Detection of Outliers in Bioequivalence Studies Data Analysis with Williams Design

A. Rasheed[1]*, S. Junaid[2] and T. Ahmad[3]

[1]Department of Research, Dow University of Health Sciences, Karachi, [2]Department of Statistics, University of Karachi, Karachi [3]Center for Bioequivalence Studies at ICCBS, University of Karachi, Karachi, Pakistan.

Abstract: Background: Drug Regulatory agencies all over the world generally discourage exclusion of outliers in a BE (BE) study; on the other hand in routine bio-statistical work we take these into the account. If the decision rules for identifying the outliers are clearly mentioned before the start of the study and laid down in protocol by the responsible biostatistician in collaboration with clinicians, the problem of outliers can be dealt smartly without jeopardizing the whole study for redoing. The purpose of this article is to introduce procedure for reliably detecting outlier subject(s) with Williams design.

Experimental: Literature review reveals many different methods for the detection of outlier values in BE studies; most of them are for BE of two treatments. For BE studies with more than two treatments use of Williams design seems imperative; but inclusion and deletion of outlying subjects may lead to profound effect on conclusion of BE which in turn may be dangerous for the health. The suggested method is an adjustment to a previously introduced method using exploratory data analysis technique such as principle component analysis and Andrews curves.

Keywords: Bioequivalence, outliers, Williams design, exploratory data analysis, principal, component, Andrews curves.

1. INTRODUCTION

1.1. Back Ground and Rationale

Bioequivalence is established when two or more Pharmaceuticals of same dosage forms (tablet, capsule, suspension etc) and of same strengths of active ingredient demonstrate statistically sound similarity in rate and extent of absorption when administered to patients or subjects. Generic products are generally required to show bioequivalence before they are permitted for distribution or sale.

For a new drug the inventor is required to submit Animal Studies, Clinical Studies and Bioavailability data to regulatory authorities where as for a generic product one has to submit only Bioequivalence data in comparison to the patent product. Bioequivalence is defined by the US FDA as follows:

"The absence of a significant difference in the rate and extent to which the active ingredient or active moiety in pharmaceutical equivalents or pharmaceutical alternatives becomes available at the site of drug action when administered at the same molar dose under similar conditions in an appropriately designed study."

*Address correspondence to this author at the Department of Research, Dow University of Health Sciences, Karachi, Karachi, Pakistan; E-mail: : abdur.rasheed@duhs.edu.pk

FDA recommends logarithmic transformation of BE measures e.g., area under the blood or plasma concentration–time curve (AUC) and maximum concentration C_{max} and requires justification if sponsors/ applicants consider that there BE study data should be statistically analyzed on the linear scale rather than on the log scale. The logarithmic transformation of BE measures, makes the distribution more symmetric and closer to the normal distribution .Therefore in this paper, we considered BE data set on both linear and logarithmic scales, in order to explain our proposed EDA techniques.

Tukey [1] proposed exploratory data analysis (EDA) a methodology for data analysis which generally employs a variety of techniques most of them are graphical. These methods are useful in detecting outliers, uncover underlying structure, maximize insight into a data set, extract important variables and determine optimal factor settings.

Enachescu and Enachescu [2] discussed the EDA technique such as Andrews curves and Principal Component Analysis (PCA). Using PCA Enachescu and Enachescu [2] mentioned that "first two principal axes in PCA span one such plane, providing a projection such that the variation in the projected data is maximized over all possible 2-D projections and applied these techniques to identify the outlying subjects in 2x2 crossover BE trial. In present work we extended these techniques to Williams design; a

special variety of Crossover/Latin square designs. With the help of our proposed approach, gathering information regarding outliers subjects and their identification in a BE data with more than two formulations becomes easily possible.

1.2. Williams Design

In crossover designs if each formulation appears in same number of times within each sequence is called 'uniform within sequence' and if each formulation appears the same number of times within each period than is called 'uniform within period'. A crossover design is called uniform if it is uniform within sequences and within periods. A Latin square, in which every treatment is occurred once and only once in each row and each column yields uniform crossover designs. In a balanced design, each of the treatments occur the same number of times in each period and the number of subjects who receive treatment i in one period and treatment j in the next period is the same for all $i \neq j$ [3].

