The Effect of Variability and Carryover on Average Bioequivalence Assessment: A Simulation Study Comparing Bioequivalence Confidence Intervals

MA. PILAR SÁNCHEZ O1, JORDI OCAÑA2, JOSEP L. CARRASCO3

1PhD. Student, Department of Statistics, Faculty of Biology, University of Barcelona, Barcelona, Spain
2Department of Statistics, Faculty of Biology, University of Barcelona, Barcelona, Spain
3Department of Public Health, Faculty of Medicine, University of Barcelona, Barcelona, Spain

* To whom correspondence should be addressed. (e-mail: jocana@ub.edu, correspondence to: Jordi Ocaña, Dep. d’Estadística. Facultat de Biologia. Universitat de Barcelona. Av. Diagonal, 645, 08028 Barcelona, Spain. Tel. +34 93402156, Fax +34 934111733.

Key words: Average bioequivalence, Confidence interval inclusion, Carryover, Crossover trial simulation

Abstract
The purpose of this work is to compare different methods of constructing bioequivalence confidence interval under 2×2 crossover designs.
The intervals under consideration were the standard “shortest”, the Westlake’s symmetric interval and two intervals described in Hsu et al., 1994 (“symmetric” and “bioequivalence specific”), that may improve the first two to a certain extent, in order to assess average bioequivalence (ABE). These intervals were evaluated by simulation under different combinations of formulation effect, sample size, carryover and simulated data variability, generating data under a 2×2 crossover design. All of them perform similarly with respect to ABE declaration, but differ in estimation accuracy. High levels of variability considerably distort the properties of all methods, including their true type I and type II error probabilities in bioequivalence studies, while the presence of carryover is less distorting and depends on the variability. We end up with some hints concerning the controversy about pretesting for carryover before performing a bioequivalence study.

1. Introduction

This paper should be considered as an appendix to Sanchez et al. (2009). The notation, the meaning of the formulae and the main concepts are based on this reference; the reader is addressed to it. While Sanchez et al. (2009) is an attempt to quantify the effect of variability and carryover on ABE studies based on the interval inclusion principle and the use of the standard “shortest” confidence interval, the aim of the present study is to review some inferential methods in ABE, all of them based on the use of confidence intervals, and to make a comparative study quantifying how their statistical properties are affected by variability and carryover. The methods under consideration were: the
“classic” shortest confidence interval (for completeness), the symmetric confidence interval proposed by Westlake (1976) and two confidence intervals described in Hsu et al. (1994). These last two intervals may represent an improvement over the standard shortest and Westlake’s intervals, but are not widely used in practice, nor taken into account in many reviews.

The work is structured as follows: In section 2 some alternative confidence intervals for ABE assessment are summarized. Section 3 describes a simulation study comparing them, especially with respect to their robustness in front to high data variability and carryover. The paper ends with a discussion in section 4.

2. Bioequivalence Confidence Intervals

In practice, the most common procedure to assess bioequivalence is based on the “confidence interval inclusion principle”, say, to declare ABE if the usual $1-2\alpha$ shortest confidence interval for the formulation effect, $\phi$,

$$I = \overline{D} \pm t_{(\alpha,N-2)} \hat{se}_\phi$$  \hspace{1cm} (1)

is fully included in the bioequivalence limits $\pm \phi_0$ (most frequently $\pm 0.223$), where $t_{(\alpha,N-2)}$ is the $1-\alpha$ quantile of a Student’s $t$ distribution with $N - 2$ degrees of freedom. $\overline{D}$ stands for the formulation effect estimator based on the intrasubject differences, $\hat{se}_\phi$ for its standard error estimator and $N = n_1 + n_2$ for the total sample size in a $2 \times 2$ crossover design where $n_j$ subjects are allocated to sequence $j = 1, 2$ (see details in Sanchez et al., 2009). This procedure defines a test of size $\alpha$.

