PharmaSUG 2013 - Paper PO07

Reliability Assessment of Image Data in Oncology and Psychology Studies

Li Zhang, Independent Consultant, Chesterbrook, PA

David Shen, Independent Consultant, Chesterbrook, PA

Gary Chen, Shire Pharmaceuticals, Chesterbrook, PA

ABSTRACT

In oncology studies, the RECIST provides the method of evaluating solid tumor response using MRI, CT scans and X-ray. By using brain-imaging-techniques like fMRI, cognitive psychology is able to analyze the relation between the physiology of the brain and mental processes. It is crucial to obtain accurate and reliable measurements in clinical studies. The analysts need to undergo comprehensive reliability assessment measures, including inter-reliability, prior to study commencement to verify measurement consistency and ensure a high level of agreement across analysts. The purpose of this study is to evaluate the proposed analysis methods to check for consistency and precise of measurements for imagine data.

INTRODUCTION

RECIST (Response Evaluation Criteria In Solid Tumors), recommended by the National Cancer Institute for sponsored trials, is a set of international standards that are used to determine efficacy for oncology studies. Typical study endpoints used to analyze RECIST data include:

- Progression Free Survival (PFS)
- Overall Survival (OS)
- Best Objective Tumor Response
- Duration of Response

Since efficacy endpoints are based on tumor assessments, clinical trials that assess the efficacy of cancer therapies are becoming increasingly dependent on imaging analysis data as surrogate endpoints. Patients with malignancy typically present with lesions that require serial magnetic resonance imaging (MRI), computed tomography (CT), X-ray and position emission tomography (PET) scans. Imaging data is important to determine whether a lesion is responding to treatment completely or partially, stable or progressive, and if so, at what rate, in evaluating response to cancer therapy.

Classical psychology infers a participant's mental states from the behavior that is shown in the scores of instruments. The present cognitive psychology developing from the information processing, introduces imagine information processing mainly computer-based techniques. By using brain-imaging-techniques like cognitive psychology is able to analyze the relation between the physiology of the brain and mental processes. Functional magnetic resonance imaging (fMRI) is an non-invasive imaging method that pictures active structures of the brain in a high spatial resolution. This picture shows the brain and its activated parts. Other Non-invasive functional neuroimaging to scan the brain includes: electroencephalogram (EEG), positron emission tomography (PET),, magnetoencephalography (MEG), optical imaging (near infra-red spectroscopy or NIRS), anatomical MRI, and diffusion tensor imaging (DTI) The findings of cognitive neuroscience are directed towards enabling a basic scientific understanding of a broad range of issues involving the brain, cognition and behavior.

Treatment decisions, particularly in reference to longitudinal assessments, are based on the information in these scans. The quality of imagine data is very important to ensure the power and precise of the outputs and results. A measurement validation in quantifying medical images is required to guarantee the consistency of measurements. In this article, various statistical techniques, including correlation analysis, linear regression, Bland-Altman method, paired *t*-test, analysis of variance (ANOVA) and ICC are tested and discussed in determining measurement consistency. The focus is on detecting a possible measurement bias and determining the robustness of the procedures.This is the paper body. This is the paper body. This is the paper body. This is the paper body. This is the paper body. This is the paper body. This is the paper body.

METHOD FOR ANALYSIS

1. Correlation Analysis

Correlation analysis is used to see if the values of two raters are associated. Simple correlation analysis can be conducted by PROC CORR, which providing Pearson correlation statistics, and probabilities for the variable.

Correlation coefficient contains information on both the strength and direction of the association between two numeric variables. If measurements are strongly associated, it is expected to have a correlation value close to 1. In contrast, if the measurements are less associated, a correlation value would be reduced to 0. Under the null hypothesis of r =0 (not associated), the significance of correlation is tested using a t-statistic with n -2 degrees of freedom.

```
proc corr data = lesion_a ;
  var measure1 measure2;
run;
```

The P-value is the probability that you would have found the current result if the correlation coefficient were in fact zero (null hypothesis). If this probability is lower than the conventional 5% (P<0.05) the correlation is called statistically significant.

Since correlation only looks at the association rather than the agreement between two measurements, correlation may inaccurately estimate the agreement of the relationship.

