The AAPS Journal, Vol. 16, No. 6, November 2014 (© 2014)
DOI: 10.1208/s12248-014-9661-0

Research Article

Reference Datasets for 2-Treatment, 2-Sequence, 2-Period Bioequivalence Studies

Helmut Schütz,[1] Detlew Labes,[2] and Anders Fuglsang[3,4]

Received 13 April 2014; accepted 25 August 2014; published online 12 September 2014

Abstract. It is difficult to validate statistical software used to assess bioequivalence since very few datasets with known results are in the public domain, and the few that are published are of moderate size and balanced. The purpose of this paper is therefore to introduce reference datasets of varying complexity in terms of dataset size and characteristics (balance, range, outlier presence, residual error distribution) for 2-treatment, 2-period, 2-sequence bioequivalence studies and to report their point estimates and 90% confidence intervals which companies can use to validate their installations. The results for these datasets were calculated using the commercial packages EquivTest, Kinetica, SAS and WinNonlin, and the non-commercial package R. The results of three of these packages mostly agree, but imbalance between sequences seems to provoke questionable results with one package, which illustrates well the need for proper software validation.

KEY WORDS: bioequivalence; crossover; software validation.

INTRODUCTION

Proof of bioequivalence is a prerequisite for approving alternative product formulations (*e.g.*, generic drug products or modification of innovator formulation). Bioequivalence concepts can also be applied to the evaluation of dosing conditions such as food effects or alternative routes of administration (such as intramuscular *versus* subcutaneous injection). For drugs that are absorbed into the general circulation and distributed to the target issue, bioequivalence studies typically—with a few exceptions—involve a comparison of pharmacokinetic profiles of the proposed drug (hereafter called Test or T; often a new formulation that is being introduced) and the reference drug (hereafter called Reference, Ref, or R) obtained following a single-dose administration. In some cases, steady-state or multiple-dose studies are conducted. The most commonly applied design is a 2-treatment, 2-period, 2-sequence crossover design, where each subject is randomized into one of two sequences, termed TR and RT, depending on which drug is administered in the first and second period, respectively. In some cases, steady-state or multiple-dose studies are conducted. Primary bioequivalence endpoints are typically the 90% confidence intervals about the geometric mean ratios (GMRs) for critical pharmacokinetic metrics such as $C_{\max}$ and AUC. To demonstrate product bioequivalence as defined by the major regulatory bodies around the world, these confidence intervals must typically fall entirely within the range 80.00–125.00% for the primary pharmacokinetic metrics (with certain exceptions such as $C_{\max}$ in Canada where only the point estimate is evaluated).

To construct these confidence intervals, the log-transformed data from a 2-treatment, 2-period, 2-sequence study (hereinafter called a 2,2,2-BE study) is typically fitted to a general linear model with factors Treatment, Sequence, Subject (Sequence), and Period.

There are, in principle, many ways to find the maximum likelihood effect estimates. The analytical solution to the normal linear model in matrix form is the effect vector b which minimizes the residual sum of squares:

$$b = (X^t X)^{-1} X^t y \qquad (1)$$

where X is the full-rank design matrix and y is the vector of observed values (typically these are log-transformed $C_{\max}$ or AUC_t values).

Deriving b directly this way involves inversion of the matrix $X^t X$ and it is well known that some algorithms for matrix inversion have poor numerical properties. To circumvent that issue, in some cases, the b vector may be estimated using algorithms that are based not on direct matrix inversion but which instead rely on a stable decomposition of X, for example QR factorisation, singular value decomposition, and more. The comprehensive statistical package R uses a QR factorization as

Electronic supplementary material The online version of this article (doi:10.1208/s12248-014-9661-0) contains supplementary material, which is available to authorized users.

[1] Consultancy Services for Bioequivalence and Bioavailability Studies, Neubaugasse 36/11, 1070, Vienna, Austria.
[2] Cooperative Clinical Drug Research and Development AG, Lindenallee 70, 15366, Hoppegarten, Germany.
[3] Fuglsang Pharma, Hiort Lorenzens Vej 6c st. tv., 6100, Haderslev, Denmark.
[4] To whom correspondence should be addressed. (e-mail: anfu@fuglsangpharma.com)

1550-7416/14/0600-1292/0 © 2014 American Association of Pharmaceutical Scientists

the default method for fitting a normal linear model. In other cases, the components of the *b* vector can be found by iterative optimisation, for instance through a simplex or Newton-based algorithm. Most commercially available packages (such as Kinetica, SAS, WinNonlin, Splus, *etc.*) do not clearly describe the type of algorithms used, the way they are implemented or the corresponding limitations associated with these implementations.

