REVIEW ARTICLE

Concordance Analysis

Part 16 of a Series on Evaluation of Scientific Publications

Robert Kwiecien, Annette Kopp-Schneider, Maria Blettner

SUMMARY

Background: In this article, we describe qualitative and quantitative methods for assessing the degree of agreement (concordance) between two measuring or rating techniques. An assessment of concordance is particularly important when a new measuring technique is introduced.

Methods: We give an example to illustrate a number of simple methods of comparing different measuring or rating techniques, and we explain the underlying principle of each method. We also give further illustrative examples from medical research papers that were retrieved by a selective literature search.

Results: Methods of comparing different measuring or rating techniques are of two kinds: those with a nominal rating scale and those with a continuous rating scale. We only discuss methods for comparing one measuring or rating technique with another one. Moreover, we point out some common erroneous approaches to concordance analysis.

Conclusion: Concordance analysis is needed to establish the validity of a new diagnostic measuring or rating technique or to demonstrate the near-equivalence of multiple measuring or rating techniques. Erroneous approaches to concordance analysis can lead to false conclusions.

▶ Cite this as:
Kwiecien R, Kopp-Schneider A, Blettner M: Concordance analysis—part 16 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2011; 108(30): 515–21.
DOI: 10.3238/arztebl.2011.0515

Institut für Biometrie und Klinische Forschung (IBKF), Westfälische Wilhelms-Universität Münster: Dr. rer. nat. Kwiecien

Abteilung Biostatistik, Deutsches Krebsforschungszentrum (DKFZ), Heidelberg: Prof. Dr. rer. nat. Kopp-Schneider

Institut für Medizinische Biometrie, Epidemiologie und Informatik (IMBEI), Universitätsmedizin der Johannes Gutenberg Universität Mainz: Prof. Dr. rer. nat. Blettner

Many diagnostic methods in medicine carry the risk of misdiagnosis. All physicians sometimes make diagnostic errors, any two physicians will sometimes disagree on a diagnosis, and even the technical measurements on which diagnoses are based are never perfectly accurate. A common feature of medical diagnosis and technical measurement is that both of them, as a rule, are erroneous, or at least susceptible to error. In this article, we will refer to persons making diagnoses, as well as to diagnostic methods and measuring techniques, as "raters," and to the diagnoses and measurements that they make as "ratings."

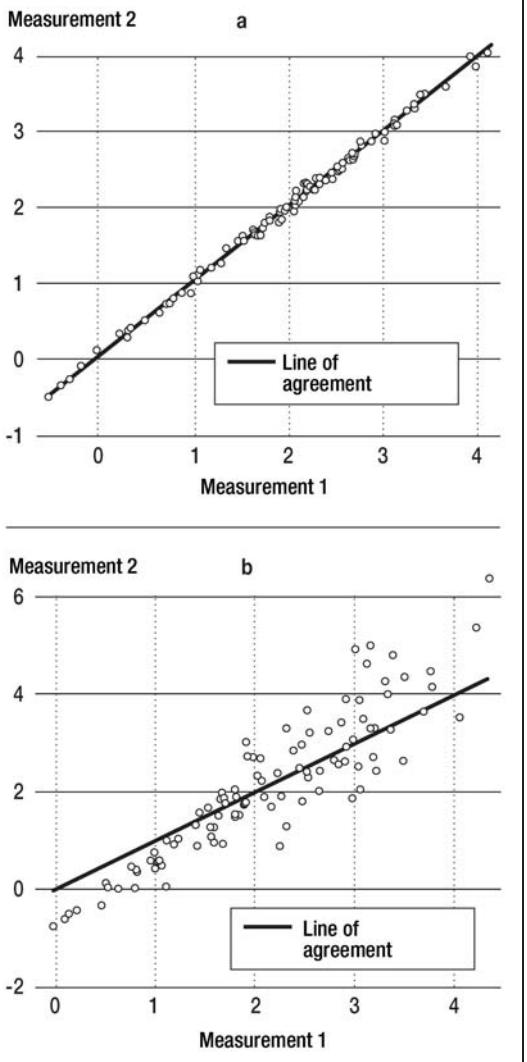
If some technique exists with which the quantity of interest can actually be measured without error, this technique is called a "gold standard." Now, suppose that a new technique is to be introduced for measuring tumor volume (for example) more readily, or with less trouble for the patient, than with the established technique (or gold standard). We will then want to know how well the measurements obtained with the new technique agree with those obtained by the old one. A common but incorrect method of comparing two measuring techniques for a quantity on a continuous scale (e.g., tumor volume) is to calculate a correlation coefficient between two sets of measurements obtained by the two techniques.