Note that in the preceding statement a $1-2\alpha$ interval (1) is associated to a test of size $\alpha$. Munk and Pflüger (1999) (in more general terms as is stated here) show that this relation between confidence and test size is associated to two conditions: convexity of the parametric region associated with the alternative hypothesis and equivariance of the confidence interval, as follows: if $I_{1-2\alpha}$ is for a $1-2\alpha$ confidence interval for $\phi$, then $I_{1-2\alpha}(d_\phi(\overline{D}), \hat{se}_\phi) = 2\phi - I_{1-2\alpha}(\overline{D}, \hat{se}_\phi)$ with respect to the transformation $d_\phi(x) = 2\phi - x$. This equivariance condition is fulfilled by (1) but relaxing this requirement in other confidence intervals may lead to $1-\alpha$ confidence intervals associated to $\alpha$ size tests of bioequivalence, according to the confidence interval inclusion rule. This is the case for the three confidence intervals described below.

Westlake (1976) introduced a confidence interval that is symmetric around zero:
\[ I_w = \left[ \bar{D} - t_2 \hat{\text{se}}_\sigma, \quad \bar{D} - t_1 \hat{\text{se}}_\sigma \right] \quad (2) \]

where \( t_1 \) and \( t_2 \) must satisfy the equations

\[
\Pr \{ t_1 < T < t_2 \} = 1 - \alpha
\]

\[
(t_1 + t_2) \hat{\text{se}}_\sigma = 2 \bar{D}
\]

which must be solved by a trial-and-error procedure. In (3) \( T \) stands for a Student’s \( t \) random variable with \( N - 2 \) degrees of freedom.

When the true formulation effect \( \phi \) is zero, the true coverage of \( I_w \) is one, and it tends to the nominal \( 1 - \alpha \) as \( \phi \) tends to infinity. When \( p = \alpha \) (not \( p = 2\alpha \)) the interval inclusion rule provides a test of size \( \alpha \).

Hsu et al. (1994) proposed two intervals with confidence level \( 1 - \alpha \), the symmetrical interval:

\[
I_S = \pm \left( \left| \bar{D} \right| + t_{(\alpha, N-2)} \hat{\text{se}}_\sigma \right) \quad (4)
\]

and the “bioequivalence specific” interval:

\[
I_* = \left[ \min \left( 0, \bar{D} - t_{(\alpha, N-2)} \hat{\text{se}}_\sigma \right), \quad \max \left( 0, \bar{D} + t_{(\alpha, N-2)} \hat{\text{se}}_\sigma \right) \right] \quad (5)
\]

Both confidence intervals, (4) and (5), have asymptotic confidence level \( 1 - \alpha \), and 100% coverage if \( \phi = 0 \). There is an inclusion relation \( (5) \subset (4) \subset (2) \). Thus, from (2) to (5) they are susceptible of providing an improvement in power. The properties of the above intervals, and their relation to \( \alpha \) level tests, are summarized in Chow and Shao (2002).

3. Simulation Study

The simulation study described in Sanchez et al. (2009) for the shortest confidence interval was also performed for each one of the remaining confidence intervals described in the previous section.

**Simulations under varying formulation effects (\( \phi \)) and no carryover effect**

We first focus on the case of no carryover, \( \kappa = 0 \), assuming a well-planned experiment with an adequate washout period.

Figure 1 illustrates the percentage of ABE approval for growing degrees of variability (from left to right, Figure 1a to Figure 1c). The sample size effect is illustrated inside each figure: dashed lines correspond to \( n = 12 \) and continuous lines to \( n = 24 \). The different confidence interval types are represented by different plotting symbols, but the performance in declaring ABE of all intervals is so similar that their
power curves are nearly indistinguishable. Only for \( n = 12 \) and \( CV = 30\% \), the Westlake interval seems to perform slightly worst than the others.

