Research Article

Reference Datasets for Bioequivalence Trials in a Two-Group Parallel Design

Anders Fuglsang,^{1,4} Helmut Schütz,² and Detlev Labes³

Received 9 October 2014; accepted 12 November 2014; published online 9 December 2014

Abstract. In order to help companies qualify and validate the software used to evaluate bioequivalence trials with two parallel treatment groups, this work aims to define datasets with known results. This paper puts a total 11 datasets into the public domain along with proposed consensus obtained via evaluations from six different software packages (*R*, *SAS*, *WinNonlin*, *OpenOffice Calc*, *Kinetica*, *EquivTest*). Insofar as possible, datasets were evaluated with and without the assumption of equal variances for the construction of a 90% confidence interval. Not all software packages provide functionality for the assumption of unequal variances (*EquivTest*, *Kinetica*), and not all packages can handle datasets with more than 1000 subjects per group (*WinNonlin*). Where results could be obtained across all packages, one showed questionable results when datasets contained unequal group sizes (*Kinetica*). A proposal is made for the results that should be used as validation targets.

KEY WORDS: bioequivalence; parallel design; software validation.

INTRODUCTION

Bioequivalence testing is a general requirement for companies developing generic medicines, testing food effects, making formulation changes, or developing extensions to existing approved medicines where absorption rate and extent to the systemic circulation determines safety and efficacy. Throughout many countries and jurisdictions, the common way of testing for bioequivalence is to compare the pharmacokinetics of the new formulation (“test”) with that of the known formulation (“reference”). Using non-compartmental analysis, the primary metrics derived in bioequivalence studies are most often the area under the concentration time curve until the last sampling point (AUC_t) and the maximum observed concentration (C_{max}) for both the test and the reference. A confidence interval is then constructed on basis of two one-sided t tests, typically at a nominal α level of 5%. The most common designs for bioequivalence testing are the two-treatment, two-sequence,

two-period randomized crossover design and the randomized two-group parallel design. The former is considerably more common than the latter and the design of choice for active ingredients whose half-lives are not prohibitively long (1–4).

To evaluate the data obtained in a bioequivalence trial, companies must use validated software but in the absence of datasets with known results, it is difficult to actually know if the software acquired correctly performs the task it is supposed to do and therefore, it is practically impossible to validate software in-house beyond installation qualification and operational qualification. On that basis, we recently published a paper in this journal with reference datasets for the two-treatment, two-sequence, two-period crossover trials and where the datasets were evaluated with different software packages (5).

Since trials with two parallel groups are the second most common type of bioequivalence studies and published datasets with known results are scarce, the purpose of this paper is to propose reference datasets for two-group parallel trials and derive 90% confidence intervals with different statistical software packages in order to establish consensus results that can be used—together with the datasets—to qualify or validate software analyzing the outcomes from parallel group bioequivalence trials. It is outside the scope of this paper to discuss more than two groups or the other design options that exist such as replicate designs.

For validation purposes, datasets should be of varying complexity in terms of imbalance, outliers, range, heteroscedasticity, and point estimate in order to cover any situation which can reasonably be expected to occur in practice. It is not the aim of this work to validate any

Electronic supplementary material The online version of this article (doi:10.1208/s12248-014-9704-6) contains supplementary material, which is available to authorized users.

¹ Fuglsang Pharma, Hiort Lorenzens Vej 6c st. tv., 6100, Haderslev, Denmark.

² BEBAC—Consultancy Services for Bioequivalence and Bioavailability Studies, Neubaugasse 36/11, 1070, Vienna, Austria.

³ Cooperative Clinical Drug Research and Development AG, Lindenallee 70, 15366, Hoppegarten, Germany.

⁴ To whom correspondence should be addressed. (e-mail: anfu@fuglsangpharma.com)

software or to advocate for or against any specific software package.

MATERIALS AND METHODS

Datasets

The datasets used in this paper seek to include small and large data sets, outliers, unequal group sizes, and heteroscedasticity as a type of stress test. The characteristics of datasets are as follows:

P1: small dataset with balance

This is the data of the first period from Clayton and Leslie (6) with balance between treatment groups, i.e., equal sizes of the test treatment (N_T) and the reference treatment (N_R): $N_T=N_R=9$.

