Software Validation

Helmut Schütz

Not only Software

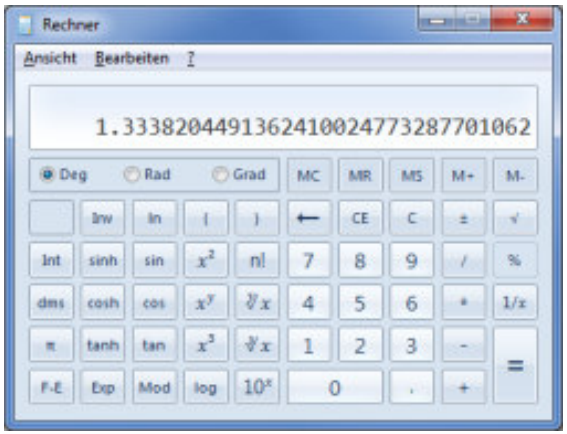
Pentium FDIV bug (1993).

- Flaw in the x86 assembly language floating point divison.
  - Example

$$\frac{4,195,835}{3,145,727} = 1.333739068902037589$$

$$\frac{4,195,835}{3,145,727} = 1.333820449136241002$$

  - Costs for replacement: $475 million.



Mostly Software

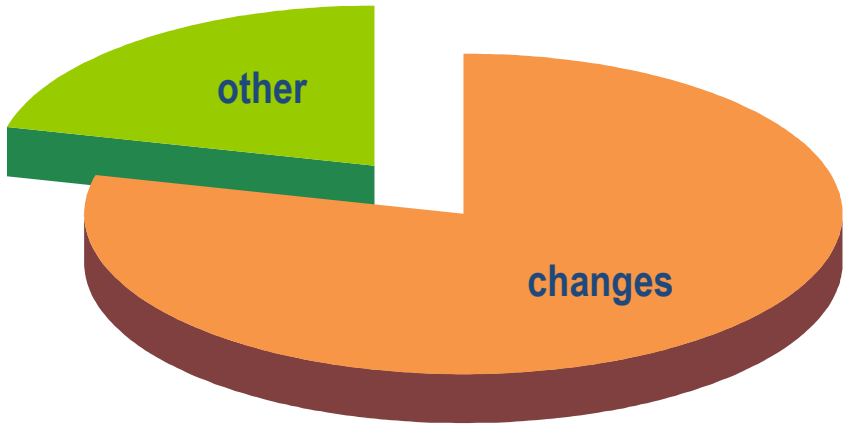
Therac-25 (1985 – 1987).

- Radiation therapy machine (Atomic Energy of Canada Ltd).
  Two operationg modes:
  - Direct electron-beam therapy.
    Low doses of high-energy (5 – 25 MeV) electrons
    over short periods of time.
  - Megavolt X-ray therapy.
    X-rays produced by colliding high-energy (25 MeV) electrons
    into a target.
- A one-byte counter in a testing routine frequently overflowed. If an operator provided manual input to the machine at the precise moment that this counter overflowed, the machine switched between operating modes. Patients received ~100 – 1,000 times the intended dose.
- Several patients injured, three died.

Mostly Software

General Principles of Software Validation (FDA 2002).

- Section 2.4: Regulatory Requirements for Software Validation
  - 242 FDA Medical Device Recalls attributed to software failures (1992 – 1998).
  - 192 (79%) caused by software defects that were introduced when changes were made to the software after its initial pro-duction and distribution.



Mostly Software

General Principles of Software Validation (FDA 2002).

- Section 2.4 (cont'd)
  - Any software […] must be validated for its intended use.
  - Computer systems must be validated to ensure accuracy, reliability, consistent intended performance, and the ability to discern invalid or altered records.
  - All […] software, even if purchased off-the-shelf, should have documented requirements that fully define its intended use, and information against which testing results and other evidence can be compared, to show that the software is validated for its intended use.

Lines of Code (LOC)

80/20-Rule.

- 80% of lines are coded within 20% of time.
  - Changing and testing is the most tedious part.
    - Average coding and testing: 10 – 50 LOC / day.
    - Software with 1 defect / 2,000 lines is considered ‘stable’.

software	year	10^6 LOC
MS-DOS	1981	0.004
Win 3.1	1992	3
Win NT 4.0	1996	12
Win XP	2001	45
Win 8	2012	60

software	year	10^6 LOC
MS Office	2013	45
Photoshop 1	1990	0.1
PS CS 6	2012	5
Mac OS X	2005	85
Linux 3.6	2012	16

Some Terms

IEEE (610, 1028), ISO, and ISTQB.

