Statistical Methods for Analytical Comparability Assessment During Process Change

Presented by NCB Comparability Working Group:

Aili Cheng (Pfizer), Bill Pikounis(JNJ), ***Irina Gershgorin(Novartis)****, Stone Wang (Novartis), Sayli Vijaykumar Pokal (Zoetis),* ***José Ramírez (Kite, a Gilead Company)***

10th IABS Statistical Workshop, November 12-14, 2024, Rockville - U.S.A

Agenda

- Introduction
- Design
- Statistical methods
- Conclusion

Introduction

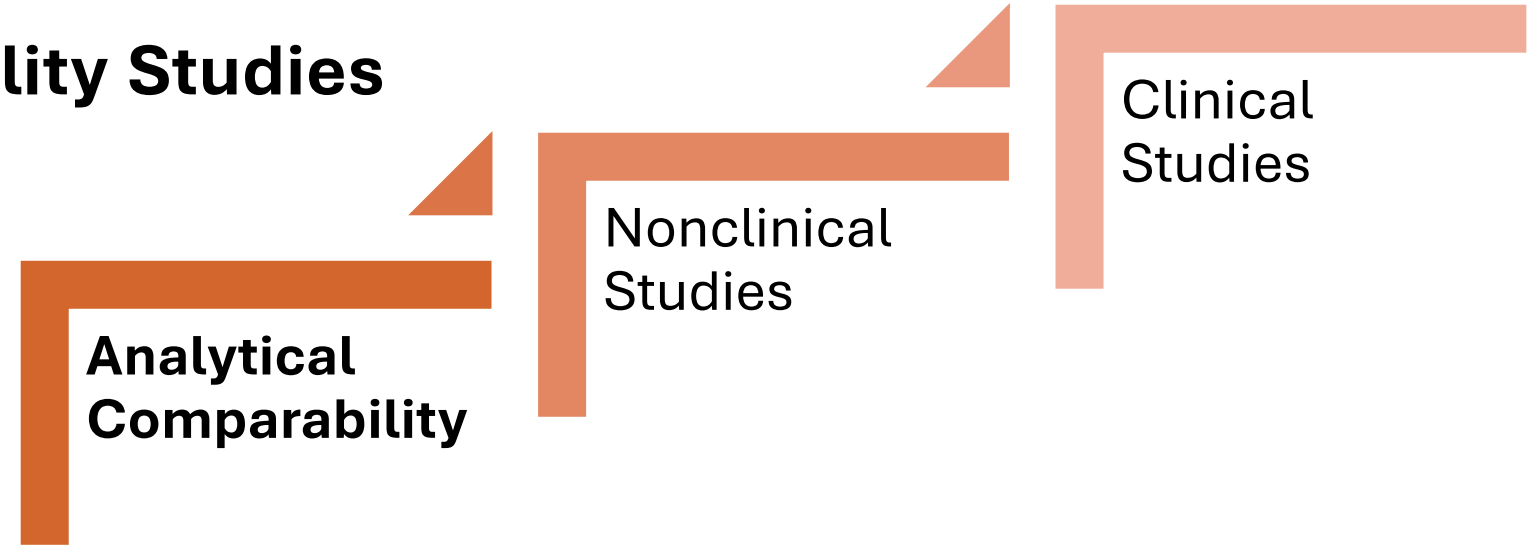
In the development of biotechnology products, changes typically occur for legitimate reasons

- Scale up
- Process improvements
- Improving quality or stability
- New manufacturing site

In these circumstances a comparability exercise is conducted to compare the pre-change and post-change products according to Guidance ICH Q5E

ICH Guideline. Q5E 2004.

Comparability Studies



- Release tests or side by side
- Characterization tests
- Stability: Real-time, accelerated, stressed

The focus of the talk is on the comparability of release or side-by-side data

"A determination of comparability can be based on a combination of analytical testing, biological assays, and, in some cases, nonclinical and clinical data. If a manufacturer can provide assurance of comparability through analytical studies alone, nonclinical or clinical studies with the post-change product are not warranted." ----ICH Q5E

Health Authority Expectations: July 2023 FDA Guidance "Manufacturing Changes and Comparability for Human Cellular and Gene Therapy Products"

