Assessment of Bias with Emphasis on Method Comparison

Roger Johnson
Department of Chemical Pathology, LabPlus, Auckland City Hospital, Auckland, New Zealand.
For correspondence: Dr Roger Johnson e-mail: rogerj@adhb.govt.nz

Summary

- Definition of bias - distinct from accuracy, bias is an average deviation from a true value.
- Method comparison - a set of specimens is assayed by both an existing method and the new candidate method, and the results compared. The following list describes the testing procedures and data handling required in a method comparison study for the assessment of bias:
  - Test material
  - Number and disposition of specimens
  - Summary of findings
  - The problem with correlation, and the difference plot
  - Statistics of difference
  - Log transformation of the difference plot
  - Statistics of difference with logs
  - Linear regression
  - Deming and Passing-Bablok models
  - The value of *r* in linear regression
  - Choice of statistics
  - Examples of suitable computer programs
- Acceptable bias criteria are discussed.
- Linearity and recovery - failing either of these criteria should serve as a warning that method comparison data may conceal an unrecognised bias.
- Finally, consideration of all steps in the assessment of bias is required to determine acceptability or not of the method comparison.

Definition of Bias

"Bias is used to express numerically the degree of trueness", trueness being "the closeness of agreement between the average value obtained from a large series of measurements and the true value".[1]

"Bias" and "inaccuracy" are often used synonymously. However, contemporary usage by ISO and CLSI[1] makes a distinction between these terms: inaccuracy relates to how closely a *single* measurement agrees with the true value whereas bias relates to how *an average of a series* of measurements agrees with the true value. In the first case, imprecision contributes to the lack of agreement whereas in the second, imprecision is minimised (ideally removed entirely) by use of an average.

Introduction

AACB members may be familiar with the paper on method evaluation by Nick Balazs and Des Geary and published in 1981,[2] fittingly as a technical report from the AACB "Scientific and Technical Committee". Their recommendations, used as a basis for method evaluation in my own laboratory for more than 20 years, addressed "inaccuracy" (what we should now call bias), making global assessments by method comparison, and separately assessing interference, linearity and recovery.

Interference is considered elsewhere in this issue (see Interference Testing in this issue). Linearity and recovery will be covered briefly but first method comparison will be discussed in more detail.

Method Comparison

Test Material

The cornerstone of many method evaluations is a method comparison in which a set of specimens is assayed by both an existing method and the new candidate method, and the results compared. For reasons of suitability and convenience, the specimens used are often excess patient specimens. In this case, they will have no known value other than that found in the existing assay which itself may have shortcomings. For this reason, it is informative to include specimens of known value which may be external quality assurance specimens, possibly from an RCPA QAP scheme or from reference sources such as the Centers for Disease Control and Prevention (CDC), or National Institute of Standards and Technology (NIST). Disadvantages of these sorts of specimen are that the matrix may be inappropriate and that costs may be significant for some materials.

Number and Disposition of Specimens

The number of specimens need not be large (for example, CLSI suggest 20,[3] although a more thorough investigation requires 40 or more[4]). The more critical aspects are that they span the range of interest and are determined with greater certainty than might be done routinely, by using multiple determinations, at least duplicates. Comparing multiple small batches with the two procedures run at the same time over several days is preferred to single larger runs,[3,4] as between-day variations can be accommodated.

Summary of Findings

The data should be displayed on an x-y plot, with the results from the existing method plotted on the x axis (conventionally the independent variable) and those from the candidate method plotted as y (Figure 1A). Inspection by eye may reveal aberrant points or nonlinear behaviour that may warrant further investigation. But beyond that, opinions differ as to how the data should be analysed.

The Problem with Correlation, and the Difference Plot

Least squares linear regression analysis (for example in Microsoft® Excel) calculates among other things a correlation coefficient, *r*, which has everything to do with scatter about the line (commonly representing imprecision) and nothing to do with agreement. Undue focus on this statistic alone as a measure of a candidate assay's worth has been rightly criticised, and caused *Annals of Clinical Biochemistry* to ban its presentation.[5] *Annals*' editors favoured the difference plot[6] in which differences between the comparison estimates are plotted against the mean of their values. The distinction in approach is illustrated in Figure 1: the *r* value shows perfect correlation in the presence of perfect disagreement. The disagreement can be hard to see in x-y plots (e.g. Part A) but the difference plot (Part B) emphasises the lack of agreement as a consistent difference. The difference plot allows a more sensitive visual review of the data than is possible with an x-y plot.

