The AAPS Journal (2024) 26:50
https://doi.org/10.1208/s12248-024-00921-x

RESEARCH ARTICLE

Group-by-Treatment Interaction Effects in Comparative Bioavailability Studies

Helmut Schütz[1,2,3] · Divan A. Burger[4,5] · Erik Cobo[6] · David D. Dubins[7] · Tibor Farkás[8] · Detlew Labes[9] · Benjamin Lang[10] · Jordi Ocaña[11] · Arne Ring[12,5] · Anastasia Shitova[13] · Volodymyr Stus[14] · Michael Tomashevskiy[15]

Received: 23 October 2023 / Accepted: 3 April 2024

Abstract
Comparative bioavailability studies often involve multiple groups of subjects for a variety of reasons, such as clinical capacity limitations. This raises questions about the validity of pooling data from these groups in the statistical analysis and whether a group-by-treatment interaction should be evaluated. We investigated the presence or absence of group-by-treatment interactions through both simulation techniques and a meta-study of well-controlled trials. Our findings reveal that the test falsely detects an interaction when no true group-by-treatment interaction exists. Conversely, when a true group-by-treatment interaction does exist, it often goes undetected. In our meta-study, the detected group-by-treatment interactions were observed at approximately the level of the test and, thus, can be considered false positives. Testing for a group-by-treatment interaction is both misleading and uninformative. It often falsely identifies an interaction when none exists and fails to detect a real one. This occurs because the test is performed between subjects in crossover designs, and studies are powered to compare treatments within subjects. This work demonstrates a lack of utility for including a group-by-treatment interaction in the model when assessing single-site comparative bioavailability studies, and the clinical trial study structure is divided into groups.

Keywords average bioequivalence · group-by-treatment interaction · Monte-Carlo simulations · regulatory guidelines

Introduction

Comparative bioavailability (BA) studies, designed to demonstrate bioequivalence (BE) between two products, are an essential part of the generic approval process (1–5), bridging an innovator's product from the formulation used in clinical phase III to the to-be-marketed formulation (6), in the case of major variations of an approved product (7), to assess potential food effects (8) or drug-drug interactions (9, 10), and dose-proportionality (6). Such studies often involve multiple groups of subjects. This division is usually necessitated by logistical constraints, such as the limited capacity

✉ Helmut Schütz
helmut.schuetz@bebac.at

1 Center for Medical Data Science of the Medical University of Vienna, 1090 Vienna, Austria
2 Faculty of Pharmacy, Universidade de Lisboa, 1649-004 Lisbon, Portugal
3 BEBAC, Neubaugasse 36/11, 1070 Vienna, Austria
4 University of Pretoria, Pretoria, South Africa
5 University of the Free State, Bloemfontein, South Africa
6 Department of Statistics and Operations Research, Universitat Politecnica de Catalunya, Barcelona, Catalunya, Spain
7 Leslie Dan Faculty of Pharmacy, Toronto, Ontario, Canada
8 Gedeon Richter Plc., Budapest, Hungary
9 Berlin, Germany
10 Boehringer Ingelheim Pharma GmbH & Co. KG, Ingelheim am Rhein, Germany
11 Department of Genetics, Microbiology and Statistics, Universitat de Barcelona, Barcelona, Catalunya, Spain
12 Hexal – a Sandoz Brand, Holzkirchen, Germany
13 Quinta-Analytica Yaroslavl, Yaroslavl, Russian Federation
14 Zakłady Farmaceutyczne Polpharma S.A., Starogard Gdanski, Poland
15 OnTarget Group, Saint Petersburg, Russian Federation

Published online: 17 April 2024

of available beds or staffing levels at a single site. In many cases, groups are admitted in a staggered manner over the course of a few days but are recruited from the same subject pool. Studies conducted across multiple sites are beyond the scope of this research.

Given the close temporal proximity and the shared subject pool, one would generally not expect a relevant group effect to be introduced. However, deviations in point estimates could indicate a true group-by-treatment interaction, meaning that the treatment effect is not independent of the group. This observation could also be a result of chance. The validity of naïvely pooling data across these staggered groups can still be questioned.

We assessed the relevance and impact of group-by-treatment interactions through simulations and a meta-study comprising over 240 well-controlled trials.

Methods and Materials

The simulations and evaluation of datasets in the meta-study were performed in R 4.3.1 (11).

Models

The following linear models of $\log_e$-transformed pharmacokinetic (PK) responses with all fixed effects were used:

(1) *group, sequence, treatment, subject*(*group*×*sequence*), *period*(*group*), *group*×*sequence, group*×*treatment*
(2) *group, sequence, treatment, subject*(*group*×*sequence*), *period*(*group*), *group*×*sequence*
(3) *sequence, subject*(*sequence*), *period, treatment*

First public information about the use of Model 1 to test for a group-by-treatment interaction for the 2-treatment 2-sequence 2-period crossover design (2×2×2) by the US Food and Drug Administration (FDA) became available in 1999 (12), where *subject*(*group*×*sequence*) was considered a random effect. It must be mentioned that due to the *group*×*treatment* term, the main effect of treatment cannot be interpreted and, hence, must not be used to assess bioequivalence. The FDA suggested testing the group-by-treatment interaction at the 0.1 level (12, 13). If significant, data of groups must not be pooled, and bioequivalence can be demonstrated in one of the groups by Model 3, provided that the group meets the minimum requirements for a complete BE study that might also lead to the paradoxical situation that BE is demonstrated in a small group but fails in larger ones. If not significant, pooled data can be analyzed by Model 2. More details were given by the FDA later (14, 15), but without specifying a level of the test.

