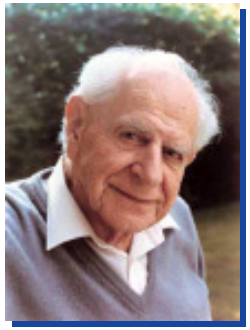
Multi-Group Studies in Bioequivalence.
To pool or not to pool?

Helmut Schütz

To bear in Remembrance...

Whenever a theory appears to you as the only possible one, take this as a sign that you have neither understood the theory nor the problem which it was intended to solve.



Karl R. Popper

Even though it's *applied* science we're dealin' with, it still is – *science*!

Leslie Z. Benet

Group Effect

Sometimes subjects are split into two or more groups

- Reasons
  - Lacking capacity of the clinical site:
    Some approaches (EMA, ASEAN States, Australia, Brazil, Egypt, Russian Federation, EEU, New Zealand) allow reference-scaling only for C_{max} – which leads to sample sizes of >100 subjects if the product is highly variable in *AUC* as well.
  - Some PIs don't trust in the test product and prefer to start the study in a small group of subjects.
- The common model for crossover studies *might* not be applicable any more.
  - Periods were performed on different dates.
  - Questions may arise whether groups can be naïvely pooled.
    - In a strict sense only valid if the GMRs of groups would be equal, *i.e.*, there is no Group-by-Treatment interaction.

Group Effect

Description

- Bolton and Bon
  - The totality of data is analyzed with a new term in the analysis of variance (ANOVA), a Treatment × Group interaction term. This is a measure (on a log scale) of how the ratios of test to reference differ in the groups. For example, if the ratios are very much the same in each group, the interaction would be small or negligible. If interaction is large, as tested in the ANOVA, then the groups cannot be combined. However, if at least one of the groups individually passes the confidence interval criteria, then the test product would be acceptable. If interaction is not statistically significant ($p > 0.10$), then the confidence interval based on the pooled analysis will determine acceptability.

Bolton S, Bon C. *Pharmaceutical Statistics. Practical and Clinical Applications.* New York: informa healthcare; Fifth edition 2009. p. 629.

Review of Guidelines

FDA 2001

- If a crossover study is carried out in two or more groups of subjects (e.g., if for logistical reasons only a limited number of subjects can be studied at one time), the statistical model should be modified to reflect the multigroup nature of the study. In particular, the model should reflect the fact that the periods for the first group are different from the periods for the second group.
- If the study is carried out in two or more groups and those groups are studied at different clinical sites […], questions may arise as to whether the results from the several groups should be combined in a single analysis.

Review of Guidelines

FDA cont'd

- No details about the analysis are given in any guidance. However, this text can be found under the FOI:
  - The following statistical model can be applied:
    - Group
    - Sequence
    - Treatment
    - Subject (nested within Group × Sequence)
    - Period (nested within Group)
    - Group-by-Sequence Interaction
    - Group-by-Treatment Interaction
  - Subject (nested within Group×Sequence) is a random effect and all other effects are fixed effects.

Review of Guidelines

FDA cont'd

- FOI cont'd
  - If the Group-by-Treatment interaction test is not statistically significant ($p \geq 0.1$), only the Group-by-Treatment term can be dropped from the model.
  - If the Group-by-Treatment interaction is statistically significant ($p < 0.1$), DBE requests that equivalence be demonstrated in one of the groups, provided that the group meets minimum requirements for a complete bioequivalence study.
  - Please note that the statistical analysis for bioequivalence studies dosed in more than one group should commence only after all subjects have been dosed and all pharmacokinetic parameters have been calculated. Statistical analysis to determine bioequivalence within each dosing group should never be initiated prior to dosing the next group; otherwise the study becomes one of sequential design.

Review of Guidelines

FDA cont'd

- FOI cont'd
  - If ALL of the following criteria are met, it may not be necessary to include Group-by-Treatment in the statistical model:
    - the clinical study takes place at one site;
    - all study subjects have been recruited from the same enrollment pool;
    - all of the subjects have similar demographics;
    - all enrolled subjects are randomly assigned to treatment groups at study outset.
  - In this latter case, the appropriate statistical model would include only the factors
    - Sequence, Period, Treatment and Subject (nested within Sequence).