We recommend that you consult with a statistician before discussing the study design and statistical approach with FDA. In general, there could be multiple appropriate statistical methods that may be used to evaluate whether data from the post-change product are within predetermined acceptable limits. To avoid errors in the design and analysis of comparability studies, a careful consideration of fundamental statistical concepts is important. For example:

- Some statistical methods may be inappropriate for a given comparison due to invalid assumptions, a need for a very large number of samples, high variability in sample data, or limited information about the population distribution. For

Health Authority Expectations

EMA Reflection paper on statistical methodology for the comparative assessment of quality attributes in drug development: July 2021

Considerations for the Development of Chimeric Antigen Receptor (CAR) T Cell Products: March 2022

FDA Draft Guidance for Industry: Statistical Approaches to Evaluate Analytical Similarity: Sep 2017, withdrawn June 2018

FDA Draft Guidance: Development of Therapeutic Protein Biosimilars: Comparative Analytical Assessment and Other Quality-Related Considerations: May 2019

Statistical Approaches to Establishing Bioequivalence: February 2001, December 2022 (Foundation is Schuirmann 1987)

FDA Draft Guidance Manufacturing Changes and Comparability for Human Cellular and Gene Therapy Products. July 2023

CGT Study Design Challenges

- In autologous cell therapy, each batch is produced from a donor
  - Variability between donors is high
  - Due to difference in genetics, cell composition, cell state, or any prior treatments
  - “A split-source design limits the impact of cellular variability by splitting individual cellular source materials into two equal portions.” (FDA 2023 draft guidance)
- Material can be taken from both healthy donor or individual patient
  - Healthy donor material is often used for ethical, logistical, and practical considerations
  - Healthy donor material is a good representative model to capture wide variability
- Tools and standardized analytical methods for both cell and gene therapy are still being developed
- Greater considerations may need to be given to storage conditions and manufacturing complexity

Study Design

Designs for cell therapy

- Paired (split donor) design:
  - Decreases variability due to batches
  - Increases power
- Often limited sample sizes (eg N=3)
- SME input regarding the scientific/practical/technical considerations of the comparability

Designs for gene therapy

- The number of batches included in the comparability assessment is often determined by the clinical trial supply needs and available material

•Historical data may be used to assess assay variability and set EAC or quality range

Reference:
- FDA Guidance Document, January 2024: Considerations for the Development of Chimeric Antigen Receptor (CAR) T Cell Products

Statistical Methods

- The risk assessment should “inform the statistical approach.” (FDA 2023 draft guidance)
- Both types of methods are mentioned as the possible methods by the FDA draft guidance and EMA reflection paper
- Both have frequentist and Bayesian versions.

453 Your risk assessment should also inform the statistical approach to comparability.
454 Higher risk attributes typically warrant a more stringent statistical analysis than lower
455 risk attributes. Side-by-side or graphical presentations (such as dot plot) to allow visual
456 comparison, in lieu of statistical analysis, may be sufficient for characterization of
457 attributes at low risk of being impacted by a manufacturing change.

Equivalence Testing

- Determines if **mean** shift between pre-change and post-change processes falls within EAC
- Two One-Sided Test (TOST) is a common technique
- Challenge:
  - EAC setting
  - # of batches are usually driven by the supply needs instead of the design

Quality Ranges via Statistical Intervals

- Mean+/-k*SD
- Range for **individual** values.
- Data structure and sample size influence method choice
- Challenge: limited pre-change batches

Equivalence Acceptance Criteria (EAC)

Health Authority Expectations: July 2023 FDA Guidance

"Manufacturing Changes and Comparability for Human Cellular and Gene Therapy Products"

629 through risk assessment, to have a potential to be impacted by the change. For
630 quantitative attributes, a comparability acceptance criterion may be defined as the largest
631 acceptable difference between the pre-change and post-change attribute (an equivalence
632 margin) or as an acceptable range for the post-change attribute (a quality range). In

EAC: The largest **acceptable** difference.

