Allowable Total Error and Limits for Erroneous Results (ATE/LER) Zones for Agreement Measurement

Yu-Hao Zhou, Jia-Jie Zang, Mei-Jing Wu, Jin-Fang Xu, and Jia He*

Department of Health Statistics, Second Military Medical University, Shanghai, China

Background: When assessing the agreement of two methods, the simple correlation coefficient, the most frequently used but easy to cause misleading interpretation of results, is still comprehensively adopted in journals of clinical chemistry or diagnostic laboratory. This study is aimed at comparing the correlation coefficient with the alternative method, which came from FDA recommended guidance for waiver assays—Allowable Total Error (ATE)/Limits for Erroneous Results (LER) zones, and fully addressing the applicable conditions of the later one. **Methods:** The data we use were both real and simulated data to calculate correlation coefficient and ATE/LER zones, respectively, when assessing the agreement. And a modified ATE/LER

zones method was brought forward to assess the agreement of repeated measurements and the repeatability of each single method. **Results:** The high correlation can only indicate that the two methods are linearly related rather than in agreement with each other. However, the ATE/LER zones can avoid this shortcoming by using certain statistical strategies considering the clinical judgment. The modified method is suitable to evaluate the repeatability of single method and the agreement of repeated measurement by two measuring methods. **Conclusion:** The ATE/LER zones method is a relatively reasonable and widely considering method for assessing the agreement. *J. Clin. Lab. Anal.* 25:83–89, 2011. © 2011 Wiley-Liss, Inc.

Key words: agreement; correlation coefficients; ATE/LER zones

INTRODUCTION

Measurement agreement has been the focus of much discussion and research with respect to its roles in clinical application. In the clinical practice, it is difficult, even impossible, to measure some medical measurements directly without any adverse effect. Under these circumstances, an alternative method might be used to avoid these shortcomings, and the agreement between the new method and the established one should be evaluated before the new one come into widely use.

There is no gold standard for some measurements in the medical analysis. A new method has to be evaluated by comparison with an established technique rather than with the true value. What should we do if we encounter this kind of problem?

Most studies use the product–moment correlation coefficient (r) between the results of the new method and the established one as an indicator of agreement (1). However, the meaning of r is usually misunderstood, which only represents the relation between two variables, not the agreement of each other. Considering the drawbacks of correlation coefficient, FDA suggested an alternative approach—Allowable Total Error (ATE)/

Limits for Erroneous Results (LER) zones to get more accurate results, based on a graphical technique and simple calculation, and the different scores of two measurements were calculated (2).

In this study, the comparison of the new approach with the traditional one for assessing agreement measurement was under careful consideration. And the practicability of the new approach, as well as its applicable conditions, was fully elaborated. The agreement measurement for repeated measurements was also discussed. Most of the analysis in our research will be illustrated by a set of measured data of free prostate-specific antigen (FPSA), which are often used to

Grant sponsor: Ministry of Science and Technology of China; Grant numbers: 2009ZX09312-025; 2008ZX09312-007.

Yu-Hao Zhou, Jia-Jie Zang, and Mei-Jing Wu contributed equally to this study.

*Correspondence to: Jia He, Department of Health Statistics, Second Military Medical University, Shanghai 200433, China.

E-mail: hejia63@yahoo.com

Received 11 June 2010; Accepted 3 November 2010

DOI 10.1002/jcla.20437

Published online in Wiley Online Library (wileyonlinelibrary.com).

evaluate the presence of prostate cancer and other prostate disorders, and two sets of hypothetical data (Table 1) generated from the measured one.

MATERIALS AND METHODS

Sample Data and Plotting Data

AIA-1800 and I2000 methods were used to measure the FPSA. In the process of measurement, a same random sequence of sample was used in the two instruments. The simulation data① were the results of I2000 measured values plus 0.003 mmol/l, which simulated the enlarged systematic error. And the simulation data② were calculated by randomly adding or subtracting 0.003 mmol/l to the I2000 measured values, which simulated the enlarged random error. In the process for the analysis of the repeated measurement data, we assumed that the data measured from the two instruments were considered as the repeated measurement data by comparison with the results of two different analyses pattern, and these data were only collected to demonstrate the statistical method and provide no evidence on the comparability of these two instruments.

