VERSITA
DOI:10.2478/bile-2013-0003

Biometrical Letters
Vol. 49 (2012), No. 1, 45-56

Comparison of clinical data based on limits of agreement

Luís M. Grilo, Helena L. Grilo

Unidade Departamental de Matemática, Instituto Politécnico de Tomar,
Estrada da Serra – Quinta do Contador, 2300-313 Tomar, Portugal,
lgrilo@ipt.pt, helenagrilo@ipt.pt

SUMMARY

Two different medical measurement methods, which usually do not produce exactly the same results, are used to analyse the serum levels of folic acid in a blood sample. We assess the (dis)agreement of the available data in order to replace the old method (the reference method, which involves a lot of human intervention) with the new one (which uses mostly machines), without causing problems in clinical interpretation. The 95% limits of agreement are estimated, before and after a logarithmic transformation, and an appropriate use of regression and a nonparametric approach are also considered. The application of these different statistical techniques is very useful and easily interpreted by medical researchers, but the results obtained do not provide confidence that the new method can be used in place of the old one for clinical purposes.

Key words: Measurement methods, graphical techniques, logarithm transformation, regression approach, nonparametric approach.

1. Introduction

During a period of time, patients with different diagnoses (e.g. anaemia, encephalopathy, HIV, lymphoma, stroke problems, etc.) had taken folic acid with the aim of improving their general health. Subsequently, blood samples were collected and the continuous variable which represents the serum levels of folic acid in the blood (nanograms per millilitre, ng/ml) was measured using two clinical methods: RIA – Radio Immune Analysis (the reference method, which involves a lot of human intervention) and IMM – Immunolite (which is newer and uses mostly machines). In Table 1 we give the available measurements using RIA and IMM for a sample size of n = 68 individuals.

Table 1. Serum levels of folic acid in the blood (ng/ml), obtained by two different clinical methods (RIA and IMM)*

Subject	RIA	IMM	Subject	RIA	IMM
1	2.86	3.5	35	4.1	4.5
2	7.9	6.57	36	1.65	2.1
3	9.7	9.14	37	7.59	6.7
4	5	4.22	38	3.61	3.48
5	1.21	2.18	39	11.17	11.1
6	3	2.46	40	4.34	3.96
7	1.72	1.6	41	5.11	4.5
8	2.16	2	42	5.31	2.89
9	2.87	3.42	43	4.23	2.14
10	7.9	4	44	3.14	3.46
11	1.34	1.47	45	12.4	8.5
12	4.2	4.29	46	6.42	5.87
13	2.1	2.3	47	2.31	2.51
14	1.4	1.65	48	17.1	12.3
15	16.4	12.1	49	1.22	1.62
16	2.3	1.97	50	2.4	1.97
17	3	2.87	51	3.17	2.74
18	1.9	2.2	52	1.82	1.76
19	5.6	3	53	4.7	3.42
20	3	3.4	54	10.4	5.5
21	10.8	11.9	55	6.6	5.87
22	3.48	3.1	56	3.2	3.69
23	5.63	4.1	57	2.69	1.72
24	4.58	3.46	58	9.9	4.89
25	3.8	4.41	59	5.3	5.93
26	4.5	4.42	60	2.3	2.4
27	1.76	0.95	61	11	8.9
28	1.65	1.38	62	19.1	11.2
29	4.82	3.07	63	2.2	2.3
30	3.2	3.1	64	4.4	3.3
31	0.91	0.62	65	1.5	1.28
32	3.82	4.39	66	8	6.9
33	1.75	2.16	67	9.3	6.79
34	5.5	6.7	68	3.1	2.45

* The data set was kindly provided by a clinical laboratory at a Portuguese hospital

We intend to evaluate how significant are the discrepancies between the measurements obtained using the two methods.

To analyse the agreement between medical measurements obtained by different clinical methods, several papers have used the Pearson correlation coefficient (which is not a measure of agreement, but a measure of association) and linear regression (which ignores the fact that both dependent and independent variables are measured with error); these statistical techniques can be misleading and inappropriate (see Altman and Bland, 1983; Bland and Altman, 2003, 1999, 1986). Thus we analyse the data set using graphical techniques[1] (involving simple statistical calculations, to determine 95% limits of agreement and confidence intervals), and also with an appropriate use of regression in order to quantify the (dis)agreement between the two methods (see Altman and Bland, 1983; Bland and Altman, 2007, 2003, 1999, 1986).

