Two-Stage Sequential Designs
Regulatory Perspective

Helmut Schütz

Disclaimer

I am not a regulator.

- I will give an overview of *my personal regulatory experiences* with sequential designs.
- However, the regulatory views are not unfounded.
  - I submitted my first TSDs (Gould's approach) to the German and French authorities in 1995. Protocols were not accepted.
  - I performed independent calculation of the *CV* in the interim (suggested by the BfArM's Joachim Röhmel) within 1995 and 2004 (successfully).
  - First of my protocols in a TSD (Potvin 'Method C') accepted by the BfArM in 2009. Study accepted in 2011. ~Ten others ever since.
  - Personal acquaintance with most of the members of the PKWP, some of the BSWP, and members of national authorities. More than 150 e-mails on the topic…
- Neither BEBAC nor myself make any representations or warranties regarding the accuracy and applicability of the content.

(Adaptive) Sequential Two-Stage Designs

Review of Guidelines.

- EMA (BE Draft Jul 2008; lines 563 – 572)
  - "[…] appropriate steps must be taken to preserve the overall type I error of the experiment."
  - "[…] both analyses conducted at adjusted significance levels (with the confidence intervals accordingly using an adjusted coverage probability which will be higher than 90%)."
- EMA (Overview of Comments on the BE Draft, Jan 2010; p. 158 – 160)
  - "It is considered that the penalty can vary by applicant's choice. Many approaches are valid. But an example will be included."
  - "A term for stage should be included in the analysis. […] This is to be treated like e.g. period effect. Important to include in the model, but the size of effect not important."

(Adaptive) Sequential Two-Stage Designs

Review of Guidelines.

- EMA (Overview of Comments on the BE Draft, Jan 2010)
  - "[…] the example chosen was the Pocock approach – very similar to the example in the flow chart* included in the comment."
    * The flow chart was Potvin's 'Method C'.
  - "In practice the company could essentially do this* under the current draft, by specifying an extreme alpha level at the first analysis, thereby taking almost no penalty. We prefer to stick to having some alpha penalty for any interim analysis, especially as it can be difficult to establish whether data are truly blind in a bioequivalence trial."
    * Blinded interim analysis according to Wittes et al. (1999) and Schwartz/Denne (2003).

(Adaptive) Sequential Two-Stage Designs

Review of Guidelines.

- EMA (BE GL, Jan 2010; p. 15 – 16)
  - "For example, using 94.12% confidence intervals* for both the analysis of stage 1 and the combined data from stage 1 and stage 2 would be acceptable, […]"
    * 94.12% CI = α_{adj} 0.0294.
    According to the Comments on the Draft: "Pocock's approach" – *not* Potvin's 'Method B'!
  - "[…] there are many acceptable alternatives and the choice of how much alpha to spend at the interim analysis is at the company's discretion."
  - "[…] pre-specified […] adjusted significance levels to be used for each of the analyses."

(Adaptive) Sequential Two-Stage Designs

Review of Guidelines.

- EMA Q&A Document Rev. 7 (Feb 2013)
  - Background:
    Concerns by the PKWP about the poolability of data and reliability of the estimated variance in the final analysis.
    - A term for the stage should be included in the ANOVA model. However, the guideline does not clarify what the consequence should be if it is statistically significant. In principle, the data sets of both stages could not be combined.

      Although the guideline is not explicit, even if the final sample size is going to be decided based on the intra-subject variability estimated in the interim analysis, a proposal for a final sample size must be included in the protocol so that a significant number of subjects (e.g., 12) is added to the interim sample size to avoid looking twice at almost identical samples. This proposed final sample size should be recruited even if the estimation obtained from the interim analysis is lower than the one pre-defined in the protocol in order to maintain the consumer risk. García-Arieta/Gordon 2012

(Adaptive) Sequential Two-Stage Designs

Review of Guidelines.

- EMA Q&A Document Rev. 7 (Feb 2013)
  - Background:
    - Agreement about an additional term in the model.
    - No agreement about a fixed final sample size and a minimum n_2 of 12.
      At least two subjects as a comprise?
  - From the perspective of TIE control it is considered that there is no minimal number of subjects to be included in the second stage of a TSD, so long as it can be demonstrated that the TIE of the study is controlled.
  - To account for the fact that the periods in the 1st stage are different from the periods in the 2nd stage, a term for period within stage is required. […], the […] ANOVA model for analysis of the combined data from a TSD would have the following terms:
    stage, sequence, sequence × stage, subject(sequence × stage), period(stage), formulation.

