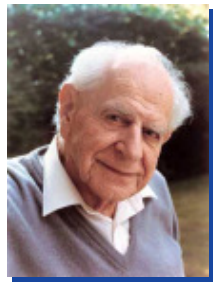
Multi-Group Studies in Bioequivalence.

To pool or not to pool?

Helmut Schütz

Remember…

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 studies are split into ≥2 groups of subjects

- Reasons
  - Limited 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 if products are 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 correct any more.
  - Periods are performed on different dates.
  - Questions may arise whether groups can be naïvely pooled. Valid only if
    - all groups have the same size and
    - GMRs of groups would be similar (no Group-by-Treatment interaction).

Review of Guidelines

FDA (Statistical Approaches Establishing BE, 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, or at the same site but greatly separated in time (months apart, for example), 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 [→ Model I] 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. [→ Model II]
  - 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. [→ Model III]
  - […] 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 […].

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).

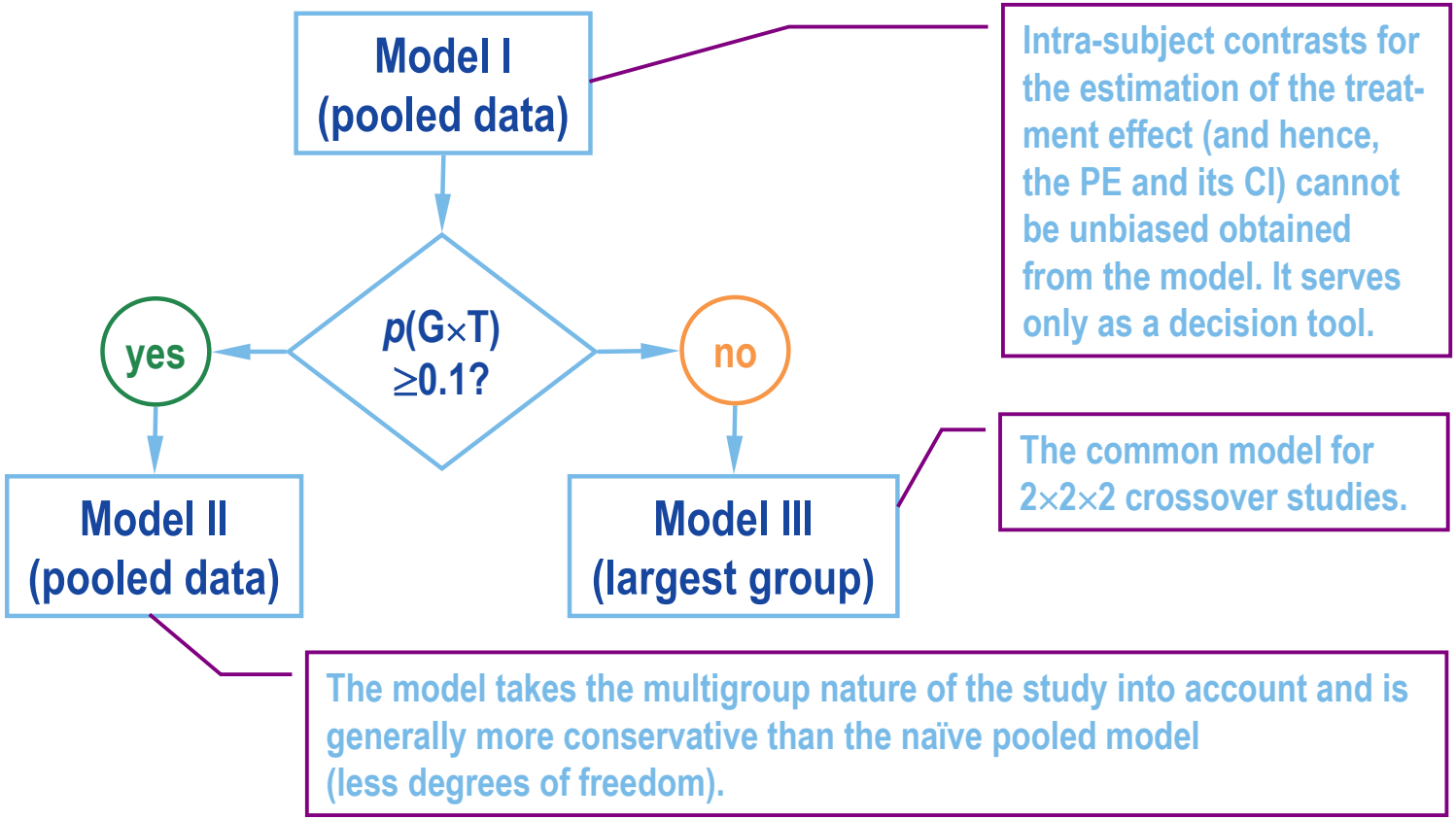
Review of Guidelines

EMA (Guideline on the Investigation of BE, 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.

