# Regression Accuracy and Power Simulations

## Version 7.0

(for Windows 8.2 and later)

January 2024

W.A. Sadler
71B Middleton Road
Christchurch 8041
New Zealand

Ph:   +64 3 343 3808
e-mail:  bill.sadler@xtra.co.nz

(formerly at Nuclear Medicine Department, Christchurch Hospital)

# Contents                                      Page

## Regression Accuracy and Power Simulations

This program (RAPS.exe, version 7.0) assesses the accuracy and statistical power of parametric linear regression using sets of *simulated* X, Y pairs. The program is designed to generate such data but users also have the option to import simulated data generated by an external process. The program incorporates ordinary (unweighted) least squares regression (OLS), iteratively reweighted least squares regression (WLS) and iteratively reweighted Deming (*1*) regression (errors in X and Y; also known as errors-in-variables regression). Significance testing is based on the usual confidence intervals for slope and intercept, but also uses the elliptically shaped joint slope, intercept confidence region. A major program aim is to contrast these two approaches; in particular the improved statistical power that can be realised in most situations by using the joint slope, intercept confidence region rather than confidence intervals.

This document gives a broad description of the program. The data simulation and numerical methods are fully described in Ref. *2*. The fine detail of using the program is given in program on-line help. A brief popup description of all controls on all dialogs can be obtained by pressing F1 when the control has focus, or by clicking the right mouse button over the control (provided it is enabled), or by using the **?** icon at the top right of the dialog. All dialogs have a Help button which provides detailed information about the dialog.

## Data

It must be emphasized that meaningful use of RAPS.exe requires computer simulated data. The program is designed to produce such data. The aim is to investigate accuracy and statistical power by randomly drawing many sets of X, Y pairs, each of a specified size (N), which have a *precisely known* underlying relationship (eg. Y = X) and *precisely known* error characteristics. In the program Accuracy dialog we are attempting to verify that the mean of the sampling slopes and the mean of the sampling intercepts equate to the values that were used to generate the data. The program performs a significance test on the difference between true and recovered values. Equally importantly, we are also seeking to verify that the expected proportion of slope and intercept confidence intervals and joint slope, intercept confidence regions enclose the true underlying slope and intercept values. A set of real method comparison X, Y pairs (even with N as large as several hundred) is unsuitable for this type of sampling experiment because the true underlying slope and intercept are unknown. Such data do of course provide an *estimate* of slope and intercept, but these values are not necessarily the true underlying values.

In practice, investigators will invariably have a reasonably precise knowledge of the X and Y ranges of values that will occur in a method or reagent lot comparison. They can also control how specimen results are distributed over those ranges. Common patterns are roughly uniform, concentrated in the central part of the range and concentrated at the lower end of the range (right skewed). Knowledge of the error (precision) characteristics of the X and Y results should be readily available (at least in a medical laboratory context) from analysis of internal QC data or precision data from a method evaluation (Ref. *3* provides one way of summarising replicated precision data). Use the program Accuracy dialog to generate simulated data that mimics as closely as possible the data that will be collected in a real life comparison, by specifying the likely range, distribution and error characteristics of the data. Assuming that accuracy is verified, the next step is to transfer the data characteristics to the program Power dialog to determine the minimum number (N) of X, Y pairs that are necessary to give a high probability of detecting the maximum deviation from the line of identity that is considered tolerable. The evaluation of statistical power (ie. necessary sample size) should precede the actual experiment.

The option of importing externally simulated data (eg. using Ref. *2* as a guideline) is mainly to provide for data distributions that differ from the uniform, centrally located and right skewed distributions offered by the program.

## Confidence Intervals and Regions

Figure 1 illustrates a sampling experiment in which OLS was applied to 50 sets of N = 20 randomly drawn X, Y pairs from a hypothetical Y = X relationship. Note the characteristic 'bow-tie' appearance (the same pattern is seen with WLS and Deming regression). Highest values of slope are associated with lowest values of intercept and *vice versa*. High values of slope and high values of intercept are not jointly observed. Likewise, low values of slope and low values of intercept are not jointly observed. Figure 1 illustrates the *correlation* of slope and intercept. For any given error characteristics, the smaller the range of X-values the higher the slope, intercept correlation (increasingly pronounced bow-tie shape). Conversely, increasing the range of X-values reduces slope, intercept correlation.

