EMA workshop
Case Study on Statistical Comparison of Stability Data - Importance of Data Quality

Franz Innerbichler
Novartis

EMA Workshop " Draft Reflection Paper on statistical methodology for the comparative assessment of quality attributes in drug development"

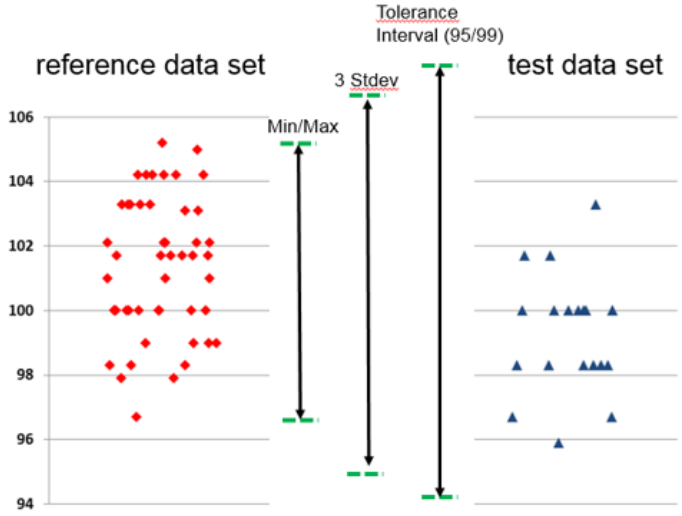
3-4 May 2018

ebe european biopharmaceutical enterprises

This is a joint industry presentation on behalf of the trade associations shown

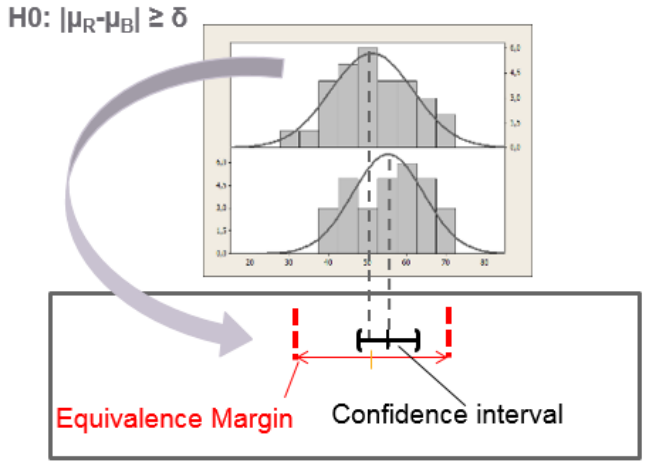
Comparison of data: Span vs Mean

Comparison of span/range:
1) Min-Max
2) +-k*sd
3) tolerance interval



applicability depends on the data and on the goals of the comparability exercise

Equivalence test on similarity of means



Comparison of data: Span vs Mean

most restrictive range test: Min-Max

$$P(\min(Y_i) < X_i < \max(Y_i)) = \frac{\binom{m-k+1}{m-k} * \binom{n+k-2}{k}}{\binom{n+m}{m}}$$

where: Y_i is the results of reference/pre, and X_i the bisimilar/post with i=1 to n for Y and i=1 to m for X.

- m is the no. of batches of the biosimilar/post
- n is the no. of batches of the reference/pre
- k is desired number of batches in the min-max range of the reference/pre
- m-k is the no. of batches out of min-max of the reference/pre
- $\binom{a}{b}$ is $\frac{a!}{b!*(a-b)!}$

5 ## in 10:
0 out: 42.8%
1 out: 33.0%
2 out: 16.5%
3 out: 6.0%
4 out: 1.5%
5 out: 0.2%

10 ## in 10:
0 out: 23.7%
1 out: 26.3%
2 out: 20.9%

with gratitude to José Ramírez, Chief Statistician at Amgen

Equivalence test on similarity of means (FDA 2017)

$$H_0 : \mu_T - \mu_R \leq -1.5\sigma_R \text{ or } \mu_T - \mu_R \geq 1.5\sigma_R$$
$$H_A : -1.5\sigma_R < \mu_T - \mu_R < 1.5\sigma_R$$

