

Comparison of analytical methods in clinical chemistry

Philippe Marquis, Service de biochimie
Centre hospitalier, Metz – France

Method comparison is performed in clinical chemistry laboratories to evaluate the agreement between two analytical methods. Data are obtained collecting samples uniformly distributed in the reportable range. They are split in half and one piece is assayed by each method.

The quality control software *MultiQC* (www.multiqc.com) includes an original method to process and plot comparison data. It gives up statistical testing for an evaluation connected to *medically allowed tolerance*. A demo of 26 slides is available at

- www.multiqc.com/MethodComparison.htm (Shockwave Flash Demo)
- www.multiqc.com/MethodComparison.exe Off-line executable file)

1. When to compare ?

Method comparison is performed whenever a new method is considered for replacing a current one. Interchanging methods may be

- Definitive if the new method has better operational qualities than the former one.
- Temporary if the new method is only an alternative method, which can replace the current one in case of failure of the routine analyzer.
- Cyclic when two analyzers are performing the same assays at different hours of the day.
- Continuous when two mirror-analyzers are simultaneously performing the same assays to increase the analytical throughput.

Method comparison is usually recommended when a new analytical method is started. But it is also useful to resort to method comparison in routine work either to troubleshoot an analytical issue or to demonstrate the permanent agreement between two analyzers within a laboratory or between two laboratories.

2. Criterion of comparison

Changing an analytical method to another one is only possible if they agree sufficiently closely. In this respect, two analytical methods can be equivalent, commutable or incompatible.

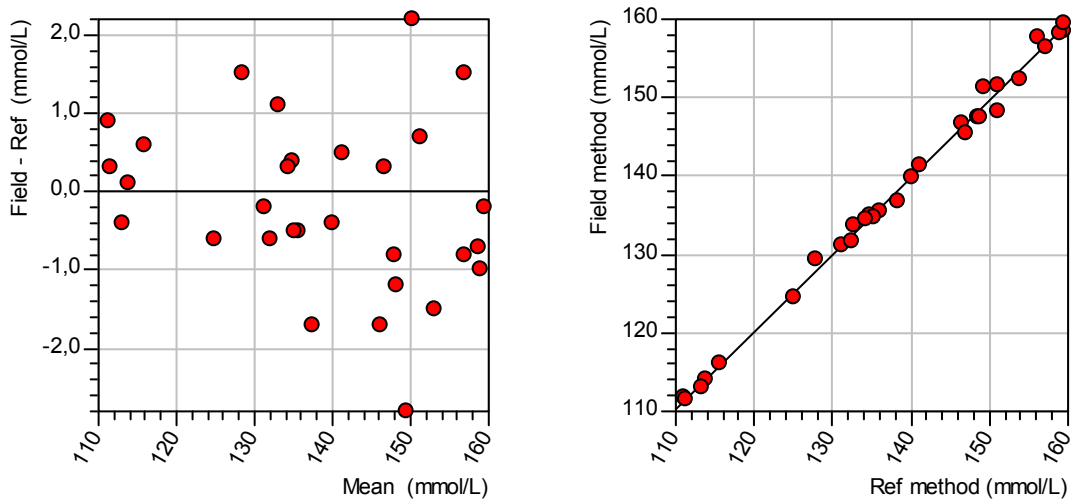
- Methods are equivalent if they give equal results within their inherent *imprecision*. They can be interchanged without loss of analytical accuracy.
- Methods are commutable if they give equal results within the *medical tolerance interval*. They may be not rigorously equivalent but they can be however interchanged without loss of diagnostic power for patients.
- Incompatible methods lead to results with differences greater than the tolerance interval.

Taking into account that our laboratories are intended to a clinical use, a departure of a new method from the current one can be accepted if it does not impair the medical diagnostic or follow up. So, the best cost-effective criterion that allows interchanging two analytical methods is *commutability* and not equivalence. Error made by replacing an analytical method by another one which is not rigorously equivalent is named the *non-equivalence error*. It is a component of total analytical error.