TABLE 1. The Measurement Data and Simulation Data of AIA1800 and I2000 Methods (mmol/l)

Subjects no.	Measurement data		Simulation data①		Simulation data②	
	AIA-1800	I2000	AIA-1800	I2000	AIA-1800	I2000
1	0.023	0.021	0.023	0.024	0.023	0.024
2	0.022	0.023	0.022	0.026	0.022	0.020
3	0.025	0.024	0.025	0.027	0.025	0.027
4	0.013	0.013	0.013	0.016	0.013	0.010
5	0.008	0.007	0.008	0.010	0.008	0.010
6	0.017	0.016	0.017	0.019	0.017	0.019
7	0.026	0.022	0.026	0.025	0.026	0.025
8	0.017	0.016	0.017	0.019	0.017	0.019
9	0.034	0.032	0.034	0.035	0.034	0.035
10	0.007	0.006	0.007	0.009	0.007	0.003
11	0.011	0.009	0.011	0.012	0.011	0.012
12	0.006	0.004	0.006	0.007	0.006	0.001
13	0.001	0.001	0.001	0.004	0.001	0.004
14	0.006	0.007	0.006	0.010	0.006	0.004
15	0.005	0.003	0.005	0.006	0.005	0.006
16	0.007	0.005	0.007	0.008	0.007	0.008
17	0.011	0.009	0.011	0.012	0.011	0.012
18	0.002	0.002	0.002	0.005	0.002	0.005
19	0.037	0.036	0.037	0.039	0.037	0.033
20	0.002	0.001	0.002	0.004	0.002	0.004
21	0.003	0.001	0.003	0.004	0.003	0.004
22	0.013	0.011	0.013	0.014	0.013	0.008
23	0.008	0.007	0.008	0.010	0.008	0.004
24	0.023	0.023	0.023	0.026	0.023	0.020

RESULTS

Inappropriate Use of the Correlation Coefficient

In most research, the following methods will be comprehensively utilized. First, the scatter plots and the regression line were drawn from the measurement data. The closer between the line of regression and the line of equality ($y = x$), the better consistency we can conclude. By this way, it is convenient for researchers to gauge the degree of agreement between measurements.

Then, we usually need to calculate the correlation coefficient (r) between the two methods. The r value equals 0.994 ($P < 0.001$) from the measured data of the two equipments (Fig. 1). The null hypothesis here is that the measurements by the two methods are not linearly related. The probability is very small and we can safely conclude that FPSA measurements by AIA1800 and I2000 methods are related. However, this high correlation does not mean that the two methods agree:

- (1) r measures the strength of a relation between two variables, not the agreement between them. The satisfied agreement can be obtained only when the points lie along the line of equality ($y = x$). Nevertheless, the perfect correlation coefficient can also be obtained if the points in Figure 1 lie along any straight line.
- (2) The measurements plus (or minus) the same value does not change the correlation coefficient, but it can affect the agreement. For example, the r for Simulation data① also equals 0.994 ($P < 0.001$), which is obtained by adding 0.003 mmol/l to all the measured value of the I-2000 method. This shows

Fig. 1. Measurement data and simulation data① with AIA-1800 and I-2000 methods with line of regression.

that the simple correlation analysis is insensitive to systematic error.

- (3) The dimension of correlation coefficient depends on the sample size and the range of measuring value in some way. The larger sample size the measurement includes, the higher correlation coefficient will be obtained, and the larger range of measuring value, the high correlation coefficient will be obtained.
- (4) Significant r value shows that the two methods are related, but it is possible that two methods designed to measure the same quantity were not related. The test of significance is irrelevant to the question of agreement (3,4).
- (5) As the hypothesis testing of r is compared with zero, even though it has the difference of significance, it also has poor significance of hypothesis testing. There might be statistical significance even if r value is very small.
- (6) Data that seem to be in poor agreement can appear quite high correlations. For example, we have perfect correlation coefficient if all the points lie along any straight line but not the line of equality. It also shows high correlation but poor agreement.

