by equation 7 (with $n - 2$ degrees of freedom), and the associated population mean by equation 6.

The *range of population means* for which the test would be significant is estimated by a method proposed and explained by Osterman *et al.,* using the following formulae (Luedtke R, personal communication) which because of their length were not printed in their paper. The formulae provide μ_1 and μ_2, which represent the points (along a spectrum of possible population means) that separate significant one-sided test results (in either direction) from nonsignificant results. Y_1 and Y_2 are the two sets of measurements. The program assumes that the population mean cannot be less than 0.

$$\mu_{1,2} = \frac{-q \pm \sqrt{q^2 - 4pr}}{2p}$$

with

$$p = a^2c^2 - ndt^2$$

$$q = 2a^2bc + 2ndt^2\overline{y}_1$$

$$r = a^2b^2 - t^2de - ndt^2\overline{y}_1{}^2$$

$$a = \sqrt{n(n-2)}$$

$$b = s^2_{Y_1}\overline{y}_2 - s_{Y_1Y_2}\overline{y}_1$$

$$c = s_{Y_1Y_2} - s^2_{Y_1}$$

$$d = s^2_{Y_1}s^2_{Y_2} - s^2_{Y_1Y_2}$$

$$e = (n-1)s^2_{Y_1}$$

$t_{n-2,0.975}$ the 97.5% quantile of the t-distribution with n-2 degrees of freedom

and

n the sample size

$s^2_{Y_1}, s^2_{Y_2}$ the sample variances of Y_1 and Y_2

$\overline{y}_1, \overline{y}_2$ the sample means of Y_1 and Y_2

$s_{Y_1Y_2}$ the sample covariance of Y_1 and Y_2

The formula for the *paired t test* will be found in most statistics textbooks, e.g. Altman (1991), p. 191.

If individual measurements are entered, the maximum permissible number of pairs is 800.

E. COMPARISON OF SUBJECTS WITH TWO OR MORE MATCHED CONTROLS ("YES-NO" VARIABLE)

This module is appropriate for the analysis of case-control studies, clinical trials and cohort studies in which each index subject (each case, experimental subject, or individual exposed to a risk or protective factor) has a fixed number (2-20) of individually matched controls, and the dependent variable is dichotomous ("yes-no"), e.g. "yes" = exposure to a risk factor (in a case-control study), the success of a treatment, or the presence of a disease (in a cohort study). It compares the findings in the index subjects and their matched controls.

The program refers to index subjects as "cases". The number of controls per case must be entered. Then each set of matched observations can be entered in a separate line, or sets with the same findings can be entered together, with their frequency. The required entries for each pattern of findings are 0 ("no") or 1 ("yes") for the "case", and the number of matched controls with "yes".

If the data are stratified, enter each stratum in turn. Click on "All strata" whenever combined results are required.

The program provides **tests for the difference** between the "cases" and their controls (exact Fisher's and mid-P tests, Mantel-Haenszel test, and Walter's test for binary data), the **odds ratio** (maximum-likelihood and Mantel-Haenszel estimates), and ***kappa.***

If *stratified data* are entered, the Walter's tests in the separate strata are combined, the *heterogeneity* of the P-values in the strata is tested, and an overall *kappa* is computed .

Tests for the difference

The program provides exact Fisher's and mid-P tests, the Mantel-Haenszel test, and Walter's test for binary data (with and without a continuity correction).

If *stratified data* are entered, the Walter's tests in the separate strata (continuity-corrected) are combined by averaging their *z* values (Stouffer *et al*. 1949: 45; DeMets 1987) and computing an overall P that controls for the stratifying variables. P-values are computed in three ways, weighting the strata by different methods: weighting them equally, by sample sizes (the number of pairs), and by the square roots of the sample sizes. In addition, a test is done for the *heterogeneity* of the P-values in the strata (Wolf 1986: 45).

Odds ratio

Maximum-likelihood and Mantel-Haenszel estimates of the odds ratio are computed, with exact (Fisher's and mid-P) and approximate confidence intervals. In occasional extreme instances, computational problems prevent the use of exact methods for the calculation of confidence intervals for the odds ratio. Approximate confidence intervals for the maximum-likelihood estimate of the odds ratio are shown only if exact intervals are not computed.

Jewell's low-bias estimator of the odds ratio (Jewell 1984) is also displayed. This serves to draw attention to the tendency for the odds ratio in a sample, especially a small one, to

overestimate the true odds ratio in the population represented. A disadvantage of the estimator is that it is affected by the direction of computation; its value when the number of case: "yes", control: "no" pairs is the numerator of the ratio is not the reciprocal of its value when this number is the denominator.

Kappa

The program computes *kappa*, which expresses the agreement among all the observations in the matched sets, and may serve to express the effectiveness of the matching procedure, since it indicates the extent to which the findings in matched sets are more similar than findings in individuals from different sets.

The probability of chance agreement is taken into account in the calculation of *kappa*. A value of 1 indicates perfect agreement (after allowing for this probability of chance agreement) between ratings; 0 indicates no agreement other than what can be attributed to chance, and a negative value indicates less than chance agreement. Fleiss *et al*. (2003) suggest that a value of 0.75 or more indicates excellent agreement, and 0.40 or less indicates poor agreement. Cicchetti and Sparrow (1981) divide Fleiss's 0.40–0.74 group into 0.60–0.74: good; and 0.40–0.59: fair. Alternative guidelines are: over 0.80, very good agreement; 0.61–0.80, good; 0.41–0.60, moderate; 0.21–0.40, fair; and 0.20 or less, poor agreement (Landis and Koch 1977, Altman 1991).

If stratified data are entered, an *overall kappa* (weighted by sample sizes) is computed.

METHODS

The program can cater for up to 20 controls per case.

Tests for the difference

The computation of *exact probabilities* uses an efficient algorithm for calculating the coefficients of the conditional distribution (Martin and Austin 1991, 1996), using code from David O. Martin's public-domain EXACTBB program.

The *Mantel-Haenszel* chi-square test for matched observations is described by Rothman (1986: 262-263: formulae 13-15 and 13-18).

The formula for *Walter's test* for binary data is formula 2 in Walter (1980); for a continuity-corrected test, 0.5 is subtracted from the absolute value of the numerator.

If *stratified data* are entered, the Walter's tests in the separate strata (continuity-corrected) are combined by averaging their *z* values (Stouffer *et al.* 1949, p. 45; DeMets 1987). Three different sets of weights are used for this purpose – weighting the test results equally, by the sample sizes in the strata, and by the square roots of the sample sizes. In addition, a heterogeneity test is performed, comparing the P-values in the strata, using the formula (Wolf 1986: 45):

$$\text{chi-square } (k - 1 \text{ d.f.}) = \sum (Zi - \text{MeanZ})^2$$

where k = number of strata,
Z_i = z value in stratum i
MeanZ = mean z value.

Odds ratio

Rothman (1986) explains the computation of maximum-likelihood (pp 254-255, 257-258) and Mantel-Haenszel point estimates (pp 256, 258: formulae 13-7 and 13-9) of the odds ratio and their approximate confidence intervals (pp 268-270: formulae 13-37 and 13-38, and pp 273-275). The computation of exact intervals uses an efficient algorithm for calculating the coefficients of the conditional distribution (Martin and Austin 1991, 1996), using code from David O. Martin's public-domain EXACTBB program.

The *low-bias estimator of the odds ratio* is computed by Jewell's formula (Jewell 1984: 431), whether the number of controls per case is fixed or variable. If there is one control per case the estimator is

$$b/(c + 1),$$

where b = number of “case Yes, control No” pairs
c = number of “case No, control Yes” pairs.

Kappa

Kappa is calculated by formulae 18.10 to 18.12 of Fleiss *et al*. (2003). To test the null hypothesis by dividing kappa by its standard error, the standard error (for an underlying zero value of kappa) is calculated by formula 18.13. The hypothesis that agreement is better than chance is tested by formula 18.14 or 18.35.

If *stratified data* are entered, an *overall kappa* (weighted by sample sizes) is computed.

F. COMPARISON OF THREE OR MORE MATCHED SAMPLES ("YES-NO" VARIABLE)

This module compares the findings in 3 to 10 related samples (each observation being matched with an observation in each other sample) where the dependent variable is dichotomous ("yes-no"). The data may be sets of observations in matched individuals, or separate sets of observations in the same individuals. The 3 to10 samples can, but need not, lie in an ordered sequence (e.g. in a trial comparing different doses).

The program may be used, for example, to analyse a clinical trial in which matched subjects receive 3 to10 different treatments, or one in which each subject receives 3 to10 different treatments, or one in which each subject receives the same treatment under 3 to10 different conditions, or an observational study comparing matched subjects who have different degrees of exposure to a risk, or are measured under different defined circumstances, or are appraised by different clinicians or interviewed by different interviewers, or a study in which the same individuals are observed under different specified conditions, or at various specified times, or are asked different specified questions.

If the samples lie in an ordered sequence, they should be numbered accordingly. If there is a reference group, it should be entered as sample 1. The pattern of findings in the members of the matched set must be entered, using 0 for "no" and 1 for "yes" (e.g., "0" for the observation in sample 1, "1" for the matched observation in sample 2, "0" for the matched observation in sample 3, etc. Matched sets can be entered individually, or sets with the same pattern of findings can be entered together, with their frequency.

If the data are stratified, enter each stratum in turn. Click on "All strata" whenever combined results are required.

The program provides **tests comparing the matched samples** (Cochran's *Q* test, Page's test for trend), pairwise (multiple) comparisons of the samples, **odds ratios**, and ***kappa***.

If *stratified data* are entered, the Cochran Q tests and Page tests in the separate strata are combined and the *heterogeneity* of the P-values in the strata is tested.

Tests comparing the matched samples

Cochran's Q test, which is an extension of the McNemar test for matched pairs, tests the null hypothesis that the probability of a "yes" result is the same in each sample, against the alternative that the relative probabilities in the different samples are consistent for all sets of related observations; that is, if in one set the probability of "yes" is larger in sample 1 than in sample 2, this is so in all sets. With three samples or small numbers (Tate and Brown 1970) a P-value that is near a borderline of significance should be treated with caution; the program provides a warning.

Page's test is appropriate if the samples fall into an ordered sequence. It is a test for the presence of a monotonic trend (Page 1963, Siegel and Castellan 1988: 184-188). The test is conservative when applied to dichotomous data, because of the large number of ties (Hollander and Wolfe 1999, p. 292).

If *stratified data* are entered, the Cochran Q tests and Page tests in the separate strata are combined by Stouffer's method (Stouffer *et al*. 1949, p. 45; DeMets 1987) to produce overall tests that control for the stratifying variables. Three different sets of weights are used for this purpose – weighting the test results equally, by the sample sizes in the strata, and by the square roots of the sample sizes. In addition, the *heterogeneity* of the P-values in the strata is tested.

Pairwise comparisons

The proportion of "yes" observations in each sample is compared with the proportion of "yes" observations in each other sample. For each comparison, the program displays the proportions and their difference (with a 95% confidence interval for the difference), and performs a significance test. Alternative P values (two-tailed) are displayed – one that is appropriate if there was an *a priori* hypothesis, and Sidak- and Boniferroni-adjusted values that take multiple testing into account and are appropriate if the comparison was not planned.

The Sidak and Bonferroni adjustments both assume that the comparisons are independent. The Sidak adjustment is slightly less "pessimistic" (Abdi 2007) - i.e., less severe, less conservative, and it has a bit more power than the Bonferroni method. So from a purely conceptual point of view, the Šídák method may be preferred). If the assumption of independence is false, both procedures "do a good job of protecting against false statements of statistical significance, but have less power to detect real differences" (GraphPad Statistics Guide 2013).

Since the confidence intervals and significance tests are based on different procedures, they do not completely correspond, especially if the number of observations is small.

Odds ratios

Odds ratios comparing each possible pair of samples are calculated.

Kappa

The program computes *kappa*, which expresses the agreement among all the observations in the matched sets, and may serve to express the effectiveness of the matching procedure, since it indicates the extent to which the findings in matched sets are more similar than findings in individuals from different sets (Fleiss *et al*. 2003: 617-618).

The probability of chance agreement is taken into account in the calculation of *kappa*. A value of 1 indicates perfect agreement (after allowing for this probability of chance agreement) between ratings; 0 indicates no agreement other than what can be attributed to chance, and a negative value indicates less than chance agreement. Fleiss *et al*. (2003) suggest that a value of 0.75 or more indicates excellent agreement, and 0.40 or less indicates poor agreement. Cicchetti and Sparrow (1981) divide Fleiss's 0.40–0.74 group into 0.60–0.74: good; and 0.40–0.59: fair. Alternative guidelines are: over 0.80, very good agreement; 0.61–0.80, good; 0.41–0.60, moderate; 0.21–0.40, fair; and 0.20 or less, poor agreement (Landis and Koch 1977, Altman 1991).

METHODS

Tests comparing the matched samples

Cochran's Q test is described by Siegel and Castellan (1988: 170-174), Daniel (1978: 241-244) and Zar (1998: 268-270).

Page's test is described by Siegel and Castellan (1988: 184-188). A large-sample approximation (formula 7.10) is used, since the available tables of critical values for small numbers are inappropriate in the presence of many ties (Hollander and Wolfe 1999: 291-292.

If *stratified data* are entered, the Cochran *Q* tests in the separate strata are combined by averaging their *z* values (Stouffer *et al*. 1949, p. 45; DeMets 1987). Three different sets of weights are used for this purpose – weighting the test results equally, by the sample sizes in the strata, and by the square roots of the sample sizes. Also, heterogeneity test is performed, comparing the P-values in the strata, using the formula (Wolf 1986: 45):

$$\text{chi-square } (k - 1 \text{ d.f.}) = \sum (Z_i - MeanZ)^2$$

where k = number of strata,
Z_i = z value in stratum i
$MeanZ$ = mean z value.

The Page tests for trend are combined in the same way, but using the signed *z* values provided by the tests, and without excluding sets that exhibit no differences between their members. The Page tests are not combined if there are 12 or fewer sets in any stratum, or 21 or fewer sets if the dependent variable has 3 categories.

Pairwise comparisons

A 95% confidence interval between the proportions of "yes" responses in two samples is estimated by the method described by Bi (2006: formula 5.1.4):

$$\text{Lower limit} = P_1 - P_2 - z^* \sqrt{\{[P_1 + P_2 - 2P_{12} - (P_1 - P_2)^2] / N\}}$$

$$\text{Upper limit} = P_1 - P_2 + z^* \sqrt{\{[P_1 + P_2 - 2P_{12} - (P_1 - P_2)^2] / N\}}$$

where P_1 and P_2 = the proportions of "yes" observations in the two respective samples
P_{12} = the proportion with "yes" observations in both of the respective samples
N = number of subjects (i.e., number of sets of matched observations)
z^* = the z value corresponding to a P value of *alpha** (i.e., the upper *alpha** point of the standard normal distribution)

$$alpha^*_j = 0.5 * [1 - (1 - alpha)^{1/c}]$$

alpha = 0.05 for a 95% confidence interval
c = the total possible number of pairwise comparisons = $k(k - 1) / 2$
N = number of sets of matched observations (e.g. number of subjects)
k = number of matched samples

McNemar tests are used to test the significance of the difference between the samples. Two-tailed P values are displayed. To compensate for the multiple testing, a Sidak-adjusted P value and a(Bonferroni-adjusted P value (i.e P multiplied by *c*) is also provided. The formula for the Sidak adjusted P value is $1 - (1 - P^c)$.

Odds ratios

For each pair of samples, the odds ratio is the number of matched pairs that have "yes" for the first sample and "no" for the second, divided by the number with "no" for the first sample and "yes" for the second.

Kappa

Kappa and its standard error are calculated by formulae 18.50 and 18.53 of Fleiss *et al*. (2003).

G1. COMPUTE *KAPPA* FOR 3 OR MORE RATINGS (NOMINAL DATA)

This module appraises the agreement between a fixed number (3 or more) of matched observations with respect to a variable with 2-10 categories. It might be used to measure the agreement between 3 or more ratings of the same individuals, e.g. by different observers or tests, or between ratings made by the same observer on different occasions.

The numbers of ratings (k = 3 or more) and the number of categories (2-10) must be entered. The findings in the set of ratings are then entered, by entering the number of ratings falling into each category (these should add up to k). Each set of ratings can be entered separately, , or sets with the same pattern of findings can be entered together, with their frequency.

The program provides the overall ***kappa***, and *kappa* values for individual categories.

If stratified data are entered, an overall value of *kappa* is computed.

Kappa

The overall *kappa* is computed, with its standard error and significance. *Kappa* values are also reported for individual categories, with their significance; but these test results should be treated with caution, since they are not based on a multiple-comparison procedure.

If *stratified data* are entered, an overall value of *kappa*, weighted by sample size, is computed.

The probability of chance agreement is taken into account in the calculation of *kappa*. A value of 1 indicates perfect agreement (after allowing for this probability of chance agreement) between ratings; 0 indicates no agreement other than what can be attributed to chance, and a negative value indicates less than chance agreement. Fleiss *et al*. (2003) suggest that a value of 0.75 or more indicates excellent agreement, and 0.40 or less indicates poor agreement. Cicchetti and Sparrow (1981) divide Fleiss's 0.40–0.74 group into 0.60–0.74: good; and 0.40–0.59: fair. Alternative guidelines are: over 0.80, very good agreement; 0.61–0.80, good; 0.41–0.60, moderate; 0.21–0.40, fair; and 0.20 or less, poor agreement (Landis and Koch 1977, Altman 1991).

METHODS

Kappa

Kappa and its standard error are calculated by formulae 18.51 and 18.53 of Fleiss *et al.* (2003).

G2. COMPUTE WEIGHTED *KAPPA* FOR 3 OR MORE RATINGS (ORDINAL DATA)

This module appraises the agreement between a fixed number (3 or more) of matched observations with respect to a variable with 3 or more ordered categories. It might be used to measure the agreement between 3 or more ratings of the same subjects, e.g. by different observers or tests, or between ratings made by the same observer on different occasions.

The numbers of ratings (k = 3–10) or more) and the number of categories (3–10) must be entered. The categories chosen by the various raters are then entered - either the ratings for each subject seprately, or the ratings for each set of subjects with an identical set of ratings (with their frequency).

The program provides the *weighted kappa*, and the *analysis of variance* on which it is based. A simple (unweighted) kappa is also displayed.

Weighted *kappa*

Weighted *kappa* measures the agreement between independent raters or ratings, using a set of ordered categories. ["Raters" and "ratings" are used synonymously in this module.] Cognisance is taken not only of complete agreements between ratings, but also of partial agreements, each combination of categories being given a weight based on their closeness. Scores of 1, 2, 3, etc. are allotted to the categories for this purpose (which assumes that the categories are more or less equally spaced along some dimension), and the weight given to each pair of observations depends on the size of the absolute difference between the scores of the categories in which the pair-mates fall. Complete agreement between two ratings is given a score of 1, and in other instances a quadratic weighting scheme is used; with weights that are inversely proportional to the square of the difference. between the two scores (Fleiss *et al.* 2003, formula 18.30). If there are 4 categories, the weight is 0.89 if the difference between scores is 1, 0.56 if it is 2, and 0 if it is 3. Quadratically-weighted *kappa* values tend to increase with the number of categories (Brenner and Kliebsch 1996).

The value of *kappa* is derived from an *analysis of variance*, since quadratically weighted kappa is equivalent to the intraclass correlation coefficient provided by such an analysis (Fleiss and Cohen 1973, Berry *et al.* 2008). The program displays the analysis of variance. This method supplies the same result as more elaborate computer-intensive methods. Different methods yield somewhat different P values for tests of the difference of *kappa* from zero (Berry *et al.* 2008),

The program also displays the *simple (unweighted) kapp*a, treating the categories as nominal – either there is agreement between the two ratings (score = 1) or there is not (score = 0).

Essenitially, both these versions of *kappa* are in general agreement with the basic formula

(*PO* – *PC*) / (1 – *PC*), where *PO* is the proportion of interrater agreement and *PC* is the proportion of agreement expected on the basis of chance alone. A large sample size (*N*) is required. As a rough rule of thumb, (*N* + 1) / *N* should be close to 1.0 (Cicchetti *et al.* 2006).