Williams [4] introduced a crossover design in which every treatment follows every other treatment the same number of times called Williams design. These Williams designs require fewer subjects than those based on the complete sets of orthogonal Latin squares [5]. In Williams design when the number of formulations are even than balance can be achieved by a single Latin square design, but when the number of formulations are odd than two Latin square designs are needed.

2. EXPERIMENTAL TECHNIQUES

In this present work we used two EDA techniques, Andrews curve and PCA to ease the problem of detecting outliers in BE studies with more than two treatments. Modified z-scores method commonly used method for outlier detection also used here to insert the fictitious outliers in original data set. In Modified z-scores method subject having absolute z-scores greater than 3.5 are labeled as outliers.

EXPLORATORY DATA ANALYSIS TECHNIQUES

2.1. Andrews Curves

Andrews [6] introduced a method of visualization for multivariate data. Each multidimensional data point is mapped into a periodic function

$$f_X(t) = X_1 / \sqrt{2} + X_2 \sin t + X_3 \cos t + ... \quad (1)$$

This graphical approach displays a point in multidimensional space by a two-dimensional curve using the function $f_x(t)$ given above in the interval $-\prod < \tau < \prod$. The advantage of this method is that it allows the inclusion of many dimensions. A collection of multidimensional points, that is, a multivariate data set, is displayed as a group of curves. In this method numbers of variables are unlimited. These curves are dependent on the order of the variables. Lower frequency terms (i.e., those that are first in the sum given in the above equation) exert more influence on the shape of the curves we can get more information about data by re-ordering the variables and viewing the resulting plot. Observations showing quite apparent different curves are considered as outliers.

2.2. Principal Component Analysis

The objective of PCA is to discover or to reduce the dimensionality of the data set and identify new meaningful underlying variables.

In PCA number of (possibly) correlated variables are transformed into (smaller) number of variables which are uncorrelated called principal components. Large amount of variability is accounted by the first PC and each succeeding PC accounts for as much of the remaining variability as possible.

PCA can be performed either by using a covariance matrix or correlation matrix both matrices are calculated from the data matrix, if one is using correlation matrix so firs variables should be standardized.

2.2.1 Eigen Analysis

Eigen analysis is a mathematical technique used in the PCA, in this technique Eigen values and Eigen vectors of a square symmetric matrix with sums of squares and cross products are calculated. The eigen-vector associated with the largest Eigen-value has the same direction as the first principal component. The Eigen-vector associated with the second largest Eigen-value determines the direction of the second principal component.

X is $p \times n$ the data matrix where (p= number of variables and n= number of observation), $\sum$ is covariance matrix obtained from the data matrix X, and Z is the standardized data matrix, R is correlation matrix obtained from the data matrix Z. λ_i is called Eigen value denotes the variance of the i-th PC (i.e., $\lambda_i = Var(ith\,PC)$ that can be calculated by setting

$|R-\lambda I|=0$ Where I is the identity matrix. $U_i=\beta^{(i)'}Z$ is called the i-th PC where $\beta^{(i)}$ is denote the i-th eigen vector that can be calculated by setting $(R-\lambda_i I)\beta^{(i)}=0$ where $\beta=\left[\beta^{(1)}\ \beta^{(2)}\ \beta^{(3)}\ ...\ \beta^{(p)}\right]$ and each of $\beta^{(i)}$ is defined as $\beta^{(i)'}=\left[\beta_1^{(i)}\ \beta_2^{(i)}\ \beta_3^{(i)}\ ...\ \beta_p^{(i)}\right]$ with $\hat{a}^{(e)'}\hat{a}^{(e)}=\hat{a}_1^{(e)^2}+\hat{a}_2^{(e)^2}+\hat{a}_3^{(e)^2}+...+\hat{a}_p^{(e)^2}=1$ The sum of the Eigen-values equals the trace of the square matrix (i.e., $tr\ (\Sigma)=\sum_{i=1}^{p}\lambda_i$) and the maximum number of Eigen-vectors equals the number of rows (or columns) of this matrix.