Chow and Liu provided equations that can be used to find the confidence interval without using matrix algebra and which can be implemented in a spreadsheet or even done with a standard calculator [5].

The formulae for calculating the confidence interval upper and lower limits on the logarithmic scale, ln(*Limits*), in a 2,2,2-BE trial are generally:

$$\ln(\text{Limits}) = (\text{In}(\text{Test}) - \text{In}(\text{Ref})) \pm t_{df,\alpha}\sqrt{0.5V\left(\frac{1}{n_{\text{TR}}}+\frac{1}{n_{\text{RT}}}\right)} \quad (2)$$

where ln(Test) and ln(Ref) are the effect estimates of test and reference, respectively, *V* is the mean squared error (error variance), $t_{df,\ \alpha}$ is the critical value of the *t* distribution at *df* degrees of freedom and at the appropriate alpha and where n_{TR} and n_{RT} are the number of subjects in sequence TR and RT, respectively.

Computationally, the ability to accurately determine confidence intervals relies on accurate determination of maximum likelihood treatment effects and these may therefore be affected by numerical precision issues, available memory, and on actual algorithmic implementations to determine, *e.g.*, variances.

For a company that wishes to validate software for the calculation of bioequivalence, it is thus not really easy to know if or when the software correctly calculates the confidence interval, and/or if it does so consistently for all datasets. One may rarely find a dataset in the public domain with a published solution, and then compare one's own software's output for the same dataset to the published result. While this provides some degree of confidence, it might not be sufficient. For example, if the published dataset one is comparing to is balanced and our software passes, then how do we now our own software is valid for imbalanced datasets? Or in the presence of outliers?

ICH guideline E9 (adopted by regulators in, *e.g.*, US, EU, and Canada) specifies:

> *The credibility of the numerical results of the analysis depends on the quality and validity of the methods and software (both internally and externally written) used both for data management (data entry, storage, verification, correction and retrieval) and also for processing the data statistically. Data management activities should therefore be based on thorough and effective standard operating procedures. The computer software used for data management and statistical analysis should be reliable, and documentation of appropriate software testing procedures should be available* [6].

Large statistical frameworks like R or SAS can perform almost any task but the user must program the software to perform the desired task, and this programming step may itself introduce some weaknesses in terms of general performance.

FDA's current thinking of the Sponsor's responsibility towards ensuring that the software used in clinical trials reads:

> *For software purchased off-the-shelf, most of the validation should have been done by the company that wrote the software. The sponsor or contract research organization should have documentation (either original validation documents or on-site vendor audit documents) of this design level validation by the vendor, and should have itself performed functional testing (e.g., by use of test data sets) and researched known software limitations, problems, and defect corrections.*

> In the special case of database and spreadsheet software that is [1] purchased off-the-shelf, [2] designed for and widely used for general purposes, [3] unmodified, and [4] not being used for direct entry of data, the sponsor or contract research organization may not have documentation of design level validation. However, the sponsor or contract research organization should have itself performed functional testing (e.g., by use of test data sets) and researched known software limitations, problems, and defect corrections [7].

Thus, functional testing is a necessity that the Sponsor must perform.

In this paper, we hence aim to introduce a range of data sets of varying complexity and with proposed known solutions in terms of the point estimates and confidence intervals. These datasets can be used for performance qualification of software intended for evaluation of 2,2,2-BE studies.

The datasets will include two standard datasets from the public domain, and simulated/manipulated datasets that vary in size and/or introduce outliers, imbalance between sequences, extreme data ranges, mixtures of the factors, and we also evaluate a dataset with a non-Gaussian error distribution.

To increase trust in the results obtained, we show results obtained *via* independent statistical packages; *i.e.*, EquivTest/PK (Statistical Solutions 2006), Kinetica (version 5.0.10, Thermo Scientific 2007), SAS (version 9.2, SAS Institute 2008) and Phoenix/WinNonlin (version 6.3.0.395, Pharsight Corp. 2012), R (3.0.2, R Foundation for Scientific Computing, 2013).