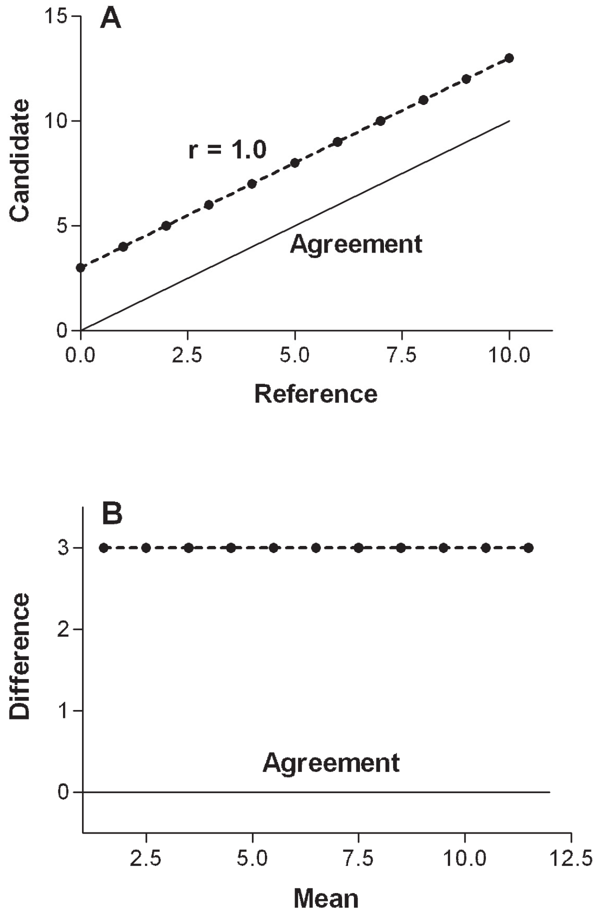

Figure 1. Method comparison. **A.** Conventional x-y plot with perfect correlation (r = 1.0); the lack of agreement between the two sets of data may go unremarked unless the line of agreement is also shown. **B.** The difference plot highlights the lack of agreement immediately and encourages statistical analysis of the difference.

Statistics of Difference

If the data as displayed in a difference plot shows even scatter at different concentrations, the difference plot is amenable to statistical analysis in which the bias between the two sets of results can be described by a mean and SEM.[5] If the 95 % confidence interval for the *mean* difference (mean ±2 SEM) includes zero, a statistician would say that there is no evidence of bias.

Log Transformation of the Difference Plot

The example shown in Figure 1 is of constant or systematic bias and is unusual in clinical chemistry. More common is the situation in Figure 2A. Here there is a progressive deviation with concentration, a proportional bias, in which the difference plot (Figure 2B) is unhelpful until the data are transformed. One possible transformation is by plotting proportional rather than absolute differences.[7] This plot has a familiar feel but the derived data are inherently non-Gaussian. The transformation preferred by statisticians is to convert the experimental data to logarithms and then proceed *exactly* as before[5] (Figure 2C).

Statistics of Difference with Logs

The log data have to be transformed back to be intelligible. The mean difference of about 0.08 in Figure 2C shows that the slope of the line is $10^{0.08}$, i.e. 1.2. If calculation of the mean ±2 SEM of the log data includes zero, an argument is provided for accepting the slope as 10^0, i.e. 1.0, and therefore without proportional bias. However, to be valid such calculations require a Gaussian distribution of data.[8]

Linear Regression

When the comparison data contain elements of both systematic and proportional bias, the difference plot whether direct or transformed can be difficult to interpret, and some form of linear regression may give a clearer result. But which model to use? Least squares linear regression (as done in Microsoft® Excel) considers error only in the "y" direction and minimises this component in constructing the line. Invariably "x" is also subject to error, a fact that becomes clear if x and y are reversed: the regression line in this case is not equivalent to the first because the errors minimised are not the same.[5]

Deming and Passing-Bablok Models

Two models often used to overcome this difficulty are those of Deming and of Passing and Bablok. The first takes into account variability in both x and y whereas the second is a non-parametric approach in which the median slope of all possible lines between individual data points is found. Both approaches have their champions.

The Value of r *in Linear Regression*

Yet another argument proposes that the regular least squares approach is valid provided that the line is sufficiently well-defined.[9] Definition in this case is judged by *r* being high (>0.975 for values spanning one decade; >0.99 for values spanning three decades because *r* is affected by the range of values). In fact, the authors suggest that if *r* is not sufficiently high (i.e. below the cut-off values mentioned), either more data need to be collected or existing data need more careful scrutiny.