The probability of chance agreement is taken into account in the calculation of *kappa*. A value of 1 indicates perfect agreement (after allowing for this probability of chance agreement) between ratings; 0 indicates no agreement other than what can be attributed to chance, and a negative value indicates less than chance agreement. Fleiss *et al*. (2003) suggest that a value of 0.75 or more indicates excellent agreement, and 0.40 or less indicates poor agreement. Cicchetti and Sparrow (1981) divide Fleiss's 0.40–0.74 group into 0.60–0.74: good; and 0.40–0.59: fair. Alternative guidelines are: over 0.80, very good agreement; 0.61–0.80, good; 0.41–0.60, moderate; 0.21–0.40, fair; and 0.20 or less, poor agreement (Landis and Koch 1977, Altman 1991).

Methods

Weighted *kappa*

Kappa is computed from an analysis of variance (Berry *et al* 2008): formula 7):

$Kappa = ICC = (MS_{BS} - MS_{MxS}) / [MS_{BS} + (M - 1)MS_{MxS} + (M(MS_M) / (N - 1)$

where ICC = intraclass correlation coefficient
M = number of raters
N = number of subjects
MS_BS = between-subjects mean square
MS_M = between-raters mean square
MS_{MxS} = residual mean square (raters x subjects interaction).

The approximate P value is based on
$F = MS_{BS} / MS_{MxS}$ with N – 1 and (M – 1)(N – 1) degrees of freedom

Simple kappa

Kappa and its standard error are calculated by formulae 18.51 and 18.53 of Fleiss *et al*. (2003).

G3. APPRAISAL OF AGREEMENT BETWEEN 3 OR MORE RANKINGS

This module appraises the agreement between three or more (up to 20) rankings of 3–9 alternative choices, expressing the judgments of 3–20 raters. The available choices may be (for example) different diagnoses or treatments. These analyses may be useful in studies of reliability and as a basis for decision-making based on the raters' preferences.

Each rater's ranking must be entered, by allocating an index letter ('A', 'B', 'C', etc.) to each of the available choices, and entering the index letters in a sequence expressing the rater's preference. Ties are not acceptable.

As indicators of the raters' agreement on specific choices, the program reports each choice's median rank, and the *agreement coefficient A* (Riffenburgh and Johnstone 2009) for each choice. As indicators of the raters' overall agreement, it computes *Kendall's coefficient of concordance* and a *"top-down" coefficient of concordance*.

Agreement coefficient A

This coefficient (Riffenburgh and Johnstone 2009) is based on the absolute differences between the ranks that different raters ascribe to a given choice, and the choice's median rank. It generally ranges from 0 (the level of agreement expected by chance) to 1 (perfect agreement); a negative value indicates less agreement than might be expected by chance. A P value is displayed.

If the top choice (according to the median ranks) has a significant coefficient, this may be regarded as justifying its acceptance as the raters' recommendation.

Coefficients of concordance

Kendall's coefficient of concordance expresses the association between sets of rankings. The coefficient ranges from 0 (no agreement) to 1 (complete agreement). The coefficient's significance is reported, using tabulated critical values if the sample is small and a chi-square test if it is large.

The *"top-down" coefficient of concordanc*e (Zar 1998: 449-450) gives emphasis to high-ranking (preferred) choices. Its significance is reported.

METHODS

Agreement coefficient A

This coefficient is computed by formula 7 of Riffenburgh and Johnstone (2009), who provide tabulated critical values for P = 0.05 and P = 0.10 (tables 3 and 4), based on analyses of all possible permutations.

Coefficients of concordance

The formula for *Kendall's coefficient of concordanc*e is provided by (*inter alios*) Zar (1998: formula 20.67). Critical values for P = 0.05 and P = 0.01 (for small samples) are provided bySiegel and Castellan (1988: Table T); harmonic interpolation is used where necessary. The chi-square test (for larger samples) uses Siegel and Castellan;s formula 9.19.

The formula for the "*top-down" coefficient of concordanc*e is provided by Zar (1998: formula 20.67) ; the corresponding chi-square test uses formula 20.79.

H. COMPARISON OF TWO GROUPS OR TWO MEASURES (FIXED NUMBER OF MATCHED NUMERICAL OBSERVATIONS)

This module compares two sets (designated "cases" and "controls") of matched numerical observations. It can be used to compare two groups – index subjects with matched controls in a case-control study, cohort study, or trial – or two measurement methods.

For a comparison of groups, each matched set must contain between 3 and 11 observations in all., comprising a fixed number of "cases" (1 to 5) and a fixed number of "controls" (1 to 10). The matched sets of observations must be entered individually, after entering the numbers of cases and controls per set. Up to 500 sets may be entered.

For a comparison of measurement methods (A and B), equal-sized sets of replicate measurements by the two methods are required (2 to5 by each method). The two methods may be applied to the same subjects or to different subjects. The program terms the measurements by method A as "cases", and those by method B as "controls". After entering the numbers of "cases" and "controls" per set (numbers which must be identical), the measurements of each subject by method A must be entered, in a separate line, followed (in the same line) by the measurements (of the same or a different subject) using method B.

The results relevant to a comparison of groups are three **tests** (Rosner's and Walter's tests and a paired *t*-test) for the difference between the mean values, approximate confidence intervals for the **difference between the mean values**, **between-sets and within-sets variances** and. the **Hodges-Lehmann procedure.**

.

The results relevant to a comparison of measurements include a **95% repeatability coefficient** and **ANOVA table** for each method; the **95% limits of agreement** between the methods and the **relationship between the difference and the mean value** (appropriate if the two methods were applied to the same subjects); and **F-tests** for the difference between the methods, for the effect of repeated measurements, and for interaction, and a repeated-measures **ANOVA tables value** (appropriate if the two methods were applied to different subjects).

Tests

Rosner's test is a generalization of the paired *t*-test that takes account of within-sets and between-sets variability (Rosner 1982). If single index subjects are compared with controls, it appraises the significance of the differences between their values. If two groups of observations are compared, it appraises the difference between the mean values in the two groups. Two P-values may be displayed. If so, these may be regarded as the bounds of the true P-value. The true P-value depends on the relative magnitude of the within-sets and between-sets variabilities (see below), as explained by an on-screen message. The test sometimes presents technical difficulties, and is omitted.

Walter's test (Walter 1980) tests the significance of the mean case-control difference weighted by the numbers of cases and controls in the set. Rosner (1982) points out that (unlike his test) Walter's test assumes zero between-sets variability, and may therefore provide a misleadingly low P-value if there is much between-sets variability.

The *paired t-test* tests the significance of the unweighted mean difference between the case and control means within each matched set. Rosner (1982) points out that (unlike his test) the paired *t* test assumes zero within-sets variability, and may therefore provide a misleadingly low P-value if there is much within-sets variability.

If the numbers of cases and controls (assumed to be the numbers of replications by two methods of measurement) are equal, *F-tests* are performed for the difference between the two methods, for differences between repeated measurements, and for interaction - i.e. for a difference between the methods in the uniformity (reliability) of repeated measurements. Each of the latter two tests is done three times - without adjustment, and with two adjustments. The adjusted tests are Fleiss's "Approximation 3", which is not appropriate in all situations, and his "Approximation 4", which is valid in all situations but may be extremely conservative (Fleiss 1985: 227). The F-tests are appropriate only if the two methods of measurement were applied to different subjects.

Difference between the mean values

The program displays the mean case-control difference and its standard error, computed separately by the Rosner and Walter procedures and for unweighted data, with approximate 90%, 95%, and 99% confidence intervals.

Hodges-Lehmann procedure

This nonparametric procedure (Sprent 1993:89-90) determines the median of the differences between two matched sets, e.g matched cases and controls (with 90%, 95%, and 99% confidence intervals). [This is not necessarily the same as the difference between the medians, or the median of the differences observed in each matched set.]

A large-sample method of analysis is used if there are over 50 matched sets.

The analysis takes account of tied differences (if the large-sample method is used), but not of variation within matched sets.

Between-sets and within-sets variances

The between-sets variance computed by Rosner's procedure (Rosner 1982) is reported. This represents the variation between matched sets, and the within-sets variance represents the variation within either the case or the control group for a specific matched set. The ratio of the two variances is an indication of the value of multiple matching. If the between-sets variance is much larger than the within-sets variance, multiple matching brings little benefit (Rosner 1982; Lee and Wilkens 1994).

95% repeatability coefficient

If the numbers of cases and controls (assumed to be the numbers of replications by two methods of measurement) are equal, the 95% repeatability coefficient is computed for each method. This expresses the expectation (with 95% confidence) of the maximum size of the absolute difference between two observations using the same method.

95% limits of agreement

If the numbers of cases and controls (assumed to be the numbers of replications by two methods of measurement) are equal, the 95% limits of agreement are computed. These (which are appropriate only if the two methods of measurement were applied to the same subjects) answer the question, "given a measurement by one method, how far might this be from a measurement by the other method?" They demarcate the bounds of the range that, with a 95% probability, includes the difference between single measurements of the same subject by the two methods. The 95% confidence intervals of the limits of agreement are estimated (the limits of agreement may be very imprecise if the sample is small).

Use of the 95% limits of agreement assumes that the differences are reasonably constant throughout the range of measurement. To check this assumption, the program displays *Spearman's coefficient of correlation between the difference and the mean level* (also appropriate only if the two methods of measurement were applied to the same subjects). The correlation coefficient may be expected to be zero if the mean difference does not change with increasing values. Even when one of the methods of measurement is a new one and the other is an accepted standard, it is preferable to examine the relationship between the difference and the mean value rather than the relationship between the difference and the standard measurement, which (as shown by Bland and Altman 1995b) is likely to be misleading.

ANOVA tables

If the numbers of cases and controls (assumed to be the numbers of replications by two methods of measurement) are equal, a one-way ANOVA table is displayed for each method, showing the between-subjects and within-subjects components of variance, as well as a repeated-measurement ANOVA for the combined data (Fleiss 1986: 220-228), which is appropriate only if the two methods of measurement were applied to different subjects.
.

METHODS

The two groups of observations are referred to as "cases" and "controls".

Tests

Rosner's test (Rosner 1982) is a generalization of the paired *t*-test that takes account of within-sets and between-sets variability. It adjusts and appraises the significance of the mean within-set difference. The test sometimes presents technical difficulties, since it requires the computation of maximum-likelihood estimates by an iterative procedure that may fail to find an appropriate (positive) root. If this difficulty is encountered (usually because of marked within-set variability) an appropriate message is displayed.

In Rosner's procedure the within-pairing variability is calculated by Rosner's formula 2.2 (Rosner 1982), and maximum likelihood estimates of the between-pairing variability and the adjusted mean case-control difference

are then computed by an iterative process, using the van Wijnsgaarden-Dekker-Brent root-solver (Press *et al.* 1989: 283-286). The adjustment takes account of the numbers of cases and controls per set, using their reciprocals. Significance is appraised by Rosner's formula 2.3, using alternative degrees of freedom when referring the test statistic (*lambda*) to the *t*-distribution, namely $N - 2R$ and $R - 1$ (where N = number of subjects and R = number of matched sets). This provides two P-values (both of which are shown if they differ appreciably), which may be regarded as the bounds of the true P-value. The true value depends on the relative magnitude of the within-sets and between-sets variabilities.

Walter's test uses formula 2.4 of Rosner (1982). This permits application of the test to situations where there are matched sets with two or more cases.

The *paired t-test* is calculated by the usual formula (see, e.g. Selvin 1991: 65, formula 2.51), except that in each matched set the two values (of case and control) are replaced by the means (of cases, if there is more than one case, and of controls, if there is more than one control).

The *F-tests* are based on a repeated-measurement ANOVA (see Fleiss 1986: 220-228). The adjustments, which Fleiss calls Approximations 3 and 4, involve changes to the degrees of freedom (Fleiss 1986: 227: formulae 8.9 and 8,10); the changed degrees of freedom are rounded off to the nearest whole number. This ANOVA is not done if the number of measurements varies for different subjects.

Difference between the mean values

In Rosner's procedure (see above), the adjusted mean case-control difference is computed by weighting the difference in each matched set by

$$1 / \{B + W.[(1 / N1) + (1 / N2)]\}$$

where

B = between-sets variance
W = within-sets variance
N1 and N2 = numbers of cases and controls in the set.

In Walter's procedure, the difference in each matched set is weighted by

$$1 / [(1 / N1) + (1 / N2)]$$

Hodges-Lehmann procedure

The differences between the values of cases and controls are calculated in the *n* matched sets. Where there are more than one case or control, their respective median values are used.

As described by Han (2008), each difference is then compared with each other difference, and for each of these $m = n(n-1)/2$ comparisons of two values, the mean of the pair of differences (Walsh value) is computed. The *m* means are then ranked in ascending order, and their median is determined. This is the point estimate of the Hodges-Lehmann median difference between cases and controls.

If $n <= 50$, a value *R* corresponding to the value of *n* is obtained from Table A12 of Conove r(1999: p. 545), using the $W_{0.005}$, $W_{0.025}$, and $W_{0.05}$ column for the 90%, 95%, and 99% confidence intervals respectively. The lower confidence limit is the Walsh value whose rank is *R* in the series, and the upper confidence limit is the Walsh value whose rank is R from the upper end of the series,

If $n > 50$, confidence intervals are estimated by a large-sample approximation (Hollander and Wolfe 1999:132-133, using the formulae provided by Han (2008), but with a correction for tied ranks (Unistat Statistics Software). The lower confidence limit is the Walsh value whose rank is *R* in the series, where $R = .za./.b$ rounded up to the nearest integer), and the upper confidence limit is the Walsh value whose rank is *R* from the upper end of the series,

where z = -1.645, -1.96, or -2.5767 (for 90%, 95%, or 99% limits respectively)

$a = \sqrt{[n(n + 1)(2n + 1) / 24 - Tee / 48]}$

$b = n(n + 1) / 4$

n = number of matched sets

Tee = the sum of $(t_i^3 - t_i)$

t_i = the number of ties in each set of tied ranks

The correction for ties (*Tee* / 48) in calculating *a* is omitted if it reduces *a* to zero or a negative value.

95% repeatability coefficient

The computation of the coefficient of repeatability is explained by Bland and Altman (1999: 149).

95% limits of agreement

The 95% limits of agreement and their confidence intervals are computed by the method explained by Bland and Altman (1999; section 5.1: formulae 5.3 and 5.10), using within-subject mean squares based on one-way analyses of variance for the two methods (Guilford and Fruchter 1986: 234-5: formulae 13.15 and 13.16).

I1. COMPARISON OF 3 TO 10 SAMPLES OR REPLICATES (FIXED NUMBER OF MATCHED NUMERICAL OBSERVATIONS)

This module is for use in studies based on dependent samples of numerical (ordinal or interval-scale) observations. It can appraise the findings in 3 to10 related samples (each member of which is matched with members of all other samples), or 3 to 10 sets of measurements of each subject. It may be used, for example, to analyse a trial in which matched subjects receive 3 to 10 different treatments, or one in which each subject receives 3 to 10 different treatments, or the same treatment under 3 to 10 different conditions, or an observational study comparing matched subjects who have different degrees of exposure to a risk factor or are measured under different defined conditions or by different observers, or a study in which the same individuals are observed under different specified conditions or at various specified times, or are asked different specified questions.

The module can be used in reliability studies in which each subject's measurements are replicated 3 to 10 times, either using the same method, or using 3 to 10 different observers or methods of measurement.

If the samples lie (or are assumed to lie) in an ordered sequence, they should be arranged accordingly. If there is a reference group, it should be entered as the first sample (A). In reliability studies, replicates may be entered in any order, unless they represent fixed instruments, observers, times, conditions, etc.

Each set of related observations must be entered separately (up to 500 sets). For 30 or fewer sets, the program reports the mean of each set ,with its standard deviation and coefficient of variation. If a normal distribution is not assumed, ranks can be entered instead of the measurements, e.g. 1 3 2 instead of 6.1 11 9; or (for ties) 1 4 2.5 2.5 instead of 6.1 11 9 9 (giving tied observations the mean of the ranks they would have if they differed slightly).

The program appraises the **differences** between the samples and provides **measures of effect** and **measures of agreemen**t and **disagreement**. Some of the procedures are *nonparametric,* and are applicable to all numerical data: *Friedman's two-way analysis of variance by ranks***,** *Quade's test for non parametric two-way analysis of variance, no**n**parametric pairwise comparisons, Kendall's concordance coefficient,* and the *Spearman's correlation coefficient*. Others are *parametric,* and are applicable only to interval-scale data with an assumed normal distribution*: *analysis of variance, F-tests,* and other *pairwise comparisons; omega-squared, eta-squared, and Fisher's F index; intraclass correlation coefficients, repeatability coefficient,* and *Spearman-Brown coefficients of reliability;* and (optionally) *tests for equivalence.* The parametric procedures are not appropriate if ranks are entered.

For stratified data, enter each stratum in turn, and click on "All strata" for combined results. If stratified data are entered, the Friedman and Page tests in the separate strata are combined and the *heterogeneity* of the P-values in the strata is tested.

* [As pointed out by Altman (1991: 330), it may not be the raw data, but the residual values (after allowing for the effects of sample membership and matched-set membership), that should be normally distributed.]

Analysis of variance

A one-way analysis of variance (single-factor within-subjects ANOVA) is performed. The analysis assumes that the subjects were selected randomly from the population they represent, that distributions are normal, and that the data in the various samples have similar variances and covariances. A significant result points to a significant difference between the means of at least two of the populations represented by the samples.

Comparison of samples

The *F test*, which is appropriate for interval-scale data with an assumed normal distribution, test the null hypothesis that there is no difference among the mean values of the various samples. It is based on the analysis of variance. A significant result points to a significant difference between the means of at least two of the populations represented by the samples. An adjusted F-test is also performed, using a usually conservative method described by Geisser and Greenhouse (1958) as an extension of the results of Box (1954). This result is appropriate if the homogeneity of variances and covariances is in doubt, but it "may be too conservative".

In addition, *Friedman's two-way analysis of variance by ranks* (Siegel and Castellan 1986: 174-183; Zar 1998: 263-267) is performed. This is an extension of the sign test, and is applicable to all numerical data. It tests the null hypothesis that the values in the different samples represent the same population median, against the alternative that at least two of the samples have different medians. *Quade's test for non parametric two-way analysis of variance* (Quade 1979, Conover 1999: 373-30), which is an extension of the Wilcoxon signed-rank test, is also performed. The Quade test may be more powerful for a small number of related values, while the Friedman test may be more powerful when the number of related values is five or more.

Assuming a normal distribution, the program provides two sets of 90%, 95%, and 99% *confidence intervals for the mean* of each sample. The first set is based on the estimated variance in the specific sample, and the second set (which has narrower intervals) is based on the within-samples variance, on the assumption that the samples have similar variances (Sheskin 2007: 1052-1053). The program also provides two sets of 90%, 95%, and 99% *confidence intervals for the difference* between each pair of sample means, one using Fisher's LSD procedure and one using the Scheffé procedure (Sheskin 2007: 1034-1035).

The *pairwise comparisons*, testing the differences between all sample means (assuming a normal distribution), use Fisher's LSD procedure and the Scheffé procedure (which is more conservative). Nonparametric pairwise tests (applicable to all numerical data) are based on the Friedman procedure, and are done only if the Friedman test reveals a significant difference ($P < 0.05$) between samples; the median of each set of matched observations is displayed, with (if the number of observations is at least 10) the interquartile range.

The multiple-comparison tests include a set of comparisons of each sample mean with that of Sample A. The Dunnett procedure (Dunnett 1964) is used for this purpose.

Optionally, *equivalence tests* are performed, testing the equivalence of the matched measurements by the procedure described by Yi et al. (2007). This requires entry of the

bounds of "equivalence", i.e., the largest difference between measurements that is to be regarded as negligible or 'acceptable'. The tests are based on a comparison of the within-subject variance with this specified difference (and also with this difference multiplied by 0.5, 0.75, 1.5, or 2). A P value under 0.05 implies good agreement (negligible variation, i.e. equivalence) at a 5% significance level/

If *stratified data* are entered, the results of the Friedman analyses of variance in the separate strata are combined by Stouffer's method (Stouffer *et al.* 1949, p. 45; DeMets 1987) to produce overall P-values that control for the stratifying variables. Three different sets of weights are used for this purpose – weighting the test results equally, by the sample sizes in the strata, and by the square roots of the sample sizes. In addition, the *heterogeneity* of the P-values in the strata is tested.