We will explain why the correlation coefficient is an unsuitable indicator of the degree of agreement (concordance) between two quantitative measuring techniques. Rather, the results obtained by them should be displayed graphically, as we will demonstrate below, so that the physician can directly assess the quality of agreement.

Measuring the agreement of a new technique with a gold standard includes determining the measuring error of the new technique. In principle, however, this is done in precisely the same way as determining the degree of agreement of two measuring techniques that are both susceptible to error.

Sometimes, one would like to assess the agreement between two raters of a nominal variable (e.g., "influenza," "flu-like illness," or "other"), or of a variable that is both nominal and ordinal (e.g., "good," "fair," or "poor"). For instance, one might like to know how closely two high-school teachers agree on gradings of term papers, or how closely two doctors agree when diagnosing patients as "healthy" or "ill."

FIGURE 1



Direct comparison of two raters with a point cloud and the diagonal line x = y; Measurement 1 vs Measurement 2 in the two examples discussed in the text: Example a above, Example b below

Our present concern is not whether the raters give correct ratings, but rather how closely they agree. The situation becomes more complicated if we want to know how closely more than two raters agree with one another; we will not discuss this any further here.

We will present descriptive methods of evaluating interrater agreement visually and quantitatively. Such methods constitute what is called concordance analysis. Our discussion will center on Bland-Altman diagrams and Cohen's kappa. We will deal with two different situations: In one situation, two raters assign a nominal rating, such as "healthy" or "ill" (dichotomous) or "influenza," "flu-like illness," or "other" (more than two alternatives), to the n members of a sample of persons or objects to be rated. In the other situation, two raters assign a numerical quantity along a continuous scale to each member of the sample.

Ratings on a continuous scale

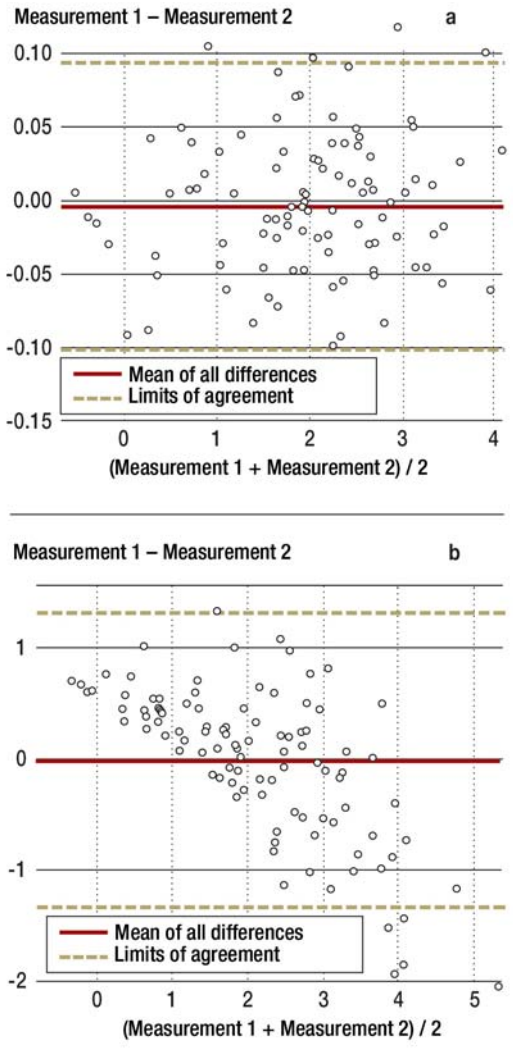
Most physical measurements are on a continuous numerical scale. Often, there is more than one technique or instrument for measuring the quantity in question, and the question arises how closely these techniques agree (1). If one wishes to introduce a new method of measuring a medical variable, one must first evaluate its validity by checking how well it agrees with an already established method, or with a gold standard.

