Combinability of Animal Data in Relative Potency Estimations

Hideaki Uehara*1, Kazuko Satoh*2, Nagisa Komokata*2, Yohei Tokita*1, Mitsue Nishiyama*2, Maiko Iida*2, Junko Ishikawa*2, Kazuo Ogawa*2 and Masahiro Yamamoto*2

*1 Kampo Scientific Strategies Division, Tsumura & Co., Tokyo 107-8521, Japan
*2 TSUMURA Research Laboratories, Kampo Scientific Strategies Division, Tsumura & Co., Ibaraki 300-1192, Japan
e-mail: uehara.hideaki@mail.tsumura.co.jp

In this article, we propose a strategy to show the combinability of multiple animal datasets in a parallel-line assay to estimate the relative potency. The following three assumptions are made in the linear fixed-effect modeling, and we examine if any of them result in nonconformance:

a) Intrasubject parallelism (parallel dose-response for each subject),

b) Intersubject homogeneity of the slopes of the mean response (averaged across study substances),

c) Intersubject homogeneity of the differences between intercepts.

For inferences about relative potency, a) is essential, and we derived a new metrics, intrasubject parallelism criterion (ISP), via the translation of aggregated individual bioequivalence criterion stated in the regulatory guidance (Food and Drug Administration, 2001). For b) and c), we used the 95% confidence interval of the $I^2$ criterion, which is commonly used to evaluate the interstudy homogeneity in a meta-analysis (Higgins and Thompson, 2002). For choosing the thresholds, we applied the conventions used in the guideline.

The proposed procedure is demonstrated in an example analysis, and its properties are evaluated through a Monte Carlo simulation. The power of our proposed intrasubject parallelism criterion was shown to be high for designs of moderate size, but the demonstration of homogeneity via $I^2$ was rather conservative.

Key words: relative potency, parallel-line assay, individual bioequivalence, combinability.
1. Background

1.1 Concept of relative potency

The relative potency (RP), or the “potency ratio” of an unknown preparation, is a pharmacological concept that is defined as “the ratio of equipotent doses of the standard preparation and the unknown preparation under the conditions of the assay” (Chapter 5.3, European Pharmacopoeia, 2008). This is typically evaluated through biological experiments, such as a parallel-line or a parallel-curve assay. The validity of these experiments is based on the assumption that the dose-response curves of the test substance and the standard substance have similar shapes, although their values on the abscissa may differ; this is known as the similarity assumption. Conventionally, the similarity between substances has been assessed via hypothesis testing of the model parameters (e.g., Finney, 1978).

Since 2008, the United States Pharmacopeial Convention (USP) has been working on a comprehensive revision to its General Chapters on bioassay, focusing on the methods for assessing similarity and the technique for combining results from multiple assays (United States Pharmacopeial Convention, 2010a, 2010b, 2010c). The equivalence test is acknowledged throughout as an effective statistical method.

Meanwhile, refined assay designs can reveal various important aspects of dissimilarity. For example, studies using multiple subjects with multiple measurements from each are generally useful for enhancing the precision of an assay, but they may reveal the intra/intersubject heterogeneities more clearly. The apparent intersubject variation in the intercepts and/or slope in a bioassay has already provoked an interest in linear mixed-effect modeling (e.g., Djira, 2010), and this should require the redefinition of RP at the population level. However, the intrasubject dissimilarity (e.g., the right-hand side of Figure 1) has a fundamental impact on the rationale of experiments, because it distorts the concept of RP at the subject level, regardless of the apparent

![Intrasubject Parallelism vs Population (only) Parallelism](image)

**Fig. 1.** Similarity and Parallelism
similarity between substances in terms of the average response curves. Unfortunately, there seems to be no established criteria for the intrasubject similarity, and none is proposed in the current revised drafts of the USP General Chapters. Considering its fundamental importance in the evaluation of RP, there is a need for a practical strategy to establish intrasubject similarity.

1.2 Motivating example

To give a practical sense of the issue, in Table 1 we display the preliminary results from a pilot ex vivo parallel-line assay. This experiment was designed to evaluate the RP of substance B (test) against that of substance A (control), using 12 animals in total. For the administration of each substance, we employed a common dose sequence of 5 levels around the initial estimate of (geometrically averaged) EC\(_{50}\) for control substance (A). Each level was separated from its adjacent by a common ratio (approximately 2\(^{1/2}\)). At sacrifice, 10 specimens are extracted from each animal and are randomly allocated to one of the 10 planned treatments (the dose levels 1 to 5 for each of the test and the control substances). Then the treatments are administered in a random order, and 10 responses under 10 different conditions are acquired per block (subject animal). In each row of Table 1, we show the simple linear regression coefficients calculated per substance for each block separately, as the summaries of intrasubject dose-response relationship. If we calculate \([\text{intercept difference}] / [\text{average slope}]\) for each row, we obtain a crude estimate of log-RP for each subject; the results are also shown in the rightmost column of Table 1. By averaging these per-subject point estimates, we obtained an overall estimate of log-RP, whose anti-log transformed value was 0.74 (95% confidence interval based on a \(t\)-test: 0.65 to 0.85).

Thus, we may conclude that the potency of substance B is 74% that of substance A.

However, here we face at least two questions:

[1] Is it appropriate to use the RP in this case? In other words, was it appropriate to assume the parallelism per subject? Note that if, for any subject, the two dose-response curves

| Subject ID | Intercept A | Intercept B | Slope A | Slope B | Crude log_{10}RP |
|------------|-------------|-------------|---------|---------|-----------------|
| 1          | 1.6291      | 1.4666      | 1.3295  | 1.6805  | -0.1080         |
| 2          | 1.5258      | 1.3366      | 1.0310  | 0.8320  | -0.2031         |
| 3          | 1.4634      | 1.4545      | 1.3744  | 1.0134  | -0.0074         |
| 4          | 1.6213      | 1.5026      | 0.6978  | 1.3223  | -0.1176         |
| 5          | 1.5407      | 1.4886      | 0.8168  | 0.4167  | -0.0845         |
| 6          | 1.5422      | 1.2405      | 1.2856  | 1.0333  | -0.2602         |
| 7          | 1.6500      | 1.4020      | 1.1511  | 1.2683  | -0.2047         |
| 8          | 1.4929      | 1.5198      | 1.2289  | 1.2268  | -0.0213         |
| 9          | 1.5362      | 1.4085      | 1.3444  | 0.8636  | -0.1157         |
| 10         | 1.5738      | 1.5283      | 1.4973  | 0.9119  | -0.0377         |
| 11         | 1.7261      | 1.5260      | 1.0376  | 1.0195  | -0.1946         |
| 12         | 1.7761      | 1.5735      | 0.7487  | 0.9411  | -0.2397         |
cross, there the log-RP estimate does not make sense, and thus the estimate obtained by averaging all subjects becomes meaningless.

Even given the appropriateness of the assumption of intrasubject parallelism, it is still not certain that the RP is homogeneous among the subjects. If it is not, further elaboration of the statistical model is warranted in order to augment the efficiency of the inference and thereby reduce the number of animals to be sacrificed.

The first of these questions the validity of the RP as a metric for this situation, and the second questions the details of the statistical model.

