Reliably Assessing Reliability with SAS® Software

P.M. Simpson, University of Arkansas for Medical Sciences, Little Rock, AR
S. Lensing, University of Arkansas for Medical Sciences, Little Rock, AR
P.R. Phillips, Australian National University, Australia
R. Hamer, UMDNJ, Robert Wood Johnson Medical School, Piscataway, NJ

Introduction

In the medical field instruments are used for monitoring, diagnosis and screening; that is, they enable measurement, classification, and selection. Non-invasive efficient tools are important. In particular, in this day of cost-effective medicine, cheaper methods are constantly being sought. However, care must be taken that the instrument used does indeed do the job for which it is used, or, if it is less effective, where its possible shortcomings lie. How then should reliability be assessed?

Protocols in which two or more measuring instruments are compared are called method comparison studies. In this context, the reliability is defined as how well the two instruments agree. Agreement relates to a range of concepts. One idea they share is the determination of how well measurements made with one instrument may be substituted for measurements made with the other instrument. Often, one method is the gold standard and the results using the new instrument must be nearly identical. Ideally, the measurements from the new instrument will also be less expensive, easier to obtain, and/or quickly available. In fact, if the tool is to be used to detect when a patient is in danger (monitoring), has a certain disease (diagnosis), or should be included in a study (screening), the similarity may only require identical classification. However, even if classification is the aim, how “far away” a person is from a cutoff value (precision) may be important. Thus, acceptable substitution of one instrument for another depends on both correct patient classification and the size of the error.

In this presentation, we will give examples in which substitution of a tool based on a high Pearson correlation coefficient may result in incorrect inclusion and exclusion of patients, as well as false positive and false negative results for efficacy. We will discuss procedures available in SAS that can be reliably used to assess reliability, focusing on method comparison studies. We will emphasize the difficulty of locating the SAS documentation for reliability statistics, as many of the reliability statistics have become available in SAS in recent years. Finally, we will indicate why, possibly, the issue of reliability may still be unresolved.

Pearson Correlation

Despite many articles discussing and criticizing reliability methodology (Dunn, 1992; Dunn, 1992), there are still widespread incorrect assessments of reliability, as shown by the prevailing use of the Pearson correlation coefficients. We will present examples of method comparison studies from the medical literature, which were published and peer-reviewed, that misuse the Pearson correlation coefficient. A list of references from which these examples were taken is available upon request.

Multiple papers cite the p-value as evidence of reliability when a significant Pearson correlation means a correlation significantly different from 0 (Chinchilli and Gruemer, 1995). The Pearson statistic will be presented when the scatter about the line is much of the time nonrandom, or the Pearson correlation will be reported when it is inflated due to a few outliers. Mathematically it can be shown that a value arbitrarily close to 1 can be obtained if there is enough separation of a few values from a group of the rest (Turner, 1997). We will describe the shortcomings of statistics that have been proposed as solutions to these problems associated with the Pearson

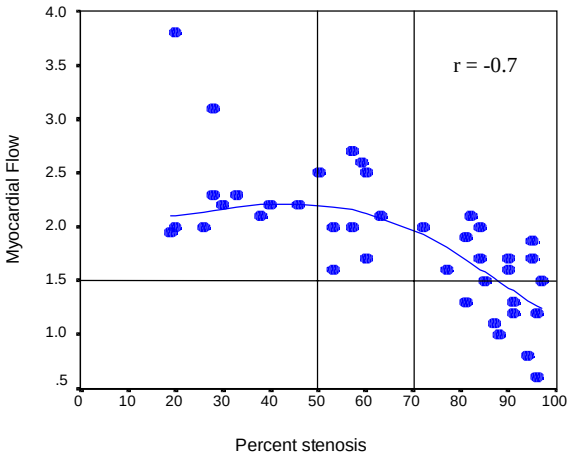

Figure 1. Myocardial blood flow by Positron Emission tomography versus with percent stenosis.

correlation coefficient.

