

APPENDIX 1: STATISTICAL ISSUES

APPENDIX: STATISTICAL ISSUES

This appendix presents the following statistical issues pertaining design and analysis of the Berlex study:

1. Validity and power of the study
 - a. Sample size considerations
 - b. Specifications of the alternative hypothesis
 - c. Adequacy of power
 - d. Conclusion
2. Analysis of adhesion data
3. Baseline and baseline-corrected analysis

1. Validity and power of the study

a. Sample size considerations:

For a given size (level of significance α) of the test, one needs 3 components to calculate the sample size for a designed experiment: the variance of the response variable to be analyzed, the specification of an alternative hypothesis and a desired power to accept the specified hypothesis when it is true.

- The variance is projected from historical data obtained from similar studies.
- The alternative hypothesis is represented by Δ , the difference from the null hypothesis that is considered scientifically meaningful.
- The power, $1 - \beta$, is generally chosen as 80% or higher.

How the sample size is calculated in a clinical trial?

In a clinical study designed to demonstrate superiority of an active treatment to a placebo, the alternative hypothesis is specified by the difference between the active treatment and placebo that is considered clinically relevant and not by the difference to be expected in the study. If the difference to be anticipated is hypothesized in calculation of the sample size, using a very large sample one can almost certainly show a statistical significance even if the observed difference is extremely small. But then what is the value of such a treatment if it does not produce a clinically relevant difference from a placebo. It is desired that the efficacy of the treatment relative to that of a placebo be both clinically meaningful and statistically significant.

What is done in a bioequivalence study?

The same principles as described above for a clinical trial should apply while designing a bioequivalence study. Only the difference is that it involves testing the following set of two null hypotheses:

$$(1) H_{01} : \frac{\mu_T}{\mu_R} \leq 0.8$$

and

$$(2) H_{02} : \frac{\mu_T}{\mu_R} \geq 1.25,$$

where μ_T and μ_R are the expected bioavailabilities for the test and reference formulations. A bioequivalence study is generally done using a crossover study. The rejection of both null hypotheses (two one-sided tests) at the 5% level implies that two formulations are bioequivalent. A power of $1 - \beta$ is required to reject these hypotheses when they are false. Equivalently, these null hypotheses can be tested using a 90% two-sided confidence interval.

The sample size calculation for a bioequivalence study requires the following three components:

- projected within-subject variance σ^2 of the log-transformed data or equivalently the coefficient of variation (CV) of the untransformed data, where

$$CV = \left[\sqrt{\exp(\sigma^2) - 1} \right] 100\%$$

- specification of the alternative hypothesis, and
- desired power

The projected CV was chosen as 25% by Berlex after examining previously conducted similar studies.

b. Specification of the alternative hypothesis

- Scientifically, the choice of alternative hypothesis for a bioequivalence study should be based on the same principles that apply to a clinical study design. One should realize that the degree of strength in bioequivalence is highest when the expected ratio $\frac{\mu_T}{\mu_R}$ is 1 and it gradually decreases as the expected ratio deviates from 1 on either side. From regulatory viewpoint, the degree of strength in bioequivalence is zero when the expected ratio ≤ 0.8 or ≥ 1.25 . It is a continuation function of the expected ratio, defined over the interval (0.8, 1.25), and attains its maximum when the expected ratio is 1. This point suggests that a small interval around the ratio of 1 should be chosen as a meaningful bioequivalence range and this concept should be utilized in choosing a specific alternative hypothesis in calculation of the sample size.
- Berlex chose the ratios 0.95 and 1.05 to define specific alternatives corresponding to the two null hypotheses. These points determined a reasonably wide interval about 1 to represent the highest points of bioequivalence intensity. In this context, Berlex regarded a difference of 5% from 1 as scientifically meaningful alternative regardless of the anticipated ratio at the completion of the study. Although it is difficult to define unique numbers for the ratio that would be agreeable to all scientists, based on sound principles of inference, statisticians use a reasonable deviation from 1 to specify the alternative hypothesis in the sample size calculation. In most cases, these numbers approximately correspond to the expected or anticipated ratios in bioequivalence studies and therefore, it seems, the scientists and statisticians have not given serious considerations to the problem of uniquely specifying the ratios that

can be used to define alternative hypotheses. This is also a reason why the sample sizes for bioequivalence studies are moderate, unless the value of CV is very large.

