22nd International Congress on Modelling and Simulation, Hobart, Tasmania, Australia, 3 to 8 December 2017
mssanz.org.au/modsim2017

Method Comparison for Interrater Reliability of an Image Processing Technique in Epilepsy Subjects

A.A. Bartolucci[a], K.P. Singh[b] and S. Bae[c]

[a] *Department of Biostatistics, University of Alabama at Birmingham, Birmingham, Alabama, 35294, USA.*
[b] *Department of Epidemiology and Biostatistics, University of Texas Health Science Center, Tyler, 75708 TX, USA*
[c] *Division of Preventive Medicine and UAB Comprehensive Cancer Center, University of Alabama at Birmingham, Birmingham, Alabama, 35294, USA.*
Email: bsejong@uab.edu

Abstract: The value of ictal SPECT in the pre-surgical evaluation of epilepsy patients is well established. SISCOM (Subtraction Ictal SPECT Co-registered to MRI) and ISAS (Ictal-Interictal Subtraction Analysis by Statistical Parametric Mapping (SPM)) are two widely evaluated ictal-SPECT processing techniques which have been shown to enhance the interpretability of raw ictal-interictal SPECT image pairs. We sought to apply three strategies of method comparison for the inter-rater reliability of the ISAS technique to simple visual analysis in terms of study localization and overall conclusiveness of this localization. We focus on the Ictal SPECT analysis with SPM (ISAS), since it was observed at the time of this study to be more robust in regard to yield of positive studies and readability.

Our study population consisted of 34 males and 33 females with a mixture of temporal lobe and extra temporal localization. Subjects were identified through medical record review at the UAB (University of Alabama at Birmingham) Epilepsy Center. Patients undergoing inpatient video-EEG monitoring as part of their pre-surgical epilepsy evaluation were intravenously injected with 20 to 40 mCi of the cerebral blood flow tracer Tc-99m hexamethylpropyleneamine-oxime at the first clinical sign of seizure onset. Injected patients were stabilized and scans were acquired within 1 to 3 hours of injection in all cases. As a second radiotracer injection and scan was performed interictally, after a minimum 24 hour EEG confirmed seizure free period. Each scan was performed on a Picker Prism 3000XP (Picker International, Bedford, OH) triple-head gamma camera equipped with low-energy, high-resolution collimators yielding an image resolution of approximately 7 mm full width at half maximum as described in detail previously in Knowlton (2004). Consecutive patients undergoing video-EEG evaluation for their refractory partial epilepsy were included if they had ictal and interictal SPECT scan data, surgically confirmed focal epilepsy, and post-operative follow up of 1 year or more. Since this was a retrospective study, no consent was required. ISAS scans were obtained from 67 consecutive patients as part of their epilepsy pre-surgical evaluation. A panel of 3 blinded experienced reviewers from different institutions evaluated each patient's SPECT data. All scans of the ISAS processing type were evaluated as a group in separate sittings. Each scan was presented in a 16 axial slice (4x4) configuration, and was evaluated based on 1) location of the SPECT abnormality, and 2) the overall localizing value of the study. Localization was identified as the area of most significant activity and its position within one of 30 pre specified brain regions of interest. The localizing value of each study was judged as degree of localization on a continuous Likert scale from 1 to 4 localizing (1- definitely localizing to 4- not localizing) which the authors have simulated from the original data. Methods of agreement were calculated among the 3 reviewers. The methods used were linear with density ellipse visualization, the Bland Altman procedure, and a Bayesian version of the linear methodology. A statistical comparison of the methods was done for reviewers 1 vs. 3 and reviewers 2 vs. 3.

***Keywords*:** *Bayesian, Bland Altman, density ellipse, method comparison*

1. INTRODUCTION

Kim and Mountz (2011) give an overview of single-photon emission computed tomography (SPECT imaging) in epilepsy. The introduction of noninvasive neuroimaging methods, such as tSPECT, positron emission tomography (PET), and magnetic resonance imaging (MRI), has played a large role in pre-surgical epilepsy evaluation. These imaging methods have become powerful tools for the investigation of brain function and an essential part of the evaluation of epileptic patients. Kim and Mountz (2011) note that of the methods utilized, only SPECT has the practical capacity to image blood flow functional changes that occur during seizures in the routine clinical setting. They further note that although functional MRI (fMRI) could, in theory, be used for this purpose, it is impractical due to patient movement during most types of seizures, a problem that is overcome by the timing and technique of SPECT imaging. SISCOM (Subtraction Ictal SPECT Co-registered to MRI) and ISAS (Ictal-Interictal Subtraction Analysis by Statistical Parametric Mapping (SPM)) are two widely evaluated ictal-SPECT processing techniques which have been shown to enhance the interpretability of raw ictal-interictal SPECT image pairs. We sought to apply three strategies of method comparison for the inter-rater reliability of the ISAS technique to simple visual analysis in terms of study localization and overall conclusiveness of this localization. We summarize our findings here.