3. How to compare ?

There are two ways to compare split samples assayed by two analytical methods.

- The *scatter-plot* : the values of the new method are plotted against the corresponding values of the current one. The mathematical relation between methods is estimated by a regression line. The disagreement between methods is measured by the departure of the regression line from the bisecting line of the plot (identity line).
- The *difference-plot* : the differences between concentrations in every split sample are plotted against the means of each pair. The disagreement between methods is measured by the deviation of the points from the horizontal nil-bias line.



Comparison of two serum sodium methods : difference-plot on the left and scatter-plot on the right (same data)

Both methods have pros and cons. They provide complementary information. CLSI (former NCCLS) has published guidelines for method comparisons where both scatter-plot and difference-plot are advised.

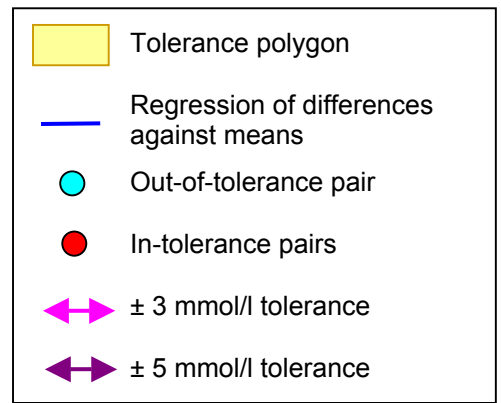
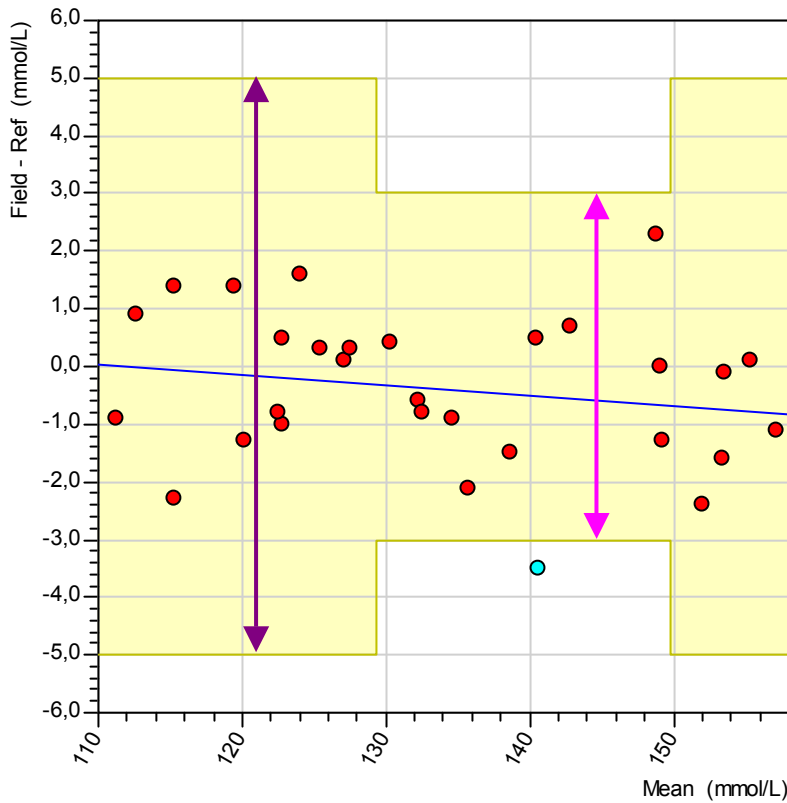
4. Tolerance polygon

Medically acceptable error is a basic figure whose knowledge is essential to a cost-effective management of quality in a laboratory. It may be an absolute or a relative allowed error. Most often in clinical chemistry, both are associated so that absolute error applies to lower concentrations and relative error applies to higher concentrations. MultiQC maintains a table of medical tolerance intervals for every analyte that it controls. The software allows sophisticated tolerance schemes because it is possible to separately define relative and/or absolute errors for low, mid and high concentrations.