Choice of Statistics

Given this array of techniques and our generally amateurish knowledge of their validity, what is the safest approach? Westgard has recommended that considering the ease with which data can be manipulated by computer, many different techniques should be applied, and: "*When in doubt about the validity of the statistical technique, see whether the choice of statistics changes the outcome or decision on acceptability*".[10]

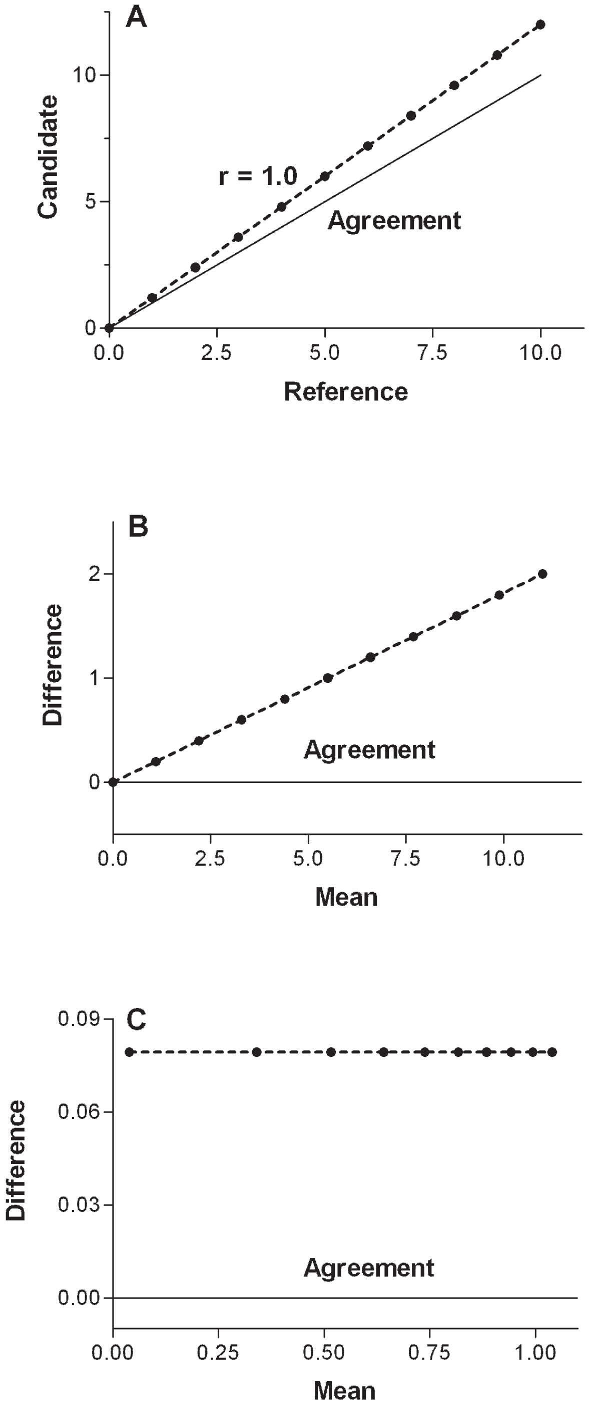

Figure 2. Method comparison. **A**. Conventional x-y plot with perfect correlation (*r* = 1.0); as in Figure 1, the lack of agreement between the two sets of data may go unremarked unless the line of agreement is also shown. **B.** The difference plot shows a lack of agreement that changes as a proportion of the mean. **C**. The data in B have been replotted after first taking logs of the data sets and then recalculating the differences and means. The constant difference shown (about 0.08) means that y differs from x by $10^{0.08}$, i.e. 1.2-fold.

Examples of Suitable Computer Programs

In our laboratory, we use a version of method comparison software now marketed as MultiQC[11] which allows easy transition between difference plot, linear regression, Deming and Passing-Bablok models, so that Westgard's advice can easily be followed. I have seen similar results from Analyse-it.[12] Both websites have instructive animations that offer advice in use of the respective programs, and both allow for a free trial of the software. A Google™ search reveals other companies that may offer similar programs.