In this section, we will present statistical methods for comparing two measuring techniques and apply them to some fictitious examples. We assume that some number n of persons or objects (perhaps 100 of them) undergo measurement with each of the two techniques, yielding a total of n pairs of measurements. As a first step, the measurements obtained by the two techniques are plotted against each other in a graph: one point is plotted for each member of the sample, its x-coordinate being the measurement obtained by the first technique and its y-coordinate the measurement obtained by the second technique. If the two techniques agree perfectly or nearly so, then all the plotted points should lie on or near the diagonal line x = y.

Two distinct and readily understandable situations are shown in *Figures 1a and 1b* (*Examples a and b*). Any pair of measurements that were precisely equal (Measurement 1 = Measurement 2) would be plotted as a point lying on the diagonal line x = y, which is drawn on both graphs. In *Example a*, the two measuring techniques agree closely; in *Example b*, however, the plot at once reveals that the difference between Measurements 1 and 2 varies ever more widely for increasing values and is greater overall than in *Example a*.

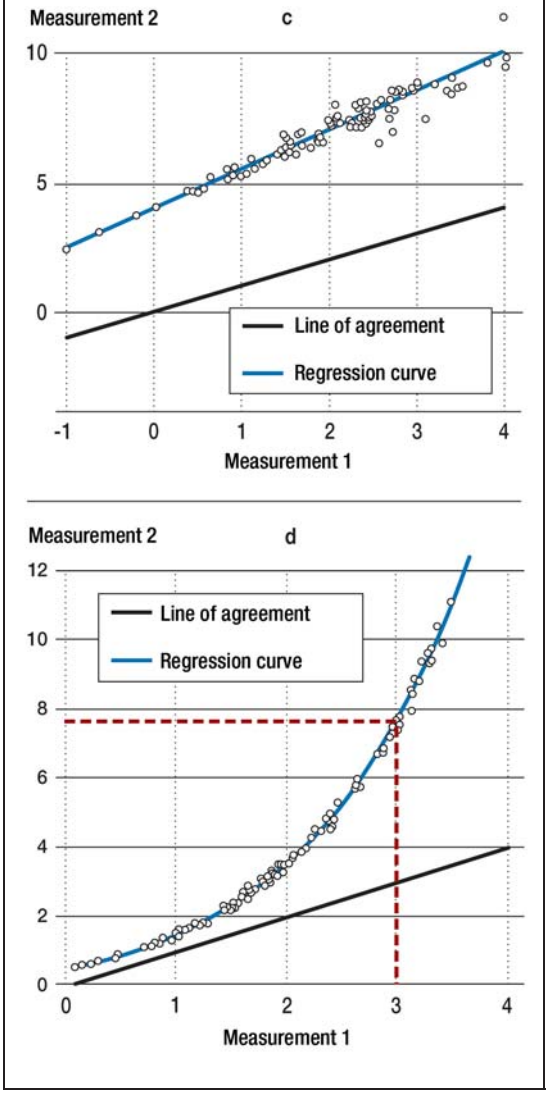
A more informative way of displaying such relationships is the so-called Bland-Altman diagram, shown for the two Examples in *Figures 2a and 2b*. As before, each pair of measurements is plotted in the x-y plane, but in a different way: The average of the two measurements is plotted as the x-coordinate, and the difference between them as the y-coordinate. In addition, the mean of all differences is plotted as a solid horizontal line, and two additional (dotted) horizontal lines are plotted above and below this line at a distance of 1.96 times the standard deviation of the differences. These two lines correspond to the so-called limits of agreement. The mean-of-all-differences line indicates a systematic deviation of the two measuring techniques for which, in general, a correction can be introduced; the limits of agreement indicate the size of further deviations that in general, are not correctable. If the quantity being measured is normally distributed, then 5% of the measured differences ought to lie beyond the limits of agreement, i.e., more than 1.96 standard deviations above or

FIGURE 2



Comparison of two raters with a Bland-Altman diagram; diagrams are shown for Example a (above) and Example b (below)

FIGURE 3



Point cloud diagrams for comparing two functionally related measuring techniques; Measurement 1 vs Measurement 2 for Example c (above) and Example d (below)

below the mean of all differences (2). The factor 2 is often used, for simplicity, instead of 1.96; the latter, however, corresponds more precisely to the 97.5% quantile of the normal distribution. In summary, the Bland-Altman diagram is a useful aid that enables a visual comparison of measuring techniques.

