BTDM, NTO
Novartis

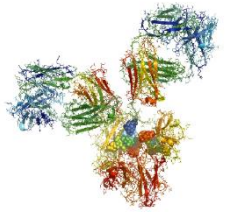
# Simulations to test properties of equivalence testing

**Franz Innerbichler, Statistical Process Expert,**
**Head of Statistical Functional Network STAMODA**
**October, 2017**
franz.innerbichler@novartis.com



NOVARTIS

# Statistics in development for mAbs



## Many pharmaceutical topics are amenable for statistical comparisons

**Comparison of processes**
Comparability
Analytical similarity
Scale down model qualification

**Comparison of PK response**
Bioequivalence trials

**Comparison of analytical methods**
Method transfer studies

## Statistical tools for comparison:

**Range approach**
mean ± k standard deviations
tolerance intervals
min-max range

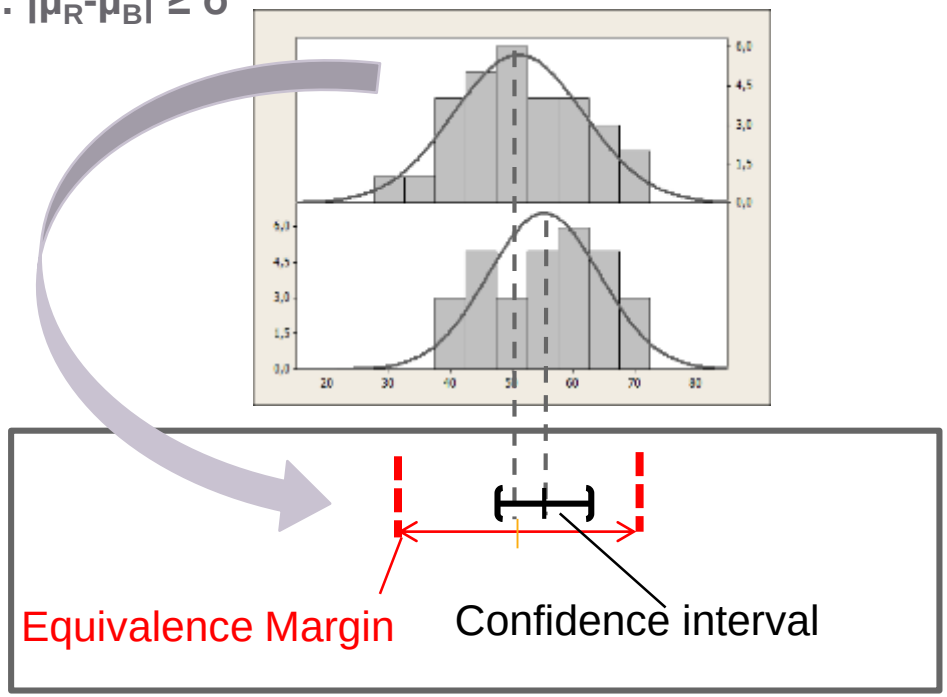
**Frequentist approach**
Equivalence test

**Others**
Bayesian statistics
non-parametric approaches

*But what is the right tool for which scientific question?*

# Equivalence test: Pharmaceutical topics and variabilities

H0: $|\mu_R - \mu_B| \geq \delta$




Components of variability

1. **Analytical method transfer**
   Analytical variability, **no** process variability
2. **Bioequivalence trials**
   Individual patient response and analytical variability**, no** process variability
3. **Analytical biosimilarity**
   analytical variability **plus** process variability

**Process variability only effects analytical similarity**



NOVARTIS

# Equivalence testing is well established for clinical trials

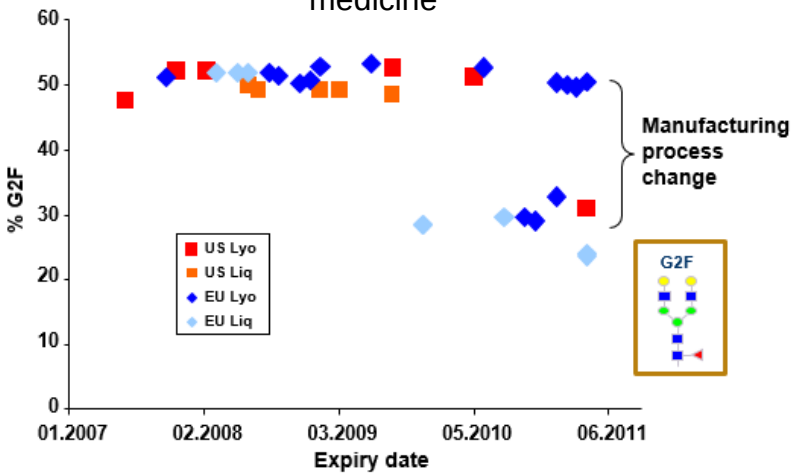
| | Equivalence testing in clinical trials | Equivalence testing in comparability / biosimilarity |
|---|---|---|
| **Mean** | The mean response to the medicine is a clinically relevant estimator of the treatment effect | Mean represents the process mean; important for the patient safety and efficacy is the range of the batches |
| **Relevance of differences** | A difference in the mean (or the variability) may indicate a **clinically relevant** difference in treatment effect | A difference in the mean or in the variability is **clinically not relevant**, as long as individual batches are within acceptable quality range (e.g. as defined by reference biologic / historical data) |
| **Data** | A priori: randomization, stratification, sample size calculation, inclusion/exclusion criteria; to get a statistically meaningful result | No independent and representative sampling possible: auto-correlation, campaign production, sample size not planned (a posteriori) |

NOVARTIS

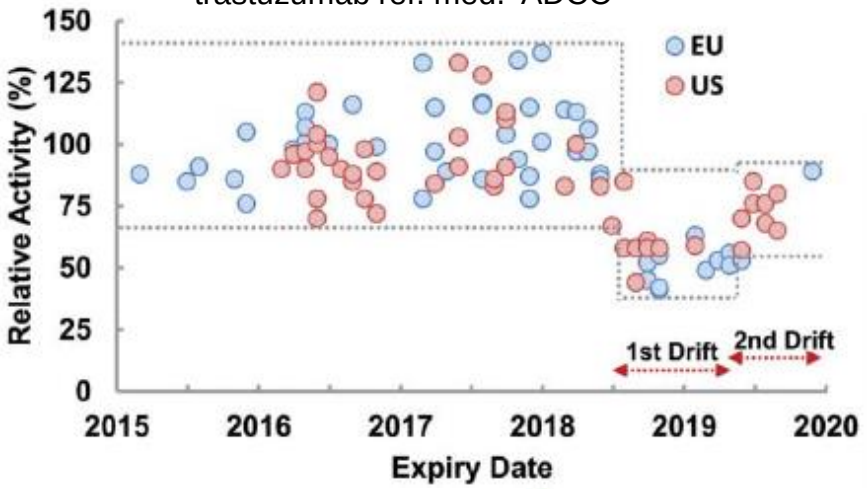
# Process variation in reference biologics

- Analytical data revealed manufacturing changes
- The processes before and after changes were evaluated and obviously deemed to be comparable[§] by regulatory agencies
- The etanercept and trastuzumab reference medicine batches were approved and administered to patients