- Error: A human action that produces an incorrect result.
- Defect: A flaw in a component or system that can cause the component or system to fail to perform its required function, *e.g.* an incorrect statement or data definition.
- Failure: Deviation of the component or system from its expected delivery, service or result.
- Example: Division by zero
  - Error: 0 as a user entry was not tested/trapped.
  - Defect: The program is (unnoticed) erroneous until data entry.
  - Failure: Runtime error during execution.

More Trems

ISO 9000 and FDA (1999).

- **Qualification.**
  The process of demonstrating the ability to fulfill specified requirements (the term ‘qualified’ is used to designate the corresponding status).
  - **Installation Q:** **[…] systems are compliant with appropriate codes and approved design intentions, and that vendor’s recommendations are suitably considered.**
  - **Operational Q:** **[…] systems are capable of consistently operating within stated limits and tolerances.**
  - **Performance Q:** **[…] meeting all release requirements for functionality and safety and that procedures are effective and reproducible.**

Qualification(s)…

Examples

- Each of the Qualification(s) should include an instruction, an expected result, and the actual result.
  - Any discrepancy between the expected result and the actual result should be tracked as a deviation.
  - Deviations should be resolved before validation is complete.
- Installation Qualification
  - The OS has the appropriate processor, RAM, etc.
  - All files required to run the system are present and access rights are granted.
  - All documentation required to train system personnel has been approved.

Qualification(s)…

Examples

- Operational Qualification
  - System security has been properly implemented.
  - All documentation required to train personnel has been approved.
  - Data entry / import accepts appropriate data and rejects inappropriate ones.
  - Data export is compliant with specifications.
  - Test datasets can be moved through an entire workflow.
  - (Technological controls for compliance with 21 CFR 11 are functioning as expected)
  - …

Qualification(s)…

Examples

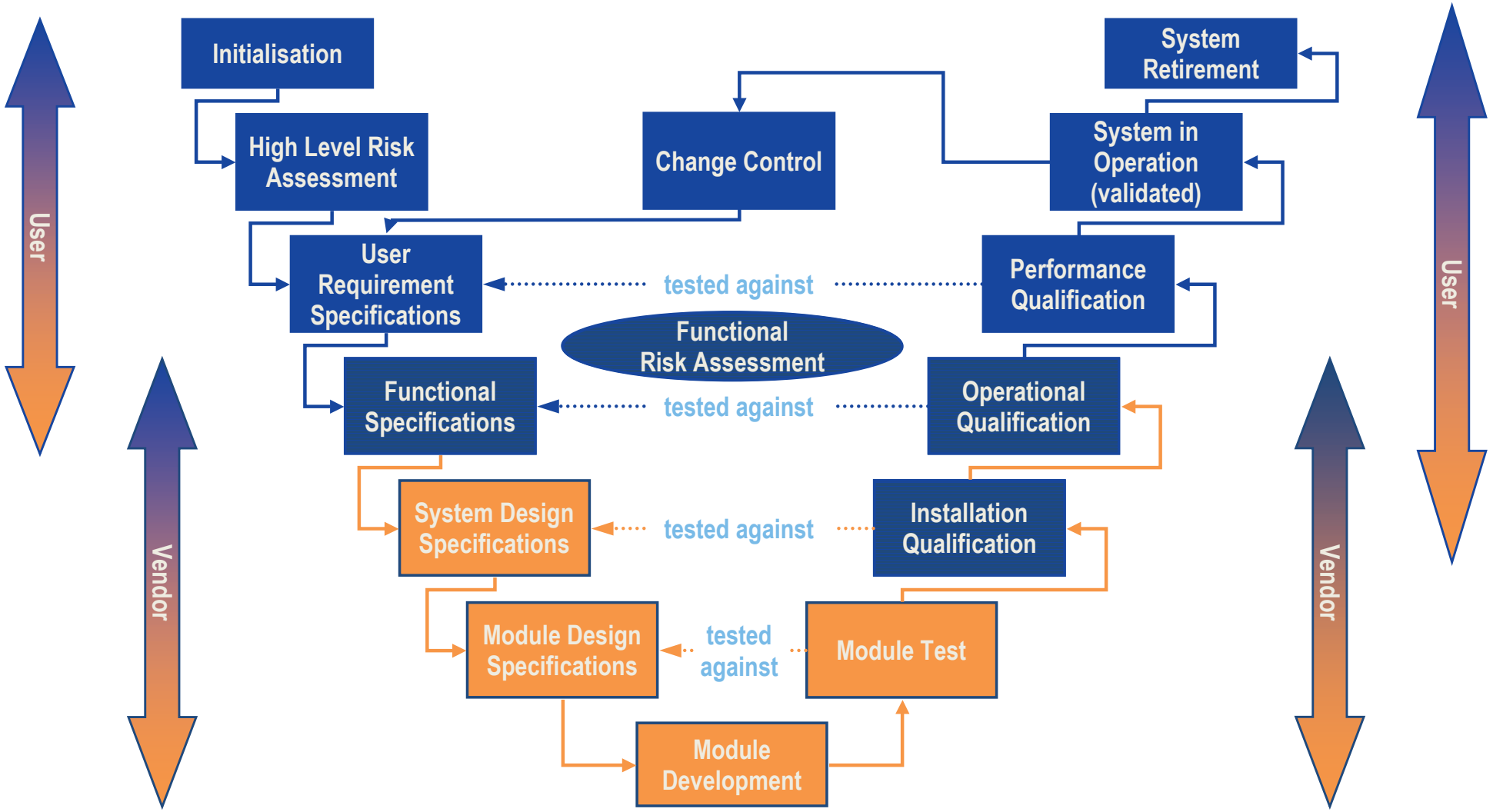
- Performance Qualification
  - Test datasets' results are within defined system requirements.
  - Concurrent independent workflows do not affect each other.
  - The system can handle multiple users without significant system lag.
  - …

