RESIDUALS AND OUTLIERS IN REPLICATE DESIGN CROSSOVER STUDIES

Robert Schall[1,2]**, Arne Ring**[3,4]**, and Laszlo Endrenyi**[5]

[1]Department of Mathematical Statistics and Actuarial Science, University of the Free State, 9300 Bloemfontein, South Africa

[2]Quintiles Biostatistics, PO Box 12603, Bloemfontein 9324, South Africa

[3]Clinical Biostatistics, Boehringer Ingelheim Pharma, 88400 Biberach an der Riss, Germany

[4]Institute for Biometry, University of Ulm, 89075 Ulm, Germany

[5]Department of Pharmacology, University of Toronto, Toronto, Ontario M5S 1A8, Canada

Key words: Bioequivalence; outlier; replicate design crossover; residual

ABSTRACT

Outliers in bioequivalence trials may arise through various mechanisms, requiring different interpretation and handling of such data points. For example, regulatory authorities might permit exclusion from analysis of outliers caused by product or process failure, while exclusion of outliers caused by subject-by-treatment interaction generally is not acceptable.

In standard 2×2 crossover studies it is not possible to distinguish between relevant types of outliers based on statistical criteria alone. However, in replicate design (2-treatment, 4-period) cross-over studies three types of outliers can be distinguished: (i) subject outliers are usually unproblematic, at least regarding the analysis of bioequivalence, and may require no further action; (ii) subject-by-formulation outliers may affect the outcome of the bioequivalence test but generally cannot simply be removed from analysis; (iii) removal of single data point outliers from analysis may be justified in certain cases.

As a very simple but effective diagnostic tool for the identification and classification of outliers in replicate design crossover studies we propose to calculate and plot three types of residual corresponding to the three different types of outliers that can be distinguished. The residuals are obtained from four mutually orthogonal linear contrasts of the four data points associated with each subject. If preferred, outlier tests can be applied to the resulting sets of residuals after suitable standardization.

1. INTRODUCTION

1.1 Outliers in Bioequivalence Studies

"Unusual values" of pharmacokinetic endpoints, either extremely large or extremely small observations, are frequently an issue in the analysis and review of bioequivalence studies. Such extreme values, or outliers, may arise through various mechanisms, including the following:

1. Product failure (coated tablet broken; single tablet with wrong drug dosage)
2. Adverse event affecting drug absorption (e.g. vomiting, diarrhoea)
3. Laboratory error / data transcription error
4. Unusual reaction of a single subject (or of a subset of subjects) to one of the formulations (so-called "subject-by-formulation interaction")

Because of the different regulatory attitude to outliers caused by different mechanisms, as noted in the next section, it will be useful to distinguish between mechanisms 1 to 3, to which we refer as outliers caused by "product or process failure", and mechanism 4, to which we refer as outliers caused by "subject-by-formulation interaction".

The potential consequences of outliers can be two-fold: outliers can bias the point estimate for the relative bioavailability of the test and reference formulation (either away from unity, which may cause failure to show bioequivalence; or towards unity, which may lead to an otherwise unjustified conclusion of bioequivalence); furthermore, outliers can inflate the standard error of the point estimate of relative bioavailability, which can lead to failure to show bioequivalence. Thus outliers may cause opposite results of the test for bioequivalence, depending on whether or not they are included in the analysis (Chow and Tse, 1990; Chow and Liu, 2000).

1.2 Regulatory Position on Outliers in 2×2 Crossover Trials

Like many other aspects of bioequivalence assessment, the statistical analysis of bioequivalence studies, including aspects of outlier handling, is highly regulated (see

EMEA, 2001; FDA, 2001 and FDA, 2003). According to regulatory guidance statements [see, for example, FDA (2001), EMEA (2006)] only documented "product or process failures" are acceptable reasons for removal of outliers from the statistical analysis of bioequivalence studies. In contrast, an unusual reaction of a subject to the drug product ("subject-by-formulation interaction") is not an acceptable reason for the removal of an outlier. "This subject could be a representative of subjects present in the general population in low numbers for whom the relative bioavailability is markedly different than for the majority of the population, and for whom the two products are not bioequivalent, even though they might be bioequivalent in the majority of the population" (FDA, 2001).

In standard 2-treatment, 2-period (2×2) crossover trials it is generally not possible, based on statistical criteria alone, to distinguish outliers caused by product/process failure from outlier caused by subject-by-formulation interaction. For this reason, removal of any outliers from the analysis of 2×2 crossover trials is "generally discouraged" by regulatory authorities (FDA, 2001) and outliers can only be verified through a re-dosing study (Braddy et al, 2008). However, in replicate design crossover studies the two types of outliers can be distinguished. "The *retest* character of these designs should indicate whether to delete an outlier or not." (FDA, 2001). Therefore, when considering the problem of outliers in bioequivalence studies it seems important to distinguish outliers potentially caused by "product or process failure" from outliers potentially caused by "subject-by-formulation interaction".

1.3 Prior Research and Focus of Present Paper

Various statistical methodologies have been developed for the identification of outliers in bioequivalence studies. Most work has focused on outlier detection in the 2×2 crossover design (Chow and Tse, 1990; Liu and Weng, 1991; Ki et al, 1995; Wang and Chow, 2003; Ramsay and Elkum, 2005; Liao, 2007). However, in recent years the use of replicate crossover designs for bioequivalence trials has become more prominent, in particular in the context of assessment of individual bioequivalence and assessment

of bioequivalence of highly variable drugs (FDA, 2001; Chow and Liu, 2000 chapters 14–15; Hauschke, Steinijans and Pigeot, 2007 chapter 9). A recent article on outliers in replicate design crossover trials by Lužar-Stiffler and Stiffler (2005) focuses on the detection of so-called "subject outliers".