G2F of etanercept reference medicine




trastuzumab ref. med. ADCC

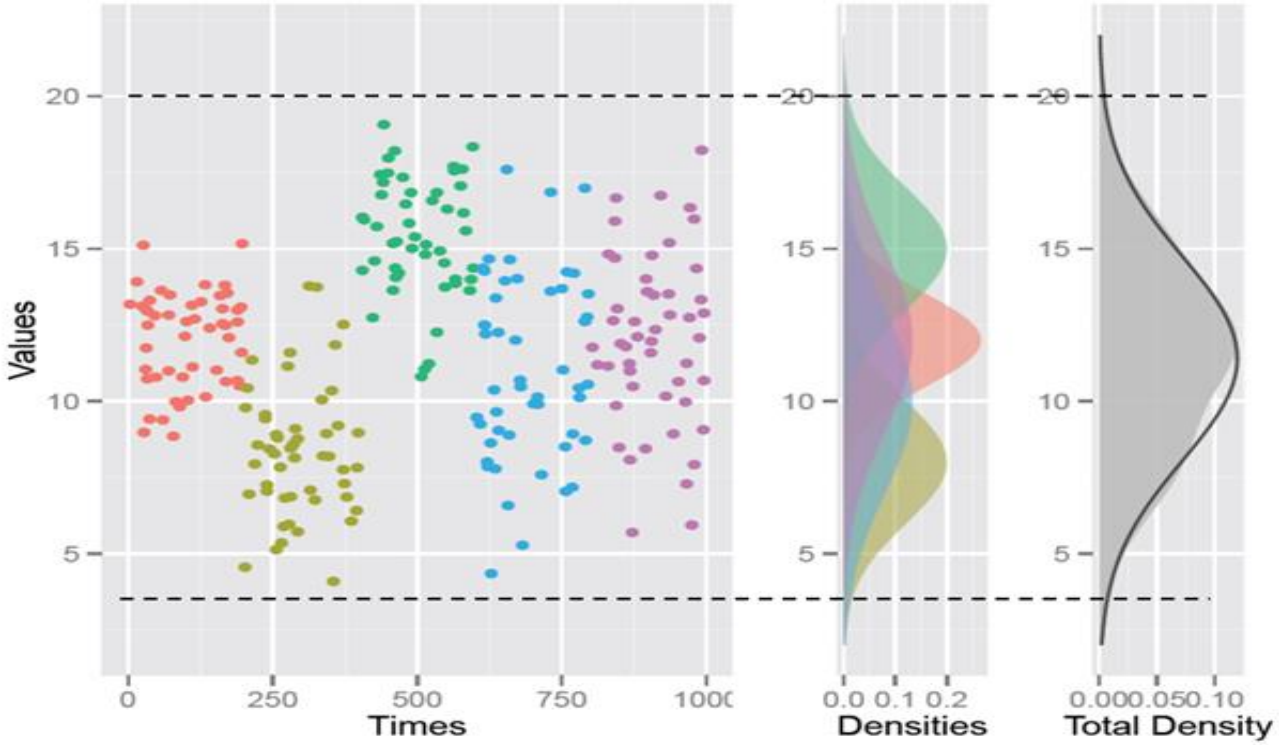



How can these process changes be detected?
By looking at the time scale

*Schiestl M, et al. Nat Biotechnol. 2011;29(4):310-312.*
*McCamish M, et al. Clin Pharmaco Ther. 2012;91(3):405-417.*
*Kim S, et al. mAbs 2017;9(4):704-714*

§ comparable in the sense of ICH Q5E

# Process variation in reference biologics



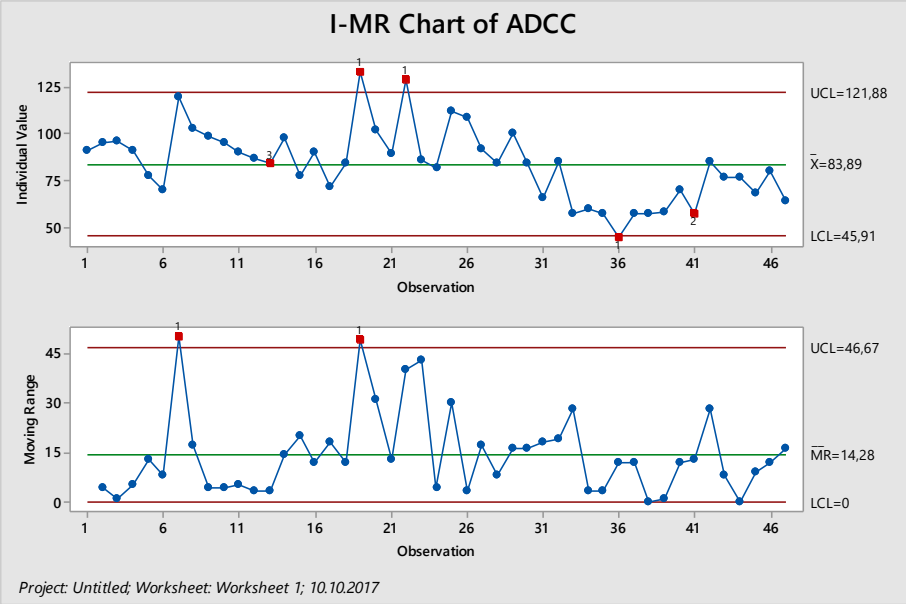

**Process shifts can hide in an overall "normal" distribution**

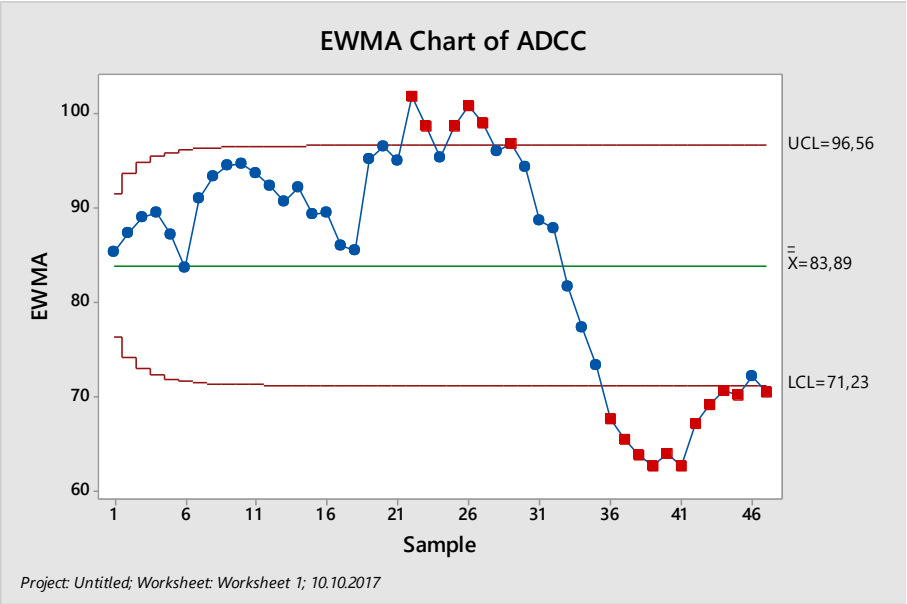
*Data: simulated random normal data resulting in multiple means*



NOVARTIS

# Variation in trastuzumab US reference biologic: mixture distributions



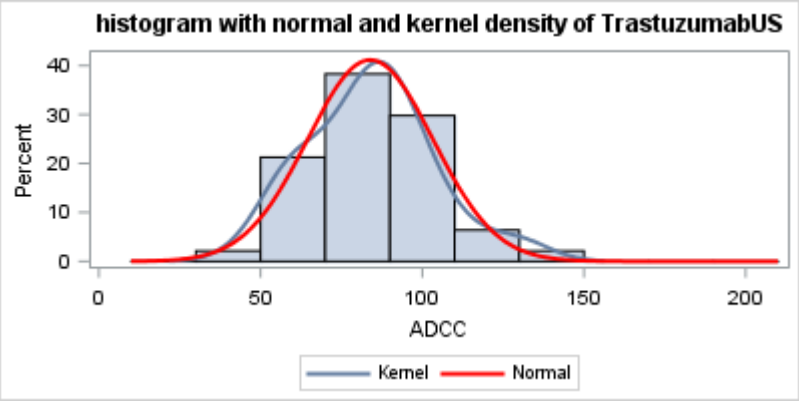



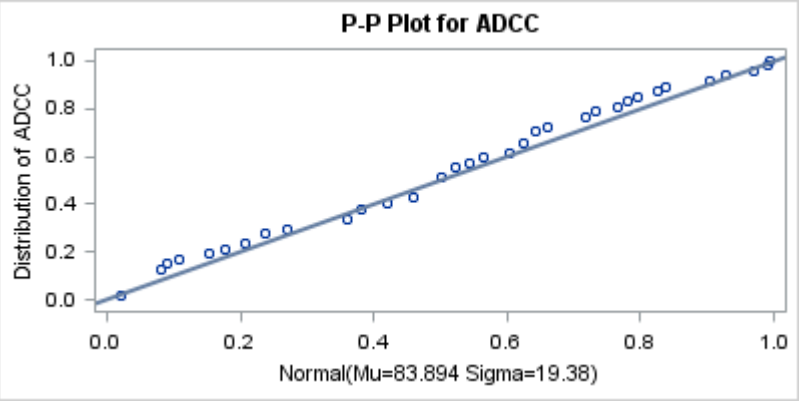

I-MR charts, as common in pharmaceutical industry, show that the process is not stable.
The EWMA chart is more alerting, that there is special cause variation.