1.3 Individual bioequivalence criterion

Although developed for different circumstances, we will refer to the individual bioequivalence criterion (Food and Drug Administration, 2001) as a prototype of our proposed intrasubject parallelism criterion, which will be presented below in Section 2.2.1. The individual bioequivalence criterion (IBC) was developed to evaluate the switchability between different formulations of the same drug, in terms of appropriate bioavailability measures (Anderson, 1993; Hauck and Anderson, 1994), e.g., the natural logarithm of the area under the curve (AUC) or the peak concentration (C\text{max}).

Let us denote the bioavailability of a standard (test) formulation that is to be observed in subject $i$ at the $j$-th administration as $Y_{R,i,j}$ ($Y_{T,i,j}$), and let us denote the subject-specific mean and intrasubject variance as $\mu_{R,i}$ ($\mu_{T,i}$) and $\sigma^2_{WR}$ ($\sigma^2_{WT}$), respectively. The switchability between the two formulations may be measured by

$$E\left\{ (Y_{R,i,j} - Y_{T,i,j})^2 - (Y_{R,i,j'} - Y_{R,i,j''})^2 \right\} = (\mu_T - \mu_R)^2 + Var(\mu_{T,i} - \mu_{R,i}) + \sigma^2_{WT} - \sigma^2_{WR},$$

(Ref. I)

where $\mu_R = E(\mu_{R,i})$ and $\mu_T = E(\mu_{T,i})$, i.e., the population means. This is the expected increase in the mean squared difference between two administrations, due to altering the formulation.

The IBC criterion of the Food and Drug Administration (FDA) is given by scaling the aforementioned expectation (Ref. I) by the intrasubject variance under reference formulation $\sigma^2_{WR}$, or the prespecified constant $\sigma^2_{W0}$ (FDA, 2001):

Reference scaled: $\frac{(\mu_T - \mu_R)^2 + \sigma^2_D + \sigma^2_{WT} - \sigma^2_{WR}}{\sigma^2_{WR}}$ (if $\sigma^2_{WR} > \sigma^2_{W0}$),

Constant scaled: $\frac{(\mu_T - \mu_R)^2 + \sigma^2_D + \sigma^2_{WT} - \sigma^2_{WR}}{\sigma^2_{W0}}$ (if $\sigma^2_{WR} \leq \sigma^2_{W0}$)

(Ref. II)

where $\mu_R = E(\mu_{R,i})$, $\mu_T = E(\mu_{T,i})$, and $\sigma^2_{D} = Var(\mu_{T,i} - \mu_{R,i})$.

The test drug is judged to have individual bioequivalence (IBE) when the 95% upper confidence limit of the IBC is less than a specified tolerable limit, $\Theta$. If the standard formulation had a very small intrasubject variance, the Reference scaled criterion is automatically inflated and the tolerable limit cannot be determined, therefore they avoided this difficulty by introducing changeover to the constant scaled criterion.
To choose $\Theta$, the FDA designated the upper limit of $|\mu_T - \mu_R|$ as $\ln(1.25)$, which is the conventional limit for testing the bioequivalence on average. Also they recommended that $\sigma^2_D$ should be less than 0.03 so that for at least 80% of the subjects, the inter-formulation difference, $\mu_{T,i} - \mu_{R,i}$, would be within the interval $(\ln(0.8), \ln(1.25))$ when $\mu_R = \mu_T$. Combining the constant-scaled criterion with $\sigma^2_W = 0.04$, they obtained the recommendation that $\sigma^2_{WT} - \sigma^2_{WR}$ should be less than 0.02, resulting $\Theta = 2.4948$. For detail, see the Appendix A of the aforementioned guidance (FDA, 2001).

2. Method

In this section, we present the details of our proposed criteria and the corresponding analysis methods. After specifying the notation in Section 2.1, we develop the details of the procedures in Section 2.2. The translation of the IBC into an intrasubject parallelism criterion is discussed in Section 2.2.1, and this is followed by a discussion about the choice of its acceptance threshold. The analysis method for complete datasets is given in Section 2.2.1.2, and the method when there are missing data is given in Section 2.2.1.3. Other methods for homogeneity evaluation are discussed in Section 2.2.2.

2.1 General notation for a parallel-line assay

Hereinafter, we will consider an ex vivo randomized-block experiment, in which $n$ animal subjects are used as the experimental blocks, to evaluate the test substance against the control. As the general basis for the statistical analysis, we will consider the following linear model:

$$Y_{i,j,k} = D_i + j \cdot a_i + (b_i + j \cdot d_i) \cdot X_{i,j,k} + \varepsilon_{i,j,k}, \quad \varepsilon_{i,j,k} \sim N(0, \sigma^2_{\varepsilon}), \quad i.i.d.,$$

where

- $i$: subject identifier ($i = 1, \ldots, n$),
- $j$: substance identifier (control : $j = 0$; test : $j = 1$),
- $k$: dose level ($k = 1, \ldots, K_{i,j}$),
- $D_i$: expected response of subject $i$ at the mean dose of the control treatment,
- $a_i$: effect of the test substance (intercept difference) in subject $i$,
- $b_i$: slope of the log-dose response in subject $i$ under the control treatment,
- $d_i$: difference in slope between substances in subject $i$,
- $X_{i,j,k}$: centralized log-dose $X_{i,j,k} = W_{i,j,k} - \bar{W}_{i,j} \cdot X_k = 1 K_{i,j} K_{i,j} X_k = 1 W_{i,j,k}$.

Here, $W_{i,j,k}$ is the log-dose of substance $j$ at dose level $k$ that is actually administered to subject $i$. Suppose that $W_{i,1,k}$ and $W_{i,0,k}$ are mutually equivalent in terms of the expected response, and that the parallelism can be assumed regarding subject $i$ ($d_i = 0$). Then the model (1) gives the subject specific log-RP (see Section 1.1) as $\log \text{RP}_i = W_{i,0,k} - W_{i,1,k} = Jpn J Biomet Vol. 37, No. 1, 2016
estimate of the regression parameter vector calculated for subject \( K \) order are given as

\[
0 = \hat{b}_i = (\hat{W}_{i,\bullet} - \hat{W}_{i,0\bullet}).
\]

Here the last term is the difference between study substances regarding the averaged experimental log-dose. Without losing generality, we focus on the situation where the experiment is designed as \( \hat{W}_{i,\bullet} = \hat{W}_{i,\bullet} \).

