COMPARISON OF MEASUREMENT METHODS USING PATIENT SAMPLES.

2018-07. Anders Kallner. Dept Clin Chem Karolinska university hospital, Stockholm Sweden.
anders.kallner@ki.se

Summary

This spreadsheet program is primarily designed for estimating the difference between two methods by comparing results of measured patient samples. It allows input of single or duplicate results. With duplicate measurements, the imprecision of the methods can also be estimated. Differences between the means of the reference and measured samples are evaluated with parametric and non-parametric methods, in addition to ordinary and Deming regression analyses. Results are displayed in a scattergram, and absolute and relative difference diagrams. Differences are demonstrated in an error grid comprising A, B and C zones. The data set can be partitioned to facilitate a detailed evaluation depending on relevant threshold values. Summary instructions are entered as comments in critical cells. All cells are protected except those for input, which have blue borders.

Short Instruction

- Download and open the spreadsheet program.
- A sample data set is provided (*2018-07 ACB Method comparison, patient samples - example data.xlsx*), consisting of a column with a sample number or ID and four columns comprising results of duplicate measurements of 40 samples using two different methods.
- Copy the five columns and paste into the spreadsheet using “Paste special”, “Values (123)”.
- Most cells will still be blank; however, results are plotted in the graphs.
- Enter the name of the component (measurand) in cell E10.
- Most cells will now display “DIV/0” and you are required to define the measuring interval (cells L22 to N24) and the assumed interval of differences (cells L27 to N28). You may find some guidance to suitable limits in cells from the graphs.
- Fine-tune limits of intervals and define what you want displayed in the graphs (Y/N in appropriate cells).
- The spreadsheet is now fully operational and you may make any changes necessary – just remember that only the blue cells are open for input.

Background

Measurement of the difference between a test and previous (comparative) method can be efficiently accomplished by comparison of paired measurements. This approach is only comparative and provides no information about whether either method is biased against an appropriate reference or conventional method. By measuring patient samples in duplicate, with concentrations covering the entire measuring interval, a good estimate of both bias and imprecision can be obtained provided the standard deviation is close to constant in the measuring interval.

Experimental design

- Select a minimum of 20 patient samples (the spreadsheet can accommodate 1,000 samples) with concentrations that cover the measuring, or the clinically relevant, interval.
- Samples with known sources of interferences such as lipidaemia, haemolysis or icterus should not be included.
- Ideally, each sample should be measured in duplicate in a random order by both the test and comparative (reference) methods. The random order is to prevent possible effects of a carry-over; however, the input data can be sorted without effect on the calculations. The program allows single measurements by either or both methods, but the outcome will be less reliable with single measurements and the imprecision of the methods cannot be estimated. Optionally, either one method or both methods can be chosen as containing duplicate or single results (cells L3 and L4, Figure 1).
- Enter the results in the columns C to F in the spreadsheet; if only single measurements are available these should be entered in the column (C) of the comparative and test (E) methods, respectively.
- Results can be copied from other suitable spreadsheets and pasted into the appropriate columns of the present spreadsheet. If this is done, use the "*Paste special*" and "*Values*" option to avoid modifying the colour and other design features of the selected cells.
- **Note that the name (E10) of the studied component and the identification (C13 and E13) of the measuring systems must be entered before calculations will start.**
- Furthermore, an interval e.g. the min and max values of the comparative method (cells N22 to N24) should be entered.
- The scales of the axes of the graphs are automatically adjusted, but if the results are numerically large it becomes important to optimise the scales by increasing the minimum value of the axis since it will automatically start at zero. The scale of the axes can be changed by standard EXCEL procedures. The scales of X and Y-axes of the scattergram shall be equal or the equal line will not have a slope of 45 ° (degrees).

Calculations

Always inspect the scattergram and difference graphs for apparent outliers before evaluating the results.