*Data extracted from ADCC chart in*
*Kim S, et al. mAbs 2017;9(4):704-714*

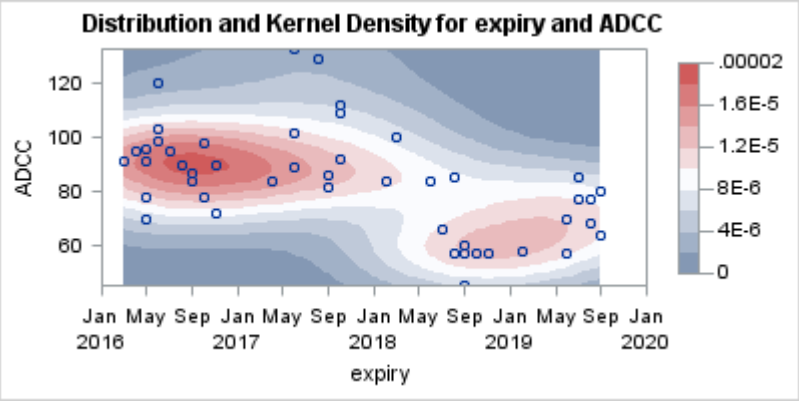
# Variation in trastuzumab US reference biologic: role of manufacturing/expiry date

Multimodality is hard to detect with traditional methods



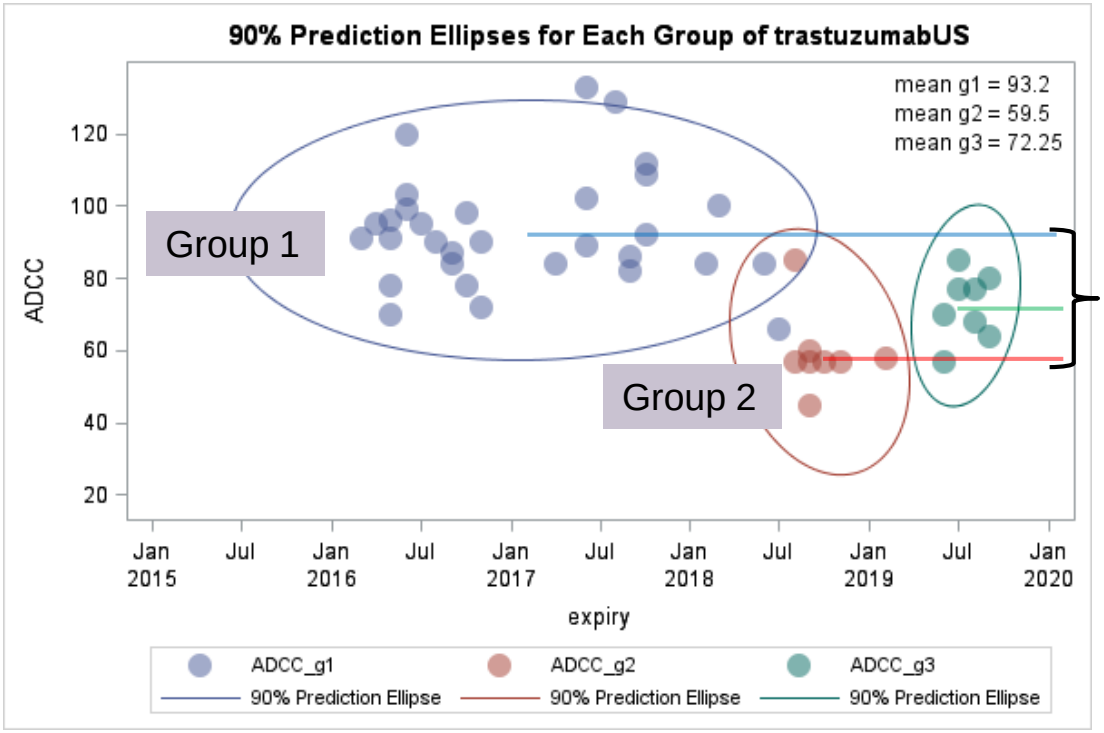




Time axis (manufacturing/expiry date) is crucial to detect mixture distribution



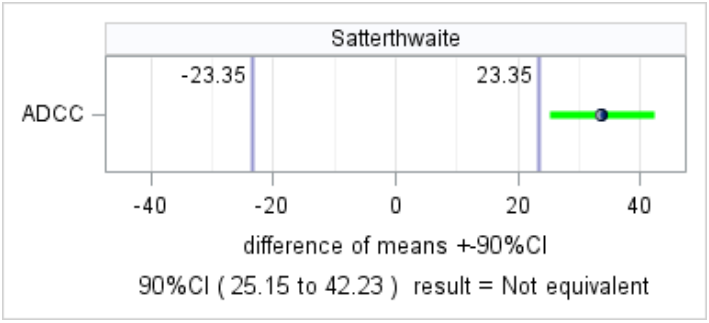

*Data extracted from ADCC chart in Kim S, et al. mAbs 2017;9(4):704-714*

# Variation in trastuzumab US reference biologic: groups not equivalent




Different means → Mixture distribution
→ autocorrelation

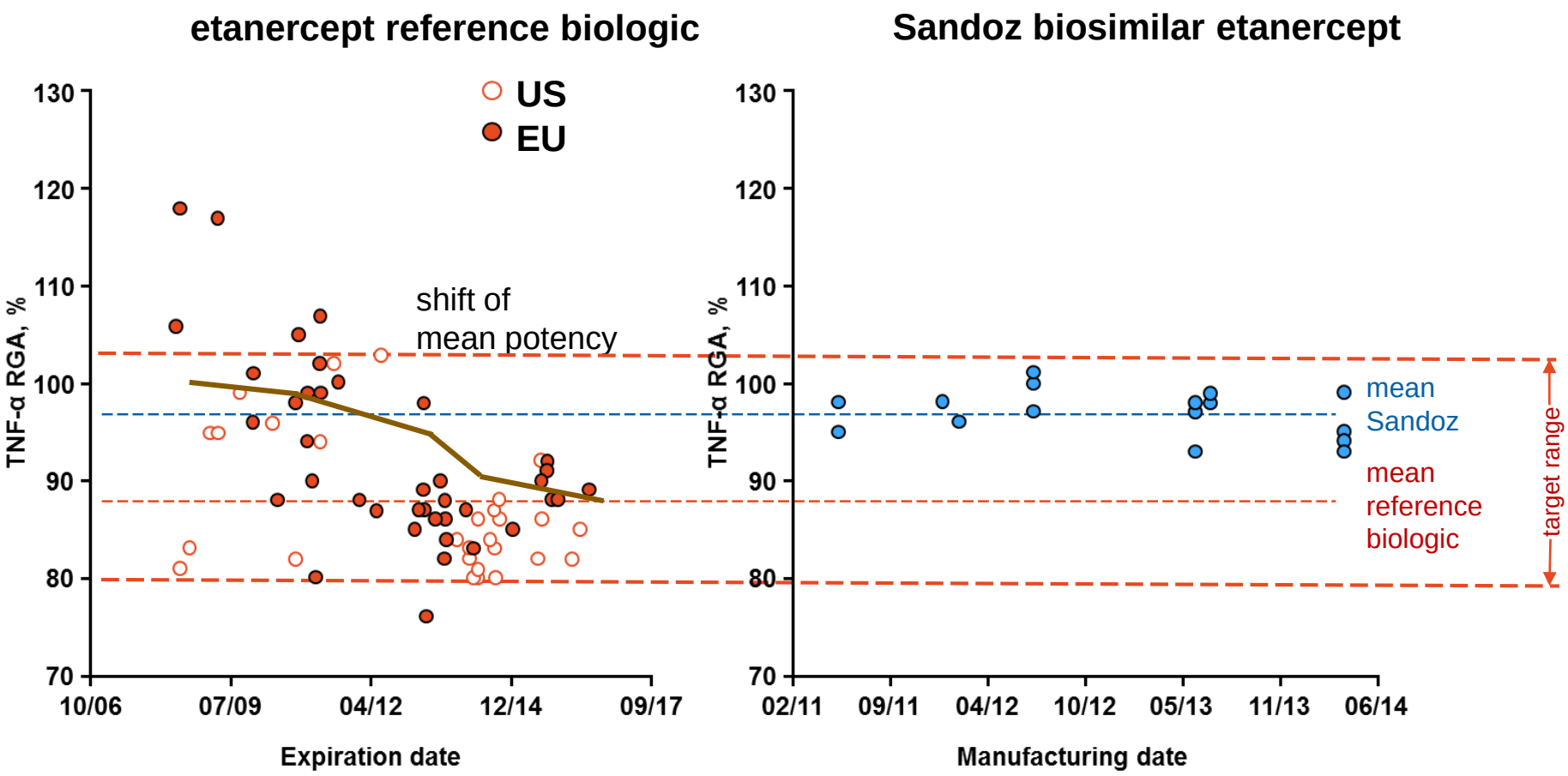
Equivalence test: group1 vs group2




A reference medicine before and after a manufacturing change may not be statistically equivalent. But, they are comparable and therefore „highly similar“ (according to ICH Q5E).