Enachescu and Enachescu [2] has mentioned that *"For normally distributed observation* U_i/λ_i *are independent* $\chi^2_{1,j}$ *variables. Consider* $\sum_{i=1}^{p}\lambda_i\chi^2_{1,j}$ *the weighted sum of square distance to zero of the projected data into principal factorial plane, with* $E\left(\sum_{i=1}^{p}\lambda_i\chi^2_{1,j}\right)=\sum_{i=1}^{p}\lambda_i=p$ *and* $Var\left(\sum_{i=1}^{p}\lambda_i\chi^2_{1,j}\right)=2\sum_{i=1}^{p}\lambda_i^2$ *. Now the Observations with a square distance greater than* m *(the rule of* 2σ *) may be considered as outliers".*

3. PROPOSED NEW TECHNIQUE

EDA techniques, Andrews curve and PCA have been used to detect outliers in BE studies with two treatments, here we advocated Andrews curve and PCA for gathering information regarding outliers in Williams design with a reference (R) and two treatments formulations (T1 and T2).

Andrews curve function is defined as

$$f(t)=R/\sqrt{2}+T_1\sin t+T_2\cos t\ \ -\pi<t<\pi\ \ i=1,2,3...12 \quad (2)$$

Each observation is projected onto a set of orthogonal basis functions represented by sines and cosines and then plotted. Thus, each sample point is now represented by a curve. Observations with identical curves show the possible outlying subjects.

In PCA, X is p x n the data matrix where (p= 3, number of formulations and n=12, number of observation for each formulation) , $\sum$ is covariance matrix obtained from the data matrix X , and Z is the standardized data matrix, *R* is correlation matrix obtained from the data matrix Z and $\lambda_1,\lambda_2\ and\ \lambda_3$ are Eigen values denoting the variances of first, second and third PC respectively. $\sum_{i=1}^{3}\lambda_i\chi^2_{1,j}$ the weighted sum of square distance to zero of the projected data into principal factorial plane, with mean $p=3$ and variance $2\sum_{i=1}^{3}\lambda_i^2$. Now the observations with a square distance greater than m (the rule of 2σ) may be considered as outliers where $m=3+2\sqrt{2\sum_{i=1}^{3}\lambda_i^2}$.

4. APPLICATION AND VALIDATION OF PROPOSED TECHNIQUE

In the present work we selected a data set of Areas Under the Curve; AUC from a bioequivalence study reported by Purich [7]. In the study twelve healthy volunteer were employed to investigate the bioequivalence of two test tablets formulations in comparison to a reference solution. Chow and Liu [8] mentioned that no assignment of sequences and periods was given by Purich [7]. Thus for the purpose of illustration Chow and Liu [8] assigned subject 1 and 2 to sequence 1; 3 and 4 to sequence 2; 5 and 6 to sequence 3; 7 and 8 to sequence 4; 9 and 10 to sequence 5; 11 and 12 to sequence 6. Table **1** gives this AUC data set after rearrangement of reference and period according to Williams design for comparing the three formulations. The numerical results given bellow are obtained with popular software SAS v 9.0.

Table 1: AUC Data Set with a Reference and Two Test Formulation (Domestic and European Tablets)

Sequence	Subject	Period I	Period II	Period III
(R, T2, T1)	1	5.68	4.21	6.83
	2	3.6	5.01	5.78
(T1,R, T2)	3	3.55	5.07	4.49
	4	7.31	7.42	7.86
(T2,T1,R)	5	6.59	7.72	7.26
	6	9.68	8.91	9.04
(T1,T2,R)	7	4.63	7.23	5.06
	8	8.75	7.59	4.82
(T2,R,T1)	9	7.25	7.88	9.02
	10	5	7.84	7.79
(R,T1, T2)	11	4.63	6.77	5.72
	12	3.87	7.62	6.74

To apply the above defined EDA techniques for determining the possible outlying subjects here AUC

data set is presented formulations wise on the linear and logarithmic scale in Table **2**.