Confusion?

General Principles of Software Validation (FDA 2002).

- Section 3.1.3: IQ/OQ/PQ
  - […] FDA and regulated industry have attempted to understand and define software validation within the context of process validation terminology.
  - While IQ/OQ/PQ terminology has served its purpose well and is one of many legitimate ways to organize software validation tasks at the user site, this terminology may not be well understood among many software professionals […].
    However, both FDA personnel and […] manufacturers need to be aware of these differences in terminology as they ask for and provide information regarding software validation.

System Life Cycle (V Model)



Esch *et al.* 2007. *Good Laboratory Practice (GLP) – Guidelines for the Validation of Computerised Systems.*

Responsibilities

Part of the SLC can be performed in close collaboration with the vendor.

- **Defining Functional Specifications and the Risk Assessment.**
- **Performing Installation and Operational Qualification.**
- **Running a large installation without a current support contract is grossly negligent.**

However, other parts are the sole responsibility of the user (*e.g.*, Performance Qualification)

Responsibilities

The ultimate responsibility in a controlled environment lies in the user's hands.

- **Full control of the SLC only possible for in-house developed software and mostly for outsourced developed one.**
- **Try to get access to the source code for independent review ('white box' validation).**
- **If not possible (vendor refuses an audit), perform a 'black box' validation.**

Responsibilities

The ultimate responsibility (cont'd).

- 'Black box' validation
  - Run datasets with certified results (*e.g.*, from NIST's Statistical Reference Datasets Project).
    - FDA (2002)
      - » Testing with usual inputs is necessary.
      - » However, testing a software product only with expected, valid inputs does not thoroughly test that software product.
      - » By itself, normal case testing cannot provide sufficient confidence in the dependability of the software product.
  - Create 'worst-case' datasets (extreme range of input, enter floating point numbers to integer fields, enter characters to numeric fields…).

Responsibilities

The ultimate responsibility (cont'd).

- Section 5.2.7 Maintenance & Software Changes (FDA 2002).
  - Corrective: Changes made to correct errors and faults.
  - Perfective: Changes made to improve the performance, maintainability, or other attributes.
  - Adaptive: Changes to make the software usable in a changed environment.
  - Sufficient […] analysis and testing should […] demonstrate that portions of the software not involved in the change were not adversely impacted (in addition to testing […] the correctness of the implemented changes).

Computer System Validation (CSV)

Analogies to a GLP study.

