Group-by-Treatment Interaction Effects in Comparative Bioavailability Studies — A Myth or Reality?

Helmut Schütz
Center for Medical Data Science of the Medical University of Vienna
Faculty of Pharmacy of the University of Lisbon
BEBAC Vienna

Background

For logistical reasons, crossover studies are sometimes performed in groups

- Limited capacity of the clinical site
  - Some jurisdictions accept reference-scaling only for C_{max} – leading to extreme sample sizes if products are highly variable in *AUC* as well
  - Even large sites might have only a limited capacity with full monitoring
- Mixed-sex groups are not feasible in certain countries (*e.g.*, India, Jordan) – studies are performed in multiple single-sex groups
- Recruitment might be an issue in large studies in patients – regularly performed in multiple groups or even several clinical sites

M13A (ICH 2024)

> “ The statistical model should take into account the multi-group nature of the BE study, *e.g.*, by using a model including terms for group, sequence, sequence × group, subject within sequence × group, period within group and formulation. The group × treatment interaction term should not be included in the model. However, applicants should evaluate potential for heterogeneity of treatment effect across groups and discuss its potential impact on the study data, *e.g.*, by investigation of group × treatment interaction in a supportive analysis and calculation of descriptive statistics by group. — implemented by April 2025: SwissMedic, FDA, EMA, MHRA, JFDA

Statistical Models

I. $\log_e(Y)$ ~ Group, Sequence, Group × Sequence,
Subject(Group × Sequence), Period(Group), Treatment,
Group × Treatment

No unbiased estimate of the treatment effect possible –
only to test the G × T interaction (M13A: supportive <u>after</u> model II)

II. $\log_e(Y)$ ~ Group, Sequence, Group × Sequence,
Subject(Group × Sequence), Period(Group), Treatment

Model to assess BE (M13A)

III. $\log_e(Y)$ ~ Sequence, Treatment, Subject(Sequence), Period

Conventional model (eventual groups not taken into account)

Supportive Analysis

Model I

- Significance of the G × T interaction term (tested at 0.05 level)
  - Crossover studies are powered to detect a potential difference in the treatment-effect (*within* subjects) and not for any group-related terms (*between* subjects)
  - Expect to detect false positives in 5% of studies
  - The *p*-value of a level α-test test is neither the probability that the null hypothesis is true nor that the alternative hypothesis is false – in frequentist statistics the outcome is dichotomous:
    Hypotheses are considered *true or false*, not something that can be represented with a probability

Supportive Analysis (cont'd)

Heterogeneity of treatment effect across groups, descriptive statistics by group

- If there are more than two groups, all *pairwise* comparisons?
- *Which* descriptive statistics?
  - Geometric means and CV?
  - Confidence intervals?

Discuss its potential impact on the study data

- What to 'discuss' – which is not trivial?