(Adaptive) Sequential Two-Stage Designs

Review of Guidelines.

- EMA Q&A Document Rev. 7 (Feb 2013)
  - To fit this model it is necessary to have in each stage at least one patient in each sequence – so a minimum of two patients in each stage of the study, but more if both happen to be randomised to the same sequence.
  - This does not supersede the requirement for at least 12 subjects overall.

Case Study 1

Potvin 'Method C' (2010 – 2011).

- Study stopped in stage 1
  - AUC: power >80%; passed BE with 90% CI.
  - C_{max} : power <80%; passed BE with 94.12% CI.
- NL: Adapting the confidence intervals based upon power is not acceptable and also not in accordance with the EMA guideline.* Confidence intervals should be selected *a priori*, without evaluation of the power. Therefore, the applicant should submit the 94.12% confidence intervals for AUC.

  * What about: "… choice of how much alpha to spend at the interim analysis is at the company's discretion."?
  - Failed to show BE of AUC with 94.12% CI.
  - Study repeated in India in a very (!) large fixed sample design.
  - Failed on C_{max}. Project cancelled.

Case Study 2

Potvin 'Method C' (2011 – 2012).

- Study passed already in stage 1
  - *CV* in the interim 30.65%, n_1 49.
  - 90% CI since power was 87.3%.
- UK, IE: Unadjusted α in stage 1 not acceptable.
  - Study passed with 94.12% CI as well (*post hoc* switch to 'Method B').
- AT: The Applicant should demonstrate that the type I error inflation, which can be expected from the chosen approach, did not impact on the decision of bioequivalence.*

  * Unofficial information: Potvin's table contains only a cell for *CV* 30% and n_1 48…
  - One million studies simulated based on the study's *CV* and n_1.
  - Empiric Type I Error 0.0494 (95% CI: 0.0490 – 0.0498).

Case Study 3

Potvin 'Method C' (2011 – 2013).

- Two studies (SD, MD) passed in stage 1; data for C_{max}
  - SD: *CV* 17.7%, n_1 15.
  - MD: *CV* 8.54%, n_1 16.
    - The SD study was performed in two groups. The fixed effects were sequence, period, treatment, group, group × treatment
    - The MD study was performed in one group. The fixed effects were sequence, period, treatment
    - In both studies REML was performed in Phoenix/WinNonlin with subject(sequence) as a random effect.
  - 90% CIs since power was >80% (would have passed with 94.12% CI as well but not reported *for educational rasons*).
- Accepted by DE (RMS), AT, DK, S, NL (CMSs) without comments.
- ES: Statistical analysis should be GLM. Please justify.
  - Studies passed with fixed-effects model according to Q&A Ref. 7.

Case Study 4

Potvin 'Method C' (2012 – 2013).

- Protocol synopsis with statistical details submitted to the Spanish Agency (2012).
  - Unofficial feedback (after consultation of AEMPS with the BSWP):
    - "Potvin's method is not valid in Europe."
- Question to the Spanish Agency (2013):
  […] we'd like to ask about the current status of TSD BE study, […] if the BE protocol with Potvin's Method C is acceptable now […].
  - Answer:
    - "Potvin's methods are not acceptable in EMA."

Rumors & Chinese Whispers (Part 1)

TSDs based on simulations

- One member of the PKWP (2015):
  - I made peace with these methods and accept studies – *if* the confidence interval is not *too* close* to the acceptance limits.
    * Remark: *How* close is "not *too* close"?
- Assessors of ES, AT (2016):
  - Kieser/Rauch (2015) showed that the adjusted α_{adj} 0.0294 used by Potvin et al. is Pocock's for *superiority*.
    The correct value for *equivalence* is 0.0304 (Jennison/Turnbull 1999).
  - Hence, all studies evaluated with a 94.12% CI in both stages are more conservative than necessary. At least these studies should not be problematic.
    - Remarks:
      One could confirm ~0.0304 for 'Method B' in simulations.
      However, it is a misconception that 0.0304 is "universally valid" for equivalence.
      Other settings (GMR, power) require *other* values – even for 'Type 1' TSDs.