Model 2 takes the multi-group nature of the study into account and provides an unbiased estimate of the treatment effect. In the Eurasian Economic Union, Model 2 is mandatory, unless a justification to use Model 3 is stated in the protocol and discussed with the competent authority (16). Health Canada and the FDA recommend mixed-effects models, where subject-related effects are random and all others are fixed (2, 12, 14). Model 3 is the standard model for bioequivalence (e.g., 4, 5) with all effects fixed (analysis of variance, ANOVA).

In Model 2, the residual degrees of freedom (*df*) is $\sum n_i - 2 - (n_G - 1)$, where n_i is the number of subjects in sequence i, and n_G is the number of groups, and in Model 3 $df = \sum n_i - 2$. In both models, the back-transformed (1-2α) confidence interval (CI) is calculated as

$$\mathrm{CI} = 100\exp\left(\overline{\log_e x_\mathrm{T}} - \overline{\log_e x_\mathrm{R}} \mp t_{df,\alpha}\sqrt{m \times \sum_{i=1}^{i=s}\frac{1}{n_i}}\right),$$

where $\overline{\log_e x_\mathrm{T}}$ and $\overline{\log_e x_\mathrm{R}}$ are the means of the $\log_e$-transformed responses of the test and reference treatments, t is the t-value for df degrees of freedom at level α (commonly 0.05), m is the design constant (e.g., $^1/_2$ in a 2×2×2 crossover design, $^3/_8$ in a two-sequence three-period full replicate design, $^1/_4$ in a two-sequence four-period full replicate design, $^1/_6$ in a three-sequence three-period partial replicate design), *MSE* is the residual mean squares error, s is the number of sequences, and n_i is the number of subjects in sequence i.

It must be mentioned that the *MSE* is generally slightly different in Models 2 and 3, whereas the point estimate (*PE*) is identical if sequences are balanced and group sizes are identical, but different in the case of imbalanced sequences and unequal group sizes. Due to the fewer degrees of freedom, the CI of Model 2 is consistently wider than that of Model 3.

It should be mentioned that in comparative BA studies, subjects are uniquely coded (17, 18). Thus, *sequence* and related nested effects — as recommended in all guidelines — lead to over-specified models and can be removed entirely, without affecting the estimated treatment effect and its associated *MSE*.

Simulation Scenarios

Monte-Carlo simulations were performed based on the fact that the mean μ follows a lognormal distribution and the variance s^2 follows a χ^2-distribution with n–2 degrees of freedom (19). We simulated 100,000 studies in each scenario using the pseudo-random number generator Mersenne-Twister (20) with a fixed seed of 123456 to support reproducibility and assessed them

for the group-by-treatment interaction. In scenarios 1–12, we simulated $2 \times 2 \times 2$ designs with two groups. In scenarios 1–10, we simulated a sample size of 48 subjects to achieve ≥ 90% power for a geometric mean ratio (GMR) = 1 and CV_w = 33.5%. This sample size was selected to align closely with the median sample size 47 in the meta-study (see below). In scenarios 11 and 12, we simulated a sample size of 80 subjects to achieve ≥ 80% power for GMR = 0.90. To simulate unequal variances of groups, variance ratios of 0.667 and 1.5 were explored.

The level of the test of the group-by-treatment interaction was set to 0.05 (21). If no true group-by-treatment interaction was simulated, the fraction of studies with $p(G \times T) \leq 0.05$ represents empirical α, whereas if a true group-by-treatment interaction was simulated, it represents empirical power. The p-values of the group-by-treatment interaction tests are expected to follow a standard uniform distribution with ∈ {0, 1} and were assessed by the Kolmogorov–Smirnov test. Supplementary graphs illustrating the distribution of these p-values for each scenario are included to complement the Kolmogorov–Smirnov test findings, and the R-script to reproduce the simulations is provided in the Online Resource.