➤ Difference-plot

For every concentration on the X axis, the error allowed for the difference within each split sample is represented by a vertical segment centered by the horizontal line of nil bias. On the whole, individual tolerance segments are merged into a polygonal area framing the nil-bias line. The top and bottom edges of this polygon are parallel for an absolute tolerance. The edges are diverging rightwards for a relative tolerance.

Interpretation of a difference-plot is easy : any point inside the tolerance polygon satisfies the tolerance conditions. Any point outside of the tolerance polygon denotes the non-commutability of the two analytical methods for the relevant sample. The final verdict is based on the percentage of non-commutable points found on the whole plot.



Concentrations	Allowed total error
< 130 mmol/l	5 mmol/l
130 to 150 mmol/l	3 mmol/l
> 150 mmol/l	5 mmol/l

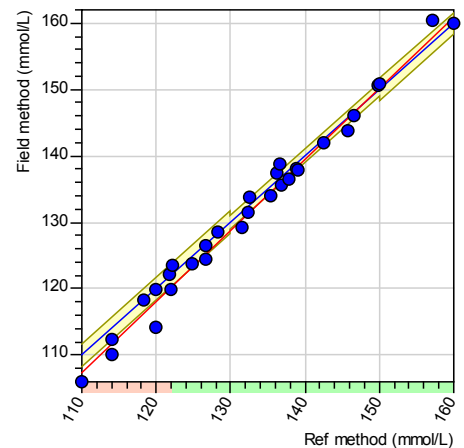
Difference-plot with its tolerance polygon : Comparison of two serum sodium methods

➤ **Scatter-plot**

Building a tolerance polygon on a scatter-plot is less straightforward. The aim is to evaluate the departure of the regression line from the identity line. In routine laboratory work, slope and intercept of regression lines are generally calculated from a set of, say, 30 to 100 samples. Random error on regression lines is thus minimized to put *non-equivalence error* in prominent position. So the error allowed for a regression line is smaller than the total error allowed for individual concentrations.

