The agreement between two diagnostic methods in binary cases: a proposal

J. C. AZZIMONTI RENZO

Facultad de Ciencias Exactas, Químicas y Naturales, Universidad Nacional de Misiones, Argentina

Azzimonti Renzo JC. The agreement between two diagnostic methods in binary cases: a proposal. Scand J Clin Lab Invest 2002; 62: 391–398.

This report can be considered as a resource in the analysis of agreement among raters, clinical tests, observers, judges or experts. The focus is on diagnostic methods. When the true results of diagnostic methods cannot be obtained, studying the agreement between them can reflect the difference between methods. Normal statistical procedures tackling this are not enough for deciding about the agreement from a clinical point of view. The clinical question is whether the new method agrees sufficiently with the old one. In binary cases, a solution for deciding about the agreement from a clinical viewpoint is introduced. The overall agreement is a duality and needs to be studied in two steps. In the first step, a condition for accepting the agreement is proposed: both methods need to have the same nosologic sensitivity and specificity. In the second step, another condition is proposed: the level of the agreement should be greater than a critical value, defined by the clinicians. When both steps show satisfactory results, the new method can replace the old one. The statistical procedures for testing both steps are presented.

Key words: Agreement table; disagreement odds; level of agreement; sensitivity; specificity; truth table

Juan Carlos Azzimonti Renzo, Av. Uruguay 2655, Posadas (3300) Misiones, Argentina. Fax. + 54 03752 422180, e-mail. arroi_pss@ciudad.com.ar

INTRODUCTION

The agreement problem concerns the comparison of two paired samples using two clinical methods in each studied individual when the true values remain unknown. The problem is in comparing the new method against the usual one (the old method). *When the new method agrees sufficiently well with the old, the old one may be replaced.* Sometimes direct measurement without there being adverse effects is difficult or impossible (e.g. cardiac stroke volume or blood pressure), so indirect methods are used; a new method has to be evaluated by comparing with an established technique rather than with the true quantity [1, 2]. This problem is analysed in the present report, but on a binary scale. There is no consensus on what statistical tests are best for assessing agreement in binary cases. The suggested methods are: (a) testing association between raters with the log odds ratio, and (b) using McNemar's test to evaluate marginal homogeneity [3].

The present report focuses on diagnostic methods. In medical practice, the diagnostic investigation starts with the patient presenting with a particular symptom or sign indicative of the presence of a particular disease, the so-called target disease. The diagnostic investigation is a consecutive (hierarchical) process always starting from the patient history and physical examination, followed by more invasive, time-consuming and costly tests such as imaging. Normally, no diagnosis is established by a single test result; each test result is judged together with other (previous) test results [4]. Sometimes, however, it is based on a single test, such as toxoplasmosis detection, Chagas' disease, etc. For the present report, this consecutive investigation is called diagnostic method.

The main objective of a diagnostic method is to be able to detect whether the illness is present or absent in the patient. Regardless of the magnitude type, the magnitude can generally be transformed in a binary case [5] by adopting a cut-off point. This adopted point separates positive from negative cases. When the disease is detected by the method the diagnosis is positive, and negative when it is not. If each individual is tested by two diagnostic methods, the set of results obtained is called paired samples. The usual presentation of the data is in the form of an agreement table (Table I). Agreement is also known as duplicate testing of the same individuals when there are two clinical methods for the same illness [6–8]. The usual way for determining the agreement between two clinical methods is to employ a statistical test. Note that a clinical criterion is not used and that responsibility for the decision is discharged onto the shoulders of the statistician.

In the statistical tests, the null hypothesis is H_0: There is agreement. If H_0 is rejected, there is validation for proving that there is not agreement. But if H_0 is not rejected, there is no validation of the agreement, and the conclusion must be: there is not enough evidence for rejecting the agreement, i.e. the existence of the agreement cannot be proved. More objections to the use of statistics for adopting the agreement from a clinical viewpoint are given in Appendix 1.

The main objective of the present report is to propose a clinical solution for the agreement problem in binary cases, i.e. to determine whether the usual diagnostic method for an illness can be replaced by a new one when the true values remain unknown. And the responsibility for this decision is based on a clinical viewpoint with the help of statistics.