P2: small dataset with imbalance

Based on dataset P1, where subjects 10...14 were removed ($N_T=9$, $N_R=4$).

P3: small dataset with an “outlier”

Based on dataset P1, where the raw data entry for subject 4 has been multiplied by 100.

P4: small dataset with balance and heteroscedasticity

This is a dataset simulated on basis of a log-normal distribution, where the simulated geometric means ratio (GMR) was 0.95 with balance between treatment groups ($N_T=N_R=20$) and heteroscedasticity ($CV_T=0.25$ and $CV_R=2.5$).

P5: moderately sized dataset with small variability and homoscedasticity

This is a dataset simulated on basis of a log-normal distribution, where the simulated GMR was 1.1 and with slight imbalance ($N_T=31$, $N_R=29$) and homoscedasticity ($CV_R=CV_T=0.05$). Although a CV of 0.05 is not realistic for any drug tested in humans; variabilities at this level are often seen with orally inhaled products (e.g., delivered dose). Applicants in Europe may choose the parallel bioequivalence model for their testing. To qualify for approval on basis of such data, the actual requirement is a 90% confidence interval with a 15% equivalence margin corresponding to an acceptance range of 85.00–117.65% when a multiplicative model is used (7).

P6: moderately sized dataset with slight imbalance

This is the dataset from Bolton and Bon (8); $N_T=24$, $N_R=26$.

P7: large dataset with imbalance and heteroscedasticity

This is a dataset simulated on basis of a log-normal distribution, where the simulated GMR was 1.2 with imbalance between treatment groups ($N_T=1000$ and $N_R=200$) and heteroscedasticity ($CV_T=0.25$ and $CV_R=2.5$).

P8: large dataset with balance and homoscedasticity

This is a dataset simulated on basis of a log-normal distribution, where the simulated GMR was 1.1 and with balance between treatment groups ($N_T=N_R=1000$) and homoscedasticity ($CV_R=CV_T=0.5$).

P9: large dataset with balance and heteroscedasticity

This is a dataset simulated on basis of a log-normal distribution, where the simulated GMR was 1.1 and with balance between treatment groups ($N_T=1000$, $N_R=1,000$) and heteroscedasticity ($CV_R=0.25$, $CV_T=2.5$).

P10: large dataset with imbalance, heteroscedasticity, and an “outlier”

Based on dataset P6, where the raw data entry for subject 138 (receiving test) was multiplied by 100.

P11: large imbalanced dataset with extreme numeric range and GMR

Based on dataset P6, where the raw data entries for subjects 1...50 (test) and subjects 1001...1050 (reference) were multiplied by 100,000.

All datasets can be downloaded as tab-delimited text files as “[Electronic Supplementary Material](#)” from the journal’s homepage. The supplementary material also presents individual box-plots of the datasets.

Evaluation of Datasets

The datasets were evaluated with

- *R* (version 3.0.2, R Foundation for Scientific Computing 2013), running under Windows 8 (64 bit).
- *OpenOffice Calc* (version 4.1.0, The Apache Software Foundation 2014), running under Windows 8. A spreadsheet was set up.
- *Phoenix/WinNonlin* (64-bit version 6.4.0.768, Pharsight, A Certara Company 2014), running under Windows 7 (64 bit).
- *EquivTest/PK* (Statistical Solutions 2006), running in a 32-bit XP Virtual PC under Windows 7 (64 bit).
- *Kinetica* (version 5.0.10, Thermo Fisher Scientific 2007), running in a 32-bit XP Virtual PC under Windows 7 (64 bit).
- *SAS* (32 bit version 9.2, SAS Institute 2008), running under Windows 7 (64 bit).

The datasets were evaluated with and without the Welch correction for degrees of freedom (ν), corresponding to either assuming heteroscedasticity (unequal variances) or homoscedasticity (equal variances), respectively, of treatment groups.