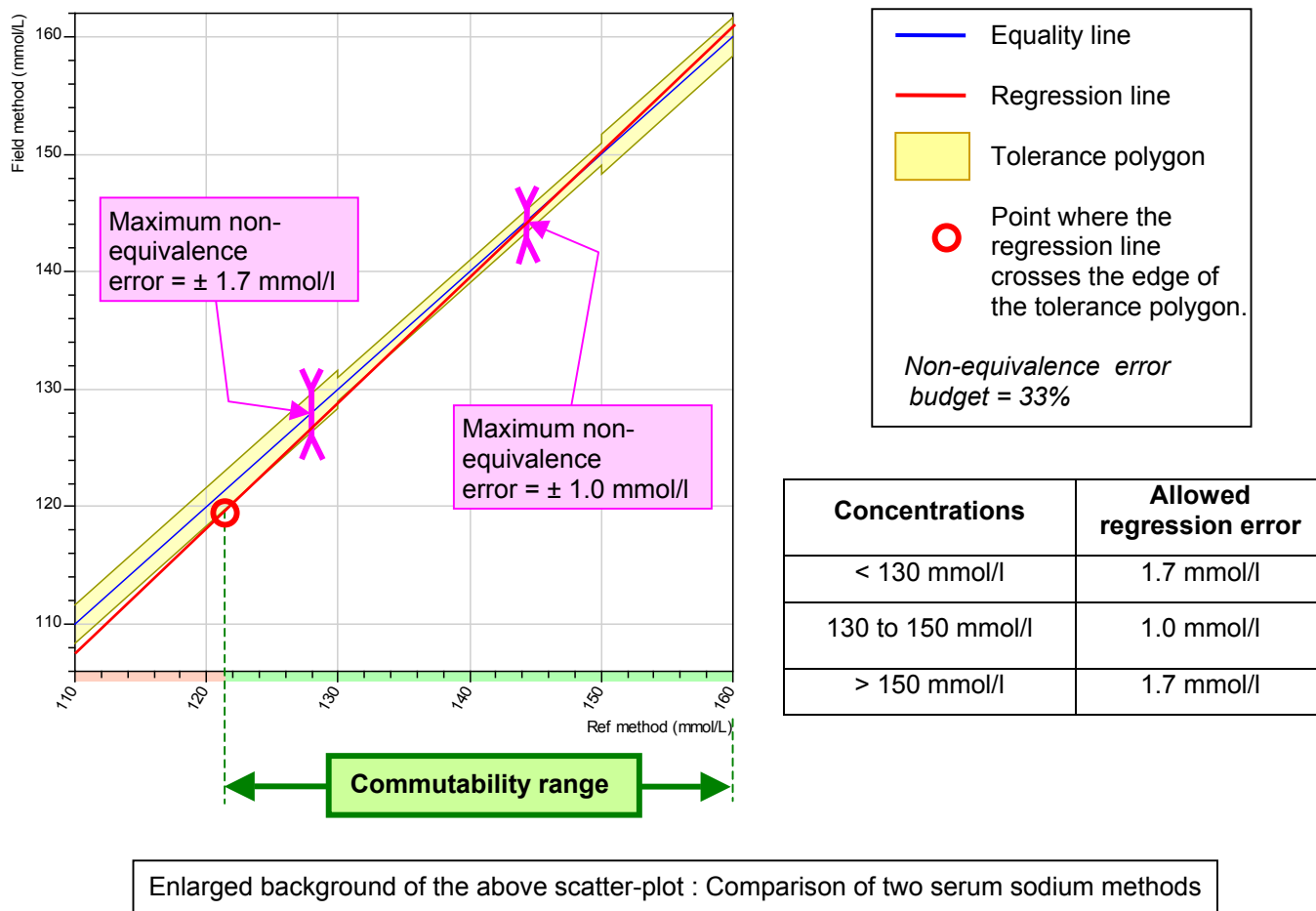
MultiQC makes use of a reduction factor named *non-equivalence error budget*. Its default value is 33%. This means that if the overall tolerance for serum cholesterol is 6%, the criterion allowing the interchange between two methods will be : non-equivalence error less than 2%. The tolerance polygon is built on a scatter-plot with this reduced tolerance. It frames the bisecting line.

Interpretation of a scatter-plot does not depend on individual points but only on the regression line. Any point of this line which is inside the tolerance polygon satisfies the commutability criterion. So the commutability range is established searching for the segment(s) of the regression line which is/are interior to the tolerance area.



Scatter-plot : Comparison of two serum sodium methods. (Background enlarged below)

MultiQC searches for the intersections of the regression line with the top and bottom edges of the tolerance area to find out the ends of all the in-tolerance segments. Then the software projects these segments onto the abscissa axis. The commutability range is outlined by a green background on the X axis. The methods under comparison are commutable if the commutability range is wider than the reportable range of the reference method.



5. Is there a best regression method for a scatter-plot ?

Practice shows that quality of data matters much more than statistical models. Anyone may make his own opinion with MultiQC which can instantaneously switch between ordinary linear regression, Deming regression, weighted Deming regression and Passing-Bablok non parametric regression. Differences between estimates of regression parameters are generally insignificant in comparison to tolerance provided that data be real laboratory data covering the whole analytical range.