ATE/LER Zones to Measure Agreement

It is most unlikely that different methods will agree 100%. We want to know by how much the new method is likely to differ from the old one; if two methods show good consistency, we can replace the old method by the new one or use the two interchangeably, for example, measure FPSA by AIA-1800 and I2000 methods. In this process, it was deemed that the two methods were clinically consistent with each other; if the results of the two methods differed by not more than 0.004 mmol/l,

we could replace the old methods by the new one because so small a difference would not affect decisions on patients' management (5). On the other hand, if the values of FPSA from the two methods could differ by 0.008 mmol/l, the new methods would be unlikely to be satisfied. But how far the difference between measurements cannot cause difficulties? This requires clinical experts and statisticians to work together to make a judgment (6).

First step: we need to examine the data. A simple plot of the results of one method against those of the other one would be drawn.

Second step: we should avoid plotting the regression line but drawing the line of equality.

Third step: divide the ATE Zones and LER Zones according to the clinical cutoff which is the maximum difference accepted clinically.

Fourth step: calculate the percentage of all the samples falling within the ATE zones and LER Zones.

From the two measurement data, results are as follows:

ATE zones

The maximal acceptable difference between the two methods of FPSA measurement is assumed as 0.004 mmol/l according to the clinical and statistical discussion which is selected for the ATE limits. We have good consistency that the lower 95% exact confidence bound should exceed 92% and 95% of observations must be within the ATE limits (2). The results can be found in Table 2.

The percentage of the samples over the entire range that fall within the ATE zone was 100% (24/24) with a lower bound of 95% confidence interval of 95.83%, as shown in Table 3.

TABLE 2. Percentage of Observations That Fall Within the ATE Zone for the Low, Middle, and High Ranges of the AIA1800 Method

Range of AIA1800 values (mmol/l)	Total number of samples	Number of samples within of ATE zones	Percent of samples within of ATE zones
0–0.010	11	11	100%
0.010–0.020	6	6	100%
0.020–0.040	7	7	100%
Total (0–0.040)	24	24	100%

TABLE 3. Percentage of Observations Over the Entire Measuring Range That Fall Within the ATE Zone With 95% Exact Lower Confidence Bound

Test	Nt (sample size)	Ntw (The number of sample within the ATE zone)	Ntw/Nt	95% confidence interval
FPSA	24	24	100%	95.83%

LER zones

The LER zone are external limits for the differences between the I-2000 and the AIA-1800 method. Results fallen into these limits may cause a risk to the safety of patient (2). None of the data points should fall within the LER zones in a surrogated method. From the data in Figure 2, the percentage of the samples over the entire range that fall within the LER zone was 0.

For the reasons above, the principle of ATE/LER zones is practically used to evaluate test method matches comparative method, and we could easily conclude that the two methods have good consistency for measurement data.

Applicable Conditions and Notifications of ATE/LER Zones

All the points scattered and fluctuated around the line of equality within ATE zones, which are in good agreement can be concluded accordingly, according to Figure 2. But it does not firmly means that two kinds of methods have good consistency when more than 95% of points fell in the ATE zones and no points lined in LER zones. Some occasions addressed as follows should be paid more attention to.

- (1) As shown in Figure 3, when nearly all the points lay in one side of ATE zones, and none in LER zones, it also reflects good consistency and high precision.
 - (a) The results demonstrated that there might be the existence of the systematic bias in this case, which is easily found when carefully observed the ATE/LER zones.
 - (b) The results also demonstrated that the difference between the two methods might be statistically

significant but not practically significant. Under such circumstances, good consistency is still deemed.