GLP study	CSV	Remarks
Study director	Validation director	Ultimate responsibility
Study plan	Validation plan	Approved/signed by SD/VD
Method description	Test scripts	Referenced to or included in plan
	Conduct	Executing according to plan and methods/scripts
	Raw data	Documented evidence of test results
Study report	Validation report	Audited by QA and approved/signed by SD/VD

Esch *et al.* 2007. *Good Laboratory Practice (GLP) – Guidelines for the Validation of Computerised Systems.*

Spreadsheets?

Radio Yerevan Jokes.

- Radio Yerevan was asked:

  Is it possible to validate M$ Excel?

- Radio Yerevan anwered:

  In principle yes, but only if you buy the source code from Mr Gates first.

EMA CPMP/CHMP/EWP (Q&A 2011–2015)

- Results obtained by alternative, validated statistical programs are also acceptable except spreadsheets because outputs of spreadsheets are not suitable for secondary assessment.

Esch *et al*. 2010. *Good Laboratory Practice (GLP) – Guidelines for the Development and Validation of Spreadsheets.*

Spreadsheets?

MS Excel 1985 – 2002.

M$ Article 828888: 'You can expect that for most users, such round off errors are not likely to be troubling in practice.'

	A	B	C	D	E	F
1	0	formula (A)	100,000,000	formula (C)	1	formula (E)
2	–1	=A$1–1	99,999,999	=C$1–1	0.99999999	=E$1–0.00000001
3	±0	=A$1	100,000,000	=C$1	1.00000000	=E$1
4	+1	=A$1+1	100,000,001	=C$1+1	1.00000001	=E$1+0.00000001
5	1	=STDEV(A2:A4)	0	=STDEV(C2:C4)	0	=STDEV(E2:E4)

In calculating the 90% CI we need the *t*-distribution (for α 0.05 and the residual degrees of freedom).

- Example: $t_{0.05,\,22} = 1.717$.
- However, in MS Excel <2007:

	A	B	C	D	E	F	G	H
1	α	df	t	formula (C)	t	workaround (E)	t	Excel 2007+
2	0.05	22	2.074	=TINV(A2, B2)	1.717	=TINV(2*A2, B2)	1.717	=T.INV(A2, B2)

Open Source Software?

In principle yes – if it's validated, *why not*?

- Since the source code is accessible, even a 'white box' validation – which *no* off-the-shelf software offers – is possible.
  - The FDA regularly uses R in M&S itself (but – as an agency – never validates anything…).
  - New releases/updates more frequent than commercial SW.
    - R & packages: 3 – 4 / year.
    - Bugs in packages: Generally corrected within one week.
  - R-packages relevant for BE:
    - Randomization: `randomizeBE` (2012)
    - NCA/BE: `bear` (2016)
    - Power and sample size: `PowerTOST` (2016)
    - Two-Stage Designs: `Power2Stage` (2015)

The R Foundation for Statistical Computing 2014. *R: Regulatory Compliance and Validation Issues.*

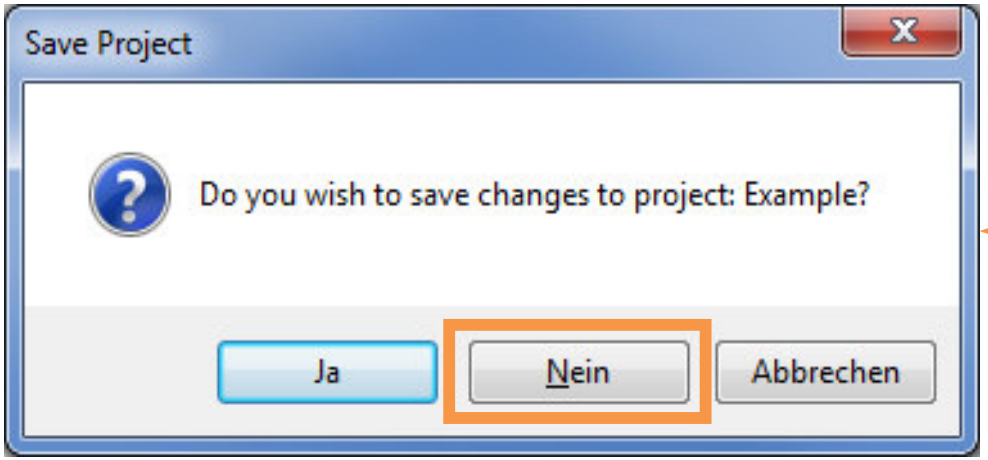
Alterations of Data possible?