Table I presents a summary of simulation scenarios, categorizing them based on multiple parameters such as group sizes (n), whether the data exhibit equal or unequal variances of groups, and the corresponding CV, the GMR for each group involved in the scenarios, and indicating the presence or absence of true group-by-treatment interaction. Below is a detailed breakdown of these scenarios:

(1) Two groups of 24 subjects each, equal variances of groups, GMR = 1 in both groups, no group-by-treatment interaction
(2) Two groups of 24 subjects each, unequal variances of groups (variance-ratio 0.667), GMR = 1 in both groups, no group-by-treatment interaction
(3) Two groups of 24 subjects each, unequal variances of groups (variance-ratio 1.5), GMR = 1 in both groups, no group-by-treatment interaction
(4) n_1 = 38, n_2 = 10,equal variances of groups, GMR = 1 in both groups, no group-by-treatment interaction
(5) Two groups of 24 subjects each, equal variances of groups, GMR = 0.95 in the first group, GMR = 1.0526 in the second group, true group-by-treatment interaction; pooled GMR = 1
(6) Two groups of 24 subjects each, unequal variances of groups (variance-ratio 0.667), GMR = 0.95 in the first group, GMR = 1.0526 in the second group, true group-by-treatment interaction; pooled GMR = 1
(7) Two groups of 24 subjects each, unequal variances of groups (variance-ratio 1.5), GMR = 0.95 in the first group, GMR = 1.0526 in the second group, true group-by-treatment interaction; pooled GMR = 1
(8) n_1 = 38, n_2 = 10, equal variances of groups, GMR = 0.95 in the first group, GMR = 1.0526 in the second group, true group-by-treatment interaction; weighted GMR = 1
(9) n_1 = 38, n_2 = 10, unequal variances of groups (variance-ratio 0.667), GMR = 0.95 in the first group, GMR = 1.0526 in the second group, true group-by-treatment interaction; weighted GMR = 1

Table I Simulation Scenarios

Scen	Design	Group size (n_1, n_2,…)	Type	CV (%)	GMR/group	|ΔGMR|*	Pooled/ weighted GMR	G×T
1	2×2×2	24, 24	=	33.5, 33.5	1.0, 1.0	1.0000	1.0000	No
2	2×2×2	24, 24	≠	29.8, 36.9	1.0, 1.0	1.0000	1.0000	No
3	2×2×2	24, 24	≠	36.9, 29.8	1.0, 1.0	1.0000	1.0000	No
4	2×2×2	38, 10	=	33.5, 33.5	1.0, 1.0	1.0000	1.0000	No
5	2×2×2	24, 24	=	33.5, 33.5	0.95, 1.0526	1.1080	1.0000	Yes
6	2×2×2	24, 24	≠	29.8, 36.9	1.0526, 0.95	1.1080	1.0000	Yes
7	2×2×2	24, 24	≠	36.9, 29.8	0.95, 1.0526	1.1080	1.0000	Yes
8	2×2×2	38, 10	=	33.5, 33.5	1.0605, 0.8	1.3256	1.0000	Yes
9	2×2×2	38, 10	≠	29.8, 36.9	1.0605, 0.8	1.3256	1.0000	Yes
10	2×2×2	38, 10	≠	36.9, 29.8	1.0605, 0.8	1.3256	1.0000	Yes
11	2×2×2	40, 40	=	30.0, 30.0	0.9, 0.9	1.0000	1.0000	No
12	2×2×2	64, 16	≠	33.0, 26.7	0.8290, 1.25	1.5078	0.9000	Yes

= Equal variances of groups

≠ Unequal variances of groups

*Absolute value of max(GMR1–i) / min(GMR1–i)

(10) $n_1 = 38$, $n_2 = 10$, unequal variances of groups (variance-ratio 1.5), $GMR = 0.95$ in the first group, $GMR = 1.0526$ in the second group, true group-by-treatment interaction; weighted $GMR = 1$

(11) $n_1 = n_2 = 40$, equal variances of groups ($CV_w = 30\%$), $GMR = 0.90$ in both groups, no group-by-treatment interaction

(12) $n_1 = 64$, $n_2 = 16$, unequal variances of groups (variance-ratio 1.5), $GMR = 0.8290$ in the first group, $GMR = 1.2500$ in the second group, true group-by-treatment interaction; weighted $GMR = 0.9000$

Meta-study

The meta-study included a total of 328 datasets of *AUC* and 331 of C_{max} from 249 comparative BA studies (BE, food effect, drug-drug interaction, dose-proportionality), 157 analytes; 242 $2 \times 2 \times 2$ designs, 33 two-sequence four-period full replicate designs, three partial replicate design, as well as 46 incomplete block designs extracted from six-sequence three-period and four-sequence four-period Williams' designs. The studies consisted of two to seven groups, with a median sample size of 47 subjects (15–176) and a median interval separating groups of six days (1 to 62 days). It should be noted that the extreme interval of 2 months in one study was due to COVID-19 restrictions. The next largest interval was 18 days. In 76.3% of the studies, the interval was 1 week or less; in 30.8%, it was only 1 or 2 days. There are more datasets than studies because some contain more than one analyte (fixed-dose combinations or parent and metabolite). The datasets were assessed by all models. Since in some of the datasets bioequivalence of C_{max} was assessed by reference-scaling or with wider fixed limits, only *AUC* targeting BE with conventional limits of 80–125% was assessed by a recently proposed method (22), where

- a "concordant quantitative interaction" was defined as when the treatment effect is overall equivalent as well as in all groups but differs in magnitude,
- a "concordant qualitative interaction" was defined as when the treatment effect is overall and in at least one group equivalent, in at least one group not equivalent, and the treatment effects in all groups are in the same direction, and
- a "discordant qualitative interaction" was defined as when the overall treatment effect is equivalent, the treatment effect in some groups is not equivalent, and the treatment effect in some groups can be in opposite directions.