After sorting the within-subject measurements in order of substance and dose level, the response vector \( Y_i = (Y_{i,0,1}, \ldots, Y_{i,0,K_0}, Y_{i,1,1}, \ldots, Y_{i,1,K_1})' = (Y_{i,0}', Y_{i,1}')' \) can be expressed as

\[
Y_i = X_i b_i + \varepsilon_i \quad \text{where} \quad X_i \left( \begin{array}{c}
X_{i,0} \\
X_{i,1}
\end{array} \right) = \left( \begin{array}{c}
j_{K_i,0} \\
j_{K_i,1}
\end{array} \right) \left( \begin{array}{c}
x_{i,0} \\
x_{i,1}
\end{array} \right).
\]

\[
b_i = \left( \begin{array}{c}
D_i \\
a_i \\
b_i \\
d_i
\end{array} \right), \quad x_{i,0} = \left( \begin{array}{c}
X_{i,0,1} \\
\vdots \\
X_{i,0,K_i,0}
\end{array} \right), \quad x_{i,1} = \left( \begin{array}{c}
X_{i,1,1} \\
\vdots \\
X_{i,1,K_i,1}
\end{array} \right)
\]

\[
\varepsilon_i = \left( \begin{array}{c}
\varepsilon_{i,0,1} \\
\vdots \\
\varepsilon_{i,0,K_i,0} \\
\varepsilon_{i,1,1} \\
\vdots \\
\varepsilon_{i,1,K_i,1}
\end{array} \right) \sim N \left( 0_{K_i,0+K_i,1}, \quad \text{diag} \left( \sigma_0^2 J_{K_i,0}, 0_{K_i,0,K_i,1} \right) \right).
\]

Here, \( 0_K \) and \( j_K \) are column vectors of length \( K \) of 0s and 1s, \( I_K \) is the identity matrix of order \( K \), and \( 0_{K,L} \) is a \( K \) by \( L \) matrix in which each of the elements is zero. The least-squares estimate of the regression parameter vector calculated for subject \( i \), and its sampling distribution, are given as

\[
\hat{b}_i = (X_i'X_i)^{-1}X_iY_i \quad \text{and} \quad \hat{b}_i|_{b_i} \sim N \left( b_i, \left( \begin{array}{c}
A_i \\
0_2
\end{array} \right) \right)
\]

\[
\text{where} \quad A_i = \left( \begin{array}{cc}
\sigma_0^2 & -\sigma_0^2 \\
\sigma_0^2 & \sigma_0^2
\end{array} \right) \quad \text{and} \quad A_i = \left( \begin{array}{cc}
\sigma_0^2 / x_{i,0} x_{i,0} & -\sigma_0^2 / x_{i,0} x_{i,0} \\
-\sigma_0^2 / x_{i,0} x_{i,0} & \sigma_0^2 / x_{i,0} x_{i,0}
\end{array} \right).
\]

Usually, the dose sequence is standardized per study substance by the protocol of the experiment \( (x_{i,0} = x_0 \text{ and } x_{i,1} = x_1, \text{ say}) \). If so, then

\[
A_i = A = \left( \begin{array}{cc}
\sigma_0^2 / x_0 x_0 & -\sigma_0^2 / x_0 x_0 \\
-\sigma_0^2 / x_0 x_0 & \sigma_0^2 / x_0 x_0
\end{array} \right) \quad \text{and} \quad B_i = B = \left( \begin{array}{cc}
\sigma_0^2 / x_0 x_0 & -\sigma_0^2 / x_0 x_0 \\
-\sigma_0^2 / x_0 x_0 & \sigma_0^2 / x_0 x_0
\end{array} \right).
\]

Here \( K_0 \) and \( K_1 \) are the lengths of \( x_0 \) and \( x_1 \).

In the following analysis, we will consider this standardized situation. Also we will assume
that \((b_i, d_i)\) are bivariate normally distributed:
\[
\begin{pmatrix}
  b_i \\
  d_i
\end{pmatrix} \sim N\left(\begin{pmatrix}
  \beta \\
  \delta
\end{pmatrix}, \begin{pmatrix}
  \sigma^2_b & \sigma_{bd} \\
  \sigma_{bd} & \sigma^2_d
\end{pmatrix}\right).
\] (5)

### 2.2 Proposed criteria and procedure

To address the issues posed above in Section 1.2, it seems relevant to review the framework of a parallel-line assay. We will begin by considering the classic fixed-effect modeling approach, which requires all of the following three conditions:

a) Intrasubject parallelism: \(d_i = 0\),
b) A common average slope: \(b_i + d_i/2 = b\), and
c) A common difference between intercepts: \(a_i = a\).

The RP defined by the population parameter becomes identical to the one given at the subject level only under an ideal situation in which all three conditions are fulfilled for all of the subjects, \(i = 1, \ldots, n\).

Since condition a) is the foundation of the RP concept, we believe that this should be examined at the start of any parallel-line assay (see Section 2.2.1).

Once intrasubject parallelism has been demonstrated, it is relevant to assess the other two conditions and choose an appropriate (possibly mixed-effect) model from among the possible candidates (Table 2; see Section 2.2.2).

When the data cast doubt on the assumptions of commonality in b) or c), it is also relevant to model the fluctuations among subjects in terms of neither \(a_i\) nor \(b_i\), but of \(\text{RP}_i\). In that case, a nonlinear mixed-effect model with a random RP may be considered.

If intrasubject parallelism is refuted, we may decide to abandon the use of the RP as the metric for the pharmacological activity of the test substance, or we may be able to use the population parameters to revise the definition of the RP, provided that this is supported by pharmacological evidence. Then, log(RP) may be estimated from the ratio of the population parameters, if we can demonstrate that the population lines are parallel.

Below, we will discuss the metrics for each of these three criteria.

#### 2.2.1 Intrasubject parallelism: Translated individual bioequivalence criterion

When the IBC was being developed, there was a focus on the following metric for the switchability to a new formulation of a drug (see Section 1.3):

| Common intercept difference | Varying intercept difference |
|-----------------------------|-----------------------------|
| Common Slope               | Linear Model (common \(a\) and \(b\)) | LME (random \(a\), common \(b\)) |
| Varying Slope              | Common RP (\(\exp(a/b)\)) | Random RP \(\exp(a/b)\) |

Table 2. Candidate models under intrasubject parallelism
\[ E \{ (Y_{R,i,j} - Y_{T,i,j'})^2 - (Y_{R,i,j''} - Y_{R,i,j'''})^2 \} = (\mu_T - \mu_R)^2 + Var(\mu_{T,i} - \mu_{R,i}) + \sigma^2_{WR} - \sigma^2_{WR}. \]

(Ref. I)

Note that this metric evaluates the switchability by examining the intrasubject fluctuation of bioavailability through repeated administrations of the drug. This is important in clinical practice in order to avoid the risk of excessive or insufficient internal exposure i.e. over or under dose by chance due to the intrasubject fluctuation of bioavailability, which may be observed as the intervisit variations in clinical pharmacokinetic studies. On the other hand in ex vivo experiments, the specimens are extracted from sacrificed animals, therefore the dose-response relationship of each animal can never be observed repeatedly; in this case, the intrasubject parallelism (or lack of it) at the time of sacrifice should be analyzed as the key property of each subject. Thus the metric defined in (Ref. I) should be simplified to
\[ E \{ (\mu_{R,i} - \mu_{T,i})^2 \} = (\mu_T - \mu_R)^2 + Var(\mu_{T,i} - \mu_{R,i}) \] in our situation. Then, according to the model given from equations (1) through (5), we newly obtain the following metrics, intrasubject parallelism criterion (ISP):
\[ ISP = \frac{(E(d_i))^2 + Var(d_i)}{Var(\tilde{b}_i|\tilde{b}_i)} = \frac{\delta^2 + \sigma^2_d}{\sigma^2_0/x'_0x_0}. \]

(6)

Note that the denominator of the ISP is determined by the common dose sequence for the control treatment. Thus, this criterion is appropriate when the experimental doses are standardized by the study protocol.