Trend of the samples

Page's test for a monotonic trend (Page 1963, Siegel and Castellan 1988: 184-188) is applicable to all numerical data. For the test to be meaningful, the samples should be entered in the sequence to be tested. The test might, for example, be a way of appraising the dose-response relationship in a trial in which different doses are given to different matched samples or to the same individuals at different times.

Measures of effect

Four measures of the magnitude of the effect – i.e., the strength of the association between the independent variable (represented by the various samples) and the dependent variable – are computed.

Omega-squared (ω^2) is an estimate of the proportion of variability of the dependent variable that is associated with variability in the independent variable, i.e. with differences between the samples (Sheskin 2007: 1049-1050). By Cohen's criteria, a value of 0.1379 or more indicates a large effect size, 0.0588 or more (but less than 0.1379) indicates a medium effect size, and 0.0099 or more (but less than 0.0588) indicates a small effect size (Sheskin 2007:1051). Cohen (1988) warns that these criteria should be used only when there is no better basis for evaluation. A zero or negative value indicate absence of an association The program computes two versions of *omega*-squared – *standard omega-squared*, which assesses the effect on total variability, and *partial omega-squared*, which is said to be more meaningful because it eliminates subject variability from the total variability (Sheskin 2007: 1050).

Eta-squared (η^2) is an alternative estimate of the proportion of variability of the dependent variable that is associated with differences between the samples. The program computes an adjusted eta-squared (Sheskin 2007: 1072), which tends to overestimate the relationship between the independent and dependent variables.

Cohen's f index (Sheskin 2007; 918) is a "standard deviation of standardized means". By Cohen's criteria, a value of 0.4 or more indicates a large effect size, 0.25 or more (but less than 0.4) indicates a medium effect size, and 0.1 or more (but less than 0.25) indicates a small effect size (Sheskin 2007: 1051).

Measures of agreement

Kendall's coefficient of concordance (which varies between 0 and 1) is based on the ranks of the observations within each related set, and expresses the degree of similarity of their ranking in different samples.

The *average Spearman's coefficient of rank correlation* between all possible pairs of rankings can vary from $-1 / (k - 1)$ to 1, where k is the number of matched observations in a set.

Intraclass correlation coefficients, which are appropriate for interval-scale data with an assumed normal distribution, are measures of agreement that express the correlation (in terms of absolute agreement) between measurements within individuals or sets of matched individuals. Six intraclass correlation coefficient (ICC) values are computed (Shrout and Fleiss 1979), with their 95% confidence intervals.

Each ICC is appropriate in a different situation. (a) The values with the rubric "two-way model with fixed raters" are appropriate in studies where the matched observations in each set represent various "unique" raters, and no inferences are made about other raters; "raters" denote the various observers, treatments, methods or conditions of observation, matched individuals, or (in a reliability study of a questionnaire or other scale) questions or other scale items, that were studied. Two such ICCs are provided. The first, which Shrout and Fleiss refer to as model 3.1, uses a single measurement as the unit of analysis, and the second (model 3,k) uses an average measurement. (b) The two ICC values reported as "two-way model with random raters" are appropriate if the raters were randomly selected from a larger population of raters and it is proposed to generalize the findings to this larger population. If analysis is based on a single measurement, this is model 2,1; if it based on an average measurement, it is model 2,k. (c) The third pair of ICC values, entitled "one-way random model", is appropriate in methodological or other studies where the measurements are replications by the same observer or using the same instrument, and the order in which they are entered does not matter (this does not apply to the other ICC values).. They apply to the use of a single measurement (model 1,1) – e.g. in studies to determine the reliability of a single measurement – or to an average measurement (model 1,k) – e.g. in studies to determine the reliability of an average measurement.

The maximum value of an ICC is 1; the lower limit is an indeterminate negative value. As a rule of thumb, it has been suggested that ICC values above 0.75 should be regarded as evidence of excellent, and values above 0.4 as evidence of good, reliability (Shoukri and Pause1999: 27).

In the appraisal of replicated measurements a low ICC may express variability of the characteristic measured, as well as low reliability of measurement; this is especially important if measurements were conducted at different times.

The *coefficient of repeatability* is applicable if replicate measurements were entered, and is appropriate for interval-scale data with an assumed normal distribution. It expresses the expectation (with 95% confidence) for the maximum size of the absolute difference between a pair of observations, assuming that repeatability is similar at all magnitudes. Approximate confidence intervals are estimated for the coefficient.

Spearman-Brown coefficients of reliability provide estimates of the effect of using the means of replicated observations. They predict what the reliability would be if two, three, four, or five replications were averaged.

Measures of disagreement

The degree of disagreement between each pair of samples is appraised by use of the information-based measure of disagreement (IBMD) between two sets of matched numerical observations proposed by Costa-Cantos *et al.* (2010), which is based on the differences between the paired observations, in relation to the magnitude of the larger value in the pair. It is applicable to ratio-scale variables (i.e., those where a zero value indicates absence of the attribute) that have positive values. The measure ranges from 0 (no disagreement) to 1 (strong disagreement).

Optionally, 95% confidence intervals are estimated for these measures of disagreement, using a bootstrap procedure (see Sheskin 2007: 532-536), which may cause a delay in the calculation if the sets of matched observations are large or numerous. If the delay is too long, the procedure can be aborted by clicking on the "Stop" button. If the data are not extensive, the confidence intervals are estimated by default.

In addition, an overall measure of the disagreement among the multiple samples (Henriquez *et al.* 2013) is reported, together with its approximate 95% confidence intervals.

METHODS

Analysis of variance

The method is described by (Zar 1998, 255-260) and, in detail, by Sheskin (2007: 1025-1031 and 1056).

Comparison of samples

The *F test* is based on the analysis of variance (Zar 1998, 255-260). Geisser and Greenhouse's alternative test is described by Sheskin (2007: 1045).

In *Friedman's two-way analysis of variance by ranks* (Siegel and Castellan 1986: 174-183; Zar 1998: 263-267), the program uses criteria for $P < 0.05$, 0.01, and 0.001 listed by Zar (1998: Table B.14) if there are 3 or 4 samples with less than 16 values in each, or 5 or 6 samples with less than 11 values in each. Otherwise significance is appraised by use of the Friedman statistic, which has an approximately chi-square distribution unless numbers are small. Also, use is made of Iman and Davenport's *F* (Iman and Davenport 1980), which is generally more powerful (Zar 1998: 264). The formula for Iman and Davenport's *F* is provided by Sprent (1993: 145) and Zar (1998: formula 12.47), with N - 1 and (k - 1)(N - 1) degrees of freedom. If the rankings in the sets are identical, *F* has a value of infinity, and

$$P = (1 / N)(k - 1),$$

where N = number of samples
k = number of sets.

In *Quade's test for non parametric two-way analysis of variance (*Quade 1979, Conover 1999: 373-30) significance is appraised by using the *F* distribution, which approximates the exact distribution of the test statistic. The F approximation becomes closer as the number of sets of related measurements increases.

In *Page's trend test* (Siegel and Castellan 1988: 184-188), Page's statistic *L* is calculated by formula 7.7, and *Z* by formula 7.10. A one-tailed P-value is computed, based on the normal distribution; but if numbers are small, (3 groups with < 21 observations in each, or 4-10 groups with < 13), the Page statistic *L* is compared with critical values for $P < 0.05$, $P < 0.01$, and $P < 0.001$ (Siegel and Castellan 1988: 354-355, Table N).

The *parametric multiple comparison tests* that compare the mean of sample A with all other sample means use the , Dunnett procedure (Zar 1998, 217-218; Dunnett 1964). The results are appraised in relation to critical values of the *Q* distribution (Zar 1998, Tables B6 and B7), and are reported as $P < 0.01$, $P < 0.05$, or not significant ($P > 0.05$). The parametric pairwise tests that compare each sample with every other sample use formulae that solve *F* by equations derived from equations 24.17 (for Fisher's LSD test) and 24.18 (for the Scheffé test) of Sheskin (2007: 1034), after substituting the observed difference between means for *CD*.

The *non parametric multiple comparison tests* are described by Siegel and Castellan (1988: 180-183); the comparisons with sample A use critical values for $P < 0.05$ and 0.01 derived from Siegel and Castellan (1988: 321, Table Aiii), and the pairwise comparisons that compare each sample with every other sample use critical levels for $P < 0.05$, $P < 0.01$, and $P < 0.001$.

If *stratified data* are entered, the results of the Friedman analyses of variance in the separate strata are combined by averaging their *z* values (Stouffer *et al*. 1949, p. 45; DeMets 1987). Three different sets of weights are used for this purpose – weighting the test results equally, by the sample sizes in the strata, and by the square roots of the sample sizes. In addition, a heterogeneity test is performed, comparing the P-values in the strata, using the formula (Wolf 1986: 45):

$$\text{chi-square } (k - 1 \text{ d.f.}) = \sum(Zi - \text{MeanZ})^2$$

where k = number of strata,
Z_i = z value in stratum i
MeanZ = mean z value.

The Page tests for trend are combined in the same way, but using the signed *z* values provided by the tests, and without excluding sets that exhibit no differences between their members. The Page tests are not combined if there are 12 or fewer sets in any stratum, or 21 or fewer sets if the dependent variable has 3 categories.

Tests of equivalence

The method is described by Yi *et al*. (2007).

$$\text{Chi-square} = SSW / (D^2 \times 1.96 \times 1.96 \times 2)$$

where SSW = within-subject variance (based on ANOVA)
D = maximum acceptable difference

The P value for the test is 1 minus the P value associated with this chi-square, with $n(k\text{-}1)$ degrees of freedom, where n = no. of sets of paired measurements and k = no. of repeated measurements (e.g. 3)

Measures of effect

These measures are computed by equations 24.25 (for standard *omega*-squared), 24.28 (for partial *omega*-squared), and 24.40 (for the adjusted *eta*-squared), of Sheskin (2007). Cohen's *f* index (Sheskin 2007: 1051)is not computed if *omega*-squared is negative.

Measures of agreement

Kendall's coefficient of concordance is derived from the Friedman statistic by formula 12.51 of Zar (1998). Its significance is tested by computing chi-square (using formula 9.19 of Siegel and Castellan 1988: 269), unless there are under 21 matched sets and under 8 samples, when use is made of the critical values in Table T of Siegel and Castellan 1988: 365).

The *average Spearman's coefficient of rank correlation* between all possible pairs of rankings is derived from Kendall's coefficient of concordance (Siegel and Castellan 1988: 262).

The following formulae (Shrout and Fleiss 1979) are used for the six intraclass correlation coefficients. Shrout-Fleiss ICC models 1,1 and 1,k are computed from a one-way random effects model ANOVA, models 2,1 and 2,k from a two-way random effects model ANOVA, and models 3,1 and 3,k from a two-way mixed effects model ANOVA.

$$\text{ICC model 1,1} = (\text{MSB} - \text{MSW}) / [\text{MSB} + (\text{k} - 1)\text{MSW}]$$

ICC model 1,k = (MSB – MSW) / MSB
ICC model 2,1 = (MSB – MSE) / [MSB + (k – 1) MSE + k(MSJ – MSE) / N]
ICC model 2,k = (MSB – MSE) / [MSB + (MSJ – MSE) / N]
ICC model 3,1 = (MSB - MSE) / [MSB + (k - 1)MSE]
ICC model 3,k = (MSB – MSE) / MSB
where MSB = between-subjects mean square
MSE = residual within-subjects mean square
MSW = within-subjects mean square
N = number of subjects
k = number of observations in matched set
Formulae for confidence intervals for the six ICC models are provided by McGraw and Wong (1996a and 1996b) in their Table 7, where they are referred to as ICC(1) and ICC(k) for Case 1, and ICC(A,1) and ICC(A,k) for Cases 2 and 3, The formulae (except those for models 2,1 and 2,k) are set out in a convenient code by Steinley and Wood (2000). Linear interpolation is used to estimate F values that are based on non-integer degrees of freedom (and 1 d.f. is substituted for <1 d.f.) in the computation of confidence intervals for models 2,1 and 2,k; the latter results may differ slightly from those provided by SPSS, which handles non-integer degrees of freedom differently.

The *Spearman-Brown prediction formula* (Fleiss 1986: 14-15: formula 1.3) for reliability (*R)* is

$$R = Nr / [1 + (N – 1)r]$$

where N = number of replicates that are averaged
r = intraclass correlation coefficient (model 1,1)
This application of the Spearman-Brown formula was suggested by its use by Solomon (2004).

Fleiss's formula 1.31 is used to estimate the number of replicates required to obtain a reliability of 0.75 or 0.8:

$$N = P(1 – r) / [r(1 – P)],$$ where P = 0.75 or 0.8

The computation of the *coefficient of repeatability* is explained by Bland and Altman (1999: 149). Approximate confidence intervals are obtained by substituting confidence limits for the within-samples variance, estimated by the method described by Zar (1998: formula 7.16), in the formulae.

Information-based measure of disagreement (IBMD)

The formula for the pairwise comparisons (Costa-Santos *et al.* 2010) is

$$\sum L_i / n$$

where $L_i = \log\{[a_i - b_i| / \max(a_i,b_i)] + 1\}.\log(2)$
or (equivalently) $L_i = \log2\{[|a_i - b_i| / \max(a_i,b_i)] + 1\}$
a_i and b_i are the observations in pair *i*
n = the number of pairs of observations
If a_i and b_i are equal, L_i is taken as 0.

The measure is not computed if any a_i or b_i is negative, or if there are over 500 sets of matched observations.

The confidence interval is obtained by a bootstrap procedure, using the basic percentile method (Efron 1981, Efron and Gong 1983) as described by Sheskin (2007: 532-536). The approximate 95% limits are the (2.5)th and (97.5)th percentiles of the distribution of the measures of disagreement (computed by the above method) in 1000 random samples of the same size as the original sample, each drawn (with replacement) from the values in the original sample. Because of resampling, repetitions of the procedure may yield slightly different results.

The random sampling in this bootstrap procedure uses a pseudo-random number generator described by Wichman and Hill (1985), which derives each number in turn from three seed numbers that it modifies for subsequent use. Initial values for the seed numbers are generated by Delphi's inbuilt random-number procedures, namely RANDOMIZE, using the system clock, and RANDOM, which generates three random numbers from which the required seed numbers are computed. Delphi's RANDOM procedure is augmented by an additional randomizing shuffle, using the algorithm of Bays and Durham, as described by Press *et al.* (1989: 215-217). The formula for each selection is

$$\text{trunc}(RM) + 1$$

where R is a random number in the range 0 < R < 1

M = the number of candidates.

The *overall IBMD* is computed by the formula provided by Henriquez *et al.* (2013). It is equivalent to a simple average of the IBMD values computed for the pairwise comparisons. Its 95% confidence intervals are estimated by averaging the lower confidence limits and the upper confidence limits, respectively, for the pairwise IBMD values.

I2. ASSESSMENT OF INTERRATER AND INTRARATER RELIABILITY

This module assesses interrater and intrarater reliability in a study that compares replicated independent ratings (a fixed number of interval-scale numerical measurements) of the same subjects made by each of two or more raters. The "raters" may be different observers, different measuring instruments, or different methods or conditions of measurement. They may be specific raters of interest, e.g. two different instruments ("fixed raters"), or they may represent a larger population of raters ("random raters").

The required entries are the number of raters , the number of ratings of each subject by each rater), and the observations. The total number of ratings of each subject cannot exceed 12.

The program computes **reliability coefficients** (*intraclass correlation coefficients*) for random and fixed raters (with their approximate confidence uintervals and with significance tests), confidence intervals for the **difference between reliability coefficients**, various measures of the **standard error of measurement** and the **minimum significant change,** and **coefficients of interrater agreement and variability.** The *analyses of variance* on which the results are based are displayed.

Reliability coefficients

Interrater reliability (the variability among the raters) and *intrarater reliability* (the variability within the raters) are expressed by *intraclass correlation coefficients*. Separate results are provided for *random raters* (for use if the results are to be generalized to other raters) and for *fixed raters* (where the results apply only to the raters that were studied). Interrater reliability is reported for random and fixed raters, and intrarater reliability for random raters and (for use if the raters are fixed) for each separate rater. The program uses the procedures described by Eliasziw *et al*. (1994), in which, in order to enhance precision, each individual measurement contributes to the estimation of both interrater and intrarater coefficients.

Lower confidence limits are reported, at three confidence levels: the 90% two-sided confidence limit (which is equivalent to the 95% one-sided confidence limit), the 95% two-sided confidence limit (which is equivalent to the 97.5% one-sided confidence limit), and the 99% two-sided confidence limit (which is equivalent to the 99.5% one-sided confidence limit). Upper confidence limits are reported for each of the above coefficients; for some data sets, the computation yields anomalous upper limits for the intrarater ICC for random raters, and these are then not reported.

Haber *et al*. (2005) have pointed out that the interrater reliability coefficient measures the total rater-related variability, and is influenced by between-subjects variability. If there are substantial differences between subjects the coefficient may be close to 1 even when there are important differences between observers.

One-tailed *significance tests* are performed, testing the null hypothesis that the ICC is below or equal to a selected minimum level of reliability that is considered acceptable, against the alternative that the ICC is greater than this minimum level (Eliasziw *et al.* 1994). A low P value rejects the null hypothesis, and suggests that the measurements have an acceptable level of reliability. The test is approximately equivalent to a comparison with the corresponding 100(1-*alpha*)% one-sided lower confidence limit. The tests are applied to the interrater ICC (for random or fixed raters) and to the intrarater ICC for each rater.

Landis and Koch (1977) have suggested the following criteria for the ICC: 0.0–0.20, slight reliability; 0.21–0.40, fair; 0.41–0.60, moderate; 0.61–0.80, substantial; and 0.81–1.00, almost perfect reliability.

If ratings by two raters are entered, the program estimates 90%, 95%, and 99% confidence intervals are estimated for the difference between their intrarater reliability coefficients, based on the confidence intervals of these coefficients.

Standard error of measurement

The standard error of measurement (SEM), or "*measurement error*", which has comparable meaning to a standard deviation, summarizes the variability among or within the raters' measurements. It is calculated separately for interrater and intrarater comparisons, and different values are reported for random raters (for use if the results are to be generalized to other raters) and for fixed raters (where the results apply only to the raters that were studied). The values that are displayed are the interrater SEM for random raters, the interrater SEM for fixed raters, the intrarater SEM for random raters, and (for use if the raters are fixed) an intrarater SEM for each rater. The interrater SEM takes account of the variability within raters' measurements as well as the variability among raters' measurements (Eliasziw *et al.* 1994).

The SEM may be used for a *significance test* to appraise whether there has been a real change in a subject's rating, in excess of interrater or intrarater variability (Eliasziw *et al*. 1994). This is done by dividing the observed change by (1.414 x SEM) and comparing the result (*z*) with critical values of the standard normal distribution; for example, if *z* exceeds 1.96, P is under 0.05. If the same rater made both measurements, the SEM to be used for this purpose is the intrarater SEM for that rater. If there were different raters, an interrater SEM (for random or fixed raters) is appropriate.

Minimum significant change

As a simple aid to the appraisal of changes in subjects' ratings, the observed change can be compared with the "minimum significant change". If the observed change exceeds this minimum significant change, it is fairly certain that a real change has occurred (Eliasziw *et al*. 1994). Various values are provided for the minimum significant change, depending on whether the ratings were made by the same rater or by different raters (fixed or random) and whether alpha is set at 0.05 or 0.01.

The minimum significant change may be referred to as a *repeatability coefficient*, expressing the expectation (with 95% or 99% confidence) of the maximum size of the absolute difference between paired observations.

Coefficients of interrater agreement and variability

Coefficients of individual agreement (*CIA*s) are computed if two raters (observers or methods) are compared. They are based on the disagreements, for each subject, between two fixed raters, in comparison with the disagreements between replicated observations by the same rater). An acceptably high coefficient (i.e., avalue close to 1 or above 1) indicates that the raters can be regarded as interchangeable; i.e., replacing one by the other does not substantially increase the disagreement between measurements made on the same subject. These coefficients should be used only if the levels of intrarater agreement are acceptable, since intrarater disagreements are the standard with which the between-rater disagreements are compared (Barnhart *et al.* 2007, Haber and Barnhart 2008, Pan *et al.* 2010). Separate coefficients of individual agreement are computed, based respectively on the assumptions that one or other of the observers or methods is a "gold standard", or that neither is a "gold standard".