The SAS System

The POWER Procedure
Equivalence Test for Mean Difference

Computed Power
Power
0.873

Prob
0.802

Result of the simulation:
80.2% of equivalence tests showed „equivalent“

The reason for the difference (80.2 vs 87.3) is the estimator of sigma on the right side of H0

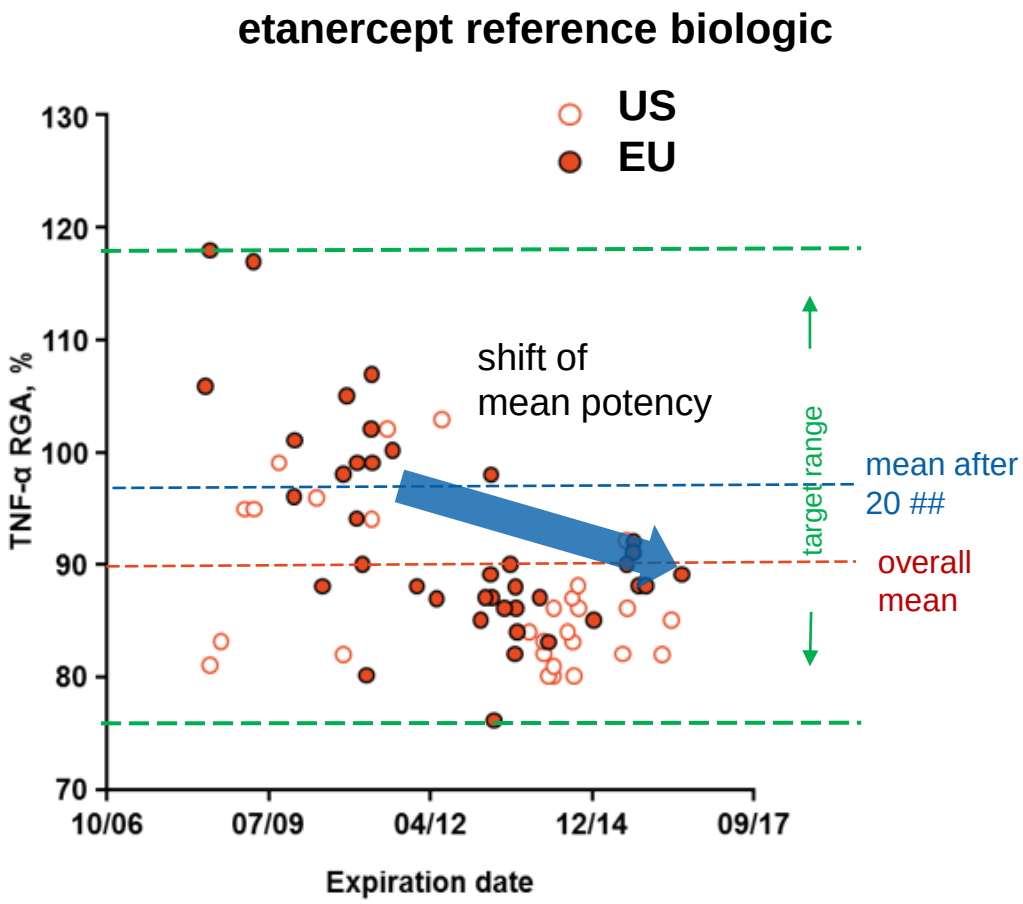
Multiplicity issue

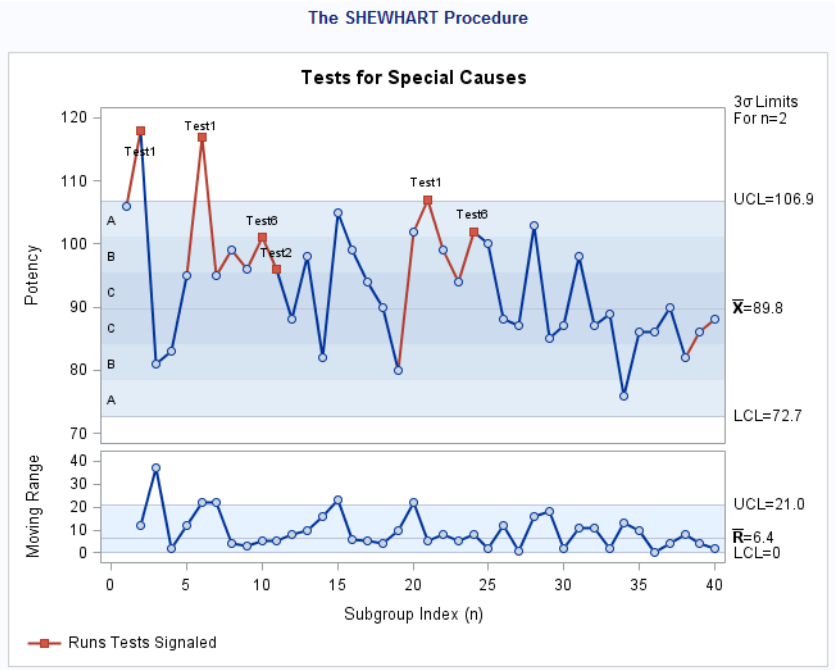
4 QAs: P(success) = 0.8^4 ≈ 41%

Burdick R, et al. The AAPS Journal.19,1,4-14, Jan2017

SAS Institute Inc., SAS Campus Drive, Cary, NC USA