In this paper we propose simple but effective residual plots for the identification of the following three types of outliers for the 2-treatment, 2-sequence, 4-period crossover design: (i) subject outliers, (ii) subject-by-formulation outliers, and (iii) single data point outliers. Our focus is not on outlier testing. However, the three sets of residuals pertaining to subject-by-formulation outliers and to single data point outliers in either the test or reference formulation are uncorrelated, and independent under normality. In principle, Grubb's outlier test can be applied to those residuals after suitable standardization.

2. TYPES OF OUTLIERS IN CROSSOVER TRIALS

2.1 Types of Outliers in 2×2 Crossover Trials

We consider the following model for the standard 2×2 crossover design:

$$y_{hij} = \mu + \zeta_h + s_{hi} + \pi_m + \tau_j + e_{hij} \tag{1}$$

Here y_{hij} is the observation for the j-th formulation (j = R [reference] or j = T [test]) on the i-th subject ($i = 1, \ldots, n_h$) in sequence $h = 1, 2$; μ the overall mean; ζ_h the fixed effect of the h-th sequence; s_{hi} the random effect of subject i in sequence h; π_m the m-th period effect ($m = 1, 2$); τ_j the j-th formulation effect; and e_{hij} the random error for formulation j and subject i in sequence h. As usual, the s_{hi} and e_{hij} are assumed mutually independent with mean zero and variances $\text{var}(s_{hi}) = \sigma_B^2$ and $\text{var}(e_{hij}) = \sigma_W^2$. Typically, y_{hij} would represent a measurement of the (log-transformed) pharmacokinetic endpoint of interest, such as log(AUC) or log($C_{\max}$).

In terms of Model (1), we can distinguish two types of outliers:

1. Subject outlier / between-subject outlier: For subject hi, both observations, y_{hiT} and y_{hiR}, are extreme. (The outlier could be modelled as a mean shift in the subject effect s_{hi}.) However, regarding the conventional mixed model analysis of average bioequivalence with complete data, a subject outlier has no consequences since a mean shift in s_{hi} has no effect on either the point or interval estimates of relative bioavailability. Thus, regarding only the analysis of bioequivalence, subject outliers could be ignored (of course, it may be interesting for reasons other than the mere analysis of bioequivalence to identify subject outliers, for example, in the identification of slow versus fast metabolizers of a drug.)

2. Single data point outlier / within-subject outlier: For subject hi, either y_{hiT} or y_{hiR}, or both (but in opposite directions), is extreme; specifically, for subject hi the within-subject difference $y_{hiT} - y_{hiR}$ is extreme. A single data point outlier could be modelled as a mean shift in either e_{iT} or e_{iR}. Single data point outliers can be problematic in that they can severely affect the results of the bioequivalence test, as pointed out above.

In summary, in standard 2×2 crossover trials, the (between) subject outlier generally is unproblematic (regarding strictly the test for bioequivalence). However, the potentially problematic within-subject outlier could be caused by any of the mechanisms outlined above: "product or process failure", and, crucially, also by "subject-by-formulation interaction", since both types of outliers can manifest as single data point (within-subject) outlier. Therefore, in 2×2 crossover trials it is usually not possible to distinguish outliers caused by product/process failure from outliers caused by subject-by-formulation interaction, *based on statistical criteria alone*. Consequently, as pointed out above, regulatory authorities do generally not accept removal of outliers from the statistical analysis of 2×2 crossover trials based *solely* on statistical criteria.

2.2 Types of Outliers in Replicate Design Crossover Trials

In the following we consider the 2-treatment, 2-sequence, 4-period replicate crossover design recommended by the relevant FDA guidance document (FDA, 2001). This design has the two sequences R T R T and T R T R; however, the methods outlined below can easily be adapted to a replicate design with the sequences R T T R and T R R T.

We consider the following model:

$$y_{hijk} = \mu + \zeta_h + s_{hi} + u_{hij} + \pi_m + \tau_j + e_{ijk} \tag{2}$$

Here y_{hijk} is the observation for the k-th replicate ($k = 1, 2$) of the j-th formulation (j = R,T) on the i-th subject ($i = 1, \ldots, n_h$) in sequence $h = 1, 2$; μ the overall mean; ζ_h the fixed effect of the h-th sequence; s_{hi} the random effect of subject i in sequence h; u_{hij} the (i, j)-th random subject-by-formulation effect in sequence h; π_m the m-th period effect ($m = 1, \ldots, 4$); τ_j the j-th formulation effect; and e_{hijk} the random error for replicate k, formulation j and subject i in sequence h.

The random terms s_{hi}, u_{hij} and e_{hijk} are assumed mutually independent with mean zero and the following variances: $\text{var}(s_{hi}) = \sigma_S^2$; $\text{var}(u_{hiT}) = \sigma_{ST}^2$ and $\text{var}(u_{hiR}) = \sigma_{SR}^2$ such that $\sigma_{BT}^2 = \sigma_S^2 + \sigma_{ST}^2$, $\sigma_{BR}^2 = \sigma_S^2 + \sigma_{SR}^2$ are the between-subject variances and $\sigma_D^2 = \sigma_{ST}^2 + \sigma_{SR}^2$ is the subject-by-formulation interaction variance component; finally $\text{var}(e_{hiTk}) = \sigma_{WT}^2$ and $\text{var}(e_{hiRk}) = \sigma_{WR}^2$ are the within-subject variances. The notation σ_{BT}^2, σ_{BR}^2, σ_{WT}^2, σ_{WR}^2 and σ_D^2 for the between- and within-subject variances is consistent with the FDA (2001) guidance document.

In terms of Model (2) three types of outliers can be distinguished:

1. Subject outlier / between-subject outlier: For subject hi, all 4 observations y_{hiT1}, y_{hiT2}, y_{hiR1}, y_{hiR2} are extreme. (The outlier could be modelled as a mean shift in the subject effect s_{hi}.) Again, regarding the conventional mixed model analysis of average bioequivalence with complete data, a subject outlier has no consequences since a mean shift in s_{hi} may inflate the estimate of between-subject variance, but does not affect

the point or interval estimate of relative bioavailability.