In *Figure 2a,* the Bland-Altman diagram for *Example a* confirms that the two measuring techniques are in close agreement. The mean-of-all-differences line is very near 0; thus, there seems to be no systematic deviation between the measured values of the two techniques. In this example, the standard deviation of all differences is roughly 0.05. Assuming that the quantity being measured is normally distributed, we can conclude that the difference between the two measurements will be less than 0.1 in 95% of cases; this difference is small in relation to the measured quantities themselves. The distance between the two limits of agreement (in other words, the width of the region of agreement) is 0.2 in this example.

When Bland-Altman diagrams are used in real-life situations to see how well two measuring techniques agree, the question whether the observed degree of agreement is good enough can only be answered in relation to the particular application for which the techniques are to be used (i.e., "good enough for what?"). Prospective users must decide how closely the measurements must agree (otherwise stated: how narrow the band between the limits of agreement must be) to be acceptable for clinical purposes. Tetzlaff et al. (1), for instance, compared magnetic

BOX 1

Calculating Cohen's kappa: an illustrative example

	Rater 2		
Rater 1	**healthy**	**ill**	**row totals**
healthy	50 (0.45)	10 (0.09)	60 (0.54)
ill	30 (0.27)	20 (0.18)	50 (0.45)
column totals	80 (0.73)	30 (0.27)	110

Let us suppose that two doctors (Raters 1 and 2) examine 110 patients for the presence of a particular disease and then state whether each patient is healthy or ill. Suppose further that Raters 1 and 2 arrive at the same diagnosis in 70 of 110 patients. The above contingency table contains all of the relevant data on the absolute and relative frequencies of agreement and disagreement in our fictitious example.

Raters 1 and 2 agreed on the diagnosis "healthy" in 45% of cases, and they agreed on the diagnosis "ill" in 18% of cases. Their probability of agreement is thus $p_0 = 70/110 = 45\% + 18\% = 63\%$. Yet, even if one rater (or both) were assigning diagnoses at random, the two of them would sometimes agree. The expected probability of agreement if this were so can be calculated from the marginal frequencies in the contingency table (which are given in the boxes marked "row totals" and "column totals"). Mathematically speaking, this is the situation called stochastic independence: One rater's judgment contains no information at all about the other rater's judgment.

For clarity in the following discussion, we will not always represent fractions and probabilities as percentages, but will sometimes write them as numbers between 0 and 1 instead: for example, 0.54, rather than 54%.

Now, if one rater were assigning diagnoses at random, we would expect agreement on the diagnosis "healthy" with probability $0.54 \times 0.73 = 0.39$, and agreement on the diagnosis "ill" with probability $0.45 \times 0.27 = 0.12$. The overall probability of agreement if one rater assigns diagnoses at random is thus $p_e = 0.54 \times 0.73 + 0.45 \times 0.27 = 0.52$ (57.2 of 110 cases). In other words, there would be agreement in 52% of cases, rather than 63%, as was actually observed. The observed probability of agreement exceeds the probability that would have been expected from random diagnosis by the amount $p_0 - p_e = 63\% - 52\% = 11\%$.

We now "norm" the excess frequency of agreement over chance in order to obtain a quantity that cannot be higher than 1. We do so by dividing the value $p_0 - p_e$, whatever it may be, by the highest value it can theoretically have, which is $1 - p_e$. In our example, $1 - p_e = 100\% - 52\%$. This theoretical highest value of $p_0 - p_e$ corresponds to the case where the raters agree 100% of the time. The normed value $\kappa_2 = (p_0 - p_e)/(1 - p_e)$ is called Cohen's kappa. In our example, Cohen's kappa has the value 11% / (100% – 52%) = 0.23.

Cohen's kappa equals 1 when the two raters agree in every case; it is 0 when they agree just as frequently as would have been expected if one rater (or both) were assigning ratings at random. Cohen's kappa hardly ever takes on its theoretical minimum value of –1.

resonance imaging (MRI) with spirometry for a specific clinical application using Bland-Altman diagrams (among other methods) and found the degree of agreement to be satisfactory.

The Bland-Altman diagram for *Example b (Figure 2b)* immediately reveals more than one limitation to the agreement of the two measuring techniques being investigated. The mean difference between the two measurements is once again near zero, but the limits of agreement are 1.4 units above and below the mean value, i.e., one can expect 95% of all measured differences to lie in the range –1.4 to +1.4. The physician must decide whether a deviation of this magnitude is acceptable. Moreover, the non-uniform distribution of the points in this diagram indicates systematic distortion (systematic bias).