Datasets for parallel designs, 3-treatment designs, and 3- or 4-period designs are all outside the scope of this paper, as are sequential designs and reference scaling. Along the same lines, we do not aim to discuss construction of ANOVAs or the use of type I *vs.* type III sums of squares since the decision regarding bioequivalence is not depending on these.

MATERIALS AND METHODS

The Datasets

All datasets discussed here can be downloaded as tab-delimited text files *via* the Electronic supplementary material on the journal's homepage.

- A Actual dataset from Sauter *et al.* 1992 [8]. Evaluated by Hauschke *et al.* 2007 [9]. Balanced with a total sample size of n=18 completers, *i.e.*, nine per sequence. Figure 1 shows the subject profiles split on sequence.
- B Actual dataset from Clayton and Leslie 1981 [10]. Evaluated by Chow and Liu 2009. Balanced with n= 18 completers. Figure 2 shows the subject profiles split on sequence.
- C Dataset with imbalance between sequences. Based on the dataset of Clayton and Leslie, from which subjects 2, 3 and 6, 8, 9 were deleted to yield a dataset with n= 13 completers: four in sequence TR and nine in sequence RT. Figure 3 shows the subject profiles split on sequence.
- D Dataset with extreme range in raw data input. Based on the dataset of Clayton and Leslie, where the data for subjects 4, 14, and 18 were multiplied by 10^6. Figure 4 shows the subject profiles split on sequence.
- E Dataset with an outlier. An aberrant value in terms of the T/R for a single subject has been introduced. Based on the dataset of Clayton and Leslie, where the data for subject 3 in period 1 has been multiplied by 100. Figure 5 shows the subject profiles split on sequence.
- F Simulated dataset with n=100 and balance between sequences, and where the residual is not normally distributed. A sine function was used to generate the dataset. Figure 6 shows the subject profiles split on sequence.
- G Simulated dataset with a high number of subjects. n= 1,000 with balance between sequences. Data were simulated using Mersenne-Twister as pseudo-random number generator and with a subsequent Box-Muller transformation to introduce an approximately normal distributed residual. Figure 7 shows the subject profiles split on sequence.
- H This is the large dataset G where
  - An extreme range in raw data input has been introduced by multiplying all raw input for subjects 70–90 by 10^6
  - Three outliers in terms of the T/R have been introduced by multiplying the Period 1 data for Subject 24, 51 (sequence TR) and 871 (sequence RT) by 10^6
  - Imbalance between sequences has been introduced by deletion of subjects 430–712

Figure 8 shows the subject profiles split on sequence.

Software and Hardware

EquivTest/PK (Statistical Solutions 2006), Kinetica (version 5.0.10, Thermo Scientific 2007), SAS (version 9.2, SAS Institute 2008), Phoenix/WinNonlin (version 6.3.0.395, Pharsight Corp. 2012), and R (3.0.2, R Foundation for Scientific Computing, 2013) were used to evaluate the datasets listed above. The authors are aware of all these softwares being actually used for regulatory submission purposes at the present day in various territories, including EU, USA, and Canada. R was executed on a Microsoft Windows 8 machine running in 64-bit mode. Kinetica and Phoenix/WinNonlin were executed on Windows XP and EquivTest/PK on Windows Vista; all in 32-bit mode. SAS was executed as 32-bit program on a machine with 64-bit Windows 7. The Electronic supplementary material includes scripts used for SAS and R. The menu-driven softwares were operated according to the manuals.

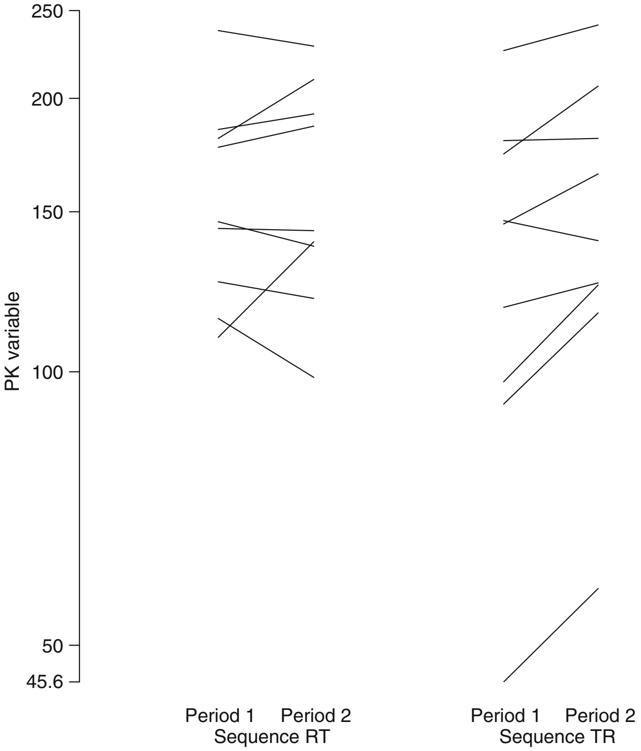

Fig. 1. Subject plots ordered by period within treatment sequence for dataset *A*

