Pitfalls in Bioequivalence

If anything can go wrong, it will. Edward A. Murphy Jr.

Study Design

- In a crossover-study the washout between treatments has to be sufficiently long
  - Pre-dose concentrations which are residuals of previous period(s) have to be avoided
  - In order to get an unbiased estimate of treatment differences the physiological state of subjects in higher period(s) has to be the same as in the (drug-naïve) first period
    - Washout (generally ≥5times the apparent half life) must not be based on an average. The distribution of half lives should be kept in mind; some subjects might show a substantially longer half life – especially if the drug is subjected to polymorphism (poor and extensive metabolizers)
    - Don't forget pharmacodynamics. If the drug is an auto-inducer (*e.g.*, coumarins) or -inhibitor (*e.g.*, imatinib) the body has to return to its original state before the next dose.

Study Design

- Drug A: $t_{½}$ 60 – 100 h (literature)
  - BA study
    - 10 mg drug A hydrochloride p.o. *vs*. i.v.
    - 12 subjects
    - 2×2×2 crossover, washout 35 days
    - Sampling until 312 hours post dose
    - LC/MS-MS, LLOQ 1 ng/mL (drug A base / plasma)
    - Results considered important for designing other studies
      - $t_{½}$ 49.9 ± 13.0 h (harmonic mean ± jackknife standard deviation)
      - In none of the samples drawn at 312 h a concentration ≥LLOQ was measured
      - Extrapolated *AUC* 10.0% (median)
        3.8% – 13.9% (minimum – maximum)

Study Design

- Drug A: $t_{½}$ 60 – 100 h (literature)
  - Comparative BA study aiming to demonstrate BE
    - 10 mg drug A hydrochloride
      (primary target T_2 *vs*. R, descriptive T_2 vs. T_1)
    - 36 subjects
    - 3×6×3 crossover (Williams' design), washout 14 days
      - Washout planned for a worst case $t_{½}$ of 66 h (covering >5 half lives)
    - Sampling until 216 hours post dose
      - No problems with extrapolated *AUC* expected (simulations)
    - GC/MS, LLOQ 0.117 ng/mL (drug A base / plasma)
  - Given that, can you imagine *what* happened – and *why*?

Study Design

- Pre-dose concentrations ≥LLOQ: n (% of subjects, geom. means)
  - Period 1: all <LLOQ
  - Period 2: 21 (58%, 0.226 ng/mL)
  - Period 3: 18 (50%, 0.222 ng/mL)
- Half lives (harmonic means)
  - Period 1: 51.68 h
  - Period 2: 54.20 h
  - Period 3: 63.03 h

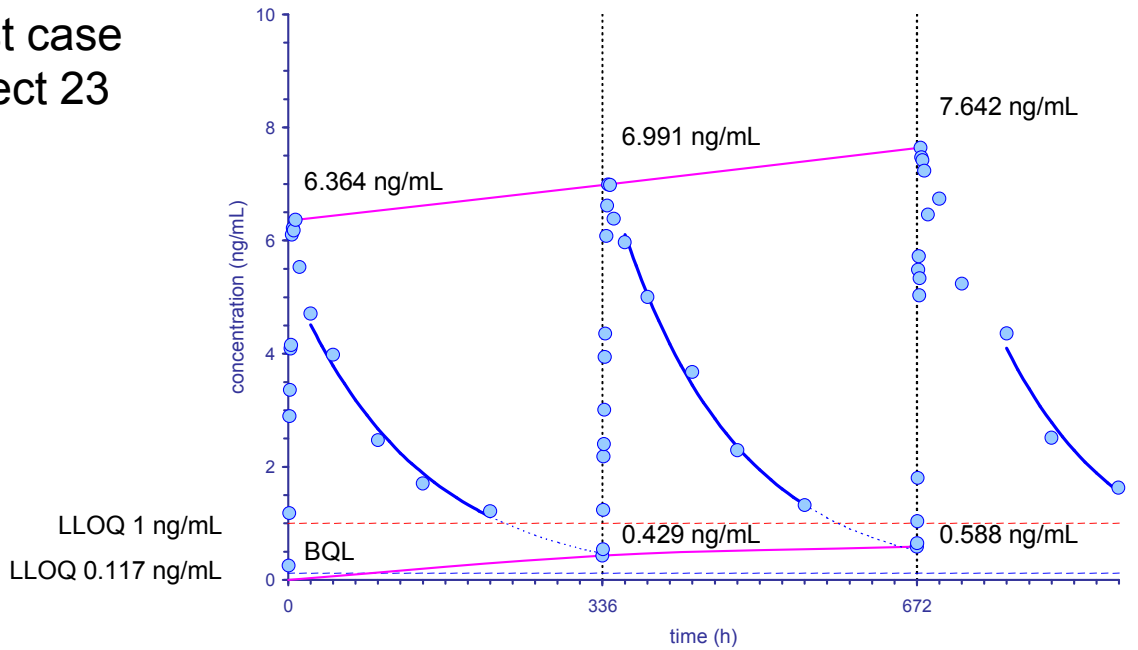
– Issues

- Improving the bioanalytical method (~9times lower LLOQ) was not a good idea
  - If we would have used the previous method we would have measured not a single (!) pre-dose concentration >LLOQ
- Shorter washout (35 days → 14) was not a good idea as well
  - Only if the estimation of λ_z is performed *blinded for* treatment different half lives in the periods (due to accumulation) become evident – even with the less sensitive method