2. Linear Regression

The correlation analysis has been previously used in measurement consistency. Alternately a linear regression can be used to determine the measurement consistency. The following regression model is used to fit measurements: Y=a + bX

When imaging data are consistent, the regression slope would be close to 1. Similarly one can test whether the intercept is close to 0 for testing a bias if one rater is systematically obtaining larger or smaller measurements compared to the other rater.

Regression is used to describe the relationship between two variables and to predict one variable from another. The linear equation can reflect the agreement between two measurements. The intercept a should equal to 0 or around 0 while slope b should equal to 1 or very close to 1 if the results from two methods are comparable.

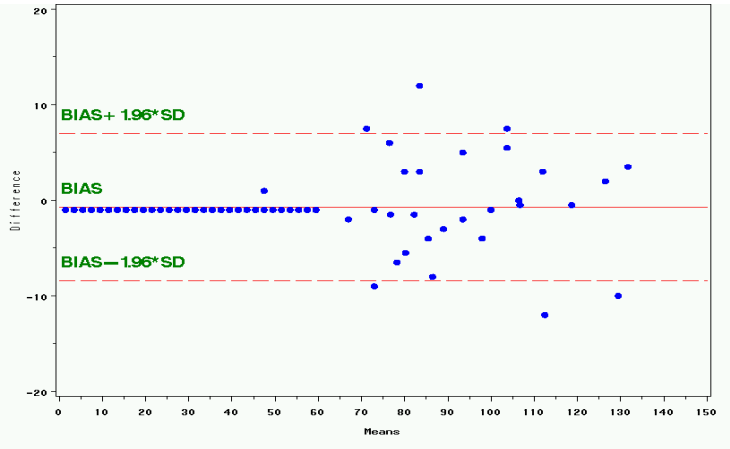
```
proc reg data=lesion_a;
  model measure1 = measure2;
run; quit;
```

The statement added to the PROC REG can explicitly test whether the slope value is 1 or not.

```
proc reg data = lesion_a;
  model measure1 = measure2;
  slope: test measure2 = 1;
run;
quit;
```

3.Bland-Altman Method

The figure shows the Bland-Altman plot for the measurements from two raters. Bland-Altman plot displays the differences in measurements on the y-axis versus the average measurement obtained from the two raters on the x-axis. Let d and S_d^2 be the mean and the variance of the difference. Bland and Altman plotted di versus the average of measurements of two raters, with the reference lines, d - 1.96*Sd, and d +1.96*Sd. The range between d - 1.96*Sd, and d +1.96*Sd. provides the “limit of agreement”. 95% of differences should lie between these two lines. This is called 95% limits of agreement method. It's simple and easy to express and interpret the data.



4. Paired *t*-Test

The weakness of the Bland-Altman method is that the measurement consistency is determined visually without statistical significance attached to the plot.

To give the statistical significance to the Bland-Altman method procedure, a paired *t*-test can be used. We test whether the measurement difference is statistically small enough using the test statistic which is distributed as the *t*-distribution with *n* - 1 degrees of freedom.

```
proc ttest data=lesion_a;
  paired measure1 * measure2;
run;
```

The paired *t*-test may indicate that there is significant inconsistency in measuring, although the scatterplots show measurement consistency. This contradiction can happen if one rater's measurements are systematically either larger or smaller than those of the other rater. When this systematic bias becomes larger than the measurement variance, this contradiction will happen.

5.ANOVA

When there are more than two raters, the paired *t*-test cannot be applied directly without significant modification. The analysis of variance (ANOVA) approach is porposed for more general cases. The strength of ANOVA is that it can be used to determine both between- and within-rater measurement consistency. If we have information about how each rater measures the same MR image consistently, we can determine who is more consistent. This additional information can be used to further train less consistent raters.

Let X_{ijk} be the kth measurement on the jth MR image by the ith rater. Then, the two-way ANOVA model is given as $X_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$.

To study the influence of the qualitative (discrete) factors on another (continuous) variable, we can use analysis of variance by PROC GLM.

```
proc glm data = lesion_b;
  class subject rater ;
  model measure  =  subject rater;
run;
quit;
```

To evaluate the additional unknown random effects of SUBEJCT to the results, we can use MIXED procedure. In the mixed model, the variances of the random-effects parameter SUBJECT assumed to impact the variability of the data become the covariance parameters for the mixed model.

```
proc mixed data= lesion_b;
  class rater subject;
```

```
  model measuret= rater /ddfm=satterth ;
  random subject;
  lsmeans rater/pdiff cl alpha=.05;
  estimate 'Measure1 vs. Measure2'  rater 1 -1;
run;
```

The RANDOM statement defines the random effects SUBJECT to constitute the vector in the mixed model. The DDFM=SATTERTH option performs a general Satterthwaite approximation for the denominator degrees of freedom. PROC MIXED also provides several different statistics suitable for generating hypothesis tests and confidence intervals. The validity of these statistics depends upon the mean and variance covariance.

