

Agreement or Prediction: Asking and Answering the Right Question

S B Tan,*^{PhD}, S B Wee,**^{FAMS, MBBS, FRCS (Edin)}, Y B Cheung,***^{PhD}

Abstract

Introduction: Measuring agreement and measuring predictive ability are similar but distinct problems. Failure to appreciate the conceptual and practical differences may lead clinical researchers to give the right answer to the wrong question. **Methods:** We illustrate the relation and difference between measuring agreement and predictive ability in a non-technical way. We provide a real example investigating the feasibility of using preoperative breast cancer tumour size measurements to estimate postoperative histological size. The intraclass correlation and R-squared are calculated to ascertain the level of agreement and predictive ability respectively. **Results:** Analysis of agreement and analysis of predictive ability serve different purposes. The optimal solution found in terms of agreement may be different from that found for prediction. **Conclusions:** A careful clarification of the goal of an investigation is important. Using an inappropriate analysis can lead to misleading results, or to results that do not really answer the research question of interest.

Ann Acad Med Singapore 2002; 31:405-7

Key words: Intraclass correlation, Linear regression, R-squared

Introduction

In an article published in this journal, Chia¹ discusses the difference between measuring association and agreement. In this paper, we extend the discussion to the difference, in terms of the concepts as well as the practical usage, between analysing agreement and prediction. In particular, we consider the different problems of finding the variable that agrees most with a “gold standard” and finding the variable that is the best predictor of this.

We illustrate the differences between these two questions of interest by using data from a study of 44 breast cancer patients. The objective of the study was to determine which of 3 preoperative methods of measurements allow for the best estimate of the true tumour size, as based on post-operative histology measurement (i.e. the “gold standard”).

Concepts of Agreement and Prediction

Chia¹ gives a useful summary of some of the ways by which the level of agreement between two variables can be assessed. He notes that agreement, unlike association, is not summarised by measures such as the Pearson and Spearman correlation coefficients, which range from –1 to 1. Having a Pearson or Spearman correlation value that is close to 1 (or –1) means that there is strong positive (or

negative) association between the two variables being compared. However, this does not necessarily mean that there is a strong agreement. Suppose a group of patients are measured for a clinical characteristic twice, one by a new instrument being investigated and one by a standard instrument. If the readings x from the new instrument are consistently twice as large as readings from the standard instrument, y , the Pearson or Spearman correlation coefficient will have value 1. However, there is clearly no agreement between them.

A more suitable measure of the level of agreement is the intraclass correlation, which ranges from 0 to 1.^{1,2} A value closer to 1 indicates a stronger agreement, or vice versa. This essentially measures how well a 45° regression line, i.e. $y = x$, fits the data. If a measurement x always equals the gold standard measurement y , the intraclass correlation will be 1. This is illustrated in Figure 1.

The 45° regression line can be seen as

$$y = 0 + 1 \times x,$$

a linear regression line, with the intercept restricted to the value 0 and the slope restricted to the value 1.

In contrast to the above, methods for finding the best model for (a linear) prediction involve determining the

* Senior Biostatistician

*** Biostatistician

Division of Clinical Trials & Epidemiological Sciences

** Senior Consultant

Department of Surgical Oncology

National Cancer Centre, Singapore

Address for Reprints: Dr Cheung Yin Bun, Division of Clinical Trials & Epidemiological Sciences, National Cancer Centre, 11 Hospital Drive, Singapore 169610.

E-mail: ctecyb@nccs.com.sg

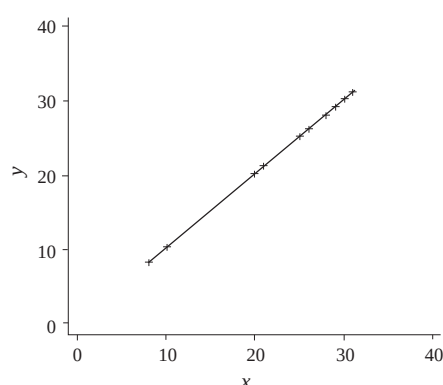


Fig. 1. Hypothetical example of good agreement. Data points indicated by “+” signs fall along a 45° line, $y = x$.