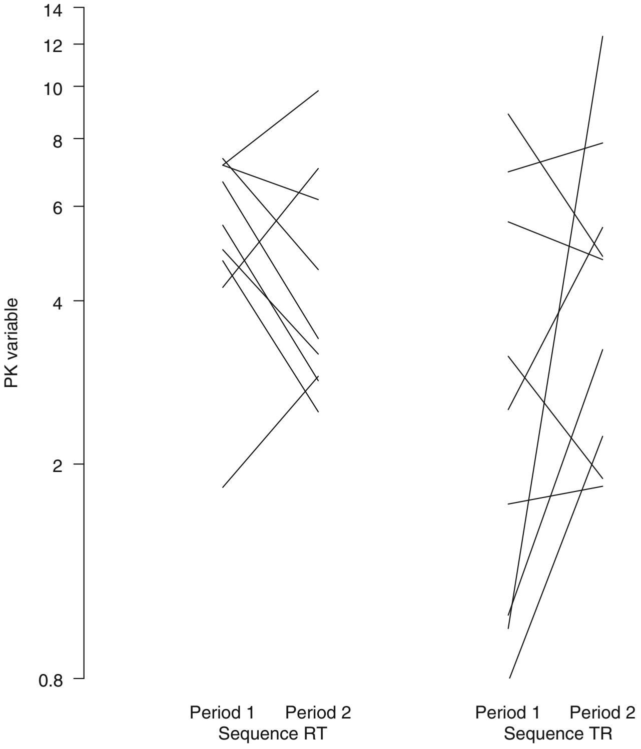

Fig. 2. Subject plots ordered by period within treatment sequence for dataset *B*

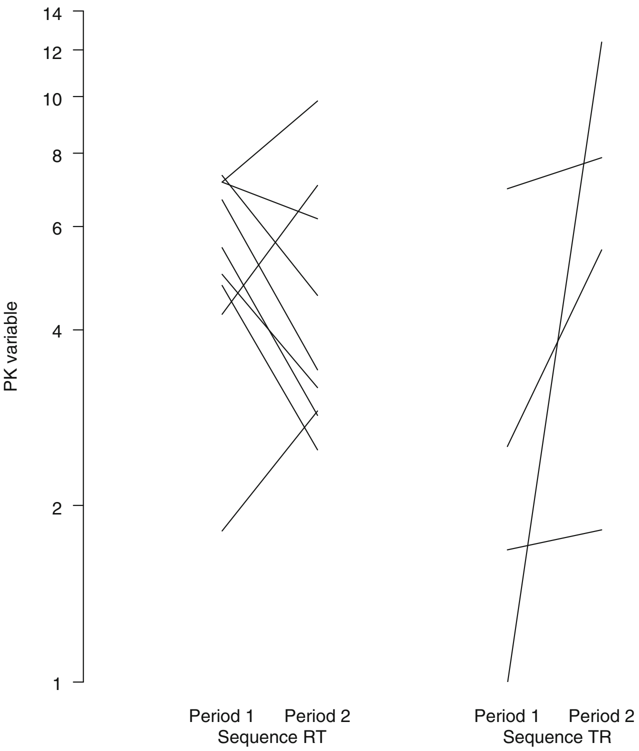

Fig. 3. Subject plots ordered by period within treatment sequence for dataset *C*. This dataset is based on Clayton and Leslie's dataset (*B*) where imbalance has been introduced by removing five subjects