- (2) As shown in Figure 4, all results in the rim of ATE zones, fluctuating along the line of equality, are not out of ATE zones. This reflects good consistency and poor precision. Under this circumstance, appropriate increased sample size or repeated measurement should be employed as the solution to increase the stability of the results.
- (3) If the differences are correlated with the mean, we should base on the different boundary values in different Settings. It has two situations as follows:
 - (a) The differences are proportional to the mean. We could accord the evolution to draw up different boundary value, for instance, it may be the line in

Fig. 3. ATE/LER zones of the simulation data①.

Fig. 2. ATE/LER zones of the measurement data.

Fig. 4. ATE/LER zones of the simulation data②.

which a slope is not equal to 1 or other regular curve.

- (b) The differences are disproportionate to the mean. We need to setup different boundary values according to different measurements, and the graphics of ATE/LER Zones will be serrated.
- (4) The ATE/LER zones division and criterion for repeated measurements of single method are different from that of comparative consistency of two methods. The detailed discussion will be made in the next section.

The Stability of Repeated Measurement by Using ATE/LER Zones

The standard ATE/LER zones cannot be directly applied when estimating the stability of a single method performed on repeated measurements. However, a modified approach can be used. For the two set of measurements data in a single method, we should estimate by the mean difference ($\bar{d}$) and the standard deviation of the difference(s). In addition, the standard deviation of the difference differs from the standard calculation methods. The hypothetical difference between repeated measurements is considered to be zero. The coefficient s is estimated by the steps of squaring all the differences, adding them up, divided by n , and taking the square root ($s = \sqrt{\sum d^2/n}$). The measurement data of the above two methods of AIA-1800 and I2000 are assumed as the repeated measurement of a single method. Then the mean difference of the two measurements data is -0.00108 mmol/l and s is 0.00161 mmol/l in this study. Most of the objects are expected to lie within the line of $y = x - 2s$ and $y = x + 2s$ (Fig. 5). If 95% of objects lie between these limits, good consistency of repeated

measurements data can be concluded (7). If the line points lie on the line of $y = x - 2s$ and $y = x + 2s$, they should be excluded from ATE zones for conservative estimation. If higher precision is needed, the limits between $y = x - 1.96s$ and $y = x + 1.96s$ should be considered.

As shown in Figure 5, now the boundary value of the modified ATE zones will be differed from the standard ATE zones criteria when measured the agreement for paired measurements; the clinical cutoff for ATE acceptance criteria have been replaced by $2s$.

As for repeated measurements data, objects lying within $y = x \pm 2s$ showed good consistency, even if there was no clinical importance, and both systematic bias and random error were small. For the repeated measurements data in our study, we get the acceptance criteria as follows:

$$y = x - 2s = x - 0.00322; \quad y = x + 2s = x + 0.00322$$

The percentage of observation fall within the ATE zones is 95.83% after calculated using the above method from which we can easily conclude that repeated measurements have a good consistency. However, if there are more than two repeated measurements, the calculations will be more complex, which is beyond the scope of this method.

When using other repeated measure data to illustrate the new method, as is shown in Figure 6, the repeated measurements of FPSA were assessed by a new technique of the I-2000 method. The mean difference is 0.00448 mmol/l and s is 0.05449 mmol/l according to the calculation mentioned above. Most of the objects are expected to lie between the line of $y = x - 0.10898$ and $y = x + 0.10898$. The percentage of observation fall within the ATE zones is 92.79%, which is not confident enough to believe that the new method has a good repeatability.

Fig. 5. ATE zone of the repeated measurements data.

Fig. 6. Repeated measures of FPSA using I-2000 method.