Study Design

- Most statisticians unblind studies *before* performing NCA, which would cover potential problems
  - Half lives (harmonic means)
    - » T_1: 54.51 h
    - » T_2: 55.99 h
    - » R: 56.73 h

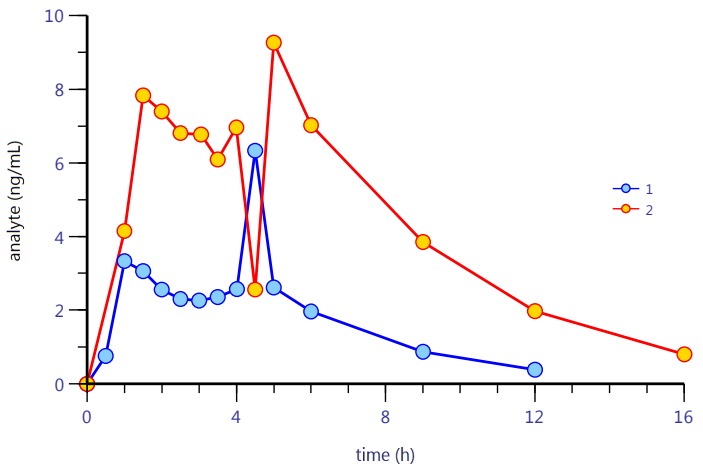
    } similar
- Worst case Subject 23

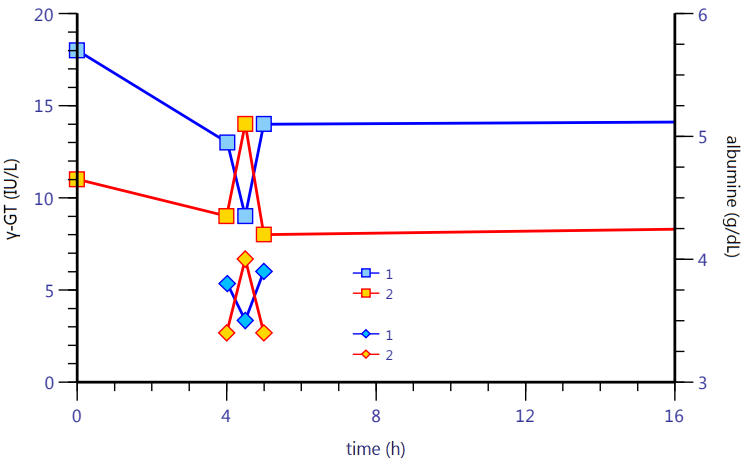


Sample handling

- Clinical phase
  - Drug B: Biphasic modified release product, pilot study
  - Suspected mix-up in the transfer from sample vials after centrifugation to (plasma) sample vials

Measurable values in clin. chemistry (limited, since anticoagulant citrate)

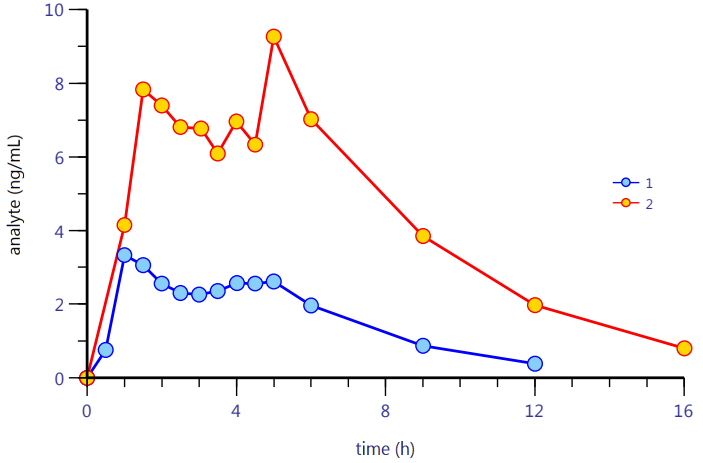


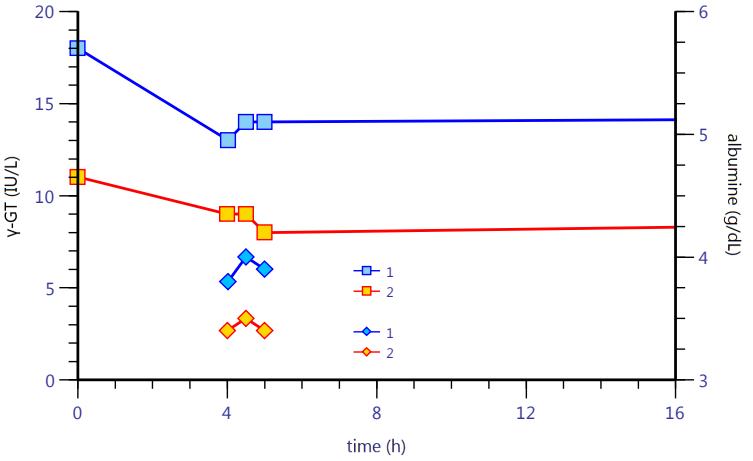


Sample handling

- Clinical phase
  - Drug B: Biphasic modified release product, pilot study
  - Exploratory: Values swapped (analyte and clin. chemistry)
  - Samples of subjects 1 & 2 both taken in the first period