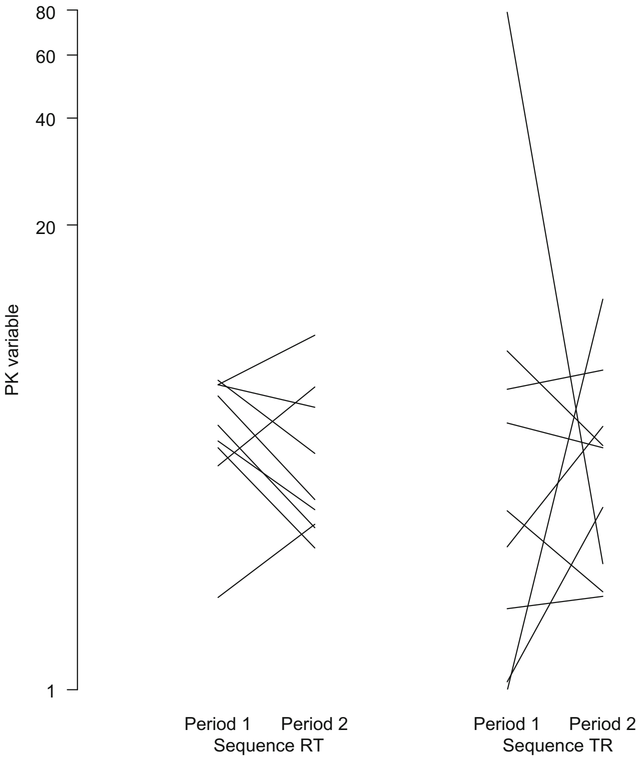

Fig. 5. Subject plots ordered by period within treatment sequence for dataset *E*. This dataset is based on Clayton and Leslie's dataset (*B*) where an outlier has been introduced

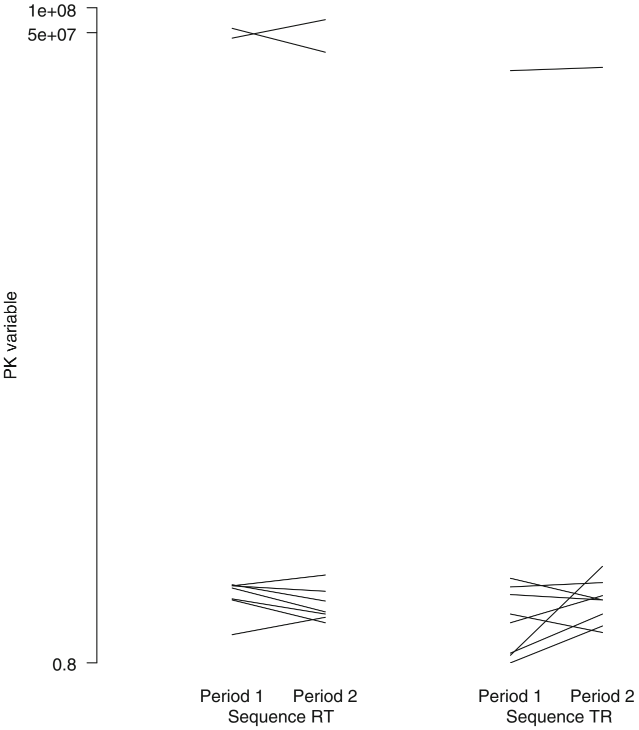

Fig. 4. Subject plots ordered by period within treatment sequence for dataset *D*. This dataset is based on Clayton and Leslie's dataset (*B*) where the raw data for three subjects was multiplied by 10^6

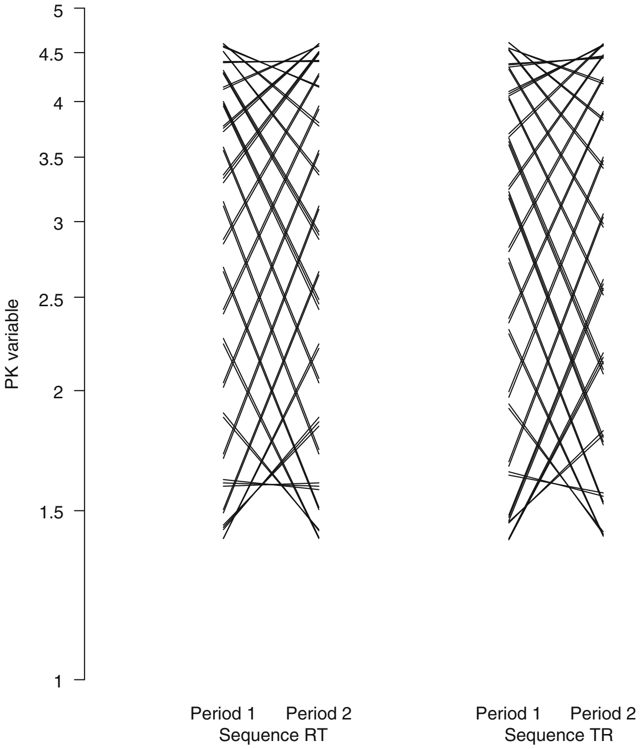

Fig. 6. Subject plots ordered by period within treatment sequence for dataset *F*. This is a purely simulated dataset with a non-Gaussian residual