Even so, however, poor agreement in a Bland-Altman diagram should not lead us to reject a new measuring technique prematurely. In *Figure 3,* two further cases (*Examples c and d*) are shown in which the two measuring techniques obviously do not agree (the plotted points lie far away from the line of agreement), yet they are nonetheless functionally related, as the regression curve shows in each case. The relation between the two techniques is linear in *Example c (Figure 3c),* nonlinear in *Example d (Figure 3d)*.

Thus, it often happens that one measurement can be accurately predicted from the other one because the two of them are clearly functionally related, even though the two measurements themselves yield very different values. In *Figure 3d,* for example, when Measurement 1 yields the value 3.0, we can use the regression curve to estimate that Measurement 2 will yield the value 7.65. The apparent lack of agreement between the two measuring techniques is thus largely correctable. Having "corrected" Measurement 2 in this way by means of the regression curve—which corresponds to our best estimate of the functional relation between the two measurements—we can compare the corrected Measurement 2 with Measurement 1 using the methods already described, e.g., a new Bland-Altman diagram. This procedure closely resembles the calibration of a measuring instrument. The determination of the functional relation itself, i.e., the generation of regression curves of the types seen in *Figure 3,* requires a variety of statistical methods, such as linear and nonlinear regression, that we cannot discuss here in any further detail.

The Pearson correlation coefficient (2) between the two measuring techniques is often considered to demonstrate a linear relationship (thus, a specific kind of functional relationship) between them. Indeed, a coefficient with a high absolute value (near 1 or –1) does indicate such a relationship. A common error, however, is to misinterpret the implications of significance tests that are applied to correlation coefficients. A finding that the correlation between two measuring techniques differs significantly from zero does not necessarily indicate that the two techniques

are in good agreement. Even the slightest, practically irrelevant relationship between two techniques could, in principle, yield a statistically significant finding of this type. A "significant" correlation actually contains no information at all about the size of the disagreement between the two types of measurement (3, 4).

Ratings on a nominal scale: Cohen's kappa

We now turn to the topic of ratings on a nominal scale. In medical research, the degree of agreement between two raters is often assessed with a measure called Cohen's kappa. Song et al. (5), for example, compared two methods of detecting bone metastases, which turned out to agree well, with a kappa value of 0.732. What Cohen's kappa measures, concisely stated, is the normed difference between the rate of agreement that is actually observed and the rate of agreement that would be expected purely by chance.

What does this mean in concrete terms? We will use a fictitious example to illustrate the use of Cohen's kappa for a dichotomous rating. The procedure for calculating Cohen's kappa is described in greater detail in *Box 1*. Let us assume that two doctors examine 110 patients for the presence of a particular disease and then state whether, in their opinion, each patient is or is not suffering from the disease (the "healthy"/"ill" dichotomy). We wish to know how closely the two doctors' diagnoses agree, i.e., how concordant they are. The 110 diagnoses of the two doctors are shown in the *Table* in *Box 1*.

The two doctors arrived at the same diagnosis in 70 of 110 cases. This figure alone, however, is not very useful in assessing concordance, because a certain number of like judgments would be expected even if one of the doctors (or both!) were assigning diagnoses entirely at random. On average, in this particular example, approximately 57 agreements would be expected by chance alone, as explained in *Box 1*. Cohen's kappa reflects the difference between this number (57) and the observed number of agreements (70), in relation to the total number of cases (110). In our example, Cohen's kappa is 0.23. The value of Cohen's kappa would be 1 if the two raters agreed in every case, i.e., if they were fully concordant; on the other hand, a value of 0 would indicate that the two raters agreed only as often as they would by chance. This would be a very discouraging finding indeed. A negative value of Cohen's kappa would indicate that the two raters agreed even less often than they would by chance, i.e., that they tended to make opposite judgments. A value of –1 would mean that the two raters arrived at opposite judgments in absolutely every case; this situation clearly arises very rarely, if ever.

The interpretation of a statistic such as Cohen's kappa is, in the end, arbitrary. Altman (2) suggested categorizing values of Cohen's kappa as shown in *Table 1*. In the example of *Box 1,* the calculated value $\kappa_2 = 0.23$ would be considered to indicate no more than a fair degree of agreement.