As can be expected from theory, all intervals define correct 5% level tests with an acceptable level of power for small or medium variabilities. An increase in the sample size from \( n = 12 \) to \( n = 24 \) increases the percentage of declaration of ABE under all coefficients de variation. Growing variabilities (for the same \( n \) and \( \phi \)) have a strong impact on power for all tests, with a clear decrease in the percentage of ABE declaration. With \( n = 12 \), when \( CV = 30\% \) and the formulation effect is around \( \phi = 1.1 \) (considerable similarity between formulations) it is unlikely to declare ABE even for very similar drugs.

With respect to coverage, the results in Figure 2 confirm the theoretical predictions. The coverage of the shortest interval \( I \) remains almost constant and closer to its nominal 90% value (independently of \( \phi \) and sample size). \( I_W, I_S \) and \( I^* \), have 100% coverage when \( \phi = 0 \), which quickly approaches 95% as \( \phi \) increases.

Figure 3 displays the results for the mean interval length. As is expected \( I \) shows the lowest value and its length remains constant as \( \phi \) increases, while the remaining intervals are wider and have increasing length. The symmetric intervals \( I_W \) and \( I_S \) behave similarly, while \( I^* \) is clearly shorter than \( I_W \) and \( I_S \) and its length increases more slowly.

All intervals, including \( I \), show a similar behavior with respect to effective length (Figure 4). This last magnitude always grows with increasing \( \phi \) values. According to Westlake’s point of view with respect to effective length, all intervals seem to be equivalent.

**Simulations under non-null carryover effects**

Figure 5 plots the percentage of ABE declaration in function of the relative amount of carryover, \( \kappa/\phi \). Negative values of this ratio correspond to carryover and formulation effects of different sign, while positive values correspond to the case where both effects have the same sign.

Figure 5 is itself an array of plots. Each row in the array (labeled “a” to “c”) corresponds to a true amount of formulation effect and each column (labeled “1” and “2”) correspond to a sample size, 12 or 24. Inside each plot, each line represents an amount of data variability, according to the previously stated pattern \( CV = 10\%, 20\% \) and 30%. 
The first row, Figure 5a.1 and Figure 5a.2, illustrates a case of clear bioequivalence. Ideally, the percentage of ABE declaration should be kept constant near 100%, independently of any disturbing effect like carryover or high variability. This is certainly the case for a low variability, CV = 10%, and \( n = 24 \), were the power line remains virtually parallel to the axis of abscises. Under the same low variability, for \( n = 12 \) there is a slight fall of power for increasing negative degrees of relative carryover. For growing variabilities, the effect of carryover becomes evident. The lines acquire a positive slope, which translates in a progressive fall of power for negative ratios \( \kappa / \phi \) and an increase of power for positive ratios. But in any case, with respect to ABE declaration, the effect of variability seems more decisive than the effect of carryover.

The second row in Figure 5 illustrates a case of true formulation effect on the equivalence limit. Ideally, the percentage of ABE declaration should lie in a 5% value. This is certainly not the case. Negative ratios \( \kappa / \phi \) imply over-conservative testing procedures, while positive ratios imply non-valid testing procedures, with a true test size exceeding the nominal 5%. The test size inflation is perceptibly high for relative carryovers exceeding 0.1. In correspondence with the low level of power when variability is high and sample size is small, now the overoptimistic nature of the test is potentiated by low variabilities and large sample sizes (the probability of declaring ABE overly exceeds 0.05) and is attenuated by high variabilities and small sample sizes.

The third row in Figure 5 illustrates a case of clear non equivalence. The percentage of ABE declaration should never be over the 5% line, but this undesirable tendency may be observed for high degrees of positive relative carryover (over 0.25), and more markedly for low variabilities and large sample sizes.