Acceptable Bias

The quotation from Westgard[10] (above) raises the question of what is acceptable bias. Clearly if no analytical goal is decided before a comparison is done, the exercise is purely descriptive. So what is an appropriate goal? Biological variation offers a realistic approach based on population data. The underlying consideration is that bias causes more than the expected 5% of a reference population's results to fall outside a pre-determined (95%) reference interval. By limiting bias to no more than a quarter of the reference group's biological variation, the proportion outside the reference interval is restricted to no more than 5.8% (a relative increase of 16% over the expected 5%), and is judged a "desirable" standard of performance.[13]

The limits on bias provided on Westgard's website[14] are for desirable performance; "optimum" and "minimum" performance standards are also recognised, respectively 50% and 150% of desirable.[13] This means that for a desirable bias of 4%, optimally it should be 2% and at worst no more than 6%.

If a new method is being introduced and the bias compared to the old method exceeds an acceptable limit, then the reference interval should be reviewed and clinicians notified that the results may be different to those previously issued.

For particular cut-points (e.g. as with plasma glucose concentration in defining diabetes), deviation at these points is of more concern than an average deviation over the full range of the assay.

Linearity

Whatever the shape of the calibration line, the expectation is that a concentration (or activity) of analyte should be matched by its assay result. Any limitation on the linearity of this relationship can be assessed by selecting a specimen with a high concentration of analyte and mixing it in linearly related proportions with one containing a low concentration: suitable mixtures contain 0, 10, 20 …. up to 100 % of the high specimen, giving 11 specimens to test. The high concentration should exceed the expected or desired upper limit of the assay to test whether as an ideal linearity extends beyond that point; the low concentration does not have to be zero although a clearer result can be expected the closer it is to zero. Analysis of these specimens should be at least in duplicate to lessen variation.

An x-y plot of concentration against proportion of high specimen is then drawn and inspected by eye when non-linearity may be evident (Figure 3A). An upper limit should be no higher than the highest concentration (or activity) that falls on the apparently linear segment. In fact it may be prudent to select an even lower concentration to allow for sub-optimal performance in routine use. Choosing a lower limit follows similar reasoning. (See article by Armbruster and Pry in this issue.)

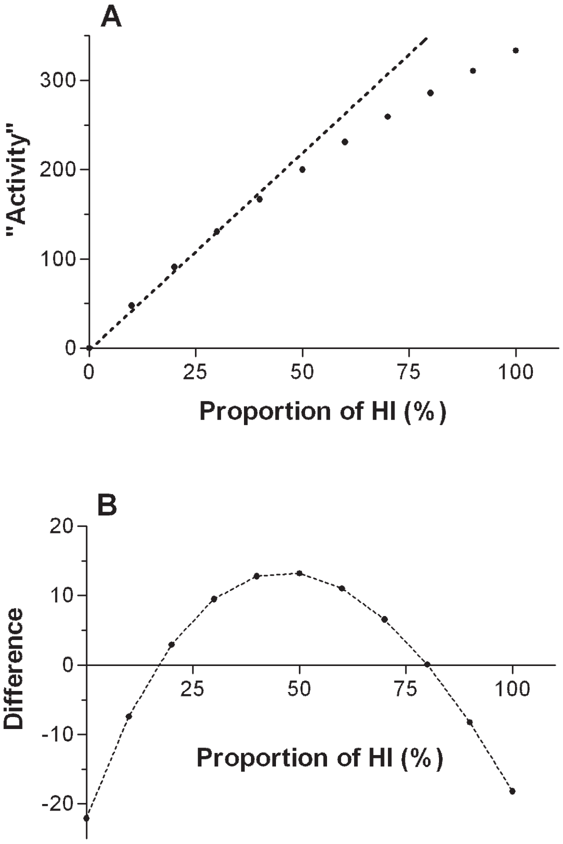

Figure 3. Assessment of linearity. **A**. Conventional x-y plot for assessing linearity by eye. The curved response might be made "linear" by restricting the measuring range to no more than 150 units. **B**. Residual plot, residuals being the individual differences between the experimental results (y values) and the results predicted from linear regression reveals the non-linear behaviour as a continuum. Note that residuals are negative at low proportions of "HI" because the regression line has a positive intercept on the y axis.

An alternative means of data analysis is by residual plot (in effect a difference plot): residuals are differences between the actual values found and those predicted for them from a least squares regression line. Curvature is suggested by the shape of this plot, or in less clear cases the sign sequence of the residuals,[9] a greater sensitivity compared with the direct plot being explained by the finer scale on the ordinate axis (Figure 3B).

Whether this performance is useful or can be made more satisfactory by using a restricted range should be considered in relation to acceptable bias (above).