TABLE I. Agreement table.

	Method 1		
Method 2	(+)	(–)	Total
(+)	a	b	a+ b
(–)	c	d	c+ d
Total	a+ c	b+ d	N

Where the frequencies are: a= the subjects that show both (+) results (positive agreement); d= the subjects that show both (–) results (negative agreement); c= the subjects that show (+) with the first method and (–) with the other one (disagreement); b= the subjects that show (–) with the first method and (+) with the other one (disagreement). Level of the agreement (raw agreement): $\lambda = (a+d)/N$ (expressed as a percentage).

PROPOSED PROCEDURE

The proposed way for deciding whether to replace the old method with a new one is by comparing sensitivity and specificity. The investigative results for sensitivity and specificity are normally called nosologic indices [9]. Comparison of these two indices is used to establish whether the old method is worse (or better) than the new one for the disease that is being considered. This procedure requires that the true diagnostic of each studied individual can be obtained. Usually, sensitivity and specificity are known only in the old method. So for studying the agreement, the same individual tested with the old method in the routine can be studied with the new method (that is the extra cost). The agreement is then cheaper and faster than a procedure based on acquisition of the true state of each individual. Note that the agreement is focused on the equivalence and does not give information about which method is the best option. The fact that the new method does not agree with the old does not mean that the new method is worse (or better) than the old.

The agreement problem is a dual concept.

Agreements and disagreements are the two opposite concepts to be studied. So the overall agreement should be analysed in two steps:

Step 1: Two diagnostic methods agree when they have the same sensitivity and specificity.

A diagnostic method can be imagined as a kind of machine that makes two types of prediction: (+) if the illness is diagnosed, and (–) if it is not. Both types of prediction can be true or false. There will therefore be four possible events, mutually independent, for each patient schematized in a truth table (see Table II) [5–7].

When one of the methods can be considered as the reference method, or the truth, the agreement table is a truth table. From this point, the agreement can be studied. Suppose that the first method is the true one, then the sensitivity and the specificity of Method 2 are: S_2= a/(a+ c) and E_2= d/(b+ d). On the other hand, if the second method is the true one, then the sensitivity and the specificity of Method 1 are: S_1= a/(a+ b) and E_1= d/(c+ d). Therefore, S_1= S_2 and E_1= E_2, if and only if b= c.

The condition b= c is studied by the statistical tests detailed in Appendix 1. The main conclusion of this Appendix is: the statistical test is not enough for deciding the replacement of the old method by the new one from a clinical point of view. So another analysis is needed:

Step 2: The level of agreement (μ) should be sufficiently large for having an acceptable agreement from a clinical point of view.

For this step, the clinicians should adopt a critical value ($\lambda_{\text{critical}}$) for the level of the agreement (or raw agreement). Then, when it is $\lambda \geq \lambda_{\text{critical}}$ the new method can replace the old, and the overall agreement will be clinically acceptable. For example, N= 100 individuals was selected randomly, and the two methods were applied. If a total of 90 agreements were found, then λ is equal to 90%. This value could be unacceptable for some kinds of cancer, HIV, etc., but it could be acceptable for other diseases. Each illness should therefore have its own critical value.

Consider now a significance test for this step. The logical idea is to compare the level of agreement (λ) against its critical value. The index λ can be considered as a probability, and the normal approximation to compare a sample value against a population value ($\lambda_{\text{critical}}$) can be used. Note that the sample size (N) will be evaluated here. The level of agreement alone is not enough to make a decision, as shown in Appendix 2. Even in cases with a high value of the level of agreement, the statistical tests presented in Appendix 1 should also be used.

As an alternative to the level of the agreement, another index is introduced in the present report: the Disagreement Odds (DO), which can be calculated as follows: If the event ***A*** is the number of disagreements, then its probability will be: p (***A***)= (b+ c)/N. The odds of the disagreements can be obtained with:

$$DO= P\ (\boldsymbol{A})/[1- p\ (\boldsymbol{A})]= (b+ c)/(a+ d)$$

Its relationship with the level of the agreement is: λ= 100/[1+ DO]% or DO= (100/λ)– 1

The reason for introducing DO is that the fraction of disagreements is more reliable than

TABLE II. Truth table and the main quality indexes.