Figure 2 illustrates a typical slope and intercept parameter space for a straight line model in the case of a narrow range of X-values (max:min ≤ 2:1). The data point in the centre of the graph represents the estimated [slope, intercept] point. The elliptical joint confidence region encompasses a region of slope, intercept values that are *jointly* reasonable for the data. Note that it does not extend into the area of large slope and large intercept values (top right) or low slope and low intercept values (bottom left). The joint confidence region takes account of the correlation of slope and intercept values whereas confidence intervals do not.
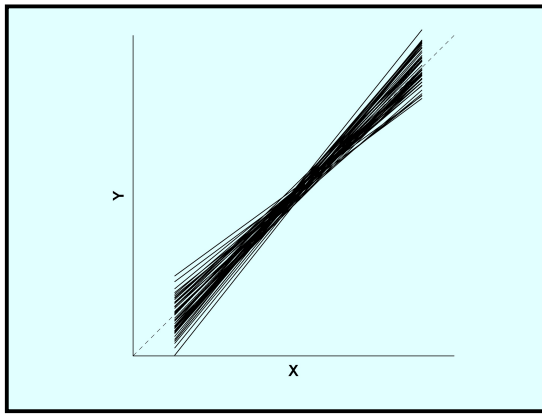
**Figure 1.** Results of OLS regression on 50 sets of X, Y values (each N = 20) randomly drawn from a hypothetical Y = X relationship. X-values were concentrated in the central part of the range and Y-values were drawn from a Gaussian distribution with SD = 10% of the range of the X-values. The dashed line is the line of identity.
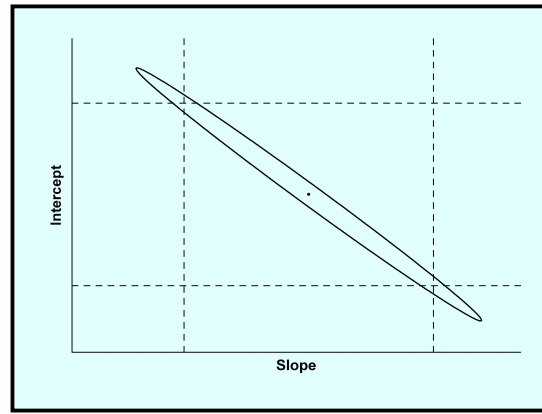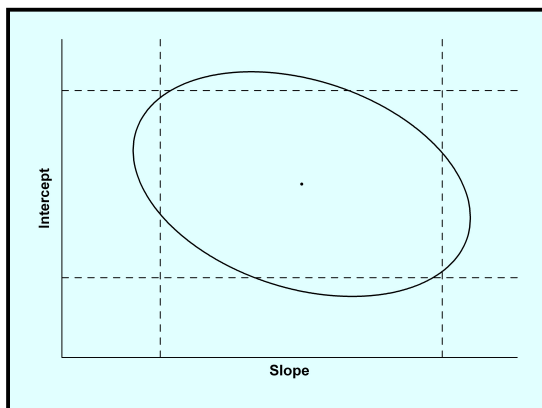


**Figure 2.** Slope, intercept parameter space typically observed with a narrow range of X-values (high slope, intercept correlation). The data point represents the estimated slope and intercept and is accompanied by slope and intercept confidence intervals (dashed lines) and corresponding joint slope, intercept confidence ellipse.

Confidence intervals and regions are identical in the sense that in sampling experiments, such as illustrated in Fig. 1, a specified proportion of intervals and regions (depending in the significance level) are expected to enclose the true underlying [slope, intercept] point (eg. the point [1, 0] in the case of Fig. 1). Either could be used as the basis of significance tests. Intuition suggests that the confidence region is preferable because it takes account of parameter correlation, but there is a further consideration. The usual purpose of significance tests is to detect differences from a hypothetical relationship (such as the line of identity, Y = X) and this is determined by the *failure* of confidence intervals or confidence regions to enclose the hypothetical [slope, intercept] point. The relative shapes in Fig 2 clearly indicate that confidence ellipses will detect smaller horizontal or vertical shifts from a target point (pure proportional bias and pure constant bias, respectively). This implies the ability to detect smaller differences with a given sample size (N) or, more importantly, require smaller sample sizes to detect a specified difference.