*Data extracted from ADCC chart in*
*Kim S, et al. mAbs 2017;9(4):704-714*

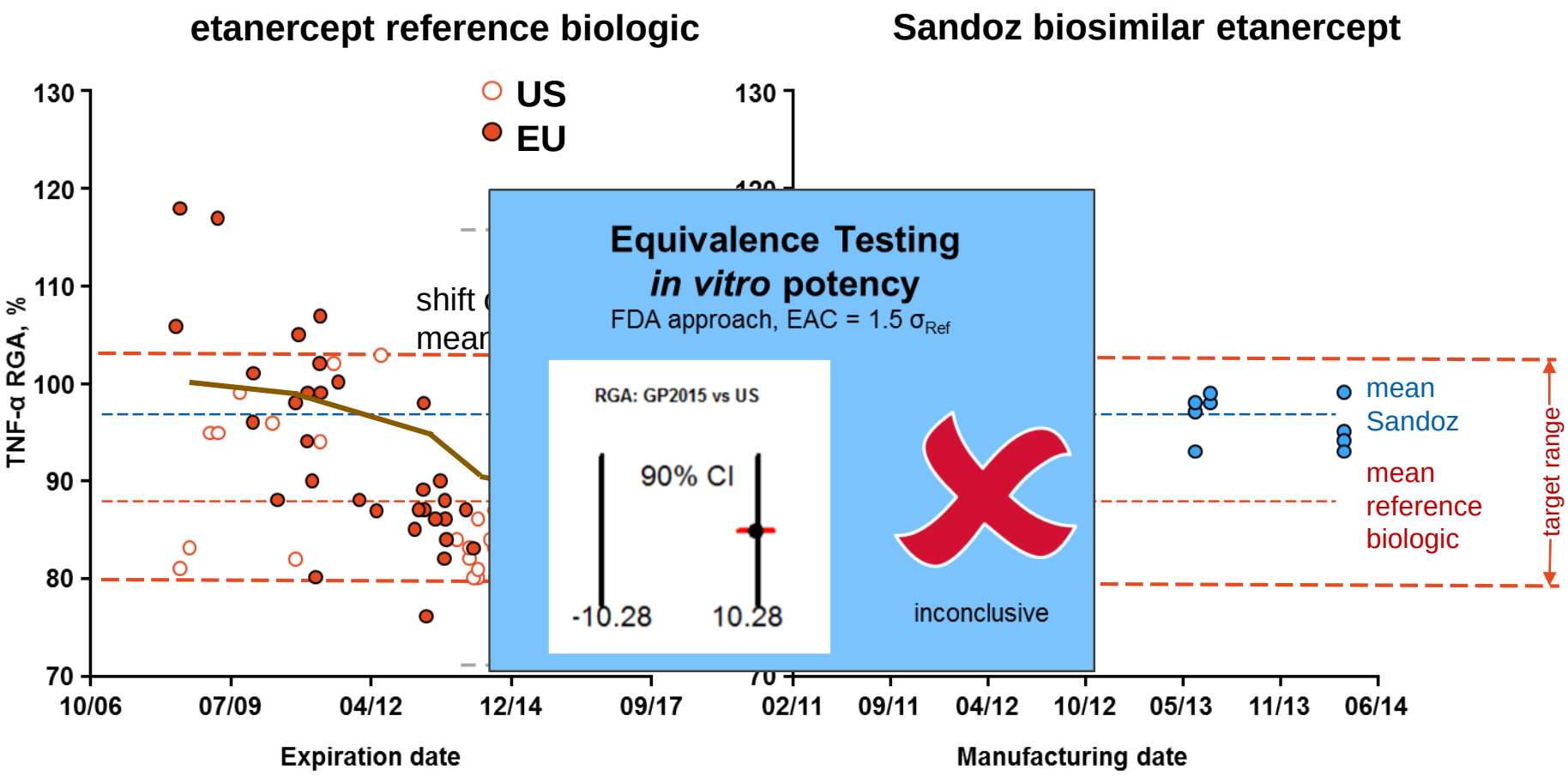
# Variation in etanercept reference biologic: Moving mean




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

NOVARTIS

# Variation in etanercept reference biologic: Moving mean



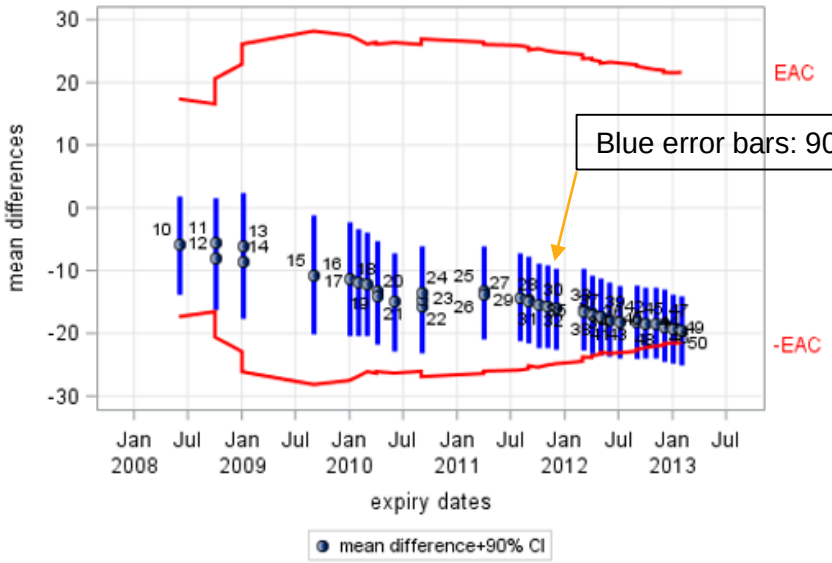

*Reference: Sandoz presentations for the July 13, 2016 Meeting of the Arthritis Advisory Committee (FDA)*

NOVARTIS

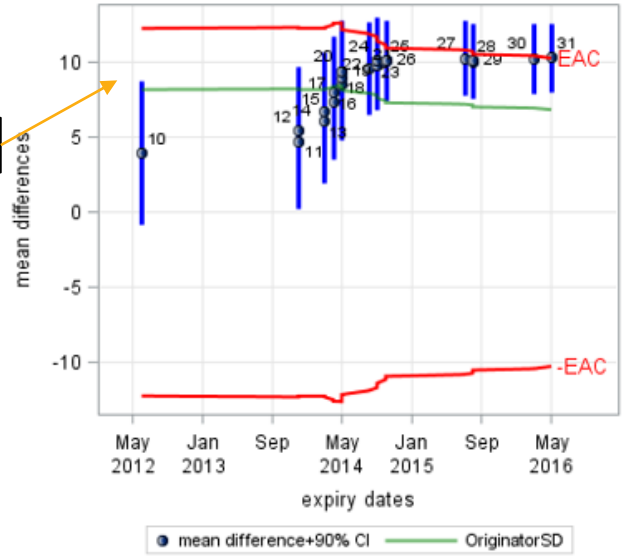
# Variation in reference biologic: not random

**Simulation:** Equivalence tests were done starting with the first 10 reference medicine batches (the biosimilar batches remained always the same), then adding one after the other refernce batch with testing for equivalence at each number of reference medicine batch

Biosimilar 1 (16 batches): data on file
Reference biologic: 10 to 50 batches
ADCC