The paradoxical effect on type I error of increased sample sizes (and small variabilities), in the sense of increasing its probability in presence of carryover, is coherent with the carryover incidence on the intervals coverage and length. Figure 6, with the same general structure than Figure 5, plots the coverage of the shortest \( I \) interval in function of the same factors discussed above. For this interval, carryover has the effect of symmetrically decreasing coverage, in a non negligible degree outside the limits \(-0.1 < \kappa / \phi < +0.1\). The incidence of carryover on coverage is more pronounced in the case of low variabilities and for large sample sizes. This is in concordance with the behavior of the mean interval length (Figure 7) which remains constant with respect to the ratio \( \kappa / \phi \). For the shortest \( I \) interval, the interval length only depends on the
standard error of the formulation effect estimation. This standard error estimation decays with growing sample sizes and decreasing data variabilities. These intervals are always centered on the formulation effect estimator which is more and more biased as the relative importance of carryover increases.

Figure 8 to Figure 10 show the same results (percentage of ABE declaration, coverage, mean length and mean effective length) for Westlake’s interval $I_W$. Similarly, Figure 11 to Figure 13 correspond to $I_S$ and Figure 14 to Figure 16 correspond to $I_\ast$. With respect to ABE declaration, all these intervals perform very similarly, with no noticeable differences with respect to the usual interval $I$. With respect to coverage, all perform partially better than $I$ in the sense of not being affected by negative ratios $\kappa/\phi$ (that is, by carryovers of opposite sign with respect to formulation effect) but being equally affected by positive ratios $\kappa/\phi$ (as is $I$). This partially better behavior with respect to coverage is associated to the much larger mean lengths of $I_W$ and $I_S$ (with respect to $I$) and to the moderately larger length of $I_\ast$.

Finally, all intervals share a very similar pattern with respect to effective length, which linearly decays in the full range of $\kappa/\phi$ ratios. This is illustrated in Figure 17 to Figure 20.

4. Discussion

All the simulation results in absence of carryover confirm the known facts about the bioequivalence methods under study and serve, mainly, as a simulation validation. There are no notable differences between interval types in ABE testing. There is no appreciable gain when BE testing is performed with a more sophisticated confidence interval in place of the straightforward standard shortest interval, and in particular these intervals do not introduce any improvement in ABE testing for high variability drugs.

The confidence intervals under study present some differences with respect to their behavior as estimation methods for the formulation effect. The symmetric intervals $I_W$ and $I_S$ are markedly lengthier (that is, more imprecise) than $I$, and their length quickly grows with growing values of the formulation effect. Admittedly, their nominal confidence level is higher (95% in front of the 90% level of the shortest interval in order to always have a 5% test) but the 95% shortest interval is still more precise, as is shown in Figure 3. The convenience of symmetrizing around zero may be justifiable in terms of criteria for bioequivalence declaration (Westlake, 1976) though this is a controversial
issue (Kirkwood, 1981, Mantel and Westlake, 1977, Wellek, 2003) but in any case it does not seem to have any practical interest in confidence interval estimation. The non-symmetrical interval \( I_* \) is more adequate. Its length also increases with growing \( \phi \) values, but slowly than \( I_W \) and \( I_S \). It may be even more precise than the 95% shortest interval for low values of the true formulation effect \( \phi \).

On the other hand, all intervals share a very similar pattern of effective length, in correspondence with their similarity as bioequivalence testing devices.

As bioequivalence testing procedures, all intervals are affected in the same way and degree by carryover. It has some incidence in the true type I and type II error probabilities, but this incidence mostly depends on the experiment variability and on the sign of the relative carryover, expressed as the ratio \( \kappa / \phi \). Negative ratios induce a progressive decay in power—that is growing type II probabilities. Positive ratios induce an apparent growth of power (e.g. Figure 5.a for \( I \) or Figure 8.a for \( I_W \)) but in correspondence with a growth of the type I error probability (e.g. Figure 5.b or Figure 8.b) which produces invalid testing procedures, no longer respecting the nominal test size.

With respect to coverage, there are obvious differences between confidence interval types—or more precisely, between \( I \) and the remaining intervals. \( I_W \), \( I_S \) and \( I_* \) partially improve \( I \) in the sense that all of them do not present any decay in coverage for negative ratios \( \kappa / \phi \). Again, this is a consequence of the kind of bias associated to \( \phi \) estimation: by construction, intervals \( I_W \), \( I_S \) and \( I_* \) always include zero. Their coverage partial stability in front to carryover is achieved at the expense of a growing length, less pronounced for \( I_* \).