Rumors & Chinese Whispers (Part 1)

TSDs based on simulations

- Another member of the PKWP asked the BSWP *which* inflation of the Type I Error would be acceptable (2015). He gave 0.0501 as an example.
  - Answer: The TIE must not exceed 0.05.
    - Remark: Rounding of the CI as required by the GL leads to acceptance of studies (regardless the design) with CLs of 79.995% and/or 125.004% – which inflates the TIE up to 0.0508. The BSWP should mind its own business.
- One assessor (PT) saw a study rejected by one of his colleagues – although BE was shown (2016).
  - When asked why, the answer was:
    - According to the BSWP Potvin's methods are not acceptable.
  - He was not aware of such a statement and asked for an official document.
    - Such a document does not exist but all statisticians in the agencies know this statement.

Rumors & Chinese Whispers (Part 1)

TSDs based on simulations

- Scientific Advice in SE (2016).
  - Simulations based on Fuglsang's 'Type 1' TSD for Parallel Groups (2014).
  - Large n_1 (up to 125/group), homo- and heterogenous variances, potentially unequal group sizes due to drop-outs.
  - With α_{adj} 0.0274 the maximum Type I Error was 0.04992.
  - Response:
    - According to the guideline, application of a TSD was accepted provided that the patient's risk is maintained at or below 5%.
    - Confirmed that the statement about Potvin's methods is not public. These types of TSDs are not proven in a strict sense.
    - However, it was acknowledged that the simulations covered a sufficient range of possible outcomes (unequal variances and drop-out rates).
    - […] the empiric type I error rate should be evaluated with the real data (*i.e.*, the actual group sizes and variances of the study).

The Assessor's Dilemma

TSDs based on simulations

- If an assessor would like to accept TSDs he/she is facing a dilemma:
  - TSDs are stated in the GL and therefore, studies are submitted.
  - The BSWP does not "like" methods based on simulations and prefers methods which demonstrate by an analytical proof that the patient's risk is preserved – which seemingly don't exist.
  - According to the BSWP even a TIE of 0.0501 is not acceptable.
  - With one million simulations the significance limit (>0.05) is 0.05036.
    - Most methods show a TIE below this limit (and some even <0.05).
    - However, with other seeds of the random number generator (slightly) different results are possible.
  - It would be desirable to assess whether a passing study (with a CI close to the AR) has a *relevant* impact on the patient's risk.
- I coded a package in R (AdaptiveBE), which currently is evaluated by assessors in Portugal and Spain.

Package AdaptiveBE

Function check.TSD()

- Required:
  - Interim data (*CV* or *MSE*, n_1, PE or CI), data of the final analysis (*CV* or *MSE*, *N*, PE or CI), adjusted alpha(s), the type of the TSD (optionally futility rules).
  - Alternatively (*i.e.*, if not given in the report) the CIs can be used to calculate the CVs and/or the PEs.
- Algorithm:
  - Based on the interim data and the study's framework simulate 1 mio studies in order to obtain the empiric Type I Error.
    - If the TIE $\leq$0.05, stop. Can accept the applicant's results.
    - If not, optimize α_{adj} with a target TIE of 0.05. Recalculate the study (interim – and optionally – final) and compare conclusions with the reported ones.
      - » If conclusions agree, accept the study (increase of the TIE not *relevant*).
      - » If not (reported passes and adjusted fails), calculate the increase of relative risk. Whether the study is accepted or not lies in the hands of the assessor.

Package AdaptiveBE

Function check.TSD()

- Example 2 of Potvin's 'Method C'
  - The maximum TIE in Table I of the paper is 0.0510 for *CV* 20%, n_1 12.
  - I used the reported *MSE*s and sample sizes. The *CV* in the interim was with 18.21% close to the location of the maximum TIE.
  - The power-calculation was done by the shifted *t*-distribution like in the paper.
- R-code:

```
library(AdaptiveBE)
check.TSD(Var1=c(0.032634, "MSE"), PE1=c(0.083960, "log"), n1=12,
          Var =c(0.045896, "MSE"), PE =c(0.014439, "log"), N =20,
          alpha0=0.05, alpha1=0.0294, alpha2=0.0294,
          type=2, GMR=0.95, pmethod="shifted")
```

Package AdaptiveBE

Function check.TSD()