In the Welch correction, the degrees of freedom are approximated as

$$\nu \approx \frac{(s_T^2/N_T + s_R^2/N_R)^2}{\frac{(s_T^2/N_T)^2}{N_T-1} + \frac{(s_R^2/N_R)^2}{N_R-1}} \quad (1)$$

where s_T^2 and s_R^2 are the variances, and N_T and N_R are the sample sizes of groups treated by the test and reference, respectively.

If no Welch correction is applied the degrees of freedom are

$$\nu = N_T + N_R - 2 \quad (2)$$

In all cases, datasets were evaluated towards construction of a 90% confidence interval (CI) around the ratio of log-transformed sample means (point estimate, PE), corresponding to a nominal α level of 5%

$$\ln(\text{limits}) = \ln(\text{PE}) \pm t_{1-\alpha, \nu} \sqrt{\text{MSE} \left(1/N_T + 1/N_R \right)} \quad (3)$$

Table I. Results obtained with the different statistical packages with Welch correction

Dataset	90% confidence interval (point estimate)			
	<i>R</i>	<i>OO Calc</i>	<i>WinNonlin</i>	<i>SAS</i>
P1	26.78–88.14 (48.58)	26.78–88.14 (48.58)	26.78–88.14 (48.58) ^a	26.78–88.14 (48.58)
P2	23.71–74.38 (41.99)	23.71–74.38 (41.99)	23.71–74.38 (41.99) ^a	23.71–74.38 (41.99)
P3	24.40–449.08 (104.67)	24.40–449.08 (104.67)	24.40–449.08 (104.67) ^a	24.40–449.08 (104.67)
P4	38.05–136.15 (71.97)	38.05–136.15 (71.97)	38.05–136.15 (71.97) ^a	38.05–136.15 (71.97)
P5	106.44–112.10 (109.23)	106.44–112.10 (109.23)	106.44–112.10 (109.23) ^a	106.44–112.10 (109.23)
P6	91.84–115.79 (103.12)	91.84–115.79 (103.12)	91.84–115.79 (103.12) ^a	91.84–115.79 (103.12)
P7	97.38–138.51 (116.14)	97.38–138.51 (116.14)	^b	97.38–138.51 (116.14)
P8	105.79–113.49 (109.57)	105.79–113.49 (109.57)	^b	105.79–113.49 (109.57)
P9	103.80–120.61 (11.89)	103.80–120.61 (11.89)	^b	103.80–120.61 (11.89)
P10	97.82–139.17 (116.68)	97.82–139.17 (116.68)	^b	97.82–139.17 (116.68)
P11	6.30–21.60 (11.67)	6.30–21.60 (11.67)	^b	6.30–21.60 (11.67)

Ninety percent confidence intervals and point estimates are given in percent of the reference rounded to two decimals as suggested by the FDA and EMA. Calculation apparently not possible in *EquivTest/PK* and *Kinetica*

^a Results obtained by the workaround (9)

^b Not applicable due to the software's limitations

where $t_{1-\alpha, \nu}$ is the one-sided critical value of Student's t distribution at ν degrees of freedom for the given α (in this paper 0.05) and MSE is the mean square error (estimated pooled variance).

R and *SAS* are script-/code-based applications. Examples of scripts used for these two software packages are uploaded as a supplementary material. They should be adapted to the user's preferences. The *OpenOffice Calc* spreadsheet is also uploaded as a supplementary material.

Phoenix/WinNonlin natively does not offer evaluations based on the Welch correction. A workaround (9,10) was

used in generating results given in Table I. The Welch correction appears to be impossible in the menu-driven software packages *EquivTest/PK* and *Kinetica*.

RESULTS

Table I presents the 90% confidence intervals with point estimates evaluated by the different software packages if equal variances are not assumed (i.e., the Welch correction is applied).

Table II presents the corresponding results if equal variances are assumed.