Table 2: AUC(0-inf) Data Set for a Reference and Two Test Formulations on Both Linear and Logarithmic Scales

Subject	Linear Scale			Logarithmic Scale		
	R	T1	T2	R	T1	T2
1	5.68	6.83	4.21	1.737	1.921	1.437
2	3.6	5.78	5.01	1.281	1.754	1.611
3	5.07	3.55	4.49	1.623	1.267	1.502
4	7.42	7.31	7.86	2.004	1.989	2.062
5	7.26	7.72	6.59	1.982	2.044	1.886
6	9.04	8.91	9.68	2.202	2.187	2.27
7	5.06	4.63	7.23	1.621	1.533	1.978
8	4.82	8.75	7.59	1.573	2.169	2.027
9	7.88	9.02	7.25	2.064	2.199	1.981
10	7.84	7.79	5	2.059	2.053	1.609
11	4.63	6.77	5.72	1.533	1.913	1.744
12	3.87	7.62	6.74	1.353	2.031	1.908

4.1. ANDREWS CURVES

The Andrews curves for this data set are

For linear scale

$$f(t)=R/\sqrt{2}+T_1\sin t+T_2\cos t \quad -\pi<t<\pi \quad i=1,2,3\ldots 12 \quad (3)$$

For logarithmic scale

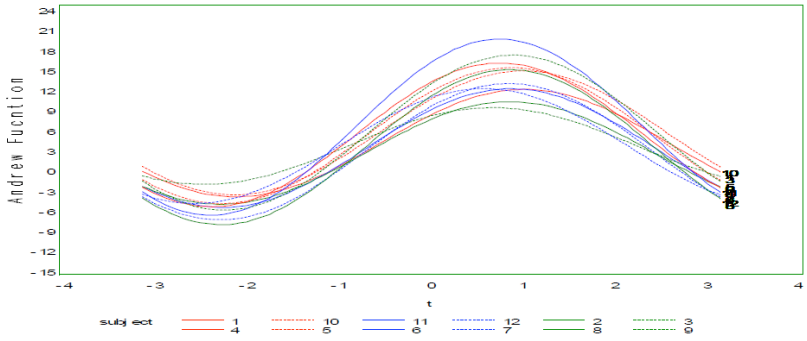
$$f(t)=\ln(R)/\sqrt{2}+\ln(T_1)\sin t+\ln(T_2)\cos t \quad -\pi<t<\pi \quad i=1,2,3\ldots 12 \quad (4)$$

From linear and logarithmic Andrew Curves given in Figures-**1a** and **1b** it is very evident that there is no curve reveals distinct behavior comparing with other curves, suggesting any subject a possible outlier.

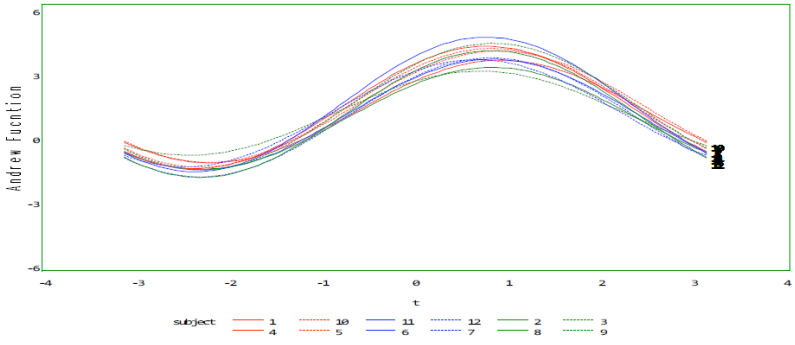
4.2. Principal Component Analysis

We employed the same data set given in the Table **1**, by using the principal components analysis we obtained the results with correlation matrix R in Table **3**.

The threshold value and the square distance to the zero for each observation are given in table 4 for the linear and logarithmic scale



1a



1b

Figure 1: Andrews curves (**a**) linear scale data; (**b**) logarithmic scale data

Table 3: Eigen-Values of the Correlation Matrix R for the Linear and Logarithmic Scales

	Linear Scale				Logarithmic Scale			
	Eigen values	Differences	Proportion	Cumulative	Eigen values	Differences	Proportion	Cumulative
1	2.0421	1.5176	0.6807	0.6807	1.9098	1.2942	0.6366	0.6366
2	0.5245	0.091	0.1748	0.8555	0.6156	0.141	0.2052	0.8418
3	0.4335		0.1445	1	0.4746		0.1582	1

Table 4: Threshold Value and the Squared Distance on Linear and Logarithmic Scales

Subject	Linear Scale		Logarithmic Scale	
	Squared distance	Threshold	Squared distance	Threshold
1	1.9756731	9.088	2.4626745	8.832
2	3.1857674		3.5669672	
3	6.0492623		7.3577339	
4	1.4066709		1.5539635	
5	0.6469274		0.8066935	
6	8.0766893		6.05619	
7	2.5784951		2.4345028	
8	1.9496287		1.7079329	
9	2.6834404		2.3762979	
10	2.0350777		2.0368382	
11	0.8310584		0.6595056	
12	1.5813093		1.9807000	

Any observation with square distance greater than corresponding threshold value may be considered as outlier. As we can see that on both scales there is no observation with square distance greater threshold.