2. METHODS

ISAS scans were obtained from 67 consecutive patients (34 males, 33 females) as part of their epilepsy pre-surgical evaluation. A panel of 3 blinded experienced reviewers from different institutions evaluated each patient's SPECT data. All scans of the ISAS processing type were evaluated as a group in separate sittings. Each scan was presented in a 16 axial slice (4x4) configuration, and was evaluated based on 1) location of the SPECT abnormality, and 2) the overall localizing value of the study. Localization was identified as the area of most significant activity and its position within one of 30 pre specified brain regions of interest. The localizing value of each study was judged as degree of localization on a continuous Likert scale from 1 to 4 localizing (1-definitely localizing to 4- not localizing). Methods of agreement were calculated among the 3 reviewers. The data here was simulated from the original data. The methods used were linear with density ellipse visualization, the Bland Altman procedure, and a Bayesian version of the linear methodology.

For the purposes of our straightforward Bayesian investigation, the task is to apply a method comparison sampling model to the data. From the model, we derive the response (rater x vs. rater y) and compute the posterior values for each of the model parameters. The authors have utilized the MCMC (Markov Chain Monte Carlo) procedure for deriving the posterior parameters of the model.

The authors have determined that a Bayesian consideration of the results would afford a coherent interpretation of the effect of the model. Thus, using a Markov Chain Monte Carlo method of parameter estimation with non-informative priors, one is able to obtain the posterior estimates. The conditions are all tested using a Bayesian statistical approach allowing for the robust testing of the model parameters under various recursive partitioning conditions of the covariates and hyper parameters which we introduce into the model. The convergence of the parameters to stable values are seen in trace plots (not shown) which follow the convergence patterns. This allows for precise estimation for determining conditions under which the response pattern will change. We give a numerical example of our results. The major platform for the theoretical development follows the Bayesian methodology for model testing with random effects for non-informative hyper parameters. We have done the basic infrastructure for the analysis using the commercially available WinBugs software employing the MCMC methodology. The BUGS language allows a concise expression of the parametric model to denote stochastic (probabilistic) relationships and to denote deterministic (logical) relationships. We also are aware of some issues with this procedure determining the Monte Carlo convergence and robustness which we will discuss.

For our purposes we use the simple model,

$$y[i]) = \beta_0 + \beta_1 x[i] \quad , i=1,\ldots,67. \quad \text{(x and, y are raters)} \quad (1)$$

with the prior input,

$$\begin{array}{lll} \beta_0 & \sim \text{normal } (0.0,1.0E\text{-}6) & \text{Prior for intercept} \\ \beta_1 & \sim \text{normal } (0.0,1.0E\text{-}6) & \text{Prior for slope} \end{array} \quad (2)$$

We investigated varying coherent prior input for tests of robustness as well.

3. RESULTS

Proceeding with this model, Table 1 gives the posterior results of the model noted in equation (1). Examining Table 1, one is able to interpret the models. The posterior estimate of β_1 is closer to one for the Reviewer 1 vs. the Reviewer 3 comparison (R1, R3) than it is for the (R2, R3) comparison indicating a slope closer to one for (R1, R3) than it is for the (R2, R3) linear model. Also, the root mean squared error is half for the (R1, R3) comparison compared to the (R2, R3) comparison. Additionally, notice the wider bootstrap posterior credible interval for the slope of the (R2, R3) model. Figure 1 is the regression result for R3 vs. R1 We've modified the program to include the 95% density ellipse. Note the very tight ellipse indicating fairly strong agreement due to the intercept being close to zero and the slope close to one. These are not definitive proof of agreement but, as we'll see later with Bland Altman, they are fairly strong. Figure 2 shows the (R2, R3) regression. Note the wide density ellipse indicating weak agreement or association between reviewers 2 and 3. When considering the ellipses together, one clearly sees the more robust shape of the (R1, R3) presentation. Doing a bootstrap methodology was another way to consider these regressions. Such was done and the results of the Bayesian bootstrapping are seen in Table 1. The Bland Altman strategy is considered more robust for our purpose here.