Review of Guidelines

Eurasian Economic Union 2016

93. Если перекрестное исследование проведено в 2 и более группах субъектов, т.е. разбиение всей выборки на несколько групп, каждая из которых начинает участие в исследовании в разные дни (например, если из логистических соображений единовременно в клиническом центре можно провести исследование с участием ограниченного числа субъектов), в целях отражения многогруппового характера исследования необходимо модифицировать статистическую модель. В частности, в модели необходимо учесть тот факт, что периоды для первой группы отличаются от периодов для второй (и последующих) группы.

Review of Guidelines

Eurasian Economic Union 2016

94. Если исследование проведено в двух и более группах и эти группы изучались в различных клинических центрах или в одном и том же центре, но были разделены большим промежутком времени (например, месяцами), возникает сомнение относительно возможности объединения результатов, полученных этих группах, в один анализ. Такие ситуации необходимо обсуждать с уполномоченным органом.
Если предполагается проведение исследования в нескольких группах из логистических соображений, об этом необходимо явно указать в протоколе исследования; при этом, если в отчете отсутствуют результаты статистического анализа, учитывающие многогрупповой характер исследования, необходимо представить научное обоснование отсутствия таких результатов.

Review of Guidelines

EMA 2010

- The study should be designed in such a way that the formulation effect can be distinguished from other effects.
- The precise model to be used for the analysis should be pre-specified in the protocol. The statistical analysis should take into account sources of variation that can be reasonably assumed to have an effect on the response variable.

Statistical Models

Proposed by the FDA

- **Model I**
  - **Fixed effects:**
    Group, Sequence, Treatment, Period(Group), Group×Sequence, Group×Treatment
  - **Random effect:**
    Subject(Group×Sequence)
  - **If the Treatment-by-Group interaction term is not significant at the 0.1 level, data of all groups can be pooled and the term dropped (*i.e.*, proceed with Model II).**
  - **If the Treatment-by-Group interaction term is significant at the 0.1 level, data must not be pooled and Model III of the largest site applied.**
  - **Intra-subject contrasts for the estimation of the treatment effect (and hence, a PE and its CI) cannot be unbiased obtained from this model. It serves only as a decision tool.**

Statistical Models

Proposed by the FDA

- **Model II**
  - Fixed effects:
    Group, Sequence, Treatment, Period(Group), Group×Sequence
  - Random effect:
    Subject(Group×Sequence)
  - The model takes the multigroup nature of the study into account and is more conservative than the naïve pooled model (three degrees of freedom less than Model III).
- **Model III**
  - Fixed effects:
    Sequence, Treatment, Period
  - Random effect:
    Subject(Sequence)
  - This is the common model for 2×2×2 crossover studies.

Statistical Models

Modification for the EEU

- All models could be evaluated with all effects fixed as well, *i.e.*, subjects are treated as fixed instead of random.
  - The decision scheme (*i.e.*, whether data can be pooled or analysis of the largest group is recommended) is applicable as well.

Low sensitivity of the test

- Between subjects factor
  - Testing at the 0.1 level proposed.
  - Can expect a false positive rate in ~10% of studies if there is not *true* G×T interaction.
    - No pooling of data allowed.
    - Substantial drop in power (BE has to demonstrated in the largest group).

Regulatory Practice

FDA

- If all conditions for pooling (2×2×2 model) fulfilled *and* stated in the SAP, acceptable.

EMA

- Implicitly accepts that pooling of groups *cannot* be reasonably assumed to have an effect on the response variable.
  - Hence, only pooling (Model III *without* a justification) applied.
  - In 37 years I came across a *single* case (biosimilar, three groups), where the MHRA required Model II.

MENA-states

- Assessment by the FDA's Model I, II, or III for groups *mandatory* – even if all conditions for pooling are fulfilled.
- Leads to rejection of studies due to false positives.