Table II. Results obtained with the different statistical packages without Welch correction

Dataset	90% confidence interval (point estimate)					
	<i>R</i>	<i>OO Calc</i>	<i>WinNonlin</i>	<i>EquivTest</i>	<i>Kinetica</i>	<i>SAS</i>
P1	27.15–86.94 (48.58)	27.15–86.94 (48.58)	27.15–86.94 (48.58)	27.15–86.94 (48.58)	27.15–86.94 (48.58)	27.15–86.94 (48.58)
P2	18.26–96.59 (41.99)	18.26–96.59 (41.99)	18.26–96.59 (41.99)	18.26–96.59 (41.99)	15.76–119.00 (41.99)	18.26–96.59 (41.99)
P3	26.35–415.71 (104.67)	26.35–415.71 (104.67)	26.35–415.71 (104.67)	26.35–415.71 (104.67)	26.35–415.71 (104.67)	26.35–415.71 (104.67)
P4	38.60–134.21 (71.97)	38.60–134.21 (71.97)	38.60–134.21 (71.97)	38.60–134.21 (71.97)	38.60–134.21 (71.97)	38.60–134.21 (71.97)
P5	106.44–112.10 (109.23)	106.44–112.10 (109.23)	106.44–112.10 (109.23)	106.44–112.10 (109.23)	106.39–112.14 (109.23)	106.44–112.10 (109.23)
P6	91.85–115.78 (103.12) ^a	91.85–115.78 (103.12) ^a	91.85–115.78 (103.12) ^a	91.85–115.78 (103.12) ^a	92.07–115.50 (103.12)	91.85–115.78 (103.12) ^a
P7	106.86–126.23 (116.14)	106.86–126.23 (116.14)	106.86–126.23 (116.14)	106.86–126.23 (116.14)	104.30–129.32 (116.14)	106.86–126.23 (116.14)
P8	105.79–113.49 (109.57)	105.79–113.49 (109.57)	105.79–113.49 (109.57)	105.79–113.49 (109.57)	105.79–113.49 (109.57)	105.79–113.49 (109.57)
P9	103.80–120.61 (111.89)	103.80–120.61 (111.89)	103.80–120.61 (111.89)	103.80–120.61 (111.89)	103.80–120.61 (111.89)	103.80–120.61 (111.89)
P10	107.20–126.99 (116.68)	107.20–126.99 (116.68)	107.20–126.99 (116.68)	107.20–126.99 (116.68)	104.59–130.16 (116.68)	107.20–126.99 (116.68)
P11	7.83–17.38 (11.67)	7.83–17.38 (11.67)	7.83–17.38 (11.67)	7.83–17.38 (11.67)	6.98–19.51 (11.67)	7.83–17.38 (11.67)

Ninety percent confidence intervals and point estimates are given in percent of the reference rounded to two decimals as suggested by the FDA and EMA

^a Accords with Bolton and Bon (8)

DISCUSSION

The conventional t test is fairly robust against violations of homoscedasticity but quite sensitive to unequal group sizes (11). Furthermore, preliminary testing for equality of variances is flawed and should be avoided (12). If assumptions are violated, the t test becomes liberal, i.e., the patient's risk might exceed the nominal level and an alternative (e.g., based on the Welch correction) is suggested (11,13). FDA's guidance states: "For parallel designs [...] equal variances should not be assumed" (14).

In *WinNonlin*, a somewhat cumbersome workaround allows a Welch correction for construction of the confidence interval, whereas in *EquivTest/PK* and *Kinetica*, there appears to be no option to evaluate two-group parallel designs with the Welch correction for unequal variances. Due to limitations of *WinNonlin* (fixed factors are restricted to 1000 levels), we were not able to process datasets P7–P11 with the published workaround.

Kinetica consistently arrived at confidence intervals that differed from the other packages if datasets had unequal group sizes (i.e., P2, P5–7, P10–P11). A similar phenomenon was noted in a previous publication presenting balanced and imbalanced reference datasets for two-treatment, two-sequence, two-period crossover bioequivalence trials (5). Although we do not have access to *Kinetica*'s source code, it is noted that *when group sizes are equal* (that is, when $N_T = N_R$), Eq. (3) can be simplified to

$$\ln(\text{limits}) = \ln(\text{PE}) \pm t_{1-\alpha, v} \sqrt{\frac{2\text{MSE}}{N_R}} \quad (4)$$

and if this version of the equation uncritically is applied to imbalanced datasets, then

1. The point estimate remains correct.
2. The width of the confidence interval will change. The use of Eq. (4) could therefore increase the chance of products *appearing* approvable (confidence intervals narrower) when $N_T < N_R$ and could decrease the chance of products appearing approvable (confidence intervals wider) when $N_T > N_R$.
3. The results coincide with *Kinetica*'s that we found for the investigated datasets. Equation (4) is stated in the user manual (15).