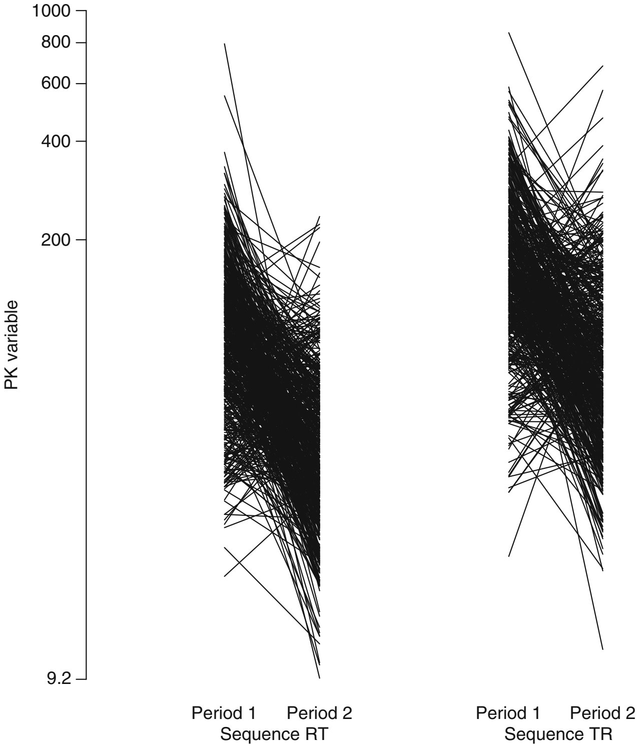

Fig. 7. Subject plots ordered by period within treatment sequence for dataset *G*. This is a large simulated dataset

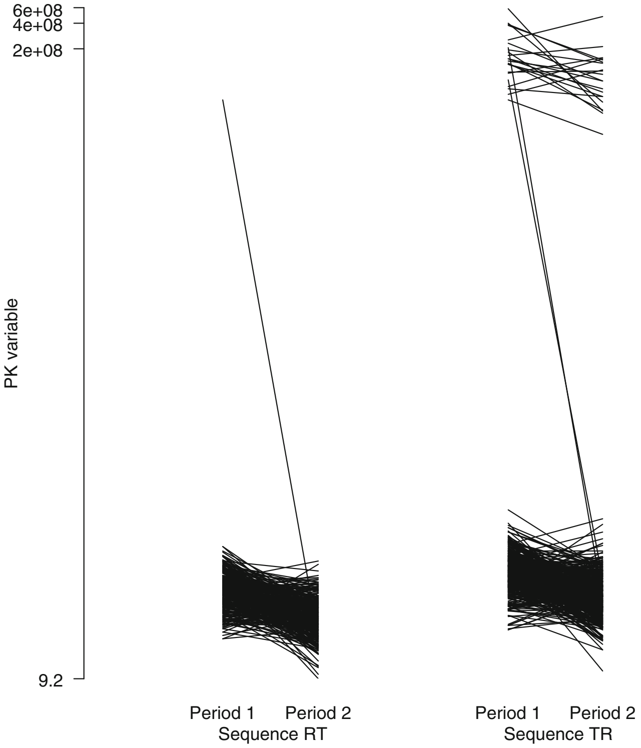

Fig. 8. Subject plots ordered by period within treatment sequence for dataset *H*. This dataset is based on the large simulated dataset (*G*) with imbalance, outliers, and an extreme range in raw data input

RESULTS AND DISCUSSION

Table I lists the datasets along with the evaluations obtained with the different statistical packages EquivTest/PK, Kinetica (5.0.10), SAS (9.2), Phoenix/WinNonlin (6.3.0.395), and R (3.0.2). Note that datasets *D* and *B* give the same results; this is expected since the within-subject treatment ratios (or the logarithmic within-subject treatment differences) are the same between the two sets.

There is perfect agreement between EquivTest, SAS, WinNonlin, and R for the datasets tested.