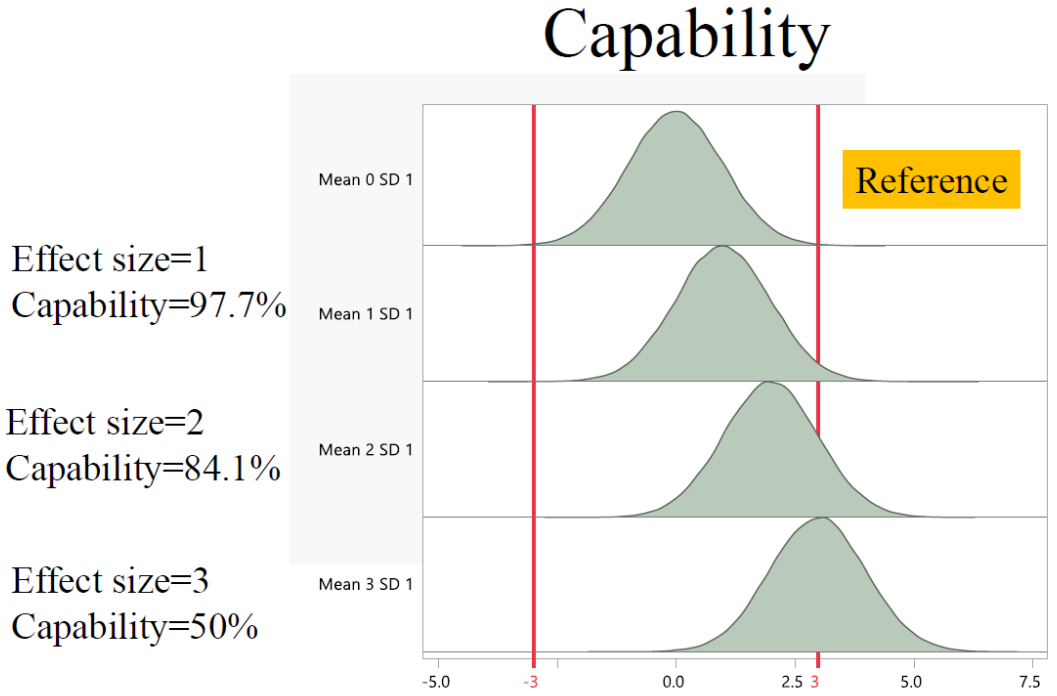
"A comparability range that excludes **any relevant difference**." (EMA)

Window of no practical relevance.

Statistical
EAC Determination for TOST Procedure (1)

Using Capability (Burdick, 2022)

- EMA recommended that acceptance criteria should be established based on similarity condition
- Used capability to define similarity condition
- Capability= defined as the probability that a value from the post-change process falls within 3-sigma limits of the reference process.
- EAC=the effect size corresponding to the “clearly unacceptable” condition which is determined jointly by SME and statisticians
- Sample size/design is determined by power analysis and EAC



$$\text{Effect Size} = \frac{|\text{Difference in means}|}{\text{Reference Standard deviation}}.$$

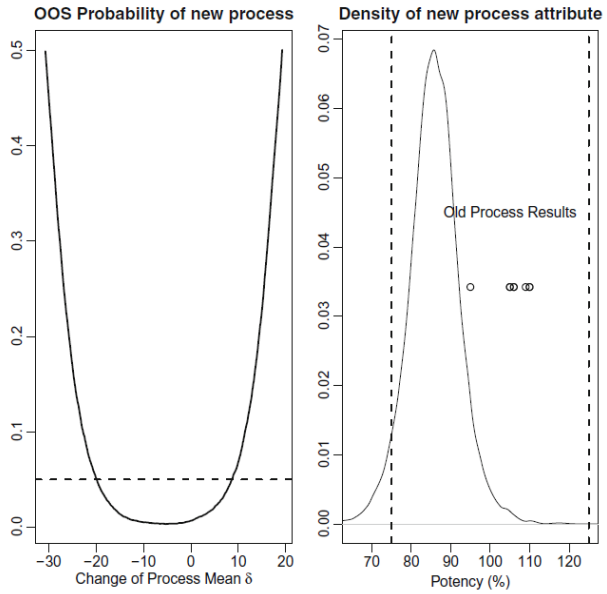
Reference:

- Burdick, R., 2022, Statistical Approaches for CMC Comparability Testing for Gene and Cell Therapy Companies, 8th IABS Statistics Workshop
- EMA reflection paper, 2021, Reflection paper on statistical methodology for the comparative assessment of quality attributes in drug development

Statistical
EAC Determination for TOST Procedure (2)

Based on OOS

- EAC is the mean shift leading to small enough OOS
- Sample size is determined by power analysis
- May not work before specifications are finalized
- Bayesian approach (see ref):
  - OOS rate based on posterior predictive distribution from comparability data and priors from historical data/relevant scientific knowledge.
  - Useful when data is scarce but prior knowledge is rich.