2. Subject-by-formulation outlier / within-subject but between-replicate outlier: For subject hi, the replicates y_{hiT1} and y_{hiT2}, jointly, are different from the replicates y_{hiR1} and y_{hiR2}; in other words, for subject hi the within-subject, between replicate difference $([y_{hiT1} + y_{hiT2}] - [y_{hiR1} + y_{hiR2}])$ is extreme. A subject-by-formulation outlier could be modelled as a mean shift in either u_{hiT} or u_{hiR}. Subject-by-formulation outliers can be problematic since they can severely affect the results of the bioequivalence test.

3. Single data point outlier / within-subject, within-replicate outlier: For subject hi, one of the four observations y_{hiT1}, y_{hiT2}, y_{hiR1}, or y_{hiR2} is extreme; that is, either one of the within-subject, within-replicate differences $(y_{hiT1} - y_{hiT2})$ or $(y_{hiR1} - y_{hiR2})$ is extreme. A single data point outlier could be modelled as a mean shift in either e_{hiT1}, e_{hiT2}, e_{hiR1} or e_{hiR2}. Single data point outliers can be problematic since they can severely affect the results of the bioequivalence test.

3. RESIDUALS

As a simple diagnostic tool for the identification and classification of outliers (into the above three types) in replicate design crossover studies we propose to calculate and plot four sets of residuals, of three types corresponding to the three different types of outliers that can be distinguished. The residuals are obtained as four (mutually orthogonal) linear contrasts of the four data points associated with each subject.

Initially, we form the following four contrasts of the four observations for each subject $(i = 1, \ldots, n_h; h = 1, 2)$:

$$\begin{aligned} c_{hi1} &= y_{hi\cdot\cdot} = (y_{hiT1} + y_{hiT2} + y_{hiR1} + y_{hiR2})/4 \\ c_{hi2} &= y_{hiT\cdot} - y_{hiR\cdot} = (y_{hiT1} + y_{hiT2})/2 - (y_{hiR1} + y_{hiR2})/2 \\ c_{hi3} &= (y_{hiT1} - y_{hiT2})/\sqrt{2} \\ c_{hi4} &= (y_{hiR1} - y_{hiR2})/\sqrt{2} \end{aligned}$$

For each subject, the four contrasts are, respectively

- c_{hi1}: the average of the four observations; extreme values of c_{hi1} would indicate a subject outlier.
- c_{hi2}: the difference of the treatment averages (over the two replicates) for subject hi; extreme values of c_{hi2} would indicate a subject-by-formulation outlier.
- c_{hi3}: the difference between the test formulation replicates; extreme values of c_{hi3} would indicate a data point outlier among the test formulation data.
- c_{hi4}: the difference between the reference formulation replicates; extreme values of c_{hi4} would indicate a data point outlier among the reference formulation data.

We note that the four contrasts c_{ghi} can be written as the following linear transformation of the four observations y_{hijk} of each subject; note that the rows of the transformation matrix are orthogonal (the correlation structure of the resulting contrasts is discussed below):

$$c_{hi} = \begin{pmatrix} c_{hi1} \\ c_{hi2} \\ c_{hi3} \\ c_{hi4} \end{pmatrix} = \begin{pmatrix} +2^{-2} & +2^{-2} & +2^{-2} & +2^{-2} \\ +2^{-1} & +2^{-1} & -2^{-1} & -2^{-1} \\ +2^{-\frac{1}{2}} & -2^{-\frac{1}{2}} & 0 & 0 \\ 0 & 0 & +2^{-\frac{1}{2}} & -2^{-\frac{1}{2}} \end{pmatrix} \begin{pmatrix} y_{hiT1} \\ y_{hiT2} \\ y_{hiR1} \\ y_{hiR2} \end{pmatrix}$$

In terms of model (2), the variances of the four contrasts are

$$\begin{aligned} \mathrm{var}(c_{hi1}) &= \sigma_S^2/2 + (\sigma_{BT}^2 + \sigma_{BR}^2)/4 + (\sigma_{WT}^2 + \sigma_{WR}^2)/8 \\ \mathrm{var}(c_{hi2}) &= \sigma_D^2 + (\sigma_{WT}^2 + \sigma_{WR}^2)/2 \\ \mathrm{var}(c_{hi3}) &= \sigma_{WT}^2 \\ \mathrm{var}(c_{hi4}) &= \sigma_{WR}^2 \end{aligned}$$

Remarks:

1. If $\sigma_D^2 = 0$ (absence of formulation-by-subject interaction variance component) and $\sigma_{WT}^2 = \sigma_{WR}^2 = \sigma_W^2$ (the within-test and within-reference variances are equal), then

$$\text{var}(c_{hi2}) = \text{var}(c_{hi3}) = \text{var}(c_{hi4}) = \sigma_W^2$$

that is, the contrasts c_{hi2}, c_{hi3} and c_{hi4} all have the same variance.

2. The contrasts c_{hi3} are independent of the contrasts c_{hi4}, because, under model (2), the random effects s_{hi} and u_{hij} cancel out when the within-formulation contrasts are formed, and the error terms e_{hiTk} are assumed independent of e_{hiRk}.

3. The contrasts c_{hi3} and c_{hi4} are uncorrelated with the contrasts c_{hi1} and c_{hi2}, because the difference $e_{hiT1} - e_{hiT2}$ is uncorrelated with the sum $e_{hiT1} + e_{hiT2}$; similarly $e_{hiT1} - e_{hiT2}$ is uncorrelated with $e_{hiT1} + e_{hiT2}$. Furthermore, under normality, the contrasts c_{hi3} and c_{hi4} are independent of the contrasts c_{hi1} and c_{hi2}.