The idea of using the joint parameter confidence region is far from new. Munson and Rodbard wrote an excellent paper (*4*) advocating the use of joint slope, intercept confidence regions in an immunoassay context. Text book authors make the same general recommendation (eg. Refs. *5, 6*). However, for one reason or another, confidence intervals appear to be entrenched as the standard mode of significance testing in medical laboratory regression analysis. This computer program is a tool for acquiring a feel for the properties of confidence regions and can also be used to formally estimate necessary sample size by simulating data whose properties mimic as closely as possible the range and error characteristics of real life comparison data. A companion computer program (*3*) estimates variance (weighting) functions and offers OLS, WLS and Deming analysis of real data, including plots of joint parameter confidence regions at user defined significance levels.

To give a more complete picture, Fig. 3 illustrates a typical parameter space when the max:min ratio of X-values is several thousand-fold (very low slope, intercept correlation). In these cases the more circular joint confidence region confers no statistical power advantage. However, Fig. 3 illustrates an extreme case that is only likely to be relevant for a small number of medical laboratory analytes measured by immunoassay.



**Figure 3.** Slope, intercept parameter space typically observed with a very wide range of X-values (low slope, intercept correlation). The data point represents the estimated slope and intercept and is accompanied by slope and intercept confidence intervals (dashed lines) and corresponding joint slope, intercept confidence ellipse.

## Imported Data

Simulated data generated by external applications can be either imported from a text file or copy/pasted via the Windows clipboard. In either case the data must be organised as a single X, Y pair on each line, and the X, Y values must be separated by a single character. Any character can be used for text file data except $0 - 9$, + (plus), - (minus) or period. A Tab character is compulsory for clipboard data. Refer to program on-line help for details of the required data properties. Because of internal storage limitations the number of imported X, Y pairs is limited to $10^6$ (ie. importing is terminated, if necessary, when that figure is reached). Accuracy simulations require a minimum of 3000 imported X, Y pairs and Power simulations a minimum of $10^5$. In the Accuracy dialog the option is given to either randomly draw X, Y pairs from the imported data or to draw *sequential* sets of X, Y pairs of a specified size. The latter could be used to compare results produced by this program with those produced elsewhere with the same data.

Data generated by the program represent an ideal situation because they are randomly drawn from Gaussian distributions with precise error properties as defined in the **Regression Type and Weighting** box. Imported data could be used to systematically assess the effect on both accuracy and power, of data drawn from non-Gaussian distributions, or the deliberate misspecification of errors, eg. generate and import data with certain error properties, then test the effect of deliberately specifying alternative error properties in the **Regression Type and Weighting** box (eg. to mimic the effect of over or underestimated method errors).

## Accuracy

Accuracy in this context refers to the reliability of recovery of the slope and intercept values that were used to generate sets of paired X, Y values and also the concordance between expected and observed frequencies of confidence intervals and confidence regions enclosing target values. Figure 4 illustrates the program Accuracy dialog. Use the **Data and Design Parameters** box to specify properties of the X, Y data that the program will randomly generate, or to import X, Y pairs generated externally. Use the **Regression Type and Weighting** box to specify error properties of the Y-variate (OLS or WLS) or both the X and Y-variates (Deming). A detailed explanation of how to specify error properties is given in program on-line help. The **Results** box includes significance tests on the differences between true and recovered slope and intercept values (see on-line help for more information) and the **Enclosure Frequencies** box summarises the observed enclosure properties of confidence intervals and joint confidence regions. The lower panel displays the current date and time. Use the adjacent buttons to produce a permanent copy of the design and outcome of a simulation run by copying an image of the dialog to the Windows clipboard for subsequent pasting into another application, such as Word, or directly printing an image of the dialog.



**Figure 4.** The program Accuracy dialog illustrating the options available, plus the results from an OLS simulation run using data generated by the program according to specifications given in the Data and Design Parameters box.