Measuring Agreement for Repeated Measurements

If we get two sets of repeated measurement results by two methods, we can calculate the mean for repeated measurement data of each method, then use these two sets of means to compare two methods by using the analysis for assessing agreement described above. In this situation, the estimation of the standard deviation of the differences will be small by using the mean differences of the repeated measurement of both methods, because some of the repeated measurement error has been neutralized. This effect should be adjusted by means of modifying standard deviation in the following method. The differences (d_1, d_2) between repeated measurements for each method should be calculated, and then the standard deviation (s_1, s_2) for the differences (d_1, d_2) is computed. The standard deviation of the difference between the means for each method (s_D) should also be worked out. Then the adjusted standard deviation of the differences, s_c , is modulated as $\sqrt{(s_1^2/4) + (s_2^2/4) + s_D^2}$ (8). In this way, it is more rational for two repeated measurement data by using the s_c instead of s_D when determining the ATE zones.

There might be some repeated data for both methods clinically. For example, we used two methods, I-2000 and I-1800 methods, to measure the CA125 value twice, respectively. Then the repeated measurements were obtained. When assessing the agreement of the two methods, we should use the s_c to form the ATE zones. The results are calculated as follows:

$$s_1 = 0.56017, \quad s_2 = 0.56425, \quad s_D = 0.36577, \\ s_c = 0.54021$$

Then two groups of line could be worked out, $y = x \pm 2s_D$ and $y = x \pm 2s_c$. The lines of latter group,

Fig. 7. Repeated measures of CA125 for two methods.

the black ones, should be adopted to set down the acceptance limits.

As is shown in Figure 7, the percentage of objects fall within the ATE zones is 95.45%, which is confident enough to believe that two methods show a good agreement. However, when assessing the agreement by using the s_D , the percentage of objects fall within the ATE zones is 86.36%, which is not confident enough to believe that two methods show a good agreement.

DISCUSSION

There are several traditional methods to assess the agreement of diagnostic test, such as correlation coefficient and regression analysis. The ATE/LER zones approach is recommended by FDA to avoid the shortcomings of the traditional methods. Furthermore, the same method may be used to analyze the repeatability of a single measurement method or to compare measurements by two observers. But neither of them is perfect to analyze the consistency of two measurements. Therefore, it will be more appropriate to assess the consistency of two methods using the combined methods.

There are many different treatments for the consistency of the diagnostic test, and why is that? Several reasons may be at work here: inappropriate methods get widely used, once these methods are published, others will read it and if they encounter similar problems, they will use the same technique—imitation is the main reason and to solve different problems, such as Bland and Altman (8) regard the consistency of the test results in any circumstance, but Myles and Cui (9) think the consistency of the results is the effect after receiving treatment of the consistency test, assessing the results is the consistency of treatment effect, so they have different treatments when encountering the same question.

Finally, assessing the agreement, Allowable Total Error and Limits for Erroneous Results (ATE/LER) zones is a relatively good method, even though analyzing process requires more factors, and it could obtain the results for more in accordance with clinical significance.

REFERENCES

1. Li Y, Li X. Evaluation on different assessment methods of quantitative measurements. *Mod Prev Med* 2007;17:3263–3266.
2. Center for Devices and Radiological Health (CDRH), US Food and Drug Administration (FDA). Guidance for industry and FDA staff: recommendations for clinical laboratory improvement amendments of 1988 (CLIA) waiver applications for manufacturers of in vitro diagnostic devices. 2008. Available online at: <http://www.fda.gov/cdrh/oivd/guidance/1171.pdf>

3. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *Statistician* 1983;32:307–317.
4. Serfontein GL, Jaroszewicz AM. Estimation of gestational age at birth: comparison of two methods. *Arch Dis Child* 1978;53:509–511.
5. Krouwer JS, Astles JR, Gutman SI, Koch DD, Levine JB, Otto CN. 2003. Estimation of total analytical error for clinical laboratory methods: approved guideline. NCCLS 0273-3099.
6. Krouwer JS. Six Sigma can be Dangerous to Your Health, Vol. 14. Berlin: Springer. 2009. p 49–52.
7. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999;8:135–160.
8. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;i:307–310.
9. Myles PS, Cui J. Using the Bland-Altman method to measure agreement with repeated measures. *Br J Anaesth* 2007;99:309–311.