	Disease (true results)		
Test results	Yes	No	Total
Positive	tp	fp	T+
(+)	*true positive*	*false positive*	
Negative	fn	tn	T–
(–)	*false negative*	*true negative*	
Total	TD	TnD	N

Where N is the number of the investigated subjects and

T+ = tp+ fp: Total of positive diagnosed subjects	Sensitivity= tp/TD= S
T– = tn+ fn: Total of negative diagnosed subjects	Specificity= tn/TnD= E
TD= tp+ fn: Total of diseased subjects	Prevalence= TD/N
TnD= fp+ tn: Total of not diseased subjects	Accuracy= (tp+ tn)/N

any statistical test. DO is a simple, understandable and useful concept for readers who prefer the use of odds rather than probabilities. By using DO or λ, clinicians can concentrate on clinical facts, rather than on mathematical issues by using an easy model of confidence intervals, and following the usual recommendations [11]. It has a simple clinical meaning, when the DO= 1/9 indicates that there will be 1 disagreement and 9 agreements in 10 samples (λ= 90%). The ideal DO value is DO= 0, because there are no disagreements, and the agreement is "perfect". When the DO= 1, this is like throwing a coin into the air to obtain the diagnostic. When the number of disagreements (b+ c) is greater than the number of agreements (a+ d), it is not reasonable to hope that the clinical agreement will be acceptable. Therefore, the assumption $0 \leq DO \leq 1$ can be adopted from a clinical viewpoint.

Under this assumption, the DO can be imagined as a proportion. Therefore the 95% confidence interval for DO can be obtained by using the normal approximation [12]. Assuming π is the real proportion of the disagreement odds in the population, from which a sample of size N was chosen randomly, and DO is the observed value ($0 \leq DO \leq 1$), then the expected value will be E (DO)= π, which can be estimated by $\pi \approx DO$. And the standard error can be estimated with SE (DO)= $[DO\ (1- DO)/(a+ d)]^{1/2}$. So, if the sample is large enough, (a+ d) > 25, the 95% confidence interval of this index can be estimated with the following critical limits:

$$\text{Upper limit} = DO + 1.96\ SE\ (DO)$$

$$\text{Lower limit} = DO - 1.96\ SE\ (DO)$$

Analogously, the 95% confidence interval for the level of the agreement is obtained with:

$$\text{Upper limit} = \lambda + 1.96\ SE\ (\lambda)$$

$$\text{Lower limit} = \lambda - 1.96\ SE\ (\lambda)$$

Where SE (λ)= $[\lambda\ (100- \lambda)/(N)]^{1/2}$ and E (λ)$\approx\lambda$

The rules for acceptance of the agreement from a clinical viewpoint are:

(1) When the critical value $DO_{critical}$ does not lie in the interval because it is smaller than the lower limit (or when $\lambda_{critical}$ does not lie in the interval because it is greater than the upper limit), then the agreement is not clinically acceptable. However, there is no statistical rejection, which means that the clinical criterion manages the problem, and the final decision about the agreement is a clinical responsibility, not a statistical one.

(2) When the critical value $DO_{critical}$ lies in the interval, or is greater than the upper limit, then the agreement is acceptable from a clinical point of view (i.e. the agreement is acceptable when the critical level of the agreement, $\lambda_{critical}$, lies in the interval or is smaller than the lower limit).

For example, in Case 1 of Table III the agreement is rejected because it does not verify Step 1. But it is verified in the last two cases of Table III. In Case 2, the DO= 0.2 and its 95% confidence interval is (0.16, 0.24). As the critical value $DO_{critical}$= 0.053 ($\lambda_{critical}$= 95%) does not lie in the interval, the conclusion is: there is

TABLE III. The conditions for the agreement tested in three cases.