Variation in etanercept reference biologic: Changes in production

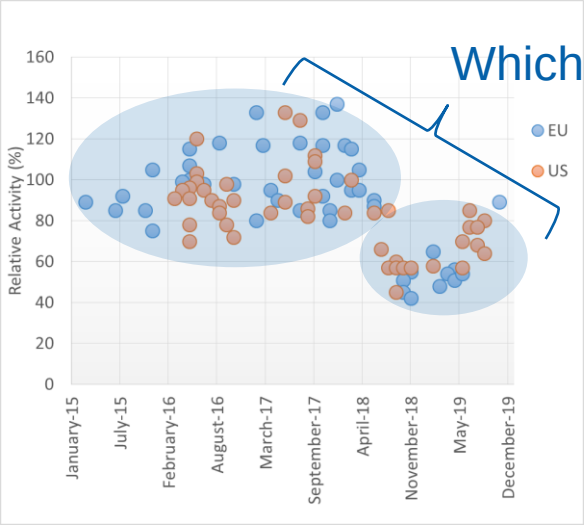




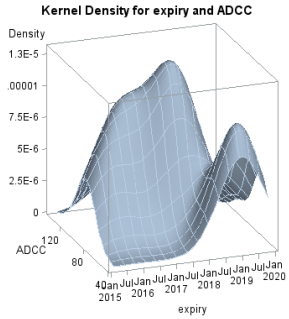
An equivalence test for the mean value is **misleading**, a comparison of ranges is more appropriate

Reference: Sandoz presentations for the July 13, 2016- Meeting of the Arthritis Advisory Committee (FDA)

Central Limit Theorem for bimodal data



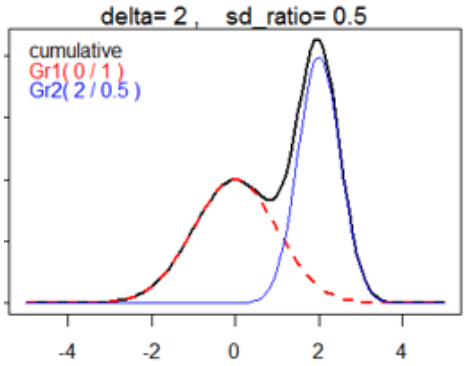
Which distribution does the mean have?