The Pearson correlation will often be reported when there is a nonlinear relationship between variables as in Figure 1, which is from a published article incorrectly claiming that myocardial blood flow by positron emission

tomography can differentiate coronary lesions of 50% to 70% stenosis from 70% to 90%.

Often studies will involve multiple institutions and the data will be combined. The data in Figure 2 is from another published article comparing two imaging methods. In Figure 2, there appears to be good agreement for the combined data as the points exhibit a random scatter around the line y = x. However, when we examine the trend according to institution, we see that different institutions exhibit different patterns. Furthermore, in this study, four observations per patient were used in the plot and in the calculation of the correlations. Repeated measurements from the same patient will usually result in an overstated correlation. It is notable that the Pearson correlation coefficient measures how close the data is to the line of best fit **not** as it should in many cases to the line of agreement y = x (Carroll et al. 1995; Linnet, 1990).

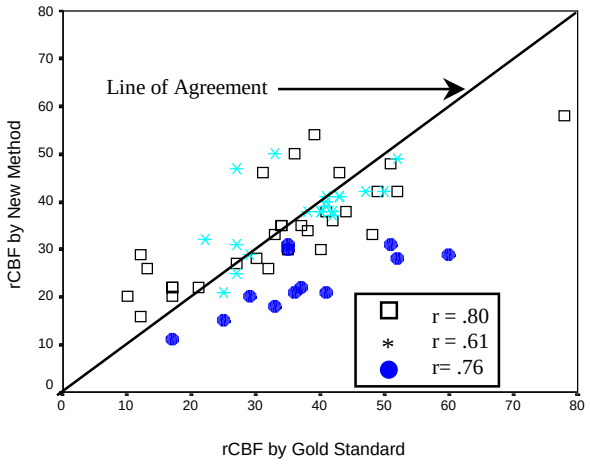

Figure 2. Comparison of two imaging methods done at three locations

Other strategies

The aforementioned examples demonstrate the inappropriate use of the Pearson correlation coefficient. However, the best statistic for measuring reliability is not clear cut. Although there are many modeling approaches for agreement, including causal modeling, it is clear that an index to rate reliability is attractive in its simplicity. An alternative statistic, the concordance correlation coefficient offers advantages because it addresses the issue of agreement. However, it can be misleading in that it summarizes the fit around the line of identity and therefore, like the Pearson correlation a value close to one may not denote lack of variability around the line. Another possibility is the intraclass correlation coefficient or its analogues (Bland and Altman, 1986; Bland and Altman, 1996). However, the value of this method depends heavily on the sample used, and without repeat measurements, estimates of precision are impossible (Giraudeau et al. 1996; Lyles and Chambless, 1995; Muller and Buttner, 1994; Vargha, 1997; Muller and Buttner, 1997; Lyles and Chambless, 1995; Shrout and Fleiss, 1979). If a cutpoint is to be used to classify patients, agreement of the classifications could be examined, using Kappa indices. Kappa is commonly used to measure reliability or agreement or nominal or ordinal variables; however it also has limitations(Wieman et al. 1989) . Modelling again may present a more complete but less simple and therefore, to many, less attractive summary of the agreement in the data. If one method is a gold standard then predictivity, using sensitivity, specificity or allied statistics should be determined (Metz and Shen, 1992). ROC analysis has many summary measures; one of which will suit every occasion (Brenner and Kliebsch, 1996; Phelps and Hutson, 1995; Somoza, 1996; Dwyer, 1997; Zhou and Gatsonis, 1996). For example if high sensitivity is required there is an index based on ROC analysis. Even though, ROC type analyses have much to offer in comparing methods, it can be argued that in paying attention to the misclassification rather than the consequences of misclassification, there may not result in an appropriate comparison (Obuchowski, 1997). The method of assessing reliability **must** reflect the medical use of the instruments.

Reliability in SAS

SAS/STAT® software has many procedures that provide reliability analyses. We will give an overview of procedures available for reliability analyses, giving primary focus to those that involve comparisons to a gold standard.