The test drug is judged to have intrasubject parallelism relative to the control when the 95% upper confidence limit of the ISP is less than the tolerable upper limit, \( \Theta \). In practice, the corresponding linearized criterion
\[ \eta = \delta^2 + \sigma^2_d - \Theta \cdot \sigma^2_0/x'_0x_0 \] is used for testing. A negative value for the upper confidence limit of \( \eta \) can be interpreted as evidence of intrasubject parallelism.

### 2.2.1.1 Choosing the tolerable upper limit for the ISP

The choice of the tolerable upper limit \( \Theta \) for the ISP is straightforward if a bioassay has already been established with abundant data. We can find the reference value for the error variance (\( \sigma^2_0 \), say) and for the control slope (\( \beta_c \)) by simply applying the model given by equation (1), for each subject in the historical data. As the tolerable limit for \( |d_i| \), a proportion (\( \pi \)) of reference slope (\( \Delta = \pi \cdot |\beta_c| \)) may be considered. In the same way as was done for the IBC, the tolerable upper limit of the ISP in equation (6) may be chosen as \( \Theta = \{ \Delta^2 + (\Delta/Z_{1-\gamma/2})^2 \}/(\sigma^2_0/x'_0x_0) \), where \( Z_p \) is the \( p \)-quantile of the standard Gaussian distribution, and \( \gamma \) is the prespecified probability that \( |d_i| \) is greater than \( \Delta \) when \( E(d_i) = 0 \). We may choose a value for \( \pi \) between \( \pi_0 = 0.3 \) and \( \pi_1 = 0.5 \), referring to USP-1032 (see Subsection 4.7 “Sample Suitability”, United States Pharmacopeial Convention, 2010a), and \( \gamma \approx 0.2 \) (see Section B.2, Appendix A, FDA, 2001).

However, we need to be careful in the use of this approach, because the null hypothesis thus given corresponds to an extreme situation, where the mean and variance are both on the margins of the tolerable values. In particular, the ISP can determine intrasubject parallelism, even if one of the two parameters exceeds the tolerable limit that was determined when choosing \( \Theta \).
2.2.1.2 ISP evaluation under a homogeneous design without missing data

If the dose sequence is standardized per study substance by the protocol of the experiment \((x_{i0} = x_0 \text{ and } x_{i1} = x_1)\), and the dataset is complete, the moment estimator of the linearized criterion \((\eta = \delta^2 + \sigma_\theta^2 - \Theta \cdot \sigma_0^2 / x_0^2)\) is given as

\[
\hat{\eta}_C = \bar{d}^2 + (1 - 1/n) s_d^2 - (1 + \Theta) \cdot s_2^2 / x_0^2 - s_1^2 / x_1^2 \quad \text{where} \quad \bar{d} = \frac{\sum d_i}{n},
\]

\[
s_d^2 = \sum_{i=1}^n (d_i - \bar{d})^2 / (n - 1) \quad \text{and} \quad \sigma_\theta^2 \sim \frac{\sigma^2}{n(K_\theta - 2)}, \chi^2_{n(K_\theta - 2)} (j = 0, 1),
\]

\[
s_1^2 = \sum_{i=1}^n Y_{i,j} (I_{K_j} - X_{i,j} X_{i,j}^{'})^{-1} X_{i,j} Y_{i,j} / \{n \cdot (K_j - 2)\} \quad (j = 0, 1).
\]

Based on the fact that \(s_d^2 \sim \frac{\text{Var}(d_i)}{n - 1} \cdot \chi^2_{n-1} \) and \(s_2^2 \sim \frac{\sigma_\theta^2}{n(K_j - 2)} \cdot \chi^2_{n(K_j - 2)} (j = 0, 1)\), as well as their mutual independence, the upper 95% confidence limit on \(\eta\) can be obtained via the method of variance estimates recovery (MOVER; Zou and Donner, 2008; see Appendix 1):

\[
\hat{\eta}_C + \sqrt{U_D + U_I + U_{R0} + U_{R1}},
\]

where \(U_D = \left(\frac{[\bar{d}^t + t_{n-1,1-\alpha} \cdot \sqrt{s_d^2 / n} - \bar{d}^2]_2}{s_d^4} \right)\), \(U_I = \left(\frac{n - 1}{n} \right)^2 \cdot s_4 \cdot \left(\frac{n - 1}{\chi^2_{n-1,1-\alpha}} - 1\right)\), and \(U_{R0} = \left(\frac{1 + \Theta}{X_{i0}^{'} X_0} \right)^2 \cdot s_5 \cdot \left(\frac{n \cdot (K_\theta - 2)}{\chi^2_{n(K_\theta - 2),1-\alpha}} - 1\right)\), and \(U_{R1} = \left(\frac{1}{X_{i1}^{'} X_1} \right)^2 \cdot s_4 \cdot \left(\frac{n \cdot (K_1 - 2)}{\chi^2_{n(K_1 - 2),1-\alpha}} - 1\right)\).

2.2.1.3 ISP evaluation under a homogeneous design with missing data

Unfortunately, even when the study is well designed, loss of data can happen unexpectedly, and as a consequence, we must analyze unbalanced data. When all the measurements were performed concurrently (e.g., at the sacrifice), data can be assumed to be “missing completely at random” (Little and Rubin, 2002), and we can handle the data as if taken from a study that was designed to be unbalanced. In this case, we may approximate the distribution of \(s_d^2\) in equation (7) by applying a method for an unbalanced 1-way random effect model (Thomas and Hultquist, 1978; see Appendix 2):

\[
\sigma_\theta^2 \sim \frac{\sigma^2}{n(K_\theta - 2)} \cdot \chi^2_{n(K_\theta - 2)}, \quad \text{where} \quad \hat{K}_j = \frac{1}{n} \sum_{i=1}^n K_{i,j} (j = 0, 1), \quad \text{the moment estimator of the linearized criterion is revised and becomes}
\]

\[
\hat{\eta}_U = \bar{d}^2 + (1 - 1/n) s_d^2 - s_0^2 \left(\frac{1}{n} \cdot \sum_{i=1}^n \frac{1}{x_{i,0}^2} + \frac{\Theta}{x_0^2} \right) - s_1^2 \left(\frac{1}{n} \cdot \sum_{i=1}^n \frac{1}{x_{i,1}^2}\right).
\]
As an approximation of its upper 95% confidence limit on \( \eta \), we use

\[
\hat{\eta}_U + \sqrt{U_D + U_I + U_{R0}^2 + U_{R1}^2},
\]

where

\[
U_{R0}^2 = \left( \frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i \theta_0 x_0} \right)^2 \cdot s_0 \cdot \left( \frac{n \cdot (K_{00} - 2)}{\chi^2_{n-(K_{00}-2),1-\alpha}} - 1 \right)^2
\]

and

\[
U_{R1}^2 = \left( \frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i x_1} \right)^2 \cdot s_1 \cdot \left( \frac{n \cdot (K_{1} - 2)}{\chi^2_{n-(K_{1}-2),1-\alpha}} - 1 \right)^2.
\]

Note that the unbalanced nature of this design jeopardizes the independence between \( \bar{d} \) and \( s_2^2 \), and therefore, in this case, the prerequisite for the MOVER method is not strictly fulfilled. For any given dataset, the appropriateness of this method will vary according to the pattern or rate of missing data, so it should be examined via Monte Carlo simulation before use.