    Suspected mix-up likely due to clin. chemistry values





Sample handling

- Clinical phase
  - Barcode system failed in the first period
  - No bail-out procedure (*e.g.*, four-eye principle)
  - Sponsor monitored plasma separation only up to two hours (when the barcode system was still operable)
  - Blinded review of data for irregular profiles
    - EMA BMV GL (2011)
      - Exclusion only possible if error documented
      - Measurements are ‘carved from stone’ (not even confirmatory reanalysis is acceptable)
      - Reanalysis of pre-dose samples if >LLOQ acceptable (why?)
    - FDA Rev.1 (Sep 2013)
      - Exclusion after repeated analysis acceptable if defined by SOP
    - FDA Draft (May 2018), ICH M10 Draft (Feb 2019)
      - Like EMA, not acceptable

Sample handling

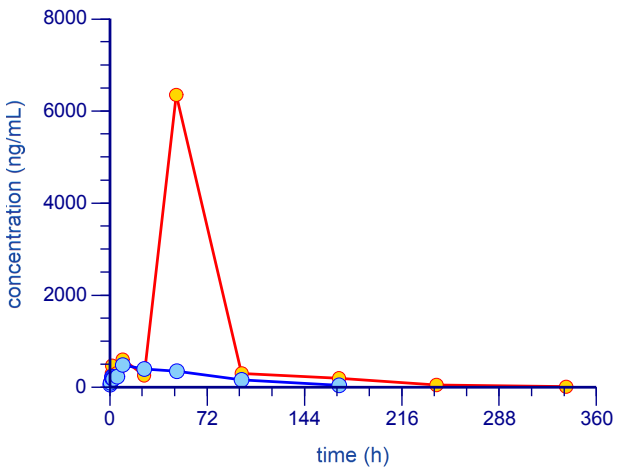
- Clinical phase
  - Drug C: Liposome encapsulated for infusion
  - Analytes
    - Encapsulated drug
    - Unencapsulated drug (*i.e.*, released from liposomes)
    - Total drug (encapsulated + unencapsulated)
    - Metabolite (formed from unencapsulated drug only)
  - Drug may be released from liposomes by
    - shear forces (infusion pump, needle with narrow diameter)
    - high temperature and extended interval until centrifugation
    - high *g* force in centrifugation
    - Only the latter two can be prevented
      - blood samples on ice, ≤ 45 minutes until centrifugation
      - stabilization by DMSO

Sample handling

- Clinical phase
  - Multi-site study in terminal cancer patients
  - Clinical staff trained about critical sample handling but
    - unfamililar procedure esp. in small sites
    - necessity of following SOPs and documentation of deviations in conformity with GCP not well understood
    - well-being of patients considered by clinical staff of oncology departments of higher priority than ‘annoying paperwork’

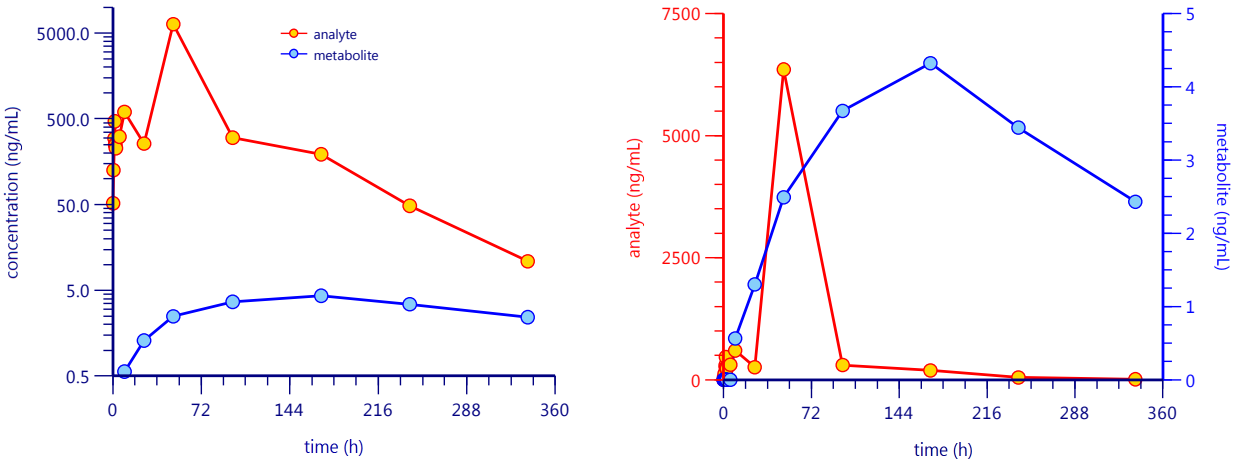
Sample handling

- Clinical phase
  - Surprises in bioanalytics
  - Extremely high concentrations of unencapsulated drug C observed in about 2% of samples
    - All suspect values confirmed in repeated analyses (against GLs!)



Sample handling

- Clinical phase
  - Extremely high concentrations of unencapsulated drug C observed in about 2% of samples
    - However, ‘normal’ concentrations of the metabolite
      - Since the metabolite can only be formed from the unencapsulated drug, the analyte’s high concentrations were considered an artifact
      - No documented improper sample handling (stabilization, temperature & time until centrifugation)