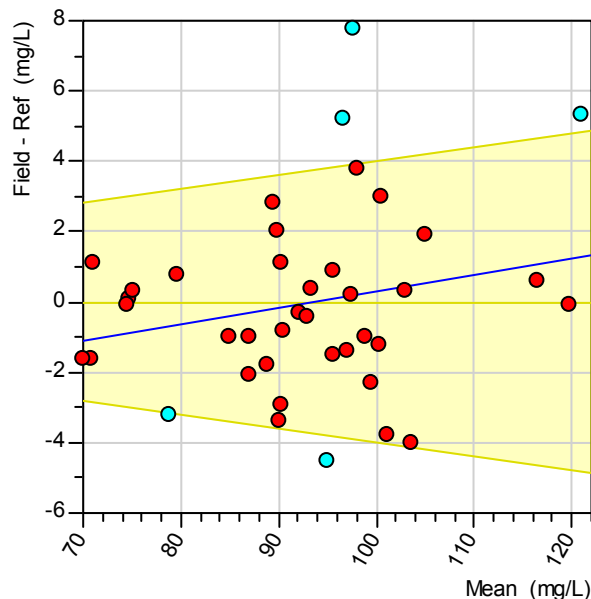
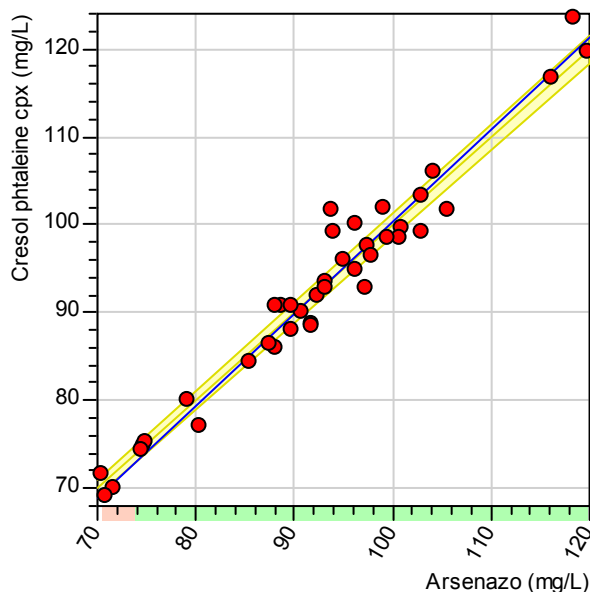
6. Discrepancies between scatter-plot and difference-plot ?

The figures below show a comparison between two serum calcium reagents : cresol phtaleine versus arsenazo. Plots are based on a medical tolerance of 4% and a non-equivalence error budget of 33%.

The *scatter-plot* (left picture) shows a commutable range of [74 to 120 mg/l]. So, the two methods should be directly commutable, at least for normal and elevated calcemias. The commutable range might be widened to lower concentrations by setting calibration correction factors, easy to calculate from the regression parameters.

There are however 5 out-of-tolerance pairs in the *difference-plot* (blue in the right picture). This bad agreement between methods cannot be explained by the imprecision of the one or the other. Capability indexes calculated by MultiQC are about 2 for each method. Both are therefore highly capable to

individually meet the medical tolerance. Why then do the differences between methods do not meet the medical tolerance ? The answer is generally referred to as “aberrant-sample bias” whose origins may be differences in specificity, matrix effects or many other unknown causes.



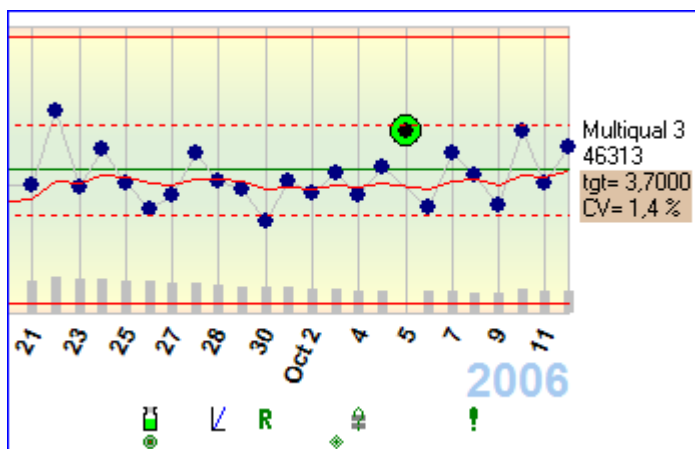
The conclusion is that the scatter-plot and the difference-plot are complementary. The former evaluates the average agreement between all of the pairs whilst the latter individually focuses on each pair.