First, decide if the calculations will be based on duplicates of the comparative (X) or test (Y) method, which is the default. If this is not the case, change the contents of the appropriate cells (L3 and L4). If the duplicate option is negated then only the data in the first column will be used and any data in columns D and F, respectively, will be disregarded. It is thus possible to include duplicate or single measurements for either of the methods. The program uses the average of the observations for graphs and regression analysis. Since routine measurement procedures usually measure samples in singlicate, it might be interesting to use duplicates for the reference and singlicate results for the routine samples; however, use duplicate measurements to identify samples with unusually discrepant results and to estimate the method imprecision.

The absolute and relative difference between observations in the independent, dependent or average results are flagged in columns, subject to value in cell J10.

One or both of the results of the test or comparative method can be deleted. If both from either method are deleted, then the corresponding results of the other method will be disregarded and the degrees of freedom adjusted. If only one (column D or F) is deleted, the remaining will be used and all averages included in the *t*-test. Note that Excel disregards an empty cell in the calculation of the average or variance.

The program calculates critical statistical parameters, Figure 1. The false rejection rate (α) can be modified; where 1 % or 5 %, correspond to a confidence level of 99 % and 95 %, respectively, and the recommended values. To determine whether there is a significant difference between the means of the two methods, a Student's dependent *t*-test is applied. This test is based on the differences between the paired observations or paired results in the first column of each group (if single measurements have been performed or chosen), or on the averages for each method. The evaluation is presented as "Significant difference" or "No significant difference".

The validity of the Student's t_{dep} requires that the *differences* are normal distributed. The validity is independent of the distribution of the results of the observations. In addition to the Student's dependent *t-test*, a Wilcoxon sign rank test is performed and the *p*-value reported and evaluated.

The number of missing data is separately reported for the test and comparative methods, respectively. Even if there are missing data, calculations are carried out correctly.

The averages of the comparative and test methods are calculated and presented in L12. Optionally, the median or the standard deviation of the distribution of the results can be displayed (N11); if the median is chosen, the interquartile interval (or any other interval as chosen in cell P10) is shown.

The coverage factor (P9) and the upper percentile (P10) can be set. Defaults are k=2 (coverage factor, approximately yielding a 95 % probability interval) and the upper percentile 75, respectively.

The relative percentage difference is calculated for each sample in relation to the results of the comparative method and the average calculated from the individual observations. A seemingly paradoxical effect can occur where a positive average bias is obtained where the relative average bias is negative, and vice versa.

The difference is calculated as the difference between the results of the test method and the comparative method, i.e. *Test* minus *Comparative*.

An ordinary linear regression (OLR) model is fitted to the data. The regression is estimated from the average of the observations (if available) or from the first observation of duplicate measurements. This can be changed by toggling between Y and N in cells L3 and L4, respectively.

If duplicate measurements have been performed, the imprecision of each method can be estimated in addition to bias. This calculation is based on the difference between the results of the duplicates (Dahlberg's formula).

$$s(x) = \sqrt{\frac{\sum_{i=1}^{N} d_i^2}{2 \times N}}$$ (N is the number of paired samples)

This approach assumes that the standard deviation is constant within the measuring interval, but it is regarded as a relatively robust method.

The orthogonal regression (Deming) allows a variation in the results of the dependent (Y) and independent (X) variables but requires that the λ ratio $(SD_{test}/SD_{comparative})^2$ is specified. If duplicate measurements have been performed, λ is estimated from these data. If the λ is large, then the Deming regression will approach that of "ordinary linear regression". The program allows input of the individual standard deviation of the X or Y variable (from other sources), e.g. if only single measurements have been carried out (L19 and L20, Figure 1). If a standard deviation is entered in L19 and L20 then their squared ratio will take precedence over the calculated λ.