In any case, carryover has some impact on all bioequivalence procedures, though modulated by variability and other factors. Using a more sophisticated confidence interval (instead of the usual “shortest” interval) does not provide any guard in front of the possible disturbing presence of carryover.

References

Sanchez O., M.P., Ocaña, J., Carrasco, J.L. The Effect of Variability and Carryover on Average Bioequivalence Assessment: A Simulation Study. *Submitted*


Figure 1

Percentage of ABE approval vs. formulation effect with CV=10, 20, 30% and $n_1=n_2=12, 24$. The thick horizontal line represents the nominal 5% test size.
Figure 2  Coverage vs. formulation effect with CV=10, 20, 30% and $n_1 = n_2 = 12$, 24. The thick horizontal line represents the nominal 90% coverage.
Figure 3a  
CV = 10%

Figure 3b  
CV = 20%

Figure 3c  
CV = 30%

Figure 3 Interval length vs. formulation effect with CV=10, 20, 30% and $n_1=n_2=12, 24$. 
Figure 4 
Effective length vs. formulation effect with CV=10, 20, 30% and \( n_1=n_2=12, 24 \).
Figure 5: % of ABE declaration in function of $\kappa/\phi$ according to CV, for some $\phi$ and sample sizes, for the 1 conf. interval.

- CV=10
- CV=20
- CV=30

$n_1=n_2=12$

$n_1=n_2=24$

Formulation Effect $= \phi = 0.095$

Formulation Effect $= \phi = 0.223$

Formulation Effect $= \phi = 0.262$
Figure 6: % Coverage in function of $\kappa/\phi$ according to CV, for some $\phi$ and sample sizes, for the $I$ conf. interval.

$n_1=n_2=12$

**Figure 6a.1**

$n_1=n_2=24$

**Figure 6a.2**

Formulation Effect $= \phi = 0.095$

Formulation Effect $= \phi = 0.223$

Formulation Effect $= \phi = 0.262$
Figure 7: Interval Length in function of $\kappa / \phi$ according to CV, for some $\phi$ and sample sizes, for the $I$ conf. interval.

$n_1=n_2=12$

$n_1=n_2=24$

Figure 7a.1

Figure 7a.2

Figure 7b.1

Figure 7b.2

Figure 7c.1

Figure 7c.2
Figure 8: % of ABE declaration in function of $\kappa/\phi$ according to CV, for some $\phi$ and sample sizes, for the $I_\mu$ conf. interval.

$n_1=n_2=12$

Figure 8a.1

$n_1=n_2=24$

Figure 8a.2

Figure 8b.1

Figure 8b.2

Figure 8c.1

Figure 8c.2
Figure 9: % Coverage in function of $\kappa/\phi$ according to CV, for some $\phi$ and sample sizes, for the $\mathcal{I}_W$ conf. interval.

- CV=10
- CV=20
- CV=30

$n_1=n_2=12$

$\phi = 0.095$

$n_1=n_2=24$

$\phi = 0.223$

$n_1=n_2=24$

$\phi = 0.262$
Figure 10: Interval Length in function of $\kappa / \phi$ according to CV, for some $\phi$ and sample sizes, for the $I_{\phi}$ conf. interval.

- $n_1 = n_2 = 12$
- $n_1 = n_2 = 24$

Formulation Effect = $\phi = 0.095$

Formulation Effect = $\phi = 0.223$

Formulation Effect = $\phi = 0.262$
Figure 11: % of ABE declaration in function of $\kappa/\phi$ according to CV, for some $\phi$ and sample sizes, for the $I_s$ conf. interval.

$\kappa/\phi$

$n_1=n_2=12$

$n_1=n_2=24$

Figure 11b.1

Figure 11b.2

Figure 11c.1

Figure 11c.2
Figure 12: % Coverage in function of $\kappa / \phi$ according to CV, for some $\phi$ and sample sizes, for the $I_s$ conf. interval.