FDA's Framework

Proposed if criteria for pooling not fulfilled



FDA's Framework

Low sensitivity of the test

- Group-by-treatment interaction is a *between* subjects factor
  - Testing at the 0.1 level proposed.
  - Can expect a false positive rate in ~10% of studies if there is *no true* G×T interaction.
    - No pooling of data allowed.
    - Substantial drop in power, since BE has to be demonstrated in the largest group.
    - Open question: What if large groups have the same size?
      - » Assumption: All should pass BE.

Regulatory Practice

FDA

- If *all* criteria for pooling fulfilled *and* the conventional 2×2×2 model stated in the protocol, acceptable.

EMA

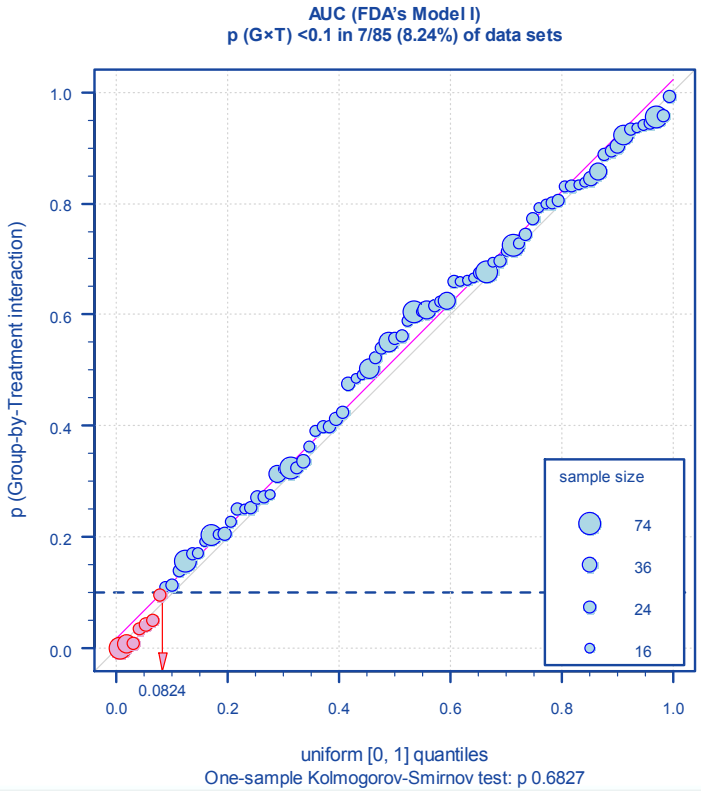
- Implicitly accepts that pooling of groups *cannot* be reasonably assumed to have an effect on the response variable.
  - Hence, only pooling (Model III *without* a justification) applied.
  - In 38 years I came across only two cases where Model II was requested (one multi-group study and one multi-site study).

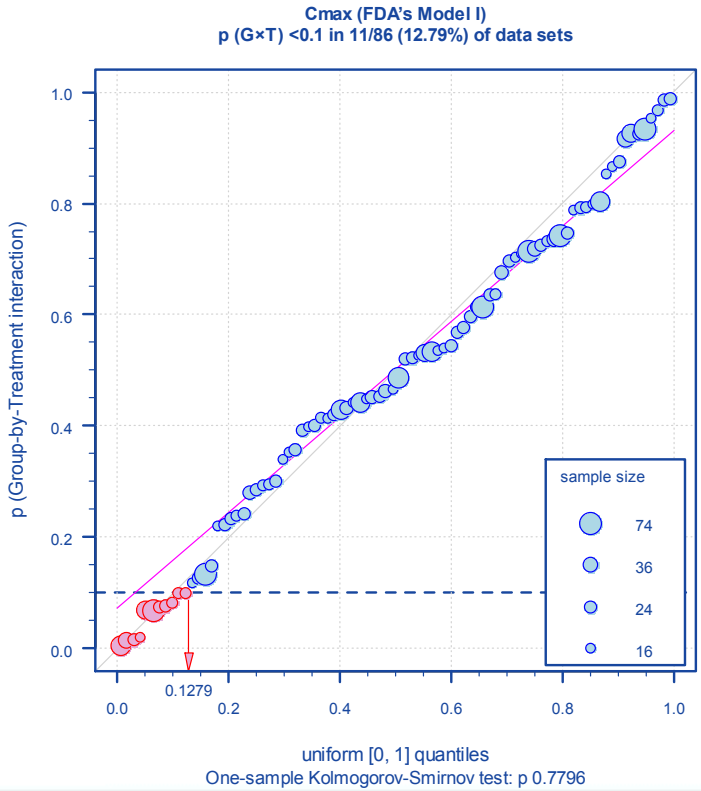
Russia, Eurasian Economic Union, MENA States

- Assessment according to the FDA's framework (Model I → II *or* III]) *preferred* – even if all criteria for pooling are fulfilled?
- Leads to failing studies due to false positives (loss of power).