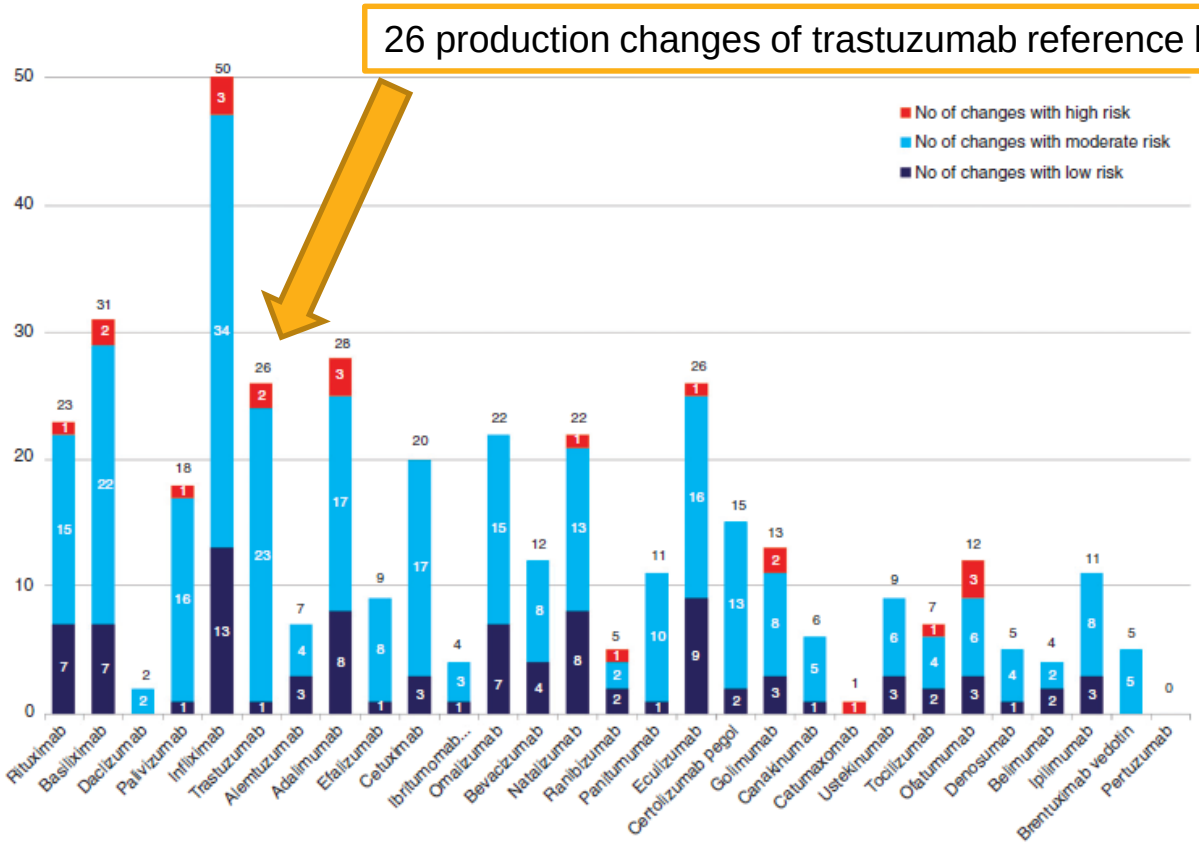

Biosimilar 2 (19 batches): data on file
Reference biologic: 10 to 31 batches
TNFalpha




Analyzing more reference batches over time poses the risk that the equivalence test renders „not-equivalent“§.

§ FDA GfI „Statistical Approaches to Evaluate Analytical Similarity“ draft guidance Sept 2017

# Many manufacturing process changes after approval




## Changes include

- Change in the supplier of a cell culture media
- New purification methods
- New manufacturing sites and conditions
- Changes in batch size

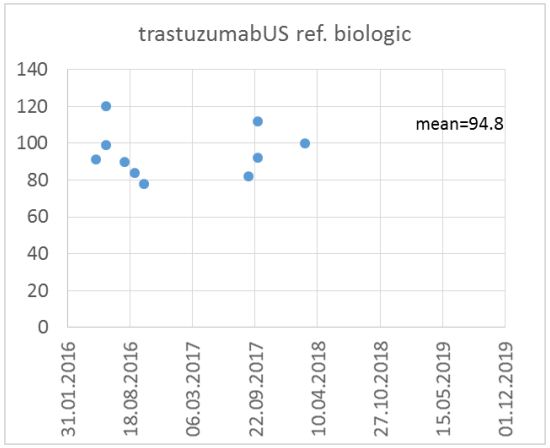
*Source: Vezer B et al, Current Medical Research and Opinion, 2016,32(5): 829-834*

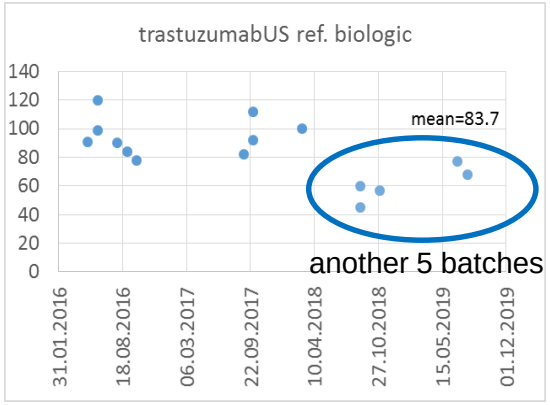


NOVARTIS

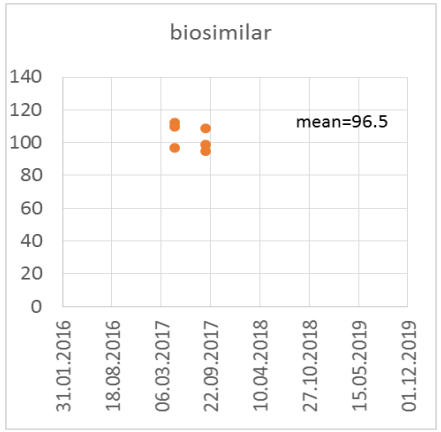
# Implications of sampling reference biologic lots across many years

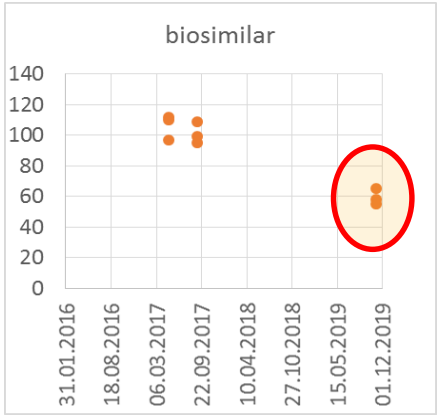
10 random batches

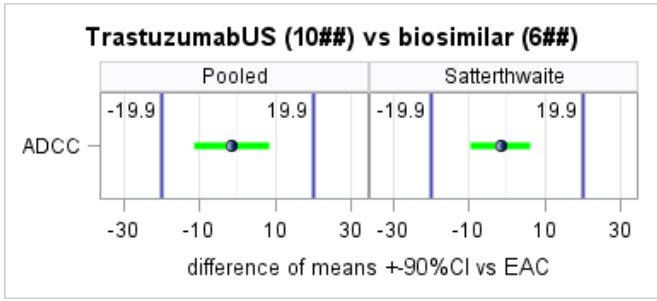


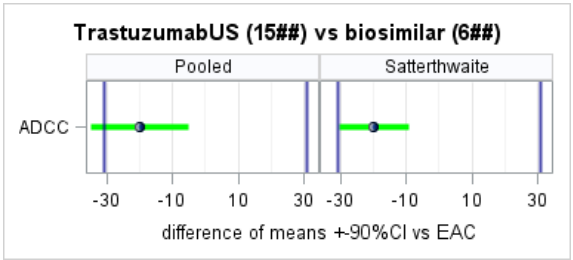



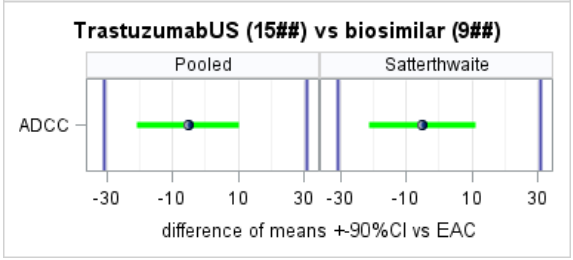

biosimilar 6 batches ~same mean




The biosimilar applicant has to develop 2 different manufacturing processes to become equivalent