Example: Phoenix/WinNonlin

Properties | Information | History

Example >> Data >> EMA full replicate

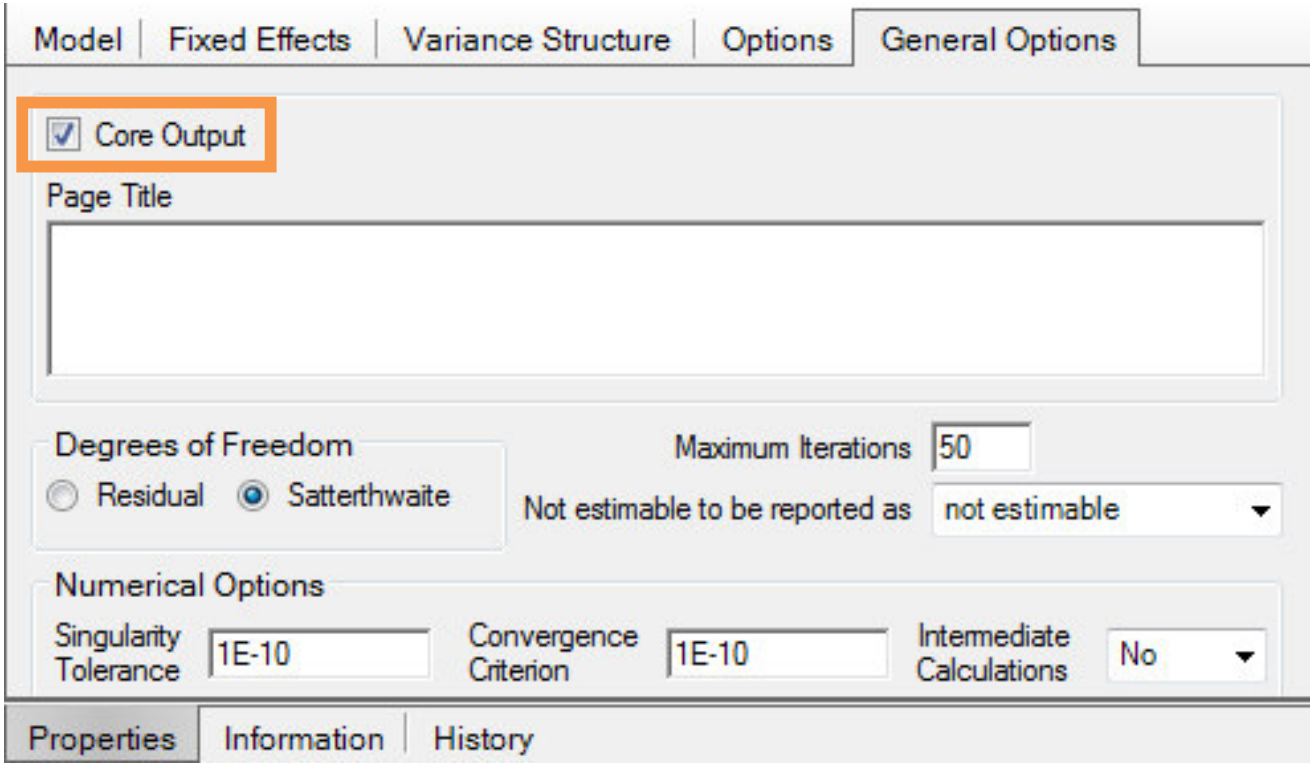
Timestamp	User	Object Name	Event	Description
2015.04.09 14:45:10 UTC	HS	Worksheet	Object Created	Object Created E:\Public\Documents\BEBAC\Phoenix Projects\EMA full replicate.xls
2015.04.09 14:45:46 UTC	HS	EMA full replicate	Value changed	F1; changed from 7.734541 to 7.5



If software allows changes without an audit trail, take measures!
PKS is 21 CFR 11 compliant…

Alterations of Data possible?

Example: Phoenix/WinNonlin



Always select the Core Output (off by default)

Alterations of Data possible?