4. If, as in Remark 1, $\sigma_D^2 = 0$ and $\sigma_{WT}^2 = \sigma_{WR}^2 = \sigma_W^2$, then the contrasts c_{hi1} and c_{hi2} are uncorrelated, and independent under normality.

The expected values of the four contrasts can be written as

$$\text{E}(c_{ghi}) = \gamma_{gh}, \qquad g = 1, \ldots, 4; \quad h = 1, 2$$

where the expected values γ_{gh} can be written in terms of the parameters of model (2) as follows:

$$\begin{aligned}
\gamma_{1h} &= \mu + (\tau_T + \tau_R)/2 + (\pi_1 + \pi_2 + \pi_3 + \pi_4)/4 + \zeta_h, \qquad \text{for } h = 1, 2 \\
\gamma_{2h} &= \begin{cases} (\tau_T - \tau_R) + (\pi_1 + \pi_3)/2 - (\pi_2 + \pi_4)/2, & \text{for } h = 1 \\ (\tau_T - \tau_R) + (\pi_2 + \pi_4)/2 - (\pi_1 + \pi_3)/2, & \text{for } h = 2 \end{cases} \\
\gamma_{3h} &= \begin{cases} (\pi_1 - \pi_3)/\sqrt{2}, & \text{for } h = 1 \\ (\pi_2 - \pi_4)/\sqrt{2}, & \text{for } h = 2 \end{cases}
\end{aligned}$$

$$\gamma_{4h} = \begin{cases} (\pi_2 - \pi_4)/\sqrt{2}, & \text{for } h = 1 \\ (\pi_1 - \pi_3)/\sqrt{2}, & \text{for } h = 2 \end{cases}$$

Thus each contrast follows a simple one-way (two-sample) layout with sample size n_1 and n_2, respectively, corresponding to the two design sequences. Residuals, therefore, can be defined by subtracting from the above contrasts the respective sequence averages:

$$\begin{aligned} r_{S_{hi}} &= c_{hi1} - c_{h\cdot 1} = y_{hi\cdot\cdot} - y_{h\cdots} \\ r_{SF_{hi}} &= c_{hi2} - c_{h\cdot 2} = (y_{hiT\cdot} - y_{hiR\cdot}) - (y_{h\cdot T\cdot} - y_{h\cdot R\cdot}) \\ r_{DT_{hi}} &= c_{hi3} - c_{h\cdot 3} = [(y_{hiT1} - y_{hiT2}) - (y_{h\cdot T1} - y_{h\cdot T2})]/\sqrt{2} \\ r_{DR_{hi}} &= c_{hi4} - c_{h\cdot 4} = [(y_{hiR1} - y_{hiR2}) - (y_{h\cdot R1} - y_{h\cdot R2})]/\sqrt{2} \end{aligned}$$

Studentized residuals $s_{S_{hi}}$, $s_{SF_{hi}}$, $s_{DT_{hi}}$, and $s_{DR_{hi}}$ are obtained by dividing the above (unstudentized) residuals by their standard error, for example,

$$s_{S_{hi}} = \frac{r_{S_{hi}}}{\sqrt{(1 - 1/n_h)\sum_{i,h} r^2_{S_{hi}}/(n_1 + n_2 - 2)}}$$

and similarly for $s_{SF_{hi}}$, $s_{DT_{hi}}$, and $s_{DR_{hi}}$.

4. RESIDUAL PLOTS

4.1 Uses of Unstudentized Residuals

A visual comparison of the variability of the residuals $r_{DT_{hi}}$ with the variability of the residuals $r_{DR_{hi}}$ will indicate the relative magnitude of the within-subject variances σ^2_{WT} and σ^2_{WR}; specifically, such a visual check will indicate whether any difference between the estimates of σ^2_{WT} and σ^2_{WR} might be due to single data point outliers.

Furthermore, a visual comparison of the variability of the residuals $r_{SF_{hi}}$ with the variability of $r_{DT_{hi}}$ and of $r_{DR_{hi}}$ will indicate the potential presence of subject-by-formulation interaction (if $r_{SF_{hi}}$ appears more variable than either $r_{DT_{hi}}$ and of $r_{DR_{hi}}$)

– see Remark 1 above. The plot of $r_{SF_{hi}}$ could also indicate whether any notable subject-by-formulation variance component might to due to subject-by-formulation outliers.

4.2 Uses of Studentized Residuals

For outlier detection, we suggest to plot the four sets of studentized residuals against the subject number, and to inspect the residuals in the following order:

1. Plot $s_{DT_{hi}}$ and $s_{DR_{hi}}$: Identify possible data point outliers among either the test or reference formulation replicates.

2. Plot $s_{SF_{hi}}$: Identify possible subject-by-formulation outliers.

3. Plot $s_{DT_{hi}}$: Identify possible subject outliers.

Note: When an outlier is detected among the residuals $s_{DT_{hi}}$ or $s_{DR_{hi}}$ for a given subject hi, one will refer back to the data, namely to the pattern of the four observations for the subject, to determine which of the two replicates is responsible for the outlier. In Section 5 we provide an example of this process.

4.3 Potential Contamination

We note the potential effects of contamination: Genuine data point outliers may create "false" subject-by-formulation outliers, since an extreme value of a single data point will not only cause the relevant residual $s_{DT_{hi}}$ or $s_{DR_{hi}}$ to be extreme, but may also cause the residual $s_{SF_{hi}}$ to be extreme. Furthermore, genuine (and false) subject-by-formulation outliers may create "false" subject outliers, through a similar process.

Therefore, we recommend to inspect the above plots successively in the order given. "First occurrence" of an extreme value will identify outlier type. That is, if a single data point outlier is identified in a plot of $s_{DT_{hi}}$ or $s_{DR_{hi}}$, then any extreme values of $s_{SF_{hi}}$ or $s_{S_{hi}}$ in the same subject are probably accounted for by the data point outlier. Similarly, if a (genuine) subject-by-formulation-outlier is identified in a plot of $s_{SF_{hi}}$,

then an extreme value of $s_{S_{hi}}$ in the same subject is probably accounted for by the subject-by-formulation-outlier.