NOVARTIS

# Simulation of manufacturing process changes

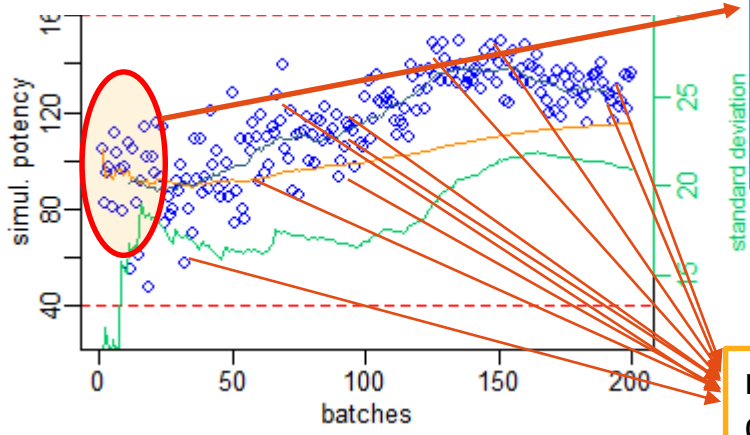
**Simulation:**

- 10 biosimilar batches vs 10 reference batches
- Simulated biosimilar has the same scale and half of dispersion as the starting settings of the reference.
- To account for reduction of process variability over time, the dispersion is reduced by 50% after 75 batches

**Advantages of simulations:**

- Scenarios can be simulated that have no (closed) mathematical solution
- Ideal for random effects
- Easier to be understood (compared to formulas and proofs)

200 batches of reference/distribution are simulated:
no of shifts: ~10
magnitude of shift: N(0,10²)
no of simulated distributions: 500




Simulated biosimilar has the same mean and sd/2 as the first batches of the „reference distribution“

randomly 10 batches drawn = reference (200 times)

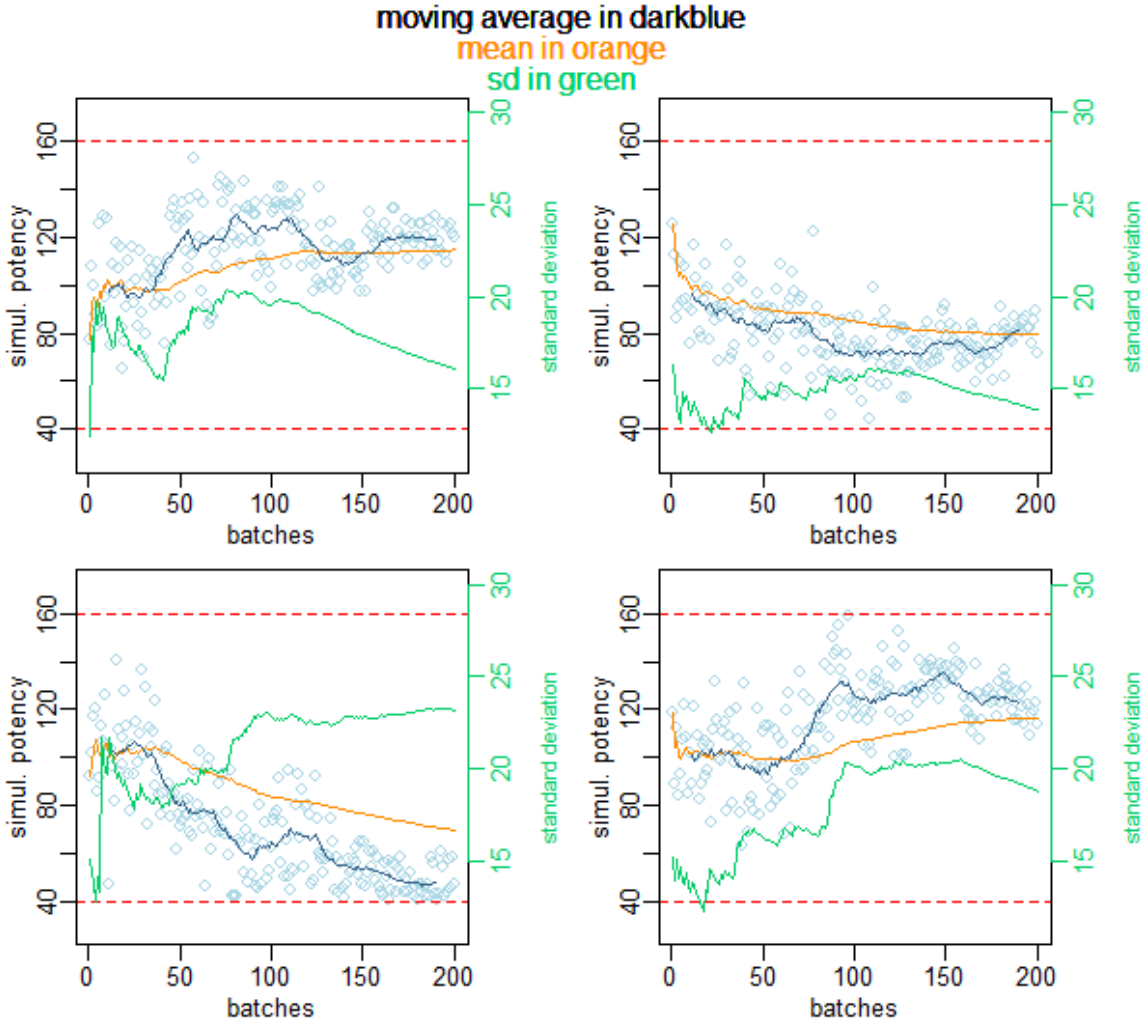
Compare them with an equivalence test (EAC=1.5*sigmaR)



NOVARTIS

# Simulation of manufacturing process changes

4 examples of distributions




**Simulation:**
Task: Frequency of concluding equivalence for 1 QA
No of ## / sim. distribution: 200
no of simulated distributions: 500
no of shifts: ~10 Bin(1, p=0.05)
magnitude of shift: $N(0,10^2)$
Upper and lower limit: 40/160

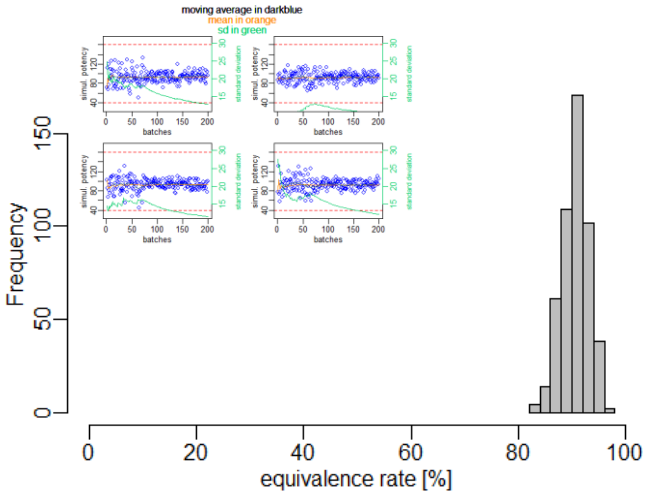
No of bootstrapped samples: 200
no of ## sampled from distribution: 10
no of biosimilar: 10 (mean and sd/2 from start settings of ref distribution)
sd(ref) 1-75=15.568 (TrastuzumabUS has a sd of 15.568 in group1)
sd(ref) 76-200 = 7.784 (simulation of reduction of process variability)
Delta at start of simulation=0
R (3.2.3) and
R Studio (0.99.878)



NOVARTIS

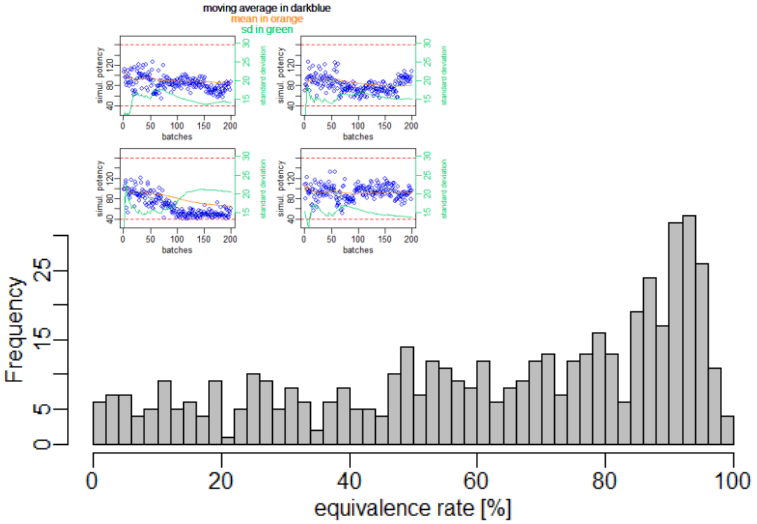
# Simulation of manufacturing process changes