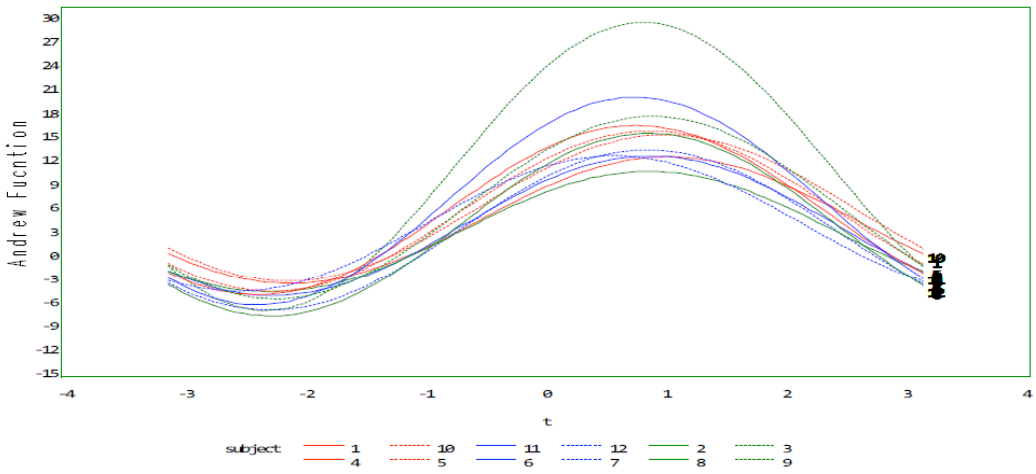
As evident from the above analysis no observation is found as an outlier in the above data set. In order to verify the proposed extended EDA techniques (Andrews curves and PCA) it was felt imperative to introduce intentionally some outlying values in the original data. Accordingly we made some changes in the original data set by replacing few values with fictitious (obvious outlier) values.

We replaced some original values with few fictitious extreme (very high and very low) values for each treatment (i.e., R, T1, and T2) which were identified as outliers by certainity by confirming them as outliers through modified z-scores method. We carried this exercise two times.

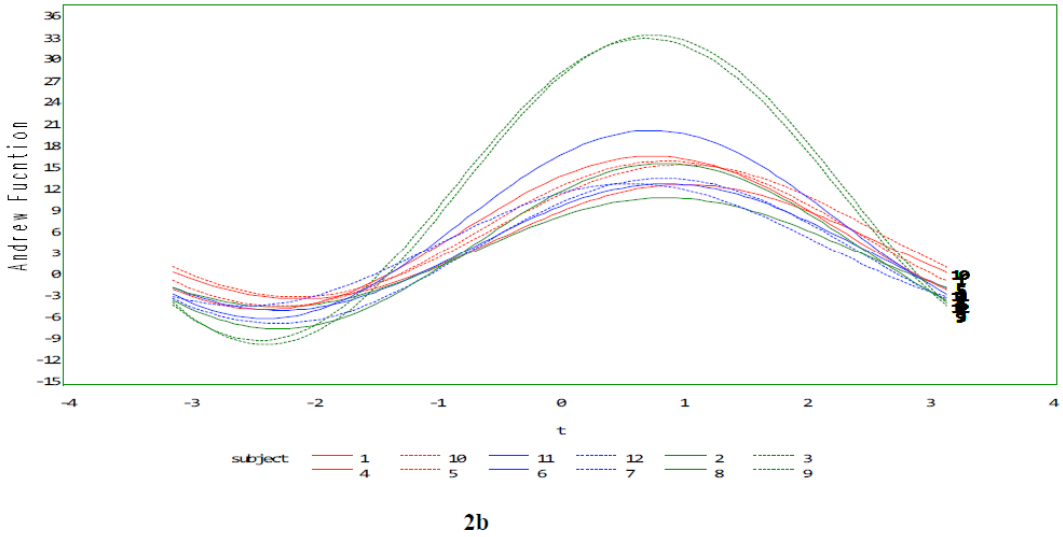
In first instance we randomly selected a subject 3 from original data set and replaced its values for all three treatments (5.07, 3.55 and 4.49) by fictitious values (15.2, 13.2 and 12.56) previously identified as outliers.