NCA

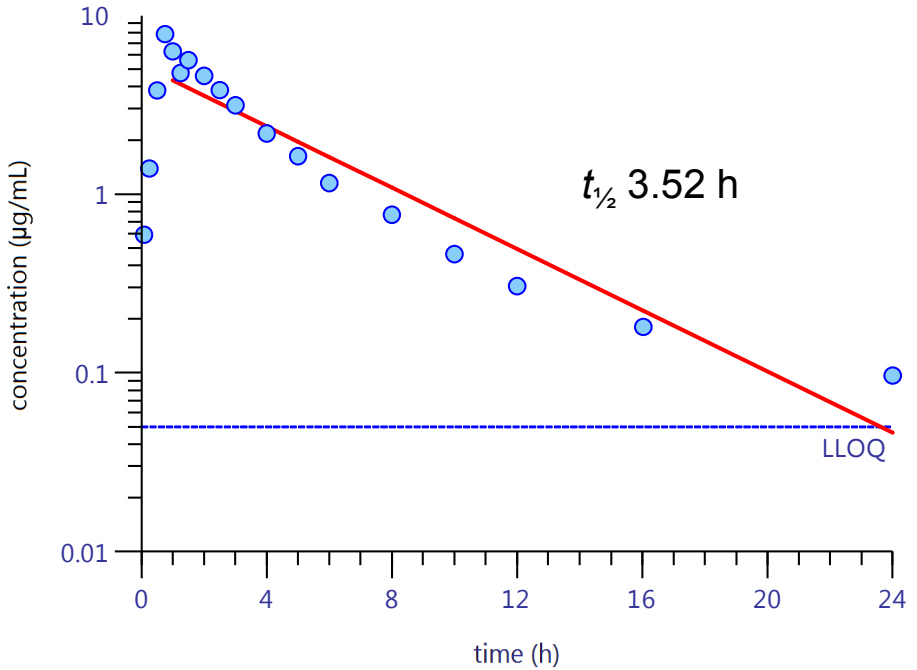
- Requirements for BA/BE studies
  - Bioanalytical method developed and validated *for the intended use*
    - Calibration range
      - LLOQ ≤5% C_{max} in any of the subjects
      - ULOQ ideally ≥C_{max} in any of the subjects
    - (In)accuracy and (im)precision
      - 15% throughout the range (20% for ligand-binding assays)
      - 20% at the LLOQ (30% for ligand-binding assays)
  - Sampling long enough to obtain reliable estimates of
    - λ_z : at least three samples in the log/linear part
    - AUC_{0-t} : covering ≥80% of $AUC_{0-\infty}$ in ≥80% of observations
    - Both are *not required* if target metric is AUC_{0-72} (IR single dose) or $AUC_{0-\tau}$ (steady state)

NCA

- Drug D: $t_{½}$ 2 – 3 h (literature)
  - BE study (500 mg D component of a three-drug FDC)
    - liquid formulations, T *vs*. R
    - 27 subjects
    - TRR|RTR|RRT partial replicate design, washout seven days
    - Sampling until 24 hours post dose
    - LC/MS-MS, LLOQ 50 ng/mL
  - Drug D passed ABE with ease
    - $t_{½}$ 3.92 ± 0.88 h (T), 4.98 ± 1.24 h (R)
    - Extrapolated *AUC* (median, minimum – maximum)
      T: 1.76% (0.87 – 3.61%), R: 2.42% (1.14 – 6.19%)
  - Sponsor developed a four-drug FDC
    - Data of the BE study should be used in a PopPK model to optimize the sampling schedule for a new study

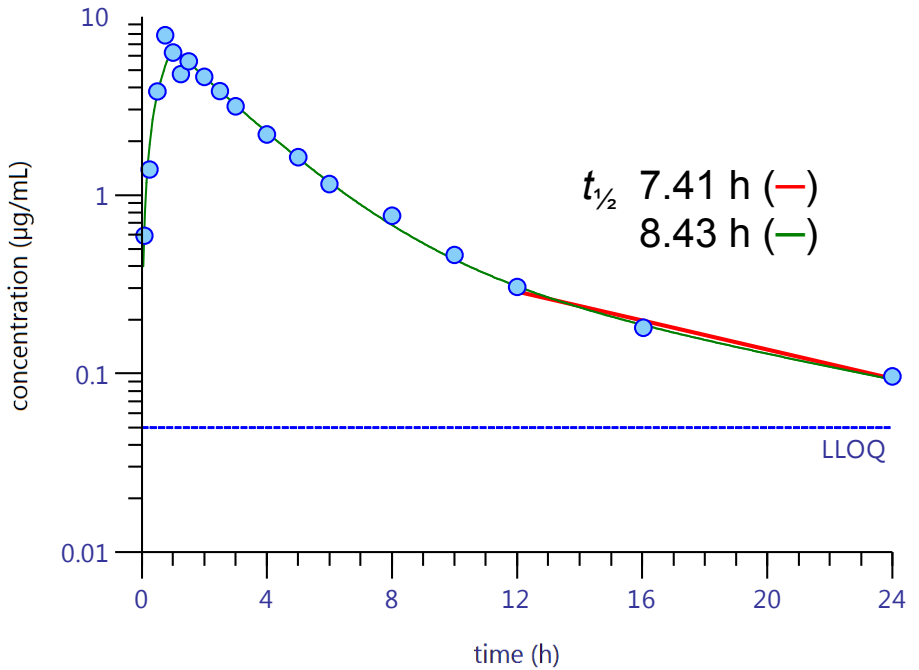
NCA

- Drug D: $t_{½}$ 2 – 3 h (literature)
  - No individual λ_z or $t_{½}$ (as well as time ranges used in estimation) given in the report, only AUC_{0-t} and $AUC_{0-\infty}$
  - Reproduced the CRO's results by trial and error. Example:



NCA

- Drug D: $t_{½}$ 2 – 3 h (literature)
  - Obviously the time range for the estimation of λ_z was wrong
    - Two-compartment model!
  - What I obtained by NCA (—) and a PK model (—)