Result without process shifts




equivalence rate ~ 83 - 97.5%

Result with process shifts in reference data




equivalence rate ~ 0.5 - 98.5%
manufacturers of reference medicine have an advantage vs manufacturer of biosimilars in case of many production changes

*Production changes of the reference biologic make biosimilar development a gamble!*

**Simulation:**
As on previous slide, but no shifts

**Simulation:**
As on previous slide
magnitude of shift: $N(0,10^2)$



NOVARTIS

# Equivalence test: probability of success is smaller than expected

1) Power calculation in SAS

**The SAS System**

**The POWER Procedure**
**Equivalence Test for Mean Difference**

| Fixed Scenario Elements | |
|---|---|
| Distribution | Normal |
| Method | Exact |
| Lower Equivalence Bound | -1.5 |
| Upper Equivalence Bound | 1.5 |
| Alpha | 0.05 |
| Mean Difference | 0.125 |
| Standard Deviation | 1 |
| Sample Size per Group | 10 |

| Computed Power |
|---|
| Power |
| 0.873 |

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

2) Simulated „power“ with random numbers from normal distribution

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

| Prob |
|---|
| 0.802 |

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

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

*Burdick R, et al. The AAPS Journal.19,1,4-14, Jan2017*
*Commentary on above: Tsong Y, et al., The AAPS Journal.19,1,15-17, Jan2017*

**Simulation:**
Frequency of concluding equivalence for 1 QA:
n1=10 with N(0;1) ; n2=10 with N(0.125;1);
Normally distributed random samples with seed 654
no of equivalence tests: 1E5
SAS IML 14.1



NOVARTIS

# Equivalence test of several tier-1 attributes (1): probability of success is smaller than expected

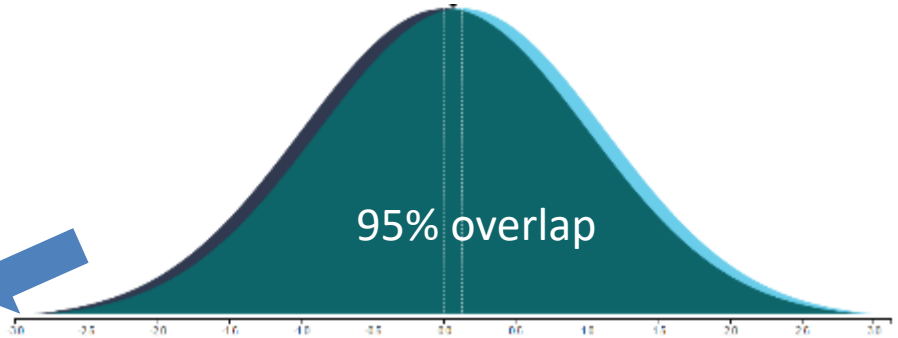
## Multiple testing

**Consider the following example**

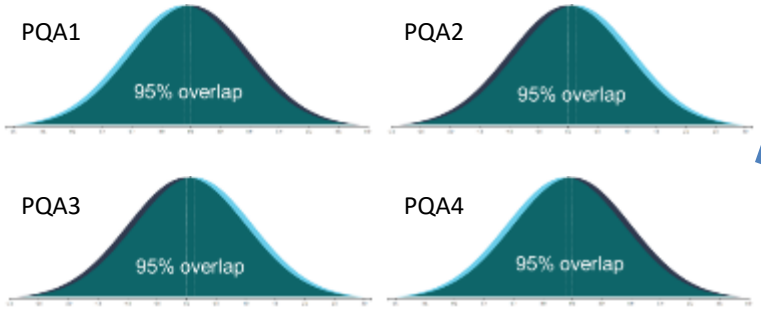
- A biosimilar medicine is developed with very high similarity to the reference biologic
- 4 PQAs assessed using equivalence testing
- The mean of each PQA is very close to the reference biologic mean, $\mu_T - \mu_R = \pm 1/8\ \sigma_R$

**1. Starting with very small differences**

Two normal distributions ($\sigma = 1$) with difference of means = 1/8 sigma




**2. Testing multiple CQAs**




The probability of successfully demonstrating equivalence (10+10 lots) for all PQAs

$$P(\text{success}) = 0.8^4 \approx 41\%$$

**3. Difficult to prove equivalence for all PQAs**

| Number of PQAs = N | Probability of demonstrating equivalence for all PQAs = $0.8^N$ |
|---|---|
| 1 | 80% |
| 2 | 64% |
| 3 | 51% |
| 4 | 41% |
| 5 | 33% |
| 6 | 26% |

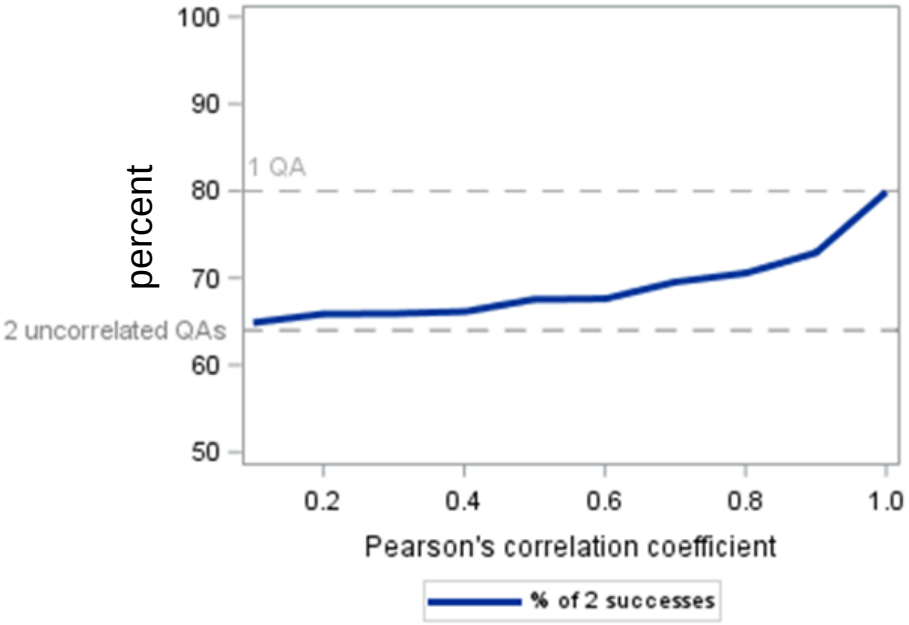
Probability of success of all tests



NOVARTIS

# Equivalence test of several tier 1 attributes (2) Correlation attributes

Is the correlation of tier-1-attributes high enough to outweigh the power reduction of multiple-testing problem?




**Simulation:** 2 QAs with random samples from normal population with increasing levels of correlation (0.1-1): n1=10 with N(0;1) ; n2=10 with N(0.125;1);
no of equivalence tests: 1E5
SAS 9.4 + SAS IML 14.1

2 tier-1 attributes need to be nearly perfectly correlated to show an overall power of 80%, as expected for a single test

**„Real life“ examples of sample correlations:**
Rituximab reference : ADCC and CDC: r = 0.29
Infliximab reference : TNFα RGA and SPR: r = 0.19

*Novartis in-house data*



NOVARTIS

# Equivalence testing is contengious for many reasons:

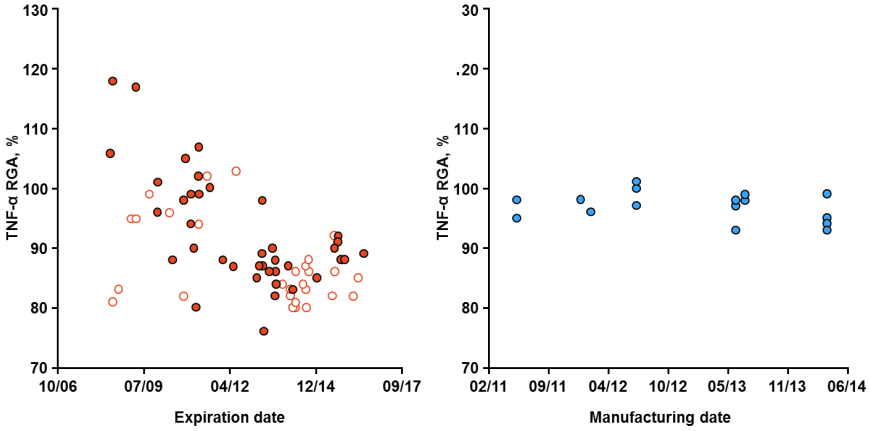
**Whenever the reference biologic undergoes shifts, it becomes a gamble to meet the equivalence test.**