The Accuracy dialog has two graphical options.  Click the **View Variance Functions** button to view the Y-variance function (OLS and WLS) and both Y and X-variance functions (Deming).  This option is also available in the Power dialog.  Click the **View Results** button to view the sampling lines estimated in the current simulation run.  They can be viewed as either the lines themselves in the usual X, Y coordinate frame, or as [Slope, Intercept] points in a parameter space plot.

A Deming regression simulation, used to produce results in Ref. *2*, is repeated here to provide examples of the Accuracy module graphs.   The Ref. *2* data were based around an automated immunoluminometric assay for the pituitary hormone thyrotropin ("Access" instrument, Beckman Coulter, Fullerton, CA, USA).  The data range was set at 0.015 – 40 mU/L.  The Y-variance function was estimated (*3*) from daily measurements of thyrotropin in five internal Quality Control specimens over a 7-month period.  A hypothetical X-variance function was constructed such that predicted variance was smaller than predicted Y-variance by a factor of 4 at the lower end of the range, and larger than predicted Y-variance by a factor of 4 at the upper end of the range, with crossover at the mid-point of the range (20mU/L).  In each case the variance function was $\sigma^2 = (\beta_1 + \beta_2 U)^J$ where $\sigma^2$ and U denote variance and mean respectively and $\beta_1$, $\beta_2$ and J are the parameters  (Y-variance function: $\beta_1 = 0.00309$, $\beta_2 = 0.04621$ and J = 2.161; X-variance function: $\beta_1 = 0.04312$,  $\beta_2 = 0.04598$ and J = 4.296).  The variance functions are shown in Fig. 5, translated to CV(%) versus Mean.  Variance function crossover appears shifted to the right because of the optional logarithmic scale on the X-axis.  Although the main intention of RAPS.exe is to mimic real method comparisons, by specifying the range, distribution and error properties of real data, this particular set up was aimed at testing estimation accuracy under a "worst-case" example of non-parallel error properties.  Despite the rather benign appearance of the CV profiles, the predicted relative increase in variance between 0.015 and 40 mU/L is 650000-fold for the Y-variance function and 10.3 million-fold for the X-variance function (typical of some classes of immunoassays).
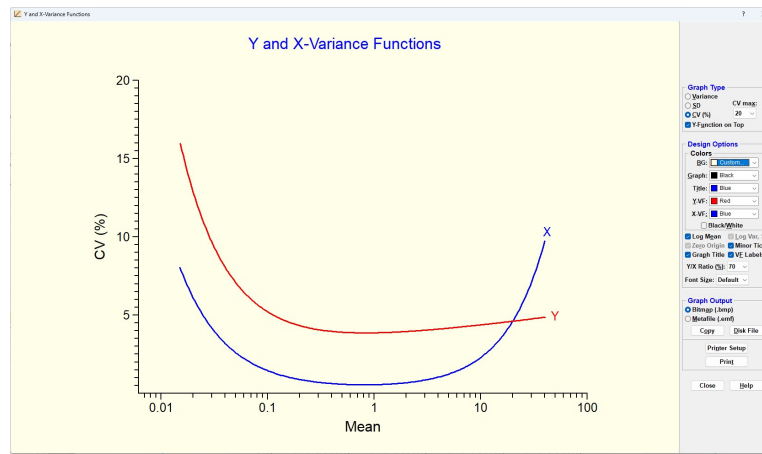


**Figure 5.** Illustration of the Y and X-variance functions used to generate data for the Deming regression simulation results in Ref. *2*.

Figures 6, 7 show the resulting parameter spaces for 10000 Deming regression sampling lines, each estimated from N = 100 X, Y pairs, with true underlying relationship Y = X.  The distributions of the [Slope, Intercept] data points for narrow and wide ranges, respectively, indicate the expected *shapes* of  the joint slope, intercept confidence regions that could have been constructed around each of the 10000 data points (compare with Figs. 2, 3, respectively, and see also the **Power** section below).
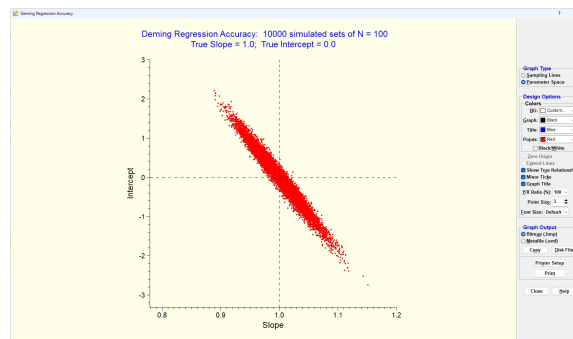