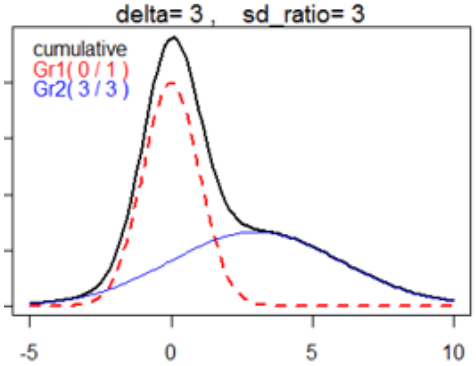


reproduced from Kim S, et al. mAbs 2017;9(4):704-714

Simulations:
2 groups with normal distribution
delta (=mean difference) from 0 to 3
sd_ratio (=sdTEST/sdREF) from ~0 to 3
Anderson-Darling test to check for normality

Result: sample size of >=30 needed for normality of the mean



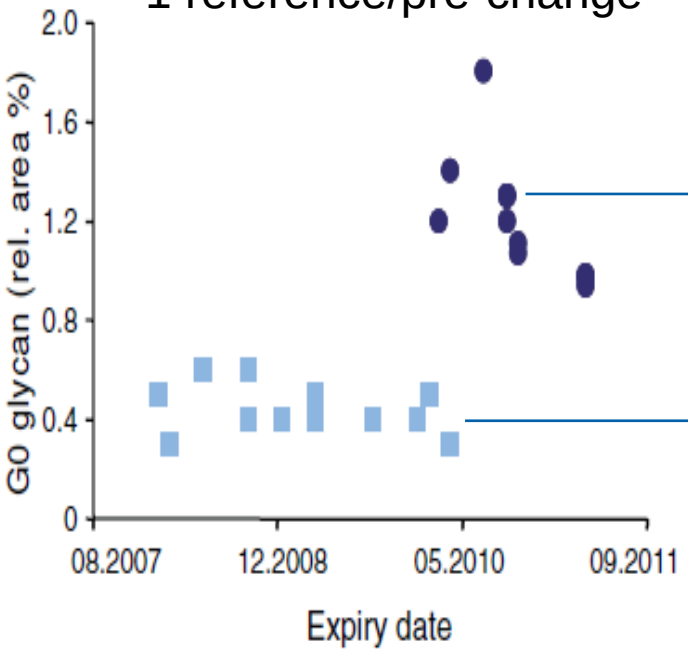


If sub-groups are non-normal?

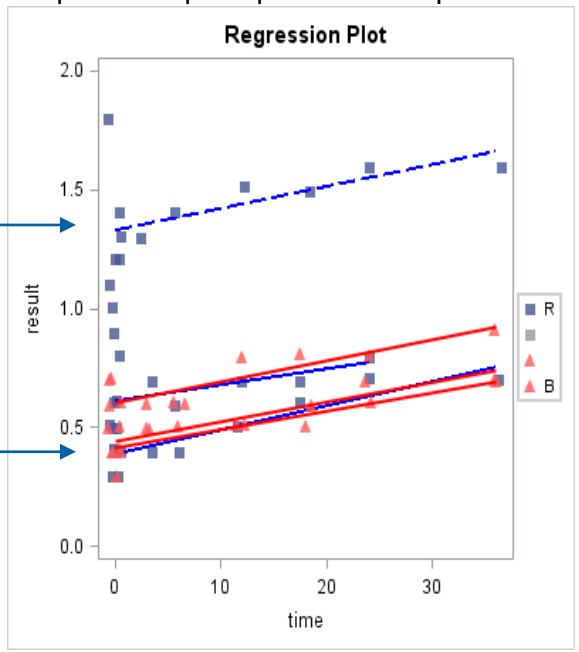
Does this mean that the equivalence test can only be applied if sample size is >30?

Changes in production: stability data are also in a wide range / common intercepts may differ