Using the CIAs, a lower limit of 0.8 for "acceptable" agreement has been suggested; this indicates that the disagreement between the raters does not exceed the disagreement between replicated observations (by the same rater) by more than 25%. In comparisons of the effects of two drugs on the same subjects (i.e. if "methods" A and B are drugs A and B), the U.S. Food and Drug Administration uses an individual bioequivalence criterion (IBC) that can be derived from the CIA by the formula IBC = [2(1 – CIA)] / CIA (Barnhart *et al.* 2007), and it recommends an upper limit of 2.495 for declaring individual bioequivalence (Food and Drug Administration 2001). This is equivalent to a lower limit of 0.445 for the CIA.

If there are the same number of replicated observations (up to 6) by each of two observers (or methods),the program also computes the *adjusted coefficient of individual equivalence* ( ), using a permutation-based nonparametric procedure, with its 90%, 95%, and 99% confidence intervals (Pan *et al.* 2011b). The suggested criterion is that a value of 0.8 or more indicates good agreement between the observers. The program uses the absolute difference between observations as its measure of disagreement; (if it used the square of the difference, the CIEA would in this instance be the same as the CIA). If the interobserver difference is smaller than the intra-observer differences the CIEA may be severely biased (M Haber, personal communication), and a warning is then displayed. Computer simulations indicate that this coefficient is robust, and relatively little affected by the degree of between-subject variability (whereas the interrater concordance coefficient tends to be inflated when there is much between-subject variability) (Pan *et al.* 2011a).

Large-sample confidence intervals are provided.

The c*oefficient of interobserver variability* (*CIV*), an index proposed by Haber *et al.* (2005), is the ratio of the interrater component of variability to the total (between and within) rater-related variability. Its maximum is 1, and it has the same value for random and for fixed raters. Unlike the intraclass correlation coefficient, it is not influenced by between-subjects variability. Its significance is reported (null hypothesis: coefficient = 0). A negative value may (like a zero value) be interpreted as maximal agreement, since it means that the interobserver variability is smaller than would be expected if the raters agreed perfectly (Haber MJ, personal communication, 2005).

The program also displays the *coefficient of interobserver agreement* (which is 1 - *CIV*) and

its significance (null hypothesis: coefficient = 1). If there are two observers, this coefficient is equivalent to the *CIA* where there is no "gold standard".

The *coefficient of excess observer variability*, which is 1 / (1 - *CIV*), is the ratio of the total observer variability to the variability that would be expected if the raters agreed perfectly. Its maximum is infinity. A value of 1 (or a value under 1, if the *CIV* is negative) indicates that there is no excess variability due to true differences between the raters.

METHODS

Ratings may be entered for up to 900 subjects.

The computation of *intraclass correlation coefficients*, the *standard error of measurement*, and the *minimum significant change* is described in detail by Eliasziw *et al.* (1994). Formulae for the intraclass correlation coefficients are on pp. 779-780, for the confidence levels on pp. 781-782, for the one-tailed *significance tests* on pp. 785 and 786, and for the standard error of measurement on p. 783. The minimum significant change is $z.\sqrt{(2)}.\mathrm{SEM}$ (using the appropriate SEM), where z = 1.96 (for *alpha* = 0.05) or 2.576 (for *alpha* = 0.01). The formulae are also provided by Hayen *et al.* (2007)

Confidence intervals for the intrarater coefficients for specific fixed raters are estimated by the formula for model ICC(1) provided by McGraw and Wong (1996a: Table 7), and set out in a convenient code by Steinley and Wood (2000).

If one of the degrees of freedom for *F* is not an integer, it is rounded off to the nearest integer. If it is below 1, it is changed to 1.

The computation of the coefficient of interobserver variability (CIV) and its derivatives is described by Haber *et al.* (2005). Formula 6.2, which provides the same numerical result as formula 3.2, is used to estimate the CIV. The *coefficient of interobserver agreement* is 1 - CIV, and the *coefficient of excess observer variability is* 1 / (1 - CIV). The significance test is described on pp. 78-79.

Coefficients of individual agreement are computed by the methods described by Haber and Barnhart 2008. For each subject, the mean squared deviation (MSD) between all pairs of observations (comparing the two raters, A and B) is computed. These values are then averaged over all subjects, to provide an overall MSD(A,B). Corresponding average values for the overall within-rater MSDs, namely MSD(A,A') and MSD(B,B'), are obtained by doubling the within-subject (residual) mean squares provided by the analyses of variance for fixed raters.
If neither A nor B is a gold standard, CIA = [MSD(A,A') + MSD(B,B') / 2] / MSD(A,B).
If A is a gold standard, CIA = MSD(A,A') / MSD(A,B).
If B is a gold standard, CIA = MSD(B,B') / MSD(A,B).

The *adjusted coefficient of individual equivalence (*CIEA) is computed by the formulae provided by Pan *et al.* (2011b). The computation is based on the mean absolute deviations between observations (MAD); if it were instead based on mean squared deviations (MSD), the coefficient would (in this instance, where there aret he same number of replicated observations for each observer) be equivalent to the CIA with no gold standard (Pan *et al* 2011a: Appendix C1).

Confidence intervals for the difference between two intrarater coefficients are computed by the MOVER (method of variance estimates recovery) technique, as described and tested by Ramasundarahettige *et al*. (2009), using their formulae 8 and 9.

Repeated-measure analyses of variance are performed (Armitage *et al.* 2002: 244-246: Example 9.2). The between-raters-within-subjects sum of squares (used for estimating the CIV) is the sum of the between-rater and interaction sums of squares (Haber *et al.* 2005). The observed mean squares are calculated in the same way for random raters and fixed raters, but the expected mean squares and estimated variance components are calculated differently (Eliasziw *et al.* 1994: Tables 2 and 3).

The procedure assumes that the subject and rater effects and the interrater and intrarater random errors have a normal distribution.

J. COMPARISON OF SUBJECTS WITH VARYING NUMBERS OF MATCHED CONTROLS ("YES-NO" VARIABLE)

This module is appropriate for the analysis of case-control studies, clinical trials and cohort studies in which each index subject (each case, experimental subject, or individual exposed to a risk or protective factor) has a variable number (1-20) of individually matched controls, and the dependent variable is dichotomous ("yes-no"), e.g. "yes" = exposure to a risk factor (in a case-control study), the success of a treatment, or the presence of a disease (in a cohort study). It compares the findings in the index subjects and their matched controls.

The program refers to index subjects as "cases". Each set of matched observations can be entered in a separate line, or sets with the same findings can be entered together, with their frequency. The required entries for each pattern of findings are 0 ("no") or 1 ("yes") for the "case", the number of matched controls with "yes", and the number of matched controls with "no".

If the data are stratified, enter each stratum in turn. Click on "All strata" whenever combined results are required.

The program provides **tests (**Mantel-Haenszel test, Walter's test for binary data), the **odds ratio** (maximum-likelihood and Mantel-Haenszel estimates, with their confidence intervals, and a low-bias estimate), and **kappa**.

If *stratified data* are entered, an overall Mantel-Haenszel test is done, the results of the Walter's tests in the separate strata are combined, the *heterogeneity* of the P-values in the strata is tested, and an overall *kappa* is computed.

Tests

The program performs a Mantel-Haenszel test (without a continuity correction) and Walter's test for binary data (with and without a continuity correction).

If *stratified data* are entered, an overall Mantel-Haenszel test is done, and the Walter's tests in the separate strata (continuity-corrected) are combined by averaging their *z* values (Stouffer *et al*. 1949: 45; DeMets 1987) and computing an overall P that controls for the stratifying variables. P-values are computed in three ways, weighting the strata by different methods: weighting them equally, by sample sizes (the number of pairs), and by the square roots of the sample sizes. In addition, a test is done for the *heterogeneity* of the P-values in the strata (Wolf 1986: 45).

Odds ratio

Maximum-likelihood and *Mantel-Haenszel estimates* of the odds ratio and their 90%, 95%, and 99% confidence intervals are computed, and Jewell's low-bias estimator of the odds ratio (Jewell 1984) is shown.

Kappa

The program computes *kappa*, which expresses the agreement among all the observations in the matched sets, and may serve to express the effectiveness of the matching procedure, since it indicates the extent to which the findings in matched sets are more similar than findings in individuals from different sets. (Fleiss *et al*. 2003: 617-618).

The probability of chance agreement is taken into account in the calculation of *kappa*. A value of 1 indicates perfect agreement (after allowing for this probability of chance agreement) between ratings; 0 indicates no agreement other than what can be attributed to chance, and a negative value indicates less than chance agreement. Fleiss *et al*. (2003) suggest that a value of 0.75 or more indicates excellent agreement, and 0.40 or less indicates poor agreement. Cicchetti and Sparrow (1981) divide Fleiss's 0.40–0.74 group into 0.60–0.74: good; and 0.40–0.59: fair. Alternative guidelines are: over 0.80, very good agreement; 0.61–0.80, good; 0.41–0.60, moderate; 0.21–0.40, fair; and 0.20 or less, poor agreement (Landis and Koch 1977, Altman 1991).

METHODS

Tests

The Mantel-Haenszel test uses formula 13-18 of Rothman (1986). If stratified data are entered,

$$\text{chi-square} = (\sum \text{Num}_i)^2 / \sum(\text{Den}_i^2)$$

where Num_i = numerator of Rothman's formula in stratum *i*
Den_i = denominator of Rothman's formula in stratum *i*

The formula for *Walter's test* for binary data is formula 2 in Walter (1980); for a continuity-corrected test, 0.5 is subtracted from the absolute value of the numerator. If *stratified data* are entered, the Walter's tests in the separate strata (continuity-corrected) are combined by averaging their *z* values (Stouffer *et al*. 1949: 45; DeMets 1987). Three different sets of weights are used for this purpose – weighting the test results equally, by the sample sizes in the strata, and by the square roots of the sample sizes. In addition, a heterogeneity test is performed, comparing the P-values in the strata, using the formula (Wolf 1986: 45):

$$\text{chi-square } (k - 1 \text{ d.f.}) = \sum(Zi - \text{MeanZ})^2$$

where k = number of strata,
Z_i = *z* value in stratum *i*
MeanZ = mean *z* value.

Odds ratio

The computation of the maximum likelihood estimate and exact intervals uses an efficient algorithm for calculating the coefficients of the conditional distribution (Martin and Austin 1991, 1996), using code from David O. Martin's public-domain EXACTBB program.

The Manterl-Haenszel estimate of the odds ratio is computed by formula 13-9 of Rothman (1986), and its confidence intervals by the procedure described on page 274 of Rothman (1986).

Kappa

Kappa and its standard error are calculated by formulae 18.44 and 18.46 of Fleiss *et al*. (2003).

K. COMPUTE *KAPPA* FOR A VARIABLE NUMBER OF RATINGS

This module appraises the agreement between a variable number (3 or more) of matched observations with respect to a "yes"-"no" (dichotomous) variable. It might be used to measure the agreement between ratings of the same individuals, e.g. by different observers or tests, or between ratings of the same individuals made by the same observer on different occasions.

The findings in the set of ratings are then entered, by entering the numbers of "yes" ratings and "no" ratings. Each set of ratings can be entered separately, or sets with the same pattern of findings can be entered together, with their frequency.

The program provides the overall ***kappa***, and *kappa* values for individual categories.

If stratified data are entered, an overall value of *kappa* is computed.

Kappa

The overall *kappa* is computed, with its standard error and significance. *Kappa* values are also reported for individual categories, with their significance; but these test results should be treated with caution, since they are not based on a multiple-comparison procedure.

For *stratified data* are entered, an overall value of *kappa*, weighted by sample size, is computed.

The probability of chance agreement is taken into account in the calculation of *kappa*. A value of 1 indicates perfect agreement (after allowing for this probability of chance agreement) between ratings; 0 indicates no agreement other than what can be attributed to chance, and a negative value indicates less than chance agreement. Fleiss *et al*. (2003) suggest that a value of 0.75 or more indicates excellent agreement, and 0.40 or less indicates poor agreement. Cicchetti and Sparrow (1981) divide Fleiss's 0.40–0.74 group into 0.60–0.74: good; and 0.40–0.59: fair. Alternative guidelines are: over 0.80, very good agreement; 0.61–0.80, good; 0.41–0.60, moderate; 0.21–0.40, fair; and 0.20 or less, poor agreement (Landis and Koch 1977, Altman 1991).

METHODS

Kappa

Kappa and its standard error are calculated by formulae 18.44 to 18.46 of Fleiss *et al.* (2003).

L1. COMPARISON OF TWO GROUPS OF VARYING NUMBERS OF MATCHED NUMERICAL OBSERVATIONS

This module is appropriate for the analysis of case-control or cohort studies, trials, comparisons of methods of measurement, or other studies that compare two groups of matched numerical variables, where some or all of the matched sets have 3 or more observations, and the numbers of observations in the two groups (in each set) may vary. The program compares the two groups of observations.

The groups are arbitrarily referred to as "cases" and "controls". Optionally, a fixed number can be specified for the cases in each matched set. A matched set may contain 2-9 observations (1-8 cases and 1-8 controls). Each set must be entered in a separate line: first the case or cases, , then a slash (/), then the control or controls, then another slash. For example, the entry for a set containing 1 case and 3 controls might be:

16.23 / 9.8 11.06 15.11 /

Up to 500 sets may be entered.

The program provides three **tests** (Rosner's and Walter's tests and a paired t-test) for the **difference between the mean values** of cases and controls, approximate confidence intervals for this difference, **between-sets and within-sets variances,** and **Hodges-Lehmann estimate** of difference between medians.

Tests

Rosner's test is a generalization of the paired *t*-test that takes account of within-sets and between-sets variability (Rosner 1982). Iit appraises the significance of the differences between the mean values in the two groups. Two P-values may be displayed. If so, these may be regarded as the bounds of the true P-value. The true P-value depends on the relative magnitude of the within-sets and between-sets variabilities (see below), as explained by an on-screen message. The test sometimes presents technical difficulties, and is omitted.

Walter's test (Walter 1980) tests the significance of the mean case-control difference weighted by the numbers of cases and controls in the set. Rosner (1982) points out that (unlike his test) Walter's test assumes zero between-sets variability, and may therefore provide a misleadingly low P-value if there is much between-sets variability.

The *paired t-test* tests the significance of the unweighted mean difference between the case and control means within each matched set. Rosner (1982) points out that (unlike his test) the paired *t* test assumes zero within-sets variability, and may therefore provide a misleadingly low P-value if there is much within-sets variability.

Difference between the mean values

The program displays the mean case-control difference and its standard error, computed separately by the Rosner and Walter procedures and for unweighted data, with approximate 90%, 95%, and 99% confidence intervals.

Hodges-Lehmann procedure

This nonparametric procedure (Sprent 1993:89-90) determines the median of the differences between two matched sets, e.g matched cases and controls (with 90%, 95%, and 99% confidence intervals). [This is not necessarily the same as the difference between the medians, or the median of the differences observed in each matched set.]

A large-sample method of analysis is used if there are over 50 matched sets.

The analysis takes account of tied differences (if the large-sample method is used), but not of variation within matched sets.

Between-sets and within-sets variances

The between-sets variance represents the variation between matched sets, and the within-sets variance represents the variation within either the case or the control group for a specific matched set. The ratio of the two variances is an indication of the value of multiple matching. If the between-sets variance is much larger than the within-sets variance, multiple matching brings little benefit (Rosner 1982; Lee and Wilkens 1994).

METHODS

Tests

Rosner's test (Rosner 1982) is a generalization of the paired *t*-test that takes account of within-sets and between-sets variability. It adjusts and appraises the significance of the mean within-set difference. The test sometimes presents technical difficulties, since it requires the computation of maximum-likelihood estimates by an iterative procedure that may fail to find an appropriate (positive) root. If this difficulty is encountered (usually because of marked within-set variability) an appropriate message is displayed.

In Rosner's procedure the within-pairing variability is calculated by Rosner's formula 2.2 (Rosner 1982), and maximum likelihood estimates of the between-pairing variability and the adjusted mean case-control difference are then computed by an iterative process, using the van Wijnsgaarden-Dekker-Brent root-solver (Press *et al.* 1989: 283-286). The adjustment takes account of the numbers of cases and controls per set, using their reciprocals. Significance is appraised by Rosner's formula 2.3, using alternative degrees of freedom when referring the test statistic (*lambda*) to the *t*-distribution, namely $N - 2R$ and $R - 1$ (where N = number of subjects and R = number of matched sets). This provides two P-values (both of which are shown if they differ appreciably), which may be regarded as the bounds of the true P-value. The true value depends on the relative magnitude of the within-sets and between-sets variabilities.

Walter's test uses formula 2.4 of Rosner (1982). This permits application of the test to situations where there are matched sets with two or more cases.

The *paired t-test* is calculated by the usual formula (see, e.g. Selvin 1991: 65, formula 2.51), except that in each matched set the two values (of case and control) are replaced by the means (of cases, if there is more than one case, and of controls, if there is more than one control).

Difference between the mean values

In Rosner's procedure (see above), the adjusted mean case-control difference is computed by weighting the difference in each matched set by

$$1 / \{B + W.[(1 / N1) + (1 / N2)]\}$$

where B = between-sets variance
W = within-sets variance
N1 and N2 = numbers of cases and controls in the set.

In Walter's procedure, the difference in each matched set is weighted by $1 / [(1 / N1) + (1 / N2)]$

Hodges-Lehmann procedure

The differences between the values of cases and controls are calculated in the *n* matched sets. Where there are more than one case or control, their respective median values are used.

As described by Han (2008), each difference is then compared with each other difference, and for each of these $m = n(n - 1)/2$ comparisons of two values, the mean of the pair of differences (Walsh value) is computed. The *m* means are then ranked in ascending order, and their median is determined. This is the point estimate of the Hodges-Lehmann median difference between cases and controls.

If $n <= 50$, a value *R* corresponding to the value of *n* is obtained from Table A12 of Conover r(1999: p. 545), using the $W_{0.005}$, $W_{0.025}$, and $W_{0.05}$ column for the 90%, 95%, and 99% confidence intervals respectively. The lower confidence limit is the Walsh value whose rank is *R* in the series, and the upper confidence limit is the Walsh value whose rank is R from the upper end of the series,

If $n > 50$, confidence intervals are estimated by a large-sample approximation (1 and Wolfe 1999:132-133, using the formulae provided by Han (2008), but with a correction for tied ranks (Unistat Statistics Software). The lower confidence limit is the Walsh value whose rank is *R* in the series, where $R = .za./.b$ rounded up to the nearest integer), and the upper confidence limit is the Walsh value whose rank is *R* from the upper end of the series,

where $z = -1.645$, -1.96, or -2.5767 (for 90%, 95%, or 99% limits respectively)

$a = \sqrt{[n(n + 1)(2n + 1) / 24 - Tee / 48]}$

$b = n(n + 1) / 4$

n = number of matched sets

Tee = the sum of $(t_i^3 - t_i)$

t_i = the number of ties in each set of tied ranks

The correction for ties ($Tee / 48$) in calculating *a* is omitted if it reduces *a* to zero or a negative value.

Between-sets and within-sets variances

These variances are computed by Rosner's procedure (Rosner 1982).

L2. COMPARISON OF TWO METHODS OF MEASUREMENT, USING REPEATED NUMERICAL OBSERVATIONS

This module is appropriate in methodological studies that compare two methods of measuring a numerical variable by applying each method to each subject more than once.

The program provides the **mean difference** between measurements by the two methods, and gives two sets of results – one applicable to studies in which the true value does not change from one set of measurements to another, and one to studies in which the true value may vary. In each instance, the program computes the **standard deviation of the difference** between the two methods, the **95% limits of agreement** between these measurements, and **ANOVA** tables. The **relationship between the difference and the mean value** is appraised**.**

The number of pairs of measurements per subject can vary, but for each subject there must be the same number of measurements (at least two) by each method. The measurements of each subject are entered in a separate line. If the true value can change between pairs of measurements, the measurements of a subject by the two methods must be entered in the same sequence, i.e., the first measurement by each method must be entered first, the second must be entered second, and so on.

Mean difference and 95% limits of agreement

The mean difference is the weighted mean of the differences between the measurements by the two methods.

The *95% limits of agreement* (Bland and Altman 1999) answer the question, "given a measurement by one method, how far might this be from a measurement by the other method?" They demarcate the bounds of the range that, with a 95% probability, includes the difference between single measurements of the same subject by the two methods.