Table 1. Posterior Parameter Estimates

Variable (parameter)	Mean	RMSE	Median	95% Credible Interval
(R1,R3): β_0	0.022	-	0.033	(0.008, 0.060)
(R1,R3): β_1	0.992	0.057	0.983	(0.97, 1.01)
(R2,R3): β_0	0.308	-	0.380	(0.06, 0.56)
(R2,R3): β_1	0.884	0.191	0.830	(0.71, 1.05)

However, keep in mind that the density ellipse is a good start for visualizing the actual graphical relationship and strength of association of two methods or two variables to each other. Correlation and regression studies are often proposed. However, correlation studies the relationship between one variable and another, not the differences, and it is not recommended as a method for assessing the comparability between methods.

Altman and Bland (B&A. 1983) proposed an alternative analysis, based on the quantification of the agreement between two quantitative measurements by studying the mean difference and constructing limits of agreement. The B&A plot analysis is a simple way to evaluate a bias between the mean differences, and to estimate an agreement interval. A good discussion and understanding of this procedure is presented in Giavarina (2015) and Bartolucci et al. (2015).

We proceed here with the Bland-Altman structure. The plot as seen in Figure 3 is the plot for the (R1, R3) method comparison. One can see that the mean solid line is the mean of the paired differences, R3-R1 (Rev_3-Rev_1) on the vertical. The horizontal axis is the average of the two reviewer ratings. The dotted lines outside the mean lines are the 95% confidence interval on the mean difference. Note how close the mean line is to the value 0 in this case. As a matter of fact, the actual mean paired difference is 0.0015 with 95% confidence interval (-0.013, 0.016), which clearly includes 0. If one were to do a pairwise test of mean differences = 0, then one would not reject the null of no difference, p=0.836. Another caveat of the Bland Altman graph is that all the paired differences fall within two standard deviations of the mean difference. Two standard deviations of the mean difference is ± 0.116. Clearly in Figure 3 all the points are within ± 0.10 of the mean. Another consideration is that one can put an upper 95% confidence interval on the upper 2 standard deviation (SD) line and a lower confidence interval on the lower 2 SD line and determine if the points fall within those limits. Clearly that is not needed in this case since all the points are within the 2 SD limit. Let us now consider the Bland Altman treatment of the (R2, R3) comparison. This is seen in Figure 4.

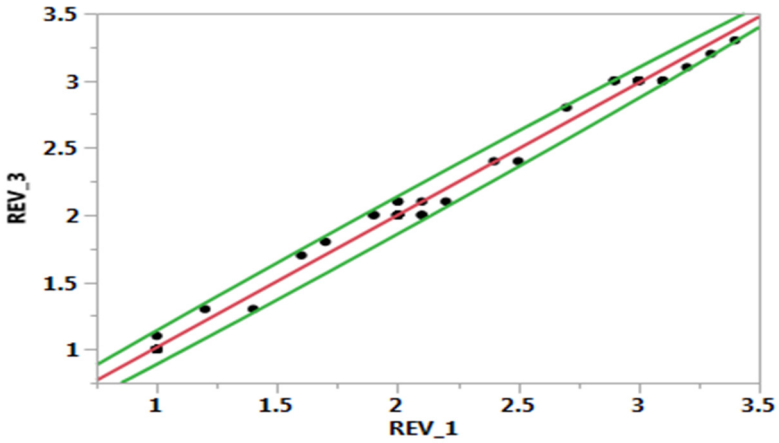

Figure 1. Regression R3 vs. R1

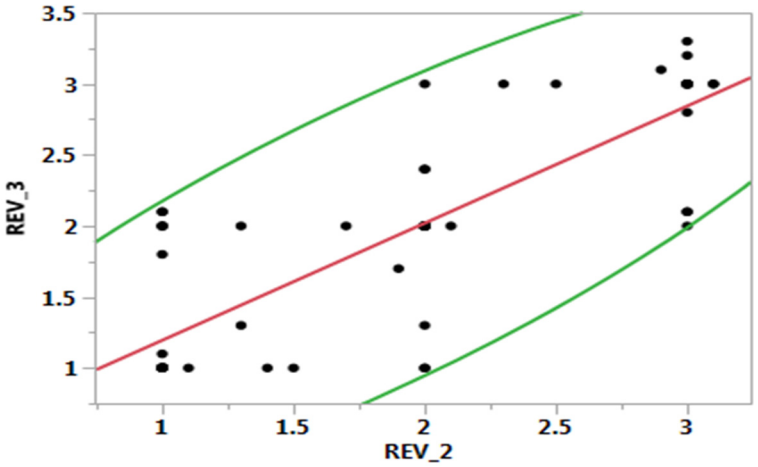

Figure 2. Regression R3 vs. R2

Figure 4 shows much greater dispersion of the difference points about the mean. The mean of the paired differences is 0.052 with 95% confidence bounds (-0.059, 0.165). This interval does contain 0 as the previous plot. However, notice the differences in CI width. For the (R1, R3) comparison, the CI width is 0.029 compared to 0.224 for the (R2, R3) pair. Clearly the (R2, R3) comparison width is about 10 times that of the (R1, R3) showing the greater precision of the (R1, R3). One will note that we did not discuss the (R1, R2) situation throughout. This was simply because those results were much the same as the (R2, R3) results. That is to say the estimate of the intercept was around 0.339 and the slope was estimate was 0.841. Clearly there was considerable agreement for R1 and R3, but not R1 to R2 or R2 to R3.