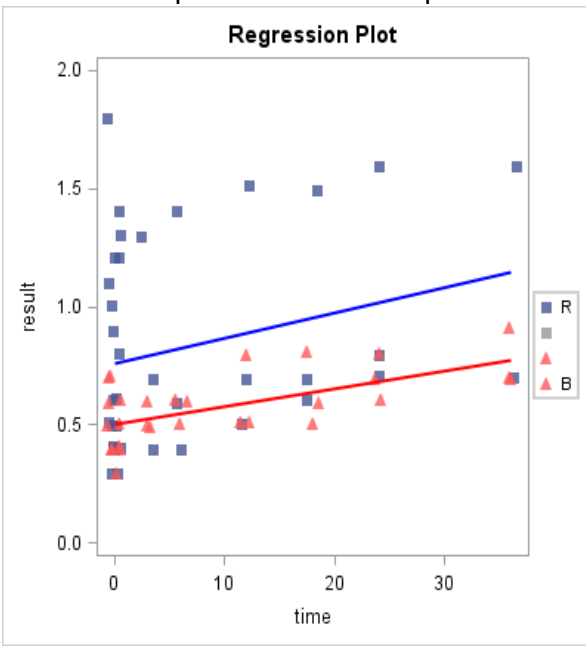
1 reference/pre-change



separate slope separate intercept



common slope common intercept



artificial dataset

Schiestl M, et al. Nat Biotechnol. 2011;29(4):310-312

in red: biosimilar/post change

Production changes may lead to a wide spread in stability data.
An equivalence test for the intercept is misleading.

A comparison of ranges is appropriate. A comparison of slopes is appropriate, if sample size is high enough.

ebe
european biopharmaceutical enterprises

Stability comparison: quality of data

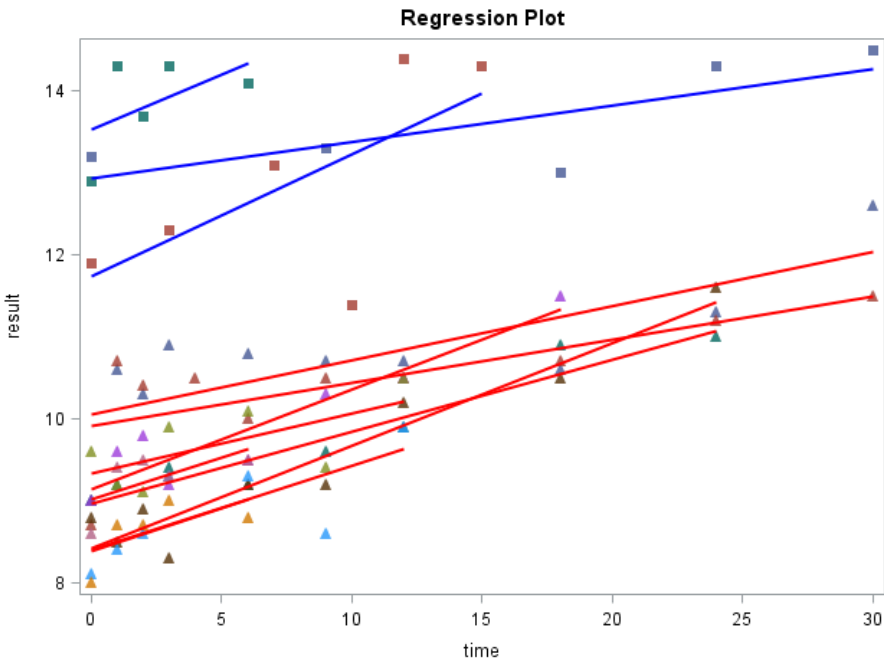
following requirements on quality of data:

1. close to normal data over time (normal distr. of residuals)
2. Quality attribute changes over time
3. Quantitative, i.e. numerical data on a continuous scale
4. sufficient resolution (data should not be over-rounded)
5. Non-censored data, e.g. „<LOQ“