Duplicate Y:s (Y/N):	N	ACB; 6.63			
Duplicate X:s (Y/N):	Y				
Student's tdep:	**2.06**	Number of samples:	48		
α %	5				
df:	47		Pairs	Obs.	
t_{crit}:	2.012	Missing Y:	0	0	
p (2-tails):	**0.045**	Missing X:	0	0	
Significant difference			Coverage factor:		1.96
Wilcoxon p:		Median or	Upper Percentile:		75
No signif diff.	**0.089**	SD (M/S):	M	25 perc	75 perc
Y-mean:	39.43	Median:	39.50	33.55	45.00
X-mean:	39.15	Median:	39.10	33.45	45.40
Median diff:	**0.283**	St dev:	0.95	**SEM:**	**0.14**
Median rel diff:	**0.72**	St dev:	2.94	**SEM:**	**0.42**
s(Y) Test:		X-min:	26.4	Y-min:	27.9
s(X) Comparison:	1.00	X-max:	51.4	Y-max:	53.0
λ $(s(Y)/s(X))^2$:	1.000				

Figure 1. Calculated statistics and options for manual input.

Report of results

In addition to the table shown in Figure 1, the results are presented in three graphs and a separate sheet of histograms.

The scattergram shows the results plotted together with the line of equality and vertical lines indicating the partitioning of the data. The operator can choose to display either or both regression lines (OLR or Deming) in the scattergram. The average and median are also displayed.

Two difference graphs are presented, one showing the absolute differences and the other the relative differences in relation to the average of the results of the two methods. If the comparative method is a reference method (i.e. with a small imprecision), then it is recommended to compare the differences to the results of the comparative method rather than the average of the two. The operator may toggle between these two conditions by changing the symbol (Y/N) in cell O28. The X-axis is changed in both graphs. A trend line of the observations is plotted, the average difference as estimated in the partitions, and the estimated allowable difference (*k* x *s(bias)* of the total bias).

The slope and intercept of the Deming and OLR and their uncertainties are displayed. The coefficient of determination and the correlation coefficient with its confidence (95%) interval is also calculated.

The spreadsheet program may be used for regression analysis of other datasets when it becomes important to verify that the slope is significantly different from zero.

The operator may choose to partition the results. The program allows three partitions of the comparative (X) values, which translates into the corresponding test (Y) values. Moreover, the tails of the distribution of the comparative method can be moved and thus the program offers adjustments for skew distributions (e.g. when the data set is truncated). The difference between the results is presented for each partition and differences between the methods are evaluated within in partition using the *t-test*, adjusted for the degrees of freedom (*df*). If duplicates have been measured, then also the imprecision in the partitions will be reported.

The scales of the axes of the graphs are automatically adjusted, but if the results are numerically large it becomes important to optimise the scales by increasing the minimum value of the axis. This modification can be achieved by standard EXCEL procedures. The scales of X and Y-axes of the scattergram should be equal in comparison of measuring methods.

Scatterplot and partitioning of data

Up to three tiers can be defined in the blue cells as shown in figure 2. The user needs to define the reporting interval in cell L22 and N24, respectively.

This allows the partitioning of the data set into three partitions and to truncate the dataset. If the upper limits of the Low and Mid partition are equal or the upper Mid is blank, then only two

partitions will be defined. The limits of the partitions are shown in all graphs as vertical lines. The number of observations in each partition may be small and the relevance of any difference or *t*-value should be carefully considered as well as the distribution (s(X), s(Y)). The number of observations in the partitions is defined as the number of observations of **the comparative method** within that particular partition.

Partitioning of Comparison results				Number:	Mean bias:	Bias %	s(X)	s(Y)	t_{dep}	p-value	Signif.	Slope	Interc	Coeff det	Display
Low:	0.0	to	30.00	3	1.60	6.0	1.2		1.658	0.239	NS	-1.31	64.40	0.961	N
Mid:	30	to	40.00	24	0.10	0.2	0.3		0.776	0.446	NS	1.07	-2.22	0.969	N
High:	40	to	55.00	21	0.24	0.5	0.3		1.503	0.148	NS	1.03	-1.12	0.962	N

Figure 2. Partitioning of the measuring interval into three tiers. The regression lines of each partition can be displayed.