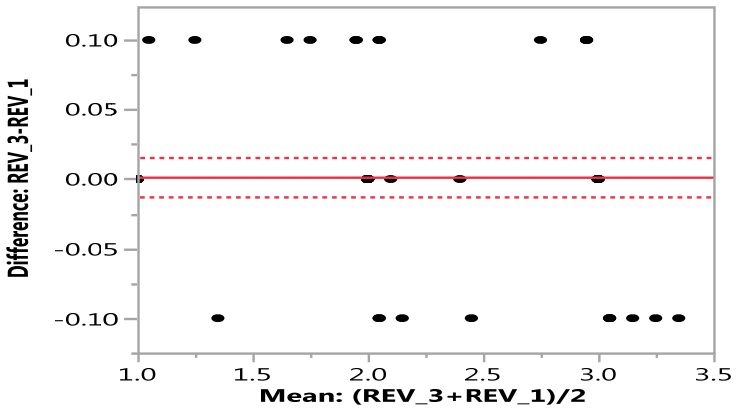

Figure 3. Bland Altman graph for R3 vs. R1

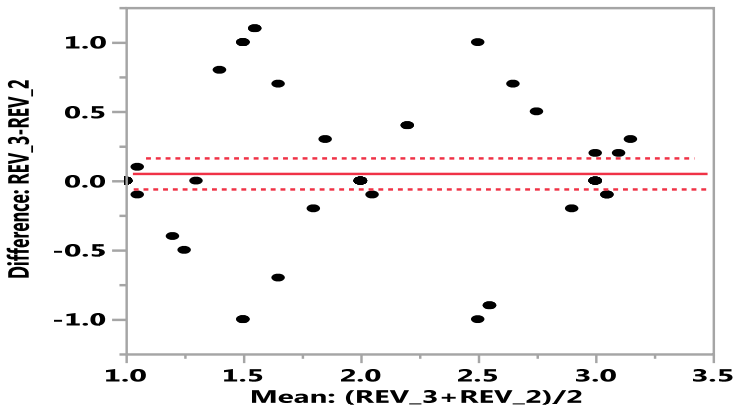

Figure 4. Bland Altman graph for R3 vs. R2

We mentioned above that we checked for the robustness of the models under certain conditions. This was done merely by varying the prior input of the means in equation (1) which varied from 0 to 1 leaving the prior variance hyper parameters fairly well dispersed. All results were clinically and biologically meaningful.

4. DISCUSSION AND CONCLUSIONS

For the purposes of our straightforward investigation the task is to apply a simple linear sampling model plus the Bland Altman procedure as well to the data. The Bayesian approach was primarily to consider the random effects. From the model, we computed the posterior intercept and slope and added bootstrap sample to adjust the credible intervals from dispersed prior input. The authors have utilized the MCMC (Markov Chain Monte Carlo) procedure for deriving the posterior parameters of the model which include the posterior means, medians

and 95% intervals. Our results suggest that there is sufficient clinical validity based on the localization strategy employed by the reviewers. The authors have demonstrated that a Bayesian consideration of the results would afford a reasonably coherent interpretation of the results assuming random effects in the model. Thus, using a Markov Chain Monte Carlo method of parameter estimation with non- informative priors, one is able to obtain the posterior estimates and credible regions of estimates of these effects. The models used were robust and convergence for the most part was tractable for the parameter estimates.

REFERENCES

Altman, D.G and, Bland, J.M., (1983). Measurement in medicine: the analysis of method comparison studies. *Statistician.* 32, 307-317.

Bartolucci, A.A., Singh, K.P., and Bae, S. (2015) *Introduction to Statistical Analysis of Laboratory Data*. John Wiley. New York.

Giavarina, D. (2015). Understanding Bland Altman analysis. *Biochemia Medica*. 25(2). 141-151.

Kim, S. and Mount, J.M. (2011), SPECT Imaging of Epilepsy: an Overview and comparison with F-18 PET. *International Journal of Molecular Imaging*. 1-9.

Knowlton, R.C., Lawn, N.D., Mountz, J.M., and Kuzniecky, R.I. (2004). Ictal SPECT analysis in Epilepsy: Subtraction and statistical parametric mapping techniques. Neurology. 63. 10-15.