We restricted the method to two groups, because more would result in a multidimensional problem. Of note, a manipulation (i.e., an undocumented interim analysis after the first group and switching Test (T) with reference (R) in the second) would be only possible if groups are separated by a long interval. Such suspected manipulation could be easily detected by plotting T/R-ratios against subject ID. Details of the datasets are given in the Online Resource.

Results

Simulations

Table II presents the result of simulation scenarios, indicating the presence or absence of a true group-by-treatment interaction by empirical α or power (i.e., the fraction of studies with a significant group-by-treatment interaction in Model 1 if no or a true group-by-treatment interaction was simulated), and *p*-values of the Kolmogorov–Smirnov test.

To provide a clearer and more synthesized understanding of our simulation results in Table II, we have categorized the key findings regardless of the study design (crossover or parallel) as follows:

(1) Simulations without group-by-treatment interaction (Scenarios 1–4, and 11): In these scenarios, where no group-by-treatment interaction was introduced, the proportion of studies detecting a statistically significant interaction was close to the anticipated significance level of approximately 0.05.

(2) Crossover design simulations with group-by-treatment interaction (Scenarios 5–10, and 12): When a group-by-treatment interaction was introduced into these simulations, the empirical power increased in relation to the

Table II Results of 100,000 Simulated Studies in each Scenario

Scen	Design	G×T	Empirical α	Empirical power	*p*(unif.)
1	2×2×2	No	0.0497	–	0.756
2	2×2×2	No	0.0497	–	0.894
3	2×2×2	No	0.0499	–	0.927
4	2×2×2	No	0.0502	–	0.584
5	2×2×2	Yes	–	0.117	$<2.2 \cdot 10^{-16}$
6	2×2×2	Yes	–	0.117	$<2.2 \cdot 10^{-16}$
7	2×2×2	Yes	–	0.117	$<2.2 \cdot 10^{-16}$
8	2×2×2	Yes	–	0.348	$<2.2 \cdot 10^{-16}$
9	2×2×2	Yes	–	0.402	$<2.2 \cdot 10^{-16}$
10	2×2×2	Yes	–	0.294	$<2.2 \cdot 10^{-16}$
11	2×2×2	No	0.0499	–	0.733
12	2×2×2	Yes	–	0.944	$<2.2 \cdot 10^{-16}$

absolute value of the difference between the population means of the two groups.

In our first simulation scenario, used as an illustrative example in Fig. 1, we observed that the interaction was detected in about 4.97% of cases, even without a true group-by-treatment interaction. This detection rate is around/similar to the upper 95% significance limit of the binomial test (0.0511), indicating a low rate of false positives. Additionally, the uniformity of the *p*-values, validated by the Kolmogorov–Smirnov test ($p = 0.756$), suggests that their distribution aligns with the expected uniform pattern under the null hypothesis.

Meta-study

In 15 (4.57%) of the *AUC* datasets and 18 (5.44%) of the C_{max} datasets, a significant ($p < 0.05$) group-by-treatment interaction was detected, which is approximately the level of the test and does not exceed the upper 95% significance limits of the binomial test (0.0731 for $n = 328$ and 0.0725 for $n = 331$). See also Figs. 2 and 3, as well as Table III. Neither concordant nor discordant interaction was detected in the eligible *AUC* datasets (Fig. 4). In the dataset with the largest interval of 62 days separating groups, the *PE* in the first group was 95.37% and in the second 100.92%. The subjects' T/R-ratio showed no trend (see the Online Resource).

Discussion

As demonstrated in the simulations, significant group-by-treatment interactions were detected at approximately the level of the test, although none was simulated. Consequently, these cases are considered false positives. When true group-by-treatment interactions were simulated, in most cases, the test failed to detect them, i.e., showed low empirical power. Only with large sample sizes and extremely different group sizes is a true group-by-treatment interaction correctly detected with sufficient power. Heteroscedasticity did not affect the results, which is not surprising since the pooled data models assume homoscedasticity.