The scatter-plot is a tool to search for differences of calibration.
The difference-plot is a tool to search for aberrant pairs.

7. Method comparison and quality control

The quality control software *MultiQC* (www.multiqc.com) includes a module to process method comparison data. This data is archived among the particular events of the relevant analyte. Among these events we can also find calibrations, changes of reagent lots, linearity checking and repeatability tests.

Analytical events can be shown at any time by clicking the relevant icon in the *events bar* located under the QC charts. So everyone can easily access the information needed to troubleshoot an out-of-control situation of a bad EQA return.



Analytical events of MultiQC

- Reagent blank
- Calibration
- New reagent lot
- Comment
- Linearity checking
- Test of repeatability
- Method comparison

8. Comments

➤ The Bland-Altman plot

Bland and Altman published a paper in *The Lancet* (1986) which popularized the difference-plot among clinical researchers. MultiQC does not implement the difference-plot exactly as it was described by Bland and Altman because it is not well adapted to clinical chemistry data :

- The reportable range of analytical methods is much wider than the range of clinical measurements. Range ratios in routine clinical chemistry generally exceed 10. So the hypothesis of homoscedasticity is far from being fulfilled. A log transformation would be always necessary, making difficult reading of the plots.
- The bias is rarely constant over the whole reportable range. Most often, differences of calibration create a proportional bias that depends on concentration. So the average bias has no practical meaning.
- Allowed error schemes are more complex in clinical chemistry than the simple agreement limits of Bland and Altman.

➤ A prerequisite to the difference-plot

The difference-plot implemented in MultiQC compares deviations between analytical methods to medical tolerance taken as criterion of judgement. This comparison is senseless if each method does not individually meet the criterion. So a prerequisite before any method comparison through a scatter-plot is to check that the capability index of each method under evaluation is greater than 1.

The famous paper by Bland and Altman in *The Lancet* (1986) provides an excellent example of such a mis-interpretation of a difference-plot. The authors compare two flowmeters, a large one and a mini one. They write that they would agree to interchange the flowmeters if these meters “were unlikely to give readings which differed by more than, say, 10 l/min”. They also write later in the paper “The coefficient of repeatability is 56.4 l/min for the mini meter. For the large meter the coefficient is 43.2 l/min”. How can Bland and Altman demand that differences between large and mini meters do not exceed 10 ml/mn whilst replicates of each meter can differ by more than 50 l/mn ! The conclusion of Bland and Altman was that the large and mini flowmeters cannot be interchanged. A more realistic end would be : both devices must be discarded because each one is unable to meet the error specification of 10 l/mn.

➤ Correlation coefficient

Correlation coefficient use is inappropriate for comparing analytical methods:

- The correlation coefficient measures the strength of the relation between two variables, not the agreement between them. Two analytical methods may be highly correlated whereas a huge bias makes them completely incompatible.
- The magnitude of the correlation coefficient is affected by the range of concentrations studied. The correlation coefficient can be made smaller by measuring samples that are similar to each other and larger by measuring samples that are very different from each other.

Correlation coefficient has no place in method comparisons because it does not answer the actual question of agreement and has an arbitrary value depending on the choice of analytical samples. Correlation coefficient is therefore not displayed by MultiQC.

9. References

- MultiQC, quality control software for clinical chemistry : www.multiqc.com
- Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement: <http://www-users.york.ac.uk/~mb55/meas/ba.htm>