2. Statistical Techniques

To measure the agreement between clinical methods RIA and IMM, we estimate, in section 2.1, the 95% limits of agreement before and after the logarithm transformation of the data. In section 2.2 we apply a more general method used when the log transformation does not entirely solve the problem of complex variation across the range of measurement.

2.1. Limits of agreement approach

Examining observations relative to the identity line (RIA = IMM) in the scatter plot of Figure 1, where method RIA is plotted on the *x*-axis and method IMM on the *y*-axis, we detect some dispersion of observations around the line which is not constant across the range of measurement (non-constant variance), and also a clear bias with the majority of observations lying to one side of the equality line (proportional bias).

---

[1] These techniques are available in the Analyse-it Method Evaluation package for Microsoft Excel.

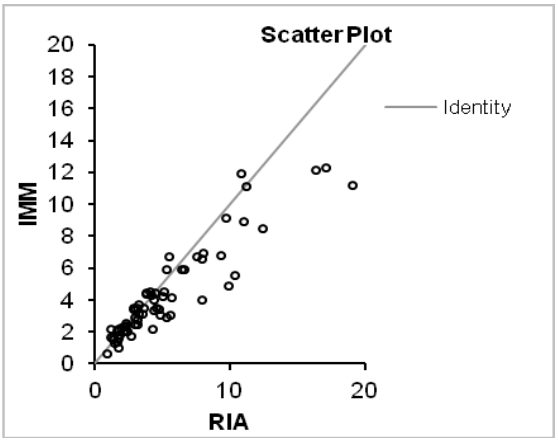

Figure 1. Serum levels of folic acid in the blood (ng/ml) measured by RIA and IMM, with the line of equality

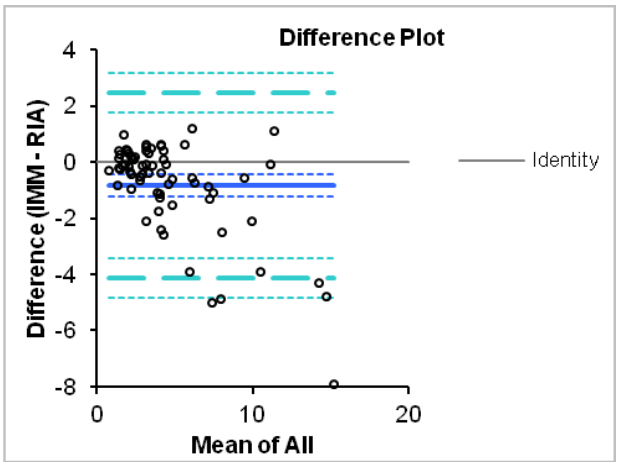

Figure 2. Serum levels of folic acid in the blood (ng/ml): difference (IMM - RIA) *vs.* average of values measured

To identify differences between these two alternative clinical methods, we also use the difference plot in Figure 2 (this informative plot shows the difference between the methods, d, on the vertical axis, plotted against the best estimation of the true value – the mean of observations from both methods, on the horizontal axis). This plot also shows 95% limits of agreement and confidence intervals for the bias and for the limits of agreement, which enables us to

analyse the relationship between the difference and the magnitude of measurement. The scatter of differences around the zero line is not constant – the differences tend to be negative, especially for high levels of folic acid. The mean and standard deviation of the differences are not constant, i.e. they depend on the magnitude of the measurement. Based on the limits of agreement we can confirm the underestimation of the IMM method. The limits seem to have a large range for low values of the mean and a small range for high values of the mean.

In Table 2 we give the mean differences, $\bar{d}$ (to estimate the bias, which ideally should be zero), and the standard deviation of differences, s'_d (to estimate the variation about $\bar{d}$), both to estimate the 95% limits of agreement $(\bar{d} \pm 1{,}96 \times s'_d)$ shown in Figure 2 (these provide an interval within which 95% of the differences between measurements by the two methods are expected to lie, if the differences are normally distributed). When we have large variation of differences, the limits of agreement are not small enough, which indicates some lack of agreement. Here we have four differences outside the limits of agreement, which corresponds to $(4/68) \times 100 \approx 5.9\% > 5\%$ of differences.

Table 2. The 95% limits of agreements

Mean differences	-0.821 ng/ml
Standard deviation of differences	1.689 ng/ml
95% limits of agreement	from -4.130 ng/ml to 2.,489 ng/ml

In Table 3 we give the standard error of $\bar{d}$ (the standard deviation of $\bar{d}$ is estimated by $s'_d / \sqrt{n}$) used to estimate the 95% confidence intervals for the bias, and the standard error of the limits of agreement, $(\bar{d} \pm 1.96 \times s'_d)$, which is about $\sqrt{3s'^2_d / n}$, to estimate the 95% confidence intervals for the limits of agreement. In this case study we may note that the 95% confidence interval for the bias does not contain zero.