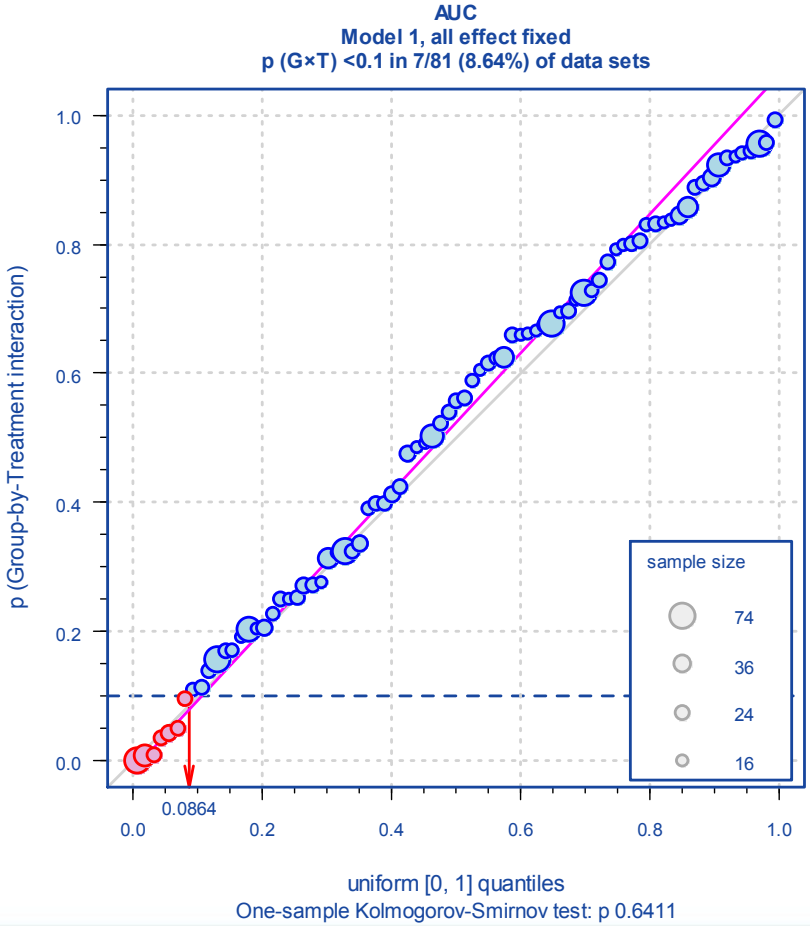
Regulatory Practice

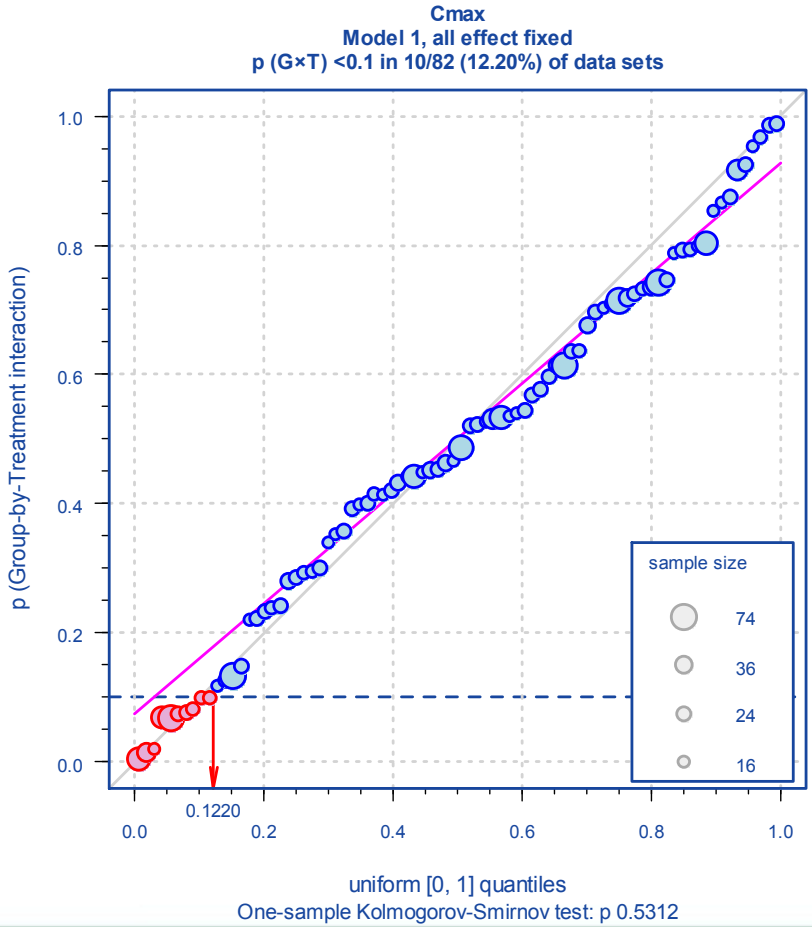
Eurasion Economic Union (GL Sections 93–94)