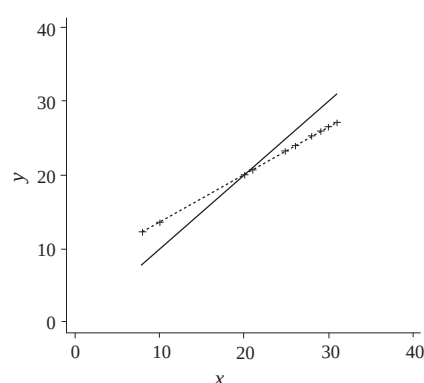


Fig. 2. Hypothetical example of poor agreement but good prediction. Data points indicated by “+” signs fall along the dotted line $y = 7 + 0.65x$. A 45° (solid) line is included for reference.

regression line with general form

$$y = a + b \times x,$$

that best fits the data points. The intercept and slope of the line are estimated from the data, instead of being pre-specified as 0 and 1 respectively. In general, this line is not constrained to be a 45° line and an example of such a line is given in Figure 2.

From this Figure, it can be seen that the data points do not fall along a 45° line. Indeed x tends to over-estimate smaller values of y and under-estimate larger ones. The intraclass correlation is 0.91, which quantifies the level of agreement. Although there is some disagreement, this is taken account of via the non-zero intercept and the non-unit slope of the linear regression line. Thus, the prediction of y based on x can potentially still be good. In fact, for this example, y can be perfectly predicted by the equation $y = 7 + 0.65 \times x$.

One way of quantifying how well the regression line fits the data is via the R-squared value.³ The R-squared value indicates the proportion of variation that is explained by the regression line. The closer the R-squared value is to 1, the better the regression line fits the data. The R-squared value is calculable for regression analysis of continuous data. There are various “pseudo R-squared” measures for logistic

regression of binary data and Poisson regression of count data, etc, which are beyond the scope of this paper. Interested readers are referred to Long and Freese.⁴ Since a 45° line is a special case of an ordinary linear regression line, if y and x indeed perfectly agree, the linear regression line will become $y = 0 + 1 \times x$, that is $y = x$.

The correct choice for the data analysis depends on what is the real question of interest to the investigators. If they are interested in finding a new measurement method as a “direct replacement” for the gold standard, they would need to estimate the level of agreement. The intraclass correlation would be a suitable measure of this. However, if the purpose is to find a way to predict the gold standard measurement, there is no reason to restrict the prediction equation to be $y = x$. Carrying out an ordinary regression analysis would be better.

We now consider an example, using data on breast cancer, to illustrate what we have discussed so far.

Example I

A retrospective study was conducted to examine the feasibility of using preoperative tumour size measurements to estimate the tumour size determined on postoperative histology (the gold standard). This is important for the purpose of patient management, as not all tumour sizes can be determined on postoperative histology, for example tumours that are removed piecemeal. A total of 44 cases of breast cancer from which information on preoperative measurements based on clinical judgement, mammography and ultrasound, together with their histologically measured tumour size were analysed.

The goal of the study, as initially expressed by the clinical investigators, was vaguely put as to find out which of the 3 preoperative methods of measurement allow for the best estimate of the true tumour size to be made. After careful discussion, it was clarified that what the investigators really wanted to do was to predict the tumour size based on histology. It was not necessary to find a measurement that could be used as a direct replacement of the histology size. As such, there was no need to restrict the analysis to a 45° line and an ordinary regression analysis would be appropriate.

Table I gives both the intraclass correlation values as well as the R-squared values for each of the three preoperative variables, when fitted individually. The mammographic measure has an intraclass correlation value that is much higher than that of the clinical measure and marginally higher than that of the ultrasound measure. It thus agrees best with the gold standard. However, ultrasound has the largest R-squared value. Thus, the ultrasound values give the best prediction of the histology values. This was what the investigators wanted to obtain.