4.4 Outlier Tests

As noted above, each of the four sets of residuals, $r_{S_{hi}}$, $r_{SF_{hi}}$, $r_{DT_{hi}}$ and $r_{DR_{hi}}$ follows a two-sample layout where the groups in question are determined by the design sequence. Therefore, extreme values in each of the four sets of residuals can be tested using outlier tests appropriate for the two-sample problem.

In the context of analysis of bioequivalence, the subject-by-formulation and data point outliers are of primary interest. The corresponding sets of residuals, namely $r_{DT_{hi}}$ and $r_{DR_{hi}}$, are mutually independent whatever the distributions of the random terms in model (2); furthermore, both $r_{DT_{hi}}$ and $r_{DR_{hi}}$ are independent of $r_{SF_{hi}}$ under normality. Outlier tests, therefore, could be based on the studentized residuals. If period and sequence effects are ignored, outlier tests for the one-sample problem, such as Grubb's test, can directly be applied to the three sets of residuals. If one wants an overall test for the maximum residual across the three sets, the distribution of the maximum residual in three independent samples needs to be determined.

5. EXAMPLE

We consider data of 31 subjects from a bioequivalence trial with a 2-treatment, 2-sequence, 4-period replicate crossover design. The replicate design was used for the trial because of the high pharmacokinetic variability of the drug. We present residual plots for the (log-transformed) pharmacokinetic variable AUC_∞, and statistical assessment of bioequivalence for the variables AUC_∞ and $\mathrm{C}_{\max}$.

5.1 Residual Plots

A plot of the studentized residuals $s_{DT_{hi}}$ (Figure 1) reveals nothing untoward, all residuals lying within a range of approximately -2.5 to $+2.5$ (the critical value for Grubb's test for a single sample of size $n = 31$ is 2.94; the dotted lines in Figures 1

– 4 indicate the critical value). However, the plot of the studentized residuals $s_{DR_{hi}}$ (Figure 2) reveals a dramatic outlier for subject 7, with $s_{DR_7} = 4.48$. In order to find out which of the two reference formulation replicates for subject 7 is responsible for the outlier, we refer to the data: the four data points for AUC_∞ are as follows (ng·h/mL): Test 1: 495.4; Test 2: 488.0; Reference 1: 581.4; Reference 2: 121.1. Clearly, the second replicate for the reference formulation is the single data point outlier.

A plot of the subject-by-formulation residuals $s_{SF_{hi}}$ (Figure 3) shows that the residual for subject 7 is the largest, apparently caused by contamination by the single data point outlier. However, all residuals, including that for subject 7, lie within the range of approximately -2.5 to $+2.5$.

Finally, a plot of the subject residuals $s_{S_{hi}}$ (Figure 4) reveals a distinct outlier for subject 31, with $s_{S_{31}} = 3.27$.

The plots of the unstudentized residuals $r_{DT_{hi}}$ and $r_{DR_{hi}}$ (Figures 5 and 6) suggest that the within-subject variability of the reference formulation is actually somewhat smaller than that of the test formulation, although the dramatic single data point outlier for the reference formulation is expected to inflate the estimate of σ^2_{WR} when the outlier is included in the analysis.

The plot of the $r_{SF_{hi}}$ residuals (Figure 7) suggests moderate excess variability over the within-subject residuals, particularly over the reference formulation. Therefore, one would expect a moderately large estimate of the subject-by-formulation variance component. There is no indication that the subject-by-formulation variance is caused by a single data point (or by a small number).

We note that plots of the residuals for C_{max} (not shown) and inspection of the data for subject 7 reveal the same picture as for AUC_∞, although the subject residual for subject 31 is not as extreme for C_{max} as it is for AUC_∞.

5.2 Assessment of Bioequivalence and Sensitivity Analysis

Table 1 presents the results of the statistical analysis of the data for AUC_∞ and C_{max}, both including and excluding the single data point outlier for subject 7. We fitted

the conventional mixed analysis of variance model (FDA, 2001). Note that the analysis excluding the single data point for subject 7, period 4 (R2 replicate) is equivalent to an analysis including all data but fitting a dummy variable representing a mean shift in the corresponding data point.

Clearly, the single data point outlier greatly inflated the estimate of the within-subject variability of the reference formulation of both AUC_∞ and C_{max}: when all data are analysed, the within-subject coefficient of variation for the reference is larger than for the test formulation, while the relationship is reversed when the single data point outlier is removed from analysis. We note that, for this example, removing the single data point outlier did not decrease the estimate of the subject-by-formulation interaction variance component.

Despite its size, the single data point outlier in subject 7 had no dramatic effect on the result of the bioequivalence test for the variable AUC_∞: whether or not the outlier is excluded from the analysis, the 90% confidence interval (CI) for the test/reference mean ratio of AUC_∞ is comfortably within the conventional bioequivalence range of $[0.8, 1.25]$. However, the situation is somewhat different for C_{max}: the analysis including the outlier very nearly fails to demonstrate bioequivalence, with an upper limit of the 90% CI of 1.248.

We also analysed the variable AUC_∞ adjusting for the potential subject outlier in subject 31, by fitting a dummy variable representing a mean shift in the corresponding subject effect. This procedure has of course no effect on either the point or interval estimate for the test/reference mean ratio, nor on the estimates of the within-subject or subject-by-formulation variance components (the slight difference in the latter is due to rounding error). However, the estimates of the between-subject variance components are greatly reduced when the subject outlier is removed (dummy variable fitted).

6. DISCUSSION

Outliers may only be removed from analysis of bioequivalence trials if they are

caused by process or product failures; when outliers are potentially caused by subject-by-formulation interaction they may not be removed from analysis. Without external documentation, in 2×2 crossover trials is not possible to distinguish outliers caused by process or product failures from outliers caused by subject-by-formulation, solely through statistical criteria.