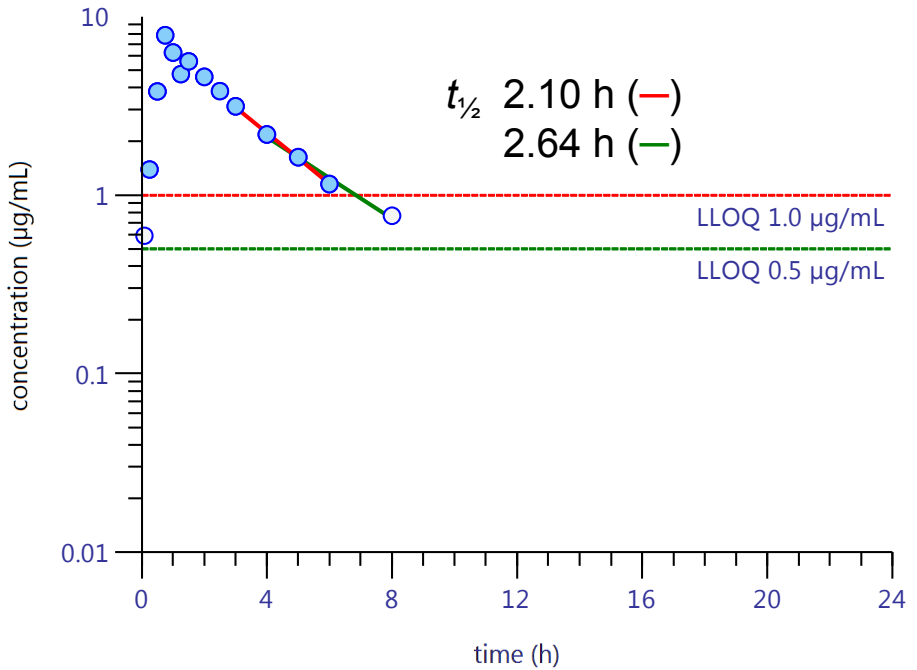


NCA

- Drug D: $t_{½}$ 2 – 3 h (literature)
  - Why? No problems with correct estimation of λ_z
    - $t_{½}$ 4.63 ± 1.07 h (T), 5.59 ± 1.19 h (R)
    - Extrapolated *AUC* (median, minimum – maximum)
      T: 2.08% (1.06 – 4.32%), R: 2.84% (1.47 – 6.19%)
  - Potential explanations
    - 'Push-the-button-pharmacokineticist' at work
      - Relied on an automatic algorithm?
      - No visual inspection of fits?
    - Anticipatory obedience?
      - The bioanalytical method was at least 10times more sensitive than ones used in the past (drug D approved in 1955)
      - Maybe the CRO wanted to avoid a single sentence in the discussion section of the report clarifying why a second phase is apparent – explaining longer half lives than the ones known from the literature

NCA

- Drug D: $t_{½}$ 2 – 3 h (literature)
  - Estimation of λ_z by bioanalytical methods with an LLOQ of 1.0 or 0.5 µg/mL explains short half lives given in the literature



NCA

- Drug D: $t_{½}$ 2 – 3 h (literature)
  - Lessons learned
    - The report should allow independent assessment
    - Good practice [1,2]
      - All raw data
      - λ_z and/or $t_{½}$ as well as time ranges used in estimation
      - All derived PK metrics
    - Desirable
      - Machine-readable data
      - Open formats (CSV, XML, CDISC, M$ XLSX) preferred over proprietary ones (SAS XPT, M$ XLS)
    - Unacceptable
      - A 500+ page PDF generated by SAS
      - As above but a scanned printout

1. Schulz H-U, Steinijans, VW. *Striving for standards in bioequivalence assessment: a review*. Int J Clin Pharm Ther Toxicol. 1991;29(8):293–8. PMID 1743802.
2. Sauter R, Steinijans VW, Diletti E, Böhm E, Schulz H-U. *Presentation of results from bioequivalence studies*. Int J Clin Pharm Ther Toxicol. 1992;30(Suppl.1):S7–30. PMID 1601535.

Statistics

- Adaptive Two-Stage Sequential Design in BE
  - EMA (2010)
    It is acceptable to use a two-stage approach […]. If this approach is adopted appropriate steps must be taken to preserve the overall type I error of the experiment […]. 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, but there are many acceptable alternatives and the choice of how much alpha to spend at the interim analysis is at the company's discretion.
  - The 94.12% CI (α 0.0294) preserves the patient's risk in simulation-based methods if and only if
    - *GMR* 0.95 and
    - target power 80%