For continuous data, causal modeling is appropriate for when both measurements are assumed to have measurement error (Bentler and Stein, 1992; Rindskopf and Rindskopf, 1986). The CALIS procedure will perform these types of analyses. When there are independent measurements and the x variable representing one instrument is assumed to have no measurement error, general linear model approaches, such as the REG, ANOVA, and MIXED procedures, can be used (Quan and Shih, 1996; Agresti, 1992). The appropriate mean square errors are used to calculate an intraclass correlation coefficient or a concordance correlation coefficient. Often, this calculation is done by hand with values taken from the output, since these statistics are not automatically generated by SAS. Several types of intraclass correlation coefficients are available in a SAS macro at http://www.sas.com/techsup/download /stat /intracc.sas . A macro to calculate confidence intervals for the intraclass correlation coefficient is available from R. Hamer, e-mail: hamer@rci.rutgers.edu. Other times, we are not fortunate enough to have a macro written by someone else, and if we want to automate a reliability procedure, we have to program it ourselves. Several lines of code may be required as in the following SAS code that will give the concordance correlation coefficient and 95% confidence limits.

```
%macro cp;
proc corr data=one cov outp=covdat noprint;
  var x1 x2;
run;
```

```
data cp;
  set covdat;
  if _type_='COV' then do;
    if _name_='X1' then call
      symput('COV12',x2);
    if _name_='X1' then call
      symput('COV1',x1);
    if _name_='X2' then call
      symput('COV2',x2);
  end;
  if _type_='MEAN' then do;
    call symput('MEAN1',x1);
    call symput('MEAN2',x2);
  end;
  if _type_='N' then call symput('N',x1);
  if _type_='CORR' and _name_='X1' then call
    symput('RHO_HAT',x2);
  pc_hat=(2*&cov12)/(&cov1+&cov2+(&mean1-
    &mean2)**2);
  z_hat=0.5*log((1+pc_hat)/(1-pc_hat));
  u2=(mean1-mean2)**2/sqrt(cov1*cov2);
  ratio1=((1-rho_hat**2)*pc_hat**2)/((1-
    pc_hat**2)*rho_hat**2;
  ratio2=(4*(pc_hat**3)*(1-
    pc_hat)*u2)/rho_hat*(1-pc_hat**2)**2;
  ratio3=(2*(pc_hat**4)*u2*u2)/
    ((rho_hat**2)*(1-pc_hat**2)**2);
  sig_hat=(ratio1+ratio2-ratio3)/(n-2);
  lower95=tanh(z_hat-1.96*(sig_hat));
  upper95=tanh(z_hat+1.96*(sig_hat));
run;

proc print;
  var cp_hat lower95 upper95;
run;

%mend;
```

Cronbach's Coefficient Alpha (Chronbach, 1951) is a measure to quantify the internal consistency of a test consistinng of a number of items. This internal consistency is one of the things that test theorists call reliability. The CORR procedure in SAS has an ALPHA option which will give this coefficient. However, this estimate of reliability is flawed in that the more items the test has the higher the coefficient will tend to be. (The variance of a score based on the number of items will tend to decrease as the number of items increases).

For binary data, the LOGISTIC procedure gives sensitivity, specificity, and area under the ROC curve, labeled "c" in the output. The cutoff value for classifying observations is 0.50 unless otherwise specified using the PPROB model option. This will not be adequate for more general situations suggested in the literature cited above.

For categorical scales, to assess agreement among 2 or more observers, the Kappa statistic is available in the FREQ Procedure. This will calculate the original (unweighted) kappa statistic and does allow for weighted elements where the weights reflect the relative spacing of ordered categories. However, it does not allow for weights which are customized. For example, if two values were considered to agree if they were within one category of each other, a weighted kappa could be calculated but not by present SAS software. . Other indices of predictive ability are Somers' D, Goodman-Kruskal Gamma, and Kendall's Taus. All these look at the number of concordant and discordant pairs. They all have problems. The major one is that, unlike Kappa, they do not adjust for the number of concordant pairs that would be expected by chance alone. In addition there are other "agree" statistics, which really do not look at agreement but aspects of association. These include test statistics like McNemar's and Cochran's Q test , McNemar's generalization from 2x2 tables. These look at a very limited aspect of agreement, namely marginal homogeneity. More complex modeling of agreement for categorical outcomes can be done using the GENMOD procedure.