**Figure 6.** Parameter space for 10000 Deming regression sampling lines, each estimated from N = 100 X, Y pairs, with true underlying relationship Y = X, and a narrow data range 14 – 28 mU/L (max:min ratio 2:1). X, Y errors are illustrated in Fig. 5.
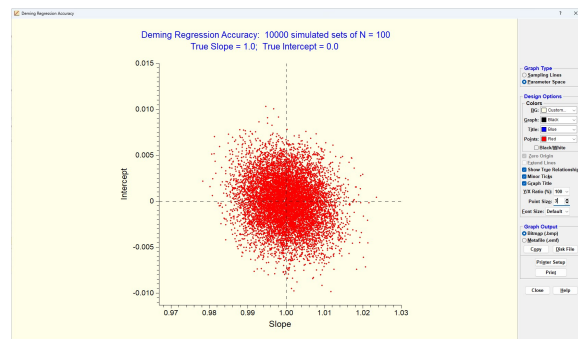
**Figure 7.** Parameter space for 10000 Deming regression sampling lines, each estimated from N = 100 X, Y pairs, with true underlying relationship Y = X, and a wide data range 0.015 – 40 mU/L (max:min ratio 2667:1). X, Y errors are illustrated in Fig. 5.
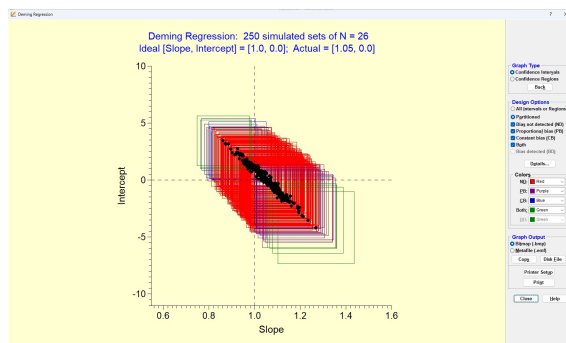
## Power

The program Power dialog, illustrated in Fig. 8, follows the same principles. Data can be either generated by the program or imported from an external source. The focus here is on the Actual slope and intercept values which are used by the program to generate sets of data. The resulting confidence intervals and regions are assessed for enclosure of the Ideal slope and intercept values. The issue is the sample size required to reliably detect a difference between Actual and Ideal values when such a difference actually exists. The program offers automatic sample size determination (as illustrated).

It is important to appreciate that since the program is based on randomly generated data, there will always be some variation between results of simulation runs of the same design. Obviously the larger the number of samples and/or sample size (N), the smaller the run-to-run variation in outcomes, and *vice versa*. Multiple runs of the same design quickly produce a feel for the variation associated with a particular design and sample size. A **V**isualise button (introduced in program version 7.0) generates new data, exactly as per the current power design, and plots confidence interval rectangles and confidence region ellipses thereby graphically illustrating the detection (statistical power) differences between them (see Figs. 9, 10).
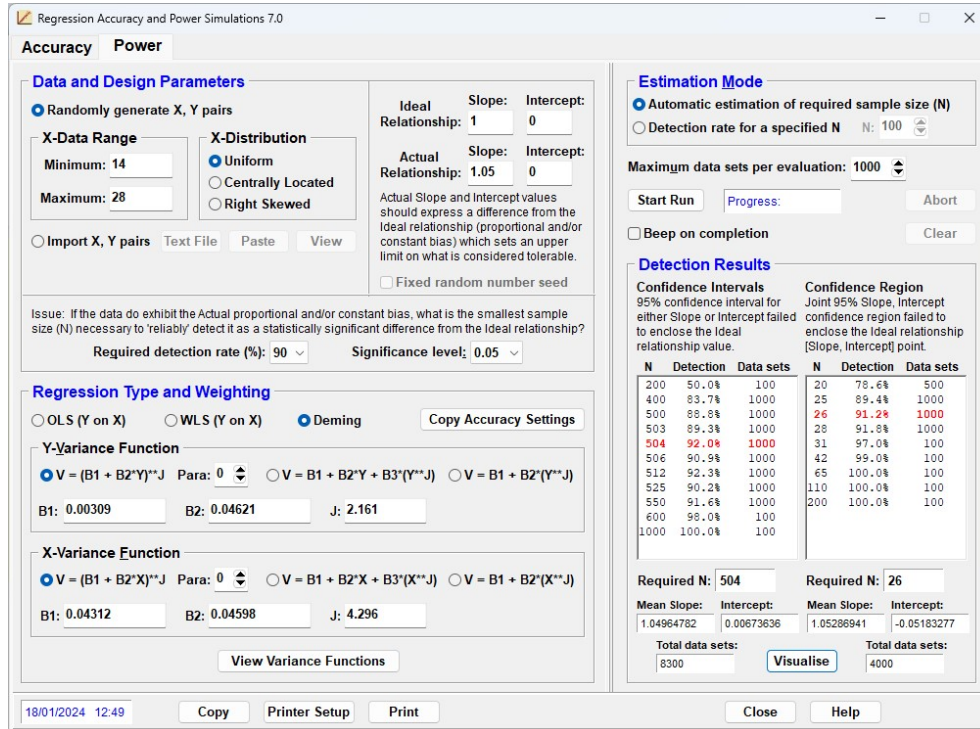