Use of the 95% limits of agreement assumes that the differences are reasonably constant throughout the range of measurement. To check this assumption, the program displays *Spearman's coefficient of correlation between the difference and the mean level.* The correlation coefficient may be expected to be zero if the mean difference does not change with increasing values. Even when one of the methods of measurement is a new one and the other is an accepted standard, it is preferable to examine the relationship between the difference and the mean value rather than the relationship between the difference and the standard measurement, which (as shown by Bland and Altman 1995b) may be misleading.

ANOVA tables

One-way ANOVA tables show the between-subjects and residual components of variance.

METHODS

The *standard deviations* and *95% limits of agreement* are computed by the methods explained by Bland and Altman (2007).

M. COMPARISON OF REPLICATE NUMERICAL MEASUREMENTS (VARYING NUMBERS)

This module appraises the agreement between matched numerical measurements, in a study where the numbers of matched measurements vary. It might be used to measure the agreement between replicate ratings of the same individuals by different observers or by the same observer on different occasions, in studies of interobserver or intraobserver reliability.

The measurements of each subject must be entered, in any order, on a separate line.

The program computes a **95% repeatability coefficient,** an **intraclass correlation coefficient** (with its 95% confidence interval) and **Spearman-Brown coefficients of reliability**, and estimates the number of replicates required to obtain a mean-rating ICC of 0.75 or 0.8.

95% repeatability coefficient

This coefficient expresses the expectation (with 95% confidence) for the maximum size of the absolute difference between a pair of observations, assuming that repeatability is similar at all magnitudes. Approximate 95% confidence intervals are estimated for the coefficient.

Intraclass correlation coefficient

The *intraclass correlation coefficient* (ICC), which is appropriate for interval-scale data with an assumed normal distribution, is a measure of agreement that expresses the correlation between measurements within individuals or sets of matched individuals. The program provides an estimate of the Shrout-Fleiss model 1,1 ICC (Shrout and Fleiss 1979), which is based on a "one-way random model"; the coefficient applies to the use of a single measurement. As a rule of thumb, it has been suggested that values above 0.75 indicate excellent, and values above 0.4 good, reliability (Shoukri and Pause1999: 27). Negative ICC values indicate that the within-subject variation is greater than the between-subject variation.

The program reports the effective average number of replicates, on which (if the numbers of replicates vary) the computations are based.

Spearman-Brown coefficients of reliability

Spearman-Brown coefficients of reliability provide estimates of the effect of using the means of replicated observations (Fleiss 1986: 14-15). They predict what the reliability would be if between 2 and 6 replications were averaged. The program also uses the formula in reverse, to estimate the number of replicates required to obtain a mean-rating ICC of 0.75 or 0.8.

METHODS

The computation of the *coefficient of repeatability* is explained by Bland and Altman (1999: 149). It is based on the within-sets variance, computed by formula 13.16 of Guilford and Fruchter (1986: 235). Approximate confidence intervals are obtained by substituting confidence limits for the within-sets variance, estimated by the method described by Zar (1998: formula 7.16), in the formula.

The formula for the *intraclass correlation coefficient* (Shrout-Fleiss ICC model 1,1, computed from a one-way random effects model ANOVA) is:

$$ICC = (MSB - MSW) / [MSB + (k - 1)MSW]$$

where MSB = between-subjects mean square
MSW = within-subjects mean square
k = effective average number of replicates per subject.

The *effective average number of replicates* is computed by formula 5 of Ebel (1951). This provides a value (introduced by Snedecor 1946: 234) that is close to the harmonic mean. The use of Ebel's procedure was suggested by Solomon's rating reliability calculator (Solomon 2004).

Formulae for confidence intervals for the ICC models are provided by McGraw and Wong (1996a and 1996b) in their Table 7, where this ICC is referred to as ICC(1). The number of ratings in the formulae, which as appropriate for studies with a fixed number of replicates, is replaced by the effective average number of replicates.

The *Spearman-Brown prediction formula* (Fleiss 1986: 14-15: formula 1.3) for reliability (*R)* is

$$R = Nr / [1 + (N - 1)r]$$

where N = number of replicates that are averaged
r = intraclass correlation coefficient

Fleiss's formula 1.31 is used to estimate the number of replicates required to obtain a reliability of 0.75 or 0.8:

$$N = P(1 - r) / [r(1 - P)]$$

where P = 0.75 or 0.8

Mis1. EFFECT OF MISCLASSIFICATION: COMPARISON OF CASES AND MATCHED CONTROLS

This module appraises the effect of misclassification (nondifferential or differential) on a comparison of cases and matched controls with respect to their exposure to a risk or protective factor. It demonstrates the effect of the sensitivity and specificity of the measure of exposure, by computing the "true" findings that would give rise to the observed findings..

The program requires entry of the observed frequencies in a paired-data 2x2 table, and estimates of the sensitivity and specificity (in cases and in controls) of the measure of exposure.

The program computes what the frequencies would be if there were no misclassification, i.e. the *"true" frequencies* that would have given rise to the observed finding, together with the *"true" odds ratio* based on the computed frequencies. Confidence intervals are displayed for the observed and "true" odds ratios.

The computed "true" results are not shown if they are unrealistic (if a "true" frequency is negative). A message is displayed saying that the observed frequencies are not compatible with the sensitivity and specificity values, and that if the entries are correct, the findings may represent sampling error.

METHODS

The program constructs a 4 x 4 matrix representing four equations that express the relationship between the observed and true (correctly classified) frequencies, and solves them by calculating the inverse of the matrix and postmultiplying this by a vector composed of the observed frequencies. The procedure, a generalization of Barron's procedure for nondifferential misclassification (Barron 1977), is described by Kleinbaum, Kupper and Morgenstern (1982: 228-236) and Greenland and Kleinbaum (1983). If the matrix is not invertible an error message is displayed..

Exact Fisher's 95% confidence intervals are computed for the odds ratios; the "true" ratio is based on the "true" frequencies, after rounding them off to the nearest integer. The intervals are computed by an algorithm described by Martin and Austin (1991) and using code from David O. Martin's public-domain EXACTBB program. Uncertainty of the sensitivities and specificities is not taken into consideration.

Mis2. EFFECT OF MISCLASSIFICATION: COMPARISON OF MATCHED EXPOSED AND UNEXPOSED SUBJECTS

This module appraises the effect of misclassification (nondifferential or differential) on a comparison of matched subjects exposed and unexposed to a risk or protective factor, where the dependent variable is a disease or some other outcome. It demonstrates the effect of the sensitivity and specificity of the measure of the outcome variable, bycomputing the "true" findings that would give rise to the observed findings..

The program requires entry of the observed frequencies in a paired-data 2x2 table, and estimates of the sensitivity and specificity (in the exposed and unexposed groups) of the measure of the outcome variable..

The program computes what the frequencies would be if there were no misclassification, i.e. the *"true" frequencies* that would have given rise to the observed finding, together with the *"true" odds ratio* based on the computed frequencies. Confidence intervals are displayed for the observed and "true" odds ratios.

The computed "true" results are not shown if they are unrealistic (if a "true" frequency is negative). A message is displayed saying that the observed frequencies are not compatible with the sensitivity and specificity values, and that if the entries are correct, the findings may represent sampling error.

METHODS

The program constructs a 4 x 4 matrix representing four equations that express the relationship between the observed and true (correctly classified) frequencies, and solves them by calculating the inverse of the matrix and postmultiplying this by a vector composed of the observed frequencies. The procedure, a generalization of Barron's procedure for nondifferential misclassification (Barron 1977), is described by Kleinbaum, Kupper and Morgenstern (1982: 228-236) and Greenland and Kleinbaum (1983). If the matrix is not invertible an error message is displayed..

Exact Fisher's 95% confidence intervals are computed for the odds ratios; the "true" ratio is based on the "true" frequencies, after rounding them off to the nearest integer. The intervals are computed by an algorithm described by Martin and Austin (1991) and using code from David O. Martin's public-domain EXACTBB program. Uncertainty of the sensitivities and specificities is not taken into consideration.

Mis3. EFFECT OF MISCLASSIFICATION: COMPARISON OF ANY TWO MATCHED GROUPS

This module appraises the effect of misclassification (nondifferential or differential) on a comparison of any two matched groups with respect to a dependent variable It demonstrates the effect of the sensitivity and specificity of the measure of the dependent variable, by computing the "true" findings that would give rise to the observed findings..

The program requires entry of the observed frequencies in a paired-data 2x2 table, and estimates of the sensitivity and specificity (in groups A and B) of the measure of the dependent variable.

The program computes what the frequencies would be if there were no misclassification, i.e. the *"true" frequencies* that would have given rise to the observed finding, together with the *"true" odds ratio* based on the computed frequencies. Confidence intervals are displayed for the observed and "true" odds ratios.

The computed "true" results are not shown if they are unrealistic (if a "true" frequency is negative). A message is displayed saying that the observed frequencies are not compatible with the sensitivity and specificity values, and that if the entries are correct, the findings may represent sampling error.

METHODS

The program constructs a 4 x 4 matrix representing four equations that express the relationship between the observed and true (correctly classified) frequencies, and solves them by calculating the inverse of the matrix and postmultiplying this by a vector composed of the observed frequencies. The procedure, a generalization of Barron's procedure for nondifferential misclassification (Barron 1977), is described by Kleinbaum, Kupper and Morgenstern (1982: 228-236) and Greenland and Kleinbaum (1983). If the matrix is not invertible an error message is displayed..

Exact Fisher's 95% confidence intervals are computed for the odds ratios; the "true" ratio is based on the "true" frequencies, after rounding them off to the nearest integer. The intervals are computed by an algorithm described by Martin and Austin (1991) and using code from David O. Martin's public-domain EXACTBB program. Uncertainty of the sensitivities and specificities is not taken into consideration.

P1. POWER OF TEST FOR DIFFERENCE BETWEEN PROPORTIONS (MATCHED PAIRS)

This module computes the power of a McNemar test for a difference between proportions observed in matched subjects, or in the same individuals (as in before-after studies, comparisons of diagnostic procedures, and crossover trials).

The program requires entry of the desired level of significance (for a one-sided or two-sided test), the sample size (the number of pairs of observations), the odds ratio to be detected, and either the expected number or the expected percentage of pairs with discrepant ("yes-no" and "no-yes") results.

Optionally, the percentage of expected losses of pairs in a projected study (nonresponses, dropouts, exclusions from the analysis, etc.) can be entered, and the sample size that is entered will be reduced accordingly before power is computed. This does of course not allow for possible bias. If the expected loss of observations is L%, the expected loss of pairs may be about $2L - [L^2 / 10000]$ %.

Results should be used with caution if samples are very small.

METHODS

Power is computed by the asymptotic unconditional method. The formula is an inversion of formula 3 of Julious *et al*. (1999), and is specified by Sahai & Kurshid (1996b: top of page 562). If an odds ratio under 1 is entered, the computation uses its reciprocal; for this purpose, an odds ratio of 0 is first converted to 0.000001.

If an expected loss rate is entered, the sample size is reduced before power is computed, and so is the expected number of discrepant pairs, if this number was entered.

P2. POWER OF TEST FOR COMPARING DISTRIBUTION OF ORDERED CATEGORIES (MATCHED PAIRS)

This module computes the power of a test (e.g., the Mann-Whitney test for paired data) for a difference between paired observations using an ordinal scale (such as "mild-moderate-severe"). The paired observations may relate to matched subjects, or to the same individuals (as in before-after studies, comparisons of diagnostic procedures, and crossover trials).

The program requires entry of the desired level of significance (for a one-sided or two-sided test), the sample size (the number of pairs of observations), and the odds ratio to be detected. The procedure assumes a proportional odds model; that is, the odds ratio is assumed to be the same, whatever cutting-point may be used when combining adjacent ordered categories to convert the frequency-distribution table into a 2x2 table.

The estimate of power is a conservative one (i.e., it underestimates power), especially if there are many categories.

Optionally, the percentage of expected losses in a projected study (nonresponses, dropouts, exclusions from the analysis, etc.) can be entered, and the sample size that is entered will be reduced accordingly before power is computed. This does of course not allow for possible bias. If the expected loss of observations is L%, the expected loss of pairs may be about $2L - [L^2 / 10000]$ %.

METHODS

The program uses an inversion of the simple "rule-of-thumb" formula recommended by Julious *et al*. (1999: formula 2) for estimating sample size for these tests.

If an expected loss rate is entered, the sample size is reduced before power is computed

P3. POWER OF TEST FOR DIFFERENCE BETWEEN MEANS (MATCHED PAIRS)

This module computes the power of a paired *t*-test for a difference between means observed in paired observations, in matched subjects, or in the same individuals (as in before-after studies, comparisons of diagnostic procedures, and crossover trials).

The program requires entry of the desired level of significance (for a one-sided or two-sided test), the sample size (the number of pairs of observations), and the difference to be detected (e.g. observation A minus observation B). In addition, the standard deviation of the differences between paired values is required. This can be entered, if its value is known or can be assumed. If not, there are two alternatives that permit computation of the standard deviation. These are: (a) entry of the within-subject mean square in an ANOVA (the residual within-subject mean square, after removal of the between-subjects component), if this is known (possibly from a published study; and (b) entry of the standard deviations of the two sets of observations, together with the correlation coefficient between the two sets (if a zero coefficient is entered, this will provide a conservative estimate of sample size).

Optionally, the percentage of expected losses in a projected study (nonresponses, dropouts, exclusions from the analysis, etc.) can be entered, and the sample size that is entered will be reduced accordingly before power is computed. This does of course not allow for possible bias. If the expected loss of observations is L%, the expected loss of pairs may be about $2L - [L^2 / 10000]$ %.

METHODS

The program uses an inversion of formula 1 of Julious *et al*. (1999).

If the standard deviation of the differences is not entered, it is computed either from the within-subject mean square, by multiplying its square root by √2 (Julious *et al*. 1999), or from the standard deviations of the two sets of observation (SD_a and SD_b) and the correlation coefficient *(r*), as

$$\sqrt{(SD_a^2 + SD_b^2 - 2rSD_aSD_b)}$$

(Sokal and Rohlf 1981: 573).

If an expected loss rate is entered, the sample size is reduced before power is computed

S1. SAMPLE SIZES: "YES-NO" DATA: DIFFERENCE (MCNEMAR TEST)

This module computes the sample size (the number of discrepant pairs of observations and the total number of pairs of observations) required for a McNemar test to detect a difference of a given magnitude between paired dichotomous ("yes"/"no") observations in matched subjects or in the same individuals (as in before-after studies, comparisons of diagnostic procedures, and crossover trials). It also computes the numbers of matched sets required for case-control studies with more than one matched control per case.

Three entry options are offered: (a) entry of the odds ratio to be detected and the expected percentage of discrepant ("yes-no" and "no-yes") pairs, or (b) entry of the odds ratio to be detected, the assumed value of the **matching factor** (see below) and the expected proportion of "yes" in the set of observations where that proportion is lower, or (c) the expected proportions of "yes" in both sets of matched observations. The first two options are preferable to the third. In addition, the required significance level and power must be entered.

If the expected proportions of "yes" in the two sets of observations are entered, the computation provides results based on the assumption that that the two sets are mutually independent. The required number of pairs that is reported is a maximal estimate, unless the matched observations are negatively correlated. The stronger the positive correlation, the more the overestimation, as demonstrated in Table 2 of Lehr (2001). If there is a negative correlation (that is, if a "yes" is likely to be associated with a "no" in the matched observation, as might occur in a paired before-after study where the first response influences the second, the computed sample sizes are underestimates. An additional "worst-case" maximal requirement is calculated, for use in such instances.

Optionally, the program will inflate sample sizes to compensate for the probability that not all the selected observations will be included in the analysis, e.g. because of failure to locate addresses, refusal to participate, or missing data. This requires entry of the expected percentage of pairs that will be lost. This inflation does of course NOT compensate for possible selection bias.. If the expected loss of observations is L%, the expected loss of pairs may be about $2L - [L^2 / 10000]$ %.

Matching factor

The matching factor is a measure of the degree to which the two sets of findings are similar because of matching. In a well-matched case-control study similarity may be expected between the exposure status of cases and their matched controls; and in a well-matched cohort study or trial, matched subjects may be expected to be similar with respect to prognostic factors affecting the outcome. The more similar the findings, the larger the sample sizes required.

The matching factor may be derived from the expected 2 x 2 table showing the paired results; it is the product of the two numbers of concordant pairs, divided by the product of the two numbers of discordant pairs. In a case-control study this is the *exposure odds ratio*,

measuring the unconditional association of the exposure status of a case with that of a matched control (Fleiss and Levin 1988, Lachin 1992).

The matching factor is 1 if the findings are independent, and is seldom much more than 2.5 (Fleiss and Levin 1988).

METHODS

If the odds ratio and expected percentage of discrepant pairs are entered, the required number of discrepant pairs is computed by formula 2 of Julious *et al*. (1999), and the required total number of pairs by formula 3 of Julious *et al*. (1999) (formula 5.4 of Sahai & Kurshid 1996b); this is an asymptotic unconditional method that has been shown to approximate satisfactorily to the results of computer simulations (Connett *et al*. 1987).

The same method is used if the expected proportions of “yes” in both sets of observations are entered, after estimating the numbers of pairs with discrepancies in each direction (*S* and *T*) from the proportions of “yes” (*P1* and *P2*), using formulae assuming an independent distribution (Royston 1993; Julious *et al*. 1999: 245):

$$S = [P1(1 - P2)] \text{ and}$$

$$T = [P2(1 - P1)]$$

and then estimating the odds ratio and proportion of discrepant pairs from *S* and *T*:

$$\text{Odds ratio} = S / T$$

$$\text{Proportion of discrepant pairs} = S + T$$

For the “worst-case” estimate,

$$S = \min(P1, 1 - P2)$$

$$T = P2 - P1 + S$$

If the matching factor is entered, sample sizes are computed by the multinomial unconditional procedure (Connor 1987; Lachin 1992: formulas 17 and 21), which is slightly conservative. If the calculated sample size is under 30, use is instead made of the local unconditional variance (Mitra 1958, Lachin 1992: formula 19), which is then more accurate. The estimated number of discordant pairs is also displayed.

Sample sizes for case-control studies with more than one matched control per case are calculated by formula 4 of Julious *et al*. (1999).

All sample sizes are rounded up to the next whole number.

If an expected non-inclusion rate (*R*%) is entered, the program multiplies computed sample sizes by

$$1 / [1 - (R / 100)]$$

before rounding them up

S2. SAMPLE SIZES: "YES-NO" DATA: AGREEMENT (*KAPPA*)

This module computes the sample size required in a study to determine *kappa* for two categories and two sets of observations.

The assumed value of kappa, the assumed proportion of "yes" findings (which is assumed to be similar in both sets of observations), and the required significance level must be entered. In addition, one of the following must be entered: (a) the required *power*; (b) the desired *width of the confidence interval* for *kappa* (if the significance level is set at 5%, this refers to the 95% confidence interval); or (c) the desired *lower confidence limit* for *kappa* (if the significance level is set at 5%, this refers to the lower 95% confidence limit). Kupper and Hafner (1989) have pointed out that sample size formulae based on confidence interval width may underestimate the sample size required to provide statistically credible results.

If power is entered, the program computes the sample sizes required to determine whether a *kappa* of the specified magnitude is significantly higher than 0.4 (taken to mean fair or good agreement) or 0.6 (taken to mean good agreement).

Optionally, the program will inflate sample sizes to compensate for the probability that not all the selected observations will be included in the analysis, e.g. because of failure to locate addresses, refusal to participate, or missing data. This requires entry of the expected percentage of pairs that will be lost. This inflation does of course NOT compensate for possible selection bias.. If the expected loss of observations is L%, the expected loss of pairs may be about $2L - [L^2 / 10000]$ %.

METHODS

If power (1 – *beta*) is entered, the program computes the sample size required to determine whether the lower [(1-*alpha*)*100]% confidence interval of the specified *kappa* exceeds 0.4 or 0.6. Computation is based on a non-centrality parameter that is derived from (1 - *beta*) and (2 x *alpha*), and entered in the sample size formula provided by Donner and Eliasziw (1992).

If the desired width of the confidence interval or the desired lower confidence level is entered, the program uses the procedure described by Donner (1999; formula 2.2).

All sample sizes are rounded up to the next whole number.

If an expected non-inclusion rate (R%) is entered, the program multiplies computed sample sizes by

$$1 / [1 - (R / 100)]$$

before rounding them up

S3. SAMPLE SIZES: "YES-NO" DATA: EQUIVALENCE TEST

This module computes the number of pairs required for a test of the equivalence of two sets of paired "yes-no" observations. This may be useful in the planning of equivalence tests in matched case-control studies, matched-control parallel trials, crossover trials, and comparisons of diagnostic or screening tests.