- As the Mylan document suggests, one can demonstrate bioequivalence of two formulations even when the true ratio is close to either of the boundaries, 0.8 or 1.25. This is true because the width of a confidence interval can be made as small as one would wish by enrolling an extremely large number of subjects in the study. For example, suppose the true ratio (expected ratio) is 1.24 and a study is designed so as to make the upper limit of a 90% confidence interval smaller than 1.25. Theoretically one can achieve this, but there are two problems with this approach. One is that the degree of strength in bioequivalence is extremely low (a large deviation from 1) and the other is that a study would be prohibitive because it would require an extremely large sample size. For example, as shown in the chart on Page 6 of the Mylan document (which is obtained from Hauschke et al¹), for an expected ratio of 1.15, 80% power and 25% CV, the study would require to enroll 110 subjects: with 2 periods it leads to an evaluation of 220 subject-periods. This is a huge investment in a crossover study for bioequivalence testing. Instead, one would prefer to do a clinical study using a parallel-group design with fewer patients to answer important questions on efficacy and safety.
- The FDA Guidance² does not specifically address this issue for the average bioequivalence. However, in Appendix C it suggests that for the studies to evaluate the population and individual BE (bioequivalence) studies the sample size should be based on simulated data. It continues saying that "the simulations should be conducted using a default situation allowing the two formulations to vary as much as 5% in the average bioavailabilities with equal variances. . ."
- The bioequivalence of two formulations is used as a surrogate criterion for therapeutic equivalence, i.e., equivalence of efficacy and safety of the two formulations. Hence, in order to have a reasonably high likelihood of therapeutic equivalence one should not consider an unacceptably wide range of the ratio $\frac{\mu_T}{\mu_R}$ as alternative hypotheses in designing a study for bioequivalence. The principle used by Berlex in choosing specific alternative hypotheses for bioequivalence sample size is consistent with practice followed in sample size calculation for therapeutic equivalence (see, for example, Dunnett and Gent 1977³, Makuch and Simon 1978⁴, Blackwelder 1982⁵, Patel and Gupta 1984⁶, and Lin 1995⁷, among others).

c. Adequacy of power

The sample size in the Berlex study was calculated with a provision of making two simultaneous comparisons. This is a multiple comparison problem and therefore requires a control of Type I error rate for testing multiple hypotheses or a control of the coverage probability of simultaneous confidence intervals. A Bonferroni procedure was suggested for multiple comparisons and consequently a 95% confidence interval was planned for testing bioequivalence of each of the pairs, Climara patch vs. Mylan patch and Climara patch vs. modified patch. At the time of designing the study it was anticipated that the

Division of Generic Drugs at FDA would require Berlex to do multiple comparisons. But when the data were analyzed, the statistician did not have a clear signal whether or not the multiple comparisons would be required. Both 95% and 90% confidence intervals were therefore presented in the Berlex study report.

If we limit to the results of conventional 90% confidence intervals, the study with 39 subjects leads to a power of at least 90%. It should be noted that a 90% power is more than adequate for bioequivalence testing. The power would be greater than 80% if 2 comparisons were made using a 95% confidence interval.

d. Conclusion:

Berlex had calculated the sample size using scientifically determined premises and the study was adequately powered to demonstrate bioequivalence. The 3-period, 6-sequence, crossover design used by Berlex is valid and adequately powered.

2. Analysis of adhesion data:

Whatever instructions subjects were given for taping a patch, the survival analysis of time to the first event (patch lift or patch fall, whichever occurred first) is valid. Cox's regression model was applied for analyzing survival data from matched pairs. Even though the adhesion was not of primary interest in this study, there is a strong statistical evidence of better adhesion with Climara patch than with Mylan patch. This additional information cannot be ignored.

3. Baseline and baseline-corrected analysis:

When an individual plasma concentration was below the limit of quantification (BLQ), it was assigned a value of 0, the lowest possible value. The BLQ was 5 pg/mL. So when a measurement is BLQ, the true, but unknown, value would be between 0 and 5 pg/mL. However, this missing value needed to be imputed. The choice of 0 as an imputed value caused the largest spread among the baseline measurements. In spite of this the variance of baseline was relatively very small. For baseline corrected C_{max}, the variance is

$$\text{var}(C_{\text{max}} - \text{baseline}) = \text{var}(C_{\text{max}}) + \text{var}(\text{baseline}) - 2\text{cov}(C_{\text{max}}, \text{baseline})$$

where var stands for the variance of a variable and cov for the covariance. The variance of baseline was about 15 pg/mL², whereas that of the C_{max} was about 12,000 pg/mL². Thus var(C_{max}-baseline) depends mostly on var(C_{max}). Furthermore, the baseline means and standard deviations (s.d.) are almost equal for all three formulations as presented on the next page. One would expect such characteristics of the distributions because of sufficiently long washout periods considered between successive treatment periods.