Example: Phoenix/WinNonlin

```
                                                          Date:     4/09/2015
                                                          Time:     17:20:50

                WINNONLIN LINEAR MIXED EFFECTS MODELING / BIOEQUIVALENCE
                                        6.4.0.768
                                    Core Version 30Jan2014


  Model Specification and User Settings
          Dependent variable : Data
                   Transform : LN
                Fixed terms  : int+Sequence+Subject(Sequence)+Period+Formulation
      Singularity tolerance  :       1e-010
      Denominator df option  : satterthwaite
```

Only in the Core Output you get a timestamp of the evaluation.
Avoid fancy Excel- or Word-Export options (if possible).

Old Hats …

Parallel Groups: Example

- Evaluation (modified data set)

Program	equal variances	unequal variances
R 2.5.0 (2007)	81.21% – 190.41%	76.36% – 202.51%
NCSS 2001 (2001)	81.21% – 190.41%	76.36% – 202.51%

- Inflated α-risk in 'conventional' *t*-test (naive pooling) is reflected in a tighter confidence interval.
- Preliminary testing for equality in variances is flawed*) and should be avoided (FDA).
- Approximations (*e.g.*, Satterthwaite, Aspin-Welch, Howe, Milliken-Johnson) are currently *not implemented* in packages 'specialized' in BE (WinNonlin, Kinetica, EquivTest/PK)!

Surprise?

*) Moser, B.K. and Stevens, G.R.; Homogeneity of variance in the two-sample means test. Amer. Statist. 46, 19-21 (1992)

informa life sciences Dissolution Testing, Bioavailability & Bioequivalence | Budapest, 24 May 2007 38

… making it to the Health News

Murphy's Law:
If anything can go wrong, it will.

Reference Datasets in BE

Different software (general purpose, specialized in BE, commercial and open source), 2×2×2 crossover.

DS	EquivTest		Kinetica		SAS		WinNonlin		R	
A	90.76	99.62	90.76	99.62	90.76	99.62	90.76	99.62	90.76	99.62
B	51.45	98.26	51.45	98.26	51.45	98.26	51.45	98.26	51.45	98.26
C	39.41	87.03	44.91	99.31	39.41	87.03	39.41	87.03	39.41	87.03
D	51.45	98.26	51.45	98.26	51.45	98.26	51.45	98.26	51.45	98.26
E	55.71	151.37	55.71	151.37	55.71	151.37	55.71	151.37	55.71	151.37
F	93.37	106.86	93.37	106.86	93.37	106.86	93.37	106.86	93.37	106.86
G	88.46	95.99	88.46	95.99	88.46	95.99	88.46	95.99	88.46	95.99
H	86.81	100.55	107.80	115.85	86.81	100.55	86.81	100.55	86.81	100.55

A, B, D – G Balanced sequences ($n_{TR} = n_{TR}$)

C, H Imbalanced sequences ($n_{TR} \neq n_{RT}$)

Schütz H, Labes D, Fuglsang A. 2014. *Reference Datasets for 2-Treatment, 2-Sequence, 2-Period Bioequivalence Studies.*
Moralez-Acelay *et al.* 2015. *On the Incorrect Statistical Calculations of the Kinetica Software Package in Imbalanced Designs.*

Reference Datasets in BE

Two-group parallel (conventional *t*-test).

DS	EquivTest		Kinetica		SAS		WinNonlin		OO Calc		R	
1	27.15	86.94	27.15	86.94	27.15	86.94	27.15	86.94	27.15	86.94	27.15	86.94
2	18.26	96.59	15.76	119.00	18.26	96.59	18.26	96.59	18.26	96.59	18.26	96.59
3	26.35	415.71	26.35	415.71	26.35	415.71	26.35	415.71	26.35	415.71	26.35	415.71
4	38.60	134.21	38.60	134.21	38.60	134.21	38.60	134.21	38.60	134.21	38.60	134.21
5	106.44	112.10	106.39	112.44	106.44	112.10	106.44	112.10	106.44	112.10	106.44	112.10
6	91.85	115.78	92.07	115.50	91.85	115.78	91.85	115.78	91.85	115.78	91.85	115.78
7	106.86	126.49	104.30	129.32	106.86	126.49	106.86	126.49	106.86	126.49	106.86	126.49
8	105.79	113.49	105.79	113.49	105.79	113.49	105.79	113.49	105.79	113.49	105.79	113.49
9	103.80	120.61	103.80	120.61	103.80	120.61	103.80	120.61	103.80	120.61	103.80	120.61
10	107.20	126.99	104.59	130.16	107.20	126.99	107.20	126.99	107.20	126.99	107.20	126.99
11	7.83	17.38	6.98	19.51	7.83	17.38	7.83	17.38	7.83	17.38	7.83	17.38

1, 3, 4, 8, 9 Equal group sizes ($n_T = n_R$)

2, 5 – 7, 10, 11 Unequal group sizes ($n_T \neq n_R$)

Fuglsang A, Schütz H, Labes D. 2015. *Reference Datasets for Bioequivalence Trials in a Two-Group Parallel Design.*

Reference Datasets in BE

Two-group parallel (Welch's test).