**Figure 8.** The program Power dialog illustrating a Deming simulation run which automatically determined the smallest sample sizes necessary to achieve 90% detection of a statistically significant difference ($p < 0.05$) when the Ideal slope is 1.0 and the Actual slope is 1.05 (ie. detection of 5% proportional bias). The weighting functions are those illustrated in Fig. 5 and the data range is that used to produce the accuracy results shown in Fig. 6 (small 2:1 max:min ratio).



**Figure 9.** Parameter space showing 250 slope, intercept data points, generated by the power design shown in Fig 8, each accompanied by the corresponding 95% confidence interval rectangle. Purple (16/250), Blue (8/250) and Green (7/250) represent bias detected (proportional, constant and both, respectively). Red = bias not detected (219/250 or 87.6%).
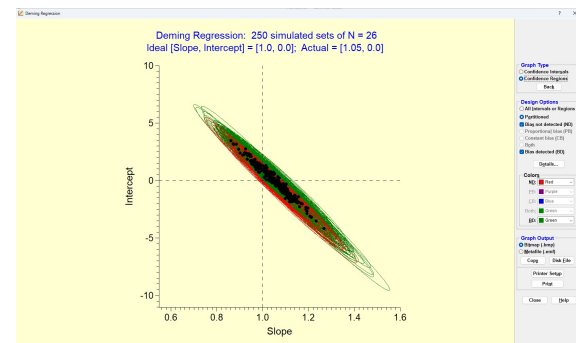


**Figure 10.** Data points from Fig. 9 now accompanied by corresponding 95% confidence ellipses. Compare the ellipse shapes with the data distribution in Fig. 6. Green = bias detected (223/250 or 89.2%), Red = bias not detected (27/250 or 10.8%).

## Program Update History

- Version 1.0 (December 2009; Ref. *2*) was used to produce the simulation results in Ref. *2*. Written with Borland Delphi version 5 and available to registered users at the website of the journal *Accreditation and Quality Assurance*.

- Version 2.0 (October 2010) was written with Embarcadero's Delphi 2010. Visual themes were introduced, on-line help was upgraded from WinHelp to HTML help and graphical displays of X, Y variance functions and Accuracy sampling results were added (see Figs. 5 – 7, for example).

- Version 3.0 (May 2011) was written with Embarcadero's Delphi XE. A formal significance test on the accuracy of recovery of underlying slope and intercept values was added (see Fig. 4) and three version 2.0 bugs were corrected; (i) Variance functions (eg. Fig. 5) always plotted correctly but warning and error messages that the program issued in some circumstances did not always match what was seen in the plot (ie. messages and plots were not always properly coordinated), (ii) In very rare circumstances, highly inaccurate power results for confidence intervals were possible after changing the **Para** spin edit value (results were so inaccurate they would almost certainly have been recognised as a program bug), (iii) Changing the **Para** spin edit value in the Power dialog occasionally resulted in inaccurate updating of the Y-axis of variance function plots. The latter two bugs were both due to a failure to reset some internal variables in the Power dialog following a variance function selection change.