The simulations underscored a crucial consideration in the context of group-by-treatment interaction testing, revealing that the smaller the true group-by-treatment interaction, the more challenging it becomes to detect. This prompts a thoughtful reflection on the definition of what is "small enough to be ignored for practical purposes." Conversely, the findings emphasize that a substantial group-by-treatment interaction is necessary for the test to be valuable in studies designed to demonstrate bioequivalence. This is

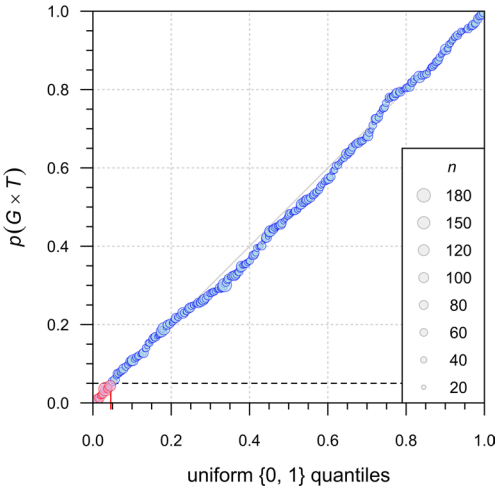

Fig. 2 *AUC*, $p(\mathrm{G}\times\mathrm{T}) = 0.0457$, $p(\mathrm{unif.}) = 0.661$ (meta-study, $n = 328$)

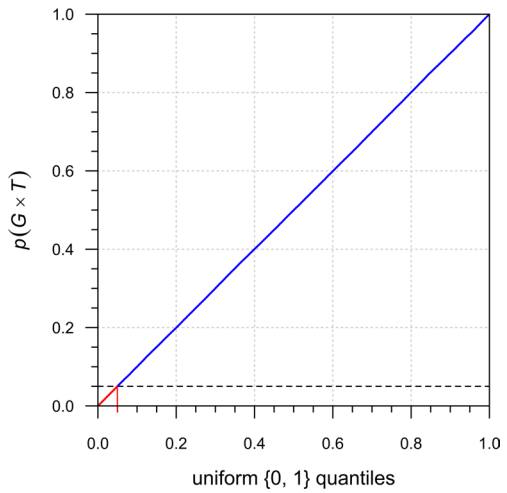

Fig. 1 $p(\mathrm{G}\times\mathrm{T}) = 0.0497$, $p(\mathrm{unif.}) = 0.756$ (simulation scenario 1)

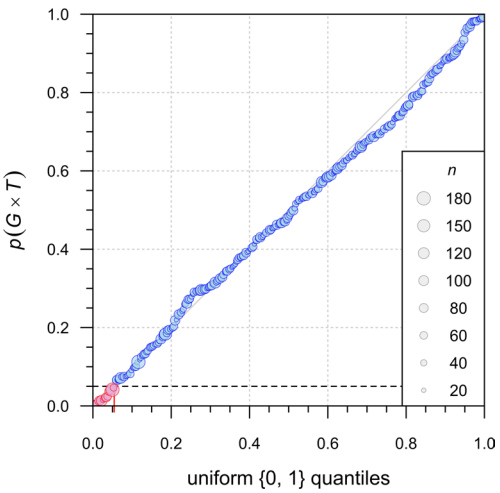

Fig. 3 C_{max}, $p(\mathrm{G}\times\mathrm{T}) = 0.0544$, $p(\mathrm{unif.}) = 0.483$ (meta-study, $n = 331$)

Table III Results of the Meta-study

PK metric	Datasets	$p(G\times T)$	signif.[a]	p (unif.)	signif.[b]	3[c]	2[d]	Loss[e]
AUC	328	0.0457	No	0.661	No	86.9%	81.7%	5.96%
C_{max}	331	0.0544	No	0.483	No	71.0%	66.2%	6.81%

[a]Above the significance limit of the binomial test at the 0.05 level (0.0731, 0.0725)

[b]Below the significance limit of the Kolmogorov–Smirnov test at the 0.05 level

[c]Passing 80–125% when evaluated by Model 3

[d]Passing 80–125% when evaluated by Model 2

[e]Relative loss in passing rate when evaluated by Model 2 compared to Model 3

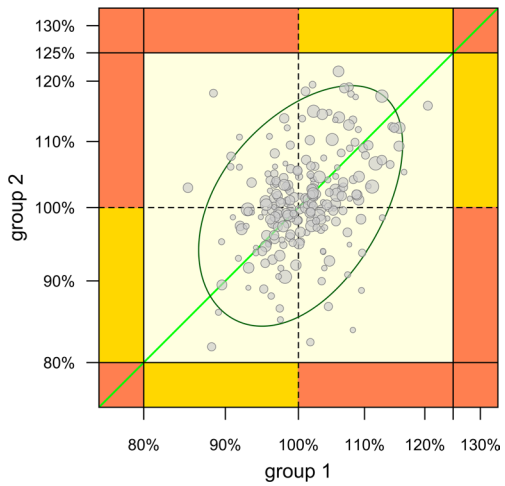

Fig. 4 *PE*s of *AUC*, analysis of interaction (22) (meta-study, $n=226$ targeting BE by Model 3; center square quantitative, yellow areas concordant qualitative, orange areas discordant qualitative, 95% confidence ellipse in green, unity line in bright green)

corroborated by the empirical power results presented in Table II and Fig. 13 of the Online Resource.

Based on the meta-analysis of well-controlled studies, it appears that significant group-by-treatment interactions are detected merely due to chance and can be considered "statistical artifacts" or false positives. Although only 226 datasets of *AUC* with two groups were eligible for a recently proposed method (22), neither concordant nor discordant interaction was detected. Testing for a group-by-treatment interaction to detect data manipulation is limited, since there is no evidence that manipulation is linked with clinic groups.