Statistics

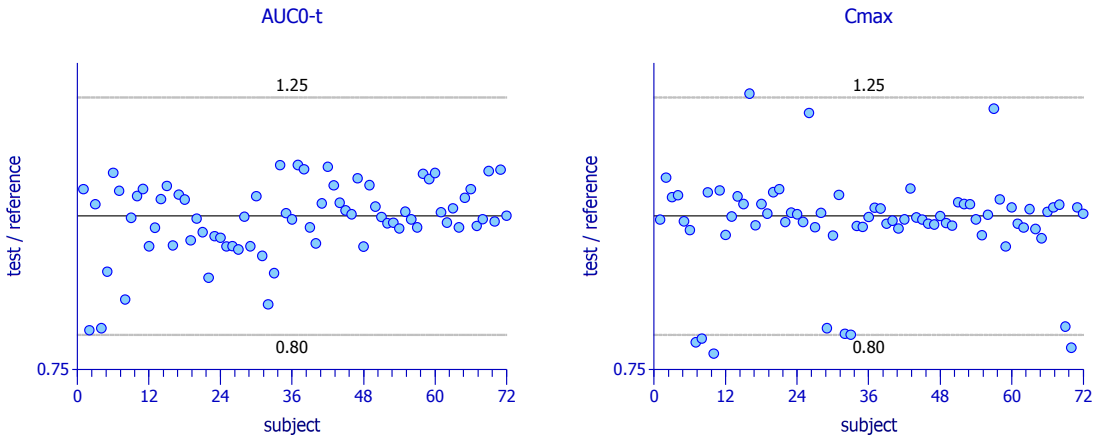
- Drug E: Adaptive Two-Stage Sequential Design
  - GMR 0.90 (≠ 0.95), target power 85% (≠ 80%), α 0.0294
  - Stage 1: n_1 24
    - Failed: PE 89.00% (94.12% CI: 77.24 – 102.54%)
    - Stage 2 with 54 subjects initiated
  - Pooled data: n_1+n_2 78
    - Passed: PE 91.00% (94.12% CI: 82.16 – 100.79%)
  - Inflated type I error (patient's risk 5.23%)
  - The study's conditions would require *more* adjustment (α 0.0279 = 94.42% CI)
    - *Post hoc* assessment based on the study's *CV*
      - Passed: PE 91.00% (94.42% CI: 82.05 – 100.92%)
      - Type I error 4.99%
      - Wider CI but conclusion agrees with the original analysis

Statistics

- Drug E: Adaptive Two-Stage Sequential Design
  - However, correct would have been to find a suitable α (0.0278) for GMR 0.90 and target power 85% already *before*, pre-specify it in the protocol, and evaluate the study with the adjusted $100(1 - 2\alpha) = 94.44\%$ CI
  - Stage 1: n_1 24
    - Failed: PE 89.00% (94.44% CI: 77.09 – 102.75%)
    - Stage 2 with 54 subjects initiated
  - Pooled data: n_1+n_2 78
    - Passed: PE 91.00% (94.44% CI: 82.05 – 100.93%)
  - Type I error controlled (patient's risk 4.99%)

Statistics

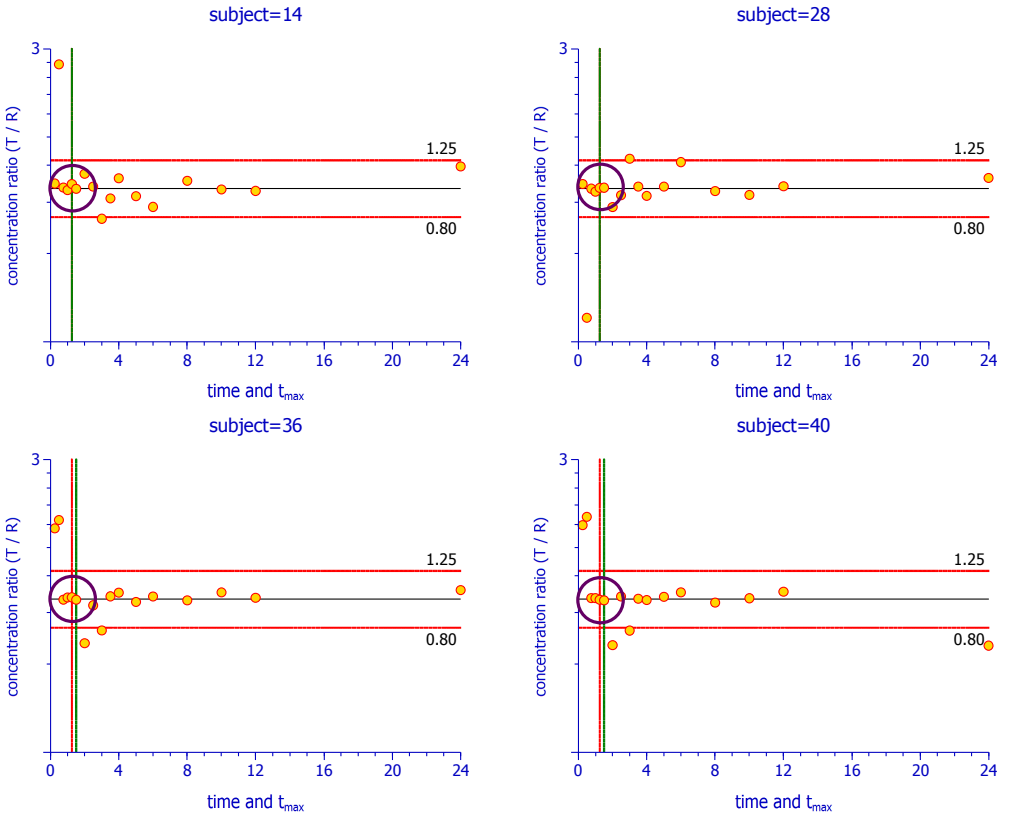
- Drug F: Documented high variability (literature, EPARs)
  - Generally a replicate design study is required (CV_{wR} of C_{max} ~40 – 50%, CV_{wR} of AUC 30 – 40%)
  - 2×2×2 crossover in 72 subjects, intra-subject CVs:
    - C_{max} 6.46%
    - AUC_{0-t} 4.87%
  - NCA and BE recalculated by ANAMED in Phoenix/WinNonlin 6.4 and myself in PHX/WNL 8.1: ‘Results’ confirmed



No obvious trend like in the 2012 GVK/Hyderabad-case!