Regression:			
Deming		±U	N
Slope:	1.00	0.02	48
Intercept:	0.25	0.68	
Coeff determ (r^2):	0.9868	r=0.993	0.988 < r < 0.996
Display (Y/N):	N		
OLR		±U	N
Slope:	0.99	0.017	48
Intercept:	0.51	0.67	
Display (Y/N):	Y	Slope sign. diff from zero	
Displ equal line (Y/N):	Y		
Displ obs (Y/N):	Y		

Figure 3. Characteristics of the regression functions and correlation coefficient. If a "Y" is entered the regression line for the Deming and Ordinal regression functions, respectively, will be shown in the scattergram.

It is important to visually inspect the scattergram (Figure 4) to determine if any of the results might be outliers. Identification of outliers is facilitated by the difference graphs and the highlighting ofthe maximal difference between duplicate values or averages in columns I and J. Although there are rules for the identification of outliers, a visual inspection is often sufficient. If a value pair is considered an outlier, it can easily be removed from the input sheet. Blank rows are accepted and the calculations are immediately updated. The removed data can be restored by using the restore button (back arrow) from the EXCEL tool bar.

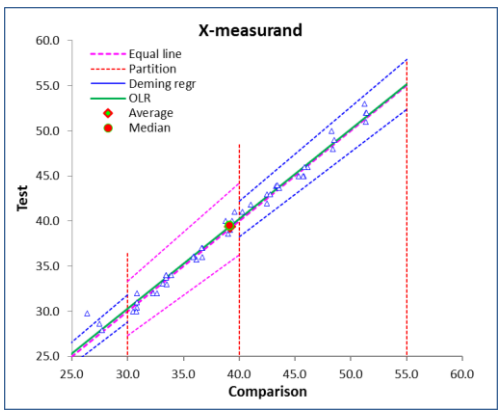

Figure 4. Scattergram with the partitions and partial error grid displayed. The equal line (dotted red) and the Deming (solid blue) regression lines are shown. The A- (5 %) are indicated in the first and third partition and the A+B zone in partition 2 (10 %).

If the absolute or relative difference is not considered constant, the importance of using the regression parameters to estimate the difference increases. If a clinically significant difference is

observed, the above procedures may not be adequate for establishing a recalibration function. In this case, to estimate the regression adequately, more samples should be measured. Alternatively, a two-point recalibration using patient material with assigned and traceable concentrations obtained by a reference method should be carried out.

Difference graphs

Originally, Bland-Altman recommended that the average of paired values should be used as the independent variable and the difference between the test and comparative method as the dependent variable. It has since been argued that if the comparative method is a reference method, then it advantageous to use this quantity as the independent variable in the difference graph. This would also more closely resemble the regression displayed in the scattergram. Accordingly, the program offers both alternatives (Cell O29). The absolute and relative differences are displayed in separate graphs. The relative differences are particularly interesting at low concentrations.

There are three alternative displays available. One with only the scatter of the differences, one with only the "tilted mountain plot" (empirical distribution function) and one with both displayed. The partitions chosen will be shown in the difference graphs in addition to several help lines. Note that the scale of the mountain plot is a secondary axis displaying the fractiles and thus always limited to 0 to 0.5.