The proposed residuals for the 2-treatment, 2-sequence, 4-period replicate crossover design can be used to identify and classify three types of outliers: (i) subject outliers, (ii) subject-by-formulation outliers, and (iii) single data point outliers in either the test or reference formulation. We recommend that those residuals are routinely calculated and plotted as a very simple and quick diagnostic check of bioequivalence data.

A necessary condition for removal of data from analysis would be the identification of an outlier as a single data point outlier, rather than a subject-by-formulation outlier. Furthermore, depending on the presumed mechanism causing the outlier, the outlier would have to be present in both AUC and C_{max} data. For example, if there is documentation that the outlier might be caused by vomiting shortly after drug administration, clearly both AUC and C_{max} would have to extremely low (indeed, the complete concentration-time profile would have to be low).

Finally, removal of data points from primary analysis should always be supported by a sensitivity analysis; thus analysis results would usually have to be presented for analyses both including and excluding the suspect data points.

REFERENCES

Braddy, A.C., Patel, D., Jackson, A.J., Davit, B., Conner, D. (2008). Statistical outliers: the significance and impact of re-dosing studies to establish bioequivalence. http:/www.aapsj.org/abstracts/AM_2008/AAPS2008-001564.PDF

Chow, S. C., Tse, S. K. (1990). Outlier detection in bioavailability/bioequivalence studies. *Statistics in Medicine* 9:549-558.

Chow, S. C., Liu, J. P. (2000). *Design and Analysis of Bioavailability and Bioequiva-*

lence Studies. Second Edition. New York: Marcel Dekker, Inc.

European Agency for the Evaluation of Medicinal Products (EMEA) (2001). The Investigation of Bioavailability and Bioequivalence. Note for Guidance. Committee for Proprietary Medicinal Products (CPMP), London.

European Agency for the Evaluation of Medicinal Products (EMEA) (2006). Questions and Answers on the Bioavailability and Bioequivalence Guideline. EMEA/CHMP/EWP/40326/2006, London.

Food and Drug Administration (FDA) (2001). Statistical Approaches to Establishing Bioequivalence. Guidance for Industry. Center for Drug Evaluation and Research (CDER), Rockville, Maryland.

Food and Drug Administration (FDA) (2003). Bioavailability and Bioequivalence Studies for Orally Administered Drug Products General Considerations. Guidance for Industry. Center for Drug Evaluation and Research (CDER), Rockville, Maryland.

Hauschke, D., Steinijans, V., Pigeot, I. (2007). *Bioequivalence Studies in Drug Development. Methods and Applications.* Chichester: Wiley.

Ki, F. Y. C., Liu, J. P., Wang, W., Chow, S. C. (1995). The impact of outlying subjects on decision of bioequivalence. *Journal of Biopharmaceutical Statistics* 5:71-94.

Liao, J. J. Z. (2007). A new approach for outliers in a bioavailability/bioequivalence study. *Journal of Biopharmaceutical Statistics* 17:393–405.

Lužar-Stiffler, V., Stiffler, C. (2005). Higher-order crossover design outlier detection. In: *Proceedings of the 27th International Conference on Information Technology Interfaces.* Zagreb: University of Zagreb, 650-655.

Liu, J. P., Weng, C. S. (1991). Detection of outlying data in bioavailability/bioequivalence studies. *Statistics in Medicine* 10:1375-1389.

Ramsay, T., Elkum, N. (2005). A comparison of four different methods for outlier detection in bioequivalence studies. *Journal of Biopharmaceutical Statistics* 15:43-

52.

Wang, W., Chow, S. C. (2003). Examining outlying subjects and outlying records in bioequivalence trials. *Journal of Biopharmaceutical Statistics* 13:43-56.

Table 1. Mixed Model ANOVA of Replicate Design Crossover Trial

Analysis Variable	Outlier removed	PE	90% CI	CV_{WT} (%)	CV_{WR} (%)	CV_D (%)	CV_{BT} (%)	CV_{BR} (%)
AUC_∞	No	1.024	(0.954, 1.100)	16.2	25.5	10.2	64.6	57.7
	Yes[1]	1.000	(0.938, 1.067)	16.8	14.2	14.4	64.5	61.9
	Yes[2]	1.024	(0.954, 1.100)	16.2	25.5	10.3	50.1	45.3
C_{max}	No	1.124	(1.011, 1.248)	31.1	32.8	14.7	57.3	52.9
	Yes[1]	1.092	(0.989, 1.206)	30.7	21.0	20.0	57.4	59.8

[1] Data point for subject 7, period 4 (R2 replicate) removed

[2] Dummy variable for mean shift in subject effect (subject 37) fitted

PE: point estimate for test/reference ratio of geometric means; CI: confidence interval; CV: coefficient of variation; $CV(\%) = 100 \cdot \sqrt{\exp(\sigma^2) - 1)}$, where σ^2 is the relevant variance component on the log-scale