Example II

There are situations in which the research purpose is primarily to assess agreement. In the course of developing a quality-of-life (QoL) questionnaire, an almost standard procedure is to check the test-retest reliability. The QoL questionnaire is administered to a group of patients twice, separated by a short period of time. The purpose of the test-retest procedure is to check whether the QoL instrument is clearly written and its components are clearly defined, so that the same patient answering the same questionnaire shortly again will give essentially the same answers. A strong agreement between the test and retest measures will support the reliability of the questionnaire. Otherwise, the developers need to examine the reasons why the instrument is not reliable and modify it accordingly. As such, the analysis is based on agreement. The intraclass correlation is a standard method for the assessment of test-retest reliability.² Neither the first nor the second measurement is more important or is a gold standard. Prediction is not relevant here.

Discussion

To make a decision on which analytic method to use requires a careful clarification of the research purpose and how the research output is expected to be used. The assessment of agreement is preferred if the goal is to find an alternative measure as a direct replacement of the gold standard. In such a case, the intraclass correlation is a useful measure of the level of agreement. However, clinicians do not always need to find a direct replacement. Most clinicians are familiar with the concept of a simple regression analysis. The use of a simple regression line is also common in clinical practice. For instance, maximum attainable heart rate can be predicted based on a linear regression line with age as the regressor. Since agreement is a restricted form of prediction, with zero-intercept and unit slope, it results in poorer predictions than that obtained using the more general form of a simple linear regression. Thus, when the purpose is to obtain the best prediction, as in the breast cancer example, one should use a prediction method such as linear regression.

The breast cancer example is used for illustration only. It can be seen that all the intraclass correlation values, as well as the R-squared values in Table I are quite low. This suggests that none of the variables are in good agreement nor make good predictors of the histologically determined tumour size. We have noted that agreement is a restricted form of simple linear regression. Indeed, simple linear regression is itself a restricted form of regression analysis, which may involve multiple regressors and non-linear relationships. In the full analysis of the data set, non-linear forms as well as multiple regression models involving all 3 preoperative estimates of tumour size were also

TABLE I: ANALYSIS OF 3 PREOPERATIVE MEASUREMENTS OF BREAST TUMOUR SIZE IN RELATION TO THE TUMOUR SIZE DETERMINED ON HISTOLOGY IN 44 PATIENTS

	Intraclass correlation	R-squared value	Regression equation
Clinical	0.2481	0.1707	$y = 9.8093 + 0.2265 \times x$
Ultrasound	0.4085	0.2245	$y = 9.3591 + 0.3064 \times x$
Mammographic	0.4097	0.2093	$y = 8.7649 + 0.3424 \times x$

considered. These models gave better predictions, based on their R-squared values. However, to keep things simple and to avoid detracting from the focus of this article, we do not report the results of these models here. (Note that in multiple regression, one may want to use an “adjusted R-squared” that penalises the complexity induced by using more variables.³) We have also not discussed alternative ways of assessing the level of agreement and prediction. Like Chia¹ and others, we stress that conclusions should not be made solely on the basis of one summary index alone.

In conclusion, it is important to make a distinction between the question of wanting to find the variable that agrees most with the outcome of interest, and wanting to find the one which is the best predictor of it. For the first question, identifying this variable will allow us to use it as a direct replacement for the outcome of interest. There would be no need to multiply or add any constant term(s) to the variable in order to get an estimate of the outcome. It offers simplicity at the expense of lower accuracy in prediction, if this is relevant.

Answering the second question provides us with the variable and predicting equation that give the most accurate prediction of the outcome of interest. In general, this would be a better estimate of the outcome than that given by $y = x$. However, the disadvantage is that we no longer have the convenience of having a direct replacement variable. As shown in the example, the optimal variable in terms of agreement and prediction may not be the same. It is therefore important and necessary for clinical investigators to clearly identify what is their real question of interest and how they are going to use the outputs.

REFERENCES

1. Chia K S. Association or agreement. *Ann Acad Med Singapore* 2000; 29:263-4.
2. Streiner D, Norman G R. *Health Measurement Scales: A Practical Guide to their Development and Use*. Oxford: Oxford Univ Press, 1989.
3. Montgomery D C, Peck E A, Vinjing G G. *Introduction to Linear Regression Analysis*. New York: John Wiley & Sons, 2001.
4. Long J S, Freese J. Scalar measures of fit for regression models. *Stat Tech Bull Reprints* 2001; 10:197-205.