The program requires entry of the desired significance level and power, the magnitude of the difference (between the proportions of "yes") that is regarded as negligible, and the expected percentage of discrepant ("yes-no" and "no-yes") pairs.

Sample sizes are computed for an equivalence test based on the performance of two one-sided tests, and for a one-sided test (e.g. for non-inferiority of a new treatment or screening test in comparison with an established one).

Optionally, the program will inflate sample sizes to compensate for the probability that not all the selected observations will be included in the analysis, e.g. because of failure to locate addresses, refusal to participate, or missing data. This requires entry of the expected percentage of pairs that will be lost. This inflation does of course NOT compensate for possible selection bias.. If the expected loss of observations is L%, the expected loss of pairs may be about $2L - [L^2 / 10000]$ %.

METHODS

The program uses the procedures described by Liu *et al*. (2002) to compute the sample sizes required to test for equivalence, on the assumption that the observed proportions of "yes" in the two sets of observations are identical. Sample sizes are computed for sample-based tests, applying a continuity correction (which increases the required sample size) unless otherwise stated. The computation without a continuity correction uses formula 7 of Liu *et al*.; the computation with a continuity correction requires an iterative process to solve an equation (Liu *et al*. 2002: 239); the van Wijnsgaarden-Bekker-Brent root-solver (Press *et al*. 1989: 283-286) is used for this purpose. Sample sizes for a one-sided test (e.g. a non-inferiority test) are computed in a similar way, with appropriate changes of significance level and power (Liu *et al* 2002: 239). If any computed sample size is too small to ensure at least one discrepant pair in each direction (applying the expected proportion of discrepant pairs *PropDP*), it is raised to $1 / (PropDP / 2)$ to meet this condition.

All sample sizes are rounded up to the next whole number.

If an expected non-inclusion rate (R%) is entered, the program multiplies computed sample sizes by

$$1 / [1 - (R / 100)]$$

before rounding them up

S4. SAMPLE SIZES: ORDERED CATEGORIES: DIFFERENCE

This module computes the number of pairs of observations required for a test to detect a given difference between paired observations using ordered categories (such as "mild-moderate-severe"). The observations may relate to matched subjects, or to the same individuals (as in before-after studies, comparisons of diagnostic procedures, and crossover trials).

The procedure used is a simple "rule-of-thumb" one, and the estimate of sample sizes is a conservative one, especially if there are many categories.*

If the majority of observations are expected to be in a single extreme category (e.g. in the "well" category of a health scale), Julious *et al*. (1999) recommend calling this category "yes" and determining the sample size needed for "yes-no" data (module S1). If there are many categories, they suggest that the data be treated as normally distributed (module S5).

The odds ratio to be detected must be entered, together with the required significance level and power. The procedure assumes a proportional odds model; that is, the odds ratio is assumed to be the same, whatever cutting-point may be used when combining adjacent ordered categories to convert the frequency-distribution table into a paired-data 2x2 table.

Optionally, the program will inflate sample sizes to compensate for the probability that not all the selected observations will be included in the analysis, e.g. because of failure to locate addresses, refusal to participate, or missing data. This requires entry of the expected percentage of pairs that will be lost. This inflation does of course NOT compensate for possible selection bias.. If the expected loss of observations is L%, the expected loss of pairs may be about $2L - [L^2 / 10000]$ %.

* A more exact estimate can be obtained by a procedure provided by the PEPI program SAMPLES, which requires entry (in addition to the odds ratio) of the expected relative distribution of positive-discrepant pairs (pairs with discrepancies consistent in direction with the odds ratio) that have different degrees of discrepancy (Julious and Campbell 1998).

METHODS

The program uses formula 2 of Julious *et al.* (1999, Appendix). The procedure is a simple "rule-of-thumb" one that estimates the number of discordant pairs needed for a two-category situation and takes this as the total number of pairs required for a comparison of ordered categories.

All sample sizes are rounded up to the next whole number.

If an expected non-inclusion rate (R%) is entered, the program multiplies computed sample sizes by

$$1 / [1 - (R / 100)]$$

before rounding them up

S5. SAMPLE SIZES: NUMERICAL DATA: DIFFERENCE (PAIRED *T* TEST)

This module computes the number of pairs of observations required for a paired *t* test to detect a difference of a given magnitude between the means of observations in matched subjects or in the same individuals (such as observations in matched pairs, before-after observations in the same individuals. or cross-over trials) (as in before-after studies, comparisons of diagnostic procedures, and crossover trials).

The difference to be detected (e.g. mean A minus mean B), and the required significance level and power must be entered. The standard deviation of the differences between paired values is also required. This should be entered if its value is known or can be assumed. Alternatively, the program can compute the standard deviation. This requires entry of either (a) the within-subject mean square in an ANOVA (the residual within-subject mean square, after removal of the between-subjects component), if this is known (possibly from a published study); or (b) the known or assumed standard deviations of the two sets of observations, together with the known or assumed correlation coefficient between the two sets (if a zero coefficient is entered, this will provide a conservative estimate of sample size).

Optionally, the program will inflate sample sizes to compensate for the probability that not all the selected observations will be included in the analysis, e.g. because of failure to locate addresses, refusal to participate, or missing data. This requires entry of the expected percentage of pairs that will be lost. This inflation does of course NOT compensate for possible selection bias.. If the expected loss of observations is *L*%, the expected loss of pairs may be about $2L - [L^2 / 10000]$ %.

Note that for a trial comparing two independent groups, each of them having paired values for each individual (e.g. before and after treatment), module H2 of COMPARE2 should be used, entering the difference to be detected between paired observations, and the standard deviations or variance of the differences between paired observations (Lachin 1981).

METHODS

If the within-subject mean square is entered, its square root is multiplied by $\sqrt{2}$ to obtain the standard deviation (S.D.) of the differences (Julious *et al*. 1999). If the S.D.s of the two sets of observation (SD_a and SD_b) and the correlation coefficient (*r*) are entered, the S.D. of the differences is calculated (Sokal and Rohlf 1981: 573) as

$$\sqrt{[SD_a^2 + SD_b^2 - 2r(SD_a)(SD_b)]}$$

The required number of pairs is estimated (for a one-sided test) by formula 2.1 of Guenther (1981), and (for a two-sided test) by the same formula using *alpha / 2* instead of *alpha* (formula 1 of Julious *et al*. 1999).

All sample sizes are rounded up to the next whole number. If an expected non-inclusion rate (*R*%) is entered, the program multiplies computed sample sizes by

$$1 / [1 - (R / 100)]$$

before rounding them up.

S6. SAMPLE SIZES: NUMERICAL DATA: AGREEMENT (INTRACLASS CORRELATION COEFFICIENT)

This module computes the sample size required in a study to measure agreement by using an intraclass correlation coefficient (ICC). It may be appropriate in a reliability study in which there are a fixed number (two or more) observations of each subject, or in studies using cluster samples of a fixed size.

The required significance level, the number of observations per subject or set, and the expected ICC must first be entered. Then two options are offered: (a) entry of the required power and the value against which the expected ICC is to be tested; in a reliability study, the latter value is the lowest acceptable ICC; choices that have been suggested (Landis and Koch (1977) are 0.4 (moderate measurement reliability, 0.6 (substantial) or 0.8 (almost perfect); in other studies, it may be zero; and (b) entry of the desired width of the confidence interval for the ICC; Kupper and Hafner (1989) have pointed out that sample size formulae based on confidence interval width may underestimate the sample size required to provide statistically credible results..

If option (a) is selected, the program uses a simple approximation (Walter *et al*. 1998) whose results have excellent agreement with exact results. It provides the sample size required to test the null hypothesis that the ICC is equal to the value against it is to be tested, against the alternative that it is higher. The method is appropriate for studies in which the ICC can be estimated from an appropriate one-way ANOVA, e.g. those in which each subject is observed by different observers, by different methods, or at different times. Between-subjects and inter-subject variation are taken into account. Walter *et al*. suggest that the method may also be a practical compromise for studies in which a two-way analysis (e.g. taking account of variation between specific observers) would be appropriate.

If option (b) is chosen, the program uses an approximation that Bonett (2002) has developed and shown to be very accurate. This method is appropriate for studies in which the ICC can be estimated from a one-way or two-way ANOVA, e.g. those in which each subject is observed by different observers, by different methods, or at different times, in which between-subjects, inter-subject, and (if necessary) between- observers or between-methods variation must be taken into account

When planning a reliability study, it may be helpful to compare the sample sizes required for different numbers of observations per subject.

Optionally, the program will inflate sample sizes to compensate for the probability that not all the selected observations will be included in the analysis, e.g. because of failure to locate addresses, refusal to participate, or missing data. This requires entry of the expected percentage of pairs or sets that will be lost. This inflation does of course NOT compensate for possible selection bias.. If the expected loss of observations is L%, the expected loss of pairs may be about $2L - [L^2 / 10000]$ %, and the maximal loss of larger sets will be $3L$%.

METHODS

If the required power and the value against which the expected ICC is to be tested are entered, the computation uses formula 12 of Walter *et al*.(1998), with the recommended addition of 0.5 if the number of observations per subject/set is 2.

If the desired width of the confidence interval for the ICC is entered, the computation uses formula 3 of Bonett (2002), with the correction suggested if the number of observations per set is 2 and the expected ICC is 0.7 or more.

All sample sizes are rounded up to the next whole number. If an expected non-inclusion rate (*R*%) is entered, the program multiplies computed sample sizes by

$$1 / [1 - (R / 100)]$$

before rounding them up

S7. SAMPLE SIZES: NUMERICAL DATA: EQUIVALENCE TEST

This module computes the number of pairs required for a test of the equivalence of the means of two sets of paired numerical observations . This may be useful in the planning of equivalence tests in matched case-control studies, matched-control parallel trials, crossover trials, and comparisons of diagnostic or screening tests.

The program requires entry of the desired significance level and power, the magnitude of the difference (between means) that is regarded as negligible, the mean of the reference set of observations, the expected absolute difference between the means of the two sets (which must be less than the maximum difference regarded as negligible).

The standard deviation of the differences between paired values is also required. This should be entered if its value is known or can be assumed. Alternatively, the program can compute the standard deviation. This requires entry of either (a) the within-subject mean square in an ANOVA (the residual within-subject mean square, after removal of the between-subjects component), if this is known (possibly from a published study); or (b) the known or assumed standard deviations of the two sets of observations, together with the known or assumed correlation coefficient between the two sets (if a zero coefficient is entered, this will provide a conservative estimate of sample size). The standard deviation of the differences (entered or computed) must be less than the mean value in the reference set.

Either set of observations may be chosen as the reference set, but in a study comparing new and established treatments the established treatment is usually selected. In such studies, a recommended definition of a negligible difference is from 0 to 20% of the mean of the reference set. The mean value must be positive. The standard deviation of the differences (entered or computed) must be less than the mean value in the reference set.

Sample sizes are computed for an equivalence test based on the performance of two one-sided tests, and for a single one-sided test (e.g. for non-inferiority of a new treatment in comparison with an established one).

Optionally, the program will inflate sample sizes to compensate for the probability that not all the selected observations will be included in the analysis, e.g. because of failure to locate addresses, refusal to participate, or missing data. This requires entry of the expected percentage of pairs that will be lost. This inflation does of course NOT compensate for possible selection bias.. If the expected loss of observations is L%, the expected loss of pairs may be about $2L - [L^2 / 10000]$ %, and the maximal loss of larger sets will be $3L$%.

METHODS

The program uses the procedure described by Chow and Wang (2001) for a crossover design using raw data. Specifically, it uses the second set of equations designated as “B1” in Appendix B. The required number of pairs is computed by an iterative process, using the van Wijnsgaarden-Dekker-Brent root-solver (Press *et al.* 1989: 283-286). The value 0.2 in Chow and Wang's equations is replaced by D/M, where D is the value entered as the maximum bound of a negligible difference, and M is the mean of the reference set. The same equations

are used to estimate the number of pairs required for a one-sided test, with appropriate changes of significance level and power (Liu *et al*. 2002: 239).

If the standard deviation of the differences is not entered, it is computed either from the within-subject mean square, by multiplying its square root by √2 (Julious *et al*. 1999), or from the standard deviations of the two sets of observation (SD_a and SD_b) and the correlation coefficient *(r*), as

$$\sqrt{(SD_a^2 + SD_b^2 - 2rSD_aSD_b)}$$

(Sokal and Rohlf 1981: 573).

All sample sizes are rounded up to the next whole number. If an expected non-inclusion rate (*R*%) is entered, the program multiplies computed sample sizes by

$$1 / [1 - (R / 100)]$$

before rounding them up

REFERENCES

Abar B, Loken E (2010) Peirce's *I* and Cohen's κ for 2x2 measures of rater reliability. Journal of Probability and Statistics 2010: Article ID 48036

Abdi H (2007) The Bonferroni and Sidak corrections for multiple comparisons. In: Salkin N (ed.) *Encyclopedia of measurement and Statistics*. Thousand Oaks (CA): Sage.

Abdi H, Williams LJ (2010) Coefficients of correlation, alienation, and determination. In: *Encyclopedia of Research Design* vol. I (Salkind NJ, ed.), Sage Publications, 171-180).

Abramson JH (2004) WINPEPI (PEPI-for-Windows) computer programs for epidemiologists. Epidemiologic Perspectives & Innovations, 2004, 1:6 (available on the Internet at www.epi-perspectives.com/content/1/1/6).

Abramson JH (2011) WINPEPI updated: computer programs for epidemiologists, and their teaching potential. Epidemiologic Perspectives & Innovations 2011, 8:1 (available on the Internet at www.epi-perspectives.com/content/8/1/1).

Abramson JH, Gahlinger PM (2001*) Computer programs for epidemiologists: PEPI version 4*. Sagebrush Press: Salt Lake City, Utah]

Agresti A (1980) Generalized odds ratios for ordinal data. Biometrics 36: 69-67.

Agresti A (1984) *Analysis of ordinal categorical data.* New York: John Wiley & Sons.

Agresti A (1996) *An introduction to categorical data analysis*. New York: Wiley.

Agresti A (1990) *Categorical data analysis*. New York: Wiley.

Agresti A, Min Y (2005) Simple improved confidence intervals for comparing matched proportions. Statistics in Medicine 24: 729-740.

Ahn C, Odom-Maryon T (1995) Estimation of a common odds ratio under binary cluster sampling. Statistics in Medicine 14: 1567-1577.

Altman DG (1991) *Practical statistics for medical research*. London: Chapman and Hall.

Altman DG (1998) Confidence intervals for the number needed to treat. British Medical Journal 317: 1309-1312

Altman DG, Andersen PK (1999) Calculating the number needed to treat for trials where the outcome is time to an event. British Medical Journal 319: 1492-1495.

Altman DG, Machin D, Bryant TN, Gardner MJ, eds. (2000*) Statistics with confidence*, 2nd edn. BMJ Books.

Armitage P, Berry G, Matthews JNS (2002*) Statistical methods in medical research*, 4th edn. Oxford: Blackwell Science.

Armitage P, Hills M (1982) The two-period crossover trial. The Statistician 31: 119-131.

Baguley T (2012) Serious stats: a guide to advanced statistics for the behavioral science. Palgrave Macmillan. Online supplement 3: Replication probabilities and prep. Aaiable on te Internet at https://docs.google.com/viewer?url=http://www.palgrave.com/psychology/baguley/students/supplements/9780230_577183_03_sup03.pdf

Barlow W, Lai M-Y, Azen SP (1991) A comparison of methods for calculating a stratified kappa. Statistics in Medicine 10: 1465-1472.

Barnett AG, van der Pols JC, Dobson AJ (2005) Regression to the mean: what it is and how to deal with it. International Journal of Epidemiology 34: 215-220.

Barnhart HX, Kosinski AS, Haber MJ (2007) Assessing individual agreement. Journal of Biopharmaceutical Statistics 17: 697-719.

Bartko JJ (1994) Measures of agreement: a single procedure. Statistics in Medicine 13: 737-745.

Basu S, Basu A (1995) Comparison of several goodness-of-fit tests for the kappa statistic based on exact power and coverage probability. Statistics in Medicine 14: 347-356.

Bennett BM, Hsu P (1960) On the power function of the exact test for the 2x2 contingency table. Biometrika 47: 393-398.

Bennett EM, Alpert R, Goldstein AC (1954) Communications through limited response questioning. Public Opinion Quarterly 18: 303-308.

Berry KJ, Johnston JE , Mielke PW Jr (2008) Weighted kappa for multiple raters. Perceptual and Motor Skills 107: 837-848.

Bhapkar VP (1966) A note on the equivalence of two test criteria for hypotheses in categorical data. Journal of thE American Statistical Association 61: 228-235.

Bjerre LM, LeLorier J (2000) Expressing the magnitude of adverse effects in case-control studies: “the number of patients needed to be treated for one additional patient to be harmed”. British Medical Journal 320: 503-506.

Bland JM (2006) How should I calculate a within-subject coefficient of variation? Available on the Internet at http://www-users.york.ac.uk/~mb55/meas/cv.htm

Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical assessment. Lancet i: 307-310.

Bland JM, Altman DG (1995a) Comparing two methods of clinical measurement: a personal history. International Journal of Epidemiology 24 (supp. 1): S7-S14.

Bland JM, Altman DG (1995b) Comparing methods of measurement: why plotting difference against standard method is misleading. Lancet 346: 1085-1087.

Bland JM, Altman DG (1996a) The use of transformation when comparing two means. British Medical Journal 312: 1153.

Bland JM, Altman DG (1996b) Measurement error proportional to the mean. British Medical Journal 313: 106.

Bland JM, Altman DG (1999) Measuring agreement in method comparison studies. Statistical Methods in Medical Research 8: 135-160.

Bland JM, Altman DG (2007) Agreement between methods of measurement with multiple observations per individual. Journal of Biopharmaceutical Statistics 17: 571-582.

Bland JM, Butland BK (undated) Comparing proportions in overlapping samples. Available on the Internet at http://www-users.york.ac.uk/~mb55/overlap.pdf

Bloch DA, Kraemer HC (1989) 2 x 2 kappa coefficients: measures of agreement or association. Biometrics 45: 269-287.

Blood E, Spratt JF (2007) Disagreement on agreement: two alternative agreement coefficients. SAS Global Forum 2007.

Bonett DG (2002) Sample size requirements for estimating intraclass correlations with desired precision. Statistics in Medicine 21: 1331-1335.

Bonett DG, Price RM (2007) Statistical inference for generalized Yule coefficients in 2x2 contingency tables. Sociological Methods & Research 35: 429-446.

Bonett DG, Price RM (2012) Adjusted Wald confidence interval for a difference of binomial proportions based on paired data.Journal of Educational and Behavioral Statistics 37:479-488.

Bowker AH (1948) A test for symmetry in contingency tables. Journal of the American Statistical Association 43: 572-574.

Box GEP (1954) Some theorems on quadratic forms applied in he study of analysis of variance problems. 1I. Effect of inequality of variance in the one-way classification. The Annals of Mathematical Statistics 25: 290-302.

Bradley EL, Blackwood LG (1989) Comparing paired data: a simultaneous test of means and variances. The American Statistician 43: 234-235.

Brennan RL, Prediger DJ (1981) Coefficient Kappa: some uses. misuses, and alternatives. Educational and Psychological Measurement 41: 687-699.

Brenner H, Gefeller O (1994) Chance-corrected measures of the validity of a binary test. Journal of Clinical Epidemiology 47: 627-633.

Brenner H, Kliebsch U (1996) Dependence of weighted kappa coefficients on the number of categories. Epidemiology 7: 199-202.

Breslow NE, Day NE (1087) *Statistical methods in cancer research, Vol. II. The design and analysis of cohort studies.* Lyon: International Agency for Research on Cancer.

Bristol DR (1989) Sample sizes for constructing confidence intervals and testing hypotheses. Statistics in Medicine 6:803-811.

Burr EJ (1964) Small-sample distributions of the two-sample Cramer-von Mises' W-square and Watson's U-square. Annals of Mathematical Statistics 35: 1091-98.

Byrt T, Bishop J, Carlin JB (1993) Bias, prevalence and kappa. Journal of Clinical Epidemiology 46: 423-429.