Interestingly, Kinetica gave results that were not in agreement with EquivTest/PK, SAS, WinNonlin and R, when we evaluated imbalanced datasets. We believe we are able to explain partially why, and this illustrates beautifully why software validation is necessary:

The formula for calculation of the confidence interval limits is given in Eq. 2.

Note that when there is balance between sequences, we have $n_{\mathrm{TR}}=n_{\mathrm{RT}}$ ($=n_{\mathrm{ps}}$ for simplicity; number of subjects per sequence), which implies:

$$n_{\mathrm{ps}} = (n_{\mathrm{TR}} + n_{\mathrm{RT}})/2 \tag{3}$$

Using Eq. 3 in combination II, we can derive a simpler expression which is valid for balanced datasets:

$$\ln(\mathrm{Limits}) = (\ln(\mathrm{Test}) - \ln(\mathrm{Ref})) \pm t_{df,\alpha}\sqrt{0.5V\left(\frac{1}{n_{\mathrm{TR}}}+\frac{1}{n_{\mathrm{RT}}}\right)}$$
$$\Rightarrow$$
$$\ln(\mathrm{Limits}) = (\ln(\mathrm{Test}) - \ln(\mathrm{Ref})) \pm t_{df,\alpha}\sqrt{V/n_{\mathrm{ps}}} \tag{4}$$

This is how the equation for calculating the confidence interval is sometimes actually presented, see, *e.g.*, Rani and Pargal [11]. But it must be emphasized that if there is imbalance between sequences, *i.e.*, $n_{TR} \neq n_{RT}$, then Eq. 4 does not give the proper result.

We speculate that it might be a possibility that Kinetica (at least on the machine used for our calculations) is using Eq. 4 on balanced as well as unbalanced datasets. The results obtained with datasets *C* and *H* are consistent with this approach given Kinetica's output of mean squared error and treatment effects, for example it would appear that for dataset *C* Kinetica gave a mean squared error of 0.3173, a critical *t* value of 1.7959, an upper confidence limit of 0.9931 and a point estimate of 0.6678, corresponding to

$$\sqrt{0.5 \times 0.3173\left(\frac{1}{n_{\mathrm{TR}}}+\frac{1}{n_{\mathrm{RT}}}\right)} = 0.2209$$

This value of 0.2209 can be obtained by setting ($n_{\mathrm{TR}}=n_{\mathrm{RT}}=6.5=(4+9)/2$) but not—within reasonable rounding—by using the corresponding integer combinations 4 and 9 which are the numbers of subjects in the two sequences. We are not aware of any relevant alternative setting or change of the software installation that can force Kinetica to give another result for datasets *C* and *H*. Our own calculations of the point estimates for the GMR, *e.g.*, obtained *via* the equations given by Chow and Liu, agree with the results of EquivTest, SAS, WinNonlin, and R. Further to this, it would appear that Kinetica's point estimates might reflect the means of

Table I. The datasets along with the evaluations obtained with the different statistical packages

	Point estimate (90% confidence interval)				
Dataset	EquivTest/PK	Kinetica	SAS	WinNonlin	R
A	95.09 (90.76, 99.62)[a]	95.09 (90.76, 99.62)[a]	95.09 (90.76, 99.62)[a]	95.09 (90.76, 99.62)[a]	95.09 (90.76, 99.62)[a]
B	71.10 (51.45, 98.26)[b]	71.10 (51.45, 98.26)[b]	71.10 (51.45, 98.26)[b]	71.10 (51.45, 98.26)[b]	71.10 (51.45, 98.26)[b]
C	58.56 (39.41, 87.03)	66.78 (44.91, 99.31)	58.56 (39.41, 87.03)	58.56 (39.41, 87.03)	58.56 (39.41, 87.03)
D	71.10 (51.45, 98.26)	71.10 (51.45, 98.26)	71.10 (51.45, 98.26)	71.10 (51.45, 98.26)	71.10 (51.45, 98.26)
E	91.83 (55.71, 151.37)	91.83 (55.71, 151.37)	91.83 (55.71, 151.37)	91.83 (55.71, 151.37)	91.83 (55.71, 151.37)
F	99.89 (93.37, 106.86)	99.89 (93.37, 106.86)	99.89 (93.37, 106.86)	99.89 (93.37, 106.86)	99.89 (93.37, 106.86)
G	92.15 (88.46, 95.99)	92.15 (88.46, 95.99)	92.15 (88.46, 95.99)	92.15 (88.46, 95.99)	92.15 (88.46, 95.99)
H	93.42 (86.81, 100.55)	107.80 (100.31, 115.85)	93.42 (86.81, 100.55)	93.42 (86.81, 100.55)	93.42 (86.81, 100.55)