TABLE 1

Categorization of values of Cohen's kappa (2)

Value of κ_k	Quality of agreement
<0.20	Poor
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Good
0.81–1.0	Very good

BOX 2

Cohen's kappa for scales with k categories (k ≥ 2)

The contingency table below applies to the more general problem of comparing two raters who use a rating scale with an arbitrary number of categories—not necessarily two, as in our discussion up to this point. The row and column totals are $a_{i.} = a_{i1} + \ldots + a_{ik}$ und $a_{.i} = a_{1i} + \ldots + a_{ki}$, respectively. The overall frequency of agreement p_0 is given by $p_0 = 1/n \times (a_{11} + a_{22} + \ldots + a_{kk})$, i.e., p_0 equals the sum of the diagonal entries in the contingency table divided by the size of the sample. Now, two stochastically independent raters would agree at a frequency p_e that we calculate as $p_e = a_{1.}/n \times a_{.1}/n + a_{2.}/n \times a_{.2}/n + \ldots + a_{k.}/n \times a_{.k}/n$, i.e., p_e equals the sum of the products of the marginal frequencies relating to each diagonal entry of the general contingency table. Cohen's kappa for k categories is defined just as it was previously for k = 2, namely by the formula $\kappa_k = (p_0\text{-}p_e)/(1\text{-}p_e)$.

	1	2	. . .	k	
1	a_{11}	a_{12}	. . .	a_{1k}	$a_{1.}$
2	a_{21}	a_{22}	. . .	a_{2k}	$a_{2.}$
. . .	. . .	. . .	. . .	. . .	. . .
k	a_{k1}	a_{k2}	. . .	a_{kk}	$a_{k.}$
	$a_{.1}$	$a_{.2}$	. . .	$a_{.k}$	n

Box 2 generalizes the foregoing discussion to nominal rating scales with any number of categories, i.e., to scales with $k \geq 2$. Until now, we have only been discussing dichotomous scales, for which, by definition, $k = 2$.

Cohen's kappa gives a quantitative assessment of how well two raters agree, but the value of Cohen's kappa itself conveys no information about the reliability of this assessment. A value determined from a small number of patients is unreliable; therefore, as in many other situations in biomedical statistics, the value of Cohen's kappa should be stated together with the associated confidence interval *(Box 3)* (6).

In practice, Cohen's kappa is often used as a one-sided test of whether the interrater agreement is strong enough to rule out random judgments by (at

BOX 3

The confidence interval for Cohen's kappa

In general, the value of any descriptive statistic (e.g, a sample mean) conveys no information about its applicability to the overall population (not just to the sample on which it is based). For this reason, descriptive statistics are usually reported with a confidence interval. An approximate 1-α confidence interval for Cohen's kappa of the overall population can be calculated with the following formula:

$$CI := \kappa_k \pm z_{1-\alpha/2} \times \sqrt{\frac{p_0 \times (1-p_0)}{n \times (1-p_e)^2}}$$

In this formula, $z_{1-\alpha/2}$ is the(1-α/2) quantile of the standard normal distribution, whose values are listed in statistical tables (9). In the present case, we are making use of an approximation to the normal distribution. The following rule of thumb is often helpful: such an approximation is acceptably accurate as long as $n \times p_0 \geq 5$ and $n \times (1-p_0) \geq 5$.

We can now calculate a confidence interval for the numerical example presented above in *Box 1* and discussed in the corresponding section of the text. The sample size is n = 110; furthermore, as we have already calculated, $p_e = 0.52$, $p_0 = 0.63$, and $\kappa_2 = 0.23$. Setting α= 5%, we now read from a statistical table that $z_{0.975} = 1.96$. Using the formula above, we find:

$$CI = 0.23 \pm 1.96 \times 0.0959$$

Equivalently stated, the 95% confidence interval for Cohen's kappa of the overall population is 0.042 to 0.418.

KEY MESSAGES

- The mere demonstration that a correlation coefficient differs significantly from 0 is totally unsuitable for concordance analysis. Such tests are often wrongly used.
- The appropriate method for concordance analysis depends on the type of scale used by the measuring or rating techniques that are to be compared.
- The point-cloud diagram, the Bland-Altman diagram, and Cohen's kappa are suitable methods for concordance analysis.
- Concordance analysis cannot be used to judge the correctness of measuring or rating techniques; rather, it shows the degree to which different measuring or rating techniques agree with each other.

least) one rater. A statistically significant test result is often wrongly interpreted as an objective indication of agreement. In fact, this finding says next to nothing about the degree of agreement between the two raters: even a very low positive value of Cohen's kappa can be statistically significant, as long as the sample on which it is based is large enough. Thus, the use of significance tests to judge concordance is a mistake.