Campbell MJ, Gardner MJ (2000) Medians and their differences. In: Altman DG, Machin D, Bryant TN, Gardner MJ, eds (2000*) Statistics with confidence*, 2nd edn. BMJ Books, pp 36-44.

Casagrande JT, Pike MC, Smith PG (1978a) The power function of the 'exact' test for comparing two binomial distributions. Applied Statistics 27:176-180.

Casagrande JT, Pike MC, Smith PG (1978b) Algorithm AS 129: The power function of the 'exact' test for comparing two binomial distributions. Applied Statistics 27:212-219.

Chatellier G, Zapletal E, Lemaitre D, Menard J, Degoulet P (1996) The number needed to treat: a clinically useful nomogram in its proper context. British Medical Journal 312:426-429.

Cheng RCH (1978) Generating beta variates with nonintegral shape parameters. Communications of the ACM 21: 317-322.

Chinn S (1990) The assessment of methods of measurement. Statistics in Medicine 9: 351-362.

Chinn S (1991) Repeatability and method comparison. Thorax 46: 454-456.

Chinn S, Heller RF (1981) Some further results concerning regression to the mean. American Journal of Epidemiology 114: 902-905.

Choi SC, Stablein DM (1982) Practical tests for comparing two proportions with incomplete data. Applied Statistics 31: 256–262.

Choi SC, Stablein DM (1988) Comparing incomplete paired binomial data under non-random mechanisms. Statistics in Medicine 7: 929-939.

Chow S-C, Wang H (2001) On sample size calculation in bioequivalence trials. Journal of Pharmacokinetics and Pharmacodynamics 28: 155-169.

Cicchetti DV, Allison T (1971) A new procedure for assessing reliability of scoring EEG sleep recordings. American Journal of EEG Technology 11: 101-109.

Cicchetti D, Bronen R, Spencer S, Haut S, Berg A, Oliver P, Tyrer P (2006) Rating scales, scales of measurement, issues of reliability: Resolving some critical issues for clinicians and researchers. Journal of Nervous and Mental Disease 194: 557-564.

Cicchetti DV, Feinstein AR (1990). High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 43: 551–558.

Cicchetti DV, Sparrow SS (1981) Developing criteria for establishing inter-rater reliability of specific items: Applications to assessment of adaptive behavior. American Journal of Mental Deficiency 86: 127-137.

Cleophas TJ, Zwinderman AH, Cleophas TF, Cleophas EP (2009) *Statistics applied to clinical trials*, 4th edn. Springer.

Connell FA, Koepsell TD (1985) Measures of gain in certainty from a diagnostic test. American Journal of Epidemiology 121: 744-753.

Connett JE, Smith JA, McHugh RB (1987) Sample size and power for pair-matched case-control studies. Statistics in Medicine 6: 53-59.

Connor RJ (1987) Sample size for testing differences in proportions for the paired-sample design. Biometrics 43: 207-211.

Cox DR(1989) Analysis of binary data, 2nd edn. New York: Chapman & Hall / CRC.

Conover WJ (1999) *Practical nonparametric statistics*, 3rd edn. New York: John Wiley & Sons.

Costa-Santos C, Antunes L, Souto A, Bernardes J (2010) Assessment of disagreement: a new information-based approach. Annals of Epidemiology 20: 555-561.

Costa-Santos C, Bernardes J, Ayres-de-Campos D, Costa A, Costa C (2011) The limits of agreement and the intraclass correlation coefficient may be inconsistent in the interpretation of agreement. Journal of Clinical Epidemiology 64: 264-269.

Cox DR, Oakes D (1984) *Analysis of survival data*. London: Chapman & Hall.

Cumming G (2005) Understanding the average probability of replication. Psychological Science 16: 1002-1004

Cumming G, Williams J, Fidler F (2004) Replication and researchers' understanding of confidence intervals and standard error bars. Understanding Statistics 3: 299-311.

D'Agostino RB (1986) Tests for the normal distribution. In: D'Agostino RB, Stephens MA (eds) *Goodness-of-fit techniques*. New York: Marcel Dekker, pp 367-419.

D'Agostino RB, Belanger A, D'Agostino RB Jr (1990) A suggestion for using powerful and informative tests of normality. The American Statistician 44: 316-321.

D'Agostino RB, Pearson ES (1973) Tests of departure from normality: empirical results for the distribution of b_2 and $\sqrt{b_1}$. Biometrika 60: 613-622.

Daly LE (1998) Confidence limits made easy: interval estimation using a substitution method. American Journal of Epidemiology 147: 783-790.

Daniel WW (1995) *Biostatistics: a foundation for analysis in the health sciences*, 6th edn. New York: John Wiley & Sons.

Darroch JN, McCloud P (1986) Category distinguishability and observer agreement. Australian Journal of Statistics 28: 371-388.

Davis CE (1976) The effect of regression to the mean in epidemiologic and clinical studies. American Journal of Epidemiology 104: 493-498.

DeMets DL (1987) Methods for combining randomized clinical trials: strengths and limitations. Statistics in Medicine 6: 341-348.

Dietz EJ (1989) Teaching regression in a nonparametric statistic course. The American Statistician 43: 35-40.

Digby PGN (1983) Approximating the tetrachoric correlation coefficient. Biometrics 39: 752-757.

Donald A, Donner A (1987) Adjustments to the Mantel-Haenszel chi-square statistic and odds ratio variance estimator when the data are clustered. Statistics in Medicine 6: 491-499.

Donner A (1984) Approaches to sample size estimation in the design of clinical trials - a review. Statistics in Medicine 3: 194-214.

Donner, Allan (1999) `Sample size requirements for interval estimation of the intraclass *kappa* statistic. Communications in Statistics, Part B -- Simulation and Computation 28: 415-429 .

Donner A, Eliasziw M (1992) A goodness-of-fit approach to inference procedures for the *kappa* statistic: confidence interval construction, significance testing and sample size determination. Statistics in Medicine 11: 1511-1519.

Donner A, Eliasziw M, Klar N (1994) A comparison of methods for testing homogeneity of proportions in teratologic studies. Statistics in Medicine 13: 1253-1264.

Donner A, Klar N (1996) The statistical analysis of *kappa* statistics in multiple samples. Journal of Clinical Epidemiology 49: 1053-1058.

Donner A, Zou GY (2010) Closed-form confidence intervals for functions of the normal mean and standard deviation. Statistical Methods in Medical Research 21: 347-359.

Dunnett CW (1964) New Tables for Multiple Comparisons with a Control. Biometrics 20: 482-491

Dunnigan K (2013) Tests of marginal homogeneity and special cases. Pharmaceutical Statistics 12: 213-216.

Durkalski VL, Palesch YY, Lipsitz SR, Rust PF (2003) Analysis of clustered matched-pair data. Statistics in Medicine 22: 2417-2428.

Ebel RL (1951). Estimation of the reliability of ratings. Biometrika 16: 407-424.

Edwards JH, Edwards AWF (1984) Approximating the tetrachoric correlation coefficient Biometrics 40: 563.

Eliasziw M, Donner A (1991) Application of the McNemar test to non-independent matched pair data. Statistics in Medicine 10: 1981-1991.

Eliasziw M, Young SL, Woodbury MG, Fryday-Field K (1994) Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example. Physical Therapy 74: 777-788.

Efron B (1981) Nonparametric standard errors and confidence intervals. Canadian Journal of Statistics 9: 139-158.

Efron B, Gong G (1987) A leisurely look at the bootstrap, the jackknife, and cross-validation. The American Statistician 37: 36-48.

Evans GT, Hoenig JM (1998) Testing and viewing symmetry in contingency tables, with application to readers of fish ages. Biometrics 54: 620-629.

Everitt BS (1977) *The analysis of contingency tables*. London: Chapman and Hall.

Fagerland MW, Lydersen S, Laake P (2013) The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. BMC Medical Research Methodology 2013: 1:391.

Fagerland MW, Lydersen S, Laake P (2014) Recommended tests and confidence intervals for paired binomial proportions. Statistics in Medicine 33: 2850–2875.

Farrington CP, Manning G (1990) Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. Statistics in Medicine 9: 1457-1494.

Feinstein AR (1995) Meta-analysis: statistical alchemy for the 21st century. Journal of Clinical Epidemiology 48: 71-79.

Feinstein AR , Cicchetti DV (1989) High agreement but low kappa: I. The problems of two paradoxes. Journal of Clinical Epidemiology 43: 543-549.

Ferguson GA (1966) *Statistical analysis in psychology and education.*, 6th edn . New York: McGraw-Hill.

Fieller EC, Hartley HO, Pearson ES (1957) Tests for rank correlation coefficients. I. Biometrika 44: 470-481.

Fieller EC, Hartley HO, Pearson ES (1961) Tests for rank correlation coefficients. II. Biometrika 48: 29-40.

Fleiss JL (1986) *The design and analysis of clinical experiments*. New York: John Wiley & Sons.

Fleiss JL, Cohen J (1973) The equivalence of weighted kappa and the intraclass correlation coefficient as measures of agreement. Educational and Psychological Measures 33: 613-619.

Fleiss JL, Cohen J, Everitt BS (1969) Large sample standard errors of kappa and weighted kappa. Psychological Bulletin 72: 323-327.

Fleiss JL, Levin B (1988) Sample size determination in studies with matched pairs. Journal of Clinical Epidemiology 41 :727-730.

Fleiss JL, Levin B, Paik MC (2003) *Statistical methods for rates and proportions*, 3rd edn. John Wiley & Sons.

Flight L, Julious SA (2015) The disagreeable behaviour of the kappa statistic. Pharmaceutical Statistics 14; 74-78.

Food and Drug Administration (FDA) (2001). *Guidance for industry: statistical approaches to establishing bioequivalence*, Food and Drug Administration, Center for Drug Evaluation and Research (CDER).

Freedman LS (1982) Tables of the number of patients required in clinical trials using the logrank test. Statistics in Medicine 1: 121-129.

Freeman PR (1989) The performance of the two-stage analysis of two-treatment, two-period crossover trials. Statistics in Medicine 8: 1421-1432.

Galbraith S, Daniel JA, Vissel B (2010) A study of clustered data and approaches to its analysis. Journal of Neuroscience 30: 10601-10608.

Gart JJ (1969) An exact test for congaing matched proportions in cross-over designs, Biometrika 56:75-80.

Gehan E (1965) A generalized Wilcoxon test for comparing arbitrarily single-censored samples. Biometrika 52: 203-223.

Geisser S, Greenhouse SW (1958) An extension of Box's results to the use of the *F* distribution in multivariate analysis. Annals of Mathematical Statistics 29: 885-891.

Giacoletti KED, Heyse J (2011) Using proportion of similar response to evaluate correlates of protection for vaccine efficacy. Statistical Methods in Medical Research: Available on the Internet: DOI 10.1177/71096228021/6299

Graham P, Bull B (1998) Approximate standard errors and confidence intervals for indices of positive and negative agreement. Journal of Clinical Epidemiology 51: 763-771.

GraphPad Statistics Guide (2013). Available on the Internet at www.graphpad.com/guides/prism/6/statistics/index.htm?stat_the_method_of_bonferroni.htm

Greenland S (1987) Variance estimators for attributable fraction estimates consistent in both large strata and sparse data. Statistics in Medicine 6: 701-708.

Greenland S (1994) Corrections. Statistics in Medicine 13: 99.

Greenland S (1999) Re: "Confidence limits made easy: interval estimation using a substitution method". American Journal of Epidemiology 149: 884.

Greenland S, Kleinbaum DG (1983) Correcting for misclassification in two-way tables and matched-pair studies. International Journal of Epidemiology 12:93-97.

Guenther WC (1981) Sample size formulas for normal theory *t* tests. The American Statistician 35: 243-244.

Gunther A, Hofler M (2006) Different results on tetrachorical correlations in Mplus and Stata – Stata announces modified procedure. International Journal of Methods in Psychiatric Research 15: 157-166.

Guilford JP, Fruchter B (1986) *Fundamental statistics in psychology and education*, 6th edn, Singapore: McGraw-Hill.

Guyatt GH, Juniper EF, Walter SD, Griffith LE, Goldstein RS (1998) Interpreting treatment effects in randomised trials. BMJ 316: 690-693.

Gwet K (2002a) Kappa statistic is not satisfactory for assessing the extent of agreement between raters. Statistical Methods for Inter-Rater Reliability Assessment No 1 (available on the Internet at http:\\www.agreestat.com/research_papers/kappa_statistic_is_not_satisfactory.pdf

Gwet K (2002b) Inter-rater reliability: dependency on trait prevalence and marginal homogeneity. Statistical Methods for Inter-Rater Reliability Assessment 2: 1–10. Available on the Internet at advancedanalyticsllc.com/irrhbk/research_papers/inter_rater_reliability_dependency.pdf

Gwet KL (2008) Computing inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical and Statistical Psychology 61: 29-48.

Gwet KL (2010) *Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among rater*s, 2nd edn. Advances Analytics. LLC, Gathersburg, MD.

Haber M, Barnhart HX, Song J, Gruden J (2005) Observer variability: a new approach in evaluating interobserver agreement. Journal of Data Science 3: 69-83.

Haber M, Barnhart HX (2008) A general approach to evaluating agreement between two observers or methods of measurement from quantitative data with replicated measurements. Statistical Methods in Medical Research 17: 151-169.

Haley DT, Thomas P, Petre M, Deocock AQ (2008) Using a new inter-rater reliability statistic. Technical report no 2008/15, Department of Computing, Faculty of Mathematics, Computing and Technology, the Open Univrersity. Abailable in the Internet at http://computing.open.ac.uk.

Halperin M, Gilbert PR, Lachin JM (1987) Distribution-free confidence intervals for Pr(X1<X2). Biometrics 43: 71-80.

Han L. (2008). SouthEast SAS Users Group. *Calculating the point estimate and confidence interval of Hodges-Lehmann's median using SAS software: SESUG 2008: The Proceedings of the SouthEast SAS Users Group, St Pete Beach, FL*, 2008.

Hayen A, Dennis RJ, Finch CF (2007) Determining the intra- and inter-observer reliability of screening tools used in sports injury research. Journal of Science and Medicine in Sport 10:201-210.

Hedges LV, Olkin I (1985*) Statistical methods for meta-analysis*. Orlando: Academic Press.

Henriquez T, Antunes L, Bernardes J, Matias M, Sato D, Costa-Santos C (2013) Information-based measure of disagreement for more than two observers: a useful tool toompare the degree of observer disagreement. BMC Medical Research Methodology 13: 47.

Higgins JPT, Green S (eds.) (2006*) Cochrane Handbook for Systematic Reviews of Interventions 4.2.6 [updated September 2006*]. In: The Cochrane Library, Issue 4, 2006. Chichester, UK: John Wiley & Sons.

Higgins JPT, Thompson SG (2002) Quantifying heterogeneity in a meta-analysis. Statistics in Medicine 21: 1539-1558.

Hildebrand DK, Laing JM, Rosenthal H (1977) *Analysis of Ordinal Data*. Beverly Hills: Sage Publications.

Hills M, Armitage P (1979) The two period cross-over clinical trial. British Journal of Clinical Pharmacology 8: 7-20.

Hirji KF, Fagerland MW (2011). Calculated unreported confidence intervals for paired data. BMC Medical Research Methodology 11: 66.

Hirji KF, Tang M-L, Vollset SE, Elashoff RM (1994) Efficient power computation for exact and mid-P tests for the common odds ratio in several 2 x 2 tables. Statistics in Medicine 13: 1539-1549.

Hoehler FK (2000) Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. Journal of Clinical Epidemiology 53: 499-503.

Hoenig JM, Morgan MJ, Brown CA (1995) Analyzing differences between two age determination methods by tests of symmetry. Canadian Journal of Fisheries and Aquatic Sciences 52: 364-368.

Hollander M, Wolfe DA (1999) *Nonparametric statistical methods,* 2nd edn. New York: John Wiley & Sons.

Hsieh FY, Bloch DA, Larsen MD (1998). A simple method of sample size calculation for linear and logistic regression. Statistics in Medicine 17: 1623-1634.

Iman RL, Davenport JM (1980) Approximations of the critical region of the Friedman statistic. Communications in Statistics - Theory and Methods A9: 571-595.

Iyengar S, Greenhouse JB (1988) Selection models and the file drawer problem. Statistical Science 3: 109-117.

Iverson GJ, Lee MD, Wagenmakers E-J (2009) Prep misestimates the probability of replication. Psychonomic Bulletin & Review 16: 424-429.

Jewell NP (1984) Small-sample bias of point estimators of the odds ratio from matched sets. Biometrics 40: 421-435.

Jolliffe IT, Stephenson DB, eds (2003) *Forecast verification: a practitioner's guide in atmospheric science*. John Wiley and Sons, Chichester.

Jones B, Kenward MG (2003) *Design and analysis of cross-over trials*, 2nd edn. Chapman and Hall (C.R.C).

Julious SA, Campbell MJ (1998) Sample size calculations for paired or matched ordinal data. Statistics in Medicine 17: 1635-1642.

Julious SA, Campbell MJ, Altman DG (1999) Estimating sample sizes for continuous, binary, and ordinal outcomes in paired comparisons: practical hints. Journal of Biopharmaceutical Statistics 9: 241-251.

Kahn HA, Sempos CT (1989) *Statistical methods in epidemiology*. New York: Oxford University Press.

Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 53: 457-481.

Kendall MG (1970) *Rank correlation methods*, 4th edn. London: Griffin.

Killeen PR (2005) An alternative to null-hypothesis significance tests. Psychological Science 16: 345-353.

Killeen PR (2010) Prep replicates: comment prompted by Iverson, Wagenmakers, and Lee (2010); Lecoutre, Lecoutre, and Poitevineau (2010); and Maraun and Gabriel (2010). Available on the Internet at: http://psycnet.apa.org/journals/met/15/2/203.pdf

Kim J S (1997) Determining sample size for testing equivalence. Medical Device and Diagnostic Industry Magazine. Available on the Internet at http://www.mddionline.com/article/determining-sample-size-testing-equivalence

Kleinbaum DG, Kupper LL, Morgenstern H (1982) *Epidemiological research: principles and quantitative method*s. New York: Van Nostrand Reinhold.

Kraemer HC (2006) Correlation coefficients in medical research: from product moment correlation to the odds ratio. Statistical Methods in Medical Research 15: 525-545.

Kraemer HC, Bloch DA (1988) Kappa coefficients in epidemiology: an appraisal of a reappraisal. Journal of Clinical Epidemiology 41:959-968.

Kraemer HC, Periyakoil VS, Noda A (2002) Tutorial in biostatistics: kappa coefficients in medical research. Statistics in Medicine 21: 2109-2129.

Kuan PF, Huang B (2013) A simple and robust method for partially matched samples using the p-values pooling approach. Statistics in Medicine (Wiley Online Library: DOI: 10.1002/sim.3758)

Kuritz SJ, Landis JR (1987) Attributable risk ratio estimation from matched-pairs case-control data. American Journal of Epidemiology 125: 324-328.

Kupper LL, Hafner KB (1989) How appropriate are popular sample size formulas? The American Statistian 43: 101-105.

Lachin JM (1981) Introduction to sample size determination and power analysis for clinical trials. Controlled Clinical Trials 2: 93-113.

Lachin JM (1992).Power and sample size evaluation for the McNemar test with application to matched case-control studies. Statistics in Medicine 11: 1239-1251.

Landis JR, Koch GG (1997) The measurement of observer agreement for categorical data. Biometrics 33: 159-174.

Lantz CA, Nebenzahl E (1996) Behavior and interpretation of the *kappa* statistic: resolution of the two paradoxes. Journal of Clinical Epidemiology 49: 431-434.

Lecoutre B. Killeen PR (2010) Replication is not coincidence; reply to Pierson, Lee, and Wagenmakers 2009 (2010) Psychonomic Bulletin & Refiew 17: 263-269.

Lecoutre B, Lecoutre M-P, Poitevineau J (2009) Killeen's probability of replication and predictive probabilities: how to comopare, use and interpret them. To apper in Psychological Method. Available on the Internet at http://halshs.archives-ouvertes.fr/docs/00/49/16/98/PDF/Lecoutre_et_al_-_Killeen_s_Probability_of_Replication.pdf

Lee J (1992) Evaluating agreement between two methods for measuring the same quantity: a response. Computers in Biology and Medicine 22: 369-371.

Lee Y, Wilkens L (1994) Comparing means based on generalized matched sampling. Psychiatry Research 54: 305-306.