**Multiplicity problem: If 2 or more tier 1 attributes are evaluated with EQ test, overall power decreases considerably.**

**EQ test performance depends on homogeneous, “normal-like” distribution. If time (manufacturing/expiry) is an important co-variable, the type-I-error is not defined.**

# Scientific considerations for analytical similarity

- Safety and efficacy within the reference biologic's variability have been demonstrated in clinical studies and by real-life experience with the reference biologic
- Every marketed batch from the reference biologic defines acceptable quality with respect to its quality characteristics
- A given quality characteristic of a reference biologic lot is acceptable for a biosimilar lot

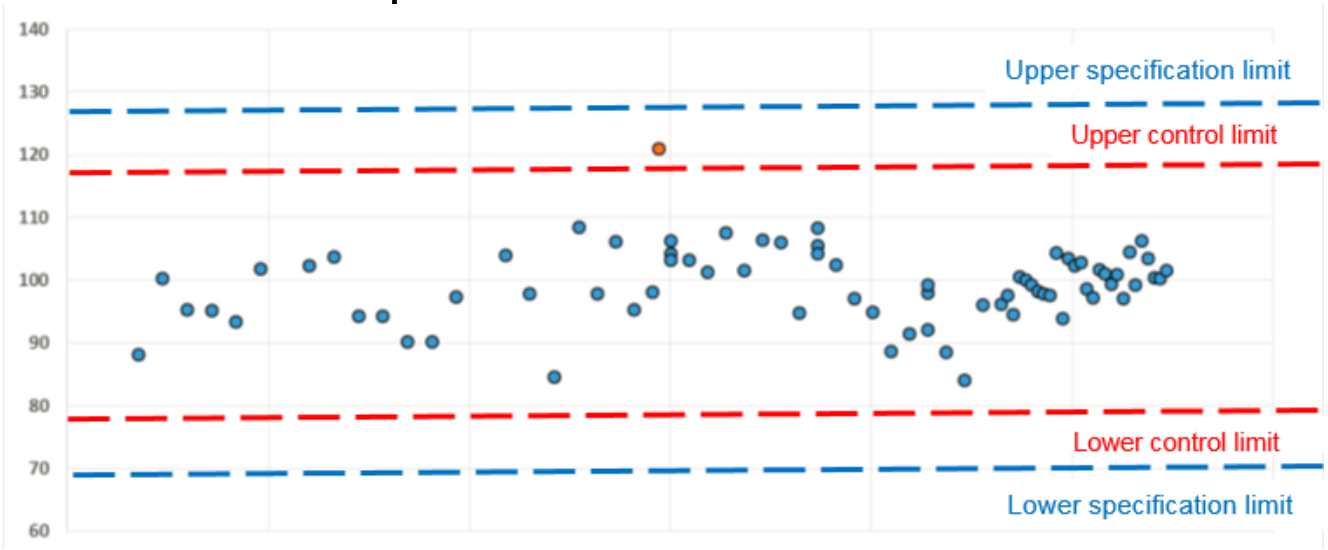



Mark McCamish & Gillian Woollett (2011) Worldwide experience with biosimilar development, mAbs, 3:2, 209-217, DOI: 10.4161/mabs.3.2.15005

NOVARTIS

# Test of means is at odds with regulations to control manufacturing

- ICH Q6b, Q5E, Q7, Q8, Q11 require compliance with ranges, they do never require consistent mean
- Thus, the mean can change over time and still represent consistent quality and clinical performance
- Acceptable ranges are proposed and justified by the manufacturer and approved by regulatory authorities
- Requirement for consistent mean for biosimilars would need a rework of the ICH guidelines → to require consistent mean in manufacturing of reference products




NOVARTIS

# Final thoughts

- EQ results should not be used as pass/fail criterion in analytical similarity assessment.
- Whenever the mean of the reference medicine changes over time, equivalence test may lead to wrong conclusions.
- Ensure fit-for-purpose, balanced, evidence-based and consistent regulatory requirements for all medicines, not for biosimilar medicines only.

Thank you very much for listening and the possibility to speak here at this great conference

# Acknowledgements


- Shu-Yi Su
- Matej Horvat
- Bernhard Schmelzer
- Thomas Stangler
- Florian Wolschin
- Andreas Seidl
- Martin Schiestl
- William Lamanna
- Uros Urleb
- Jens Schletter
- and many other colleagues from Novartis BTDM and Sandoz Biopharmaceuticals


NOVARTIS

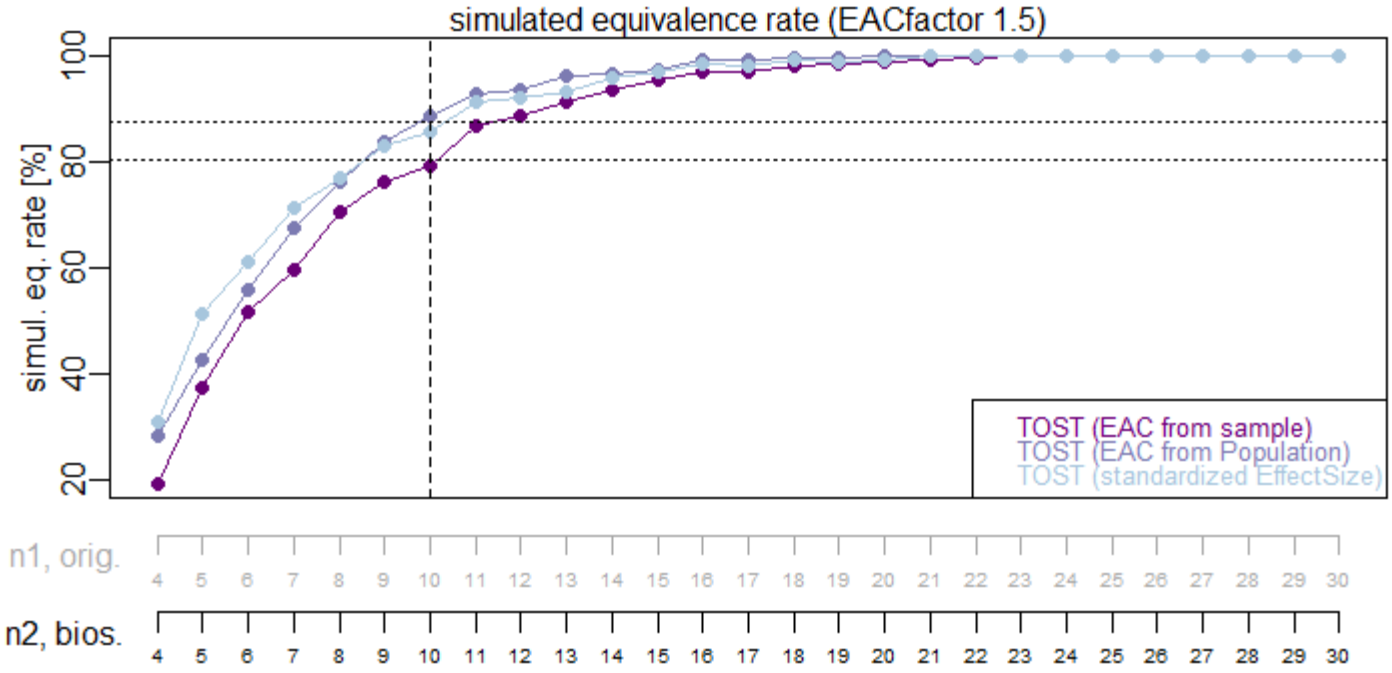
# Backup

# Comparison: sample, sigma, standardized Effect size

R. Burdick: Construct the confidence interval directly for the effect size:

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

$$H_A : -1.5 < \frac{\mu_T - \mu_R}{\sigma_R} < 1.5$$






NOVARTIS