DS	SAS		WinNonlin*		OO Calc		R	
1	26.78	88.14	26.78	88.14	26.78	88.14	26.78	88.14
2	23.71	74.38	23.71	74.38	23.71	74.38	23.71	74.38
3	24.40	449.08	24.40	449.08	24.40	449.08	24.40	449.08
4	38.05	136.15	38.05	136.15	38.05	136.15	38.05	136.15
5	106.44	112.10	106.44	112.10	106.44	112.10	106.44	112.10
6	91.84	115.79	91.84	115.79	91.84	115.79	91.84	115.79
7	97.38	138.51	NA		97.38	138.51	97.38	138.51
8	105.79	113.49	NA		105.79	113.49	105.79	113.49
9	103.80	120.61	NA		103.80	120.61	103.80	120.61
10	97.82	139.17	NA		97.82	139.17	97.82	139.17
11	6.30	21.60	NA		6.30	21.60	6.30	21.60

* Workaround required in WinNonlin; limited to 1,000 subjects.

Welch's test not implemented in EquivTest and Kinetica.

1, 3, 4, 8, 9 Equal group sizes ($n_T = n_R$)

2, 5 – 7, 10, 11 Unequal group sizes ($n_T \neq n_R$)

Fuglsang A, Schütz H, Labes D. 2015. *Reference Datasets for Bioequivalence Trials in a Two-Group Parallel Design.*

Likely Cause of Kinetica's Defects

2×2×2 crossover

- Calculation of the confidence interval (CI):

$$CI = e^{\log(\bar{x}_T - \bar{x}_R) \pm t_{1-\alpha,\, n_{RT}+n_{TR}-2} \sqrt{\frac{MSE}{2}\left(\frac{1}{n_{TR}} + \frac{1}{n_{RT}}\right)}}$$

- *Only* if sequences are balanced ($n_{TR} = n_{TR}$) a simplified formula based on the total sample size *N* is correct:

$$CI = e^{\log(\bar{x}_T - \bar{x}_R) \pm t_{\alpha,\, n_{RT}+n_{TR}-2} \sqrt{\frac{2MSE}{N}}}$$

Likely Cause of Kinetica's Defects

Two-group parallel

- Calculation of the confidence interval (CI):

$$CI = e^{\log(\bar{x}_T - \bar{x}_R) \pm t_{1-\alpha,\, n_T+n_R-2} \sqrt{\frac{MSE}{2}\left(\frac{1}{n_T}+\frac{1}{n_R}\right)}}$$

- According to the manual Kinetica uses a 'simplified' formula – but the sample size of subjects receiving the reference [*sic*] treatment in the denominator:

$$CI = e^{\log(\bar{x}_T - \bar{x}_R) \pm t_{1-\alpha,\, n_T+n_R-2} \sqrt{\frac{2MSE}{n_R}}}$$

Software Validation

Thank You!
Open Questions?

Helmut Schütz

BEBAC

Consultancy Services for
Bioequivalence and Bioavailability Studies
1070 Vienna, Austria

helmut.schuetz@bebac.at

To bear in Remembrance...

A refund for defective software might be nice, except it would bankrupt the entire software industry in the first year.

Andrew S. Tannenbaum

If debugging is the process of removing bugs, then programming must be the process of putting them in.

Edsger W. Dijkstra

I have stopped reading Stephen King novels. Now I just read C code instead.