When the datasets of the meta-study were evaluated by Model 2, about 6.4% less than with Model 3 passed the conventional limits for BE of 80.00–125.00%. This difference can be attributed to potential bias in the estimation of the treatment effect introduced by group-related terms (i.e., *subject*(*group*×*sequence*), *period*(*group*), *group*×*sequence*) and fewer degrees of freedom leading to slightly wider confidence intervals. However, this observation might not only be due to fewer degrees of freedom, but also mainly due to different residual errors and imbalanced sequences together with unequal group sizes. This finding is similar to another meta-study (23), where fewer studies passed with the carryover term in the model than without the term. It is impossible to predict whether the additional group terms by Model 2 can "explain" part of the variability, i.e., its residual *MSE* may be smaller or larger than that of Model 3.

In light of these results, we consider that Model 1 originally proposed by the FDA (12, 13) as a pre-test should be avoided due to the risk of type I error inflation. Well-known examples where a pre-test inflates the type I error are assessing variance homogenicity (24) and testing for a sequence effect in comparative bioavailability (25, 26). For this reason, our recommendation is to use Model 2 (or 3) instead. This investigation is reminiscent of the discussion of the subject-by-formulation variance component, with a similar result: The estimate for this variance component was positively biased, leading to substantial false-positive tests (27). In analogy, none of the published adaptive sequential methods contains a "poolability criterion" (28–34). Instead, data are always pooled, regardless of the results of the stages. As recently recommended, the planned model and procedures should be unambiguously stated in the protocol (5, 14, 15). Subgroup results should always be interpreted cautiously (35). In order to increase power, Bayesian shrinkage analysis of subgroups (36) must only be applied if specified *a priori* and not *post hoc* (i.e., after detecting a significant group-by-treatment interaction). Data-driven *post hoc* analysis is also discouraged by the International Council of Harmonisation (5).

It must be mentioned that in frequentist statistics, the outcome of any level α-test is dichotomous: The null hypothesis is either rejected or not rejected, not something that can be represented with a probability. It is a common fallacy to regard the p-value as the probability that the null hypothesis is true — or the alternative hypothesis is false (37, 38). It is well known that the more sophisticated interaction terms have a higher standard error than those of the main effects. Moreover, even more so in this case, since they involve a comparison between subjects instead of the main comparison which is within subjects, with a lower residual variance. On the other hand, the main analysis in an equivalence study is based on a Neyman-Pearson (NP) test, designed

with a review of the evidence (either published or not) in favor of $\{\theta_1, \theta_2\}$, the limits of the "not clinically relevance" margin. That is, the alternative hypothesis H_1 is an interval $\theta_1 < \mu_T - \mu_R < \theta_2$. Furthermore, the sample size has been determined to obtain the desired power, taking into account the standard error of the estimator of the main comparison. This higher standard error leads to a lower power for the interaction, which, added to the lack of prior support for Δ, explains the results obtained in this study, which summarizes the well-known joke "Enjoy your unexpectedly significant results, … because you will not see them again."

In order to recapitulate, the standard significance test lacks both power and prior support for H_1, leading to (39) — which respects the NP test. Therefore, we must distinguish between NP and significance testing (39), as well as remember the advice about lack of power and prior support for H_1 (40).

Conclusion

Testing for a group-by-treatment interaction is neither useful nor appropriate. When a group-by-treatment interaction does not exist in the data, it will incorrectly be detected at the level of the test. Even when a true group-by-treatment interaction exists, it will likely not be detected — except in the case of large sample sizes and extremely different group sizes — because in crossover designs, T *vs.* R is tested with a greater sample size than the G×T interaction; in the former, all subjects are used, whereas, for G×T, the subjects are split into groups and tested between them. Since the test has low power but will be significant at the α level even in the absence of true group-by-treatment interaction, it is not in any way clear how this test could contribute to regulatory decision-making. This work demonstrates a lack of utility for including a group-by-treatment interaction in the model for assessment of single-site comparative bioavailability studies when the clinical trial study structure is divided into groups for logistical reasons. The authors thus see no particular merit in this test for regulatory submissions anywhere.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1208/s12248-024-00921-x.

Acknowledgements The 27 companies from 17 countries provided data for the meta-study. We would like to thank Susana Almeida (Medicines for Europe, Belgium), Anders Fuglsang (Fuglsang Pharma, Denmark), Jiří Hofmann (Zentiva, Czech Republic), Eduard Molins Lleonart (AstraZeneca, Spain), Paulo Paixão (University of Lisbon, Portugal), and Barbara Schug and Ralph-Steven Wedemeyer (SocraTec, Germany) for fruitful discussions while developing the concept of this work.

Author Contribution The concept was developed by HS, DL, AS, and MT. HS was responsible for the simulations and assessment of the meta-study. HS drafted the manuscript, and all authors revised the manuscript critically for intellectual content and approved the final version.