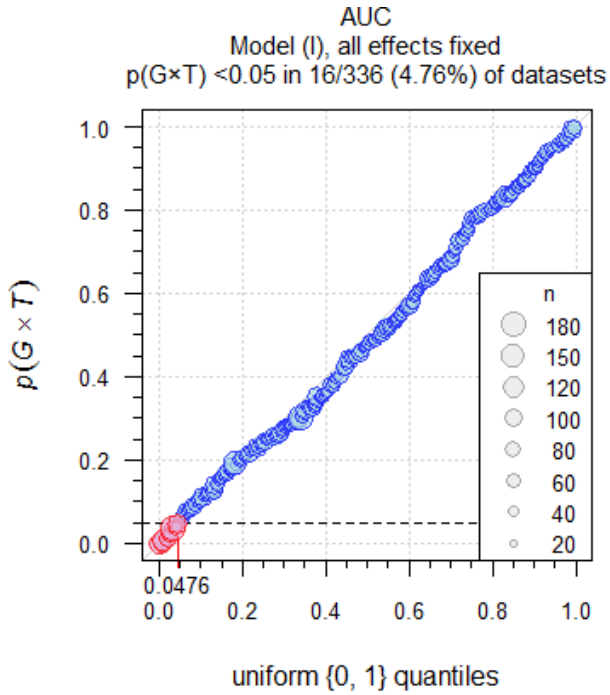
Meta–Study: Data, expectations

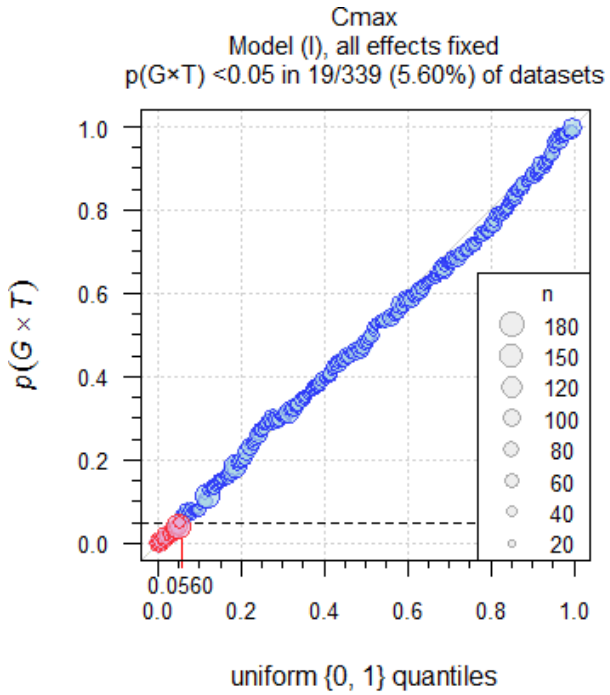
255 comparative BA studies (336 datasets of *AUC*, 339 of C_{max}) assessed for the G×T interaction

- 163 analytes, 2 – 7 groups, $\tilde{x}$ sample size 47 (15 – 176), $\tilde{x}$ interval separating groups 6 days (77% one week or less; 31% 1 or 2 days)
  - If true Group-by-Treatment interactions ($p(G \times T) < 0.05$) exist, we would observe them in *more* than 5% of datasets (*i.e.*, above the false positive rate)
  - *p*-values of the G × T-tests should follow the standard uniform distribution – we can assess this hypothesis by the Kolmogorov–Smirnov test

* Schütz H, Burger DA, Cobo E, Dubins D, Farkás T, Labes D, Lang B, Ocaña J, Ring A, Shitova A, Stus V, Tomashevskiy M. *Group-by-Treatment Interaction Effects in Comparative Bioavailability Studies*. AAPS J. 2024; 26(3): 50. https://doi.org/10.1208/s12248-024-00921-x

Meta–Study: Results





Both PK metrics

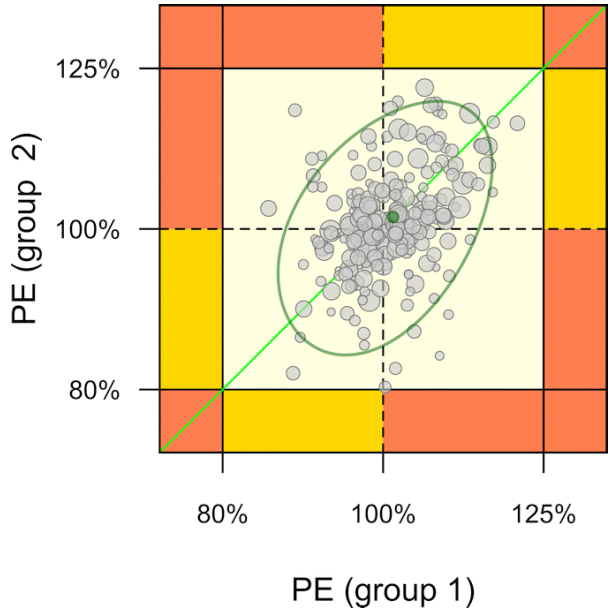
- *p*(G×T) at approximately α (FPR?)
- Null hypothesis of standard normality of *p*(G×T) not rejected

Method and results questioned*

* Sun W, Alosh M, Schuirmann DJ, Grosser S. *Letter to the Editor on "Group-by-Treatment Interaction Effects in Comparative Bioavailability Studies".* AAPS J. 2024; 26(5): 101

Meta–Study: Results (cont'd)

In all 230 passing *AUC* datasets with two groups interactions were concordant quantitative[1] (*i.e.*, the treatment effects were equivalent overall and in both groups, but differed in their magnitude[2])

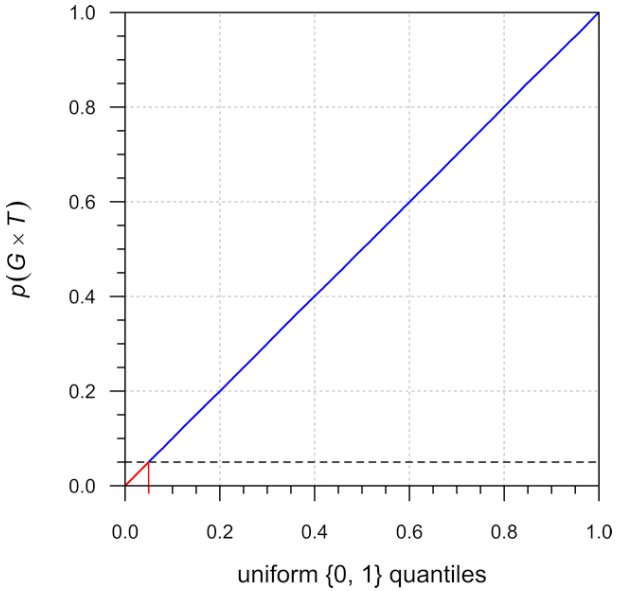


1 Schütz H, Burger DA, Cobo E, Dubins D, Farkás T, Labes D, Lang B, Ocaña J, Ring A, Shitova A, Stus V, Tomashevskiy M. *Rejoinder to the 'Letter to the Editor' on "Group-by-Treatment Interaction Effects in Comparative Bioavailability Studies"*. AAPS J. 2025; 27(1): 14. https://doi.org/doi:10.1208/s12248-024-01008-3