Richard O'Keefe

References

EMA

- Q & A (2011): *Positions on specific questions addressed to the EWP therapeutic subgroup on Pharmacokinetics* URL
- *Annex III to Procedure for Conducting GCP Inspections requested by the EMEA: Computer Systems* (2007) URL

US FDA

- Division of Field Investigations (1987): *Technical Reference on Software Development Activities* URL
- CDRH-ODE (1999): *Gfl, FDA Reviewers and Compliance on Off-The-Shelf Software Use in Medical Devices* URL
- CDRH, CBER (2002): *General Principles of Software Validation* URL
- OC (2007): *Computerized Systems Used in Clinical Investi-gations* URL

PIC/S. 2007: *Good Practices for Computerised Systems in Regulated "GxP" Environments* URL

Switzerland, Federal Office of Public Health, AGIT

- *Validation of computerised systems in GLP* (2007) URL
- *Management of electronic SOPs in GLP* (2006) URL
- *Acquisistion and processing of electronic raw data in a GLP environment* (2006) URL
- *Change Management and Risk Assessment of validated computerised systems in a GLP environment* (2012) URL

International Software Testing Qualifications Board. *Standard Glossary of Terms used in Software Testing* (2014) URL

OECD. 1995: *The Application of the Principles of GLP to Computerised Systems* URL

International Society for Pharmaceutical Engineering: *GAMP®*

- *Good Practice Guide: Risk-Based Approach to Electronic Records and Signatures* (2005)
- *Good Practice Guide: Electronic Data Archiving* (2007)
- *A Risk-Based Approach to Compliant GxP Computerized Systems* (2008)
- *Good Practice Guide: A Risk-Based Approach to Operation of GxP Computerized Systems – A Companion Volume to GAMP® 5* (2010)
- *Good Practice Guide: A Risk-Based Approach to GxP Compliant Laboratory Computerized Systems* (2nd edition 2012)

Chamberlain R. *Computer Systems Validation for the Pharmaceutical and Medical Device Industry*. Alaren Press, Libertyville, IL, 1991: ISBN 0-9631489-0-7

Stokes et al. *Good Computer Validation Practices*. Interpharm, Buffala Grove, IL, 1994: ISBN 0-935184-55-4

Olivier D. *Conducting Software Audits, Auditing Software for Conformance to FDA Requirements*. Computer Application Specialists, San Diego, CA, 1994.

Beizer B. *Black Box Testing. Techniques for Functional Testing of Software and Systems*. Wiley, New York, NY, 1995: ISBN 0-471-12094-4

Fry JD, Drew RT. *Creation and Implementation of a Computer Validation Program.* Drug Info J. 1992: 26(1); 103–8. DOI 10.1177/009286159202600111

Leveson N. *Medical Devices: The Therac-25.* 1995. URL

AGIT– Working Group on Information Technology. *Good laboratory practice (GLP). Guidelines for the archiving of electronic raw data in a GLP environment.* Qual Assur J. 2003: 7(4); 262–9. DOI 10.1002/qaj.244

Hassler S, Donzé G, Esch PM, Eschbach B, Hartmann H, Hutter L, Timm U, Saxer HP. *Good Laboratory Practice (GLP) – Guidelines for the Acquisition and Processing of Electronic Raw Data in a GLP Environment.* Qual Assur J. 2006: 10(1); 3–14. DOI 10.1002/qaj.356

Esch PM, Donzé G, Eschbach B, Hassler S, Hutter L, Saxer HP, Timm U, Zühlke R. *Good Laboratory Practice (GLP) – Guidelines for the Validation of Computerised Systems.* Qual Assur J. 2007: 11(3–4); 208–20- DOI 10.1002/aqj.426

Esch PM, Moor C, Schmid B, Albertini S, Hassler S, Donzé G, Saxer HP. *Good Laboratory Practice (GLP) – Guidelines for the Development and Validation of Spreadsheets.* Qual Assur J. 2010: 13(3–4); 41–56. DOI 10.1002/qaj.466

Microsoft Knowledge Base – Article ID: 82634. *Description of the STDEV function in Excel 2003 and in later versions of Excel.* Last Review: September 19, 2011; Rev. 4.0

The R Foundation for Statistical Computing. 2014. *R: Regulatory Compliance and Validation Issues. A Guidance Do-cument for the Use of R in Regulated Clinical Trial Environments*. Vienna, Austria. URL

Schütz H, Labes D, Fuglsang A. *Reference Datasets for 2-Treatment, 2-Sequence, 2-Period Bioequivalence Studies.* AAPS J. 2014: 16(6); 1292–7. DOI 10.1208/s12248-014-9661-0

Fuglsang A, Schütz H, Labes D. *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

Moralez-Acelay S, de la Torr 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