	Mylan Patch	Modified Patch	Climara
Mean (pg/mL)	3.4	3.2	3.4
s.d. (pg/mL)	3.8	3.8	3.9

Because of the almost identical distributions associated with three baselines the relative contribution of baseline to a confidence interval for the treatment contrast using a baseline-corrected method is expected to be extremely small. This is the reason why with and without baseline corrected methods give almost identical confidence intervals.

Further, the pairwise correlations for three baselines are high. Almost identical means and s.d.'s along with high pairwise correlations imply that the within-subject variability is very small. In this situation, multiple baselines for each formulation will not have any appreciable advantage over a single baseline. Consequently, planning three baselines for each period in a crossover design and then using their average as has been suggested by Mylan would have very little advantage over a single baseline measurement.

References:

1. Hauschke, D., Steinijs, V.W., Diletti, E., and Burke, M. 1992. Sample size determination for bioequivalence assessment using a multiplicative model. *J. Pharmacokinetics Biopharmaceutics* 20:557-561.
2. FDA Guidance: Statistical Approaches to Establishing Bioequivalence, Center for Drug and Evaluation and Research, Food and Drug Administration, January 2001.
3. Dunnett, C.W. and Gent, M. 1977. Significance testing to establish equivalence between treatments with special reference to data in the form of 2×2 tables. *Biometrics* 33:593-602.
4. Makuch, A. and Simon, R. 1978. Sample size requirements for evaluating a conservative therapy. *Cancer Treatment Report* 7:1037-1040.
5. Blackwelder, W.C. 1982. Proving the null hypothesis in clinical trials – a review. *Controlled Clinical Trials* 3:345-354.
6. Patel, H.I. and Gupta, G.D. 1984. A problem of equivalence in clinical trials. *Biometrical Journal* 26:471-474.
7. Lin, S.C. 1995. Sample size for therapeutic equivalence based on confidence interval. *Drug Information Journal* 29:45-50.

Appendix 2 – Division of Scientific Investigations Memo

MEMORANDUM

DEPARTMENT OF HEALTH AND HUMAN SERVICES
PUBLIC HEALTH SERVICE
FOOD AND DRUG ADMINISTRATION
CENTER FOR DRUG EVALUATION AND RESEARCH

DATE: August 8, 2002

FROM: Michael F. Skelly, Ph.D.
Pharmacologist

THROUGH: C. T. Viswanathan, Ph.D. CTV AM 8, 02
Associate Director - Bioequivalence
Division of Scientific Investigations (HFD-48)

SUBJECT: Review of EIRs Covering Docket 02P-0029/CP1, a Citizen
Petition concerning Climara[®] TD (estradiol
transdermal), sponsored by Berlex/3M

TO: Dale P. Conner, Pharm.D.
Director
Division of Bioequivalence (HFD-650)

At the request of HFD-650, the Division of Scientific Investigations conducted audits of the clinical and analytical portions of the following bioequivalence study.

Protocol #304100: "Bioequivalence comparison of 17 β -estradiol, estrone, and estrone sulfate from generic transdermal estradiol system and from a modified 3M transdermal system with that from Climara transdermal estradiol system for 0.1 mg/day patches applied over 7 days"

The clinical portion of the study was conducted at Bio-Kinetic Clinical Applications, Inc., in Springfield, MO. The analytical portion of the study was conducted at AAI Deutschland GmbH & Co., in Neu Ulm, Germany.

Following the inspections at Bio-Kinetic Clinical Applications (4/30-5/1/02) and AAI (4/29-5/3/02), no Form FDA-483 was issued. There were no objectionable findings from either inspection.

Conclusions:


Following the above audits, the Division of Scientific Investigations recommends that the study data from Protocol #304100 be accepted for review.

02P-0029

M. 1

Page 2 - Bioequivalence Inspection: Citizen Petition 02P0029/CP1

After you have reviewed this transmittal memo, please append it to the original Citizen Petition.


Michael F. Skelly, Ph.D.

Final Classifications:

NAI - Bio-Kinetic Clinical Applications, Springfield, MO
NAI - AAI Deutschland, Neu Ulm, Germany