Recovery

Measurement of recovery involves the assay of exogenous analyte in the specimen matrix. In essence, a base material (e.g. serum) is assayed before and after addition of a known amount of analyte (often called spiking). The difference in concentration between these measurements should equate to the known amount added.

This test is useful in deciding whether calibrators need to be made in a matrix more closely resembling the specimens to be analysed, or whether some interference that needs further investigation is present. It is not straightforward to do, however. It requires: analyte in a suitably concentrated form so that addition causes minimal disruption of the base material (≤10% by volume and with consideration of the solvent used); measurements with low imprecision (duplicate determinations at least); and addition sufficient to make a measurable difference without exceeding the assay range. The last of these requirements may be difficult to achieve when using random specimens, possibly with unknown content of analyte.

Experience suggests that the arithmetic associated with this sort of experiment can be demanding unless the base material is subject to a blank (solvent) addition to account for the dilution that occurs with spiking. Subtraction of the value in the adjusted base from that in the spiked material gives the added concentration directly and hence the amount of the spike recovered. Recovery is then [Final (Spike) Concentration - Initial (Base) Concentration]/Added Concentration.

Acceptability of recovery can be judged against Logan's criteria,[15] although nowadays we should take note of acceptable bias too. Exceeding either of these criteria should serve as a warning that method comparison data may conceal an unrecognised bias.

Summary of steps in assessment of bias

1. Criteria of acceptable performance established
2. Comparison of test method with reference method using patient material ± reference material
3. x-y Plot of data with examination by eye
4. Consideration of difference plot and statistics of difference
5. Consideration of regression analysis and statistics of regression
6. Test of interference
7. Test of linearity
8. Test of recovery
9. Judgement of acceptability

Competing Interests: None declared.

References

1. Tate J, Panteghini M. Standardisation – The theory and the practice. Clin Biochem Rev 2007;28:127-30.
2. Balazs ND, Geary TD. Guidelines for the selection and evaluation of analytical methods. Clin Biochem Rev 1980;1:51-7.
3. Clinical and Laboratory Standards Institute. User verification of performance for precision and trueness; approved guideline - second edition. CLSI document EP15-A2. Wayne, PA, USA: CLSI; 2005.
4. Clinical and Laboratory Standards Institute. Method comparison and bias estimation using patient samples; approved guideline - second edition. CLSI document EP9-A2. Wayne, PA, USA: CLSI; 2002.
5. Hollis S. Analysis of method comparison studies. Ann Clin Biochem 1996;33:1-4.
6. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307-10.
7. Pollock MA, Jefferson SG, Kane JW, Lomax K, MacKinnon G, Winnard CB. Method comparison - a different approach. Ann Clin Biochem 1992;29:556-60.
8. Twomey PJ. How to use difference plots in quantitative method comparison studies. Ann Clin Biochem 2006;43:124-9.
9. Stockl D, Dewitte K, Thienpont LM. Validity of linear regression in method comparison studies: is it limited by the statistical model or the quality of the analytical input data? Clin Chem 1998;44:2340-6.
10. Westgard JO. Points of care in using statistics in method comparison studies. Clin Chem 1998;44:2240-2.
11. MultiQC, Medical Laboratory Quality Control Software. www.multiqc.com (Accessed 27 December 2007).
12. Analyse-it® Statistical analysis add-in software for

Microsoft Excel. www.analyse-it.com (Accessed 27 December 2007).

13. Fraser CG. Biological Variation: From Principles to Practice. Washington DC, USA: AACC Press; 2001.p. 52-5.

14. Ricós C, García-Lario J-V, Alvarez V, Cava F, Domenech M, Hernández A, et al. Biological variation database, and quality specifications for imprecision, bias and total error. The 2006 update. http://www.westgard.com/guest32.htm (Accessed 27 December 2007).

15. Logan JE. Evaluation of commercial kits. CRC Crit Rev Clin Lab Sci 1972;3:271-89.

Appendix: An example of data handling and interpretation in method comparison studies.
Please see (http://www.aacb.asn.au/web/Resources/Tools/).

The data are presented in a Microsoft Excel spreadsheet format under several tabs best viewed in the order given:

Data – representative comparison data and scatter plot
Difference – calculation and presentation of the difference plot
log Difference – log transformation of the data and calculation of a log difference plot
Statistics of log Difference – calculation of error and limits of error
Regression – least squares presentation with limits on slope and intercept
Summary – comparison with acceptable bias of statistical data calculated here, and for Deming and Passing-Bablok models