Figure 1: Studentized s_{DT} Residual

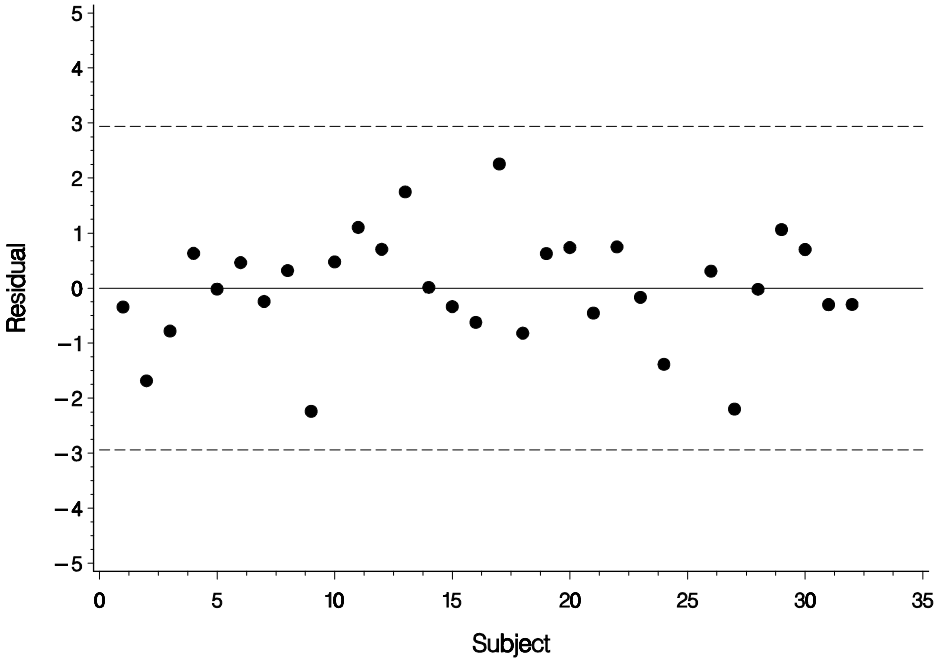

Figure 2: Studentized s_{DR} Residual

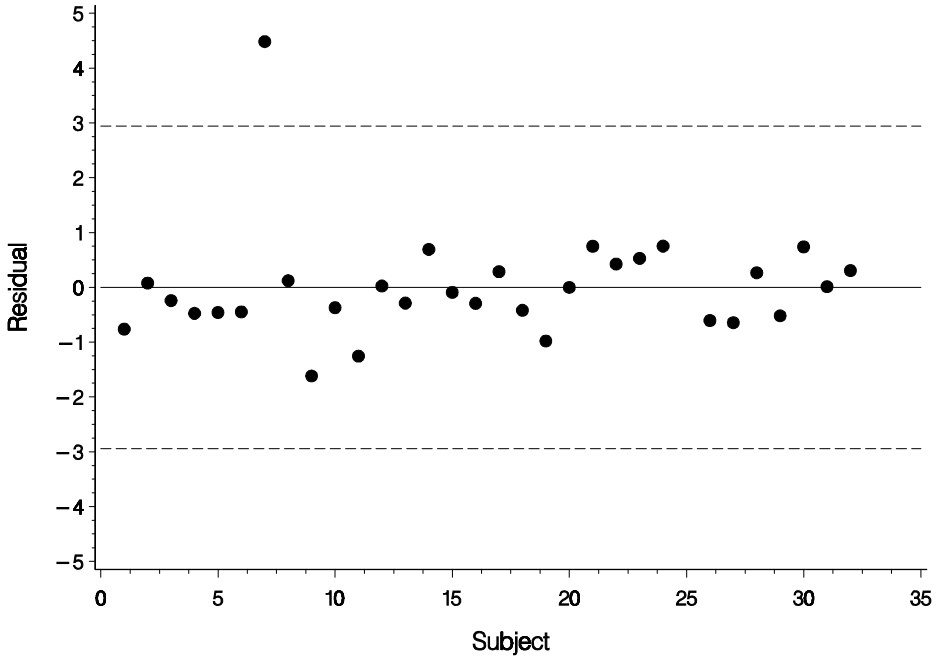

Figure 3: Studentized s_{SF} Residual

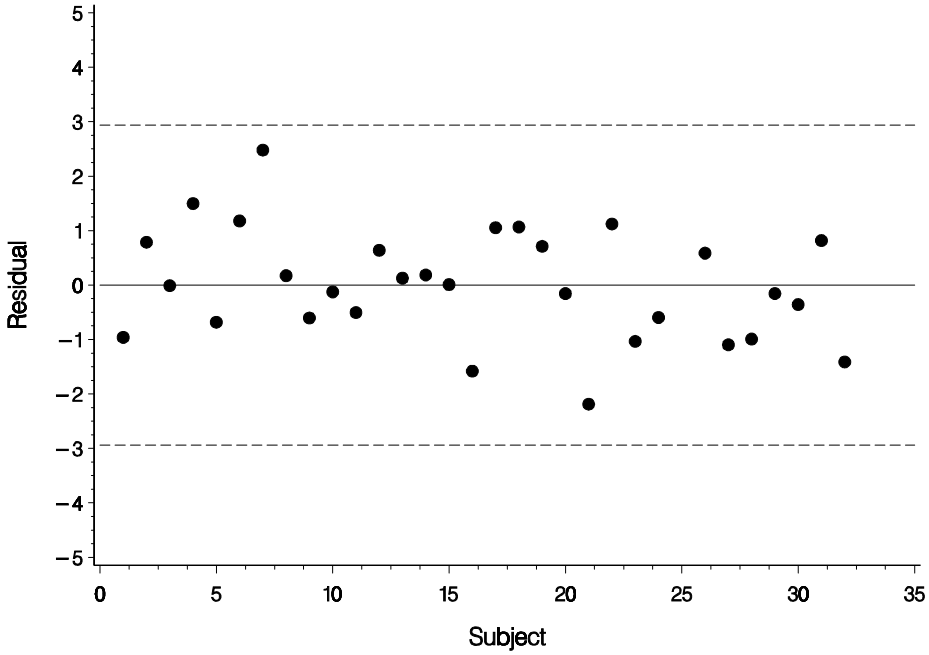

Figure 4: Studentized s_S Residual