Table 5: Threshold Value and the Squared Distance for Fictitious Data Sets 1 and 2 from Method of Principal Component Analysis

Subject	Data Set # 1		Data Set # 2	
	Squared distance	Threshold	Squared distance	Threshold
1	2.0180746	10.431	1.4646404	11.072
2	2.8926605		2.277750	
3	19.067802		11.10280	
4	0.2055743		0.1342879	
5	0.0750083		0.2062721	
6	1.9922433		0.3078588	
7	2.6616014		1.9995872	
8	0.6348092		0.4616749	
9	0.4082725		12.259875	
10	0.9686317		0.6733027	
11	1.1422837		1.1532807	
12	0.9330383		0.9586664	



2a



2b

Figure 2: (**a**) Andrews curves for the linear scale fictitious data set 1; (**b**) Andrews curves for the linear scale fictitious data set 2

In second instance we selected two random subjects 3 and 9 from original data set and replaced their values by fictitious values previously identified as outliers. Subject 3 values (5.07, 3.55 and 4.49) replaced by (15.95, 13.56 and 16.12) and subject 9 values (7.88, 9.02 and 7.25) replaced by (15.98, 14.80 and 15.7).

On these two change data sets we applied the proposed EDA techniques to confirm the validity of these techniques that whether these techniques identify the outliers in these two data sets.

We are glad to report that both proposed EDA techniques Andrews curves and PCA correctly identified outliers in both fictitious data sets on linear scale and logarithmic scale. In Figure **2** (**2a** and **2b**) subject 3 for fictitious data set 1 and subject 3 and 9 for fictitious data set 2 showing clearly different behavior and confirming themselves to be outliers and similarly in Table **5** the Threshold values and the squared distance for fictitious data sets 1 and 2 can be seen. In Table **5** Subject 3 in fictitious data set 1 and subjects 3 and 9 in fictitious data set 2 reveal squared distances are greater than the threshold values.

CONCLUSION

Through this work we report and recommend an extended exploratory data analysis techniques for identification of outliers in a Williams design data set generated during bioequivalence evaluation. In present

research for identification of outliers we successfully applied the EDA techniques, Andrews curves and principal component analysis for the bioequivalence data set with more than two treatments.

REFERENCES

[1] Tukey J-W. Exploratory data analysis. Addison-Wesley, Reading, MA. 1977.

[2] Enachescu D, Enachescu C. A new approach for outlying records in bioequivalence trials. Proceedings of the 13th International Conference on Applied Stochastic Models and Data Analysis, Vilnius, Lithuania 2009; 250-257.

[3] Jones B, Kenward M. Design and Analysis of Cross-Over Trials. (2nd edn). Chapman & Hall: London, 2003.

[4] Williams EJ. Experimental designs balanced for the estimation of residual effects of treatments. Australian Journal of Scientific Research 1949; 2: 149-168.

[5] Wang B-S, Wang X-J, Gong L-K. The Construction of a Williams Design and Randomization in Cross-Over Clinical Trials Using SAS. Journal of statistical software 2009; 29.

[6] Andrews D. Plots of high-dimensional data. Biometrics 1972; 28: 125–136.

[7] Purich E. Bioavailability/Bioequivalence Regulations: An FDA Perspective. In K.S. Albert (ed.), Drug Absorption and Disposition: Statistical considerations, American Pharmaceutical Association, Academy of Pharmaceutical Sciences, Washington, DC, 1980, pp. 115-137.

[8] Chow S-C, Liu J-P. Design and analysis of bioavailability and bioequivalence studies. (2nd edn). Dekker: New York, 2000.

https://doi.org/10.6000/1927-5951.2011.01.01.12