- **Part of the output:**

```
TIE for specified α: 0.05048 (>0.05)
                     Applied adjustment is not justified.
Final analysis of pooled data (specified α2 0.0294)
═══════════════════════════════════════════════════════

94.12% CI: 88.45–116.38% (BE concluded)

Adjusted α 1, 2    : 0.050|0.02855, 0.02855
Adjusted CIs       : 90.00%|94.29%, 94.29%
TIE for adjusted α : 0.04994 (n.s. >0.05)
Final analysis of pooled data (adjusted α2 0.02855)
═══════════════════════════════════════════════════════

94.29% CI: 88.35–116.50% (BE concluded)
```

```
┌──────────────────────────────────────────┐
│ Since conclusions of both analyses agree, │
│ can accept the original analysis.         │
└──────────────────────────────────────────┘
```

Package AdaptiveBE

Function check.TSD()

- It was difficult to fabricate an example where the original evaluation would pass and the optimized fail, *i.e.*,
  a borderline case where the CI was *"too close"* to the AR.
  - The maximum TIE reported in any of the publications is 0.0518 (Montague's 'Method D', *CV* 20%, n_1 12).
  - I used the interim *CV* and n_1, a PE_1 of 0.92, and in the final analysis a higher *CV* (22.3%), a worse PE (0.88), and one drop-out in the second stage (*N* 45).
  - The power-calculation was done by the shifted *t*-distribution like in the paper.
- R-code:

```
library(AdaptiveBE)
check.TSD(Var1=c(0.200, "CV"), PE1=c(0.92, "lin"), n1=12,
          Var =c(0.233, "CV"), PE =c(0.88, "lin"), N =45,
          alpha0=0.05, alpha1=0.028, alpha2=0.028,
          type=2, GMR=0.90, pmethod="shifted")
```

Package AdaptiveBE

Function check.TSD()

- **Part of the output:**

```
TIE for specified α: 0.05173 (>0.05)
                     Applied adjustment is not justified.
Final analysis of pooled data (specified α2 0.028)
═══════════════════════════════════════════════════════
94.40% CI: 80.00-96.80% (BE concluded)

Adjusted α 1, 2    : 0.050|0.02696, 0.02696
Adjusted CIs       : 90.00%|94.61%, 94.61%
TIE for adjusted α : 0.05001 (n.s. >0.05)
Final analysis of pooled data (adjusted α2 0.02696)
═══════════════════════════════════════════════════════
94.61% CI: 79.93-96.88% (failed to demonstrate BE)
```

```
┌──────────────────────────────────────────┐
│ Accepting the reported analysis could    │
│ increase the relative consumer risk by ~3.5%. │
└──────────────────────────────────────────┘
```

Rumors & Chinese Whispers (Part 2)

Simulations *vs.* "analytical proof"

- In principle regulators prefer methods where the control of the TIE can be shown analytically.
  - Promising zone approach (Mehta/Pocock 2011).
    Wrong: Superiority / parallel groups / equal variances.
    Critized by Emerson et al. (2011).
  - Inverse normal method (Kieser/Rauch 2015).
    Wrong: Not a proof but a claim. *Slight* inflation of the TIE (0.05026) in the supplementary material's simulations.
  - Repeated confidence intervals (Bretz et al. 2009). Adapted for bioequivalence (König et al. 2014, 2015).
    Correct. But only two posters about BE so far (not published in a peer-reviewed journal).
- Either there is a proof (but *not* for the conditions in BE) or it is not published yet.

Rumors & Chinese Whispers (Part 2)

Simulations *vs.* "analytical proof"

- Summer Symposium *'To New Shores in Drug Development Implementing Statistical Innovation'*, Vienna, 27 Juni 2016
  - Most proofs start with …

    "Let us assume parallel groups of equal sizes and normal distributed data with means of 0 and variances of 1"

    … followed by some fancy formulas.

    Do these cases ever occur in *reality*? Peter Bauer

Two-Stage Sequential Designs
Regulatory Perspective

Thank You!

Open Questions?

Helmut Schütz

BEBAC

Consultancy Services for
Bioequivalence and Bioavailability Studies
1070 Vienna, Austria

helmut.schuetz@bebac.at

References

Pocock SJ. *Group sequential methods in the design and analysis of clinical trials.* Biometrika. 1977;64:191–9.