Point estimates and 90% confidence intervals are given in percent with two decimals in accordance with the rounding principle suggested by the FDA and EMA

[a] Accords with section 6.2 of Hauschke et al

[b] Accords with section 6.9.1 of Chow & Liu

logarithmized raw treatment values rather than using the treatment effects that arise from minimizing the sums of squares.

This could explain why Kinetica's results with the installation used here are in agreement with SAS, R, EquivTest, and WinNonlin for balanced datasets, and also explains the point estimates and confidence intervals obtained with Kinetica for the imbalanced datasets.

While we do not aim to recommend a specific statistical package for evaluation of bioequivalence it can be mentioned that the comprehensive statistical framework R is non-commercial (free of charge), available for a variety of operative systems, and recently FDA have moved in a direction that specifically suggests that the use of validated R code is acceptable for submissions purposes [12]. The Electronic supplementary material includes a print of an R session with code used to evaluate the datasets introduced in this paper, as well as the results obtained for these datasets.

CONCLUSION

We propose the datasets used in this paper as a basis for software validation/qualification among companies analysing 2-treatment, 2-sequence, 2-period bioequivalence studies, and that the results obtained with EquivTest/PK, SAS, WinNonlin, and R are considered as validation goalposts. The results we obtain with Kinetica do not agree with these results for the imbalanced datasets, illustrating well why it should not be assumed that software that is installed and executing without warnings and errors and even giving (presumably) correct results for *some* published datasets will give correct results for *all* other datasets. The authors would appreciate feedback from other users of Kinetica, in particular, we would appreciate feedback re: Kinetica's performance for datasets *C* and *H*. It should be emphasized that the primary purpose of this paper is not in its own right to compare performance between different packages, or to judge the individual packages' fitness for bioequivalence evaluation purposes. It is not in any way implied in this paper that Kinetica generally miscalculates imbalanced datasets, but by way of the results obtained with Kinetica for the datasets introduced here we conclude that the installation of Kinetica used was not adequately performance qualified.

The datasets introduced in this paper can all be downloaded as Electronic supplementary material from the journal's website.

REFERENCES

1. European Medicines Agency, Committee for Human Medicinal Products. Guideline on the Investigation of Bioequivalence. CPMP/EWP/QWP/1401/98 Rev. 1/ Corr. 2010.
2. US Food and Drug Administration. Bioavailability and Bioequivalence Studies for Orally Administered Drug Products — General Considerations. 2003.
3. World Health Organization. Multisource (generic) pharmaceutical products: guidelines on registration requirements to establish interchangeability. In: Fortieth report of the WHO Expert Committee on Specifications for Pharmaceutical Preparations. Geneva, World Health Organization. WHO Technical Report Series, No. 937, 2006, Annex 7.
4. Health Canada, Therapeutic Products Directorate. Conduct and Analysis of Comparative Bioavailability Studies. 2012.
5. Chow S-C, Liu J-P. Design and analysis of bioavailability and bioequivalence studies. 3rd ed. Boca Raton: CRC; 2009.
6. International Conference on Harmonization. Statistical Principles for Clinical Trials, guideline E9. 1998.
7. US Food and Drug Administration. Computerized Systems Used in Clinical Trials. 1999.
8. Sauter R, Steinijans VW, Diletti E, Böhm E, Schulz H-U. Presentation of results from bioequivalence studies. Int J Clin Pharmacol Ther Toxicol. 1992;30:S7–30.
9. Hauschke D, Steinijans V, Pigeot I. Bioequivalence studies in drug development. Methods and applications. Chichester: Wiley; 2007.
10. Clayton D, Leslie A. The bioavailability of erythromycin stearate versus enteric-coated erythromycin base when taken immediately before or after food. J Int Med Res. 1981;9:470–7.
11. Rani S, Pargal A. Bioequivalence: an overview of statistical concepts. Ind J Pharmacol. 2004;36:209–16.
12. Soukup M. Using R: perspectives of a FDA statistical RevieweR. Presentation at useR! conference, Iowa State University, August 8–10, 2007.