One very important procedure that is often overlooked in reliability analyses is the SAS/GRAPH® software; visualizing the data is often the best way to characterize the reliability or lack of reliability between measures. A simple scatter plot of measurements from a new instrument against those from gold standard can be very insightful. Sometimes when the range of values is very narrow, then plots of the differences from the gold standard versus the gold standard value can indicate lack of agreement when a simple scatter plot does not. When there is no gold standard, then plots of the differences versus the average value of the two instruments, which is the best estimate of the true value in this situation, are very helpful.

SAS/QC®, aimed at quality control issues, has extensive graphing capabilities to look at reproducibility and reliability in the GAGE application. Intra and inter method variability can be assessed also. It should be noted that the RELIABILITY procedure performs failure time (survival) analyses, which are conceptually different from our definition of reliability analyses.

Many of the reliability statistics are just available within recent years in SAS, and therefore, they are only found in the *Changes and Enhancements* manuals. Reliability analyses are performed using the agreement statistics, such as Kappa, in the FREQ Procedure and using the GENMOD and MIXED Procedures; these particular methods are found only in the SAS/S*TAT*® *Software Changes and Enhancements* (SAS, 1996) and not the *SAS/STAT*® *User's Guide* (SAS, 1989). The Pearson correlation coefficient is easily found in the SAS/STAT® *User's Guide* which is where the novice SAS user or data analyst is likely to begin and unfortunately end their search for available reliability statistics.

Even with the reliability statistics available in SAS, no procedure will automatically produce an intraclass correlation or a concordance correlation coefficient, but it is up to us to correctly calculate one using a general linear models procedure. Furthermore, SAS only provides limited ROC analyses, reporting only simple statistics. Nonparametric ROC analyses and analyses to compare areas under the curve are not automatically available in SAS.

Conclusion

The incorrect use of the Pearson correlation coefficient may be attributed to both the readily availability of the statistic and the lack of understanding about distinguishing a statistically significant correlation from a meaningful predictive value. The issue of reliability may be considered unresolved as no single approach or algorithm really can be universally recommended. The use of any instrument mandates the examination of reliability both in selection of experimental design and in type of analysis.

References

Agresti, A. (1992) Modelling Patterns of Agreement and Disagreement. *Statistical Methods in Medical Research* **1**, 201-218.

Bentler, P.M. and Stein, J.A. (1992) Structural Equation Models In Medical Research. *Statistical Methods in Medical Research* **1**, 159-181.

Bland, J.M. and Altman, D.G. (1986) Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement. *The Lancet* 307-309.

Bland, J.M. and Altman, D.G. (1996) Measurement Error and Correlation Coefficients. *British Journal of Medicine* **313**, 41-42.

Brenner, H. and Kliebsch, U. (1996) Dependence of Weighted Kappa Coefficients on the Number of Categories. *Epidemiology* **7**, 199-202.

Carroll, R.J., Rupport, D. and Stefanski, L.A. (1995) Regression and Attenuation. In: Chapman and Hall, (Eds.) *Measurement Error in Nonlinear Models*, pp. 21-39. London:

Chinchilli, V.M. and Gruemer, H.D. (1995) The Correlation Coefficient in the Interpretation of Laboratory Data. *Clinica Chimica Acta* **236**, 113-117.

Chronbach, L.J. (1951) Coefficient Alpha and the internal structure of tests. *Psychometrika* **16**, 297-334.

Dunn, G. (1992) Design And Analysis Of Reliability Studies. *Statistical Methods in Medical Research* **1**, 123-157.