Case 1	Method 1		
Method 2	Yes	No	
+	180	22	202
–	10	188	198
	190	210	400

z^2= 3.78 <Gadj= 4.54
Statistical rejection
$DO_{critical}$= 0.053

Agreement rejected

$\lambda_{critical}$= 95%

Case 2	Method 1		
Method 3	Yes	No	
+	160	36	196
–	30	174	204
	190	210	400

z^2= 0.38 – Gadj= 0.54
DO= 0.2
DO 95% CI (0.16, 0.24)
$DO_{critical}$ <0.16

λ 95% CI (80, 87)%
$\lambda_{critical}$ > 87%
Agreement rejected

Case 3	Method 1		
Method 4	Yes	No	
+	185	10	195
–	5	200	205
	190	210	400

z^2= 1.07 – Gadj= 1.64
DO= 0.04
DO 95% CI (0.02, 0.06)
$DO_{critical}$ lies in the interval

λ 95% CI (94.4, 98.0)%
$\lambda_{critical}$ lies in the interval
Agreement accepted

statistical agreement between Method 3 and Method 1, but it is not enough to be acceptable from a clinical viewpoint. In Case 3 the results are: DO= 0.04 with a 95% CI (0.02, 0.06). Using the level of agreement the results are λ= 96.25% with a 95% CI (94.4, 98.0). As the critical value lies in the interval, the clinical conclusion is: the agreement is acceptable between both methods. So the new Method 4 can replace the old one.

An algorithm for solving all the calculations of this new procedure is available free at www.medal.org (English) or at www.bioestadistica.com.ar (Spanish). Only the four data for the table and the *clinical criterion* should be introduced for obtaining the final decision about agreement, so clinicians can readily make small changes in the clinical criterion to see what happens with the agreement. This will be helpful in research for the adoption of the critical value for each disease.

CONCLUSIONS

Appendix 1 shows why the usual statistical approach alone is not enough for solving the problem of agreement and Appendix 2 why the use of a clinical index alone is not enough either. The use of two strategies is therefore necessary to solve the problem. For this reason the present proposal can be summarized in two steps:

Step 1: Verify that sensitivity and specificity are similar in both methods by using the G-test. When the agreement verifies this condition, the next step should be performed. If it does not, be careful with certain exceptions, as explained in Appendix 2.

Step 2: Verify the clinical condition: the obtained DO is not greater than $\mathrm{DO}_{\text{critical}}$, or the obtained λ is not smaller than $\lambda_{\text{critical}}$.

When these two conditions are verified, the agreement will be acceptable from a clinical viewpoint.

This approach can only be used if the diagnostic method (or a single clinical test) can be transformed in a binary case. For example, in a positive case the treatment is necessary for the patient, and in a negative case the treatment is unnecessary.

When the true state of the patient can be obtained, a truth table is the best option to use for analysing the performance (quality) of the diagnostic methods. When the true values cannot be obtained, the proposed procedure can be used to see if a new method can replace the old, but not for deciding which is the best option. A clinical magnitude can usually be transformed in a binary case by adopting a cut-off point to separate a positive case from a negative one. So the present procedure is more general than the usual one because it can be used in almost any illness (or clinical tests).

Briefly, the main assumptions for having acceptable clinical agreement between two paired methods are:

- It is possible to transform the results of a diagnostic method into a binary case.
- The sensitivity and specificity of both methods should be equal.
- The disagreement odds should be delimited ($0 \leq \mathrm{DO} \leq 1$) to be imagined as a proportion. Note that this delimitation is not needed for λ.
- The disagreement odds should be smaller, or at most equal, to a clinical critical value, i.e. the level of the agreement should be greater, or at most equal to $\lambda_{\text{critical}}$.

The main advantages are:

- The final decision about agreement is based on a clinical criterion rather than a statistical one.
- The procedure is more reliable than any of the usual statistical tests.
- This is a new procedure for deciding if a new method can replace the old in binary cases.
- It is a more general possible application in the clinical field.
- The true values, or reference methods, are not necessary. So the cost diminishes.
- The problem of spectrum and selection bias [13] is avoided because the same individuals are tested twice.
- It could be the solution for comparing two reference methods.