Small Meta-Analysis

86 studies (60 analytes, sample sizes 15 – 74, 2 – 4 groups, interval between groups 1 – 18 days, median 3 days)





Yes, but …

… is it real?

- In the small meta-analysis significant G×T in ~10% of studies.
  - Close to the false positive rate.
  - No dependency of G×T with interval between groups found.
  - Loss in power compared to naïve pooling: 1.2% (*AUC*) and 5.8% (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 or III based on G×T observed in Model I likely inflates the Type I Error (Biosimilars Forum, Budapest 2017).

Recommendation

- Give a justification for Model III or use Model II *without* a pre-test.

EMA

Q & A document (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. […] 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.

Deficiency letter (EMA 2018) multi-site study

- A sub-set of patients cannot be selected for the BE analysis on the basis of tests for a treatment-by-site interaction. It is questioned whether this approach produces an unbiased estimate as the chosen group may no longer be representative of the initial intended study population.
  The Type I error is not controlled when the procedure […] is only related to the finally selected population/model.
- In a multi-site study Model II is preferred, and that should be used for the BE analysis regardless of the results of any interaction tests.

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.
  - 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).
  - Let us assume that there are no dropouts and pooling is not allowed (significant G×T 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.

Multi-Group Studies in BE.
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

Backup Slide (Power)

100,000 simulated 2×2×2 studies

- No period- and sequence-effects, *GMR* 0.95, two groups n_1:n_2 = 1:1

		low variability		moderate variability	
		C_{max}	*AUC*	C_{max}	*AUC*
Passed BE	Within-subject *CV*	7.5 – 22.5%	5 – 15%	15 – 30%	10 – 20%
	Target power for C_{max} (*n*)	80% (12 – 24)		90% (16 – 52)	
	Model I: *p* (G×T) <0.1	9.86%	10.03%	10.07%	10.03%
	Model III: pooled	89.80%	99.43%	91.64%	99.79%
	Model II: without pre-test	89.54%	99.40%	91.63%	99.78%

- Conclusions
  - Significant G×T-interaction at ~ the level of the test.
  - Loss in power by applying Model II compared to Model III:
    - Very low for small sample sizes.
    - Negligible for moderate sample sizes.

Backup Slide (Power)

100,000 simulated 2×2×2 studies

- No period- and sequence-effects, *GMR* 0.95, two groups n_1:n_2 ~ 3:1

		low variability		moderate variability	
		C_{max}	*AUC*	C_{max}	*AUC*
Passed BE	Within-subject *CV*	7.5 – 22.5%	5 – 15%	15 – 30%	10 – 20%
	Target power for C_{max} (n)	80% (14 – 24)		90% (16 – 52)	
	Model I: p (G×T) <0.1	9.86%	9.87%	10.10%	10.05%
	Model III: pooled	90.02%	99.36%	90.92%	99.76%
	Model II: without pre-test	89.83%	99.33%	90.83%	99.75%

- Conclusions
  - Similar to equal group sizes.

Backup Slide (Type I Error)

1,000,000 simulated 2×2×2 studies

- FDA's framework, *GMR* 1.25 (H_0), two groups n_1:n_2 ~ 1:1

		low variability		moderate variability	
		C_{max}	*AUC*	C_{max}	*AUC*
	Within-subject *CV*	7.5 – 22.5%	5 – 15%	15 – 30%	10 – 20%
	Target power for C_{max} (n)	80% (14 – 24)		90% (16 – 52)	
	Model I: p (G×T) <0.1	10.00%	9.97%	9.98%	9.97%
Empiric Type I Error	Model III:[1]	1.29%	1.28%	1.09%	1.07%
	Model II:[2]	4.48%	4.54%	4.50%	4.50%
	Aggegate (III and II)	5.77%	5.81%	5.59%	5.70%

- Conclusions
  - Significant inflation of the Type I Error.

[1] If p (G×T) <0.1 in Model I. Largest group or must pass both if groups are equally sized.
[2] If p (G×T) ≥0.1 in Model I. Pooled data.

Backup Slide (Type I Error)

1,000,000 simulated 2×2×2 studies

- FDA's framework, *GMR* 1.25 (H_0), two groups n_1:n_2 ~ 3:1

		low variability		moderate variability	
		C_{max}	*AUC*	C_{max}	*AUC*
	Within-subject *CV*	7.5 – 22.5%	5 – 15%	15 – 30%	10 – 20%
	Target power for C_{max} (n)	80% (14 – 24)		90% (16 – 52)	
	Model I: p (G×T) <0.1	9.98%	9.98%	9.99%	9.95%
Empiric Type I Error	Model III:[1]	1.18%	1.18%	1.20%	1.17%
	Model II:[2]	4.47%	4.55%	4.54%	4.52%
	Aggegate (III and II)	5.65%	5.73%	5.74%	5.69%

- Conclusions
  - Significant inflation of the Type I Error.

[1] If p (G×T) <0.1 in Model I. Largest group or must pass both if groups are equally sized.
[2] If p (G×T) ≥0.1 in Model I. Pooled data.