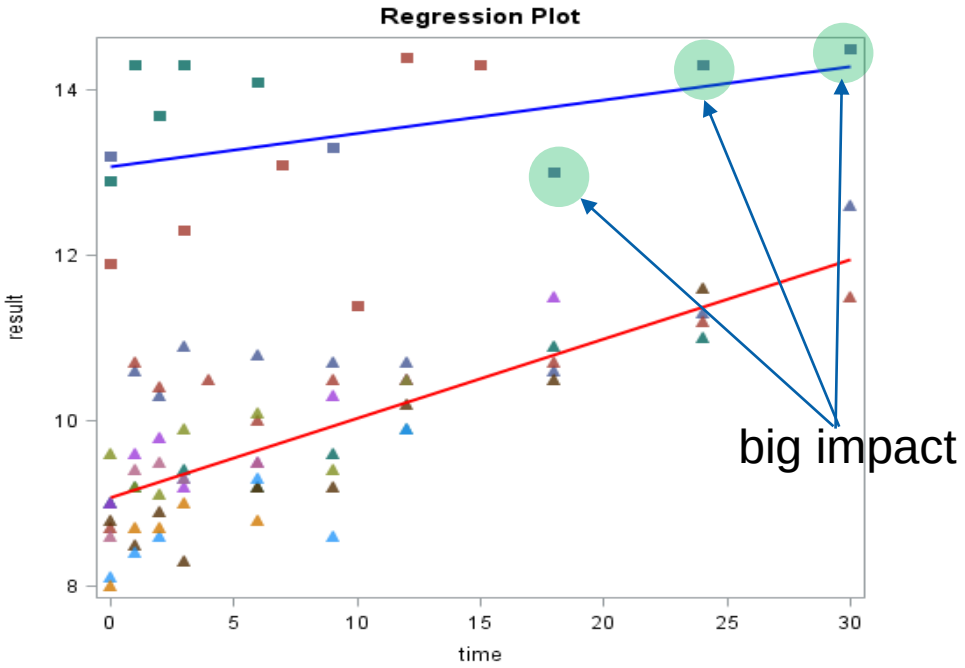
If the requirements are not fulfilled, only descriptive tools for data comparison can be used.

Stability comparison: sparse data at the end of timeline have high weight on slope estimation

Separate Slope Separate Intercept (SSSI)



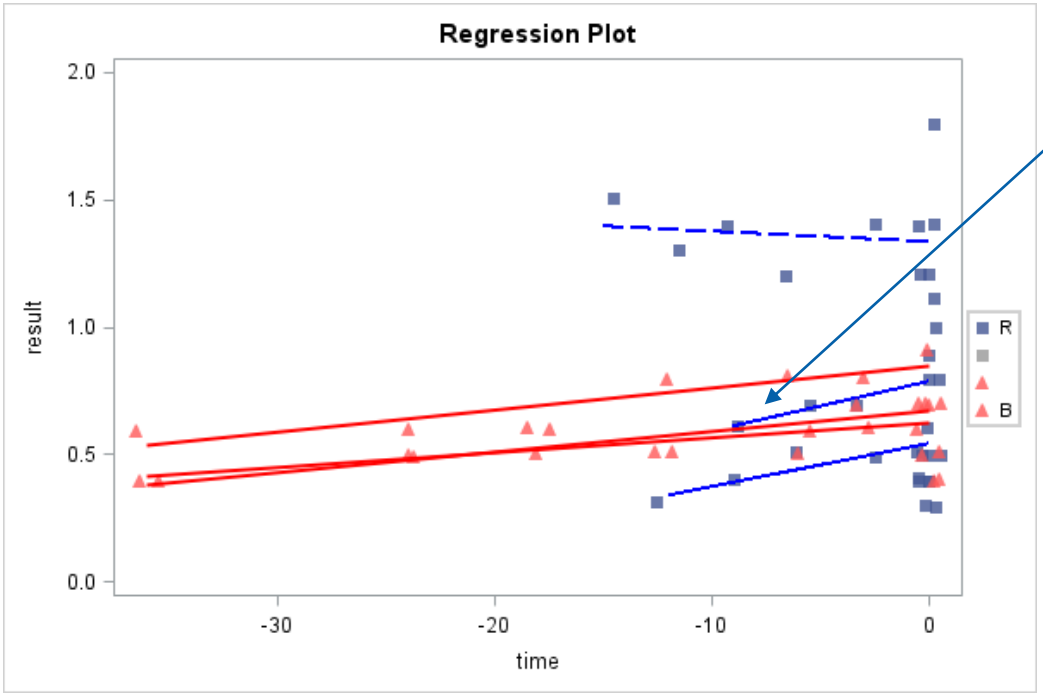
Common Slope Common Intercept (CSCI)/group



An inferential comparison of slopes is feasible, if:
1) test results are equally spread over time
2) no. of batches and no. of timepoints are sufficient

Stability comparison: special case biosimilarity

Separate Slope Separate Intercept (SSSI)



No release data are available for the reference product.