Statistics

- Drug F: Documented high variability (literature, EPARs)
  - Most dubious cases



t_{max} of drug F reported in the literature with 1–2 h.

--- t_{max} (R)
--- t_{max} (T)

Suspicion
Were bioanalytics unblinded and in the area of the expected t_{max} the "R-samples" extracted – or even just injected – twice instead of the "T-samples"?

No smoking gun found in inspection (2019).

Statistics

- Sample size estimation
  - EMA NfG (2001)
    - The number of subjects […] is determined by
      - the error variance associated with the primary characteristic to be studied as estimated from a pilot experiment, from previous studies or from published data,
      - the significance level desired,
      - the expected deviation from the reference product compatible with bioequivalence (Δ) and
      - the required power.
  - EMA IR GL (2010)
    - The number of subjects to be included in the study should be based on an appropriate sample size calculation

– *MSE*, *CV*

– *p* of type I error (α)

– T/R-ratio

– *p* of type II error (β); power = 1 – β

Statistics

- Sample size estimation *not* calculation
  - The variability is an estimate (previous studies, literature) or an assumption, the T/R-ratio an assumption, the power based on a desire (driven by the applicant's budget; although extremely highly powered studies should be rejected by the IEC)
  - The patient's risk (generally 5%) and acceptance limits (generally 80.00 – 125.00%) are fixed by the authority
- The myth of *post hoc* (aka *a posteriori*) power
  - The outcome of a comparative BA study is dichotomous
    - *Either* the study demonstrated BE *or* not
    - Calculation of *post hoc* power is futile
      - A high value does not further support BE; it only shows that expected values were not exactly realized in the study
      - A low value does not invalidate the conclusion since the patient's risk is not affected (α is independent from β)

Statistics

- 2×2×2 crossover, 71 eligible subjects
  - From the study report (SAS, code not given)
    - CV_w 23.08%
    - Failed on C_{max} PE 119.84% (90% CI: 112.44 – 127.73%)
    - Power 100.0%
  - If power (probability to pass BE!) really is 100%, why did the study fail?
  - Power can be estimated with the R package PowerTOST [3]

```
library(PowerTOST)
round(100*power.TOST(alpha=0.05, CV=0.2308, theta0=1.1984, n=71), 1)
```

    gives

```
[1] 29.0
```

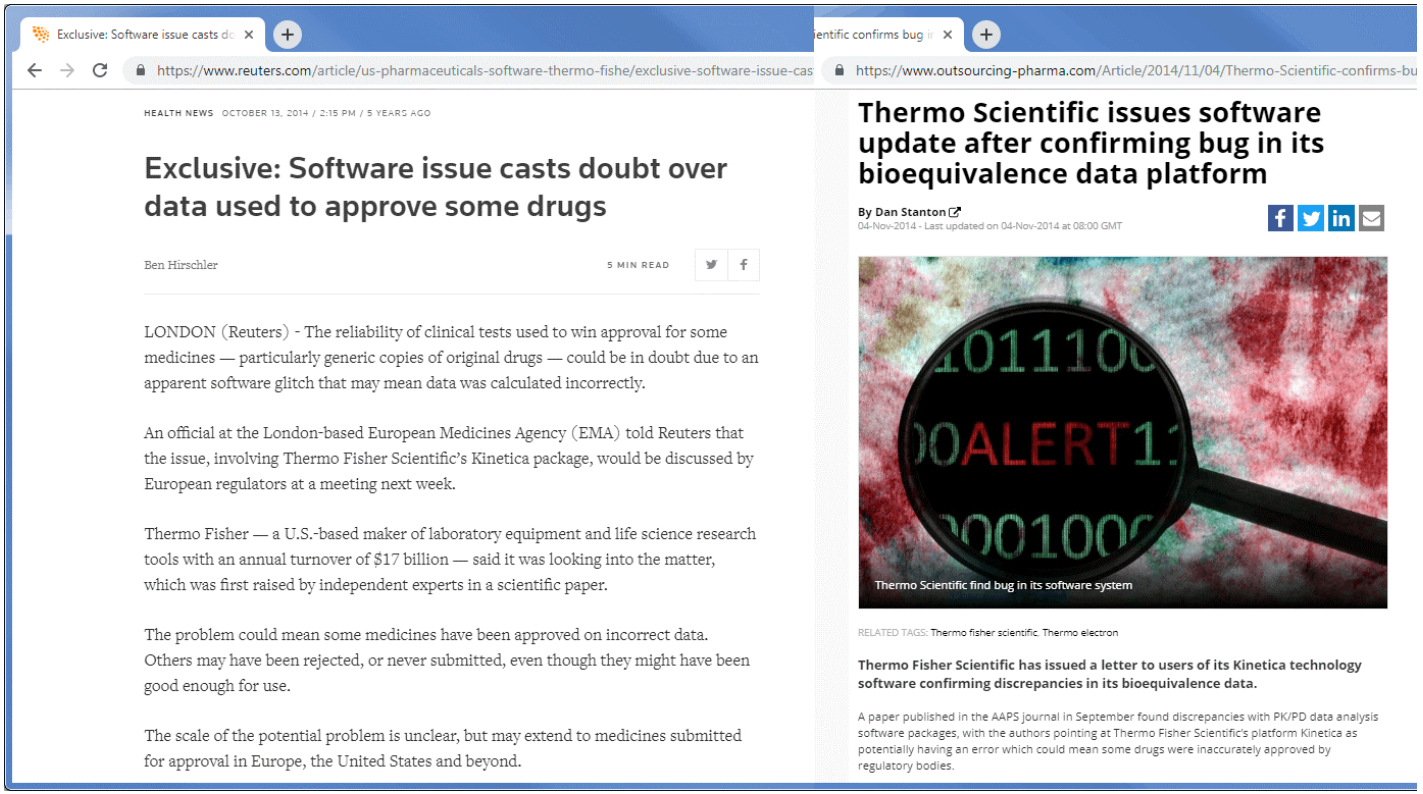
- Power is not of a regulatory concern but demonstrates a lack of statistical knowledge