Funding Open access funding provided by Medical University of Vienna.

Declarations

Conflict of Interest The authors declare no competing interests.

References

1. US Food and Drug Administration, CDER. Guidance for industry. Bioequivalence studies with pharmacokinetic endpoints for drugs submitted under an ANDA. Silver Spring, August 2021. https://www.fda.gov/media/87219/download. Accessed 22 October 2023.
2. Health Canada, Guidance Document. Conduct and analysis of comparative bioavailability studies. Ottawa. Revised Date: 2023/01/30. https://www.canada.ca/content/dam/hc-sc/documents/services/drugs-health-products/drug-products/applications-submissions/guidance-documents/bioavailability-bioequivalence/conduct-analysis-comparative.pdf. Accessed 22 October 2023.
3. Davit B, Braddy AC, Conner DP, Yu LX. International guidelines for bioequivalence of systemically available orally administered generic drug products: a survey of similarities and differences. AAPS J. 2013;15(4):974–290. https://doi.org/10.1208/s12248-013-9499-x.
4. EMA, CHMP. Guideline on the investigation of bioequivalence. London; 2010. https://www.ema.europa.eu/documents/scientific-guideline/guideline-investigation-bioequivalence-rev1_en.pdf. Accessed 22 October 2023.
5. ICH. Bioequivalence for immediate release solid oral dosage forms. M13A. Draft version. 20 December 2022. https://database.ich.org/sites/default/files/ICH_M13A_Step2_draft_Guideline_2022_1125.pdf. Accessed 22 October 2023.
6. US Food and Drug Administration, CDER. Guidance for industry. Bioavailability studies submitted in NDAs or INDs — general considerations. Silver Spring, April 2022. https://www.fda.gov/media/121311/download. Accessed 22 October 2023.
7. US Food and Drug Administration, CDER. Guidance for industry. SUPAC-IR: immediate-release solid oral dosage forms: scale-up and post-approval changes: chemistry, manufacturing and controls, in vitro dissolution testing, and in vivo bioequivalence documentation. Rockville, November 1995. https://www.fda.gov/media/70949/download. Accessed 22 October 2023.
8. US Food and Drug Administration, CDER. Guidance for industry. Food-effect bioavailability and fed bioequivalence studies. Rockville, December 2002. https://www.fda.gov/files/drugs/published/Food-Effect-Bioavailability-and-Fed-Bioequivalence-Studies.pdf. Accessed 22 October 2023.