APPENDIX 1: STATISTICAL PROCEDURES

(1) *McNemar's test*: The null hypothesis is that the expectations of b and c are equal. So, McNemar's test with continuity correction (Yates) included can be used [7, p. 121]:

$$z^2 = (|b - c| - 1)^2/(b + c)$$

Where z^2 may be regarded as a $\chi^2_{(\alpha;1)}$ variate. So when $z^2 > 3.841$ the null hypothesis should be rejected. This can be considered as statistical proof of the disagreement between the two studied methods. But there are two main objections:

(a) When $b \approx c$ it always results in $z^2 \approx 0$ and the H_0 will not be rejected.

For example, if b= c= 199 and N= 400, there will be 398 disagreements in 400 cases. The same thing happens when b= c= 1, statistically. But it is not the same thing to have 398 than to have 2 disagreements in 400 cases from a clinical viewpoint.

(b) The sample size N is not taken into account.

For example, if b= 25 and c= 10 in 400 samples, the H_0 is rejected because $z^2 = 5.6 > 3.841$. The same thing happens when b= 25 and c= 10 in 400 million samples, statistically. But it is not the same thing to have 35 disagreements in 400 samples, than in 400 million samples, clinically.

(2) *Cochran Q-test*: This test is equivalent to McNemar's test without continuity correction [10].

$$Q = (b - c)^2/(b + c)$$

Note that in these two tests the same objections (a) and (b) can be made.

(3) *McNemar's G-test*: This test could be a better option than the previous ones because it has more power to detect slight differences [6]. This problem is also known as *individual tested twice*, and is based on a multinomial distribution. The natural logarithm of the ratio between the two probabilities (the observed and the expected one), or the likelihood ratio, is G/2. Where the statistic G is approximately distributed as a $\chi^2_{(\alpha;1)}$. The equation to calculate G is:

$$G = 2\ \{b \ln[2b/(b + c)] + c \ln[2c/(c + b)]\}$$

This test has the Williams' correction for continuity [6] given by the factor $q = 1 + 1/(b + c)$, and the adjusted value will be Gadj= G/q. But the same objections (a) and (b) can be made.

Briefly, to analyse the agreement all these tests are questionable from a clinical point of view.

To illustrate these procedures, an example (Case 1, Table III) is presented. Method 1 is the old one and Method 2 is the new one. There are 32 cases of disagreement between both methods. Therefore, by applying McNemar's test, $z^2 = 3.78$, which is not significant. While by applying the Cochran Q-test, Q= 4.50, which is significant. The same occurs with McNemar's G-test, where Gadj= 4.54. This example is a "border" case, where the statistical decision is not clear. On the one hand, McNemar's test does not have enough evidence to reject the agreement, but it is very near to the 95% significance limit ($\chi^2_{(0.95;\ 1)} = 3.841$). On the other hand, the Cochran Q-test and McNemar G-test give the necessary evidence for rejection of the agreement. Therefore, the statistical conclusion should be: the agreement should be rejected because the adjusted G-test is the most powerful statistical test.

In Case 2, Table III, there is not enough evidence to reject the agreement ($z^2 = 0.38$ and Gadj= 0.54), as also in Case 3 ($z^2 = 1.07$ and Gadj= 1.64). Then, the agreement in Cases 2 and 3 should not be rejected from a statistical viewpoint. But in Case 2 there is not enough evidence to be accepted from a clinical viewpoint.

In conclusion: the statistical tests are not enough for deciding on replacement of the old method by the new one from a clinical viewpoint, i.e. the evidence is enough just for rejecting, and not for accepting, the agreement, plus the objections (a) and (b).

APPENDIX 2: THE NEED FOR THE TWO STEPS IN THE DUAL VISION PROCEDURE

The duplicate strategy is not superfluous, because if the level of agreement is high the sensitivities and specificities could be different, according to the statistical results. For example:

(a) *When the agreement is rejected statistically by using the G-test*, a high level of the agreement does not mean that the sensitivities and specificities are similar. For example for $\lambda = 90\%$.