3. Labes D, Schütz H, Lang B. *PowerTOST: Power and Sample Size Based on Two One-Sided t-Tests (TOST) for (Bio)Equivalence Studies*. 2018; R package version 1.4-7.

Software

- Validation mandatory
  - Common life cycle model should be followed
    - Installation Qualification Vendor (+ User)
    - Operational Qualification User (+ Vendor)
    - Performance Qualification User
  - White-box validation of commercial software *impossible* (source code not accessible)
    - Only black-box validation possible
      - Cross-validation with results of reference data sets obtained by other software
  - White-box validation of open-source software *possible* (by definition)
    - Possible ≠ easy; requires an expert coder
    - However, black-box validation possible as well

Software



HEALTH NEWS OCTOBER 13, 2014 / 2:15 PM / 5 YEARS AGO

Exclusive: Software issue casts doubt over data used to approve some drugs

Ben Hirschler 5 MIN READ

LONDON (Reuters) - The reliability of clinical tests used to win approval for some medicines — particularly generic copies of original drugs — could be in doubt due to an apparent software glitch that may mean data was calculated incorrectly.

An official at the London-based European Medicines Agency (EMA) told Reuters that the issue, involving Thermo Fisher Scientific's Kinetica package, would be discussed by European regulators at a meeting next week.

Thermo Fisher — a U.S.-based maker of laboratory equipment and life science research tools with an annual turnover of $17 billion — said it was looking into the matter, which was first raised by independent experts in a scientific paper.

The problem could mean some medicines have been approved on incorrect data. Others may have been rejected, or never submitted, even though they might have been good enough for use.

The scale of the potential problem is unclear, but may extend to medicines submitted for approval in Europe, the United States and beyond.

Thermo Scientific issues software update after confirming bug in its bioequivalence data platform

By Dan Stanton
04-Nov-2014 - Last updated on 04-Nov-2014 at 08:00 GMT

RELATED TAGS: Thermo fisher scientific, Thermo electron

Thermo Fisher Scientific has issued a letter to users of its Kinetica technology software confirming discrepancies in its bioequivalence data.

A paper published in the AAPS journal in September found discrepancies with PK/PD data analysis software packages, with the authors pointing at Thermo Fisher Scientific's platform Kinetica as potentially having an error which could mean some drugs were inaccurately approved by regulatory bodies.

4. Schütz H, Labes D, Fuglsang A. *Reference Datasets for 2-Treatment, 2-Sequence, 2-Period Bioequivalence Studies.* AAPS J. 2014;16(6):1292–97. doi:10.1208/s12248-014-9661-0.
5. Moralez-Acelay S, de la Torre de Alvarado JM, García-Arieta A. *On the Incorrect Statistical Calculations of the Kinetica Software Package in Imbalanced Designs.* AAPS J. 2015;17(4):1033–4. doi:10.1208/s12248-015-9749-1.
6. Fuglsang A, Schütz H, Labes D. 2015. *Reference Datasets for Bioequivalence Trials in a Two-Group Parallel Design.* AAPS J. 2015;17(2):400–4. doi:10.1208/s12248-014-9704-6.

Software

- Reference data-sets in the public domain which allow users to PQ their software installations

design	sequences/ groups	vari-ances	R	SAS	Phoenix/ WinNonlin	JMP	STaTa	OO Calc	SPSS	Kine-tica	Equiv-Test	Thoth-Pro	Statis-tica
2×2×2 Xover [4,5]	balanced	identical	☑	☑	☑	☑	☑	☑	NT	☑	☑	☑[a]	NT
	imbalanced		☑	☑	☑	☑	☑	☑	NT	☒	☑	☒	NT
2 groups parallel [6]	equal	equal	☑	☑	☑	☑	☑	☑	NT	☑	☑	–	NT
		unequal	☑	☑	☑	☑	☑	☑	NT	–	–	–	NT
	unequal	equal	☑	☑	☑	☑	☑	☑	NT	☒	–	–	NT
		unequal	☑	☑	☑[b]	☑	☑	☑	NT	–	–	–	NT
replicate, scaling [7]	balanced, imbalanced, incomplete	equal, unequal	☑	☑	☑	☑	☑	NT	☑	–	–	–	☑

☑ passed NT Not tested (yet)
☒ incorrect – Not implemented (*i.e.*, design cannot be evaluated)

a. Limited to 100 subjects
b. Limited to 1,000 subjects / group

7. Schütz H, Tomashevskiy M, Labes D, Shitova A, González-de la Parra M, Fuglsang A. *Reference Datasets for Studies in a Replicate Design intended for Average Bioequivalence with Expanding Limits.* Manuscript in preparation 2019.
8. Schütz H, Tomashevskiy M, Labes D. *replicateBE: Average Bioequivalence with Expanding Limits (ABEL).* 2019; R package version 1.0.8. https://cran.r-project.org/package=replicateBE.