- Version 4.0 (January 2013) was written with Embarcadero's Delphi XE3. The program now required Windows 2000 or later, a screen resolution of 1024 x 768 pixels or better and 16-bit colors or higher. Buttons were added to automate the transfer of **Regression Type and Weighting** settings between the Accuracy and Power dialogs (both directions). The import of externally generated data was extended from Text file only to include pasting from the Windows clipboard. Variance function plots in terms of Variance or SD versus Mean were updated to allow selection of a logarithmic scale on the Y-axis.

- Version 5.0 (July 2020) was written with Embarcadero's Delphi 10.1 Berlin. The program now requires Windows XP or later. The user interface controls have been spread out and reorganised. The program has also been brought into line with a companion program (*3*) by incorporating a 4-parameter turning point variance (weighting) function, and by allowing negative X, Y values (which are constrained to be plausible observations from the detection limit region of a measurement system). The negative values update anticipates possible future developments because left-censoring to zero or to higher values introduces biases into method comparison analyses (*7*).

- Version 6.0 (April 2023) was written with Embarcadero's Delphi 11.2 Alexandria. The program was rebuilt to ensure that it performs correctly under Windows 11. When importing data (see p. 3) users can view and check the contents of a text file or the Windows clipboard prior to the commencing the import process. A new window now allows users to view and check the data at any time after importing. The more important changes correct two bugs related to imported X, Y-values. The program requires imported X-values to have a max:min ratio $\geq 1.2$. The upgrade to version 5.0 (see previous paragraph) allowed negative X-values but, regrettably, this was not taken into account when testing the max:min ratio, ie. a single negative X-value automatically produced a negative max:min ratio and therefore rejection of what might have been perfectly valid data. Second, if the program detected non-numeric characters within imported data it flagged that 'value' for rejection by assigning a value -1 (ie. a value that was interpreted as invalid prior to version 5.0, but incorrectly accepted as a legitimate value from version 5.0). Now corrected.

- Version 7.0 (January 2024) was written with Embarcadero's Delphi 11.2 Alexandria. A new **Visualise** button has been added to the program Power module (see Figs. 8 – 10 on p. 5).

## Program Delivery, Installation and Removal

The Australasian Association of Clinical Biochemists has kindly allowed the program installation file to be downloaded from its website (www.aacb.asn.au). Proceed to Resources > Tools and Links which contains a link to the Regression Accuracy and Power Simulations program. Click the installation file link to download file RAPS70Install.zip (4034 Kb) and extract the installation file RAPS70Install.exe (4173 Kb). Double click it to launch the installation process. C:\RAPS is suggested as the program home directory but you are free to specify an alternative location. RAPS.exe (3134 Kb) and on-line help file RAPSHelp.chm (65 Kb) are installed into the specified home directory and a shortcut icon is placed on the desktop.

Use the Add/Remove Programs icon (or equivalent) in the system Control Panel to remove the program icon and the installed files.

## References

1. Deming WE. Statistical adjustment of data, New York: John Wiley & Sons (1943). Reprinted, New York: Dover Publications Inc., pp. 184 (1964).

2. Sadler WA. Joint parameter confidence regions improve the power of parametric regression in method-comparison studies. Accred Qual Assur, 15:547-554 (2010).

3. Sadler WA. Variance Function Program. Australasian Association of Clinical Biochemists website (www.aacb.asn.au). Program available from the link Resources > Tools and Links.

4. Munson PJ, Rodbard D. Statistical comparison of two assay methods (weighted linear regression with errors in x and y). In: Wilson DW, Gaskell SJ, Kemp KW, eds., Quality Control in Clinical Endocrinology, Cardiff: Alpha Omega, pp. 89-98 (1979).

5. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical Recipes in Fortran, 2nd ed., Cambridge: Cambridge University Press, pp. 660-662 (1992).

6. Draper NR, Smith H. Applied Regression Analysis, New York: John Wiley & Sons, pp. 64-69 (1966).

7. Sadler WA. Methods comparison biases due to differing uncertainties and data censoring. Ann Clin Biochem, 56:608-612 (2019).