Lee W-C (1999) Selecting diagnostic tests for ruling out or ruling in disease: the use of the Kullback-Leibler distance. International Journal of Epidemiology 28: 521-525.

Lehr RG (2001) Some practical considerations and a crude formula for estimating sample size for McNemar's test. Drug Information Journal 35 :1227-1233

Lilliefors HW (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. Journal of the American Statistical Association 62: 399-402.

Lin L I-K (1989) A concordance correlation coefficient to evaluate reproducibility. Biometrics 45: 255-268.

Lin L I-K (2000) A note on the concordance correlation coefficient. Biometrics 56: 324-325.

Lin L I-K, Chinchilli V (1997) Rejoinder to the letter to the editor from Atkinson and Nevill. Biometrics 53: 777-778.

Lin L, Hedayat AS, Wu W (2012*) Statistical tools for measuring agreement*. Springer.

Liu G (2000) Sample size for epidemiologic studies. In: Gail MH, Benichou J (eds*) Encyclopedia of epidemiologic method*s, Chichester: John Wiley and Sons, pp. 777-794.

Liu J-P, Hsueh H-M, Hsieh E, Chen JJ (2002) Tests for equivalence or non-inferiority in paired binary data. Statistics in Medicine 21: 231-245.

Locatelli I, Rousson V (2014) Using an intraclass odds-ratio as an alternative to kappa to assess the inter-rater reliability of binary measurements. Book of Abstracts COMPSTAT 2014 21st International Conference on Computational Statistics. Available on the Internet at http://www.compstat2014.org/auxil/Book-of-Abstracts-COMPSTAT2014.pdf

Locatelli I, Rousson V (2016) Asessing interrater agreement on binary ,easurementsvia intraclass odds ratio. Biometrical Journal, Early View. DOI: 10.1002/bimj.201500109

Lombard M, Snyder-Duch J, Bracken CC (2004) Practical resources for assessing and reporting intercoder reliability in content analysis research projects. Available on the Internet at http:\\www.slis.indiana.edu/faculty/hrosenba/www/Research/methods/lombard_reliability.pdf

Luedtke R, personal communication.

Lui K-J (1996) Notes in case-control studies with matched pairs under inverse sampling. Biometrical Journal (1996) 38: 681-693

Lui K-J (2001a) Interval estimation of the attributable risk in case-control studies with matched pairs. Journal of Epidemiology and Community Health 55: 885-890.

Lui K-J (2001b) Notes on testing equality in dichotomous data with matched pairs. Biometrical Journal 43: 313-321.

Lui K-J (2004*) Statistical evaluation of epidemiological risk*. Chichester: John Wiley & Sons.

Machin D, Gardner MJ (2000) Time to event studies. In: Altman DG, Machin D, Bryant TN, Gardner MJ, eds (2000*) Statistics with confidence*, 2nd edn. BMJ Books, pp 93-194.

MacLure M, Willett WC (1987) Misinterpretation and misuse of the *kappa* statistic. American Journal of Epidemiology 126: 161-169.

Mainland D (1963) *Elementary medical statistics*, 2nd edn. Philadelphia and London: W.B.Sanders Co

Mantel N (1963). Chi-square tests with one degree of freedom: extensions of the Mantel-Haenszel procedure. American Statistical Association Journal 58: 690-700.

Martin D, Austin H (1991) An efficient program for computing conditional maximum likelihood estimates and exact confidence limits for a common odds ratio. Epidemiology 2: 359-362.

Martin DO, Austin H (1996) Exact estimates for a rate ratio. Epidemiology 7: 29-33.

Martin, AA, Femia MP (2004). Delta: a new measure of agreement between two raters. British Journal of Mathematical and Statistical Psychology 57: 1–19.

Martin AA, Femia MP (2008) Chance corrected measures of reliability and validity in 2 x 2 tables. Communications in Statistics – Theory and Methods 37: 760-772

Maxwell AE (1970) Comparing the classification of subjects by two independent judges. British Journal of Psychiatry 116: 651-655.

Maxwell WE (1977) Coefficients of agreement between observers and their interpretation. British Journal of Psychiatry 130: 79-83.

May WL, Johnson WD (1997) The validity and power of tests for equality of two correlated proportions. Statistics in Medicine 16: 1081-1096.

McCarthy WF (2007) Adjustment to the McNemar's test for the analysis of clustered matched-pair data. Cobra preprint series, paper 29 (available on the Internet at http://biostats.bepress.com/cgi/viewcontent.cgi?article=1056&context=cobra).

McGraw KO, Wong SP (1996a) Forming inferences about some intraclass correlation coefficients. Psychological Methods 1: 30-46.

McGraw KO, Wong SP (1996b) Correction to McGraw and Wong (1996). Psychological Methods 1: 390.

Mee RW, Chua TC (1991) Regression toward the mean and the paired sample *t* test. The American Statistician 45: 39-42.

Mehta C, Patel N (1991*) StatXact Statistical software for exact nonparametric inference: User manual version 2*. Cambridge MA: Cytel Software Corporation.

Miller KM, Looney SW (2012) A simple method for estimating the odds ratio in matched case-control studies with incomplete paired data. Statistics in Medicine 31: 3299-3312.

Mitra SK (1958) On the limiting power function of the frequency chi-square test. Annals of Mathematical Statistics 29: 1221-1233.

Mizuno S, Yamaguchi T, Fukushima A, Matsuyama Y, Ohashi Y (2005) Overlap coefficient for assessing the similarity of pharmacokinetic data between ethnically different populations. Clinical Trials 2: 174-181.

Morris JA, Gardner MJ (2000) **Epidemiologica**l studies. In: Altman DG, Machin D, Bryant TN, Gardner MJ. *Statistics with confidence*, 2nd edn. BMJ Books.

Morrison AS (1979) Sequential pathogenic components rates. American Journal of Epidemiology 108: 709-718.

Morton V, Torgerson DJ (2005) Regression to the mean: treatment effect without the intervention. Journal of Evaluation in Clinical Practice 11: 59-65.

Mueller R, Buttner P (1994) A critical discussion of intraclass correlation coefficients. Statistics in Medicine 13:2465-2476.

Nagelkerke NJD, Hart AAM, Oosting J (1986) The two period binary response cross-over trial. Biometrics 7: 863-869.

Nam J-M (1997) Establishing equivalence of two treatments and sample size requirements in matched-pair design. Biometrics 53: 1422-1430.

Newcombe RG (1998a) Interval estimation for the difference between independent proportions: comparison of eleven methods. Statistics in Medicine 17: 873-890.

Newcombe RG (1998b) Improved confidence intervals for the difference between binomial proportions based on paired data. Statistics in Medicine 17: 2635-2650.

Newcombe RG. Altman DG (2000) Proportions and their differences. In: Altman DG, Machin D, Bryant TN, Gardner MJ. *Statistics with confidence*, 2nd edn. BMJ Books, pp. 45-56.

NIWA (National Institute of Water and Atmospheric Research) (2009) Lin's concordance. Available on the Internet at http://www.niwa.co.nz/our-services/online-services/statistical-calculators/lins-concordance

O'Brien PC, Fleming TR (1987) A paired Prentice-Wilcoxon test for censored paired data. Biometrics 43: 169-180.

Obuchowski NA (1998) On the comparison of correlated proportions for clustered data. Statistics in Medicine 17: 1495-1507.

Oden NJ (1990) Estimating kappa from binocular data. Statistics in Medicine 10: 1303-1311.

Orwin R (1983) A fail-safe N for effect size in meta-analyses. Journal of Educational Statistics 8: 157-159.

Ostermann T, Willich SN, Luedtke R (2008) Regression toward the mean – a detection method for unknown population mean based on Mee and Chua's algorithm. BMC Medical Research Methodology 8: 52.

Overall JE (1990) Comment. Statistics in Medicine 9:379-82.

Page, E B (1963). Ordered hypotheses for multiple treatments: A significance test for linear ranks. Journal of the American Statistical Association 58: 216–30.

Pan Y, Gao J, Haber M, Barnhart X (2010) Estimation of coefficients of individual agreement (CIAs) for quantitative and binary data using SAS and R. Computer Methods and Programs in Biomedicine 98: 214-219.

Pan V, Haber M, Barnhart HX (2011a) A new permutation-based method for assessing agreement between two observers making replicated binary readings. Statistics in Medicine 30: 839-853.

Pan V, Haber M, Gao J, Barnhart HX (2011b) A new permutation-based method for assessing agreement between two observers making replicated quantitative readings. Statistics in Medicine: published on-line in Wiley Online Library DOL 10.1002/sim.5323.

Parzen M, Lipsitz S, Metters R, Fitzmaurice G (2010) Correlation when data are missing. Journal of the Operational Research Society 61: 1049-1056.

Pike MC (1972) Contribution to the discussion on the paper by Peto R and Peto J: Asymptotically efficient rank invariant test procedures. Journal of the Royal Statistical Society, Series A, 135:201-203.

Peat JK, Unger WR, Combe D (1994) Measuring changes in logarithmic data, with special reference to bronchial responsiveness. Journal of Clinical Epidemiology 47: 1099-1108.

Peirce CS (1884) The numerical measure of the success of predictions. Science 4: 453-454.

Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N. McPherson K, Peto J, Smith PG (1977) Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. British Journal of Cancer 35: 1-39.

Pike MC, Casagrande J, Smith PG (1975) Statistical analysis of individually matched case-control studies in epidemiology: factor under study a discrete variable taking multiple values. British Journal of Preventive and Social Medicine 29: 196-201.

Pitman, E. J. G. (1939). A note on normal correlation. Biometrika, 31, 9-12.

Prentice RL (1978) Linear rank tests with right censored data/. Biometrika 65: 167-179.

Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1989*) Numerical recipes in Pascal: the art of scientific computing*. Cambridge: Cambridge University Press.

Quade D (1979), Using weighted rankings in the analysis of complete blocks with additive block effects. Journal of the American Statistical Association, 74: 680–683.

Ramasundarahettige CF, Donner A, Zou GY (2009) Confidence interval construction for a difference between two dependent intraclass correlation coefficients. Statistics in Medicine 28: 1041-1053.

Riffenburgh RH, Johnstone PA (2009) Measuring agreement about ranked decision choices for a single subject. International Journal of Biostatistics 5(1): article 17.

Rockhill B, Newman B, Weinbert C (1998) Use and misuse of population attributable fractions. American Journal of Public Health 88: 15-19.

Roebruck P, Kuhn A (1995) Comparison of tests and sample size formulae for proving therapeutic equivalence based on the difference of binomial probabilities. Statistics in Medicine 14: 1583-1594.

Rom DM, Hwang E (1996) Testing for individual and population equivalence based on the proportion of similar responses. Statistics in Medicine 15: 1489-1505.

Rosner B (1982) A generalization of the paired *t*-test. Applied Statistics 31: 9-13.

Rothman KJ (1986) *Modern epidemiology*. Boston: Little, Brown & Co.

Rothman KJ, Greenland S (1998) *Modern epidemiology*, 2nd edn. Philadelphia: Lippincott-Raven.

Royston JP (1983) A simple method for evaluating the Shapiro-Francia W' test for non-normality, Statistician 32: 297.

Royston P (1993) A pocket-calculator algorithm for the Shapiro-Francia test for non-normality: an application to medicine. Statistics in Medicine 12: 181-184.

Sackett DL, Richardson WS, Rosenberg W, Haynes RB (1997) *Evidence-based medicine: how to practice and teach EBM*. New York: Churchill Livingstone.

Sahai H, Khurshid A (1996a) Formulae and tables for the determination of sample sizes and power in clinical trials for testing differences in proportions for the two-sample design: a review. Statistics in Medicine 15: 1-21.

Sahai H, Khurshid A (1996b) Formulae and tables for the determination of sample sizes and power for testing differences in proportions for the matched-pair design: a review. Fundamental and Clinical Pharmacology 20: 554-563.

Salmi LR (1986) Re: Measures of gain in certainty from a diagnostic test (letter). American Journal of Epidemiology 123: 1121-1122.

Samsa GP (1996) Sampling distributions of p_pos and p_neg. Journal of Clinical Epidemiology 49 :917-919.

Sanabria P, Killeen PR (2007) Better statistics for better decisions: rejecting null hypotheses statistical tests in favor of replication statistics. Psychology in the Schools 44: 471-481.

Satten GA, Kupper LL (1990) Sample size requirements for interval estimation of the odds ratio. American Journal of Epidemiology 131:177-184.

Schoenfeld DA (1983) Sample size formula for the proportional-hazards regression model. Biometrics 39: 499-503.

Schouten HJA (1993) Estimating kappa from binocular data and comparing marginal probabilities. Statistics in Medicine 12: 2207-2217.

Schouten H, Kester A (2010) A simple analysis of a simple crossover trial with a dichotomous outcome measure. Statistics in Medicine 29: 193-198.

Schuirmann DJ (1987) A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. Journal of Pharmacokinetics and Biopharmaceutics 14: 657-580).

Scott WA (1955) Reliability of content analysis: The case of nominal scale coding. Public Opinion Quarterly 19:321-325.

Selvin S (1996) *Statistical analysis of epidemiologic data*, 2nd edn. New York: Oxford University Press.

Sen PK (1968) Estimates of the regression coefficient based on Kendall's tau. Journal of the American Statistical Association 63: 1379-1389.

Senn S (2002*) Cross-over trials in clinical research,* 2nd edn. John Wiley and Sons.

Sheskin DJ (2007) *Handbook of parametric and nonparametric statistical procedures*, 4th edn. Chapman & Hall/CRC.

Shiue W-K, Bain, LJ (1982) Experiment size and power comparisons for two-sample Poisson tests. Applied Statistics 31: 130-134.

Shoukri MM (2000) Agreement, measurement of. In: Gail MH, Benichou J (eds.) *Encyclopedia of epidemiologic methods*, pp 35-49. John Wiley and Sons.

Shoukri MM, Donner A (2001) Efficiency considerations in the analysis of inter-observer agreement. Biostatistics 2: 323-336.

Shoukri MM, Pause CA (1999) *Statistical methods for health sciences*, 2nd edn. Boca Raton: CRC Press.

Shrout PE, Fleiss JL (1979). Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 86: 420-428

Siegel S, Castellan NJ Jr (1988) *Nonparametric statistics for the behavioral sciences*, 2nd edn. New York: McGraw-Hill.

Sim J, Wright CC (2005) The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Physical Therapy 85: 257-268.

Simon R (1986) Confidence intervals for reporting results of clinical trials. Annals of Internal Medicine 105: 429-435.

Sinclair JC, Bracken MB (1994) Clinically useful measures of effect in binary analyses of randomized trials. Journal of Clinical Epidemiology 47: 881-889.

Smeeth L, Haines A, Ebrahim S (1999) Numbers needed to treat derived from meta-analyses - sometimes informative, usually misleading. British Medical Journal 318: 1548-51.

Snedecor GW (1946*) Statistical methods*, 4th edn. Ames, Iowa: Iowa State College Press.

Snedecor GW, Cochran WG (1980) *Statistical methods*, 7th edn. Iowa State University Press, Ames, Iowa.

Sokal RF, Rohlf FJ (1981) *Biometry*, 2nd edn. New York: W.H. Freeman.

Solomon DJ (2004). The rating reliability calculator. BMC Medical Research Methodology 4: 11.

Sprent P (1993) *Applied nonparametric statistical methods*, 2nd edn. London: Chapman & Hall.

Stegman J, Lucking A (2005).*Assessing reliability on annotations (1): theoretical considerations.* Collaborative Research Centre, Bieleveld University.

Steinley D, Wood P (2000). ICC.sas – program to calculate intra-class correlations and confidence intervals.. Internet document: http://www.missouri.edu/~marc/icc2.sas

Stine RA, Heyse JF (2001) Non-parametric estimates of overlap. Statistics in Medicine 20; 215-236.Goodman SN, Berlin JA (1994) The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Annals of Internal Medicine 121:200-206, appendix.

St Laurent RT (1998) Evaluating agreement with a gold standard in method comparison studies. Biometrics 54: 537-545.

Stouffer SA, Suchman EA, De Vinney LC, Star SA, Williams RM Jr (1949*) The American soldier: adjustment during army life, vol. 1*. New Jersey: Princeton University Press.

Stuart A (1955) A test for homogeneity of the marginal distributions in a two-way classification. Biometrika 42: 412-416.

Tang M-L, Ling M-H, Tian G-L (2009) Exact and approximate unconditional confidence intervals for proportion difference in the presence of incomplete data. Statistics in Medicine 28: 625-641 .

Tate MW, Brown SM (1970) Note on the Cochran Q test. Journal of the American Statistical Association 65:155-160.

Theil H (1950) A rank-invariant method of linear and polynomial regression analysis. III. Koninklijke Nederlandse Akademie Van Wetenschappen, Proceedings, Series A, 53: 1397-1412.

Thode HC Jr (1997) Power and sample size requirements for tests of differences between two Poisson rates. The Statistician 46: 227-230.

Thompson WF, Walter SD (1988a) A reappraisal of the kappa coefficient. Journal of Clinical Epidemiology 41 :949-958.

Thompson WF, Walter SD (1988b) *Kappa* and the concept of independent errors. Journal of Clinical Epidemiology 41 :949-958.

Trafimow D (2015) The attenuation of correlation coefficients: a statistical literacy issue. Teaching Statistics, early view (DOI: 10.1111/test.12087)

Uebersax J S (2000) The tetrachoric and polychoric correlation coefficients. Available on the Internet at www.john-uebersax.com/stat/mcnemar.htm

Uebersax J (2006) McNemar tests of marginal homogeneity. Available on the Internet at http://www.john-uebersax.com/stat/mcnemar.htm

Ulrich R, Wirtz, M (2004) On the correlation of a naturally and an artificially dichotomized variable. British Journal of Mathematical and Statistical Psychology 57: 235-251.

UniStat Statistical Software.: 6.4.2. Paired Samples. Available on the Intedrnet at http://www.unistat.com/guide/nonparametric-tests-paired-samples/

UniStat Statistical Software. Paired Sample in Excel with U NISTATs. Available on the Intedrnet at http://www.unistat.com/642/

Vanbelle S (2009) *Agreement between raters and groups of raters*. Ph.D. dissertation . Departement de Mathematique, Universite de Liege.

Vickers AJ, Altman DG (2001) Analysing controlled trials with baseline and follow up measurements. British Medical Journal 323: 1123-1124.

Walter SD (1980) Matched case-control studies with a variable number of controls per case. Applied Statistics 28: 172-179.

Walter SD (2001) Number needed to treat (NNT): estimation of a measure of clinical benefit. Statistics in Medicine 20: 3947-3962.

Walter SD, Eliasziw M, Donner A (1998) Sample size and optimal designs for reliability studies. Statistics in Medicine 17: 101-110.

Whitehead J (1993) Sample size calculations for ordered categorical data. Statistics in Medicine 12: 2257-2271.

Wichman BA, Hill ID (1985) Algorithm AS183. An efficient and portable pseudo-random number generator. In: *Applied Statistics Algorithms* (ed. P Griffiths, ID Hill). London: Ellis-Horwood Ltd for the Royal Statistical Society.

Wikipedia, the free encyclopedia. Available on the Internet at http://en.wikipedia.org/wiki/Spearman-Brown_prediction_formula

Wilson EB (1927). Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association 22: 209-212.

Wolf FM (1986) *Meta-analysis: quantitative methods for research synthesis.* Beverly Hills: Sage Publications.

Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL (2013) A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. BMC Medical Research Methodology 13:6.

Woodward M (1999) Epidemiology: study design and data analysis, 2nd edn. CRC Press.

Woolson RF, Bean JA, Rojas PB (1986) Sample size for case-control studies using Cochran's statistic. Biometrics 42: 927-932.

Woolson RF, O'Gorman TW (1992) A comparison of several tests for censored paired data. Statistics in Medicine 11: 193-208.

Yanagawa T, Tango T, Hiejima Y (1994) Mantel-Haenszel-type tests for testing equivalence or more than equivalence in comparative clinical trials. Biometrics 50: 859-864; erratum note in Biometrics 1995 51:392

Yi Q, Wang PP, He Y (2009) Reliability analysis for continuous measurements: equivalence test for agreement. Statistics in Medicine 27: 2816-2825.

Yudkin PL, Stratton IM (1996) How to deal with regression to the mean in intervention studies. Lancet 347: 241-243.

Zar JH (1998) Biostatistical analysis, 4th edn. Prentice Hall.