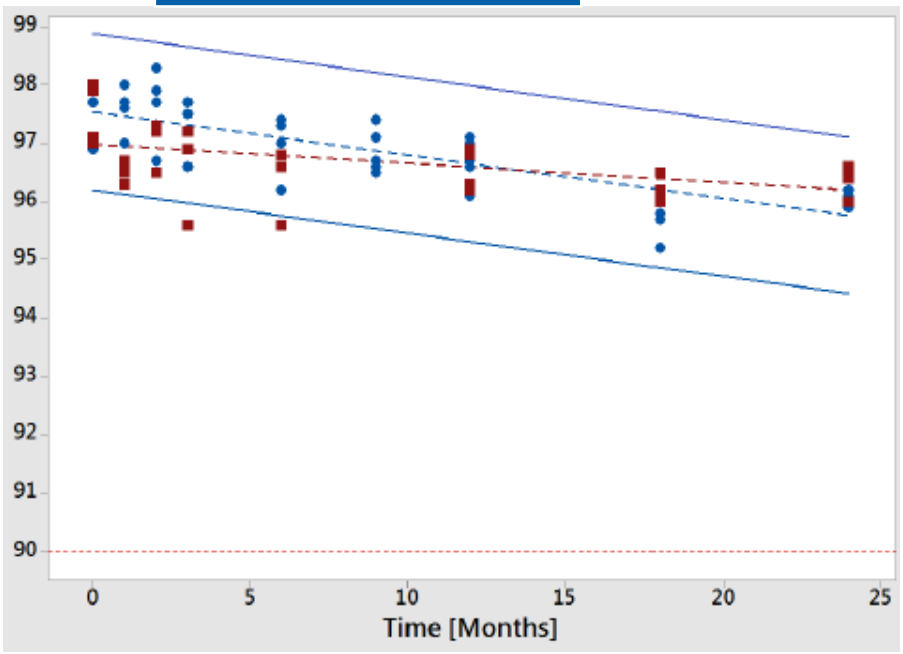
Time to expiry short.

blue: reference
red: biosimilar

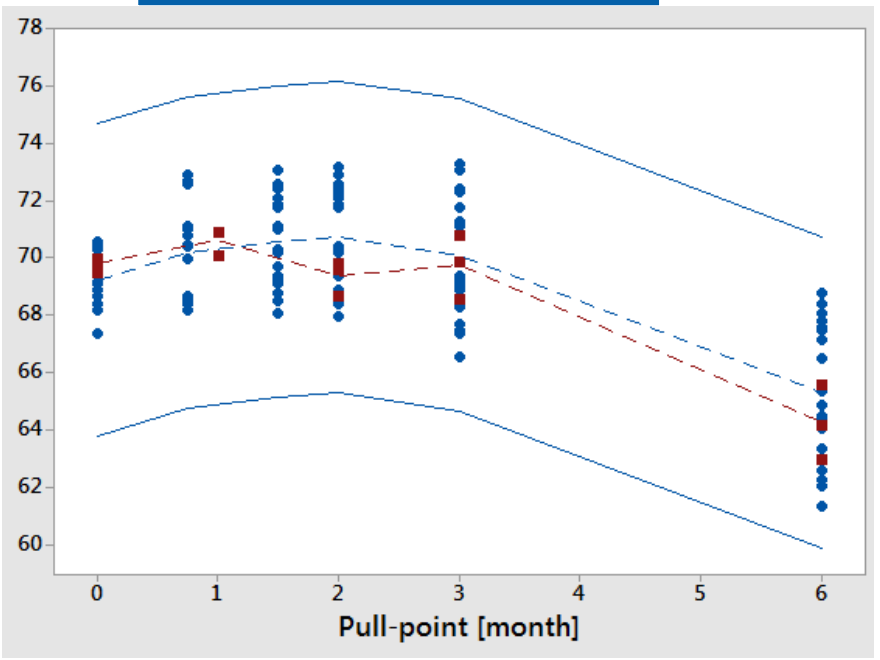
Inferential comparison of stability data (at intended storage) in analytical similarity can be misleading.
→ comparison of ranges