- If single center within limited time frame:
  - Model II
    - Described in the study report.
    - Remark: • Why not already in the study protocol?
  - Model III (pooled)
    - Justification given in the study report.
    - Remarks: • Why not already in the study protocol?
      • If no justification given, rejection of the study very likely!
      • Regularly justification is not accepted! Why?
- If large interval (*e.g.*, months apart) or different centers:
  - Discuss with the regulatory body in advance.
    - Question: • What are your experiences?

Meta-Analysis

82 studies (58 analytes, sample sizes 15 – 74, 2 – 4 groups, interval between groups 1 – 18 days)





Yes, but …

… is it real?

- In the small meta-analysis significant G×T in ~10% of studies.
  - False positives?
  - No dependency of G×T with interval between groups found.
  - Loss in power compared to naïve pooling: 1.2% (*AUC*) and 4.9% (C_{max}).

Common problems with significance testing

- Significance ≠ relevance.
- Pre-tests (like Grizzle's for sequence / unequal carry-over) are problematic (Freeman 1989).
- The decision to use Model II based on G×T in Model I likely inflates the Type I Error (Biosimilars Forum, Budapest 2017).

Recommendation

- Use Model II *without* a pre-test or give a justification for Model III.

Not for the EMA

Q & A document (EMA 2015)

- In the context of Two-Stage Designs
  - A model which also includes a term for a formulation*stage interaction would give equal weight to the two stages, even if the number of subjects in each stage is very different. The results can be very misleading hence such a model is not considered acceptable. Furthermore, this model assumes that the formulation effect is truly different in each stage. If such an assumption were true there is no single formulation effect that can be applied to the general population, and the estimate from the study has no real meaning.

Splitting

Large studies – limited capacity of the clinical center

- **Suggestions**
  - Find a larger CRO – even if more expensive!
  - If you have to split the estimated sample size into groups:
    - Dose subjects within a limited time frame.
    - 'Staggered approach' prefered, *e.g.*, the groups only days apart.

      Group I : Period 1 (w1 Mo – We) → washout → Period 2 (w2 Mo – We)

      Group II: Period 1 (w1 Th – Sa) → washout → Period 2 (w2 Th – Sa)

    - 'Stacked approach' is suboptimal.

      Group I : Period 1 (w1 Mo – We) → washout → Period 2 (w2 Mo – We)

      Group II: Period 1 (w3 Th – Sa) → washout → Period 2 (w4 Th – Sa)

    - *Do not* split groups into equal sizes!
    - Perform at least one in the maximum capacity of the clinical center.

Splitting

Large studies – limited capacity of the clinical center

- Example
  - *CV* of *AUC* 30% (no scaling allowed), *GMR* 0.90, target power 90%, 4-period full replicate design (reference-scaling of C_{max} intended). Estimated sample size 54.
  - Maximum capacity 24 beds.
    - Option 1: Equal group sizes (3 × 18).
    - Option 2a: Two groups with the maximum size (24), the remaining one 6.
    - Option 2b: One group 24, the remaining ones as balanced as possible (16 | 14).
  - Which one would you prefer – and *why*?
  - Let us assume that there are no dropouts and pooling is not allowed (significant Group-by-Treatment interaction). Expected power:
    - Option 1: 51% in each of the three groups.
    - Option 2a: 62% in the two large groups (n = 24 each).
    - Option 2b: 62% in the largest group.