### 2.2.2 Second and third criteria: homogeneity of mean slope and difference in intercepts

To evaluate the intersubject homogeneity of these two measures, we consider the \( I^2 \) criterion (Higgins and Thompson, 2002), a heterogeneity metric that is widely used in meta-analyses. It is defined as \( I^2 = \tau^2 / (\tau^2 + \sigma^2) \), where \( \tau^2 \) is the between-study heterogeneity, and \( \sigma^2 \) is the sampling variance of the estimates from each study. Various reference values have been tentatively assigned adjectives: \( I^2 < 30\% \) (mild) and \( I^2 > 50\% \) (notable), from Higgins and Thompson (2002); \( I^2 \approx 25\% \) (low), \( I^2 \approx 50\% \) (moderate), and \( I^2 \approx 75\% \) (high), from Higgins et al. (2003). However, undue reliance on these criteria has been criticized, because \( I^2 \) is dependent on the precision of the studies being analyzed (Rücker et al., 2008).

To use this method in our situation, we substitute the subject animals for the studies in the original \( I^2 \) metrics. To choose the reference values, we again refer to the parameters given by the historical data. Following the previous section, we choose \( \tau_0^2 = (\pi_0 \cdot \beta / Z_{1-\gamma_0/2})^2 \) as the reference value for negligible heterogeneity in terms of the mean slope. Furthermore, we choose \( \tau_1^2 = (\pi_1 \cdot \beta / Z_{1-\gamma_1/2})^2 \) as the reference value for crucial heterogeneity. As in 2.2.1.1, we use \( \pi_0 = 0.3 \) and \( \pi_1 = 0.5 \). We use \( \gamma_0 = \gamma_1 = 0.05 \) as the convention. Under the model given in equation (1), we have \( \hat{\phi}_3^2 = \tau_0^2 / (\tau_0^2 + s^2) \) and \( \hat{\phi}_2^2 = \tau_1^2 / (\tau_1^2 + s^2) \) as the \( I^2 \) reference values. Here, if the protocol specifies the common dose sequence, then \( s^2 \) in the denominator is given as \( \text{Var}(\bar{b}_i + \hat{d}_i/2|\bar{b}_i) \) from equation (4). Otherwise, we use equation (9) in Higgins and Thompson (2002), which is

\[
s^2 = \frac{(n-1) \cdot \sum_{i=1}^{n} w_i}{\left( \sum_{i=1}^{n} w_i \right)^2 - \sum_{i=1}^{n} w_i^2},
\]

where \( w_i = \text{Var}(\bar{b}_i + \hat{d}_i/2|\bar{b}_i) \).

We also use these thresholds as reference values for the intercept difference. At the design phase of a study, the intrasubject sampling variance (the second term in the denominator of
$I^2$ is often estimated with homoscedasticity assumed, unless the historical data inform us of heterogeneity in the test substance. We believe these thresholds are useful for facilitating the interpretation of heterogeneity, at least from the perspective of the study designer.

3. Analysis of Example Data

We applied our proposed procedures to the example data shown in Section 1.2. Because there were no missing data, we can apply the formulae given in Section 2.2.1.2. The computations were performed using R version 3.2.2 (R Core Team, 2015), using the lmList function from the nlme package (Pinheiro et al., 2015).

Figure 2 shows the normal Q-Q plot of per-subject residuals after fitting the model given by equation (1). There seems to be no large discrepancy from normality; some subjects exhibited slightly smaller variance (e.g., subjects 4, 11, and 12). From Table 1, the average of the control

![Fig. 2. Normal Q-Q plot of residuals in the per-subject estimation](image-url)
slope estimates was 1.1286 (95% confidence interval: 0.9612 to 1.2960), which was larger than $\beta_C = 0.9$, the estimate from historical data. The 95% confidence interval of the mean slope difference was $-0.3149$ to $0.1459$. Other statistics were $s_0^2 = 0.0131$, $s_1^2 = 0.0215$, $\bar{d} = -0.0845$, and $s_d^2 = 0.1315$.

Using $\beta_C = 0.9$, $\sigma_0^2 = 0.01$ (which are estimates from past experiments), $\pi = 0.5$, and $\gamma = 0.2$, the ISP threshold $\Theta$ is 7.3809. The mean slope difference and its 95% confidence interval was $-0.3149$ to $0.1459$. Other statistics were $s_0^2 = 0.0131$, $s_1^2 = 0.0215$, $\bar{d} = -0.0845$, and $s_d^2 = 0.1315$.

In the analysis below, we will assume intrasubject parallelism, which is supported by the data to some extent. The crude RP estimate 0.74 (0.65 to 0.85) given in Section 1.2 is now interpretable. Below, we will examine the need to use a mixed-effect model by examining the $I^2$ statistics.

With $\gamma_1 = \gamma_0 = 0.05$, the heterogeneity reference values are $\tau_0^2 = 0.0190$ and $\tau_1^2 = 0.0527$, which gives the $I^2$ reference values of $i_0^2 = 0.4623$ ("negligible") and $i_1^2 = 0.7049$ ("crucial"). The $I^2$ statistics of intrasubject intercept difference ($I^2_a$) and average slope ($I^2_b$) were 0.3083 (95% confidence interval: 0.0000 to 0.6508) and 0.2414 (0.0000 to 0.6136). In both cases, the heterogeneity seemed mild, at most.

Because of the wide confidence intervals obtained for the $I^2$ statistics, the validity of using the classic parallel-line assay procedure is uncertain. Therefore, we fitted a nonlinear mixed-effect model that assumes a common slope ($b_i = b$) and a bivariate normal distribution for $D_i$ and log(RP$_i$). The result demonstrated that about 8% of the interblock coefficient of variation (CV) can be explained by the variation in RP$_i$, and the estimated population geometric mean of the RP was 0.75 (95% confidence interval: 0.67 to 0.84). This was close to the estimate given by the classic parallel-line assay procedure (i.e., the fixed-effect parallel-line analysis): RP = 0.75 (Fieller’s 95% confidence interval: 0.67 to 0.83). From these results, we can conclude that in this case, the difference between the fixed- and mixed-effect models appears to be negligible, and the integrated analysis shortened the confidence interval of RP by 15% to 20%.