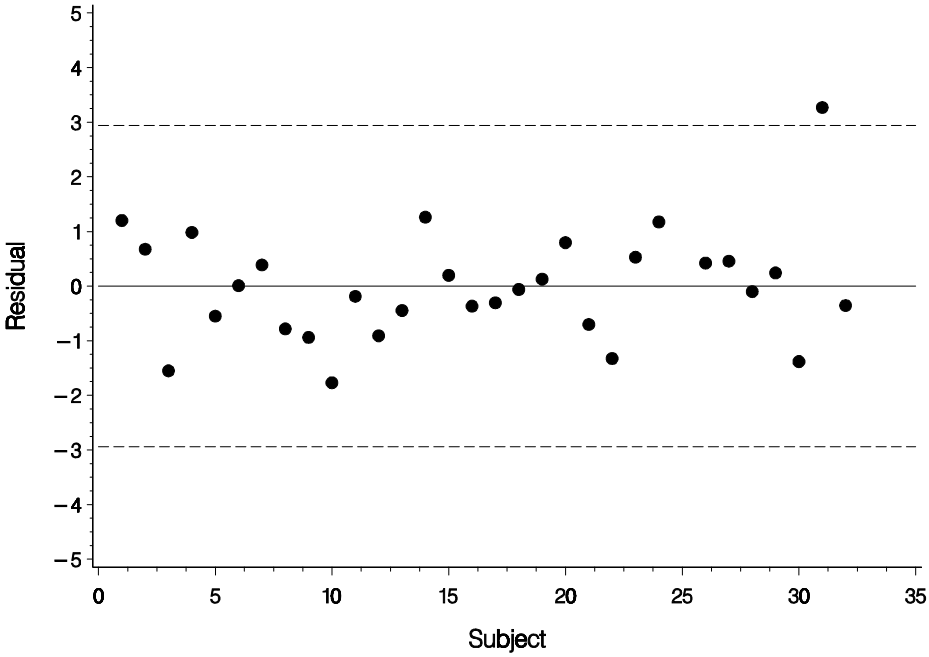

Figure 5: Unstudentized r_{DT} Residual

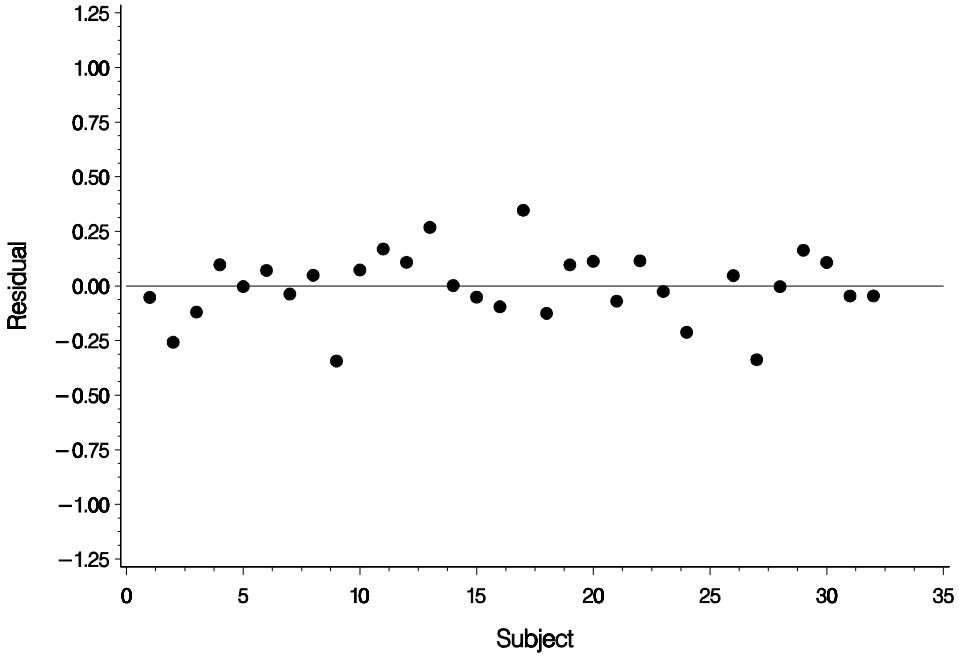

Figure 6: Unstudentized r_{DR} Residual

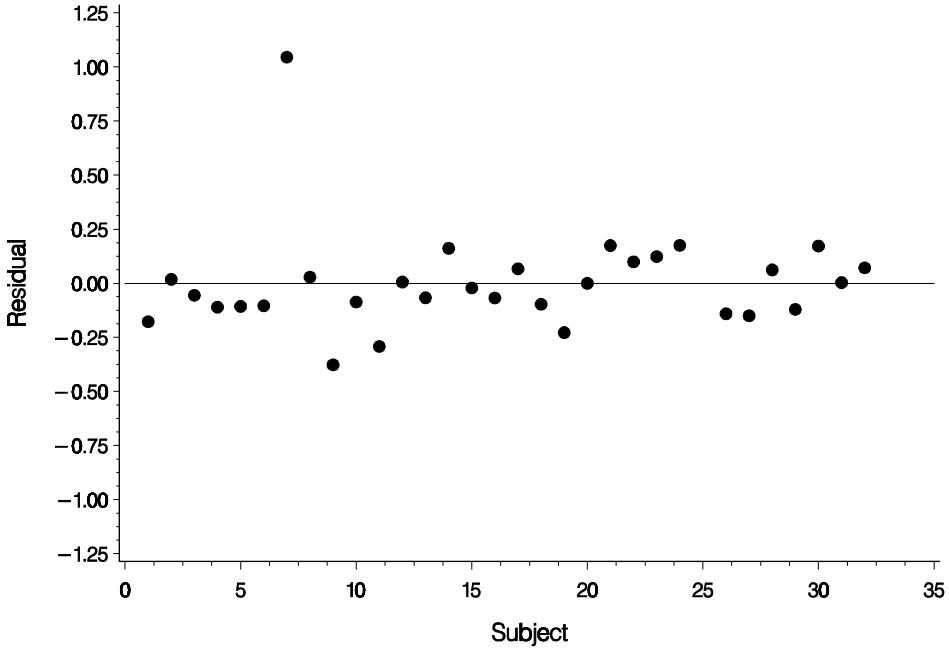

Figure 7: Unstudentized r_{SF} Residual