9. US Food and Drug Administration, CDER. Guidance for industry. Clinical drug interaction studies — cytochrome P450 enzyme- and transporter-mediated drug interactions. Silver Spring, January 2020. https://www.fda.gov/media/134581/download. Accessed 22 October 2023.
10. European Medicines Agency, CHMP. Guideline on the investigation of drug interactions. Revision 1. London, 21 June 2012. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-investigation-drug-interactions-revision-1_en.pdf. Accessed 22 October 2023.
11. R Core Team. R: a language and environment for statistical computing. Vienna, Austria, 2023. https://www.r-project.org/. Accessed 22 Oct 2023.
12. US Food and Drug Administration, CDER. ANDA 077570. Bioequivalence reviews. 2008. Control document #98–392 regarding the Group-by-Treatment interaction discussion. Rockville, September 10, 1999. https://www.accessdata.fda.gov/drugsatfda_docs/anda/2008/077570Orig1s000BioeqR.pdf. Accessed 22 October 2023.
13. Bolton S, Bon C. Pharmaceutical statistics. Practical and clinical applications. 5th ed. New York: Informa Healthcare; 2010. p. 629.
14. US Food and Drug Administration, CDER. Guidance for industry. Statistical approaches to establishing bioequivalence. Revision 1. Silver Spring, December 2022. https://www.fda.gov/media/163638/download. Accessed 22 October 2023.
15. Sun W. Bioequivalence studies in multiple groups. Presentation at: SBIA. A deep dive: FDA draft guidance on statistical approaches to establishing bioequivalence. Silver Spring, March 14, 2023. https://www.fda.gov/media/167459/download. Accessed 22 October 2023.
16. Council of the Eurasian Economic Community. On the approval of the rules for conducting research of bioequivalence of drugs within the framework of the Eurasian Economic Union. November 3, 2016; amended September 4, 2020.
17. Westlake WJ. Design and evaluation of bioequivalence studies in man. In: Blanchard J, Sawchuk RJ, Brodie BB, editors. Principles and perspectives in drug bioavailability. Basel: Karger; 1979. p. 192–210.
18. Westlake WJ. Bioavailability and bioequivalence of pharmaceutical formulations. In: Peace KE, editor. Biopharmaceutical statistics for drug development. New York: Marcel Dekker; 1988. p. 336–7.
19. Zheng C, Wang J, Zhao L. Testing bioequivalence for multiple formulations with power and sample size calculations. Pharm Stat. 2012;11:334–41. https://doi.org/10.1002/pst.1522.
20. Matsumoto M, Nishimura T. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Trans Model Comput Simul. 1998;8:3–30. https://doi.org/10.1145/272991.272995.
21. Medicines for Europe. Second bioequivalence workshop. Session 2 – ICH M13 – bioequivalence for IR solid oral dosage forms. Brussels; 2023.
22. Sun W, Schuirmann D, Grosser S. Qualitative versus quantitative treatment-by-subgroup interaction in equivalence studies with multiple subgroups. Stat Biopharm Res. 2022. https://doi.org/10.1080/19466315.2022.2123385.
23. D'Angelo G, Potvin D, Turgeon J. Carry-over effects in bioequivalence studies. J Biopharm Stat. 2001;11(1&2):35–43. https://doi.org/10.1081/bip-100104196.
24. Zimmerman DW. A note on preliminary tests of equality of variances. Br J Math Stat Psychol. 2004;57(1):173–81. https://doi.org/10.1348/000711004849222.
25. Freeman PR. The performance of the two-stage analysis of two-treatment, two-period crossover trials. Statist Med. 1989;8(12):1421–32. https://doi.org/10.1002/sim.4780081202.
26. Senn S, D'Angelo G, Potvin D. Carry-over in cross-over trials in bioequivalence: theoretical concerns and empirical evidence. Pharm Stat. 2004;3:133–42. https://doi.org/10.1002/pst.111.
27. Endrényi L, Taback N, Tóthfalusi. Properties of the estimated variance component for subject-by-formulation interaction in studies of individual bioequivalence. Statist Med. 2000;19:2867–78. https://doi.org/10.1002/1097-0258(20001030)19:20%3C2867::aid-sim551%3E3.0.co;2-j.
28. Potvin D, DiLiberti CE, Hauck WW, Parr AF, Schuirmann DJ, Smith RA. Sequential design approaches for bioequivalence studies with crossover designs. Pharm Stat. 2008;7:245–62. https://doi.org/10.1002/pst.294.
29. Montague TH, Potvin D, DiLiberti CE, Hauck WW, Parr AF, Schuirmann DJ. Additional results for 'Sequential design approaches for bioequivalence studies with crossover designs.' Pharm Stat. 2011;11:8–13. https://doi.org/10.1002/pst.483.
30. Fuglsang A. Sequential bioequivalence trial designs with increased power and controlled type I error rates. AAPS J. 2013;15:659–61. https://doi.org/10.1208/s12248-013-9475-5.
31. Fuglsang A. Sequential bioequivalence approaches for parallel designs. AAPS J. 2014;16:373–8. https://doi.org/10.1208/s12248-014-9571-1.
32. Maurer W, Jones B, Chen Y. Controlling the type 1 error rate in two-stage sequential designs when testing for average bioequivalence. Statist Med. 2018;37(10):1–21. https://doi.org/10.1002/sim.7614.
33. Lee J, Feng K, Xu M, Gong X, Sun W, Kim J, Zhang Z, Wang M, Fang L, Zhao L. Applications of adaptive designs in generic drug development. Clin Pharm Ther. 2020;110(1):32–5. https://doi.org/10.1002/cpt.2050.
34. Molins E, Labes D, Schütz H, Cobo E, Ocaña J. An iterative method to protect the type I error rate in bioequivalence studies under two-stage adaptive 2×2 crossover designs. Biom J. 2021;63(1):122–33. https://doi.org/10.1002/bimj.201900388.
35. Kent DM, Paulus JK, van Klaveren D, D'Agostino R, Goodman S, Hayward R, Ioannidis JPA, Patrick-Lake B, Morton S, Pencina M, Raman G, Ross JS, Selker HS, Varadhan R, Vickers A, Wong JB, Steyerberg EW. The predictive approaches to treatment effect heterogeneity (PATH) statement. Ann Intern Med. 2020;172(1):35–45. https://doi.org/10.7326/M18-3667.
36. Alosh M, Fritsch K, Huque M, Mahjoob K, Pennello G, Rothmann M, Russek-Cohen E, Smith F, Wilson S, Yue L. Statistical considerations on subgroup analysis in clinical trials. Stat Biopharm Res. 2015;7(4):286–304. https://doi.org/10.1080/19466315.2015.1077726.
37. Press WH, Flannery BP, Teukolsky SA, Vetterling WT. Numerical recipes in Fortran 77: the art of scientific computing. 2nd ed. Cambridge: Cambridge University Press; 1992. p. 603.
38. Wasserstein RL, Lazar NA. The ASA's statement on *p*-values: context, process, and purpose. Am Stat. 2016;70(2):129–33. https://doi.org/10.1080/00031305.2016.1154108.
39. Hubbard R, Bayarri MJ. Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. Am Stat. 2003;57(3):171–82. https://doi.org/10.1198/0003130031856.
40. Cortés J, Casals M, Langohr K, González JA. Importance of statistical power and hypothesis in P value. Med Clin (Barc). 2016. https://doi.org/10.1016/j.medcli.2015.10.011.