4. Simulation

To examine the operational characteristics of our proposed procedures, we conducted a Monte Carlo simulation of a five-dose parallel-line assay with 10,000 iterations. We chose a design similar to that discussed above, with a common dose ratio of $2^{1/2}$ and the following conditions:

- Centralized log-concentration: $x = \log_{10} 2 \cdot (-1, -1/2, 0, 1/2, 1)'$ (x'x ≈ 0.2265),
- Control slope: $\beta = 1$,
- Error Variance: $\sigma_0^2 = \sigma_1^2 = \sigma^2 = 0.01$,
- Tolerable slope difference: $\Delta = 0.3, 0.5$ ($\pi = 0.3, 0.5$)
Tolerable risks: \( \gamma = 0.2 \) and \( \gamma_0 = \gamma_1 = 0.05 \),
Number of subjects: \( n = 12, 24, \) or \( 48 \),
Variance components:
\[
\begin{align*}
    a_i & \sim N(0, \sigma_a^2) \quad \text{where} \quad \sigma_a^2 \in \left\{ 0, \frac{\sigma^2}{K}, 2 \cdot \frac{\sigma^2}{K} \right\}, \\
    b_i & \sim N(\beta, \sigma_b^2) \quad \text{where} \quad \sigma_b^2 \in \left\{ 0, \frac{\sigma^2}{4 \cdot x'x}, \frac{\sigma^2}{2 \cdot x'x} \right\}, \\
    d_i & \sim N(\delta, \sigma_d^2) \quad \text{where} \quad \sigma_d^2 \in \left\{ 0, \left( \frac{0.3 \cdot \beta}{Z_{1-\gamma/2}} \right)^2, \left( \frac{0.5 \cdot \beta}{Z_{1-\gamma/2}} \right)^2 \right\}.
\end{align*}
\]

Note that the three levels of \( \sigma_a^2 \) correspond to \( I^2_a \) values of 0, 0.33, and 0.50, and when \( \sigma_d^2 = 0 \), the three levels of \( \sigma_b^2 \) correspond to \( I^2_b \) values of 0, 0.33, and 0.50. The computations were performed again by using R version 3.2.2 (R Core Team, 2015), using the \texttt{lmList} function from the \texttt{nlme} package (Pinheiro et al., 2015).

The ISP testing size is summarized in Table 3. The type 1 error rates were well maintained at the nominal value, which is the size of testing at the extreme case (see the comment at the end of Section 2.2.1).

The simulated statistical powers of ISP are summarized in Table 4a (for \( \Delta = 0.3 \)) and Table 4b (for \( \Delta = 0.5 \)). The ISP retained its power, even when the variance of \( d_i \) and the mean difference in slope exceeded negligible values (for example \( \delta = 0.3 \) and \( \sigma_d = 0.3/Z_{1-\gamma/2} \) = 0.2341 in Table 4b, where \( \Delta = 0.5 \) was used to choose \( \Theta \)). As stated above, this is because the null hypothesis of ISP is doubly extreme (i.e., the average and the variance of intrasubject slope difference both reach their limits). Therefore, when the ISP is found to be significant, it only indicates the unlikeliness of this combined extremity. Even when one of the two parameters (the mean or the variance of the intrasubject slope difference) is within the tolerable range and the other is not, the ISP may give significant results for a sufficiently large sample size.

Figure 3 shows the cumulative distribution of the upper limit of the 95% confidence interval of \( I^2_a \). The continuous, dashed, and dotted lines correspond, respectively, to the results under zero, mild (\( I^2 = 0.33 \)), and moderate (\( I^2 = 0.50 \)) heterogeneities in the intercept difference. When there was no heterogeneity, about 40% of the results fell within the region of negligible heterogeneity (left of the vertical dotted line) when \( n = 12 \), around 80% fell within that region when \( n = 24 \), and more than 90% when \( n = 48 \). When the intersubject heterogeneity was mild or moderate, these values fell to less than 50%.

Figure 4 shows the cumulative distribution of the upper limit of the 95% confidence interval of \( I^2_b \); see Figure 3 for the meaning of the various lines.

The upper panel summarizes the results obtained when the intrasubject parallelism was held intact, and the lower panel summarizes the results when it was broken by random fluctuations.
Table 3.  Empirical size of ISP*

| $\sigma_d^2$ | $\sigma_b^2$ | n   | $\Delta$  |
|------------|-------------|-----|-----------|
| **          | ***         | 0.3 | 0.5       |
| 0           | 0           | 12  | 0.0517    | 0.0475 |
| 0           | 0           | 24  | 0.0522    | 0.0542 |
| 0           | 0           | 48  | 0.0513    | 0.0500 |
| 0           | 1           | 12  | 0.0487    | 0.0508 |
| 0           | 1           | 24  | 0.0468    | 0.0475 |
| 0           | 1           | 48  | 0.0518    | 0.0551 |
| 0           | 2           | 12  | 0.0469    | 0.0544 |
| 0           | 2           | 24  | 0.0523    | 0.0541 |
| 0           | 2           | 48  | 0.0512    | 0.0530 |
| 1           | 0           | 12  | 0.0525    | 0.0527 |
| 1           | 0           | 24  | 0.0527    | 0.0531 |
| 1           | 0           | 48  | 0.0524    | 0.0524 |
| 1           | 1           | 12  | 0.0489    | 0.0545 |
| 1           | 1           | 24  | 0.0519    | 0.0552 |
| 1           | 1           | 48  | 0.0516    | 0.0532 |
| 1           | 2           | 12  | 0.0477    | 0.0507 |
| 1           | 2           | 24  | 0.0531    | 0.0541 |
| 1           | 2           | 48  | 0.0551    | 0.0508 |
| 2           | 0           | 12  | 0.0498    | 0.0493 |
| 2           | 0           | 24  | 0.0508    | 0.0542 |
| 2           | 0           | 48  | 0.0490    | 0.0514 |
| 2           | 1           | 12  | 0.0501    | 0.0493 |
| 2           | 1           | 24  | 0.0508    | 0.0521 |
| 2           | 1           | 48  | 0.0514    | 0.0546 |
| 2           | 2           | 12  | 0.0517    | 0.0533 |
| 2           | 2           | 24  | 0.0491    | 0.0561 |
| 2           | 2           | 48  | 0.0526    | 0.0490 |

* $\Theta = \{\Delta^2 + (\Delta/\tilde{Z}_{1-\gamma/2})^2\}/(\sigma_d^2/\chi'\chi)$, $\delta = \Delta, \sigma_d = \frac{\Delta}{\tilde{Z}_{1-\gamma/2}}$

** 0: $\sigma_d^2 = 0$, 1: $\sigma_d^2 = \sigma^2/K$, 2: $\sigma_d^2 = 2\cdot\sigma^2/K$

*** 0: $\sigma_b^2 = 0$, 1: $\sigma_b^2 = \sigma^2/(4\cdot\chi'\chi)$, 2: $\sigma_b^2 = \sigma^2/(2\cdot\chi'\chi)$

($\sigma_d = 0.3$). The former results resemble those shown in Figure 4, but the latter are shifted to the right. More than 60% of the results were outside of the negligible region (left of the vertical dashed line close to the center), even when the control slope had a fixed common value. Thus, the deviation from intrasubject parallelism is also reflected in the heterogeneity of the average slope.