Table 3. The 95% confidence intervals for bias and for the limits of agreement

Standard error of mean of differences	0.205 ng/ml
95% confidence interval for the bias (for $n - 1 = 67$ degrees of freedom $t \approx 1.99$)	from -1.228 ng/ml to -0.413 ng/ml
Standard error of limits of agreement	0.355 ng/ml
95% confidence interval for the lower limit of agreement	from -4.836 ng/ml to -3.424 ng/ml
95% confidence interval for the upper limit of agreement	from 1.783 ng/ml to 3.195 ng/ml

We use the histogram of differences, Figure 3, to check the assumption of normality. The distribution of differences is skewed, and therefore does not match the normal curve (which can happen when there is a relation between differences and mean). Nevertheless we can estimate the limits of agreements, because this fact is not a serious problem in this context (Bland and Altman, 2003, 1999).

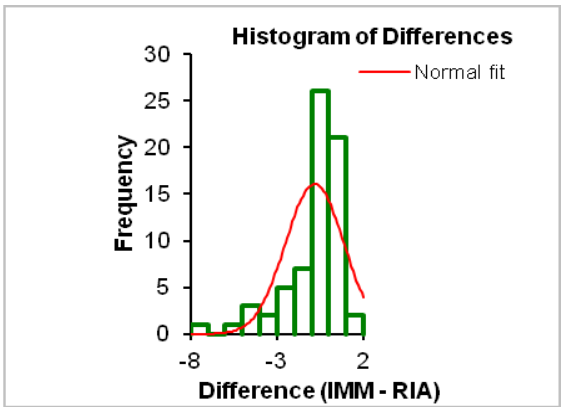

Figure 3. Histogram of differences (IMM - RIA) with normal curve

When the difference between the measurements by the two clinical methods is related to the magnitude of the measurement, which is a common situation, we should try to remove this relationship. We use a simple logarithmic transformation of the data, which allows the results to be interpreted in relation to the original data. We can back-transform the limits of agreement from log-transformed data to obtain limits related to the ratios of measurements by the two methods (Bland and Altman, 1999). Figures 4 and 5 show that the log-

transformed data bring some improvement, although the relation between the difference and the mean still remains.

In Table 4 we give the mean and the standard deviation of differences, used to estimate the 95% limits of agreement, after log transformation. To get the limits of agreement on the original scale, we take the anti-logs of these limits, obtaining 0.521 and 1.500.

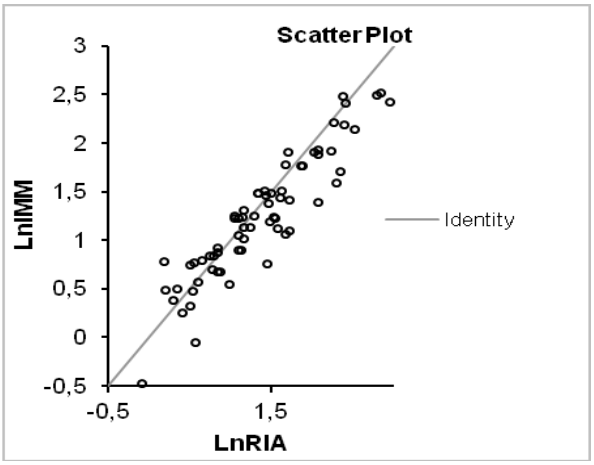

Figure 4. Measurements of folic acid in the blood after log transformation, with the identity line.

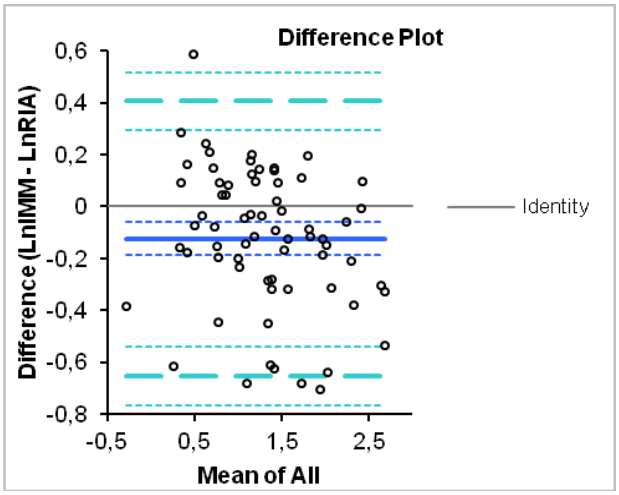

Figure 5. Difference between folic acid measurements plotted against average after log transformation, with 95% limits of agreement

Table 4. The 95% limits of agreement, after log transformation

Mean differences	-0.123
Standard deviation of differences	0.270
95% limits of agreement	from -0.652 to 0.405

Figure 6 shows that, as expected, the distribution of the differences, after log transformation, is approximately normal.