2 Sun W, Schuirmann D, Grosser S. *Qualitative versus Quantitative Treatment-by-Subgroup Interaction in Equivalence Studies with Multiple Subgroups*. Stat Biopharm Res. 2022; 15(4): 737–47. https://doi.org/10.1080/19466315.2022.2123385

Simulations

$CV_1 = CV_2 = 33.5\%$
$n_1 = n_2 = 24$
(close to sample size $\tilde{x}$ 47 of the meta–study)
$GMR_1 = GMR_2 = 1$
(no true G×T interaction)
100,000 simulated studies

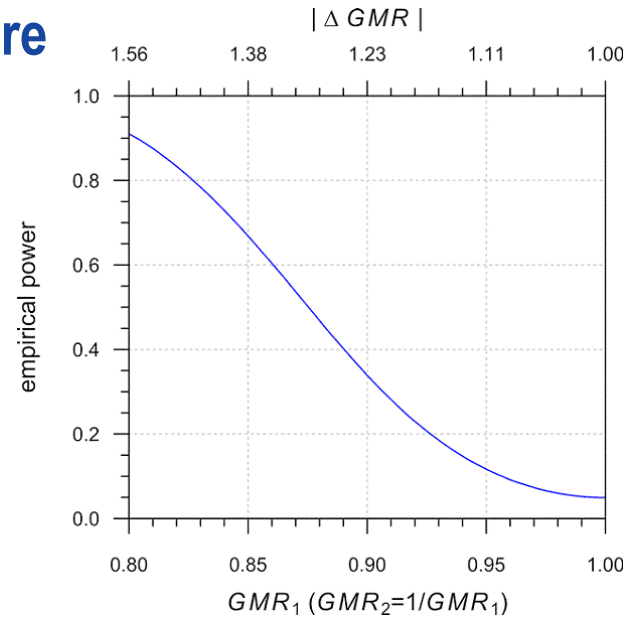


Results

- p(G×T) = 0.04967 (≤0.0511); close to false positive rate
- p(unif.) = 0.7563; null hypothesis of standard normality of p(G×T) not rejected

Simulations

Same conditions as before (*CV*, *n*),
$GMR_2 = 1 / GMR_1$
8 million simulated studies



Results

- Crossover studies are powered for the within-subject variance
- Since G×T is a between-subject effect, power of the test is poor – unless the *GMR*s of groups differ by a large amount

GMR_1	GMR_2	power
0.80	1.25	0.91
0.88	1.14	0.50
1.00	1.00	0.05

Conclusions

Based on the meta–study and the simulations

- If the G×T test yields a statistically significant ('positive') result, it is frequently false
  - There is no strong evidence that there is a true G×T interaction in trials of the meta–study
  - This means also that it is likely (but not proven) that most of the time when the G×T test is positive, it is simply statistical noise and not positive because of a *bona fide* group-by-treatment interaction
  - Investigation of a 'root cause' – as suggested in the M13A draft – is futile and thus, was removed in the final guideline

Conclusions (cont')

Based on the meta–study and the simulations

- The true G×T interaction is unknown in real studies – which is trivial
  - It is impossible to ascertain *any* true effect of *any* given model in real studies
    - We suggest that those who postulate the existence of a such a purported group-by-treatment interaction should be prepared to bear the burden of proof
    - Extraordinary claims require extraordinary evidence; if agencies have substantiated instances of a genuine G×T interaction accompanied by a plausible causative mechanism, they should be published for a second assessment

G×T Interaction Effects in Comparative BA Studies

Thank You!

Helmut Schütz

Center for Medical Data Science, Medical University of Vienna
1090 Vienna, Austria
helmut.schuetz@muv.ac.at
Faculty of Pharmacy of the University of Lisbon
1649-003 Lisbon, Portugal
helmut@edu.ulisboa.pt
BEBAC
1070 Vienna, Austria
helmut.schuetz@bebac.at

Acknowledgments: Anders Fuglsang, Jiří Hofmann, Anže Kastelic, Barbara Schug, Ralph-Steven Wedemeyer and the 28 companies of 18 countries providing data for the meta–study