Case A λ= 90%

	Method 1		
Method 2	+	–	
+	400	10	410
–	90	500	590
	490	510	1000

Assuming that Method 1 is the true one, then S_2 and E_2 can be obtained. When Method 2 is the true one, then S_1 and E_1 can be calculated. So,

S_2= 0.82 and S_1= 0.98, which are different (z= 7.57) statistically

E_2= 0.98 and E_1= 0.85, which are different (z= 7.65) statistically

Case B λ= 90%

	Method 1		
Method 2	+	–	
+	700	90	790
–	10	200	210
	710	290	1000

Analogously to the previous case:

S_2= 0.99 and S_1= 0.89, which are different (z= 7.74) statistically

E_2= 0.69 and E_1= 0.95, which are different (z= 7.25) statistically

In Case A the level of agreement is high, but the sensitivity falls from 0.98 to 0.82 when Method 2 replaces Method 1. This could be dangerous for a patient when the disease can be cured if it is detected on time, as well as any disease where a false-negative is more dangerous for the patient than a false-positive, such as myocardial infarct, uterine cancer, etc. Therefore, in spite of the high level of agreement, the new method should not replace the old due to clinical reasons, and this fact is detected by the statistical test, not by the level of agreement.

In Case B, the level of agreement is the same (90%), but the specificity falls from 0.95 to 0.69. This could be dangerous in the opposite case, when the illness is incurable and the worst mistake is a false-positive, such as tertiary syphilis for end-stage disease, irreversible cancer, etc., as well as any disease where a false-positive is more dangerous for the patient than a false-negative. Again, this rejection is detected by the G-test, but is not by seeing DO or λ.

There could be *exceptions*, however; for example, if the sample is too large and the G-test is significant, the sensitivities and specificities might not vary so much (imagine the above 100 disagreements, but in 1 million samples instead of 1000). Then, when the sensitivities and specificities are too high, the level of agreement will result too high, and the G-test results can be ignored. So, in these cases more incisive analyses (Case A and Case B) should be done.

(b) *When the agreement is not rejected statistically*, and when the level of the agreement is high, then the sensitivities and specificities should be similar, as is shown in Case 2 and Case 3 of Table III, where no significant differences are detected between both sensitivities and both specificities in each case.

The conclusion of this appendix is: The analysis of the level of agreement is not enough for deciding on the replacement, because the agreement problem is a duality: *Both steps are needed.*

ACKNOWLEDGEMENTS

To Edward de Vere (1550–1604), Oxford – *in memoriam*. I thank my students, reviewers and Dr. J. Sviberly from Medal.org.

REFERENCES

1 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; i: 307–10.

2 Bland JM, Altman DG. Statistical methods for assessing agreement between measurements. Biochimica Clinica 1987; 11: 399–404.

3 Uebersax JS. Statistical methods for rater agreement. 2002. (http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm)

4 Moons KGM, Grobbee DE. Diagnostic studies as multivariable, prediction research. JECH 2002; 56: 337–8.

5 Azzimonti Renzo JC. Bioestadística aplicada a Bioquímica y Farmacia. UNaM Ed. Universitaria, 2001. (http://www.fceqyn.unam.edu.ar/bio)

6 Sokal RR, Rohlf J. Biometry, 2nd ed. W. Freeman & Co.; 1981.

7 Armitage P, Berry G. Statistical methods in medical research. Oxford: Blackwell Scientific Publications; 1987.

8 Riffenburgh RH. Statistics in medicine. London: Academic Press; 1999.

9 Feinstein AR. Misguided efforts and future challenges for research on "diagnostic tests". JECH 2002; 56: 330–2.

10 Conover WJ. Practical nonparametric statistics, 3rd ed. New York: John Wiley & Sons; 1999.

11 Gardner MJ, Altman DG. Confidence interval rather than P values: estimation rather than hypothesis testing. Br Med J 1986; 292: 746–50.

12 Johnson NL, Leone FC. Statistics and experimental design, 2nd ed. New York: J. Wiley & Sons; 1977.

13 Knottnerus JA, van Weel C, Muris JWM. Evaluation of diagnostic procedures. Br Med J 2002; 324: 477–80.

Received: 15 February 2002
Accepted: 11 July 2002