- $n_1 = n_2 = 12$
  - CV=10
  - CV=20
  - CV=30

- $n_1 = n_2 = 24$

Formulation Effect = $\phi = 0.095$

Formulation Effect = $\phi = 0.223$

Formulation Effect = $\phi = 0.262$
Figure 13: Interval Length in function of $\kappa/\phi$ according to CV, for some $\phi$ and sample sizes, for the $I_s$ conf. interval.

- $CV=10$
- $CV=20$
- $CV=30$

$n_1=n_2=12$
Figure 13a.1

$n_1=n_2=24$
Figure 13a.2

Formulation Effect = $\phi = 0.095$

Formulation Effect = $\phi = 0.223$

Formulation Effect = $\phi = 0.262$
Figure 14: % of ABE declaration in function of $\kappa/\phi$ according to CV, for some $\phi$ and sample sizes, for the $I_*$ conf. interval.

- CV=10
- CV=20
- CV=30

$n_1=n_2=12$

Figure 14a.1

$n_1=n_2=24$

Figure 14a.2

Formulation Effect = $\phi = 0.095$ 

Figure 14b.1

Formulation Effect = $\phi = 0.223$

Figure 14b.2

Formulation Effect = $\phi = 0.262$

Figure 14c.1

Figure 14c.2
Figure 15: % Coverage in function of $\kappa/\phi$ according to CV, for some $\phi$ and sample sizes, for the $I_*$ conf. interval

- $CV=10$
- $CV=20$
- $CV=30$

$n_1=n_2=12$

Figure 15a.1

Figure 15a.2

$n_1=n_2=24$

Figure 15b.1

Figure 15b.2

Figure 15c.1

Figure 15c.2
Figure 16: Interval Length in function of \( \kappa/\phi \) according to CV, for some \( \phi \) and sample sizes, for the \( I_* \) conf. interval

\[ n_1 = n_2 = 12 \]

\[ n_1 = n_2 = 24 \]

\[ \phi = 0.095 \]

\[ \phi = 0.223 \]

\[ \phi = 0.262 \]
Figure 17: Effective Length in function of $\kappa / \phi$ according to CV, for some $\phi$ and sample sizes, for the $I$ conf. interval.

- $n_1 = n_2 = 12$
- $n_1 = n_2 = 24$

Figure 17a.1

Figure 17a.2

Figure 17b.1

Figure 17b.2

Figure 17c.1

Figure 17c.2
Figure 18: Effective Length in function of $\kappa / \phi$ according to CV, for some $\phi$ and sample sizes, for the $I_W$ conf. interval.

$n_1=n_2=12$

Figure 18a.1

$\phi = 0.095$

Figure 18a.2

$n_1=n_2=24$

Figure 18b.1

$\phi = 0.223$

Figure 18b.2

Figure 18c.1

$\phi = 0.262$

Figure 18c.2
Figure 19: Effective Length in function of $\kappa/\phi$ according to CV, for some $\phi$ and sample sizes, for the $I_S$ conf. interval.

- $n_1=n_2=12$
- $n_1=n_2=24$

**Figure 19a.1**

Formulation Effect = $\phi = 0.095$

**Figure 19a.2**

Formulation Effect = $\phi = 0.223$

**Figure 19b.1**

Formulation Effect = $\phi = 0.262$

**Figure 19b.2**

Formulation Effect = $\phi = 0.095$
Figure 20: Effective Length in function of $\kappa/\phi$ according to CV, for some $\phi$ and sample sizes, for the $I_*$ conf. interval.

$\kappa/\phi$

$\phi$ = 0.095

$\phi$ = 0.223

$\phi$ = 0.262

$n_1 = n_2 = 12$

$n_1 = n_2 = 24$

Figure 20a.1

Figure 20a.2

Figure 20b.1

Figure 20b.2

Figure 20c.1

Figure 20c.2