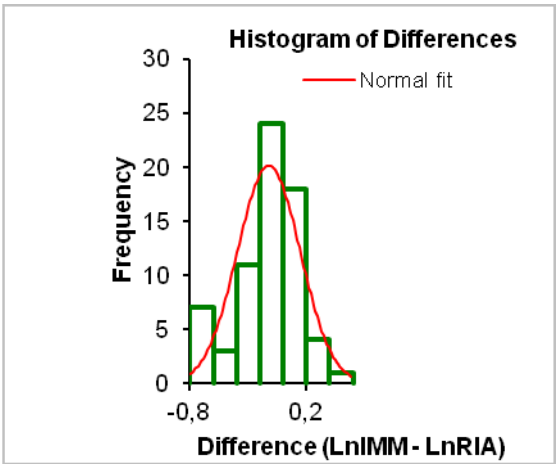

Figure 6. Histogram of the differences after log transformation, with normal curve

2.2. Regression approach

According to Bland and Altman (1999) we should apply a regression approach to evaluate the agreement when the relationship between differences and the size of measurement remains after the log transformation. Thus let D be the difference between the two methods and A the average of both methods (RIA and IMM); the regression of differences on average gives a highly significant relationship (p-value = 0.00):

$$\hat{D} = 0.7162 - 0.3321A$$

and can be used to model the relationship between mean differences and the magnitude of the serum levels of folic acid.

To model the relationship between the standard deviation of the differences and the magnitude of the levels of folic acid, we regress the absolute values of the residuals (R) on A,

$$\hat{R} = 0.2085 + 0.1469A$$

which is a statistically significant regression (p-value = 0.00).

Considering a normal distribution with mean zero and variance σ^2, it is shown that the mean of the absolute values is $\sigma\sqrt{\frac{2}{\pi}}$, which follows a half-normal distribution. Therefore, the predicted standard deviation of the differences (S_D) is the product of the fitted values by $\sqrt{\frac{\pi}{2}}$,

$$\hat{S}_D = 0.261316 + 0.184112A.$$

Taking into account the above regression equations we obtain the 95% limits of agreements,

$$\hat{D} \pm 1.96 \times \hat{S}_D.$$

Then, for this sample, we calculate:

$$Lower\ Limit = (0.7162 - 0.3321A) - 1.96 \times (0.261316 + 0.184112A)$$
$$Upper\ Limit = (0.7162 - 0.3321A) + 1.96 \times (0.261316 + 0.184112A)$$

Based on this regression approach, the fit is greatly improved, particularly for high levels of folic acid, as shown in Figure 7. However, although all the observations lie between the 95% limits of agreement, we still indentify a bias and an increase in the variance together with the magnitude of the observations.

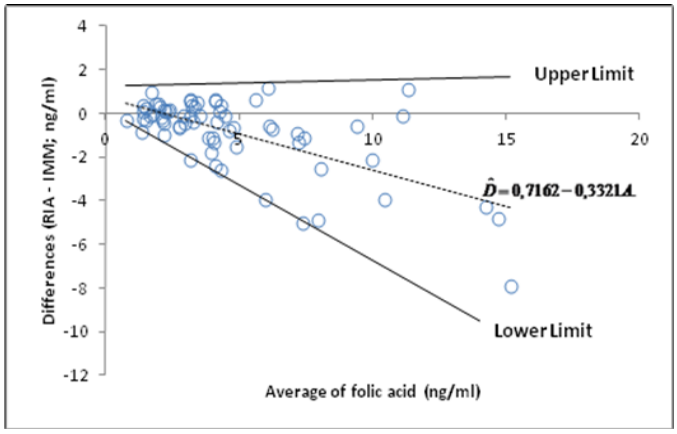

Figure 7. Limits of agreement for differences in folic acid in the blood, measured by the RIA and IMM methods (ng/ml), based on regression