Gould LA. *Group sequential extension of a standard bioequivalence testing procedure.* J Pharmacokinet Biopharm. 1995;23:57–86. DOI 10.1007/BF02353786

Haybittle JL. *Repeated assessment of results in clinical trials of cancer treatment.* Br J Radiol. 1971;44:793–7. DOI 10.1259/0007-1285-44-526-793

Peto R et al. *Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. analysis and examples.* Br J Cancer. 1977;35:2–39. DOI 10.1038/bjc.1977.1

O'Brien PC, Fleming TR. *A multiple testing procedure for clinical trials.* Biometrics. 1979;35:549–56.

Lan KG, DeMets DL. *Discrete sequential boundaries for clinical trials.* Biometrika. 1983;70:659–63.

Jennison C, Turnbull BW. *Equivalence tests.* In: Jennison C, Turnbull BW, editors. *Group sequential methods with applications to clinical trials.* Boca Raton: Chapman & Hall/CRC; 1999. p. 142–57.

Wittes J et al. *Internal pilot studies I: type I error rate of the naive t-test.* Stat Med. 1999;18(24):3481–91. DOI 10.1002/(SICI)1097-0258(19991230)18:24<3481::AID-SIM301>3.0.CO;2-C

Schwartz TA, Denne JS. *Common threads between sample size recalculation and group sequential procedures.* Pharmaceut Statist. 2003;2:263–71. DOI 10.1002/pst.068

Potvin D et al. *Sequential design approaches for bioequivalence studies with crossover designs.* Pharmaceut Statist. 2008;7(4):245–62. DOI 10.1002/pst.294

Bretz F, König F, Brannath W, Glimm E, Posch M. *Tutorial in biostatistics: Adaptive designs for confirmatory clinical trials.* Stat Med. 2009;28(8):1181–217. DOI 10.1002/sim.3538

Mehta CR, Pocock SJ. *Adaptive increase in sample size when interim results are promising: a practical guide with examples.* Stat Med. 2011;30(28):3267–84. DOI 10.1002/sim.4102

Emerson SS, Levin GP, Emerson SC. *Comments on 'Adaptive Increase in Sample Size when Interim Results are Promising: A Practical Guide with Examples'.* Stat Med. 2011;30(28):3285–301. DOI 10.1002/sim.4271

Montague TH et al. *Additional results for 'Sequential design approaches for bioequivalence studies with crossover designs'.* Pharmaceut Statist. 2012;11(1):8–13. DOI 10.1002/pst.483

García-Arieta A, Gordon J. *Bioequivalence Requirements in the European Union: Critical Discussion.* AAPS J. 2012;14(4):738–48. DOI 10.1208/s12248-012-9382-1

Davit B et al. *Guidelines for Bioequivalence of Systemically Available Orally Administered Generic Drug Products: A Survey of Similarities and Differences.* AAPS J. 2013;15(4):974–90. DOI 10.1208/s12248-013-9499-x

Golkowski D, Friede T, Kieser M. *Blinded sample size reestimation in crossover bioequivalence trials.* Pharmaceut Stat. 2014;13(3):157–62. DOI 10.1002/pst.1617

Schütz H. *Two-stage designs in bioequivalence trials.* Eur J Clin Pharmacol. 2015;71(3):271–81. DOI 10.1007/s00228-015-1806-2

Kieser M, Rauch G. *Two-stage designs for crossover bioequivalence trials.* Stat Med. 2015;34(16):2403–16. DOI 10.1002/sim.6487

König F, Wolfsegger M, Jaki T, Schütz H, Wasmer G. *Adaptive two-stage bioequivalence trials with early stopping and sample size re-estimation.* 35th Annual Conference of the International Society for Clinical Biostatistics. Vienna: August 2014. DOI 10.13140/RG.2.1.5190.0967

König F, Wolfsegger M, Jaki T, Schütz H, Wasmer G. *Adaptive two-stage bioequivalence trials with early stopping and sample size re-estimation.* Trials. 2015;16(Suppl 2):P218. DOI 10.1186/1745-6215-16-S2-P218

Labes D, Schütz H. *Power2Stage: Power and Sample-Size Distribution of 2-Stage Bioequivalence Studies.* R package version 0.4-3. 2015. https://cran.r-project.org/package=Power2Stage