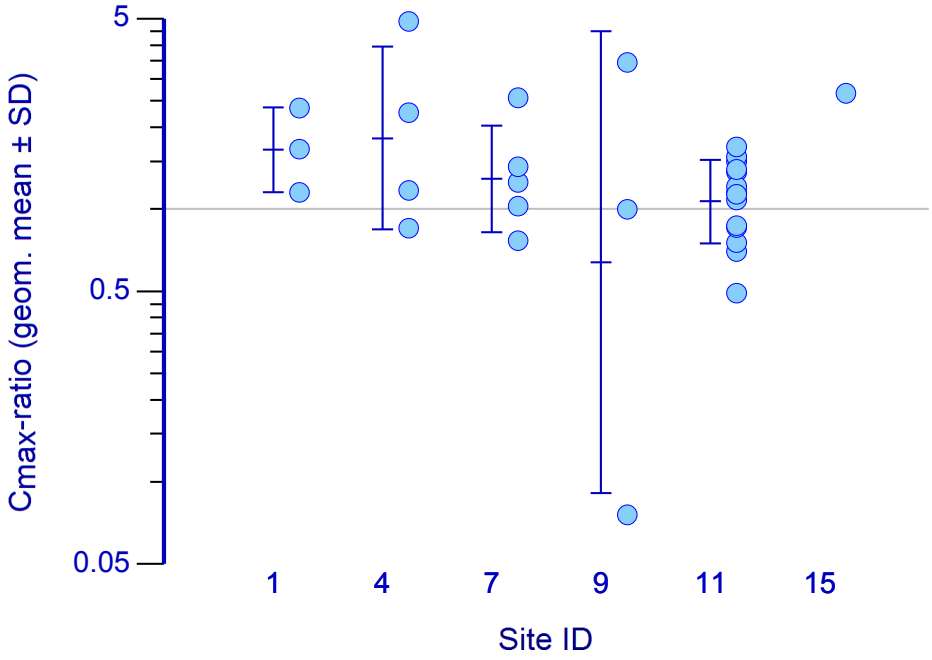
Off Topic: Multi-Center Studies

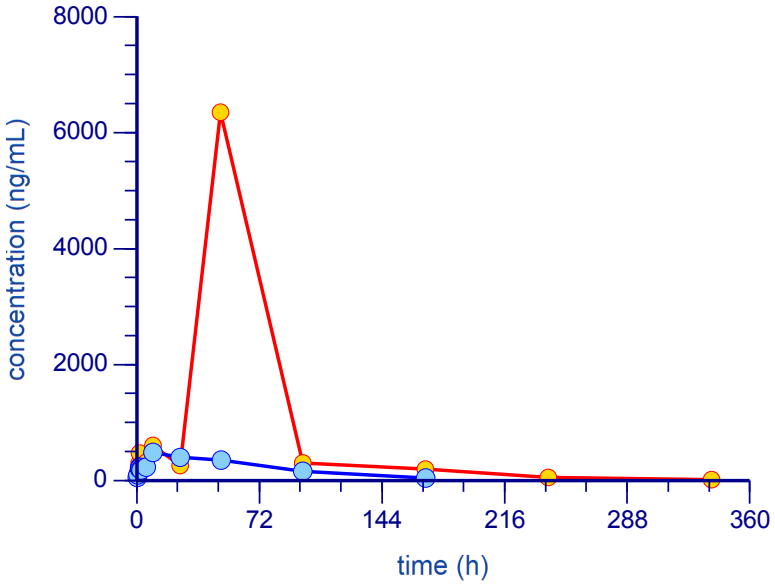
Sometimes (*e.g.*, anti-cancer drugs in patients) multi-center studies cannot be avoided

- Models similar to group-effect models can be used.
  - Replace all group-terms by center-terms.
  - If ever possible do not split centers further into groups.
    - Cave: No commonly accepted statistical model exists.
    - Whatever one statistician proposes might not be accepted by another…
- Make sure that all centers can deliver data of similar quality.
  - Equipment, training of staff, procedures.
  - Sample handling, storage, shipment.
  - Only one bioanalytical laboratory!

Nasty Example

Sloppy handling – even in only 2% of samples – can lead to serious troubles.





Multi-Group Studies in Bioequivalence.
To pool or not to pool?

Thank You!
Open Questions?

Helmut Schütz
BEBAC
Consultancy Services for
Bioequivalence and Bioavailability Studies
1070 Vienna, Austria
helmut.schuetz@bebac.at