2.3. Nonparametric approach

For the cases where the between-method differences do not have normal distribution and/or there are one or more extreme discrepancies between the clinical methods, a simple nonparametric approach is also mentioned by Bland and Altman (1999). Here, based on the sample size (n = 68) we consider the percentiles 5 and 95, which are superimposed on the scatter diagram in Figure 8. The proportion of differences between these two percentiles is

$$\hat{p} = \frac{60}{68} \times 100 \approx 88.2\%$$

and, with the estimated proportion, we construct a 95% binomial confidence interval, which contains the true proportion parameter 95% of the times the procedure for constructing the confidence interval is employed. The common formula for a binomial interval relies on approximating the binomial with a normal distribution, which is justified by the central limit theorem. So applying the general formula

$$\hat{p} \pm 1.96 \times \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}} = 0.882 \pm 0.077$$

we can conclude that between 80.1 and 95.9 percent of differences lie between percentiles 5 and 95 or, in other words, with a margin of error of 0.077 we have 88.2% of differences between the 5th and 95th percentiles.

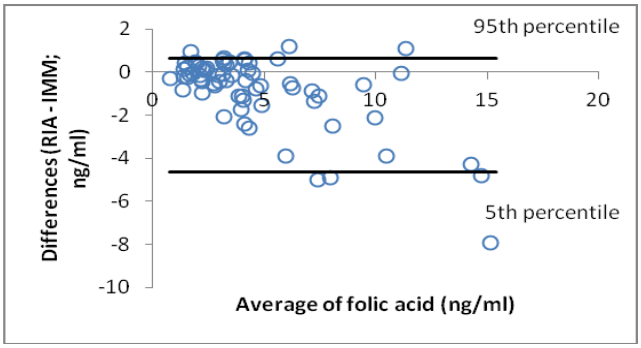

Figure 8. Difference between folic acid measurements plotted against average, with percentiles 5 and 95

3. Conclusions and final remarks

Based on graphical analysis and analytical results (range of the estimated limits of agreements, variation of differences and percentage of differences outside the agreement limits, value of bias and 95% confidence intervals), we are not confident that the new clinical method (IMM) can be used in place of the old (RIA) for clinical purposes. We have to emphasize that the values obtained with one method, in several cases, almost double the values using the other method.

Although the hypothesis tests may not seem so appropriate for assessing the (dis)agreement between the two methods, based on *p*-values obtained with the application of the *t*-Student test for paired samples, we reject the null hypotheses (i.e. that the mean of both methods are significantly different). Even taking the logarithm of the variables to obtain approximately normal variables (because the distributions of variables are skewed and the assumption of normality, according to the Kolmogorov-Smirnov test, is not valid), the decision is the same. Using the nonparametric Wilcoxon test, based on paired samples to compare the two population medians, the decision is still to reject the null

hypotheses (whether or not four outliers representing large discrepancies are excluded).

To evaluate the degree of agreement of medical measurements there are other interesting procedures for cases where repeated measurements are available (Carstensen *et al.*, 2008) or under non-standard conditions (Choudhary and Tony-NG, 2006), but these techniques are relatively more complex and therefore less attractive to medical researchers than those used in this study, which are easier to apply and give results that are simpler to interpret and very useful in practice.

The decision about what is an acceptable agreement is a clinical judgement. Clinicians believed that, in spite of some inevitable lack of agreement, both clinical methods (RIA and IMM) were interchangeable, but the results obtained here cast doubt on that conclusion.

REFERENCES

Altman D., Bland J. (1983): Measurement in medicine: the analysis of method comparison studies. Statistician 32: 307-17.

Bland J., Altman D. (2007): Agreement Between Methods of Measurement with Multiple Observations Per Individual. Journal of Biopharmaceutical Statistics 17: 571-582.

Bland J., Altman D. (2003): Applying the Right Statistics: Analyses of Measurement Studies. Ultrasound in Obstetrics and Gynecology 22: 85-93.

Bland J., Altman D. (1999): Measuring agreement in method comparison Studies. Statistical Methods in Medical Research 8: 135-160.

Bland J., Altman D. (1986): Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet: 307-310.

Carstensen B., Simpson J., Gurrin L. (2008): Statistical Models for Assessing Agreement in Method Comparison Studies with Replicate Measurements. The International Journal of Biostatistics 4(1): Article 16.

Choudhary P.K., Tony-NG H.K. (2006): Assessment of Agreement under Nonstandard Conditions Using Regression Models for Mean and Variance. Biometrics 62: 288-296.