While it is outside the scope of this paper to specifically debate any specific software's general fitness for bioequivalence evaluations, the fact that different packages arrive at different results with identical datasets underscore the need for proper validation. The datasets and evaluations presented here are relatively simple; in various areas of drug development, much more sophisticated statistical models with many more numerical settings are available. Examples include mixed effect models used for longitudinal data or replicated bioequivalence trials, and survival or time-to-event trials evaluated by Kaplan-Meier derivatives, to mention but a few. To help companies qualify or validate their software for such

evaluations, we call for other authors to publish datasets of varying complexity with known results that can be confirmed by multiple statistical packages.

CONCLUSION

This paper releases 11 datasets into the public domain along with proposed consensus results in order to help companies qualify and validate their statistical software for evaluation of bioequivalence trials with two treatment groups. The datasets are evaluated with different statistical packages with and without the assumption of equal variances for the construction of the 90% confidence interval. Not all packages are able to use a Welch correction for unequal variances when analyzing the datasets and not all packages can handle the largest datasets. In addition, one package seemed to arrive at results that stand in contrast to others. We propose the results obtained with *R*, which are identical to those obtained with *SAS* (and *WinNonlin*, where possible) to be targeted by companies qualifying or validating their software. The datasets are available as supplementary material.

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. Revision 1. 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, Annex 7. 2006.
4. Health Canada, Therapeutic Products Directorate. Conduct and analysis of comparative bioavailability studies. 2012.
5. Schütz H, Labes D, Fugsang A. Reference datasets for 2-treatment, 2-sequence, 2-period bioequivalence studies. The AAPS J. 2014;16:1292–7. doi:10.1208/s12248-014-9661-0.
6. Clayton D, Leslie A. The bioavailability of erythromycin stearate versus enteric-coated erythromycin base when taken immediately before and after food. J Int Med Res. 1981;9:470–7. doi:10.1177/030006058100900608.
7. European Medicines Agency, Committee for Medicinal Products for Human Use. Guideline on the requirements for clinical documentation for orally inhaled products (OIP) including the requirements for demonstration of therapeutic equivalence between two inhaled products for use in the treatment of asthma and chronic obstructive pulmonary disease (COPD) in adults and for use in the treatment of asthma in children and adolescents. CPMP/EWP/4151/00 Rev.1. 2009.
8. Bolton S and Bon C. Statistical consideration: alternate designs and approaches for bioequivalence assessments. In: Kanfer I and Shargel L, editors. Generic Drug Product Development. Bioequivalence Issues. New York: Informa Healthcare; 2008. p. 123–141.

9. Hughes L. Adjusting for Unequal Variances in WNL Bioeq. Certara Forums, Phoenix WNL Basics, 1 July 2013 <http://www.certara.com/forums/topic/453-adjusting-for-unequal-variances-in-wnl-bioeq/> accessed 5 November 2014; free registration required
10. Anon. Phoenix WinNonlin 6.4 User's Guide, Pharsight 2014, p. 86.
11. Wang H, Chow S-C. A practical approach for comparing means of two groups without equal variance assumption. *Statist Med.* 2002;21:3137–51. doi:10.1002/sim.1238.
12. Moser BK, Stevens GR. Homogeneity of variance in the two-sample means test. *Am Stat.* 1992;46:19–21.
13. Ruxton GD. The unequal variance *t*-test is an underused alternative to Student's *t*-test and the Mann–Whitney *U* test. *Behav Ecol.* 2006;17:688–90. doi:10.1093/beheco/ark016.
14. US Food and Drug Administration. Statistical approaches establishing bioequivalence. 2001.
15. Anon. Kinetica 5.0 User Manual Revision Number 1.00. Thermo Fisher Scientific Inc. 2008. p. 667.