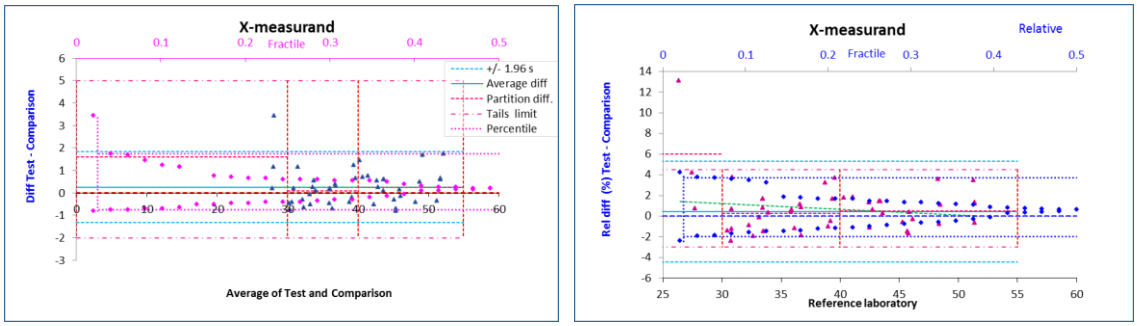

Figure 5. Difference graphs The absolute differences are plotted against the average of the two measurements, the relative, for illustration, against the independent variable. The vertical lines represent the partitions of the data and the horizontal lines allowable differences as defined in the tables. The tilted mountain plot revels a right skewness of the differences (towards larger values. A regression line is shown to illustrate a value-related difference between observations. This is particularly noticeable in the relative difference plot (right)

Evaluation

The statistics (Figure 1) show if the difference between measurements by the two methods is statistically significant according to the calculated t_{dep}. Depending on the degrees of freedom (*df*) and the false rejection rate (α), the *p*-value is calculated (i.e. the smaller the *p*-value the more likely that the difference observed between the two methods is not due to random sampling error). This evaluation is based on Student's *t*-test and assumes that the difference between the results is approximately normal distributed. If this is not the case, the evaluation may not be valid.

The result of a Wilcoxon sign rank test is also presented; the studentized *p*-value and its interpretation are calculated. If the two *p*-values (according to Student and Wilcoxon) are widely different, it is more likely that the Wilcoxon *p*-value is more reliable (conservative) but both methods have limitations in applicability.

The averages and standard deviations of the patient results and the average of the absolute and relative bias are calculated as described above and shown in the table.

From a medical point of view, the statistical significance of a difference may not be interesting and the evaluation should rather focus on the average difference (i.e. the effect size or magnitude of the difference).

The difference graphs display a number of detailed results. Thus, the average difference of each of the partitions and the total dataset are displayed. The standard deviation of the differences is calculated and displayed considering the coverage factor defined above.

If the trend line in either the absolute or relative difference plot is horizontal, this indicates that the absolute or relative difference is constant within the measuring interval. If constant then the difference is a useful and valid estimate in the measuring interval. If not constant it may be helpful to estimate the difference in partitions based on the results by the comparative method. This can be viewed as a limited "uncertainty profile".

Error grid

In the evaluation of differences between results of measurements, the null hypothesis is that there is no difference. The null hypothesis significance testing therefore focuses on the probability that this the null hypothesis is true. If the probability is less than a particular value (arbitrary and conventionally set to 5%), the probability of alternative hypothesis is believed to be (100-5) % i.e. 95%. Thus, the outcome of the comparison has become a dichotomous choice. The significance of the size of a difference is a different matter and can be formulated as the probability that the difference is within a predetermined interval. The null hypothesis is formulated differently and is usually as two different criteria. The mathematical solution to this is not presented in the program but a graphical presentation in the scattergram and a mathematical summary. This is generally recognized as an error grid and expresses the number of observations that are found with a given sector from the equal line or a regression line. A commonly used criterion in method comparisons is that 95% of the observations shall be found within an A-zone with a width of ±5 %. The program allows defining the A-zone and the limit of the C-zone outside of which no results shall occur. The zones can be defined in relation to the equal line or any of the regression lines. The outcome is reported in a table that can optionally be expressed in absolute or relative numbers. The number of observations outside the defined confidence limit of the average differences are separately reported and identifies as above or below the average difference.

Histograms

Distributions of the results of the two methods and their differences are presented histograms in a separate worksheet. A bin size is calculated on the measuring interval but can be manually modified. The histograms also display a Gaussian distribution graph based on the calculated average and standard deviation of the observations. The distributions of the results are summarized in a boxplot.

The normality of the differences (important for the validity of the Student's test) is visualized in a Q:Q plot. Detailed properties of the found differences are shown in a table.