6. Intra-Class Correlation (ICC)

Another way of performing reliability testing is to use the intra-class correlation coefficient (ICC) The range of the ICC may be between 0 and 1. The ICC will be high when there is little variation between the scores given to each item by the raters, e.g. if all raters give the same, or similar scores to each of the items. The ICC is an improvement over Pearson's r and Spearman's ρ, as it takes into account of the differences in ratings for individual segments, along with the correlation between raters. ICC assesses rating reliability by comparing the variability of different ratings of the same subject to the total variation across all ratings and all subjects. The theoretical formula for the ICC is:

$$ICC = \frac{\sigma^2(b)}{\sigma^2(b) + \sigma^2(w)}$$

where $\sigma^2(w)$ is the pooled variance within subjects, and $\sigma^2(b)$ is the variance of the trait between subjects. It is easily shown that $\sigma^2(b) + \sigma^2(w)$ = the total variance of ratings, i.e., the variance for all ratings, regardless of whether they are for the same subject or not. Hence the interpretation of the ICC as the proportion of total variance accounted for by within-subject variation.

The equation would apply if we knew the true values, $\sigma^2(w)$ and $\sigma^2(b)$. But we rarely do, and must instead estimate them from sample data. $\sigma^2(b)$ is the variance of *true trait levels* between subjects. Since we do not know a subject's true trait level, we estimate it from the subject's mean rating across the raters who rate the subject. Each mean rating is subject to sampling variation, the deviation from the subject's true trait level, or its surrogate, the mean rating that would be obtained from a very large number of raters. Since the actual mean ratings are often based on two or a few ratings, these deviations are appreciable and inflate the estimate of between-subject variance. We can estimate the amount and correct for this extra, error variation. If all subjects have k ratings, then the extra variation is estimated as (1/k) $s^2(w)$, where $s^2(w)$ is the pooled estimate of within-subject variance. When all subjects have k ratings, $s^2(w)$ equals the average variance of the k ratings of each subject (each calculated using k-1 as denominator). To get the ICC we can (1).Estimate $\sigma^2(b)$ as [$s^2(b)$ - $s^2(w)/k$], where $s^2(b)$ is the variance of subjects' mean ratings; (2). Estimate $\sigma^2(w)$ as $s^2(w)$; (3). Apply the equation.

SAS can use one-way ANOVA to calculate the between subject variation $s^2(b)$ and within subject variation $s^2(w)$, thus calculate the ICC and its confidence limits based on the above equations. A confidence interval gives an estimated range of ICC plausible values which is more informative. Recall ICC bounds between 0 and 1, it is not normally distributed and its variance is not constant. The sampling variance depend upon the ICC value. As ICC approaches 1, all sample values close to the parameter and the sampling variance approaches zero. So the CI can not be computed directly. Fisher developed a transformation which tends to normalize the distribution and stabilize the variance:

$$z = \tfrac{1}{2} \ln\left(\frac{1 + ICC}{1 - ICC}\right)$$

For the transformed z, the approximate variance V(z) is independent of the correlation. Fisher transformation makes it possible for the computation of confidence interval indirectly by 3 steps:

Step 1: Fisher transformation.

Step 2: The two-sided confidence limits are computed by the standard normal distribution.

Step 3: These computed confidence limits are then transformed back to derive the confidence intervals for ICC.