Linear vs non-linear trend: Comparison of ranges

linear degradation



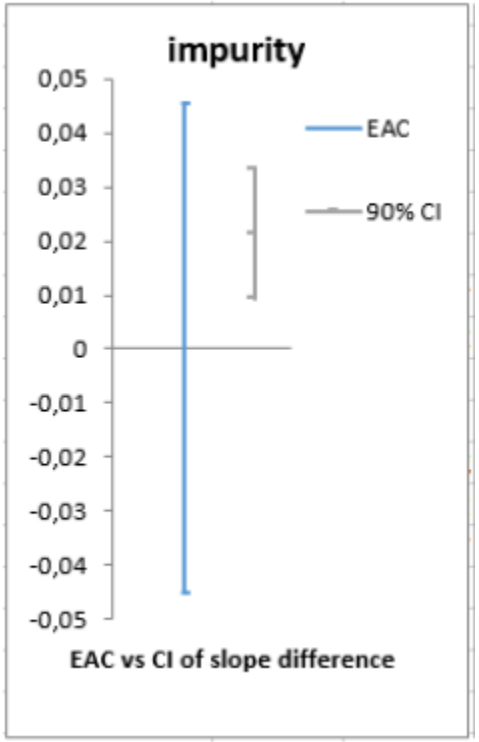
non-linear degradation



equivalence test on slopes is complicated, if the trend is not linear → range test is appropriate (e.g. 3*averaged sd)

Equivalence test for slopes

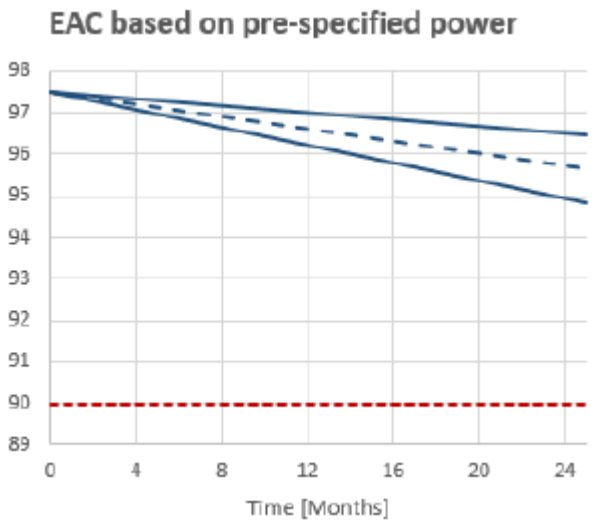
Slopes of the linear model show the trend over time on average, thus representing a mean.



for slopes: equivalence test is appropriate to show that the trend over time of both products is the same.

Difficulty: setting of EAC

example:



Sample size needs to be sufficient

Stability comparison: different readouts and different conditions

Statistical tool depending on readout:
1) single results or slope
2) trend / no trend
3) linear/non-linear trend
4) minimum number of batches and time points

Conclusion

An equivalence test is applicable, if:

- The mean is the readout of interest
- The slopes (no intercepts!) are the readout of interest
- There is no special cause variation in the data (e.g. mostly/only analytical variability of batch results)
- Multiplicity problem is controlled
- Sample size is high

A comparison of ranges is applicable in all other cases:

- Ranges or single results are the readout of interest
- More than one mode is expected to be in a group of data (i.e. production data)
- Variability of reference is much higher than that of the biosimilar (i.e. intercepts)
- Stability trend is not linear (non-normal residuals with linear model)
- Sample size of one or both groups is small
- Stability results clustered/not spread over time range

Conclusion for analytical similarity

biosimilarity comparison has special features:

- manufacturing date of the reference batches not known →
- differing start date of stability until end of shelf-life →
- time of stability is short and thus slope is variable for reference
- high variability and many production changes in the reference data

A descriptive range comparison is recommended for analytical similarity of stability data.

Acknowledgement

Schmelzer Bernhard, for providing results on the calculations of comparison of stability data and many other colleagues from EBE and Vaccines Europe

Thank you