Cohen's kappa can be further refined and generalized. When comparing ordinal ratings, one may wish to give different weights to the differences between two consecutive ratings (e.g., so that the difference between "good" and "excellent" counts more than that between "fair" and "poor"). A weighted kappa is used for this purpose. Interrater agreement can be assessed in other situations, too, e.g., with more than two raters (7).

Sensitivity and specificity are often used to compare a dichotomous rating technique with a gold standard (8). These two statistics describe the degree of agreement between the technique in question and the gold standard, in each of the two subpopulations that the gold standard defines. In contrast, Cohen's kappa is a single quantity that provides a global assessment of the agreement between the technique in question and the gold standard.

Discussion

Statistical methods of assessing the degree of agreement between two raters or two measuring techniques are used in two different situations:

- ratings on a continuous scale, and
- categorical (nominal) ratings.

In the first situation, it is advisable to use descriptive and graphical methods, such as point-cloud plots around the line of agreement and Bland-Altman diagrams. Although point clouds are more intuitive and perspicuous, Bland-Altman diagrams enable a more detailed analysis in which the differences between the two raters are assessed not just qualitatively, but also quantitatively. The limits of agreement in a Bland-Altman diagram may be unsuitable for assessing the agreement between two measuring techniques if the differences between measured values are not normally distributed. In such cases, empirical quantiles can be used instead.

The distribution of the differences between two measured values can be studied in greater detail if, as first step, these differences are plotted on a histogram (3). In many cases, when the two measuring techniques are linked by a good linear (or other functional) relationship, it will be possible to predict one of the measurements from the other one, even if the two techniques yield very different results at first glance. The Pearson correlation coefficient is a further type of descriptive statistic; it indicates the presence of a linear relationship. A significantly nonzero correlation coefficient, however, cannot be interpreted as implying that two raters are concordant, as their ratings may still deviate from each other very strongly even when a significant correlation is present.

Cohen's kappa is a suitable tool for assessing the degree of agreement between two raters for categorical (nominal) ratings. A confidence interval for Cohen's kappa can be calculated as well.

Conflict of interest statement
The authors declare that no conflict of interest exists.

Manuscript submitted on 22 November 2010; revised version accepted on 11 May 2011.

Translated from the original German by Ethan Taub, M.D.

REFERENCES

1. Tetzlaff R, Schwarz T, Kauczor HU, Meinzer HP, Puderbach M, Eichinger M: Lung function measurement of single lungs by lung area segmentation on 2D dynamic MRI. Acad Radiol. 2010; 17: 496–503.
2. Altman DG: Practical statistics for medical research. 1st edition. Oxford: Chapman and Hall 1991; 1–611.
3. Altman DG, Bland JM: Measurement in medicine: the analysis of method comparison studies. The Statistician 1983; 32: 307–17.
4. Bland JM, Altman DG: Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1: 307–10.
5. Song JW, Oh YM, Shim TS, Kim WS, Ryu JS, Choi CM: Efficacy comparison between (18)F-FDG PET/CT and bone scintigraphy in detecting bony metastases of non-small-cell lung cancer. Lung Cancer 2009; 65: 333–8.
6. du Prel JB, Hommel G, Röhrig B, Blettner M: Confidence interval or p-value? Part 4 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2009; 106(19): 335–9.
7. Bortz J, Lienert G A, Boehnke K: Verteilungsfreie Methoden in der Biostatistik. 3rd Edition. Heidelberg: Springer 2008; 1–929.
8. Hilgers R D, Bauer P, Scheiber V: Einführung in die Medizinische Statistik. 2nd edition. Heidelberg: Springer 2007.
9. Altman DG, Machin D, Bryant TN, Gardner MJ: Statistics with confidence. 2nd edition. London: BMJ Books 2000.

Corresponding author
Dr. rer. nat. Robert Kwiecien
Institut für Biometrie und Klinische Forschung (IBKF)
Westfählische Wilhelms-Universität Münster
Albert-Schweitzer-Campus 1 – Gebäude A11
D-48149 Münster, Germany
robert.kwiecien@ukmuenster.de