SAS macro ICC_SAS implement this process in the following:

```
%macro Icc_sas(ds=, response=, subject=);
ods output OverallANOVA =all;
proc glm data=&ds;
  class &subject;
  model &response=&subject;
run;
data Icc(keep=sb sw n R R_low R_up);
  retain sb sw n;
  set all end=last;
  if source='Model' then sb=ms;
  if source='Error' then do;sw=ms; n=df; end;
  if last then do;
     R=round((sb-sw)/(sb+sw), 0.01);
     vR1=((1-R)**2)/2;
     vR2=(((1+R)**2)/n +((1-R)*(1+3*R)+4*(R**2))/(n-1));
     VR=VR1*VR2;
     L=(0.5*log((1+R)/(1-R)))-(1.96*sqrt(VR))/((1+R)*(1-R));
     U=(0.5*log((1+R)/(1-R)))+(1.96*sqrt(VR))/((1+R)*(1-R));
     R_Low=(exp(2*L)-1)/(exp(2*L)+1);
     R_Up=(exp(2*U)-1)/(exp(2*U)+1);
     output;
  end;
run;
proc print data=icc noobs split='*';
  var r r_low r_up;
  label r='ICC*' r_low='Lower bound*' r_up='Upper bound*';
  title 'Reliability test: ICC and its confidence intervals';
run;
%mend;
```

Note: The above macro has three parameters: ds is the input dataset; response is the measurement of interest; subject is the subject id variable. The input dataset should have two observations from two raters for each subject.

DISCUSSION

In clinical trials, the validity of the study conclusion is limited by the reliability of the outcome that is measured. For this reason, it is essential for accurate quantitative, longitudinal measurement in a reliable and reproducible manner, which can provide increased reliability and consistency, thereby attaining greater confidence levels in the generated data.

To address this problem, the repeated measures of tumors from MR images by two trained raters were obtained. These data were used to assess the consistency of measurements by the methods mentioned above.

The correlation analysis has difficulty detecting the inconsistency between measurements. This is due to the fact that the correlation coefficient shows the degree of association, not the degree of consistency. The correlation analysis is very sensitive to outliers. As a result, we can infer that the correlation analysis is not sufficient as the measurement consistency analysis.

A linear regression should not be used either because it is similar to the correlation analysis.

Although the Bland-Altman plot provides the degree of bias, it is not easy to infer about the measurement consistency based on these plots. Bland-Altman method is only a visualization technique, it lacks the numeric quantification. The method does not provide a decision based on the quantified degree of consistency. In conclusion, Bland-Altman method is not appropriate as a single technique for determining measurement consistency.

The paired t-test provides quantification for the Bland-Altman method, and it has a good performance in detecting measurement bias. The paired *t*-test can detect measurement bias between two raters fairly well in most cases, but it would not be applicable when there are more than two raters. When most of the measurements of one rater are consistently larger or smaller than those of the other rater, the paired t-test tends to fail.

ANOVA extends the paired *t*-test method and the analysis shows a good performance in detecting measurement bias. ANOVA provides the good performance in all cases studied and showed accurate analysis results in determining the measurement consistency. In addition, it provides the additional information of within-rater consistency.

ICC is defined as the correlation of one measuring variable between two or more members with groups. Applied to the context of tumor imaging, ICC can be used as an indicator of scanning reliability or consistency across raters. For the continuous data, ICC provides good measure of reliability. It is based on the ANOVA method under a random effect model.

CONCLUSION

The measurement consistency problem occurs in image data analysis universally, and it is of broad interest to researchers in diverse medical imaging disciplines. RECIST guideline has significantly improved the quality of clinical trial. It has also presented a number of challenges to those responsible for the evaluation of cancer therapies. In some other studies, such as psychiatric and neurological studies of Schizophrenia, Alzheimer's disorders or Multiple Sclerosis, it is also necessary to have accurate measurements of MRI images.

We have described the strength and the weakness of each technique for several major statistical approaches that have been used to check measurement consistency. When comparing techniques, our main focus is on detecting the measurement bias and determining robustness to outliers.

It can be concluded that using only one method may be insufficient and that several methods should be applied and compared. A good rule to follow is not to limit measurement consistency assessment on only one method but rather to apply several methods and put the results together.

REFERENCES

L. Liu "Reliability analysis: Calculate and Compare Intra-class Correlation Coefficients (ICC) in SAS", Proceedings of NESUG 2007.

D. Shen, Z. Lu "Computation of Correlation Coefficient and Its Confidence Interval in SAS® ", Proceedings of SUGI31 (Paper 170-31), 2006

CONTACT INFORMATION

Gary Chen, Ph.D, Associate Director of Statistical Programming

Shire Phramaceuticals

735 Chesterbrook Boulevard

Chesterbrook, PA 19087

Work Phone: 484-595-8268

E-mail: gchen@shire.com

Web: www.shire.com