5. Discussion

Conventional bioassay methods are inadequate for assessing the similarity of dose-response curves at the subject level. Indeed, intrasubject parallelism is often evaluated by significance
Table 4a. Power of ISP with $\Delta=0.3^*$

| $\delta$ | $n$ | $\sigma^2$ | $\Theta = \frac{\Delta}{\sigma_0^2}$ | $\sigma_\nu^2$ | $\sigma_\psi^2$ |
|----------|-----|------------|-----------------|-------|--------|
| 0.0      | 12  | 0.5843     | 0.0336          | 0.2081| 0.1510 |
| 0.0      | 24  | 0.8903     | 0.0330          | 0.3734| 0.2335 |
| 0.0      | 48  | 0.9963     | 0.0355          | 0.6526| 0.3835 |
| 0.0      | 12  | 0.5809     | 0.0373          | 0.2068| 0.1552 |
| 0.0      | 48  | 0.9970     | 0.0333          | 0.6525| 0.3906 |
| 0.0      | 12  | 0.5773     | 0.0397          | 0.2146| 0.1517 |
| 0.0      | 24  | 0.8938     | 0.0374          | 0.3787| 0.2323 |
| 0.0      | 48  | 0.9963     | 0.0332          | 0.6535| 0.3763 |
| 1.0      | 12  | 0.5796     | 0.0331          | 0.2034| 0.1532 |
| 1.0      | 24  | 0.8863     | 0.0363          | 0.3835| 0.2343 |
| 1.0      | 48  | 0.9964     | 0.0351          | 0.6498| 0.3697 |
| 1.0      | 12  | 0.5752     | 0.0359          | 0.2105| 0.1536 |
| 1.0      | 24  | 0.8843     | 0.0388          | 0.3838| 0.2352 |
| 1.0      | 48  | 0.9964     | 0.0353          | 0.6534| 0.3711 |
| 2.0      | 12  | 0.5764     | 0.0357          | 0.2095| 0.1502 |
| 2.0      | 24  | 0.8889     | 0.0381          | 0.3805| 0.2390 |
| 2.0      | 48  | 0.9969     | 0.0345          | 0.6525| 0.3796 |

* $\Theta = \frac{(\Delta^2 + (\Delta/\sqrt{\gamma})^2)}{(\sigma_0^2/\phi^2)}$
Table 4b. Power of ISP with $\Delta = 0.5^\ast$

| $\delta$ = 0.0 | $\delta$ = 0.3 | $\delta$ = 0.5 |
|----------------|----------------|----------------|
| $\sigma_0^2$  | $\sigma_1^2$  | n              | $\sigma_0^2$  | $\sigma_1^2$  | n              | $\sigma_0^2$  | $\sigma_1^2$  | n              |
| 0 0 12 0.9715 0.7771 0.3539 0.7792 0.4821 0.1802 0.2878 0.1427 |
| 0 0 24 0.9997 0.9800 0.6395 0.9691 0.7815 0.3056 0.4509 0.2105 |
| 0 0 48 1.0000 0.9998 0.9134 0.9998 0.9705 0.5350 0.6950 0.3154 |
| 0 1 12 0.9663 0.7704 0.3492 0.7861 0.4824 0.1875 0.2785 0.1423 |
| 0 1 24 0.9998 0.9792 0.6392 0.9744 0.7705 0.3139 0.4540 0.2132 |
| 0 1 48 1.0000 0.9999 0.9118 0.9998 0.9688 0.5244 0.7034 0.3216 |
| 0 2 12 0.9706 0.7692 0.3588 0.7853 0.4827 0.1859 0.2741 0.1353 |
| 0 2 24 1.0000 0.9972 0.6358 0.9726 0.7706 0.3177 0.4491 0.2026 |
| 0 2 48 1.0000 1.0000 0.9161 0.9999 0.9697 0.5324 0.6953 0.3215 |
| 1 0 12 0.9701 0.7639 0.3472 0.7810 0.4846 0.1888 0.2817 0.1476 |
| 1 0 24 0.9998 0.9800 0.6412 0.9726 0.7732 0.3170 0.4628 0.2058 |
| 1 0 48 1.0000 0.9999 0.9105 0.9999 0.9711 0.5228 0.6993 0.3165 |
| 1 1 12 0.9709 0.7717 0.3596 0.7817 0.4877 0.1798 0.2793 0.1357 |
| 1 1 24 0.9999 0.9804 0.6346 0.9742 0.7735 0.3118 0.4445 0.2063 |
| 1 1 48 1.0000 0.9999 0.9140 0.9999 0.9665 0.5294 0.6940 0.3211 |
| 1 2 12 0.9720 0.7764 0.3494 0.7781 0.4806 0.1885 0.2722 0.1421 |
| 1 2 24 1.0000 0.9799 0.6360 0.9725 0.7748 0.3128 0.4524 0.2106 |
| 1 2 48 1.0000 0.9997 0.9083 1.0000 0.9720 0.5286 0.7037 0.3163 |
| 2 0 12 0.9704 0.7726 0.3587 0.7836 0.4874 0.1884 0.2779 0.1447 |
| 2 0 24 0.9998 0.9785 0.6408 0.9755 0.7773 0.3042 0.4618 0.2052 |
| 2 0 48 1.0000 1.0000 0.9162 0.9999 0.9673 0.5211 0.7072 0.3275 |
| 2 1 12 0.9684 0.7682 0.3512 0.7810 0.4829 0.1860 0.2861 0.1428 |
| 2 1 24 1.0000 0.9771 0.6459 0.9713 0.7684 0.3090 0.4588 0.2005 |
| 2 1 48 1.0000 1.0000 0.9076 0.9997 0.9699 0.5230 0.6982 0.3144 |
| 2 2 12 0.9710 0.7730 0.3536 0.7790 0.4861 0.1857 0.2821 0.1358 |
| 2 2 24 0.9999 0.9803 0.6393 0.9737 0.7723 0.3116 0.4563 0.2069 |
| 2 2 48 1.0000 0.9999 0.9137 0.9997 0.9699 0.5278 0.7044 0.3204 |

$^\ast \Theta = (\Delta^2 + (\Delta/Z_{1-\gamma}/2)^2)/(\sigma_0^2/\mathbf{x}'\mathbf{x})$

a better understanding of the data, one may consider combining it with other disaggregated measures, e.g., testing the equivalence of the average slopes. Indeed, the FDA’s population bioequivalence (PBE) and IBE guidelines mandate the evaluation of the average bioequivalence (see the final paragraph in Section IV.C, FDA, 2001). If we employ this approach, we need to investigate the operation characteristics of the combined procedure as well.

We also considered an evaluation of the intersubject heterogeneities of intercept difference and average slope within subject. Note that they may reflect the between-subject fluctuations in RP, given intrasubject parallelism. We employed the commonly used $I^2$ statistics for the
metrics, primarily because of their ease of interpretation. Because of the conservative nature of their confidence intervals, they are unlikely to exclude the necessity of using a mixed-effect model. The thresholds for these evaluations were chosen to mimic the FDA’s PBE/IBE guidelines, but we changed the $\gamma$ parameter from 0.2 to 0.05; this was done arbitrarily, and the results may suggest that there is room for improvement.

The proposed methods employ several assumptions, and, needless to say, their appropri-
ateness for a given dataset must be examined before they are applied. In equation (1), we assumed that the errors were normally distributed and that there was homoscedasticity among the animals and concentration levels of each substance. However, this can be uncertain when the assessment of a test substance is still at an early stage, and so it is important to further examine the robustness of these methods. We expect that the series of articles by Novick et al. (2009a, 2009b, 2009c) will provide a good model for this kind of evaluation.