Dunn, G. (1992) Design and Analysis of Reliability Studies. *Statistical Methods in Medical Research* **1**, 123-157.

Dwyer, A.J. (1997) In Pursuit of a Piece of the ROC. *Radiology* **202**, 621-625.

Giraudeau, B., Mallet, A. and Chastang, C. (1996) Case Influence on the Intraclass Correlation Coefficient Estimate. *Biometrics* **52**, 1492-1497.

JMP Statistics and Graphics Guide 1995. [Computer Program]. Cary, NC: SAS Institute, Inc.

Linnet, K. (1990) Estimation of the Linear Relationship Between the Measurements of Two Methods with Proportional Errors. *Stat Med* **9**, 1463-1473.

Lyles, R.H. and Chambless, L.E. (1995) Effects of Model Misspecification in the Estimation of Variance Components and Intraclass Correlation for Paired Data. *Statistics in Medicine* **14**, 1693-1706.

Lyles, R.H. and Chambless, L.E. (1995) Effects of Model Misspecification in the Estimation of Variance Components and Intraclass Correlation for Paired Data. *Stat Med* **14**, 1693-1717.

Metz, C.E. and Shen, J.H. (1992) Gains in Accuracy from Replicated Readings of Diagnostic Images: Prediction and Assessment in Terms of ROC Analysis. *Med Decis Making* **12**, 60-75.

Muller, R. and Buttner, P. (1994) A Critical Discussion of Intraclass Correlation Coefficients. *Stat Med* **13**, 2465-2476.

Muller, R. and Buttner, P. (1997) Authors' Reply to "A Critical Discussion of Intraclass Correlation Coefficients". *Stat Med* **16**, 821-823.

Obuchowski, N.A. (1997) Testing for Equivalence of Diagnostic Tests. *American Journal of Roentgenology* **168**, 13-17.

Phelps, C.E. and Hutson, A. (1995) Estimating Diagnostic Test Accuracy Using a "Fuzzy Gold Standard". *Med Decis Making* **15**, 44-57.

Quan, A.H. and Shih, W.J. (1996) Assessing Reproducibility by the Within-Subject Coefficient of Variation with Random Effects Models. *Biometrics* **52**, 341-353.

Rindskopf, D. and Rindskopf, W. (1986) The Value of Latent Class Analysis in Medical Diagnosis. *Stat Med* **5**, 21-27.

SAS 1989. SAS/STAT® User's Guide. Cary, NC: SAS Institute, Inc.

SAS 1996. SAS/STAT® Software: Changes and Enhancements through Release 6.11. Cary, NC: SAS Institute, Inc.

Shrout, P.E. and Fleiss, J.L. (1979) Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychol Bull* **86**, 420-428.

Somoza, E. (1996) Eccentric Diagnostic Tests: Redefining Sensitivity and Specificity. *Med Decis Making* **16**, 15-23.

Turner, D.W. (1997) A Simple Example Illustrating a Well-Known Property of the Correlation Coefficient. *Am Statistician* **51**, 170

Vargha, P. (1997) Letter to the Editor. "A Critical Discussion of Intraclass Correlation Coefficients". *Stat Med* **16**, 821-823.

Wieman, T.J., Huang, K.C., Tsueda, K., Thomas, M.H., Lucas, L.f. and Simpson, P. (1989) Peripheral Somatic Sensory Neuropathy and Skin Galvanic Response in the Feet of Patients with Diabetes. *Sugery, Gynecology and Obstetrics* **168**, 501-506.

Zhou, X.H. and Gatsonis, C.A. (1996) A Simple Method for Comparing Correlated ROC Curves Using Incomplete Data. *Statistics in Medicine* **15**, 1687-1693.

Pippa Simpson, Ph.D.
Arkansas Children's Hospital
PEDS/CARE, So. Campus Room 301
Little Rock, Arkansas 72202
(501) 320-6631
PSimpson@care.ach.uams.edu