As an alternative to the FDA’s IBE, there have been several attempts to apply the tolerance interval approach (Brown, Iyer, and Wang, 1997; Esinhart and Chinchilli, 1994a, 1994b; also see Zhang et al., 2009, and Section 6.5 in Krishnamoorthy and Mathew, 2009). Similarly, we may also consider the development of tolerance interval criterion for the intrasubject parallelism. This seems suitable for a confirmatory analysis because of the ease of interpretation; therefore, we are now developing this as an alternative to the ISP.

We have primarily followed the approach used in the FDA PBE/IBE guidelines when choosing thresholds. As an alternative, we may also consider the tolerance interval of the key parameter estimated from the historical database. In this approach, the quantitative sufficiency of the accumulated historical data needs to be examined.

From a practical point of view, when designing a study, the first thing we may want is an established formula for estimating the sample size. However, it may be more productive to consider a strategy for integrated analysis of multiple datasets from different studies with possibly different dose levels. To address the between-assay differences, a two-way nested model (possibly an unbalanced) will be employed (the subject animals are nested within each study). Although the ISP was developed for homogeneous designs, the aforementioned alternative tolerance interval approach may be expanded to adapt to this development. In the future, to improve the precision of the estimates, we should also consider the use of a measurement level covariate, since this may help to reduce the number of subjects required for the assay.

Our proposed criteria are based on conventional statistical methods that are readily available for other uses; thus, their implementation is straightforward. They may also be used as general metrics of the assay quality in randomized-block experiments that are similarly designed. Development of our methods for other types of assays (e.g., parallel-curve assay) is also warranted. We believe that all these ideas can offer important perspectives for future developments of methods for combining animal data.

Acknowledgements

We thank the two referees for their productive comments.
Combinability of Animal Data in Relative Potency Estimations

REFERENCES

Anderson S. (1993). Individual Bioequivalence: A Problem of Switchability (with discussion). *Biopharmaceutical Reports* 2, 1–11.

Brown E.B., Iyer H.K. and Wang C.M. (1997). Tolerance intervals for assessing individual bioequivalence. *Statistics in Medicine* 16, 803–820.

Djira G.D. (2010). Relative potency estimation in parallel-line assays — method comparison and some extensions. *Communications in Statistics - Theory and Methods* 39, 1180–1189.

Esinhart J.D. and Chinchilli V.M. (1994a). Sample Size Considerations for Assessing Individual Bioequivalence Based on the Method of Tolerance Interval. *International Journal of Clinical Pharmacology and Therapeutics* 32, 26–32.

Esinhart J.D. and Chinchilli V.M. (1994b). Extension to the Use of Tolerance Intervals for the Assessment of Individual Bioequivalence. *Journal of Biopharmaceutical Statistics* 4, 39–52.

Finney D.J. (1978). *Statistical Methods in Biological Assay*, 3rd ed. Charles Griffin and Company Ltd., England.

Food and Drug Administration (2001). *Guidance for Industry: Statistical Approaches to Establishing Bioequivalence*.

Graybill F. and Wang C.M. (1980). Confidence Intervals on Nonnegative Linear Combinations of Variances. *Journal of the American Statistical Association* 75, 869–873.

Hauck W.W. and Anderson S. (1994). Measuring Switchability and Prescribability: When is Average Bioequivalence Sufficient? *Journal of Pharmacokinetics and Biopharmaceutics* 22, 551–564.

Higgins J.P. and Thompson S.G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 21, 1539–1558.

Higgins J.P., Thompson S.G, Deeks J.J. and Altman D.G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal* 327, 557–560.

Howe W.G. (1974). An Approximate Confidence limits on the mean of X+Y where X and Y are two tabled independent random variables. *Journal of the American Statistical Association* 69, 789–794.

Hyslop T., Hsuan F. and Holder D.J. (2000). A small sample confidence interval approach to assess individual bioequivalence. *Statistics in Medicine* 19, 2885–2897.

Krishnamoorthy K. and Mathew T. (2009). *Statistical Tolerance Region: Theory, Applications, and Computation*. Wiley.

Little and Rubin (2002). *Statistical Analysis with Missing Data*. 2nd ed. Wiley, New Jersey.

Jpn J Biomet Vol.37, No.1, 2016
Appendices

A. MOVER method

Suppose we have a set of target parameters $\theta_1, \ldots, \theta_g$, accompanied by mutually independent estimators $\hat{\theta}_1, \ldots, \hat{\theta}_g$ and the corresponding confidence intervals (at a common confidence level), $(l_1, u_1), \ldots, (l_g, u_g)$. Zou and Donner (2008) proposed the following formulae for calculating the confidence interval $(L, U)$ for a linear combination of target parameters $\sum_{i=1}^{g} c_i \theta_i$:

$$L = \sum_{i=1}^{g} c_i \hat{\theta}_i - \sqrt{\sum_{i=1}^{g} c_i^2 (\hat{\theta}_i - l_i^*)^2} \quad \text{with} \quad l_i^* = \begin{cases} l_i & \text{if } c_i > 0, \\ u_i & \text{if } c_i < 0, \end{cases}$$

and

$$U = \sum_{i=1}^{g} c_i \hat{\theta}_i + \sqrt{\sum_{i=1}^{g} c_i^2 (\hat{\theta}_i - l_i^*)^2} \quad \text{with} \quad u_i^* = \begin{cases} u_i & \text{if } c_i > 0, \\ l_i & \text{if } c_i < 0, \end{cases}$$

\( (A.1) \)
Combinability of Animal Data in Relative Potency Estimations

\[ U = \sum_{i=1}^{g} c_i \hat{\theta}_i + \sqrt{\sum_{i=1}^{g} c_i^2 (\hat{\theta}_i - u_i^*)^2} \text{ with } u_i^* = \begin{cases} u_i & \text{if } c_i > 0, \\ I_i & \text{if } c_i < 0, \end{cases} \]

See also Graybill and Wang (1980) and Ting et al. (1990).

B. Methods for the unbalanced 1-way random effect model

Let us consider the unbalanced 1-way experiment, where \( n_i \) responses are measured in the \( i \)-th block, whose effect is regarded as random. Denote the \( j \)-th observation from block \( i \) as \( Y_{i,j} \), which is modeled as

\[ Y_{i,j} = \mu + \tau_i + \epsilon_{i,j}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, n \]

where \( \tau_i \sim N(0, \sigma^2_\tau) \) and \( \epsilon_{i,j} \sim N(0, \sigma^2_e) \).

Define \( \tilde{n} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{n_i}, \tilde{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{i,j}, \tilde{Y} = \frac{1}{n} \sum_{i=1}^{n} \tilde{Y}_i, \ SS_Y = \sum_{i=1}^{n} (\tilde{Y}_i - \tilde{Y})^2, \ SS_e = \sum_{i=1}^{n} \sum_{j=1}^{n_i} (Y_{i,j} - \tilde{Y}_i)^2 \) and \( N = \sum_{i=1}^{n} n_i \). Thomas and Hultquist (1978) proved that the distribution of \( SS_Y / (\sigma^2_\tau + \tilde{n}\sigma^2_e) \) can be approximated by a chi-square distribution with \( n - 1 \) degrees of freedom. Note that, in general, \( SS_Y \) and \( \tilde{Y} \) are correlated.