"The acceptance region is defined as the interval of allowable shift with OOS probability less than 0.05. For relative potency, the acceptance criteria *(*− 19.9*,* 8.7*)*. If the mean potency change is −19.9, the posterior probability of failing the specifications of 75–125% is 5% for the new process."

Reference: Yu, B., Zeng, L., Yang, H., 2017, A Risk-Based Approach to Setting Acceptance Criteria for Pharmaceutical Process Comparability, J Pharm Innov (2017) 12:357–366, https://doi.org/10.1007/s12247-017-9295-8

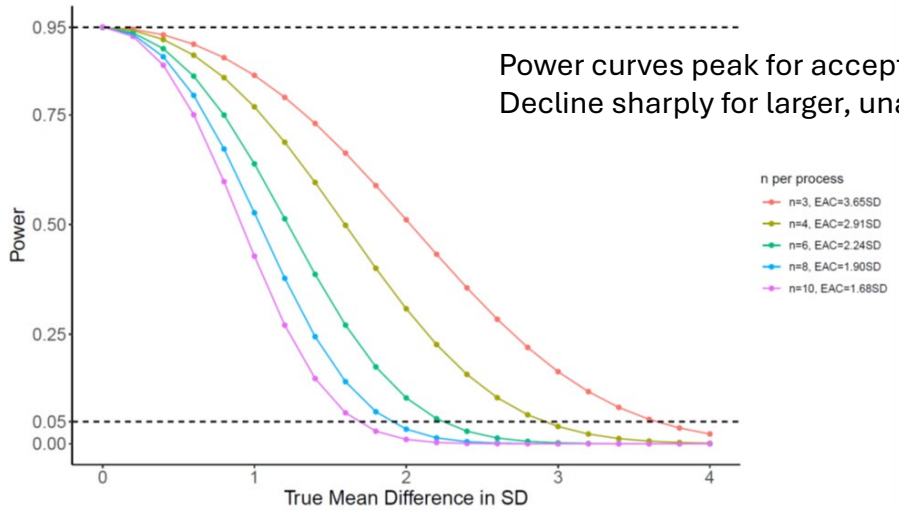
Statistical
EAC Determination* for TOST Procedure (3)

***Evaluation**

Based on given sample size

- Used when the number of batches are driven by the supply needs
- EAC= $c\sigma$, σ denotes the standard deviation (SD) representing process and analytical variability. The constant, c, is selected to ensure a high likelihood (e.g., >95%) of correctly claiming equivalence with the available sample size if the true mean difference is zero or smaller than a scientifically justified value.

Sample size (# of batches per process)	c
3	3.65
4	2.91
5	2.51
6	2.24
7	2.05
8	1.90
9	1.78
10	1.68



Power curves peak for acceptable mean shifts
Decline sharply for larger, unacceptable shifts

Reference: Chapter 12, Development of Gene Therapies: Strategic, Scientific, Regulatory, and Access Considerations, edited by Avery McIntosh, Oleksandr Sverdlov 2024

Varieties of Equivalence Test

Parameter of interest	Commonly used EAC	Benefits
Mean difference	$c\sigma$	Commonly available in statistical packages
Effect size=mean difference to sigma ratio	1, 1.5, 2, 3	Eliminates need to replace true σ with sample estimate in EAC leading to better Type I error control.* The same EAC could apply to different attributes
Mean ratio	(0.8, 1.25)	Ratio is unitless. The same EAC could apply to different attributes. Bayesian approach has been presented by Bill Pikounis*

Both frequentist and Bayesian versions are available

Reference:

- Richard K. Burdick, Neal Thomas & Aili Cheng (2017) Statistical Considerations in Demonstrating CMC Analytical Similarity for a Biosimilar Product, Statistics in Biopharmaceutical Research, 9:3, 249-257, DOI: 10.1080/19466315.2017.1280412
- Pikounis, B., 2024, Comparability with Statistical Rigor in Manufacturing Development, Bayes 2024 Conference, Rockville Maryland

Quality Ranges via Statistical Intervals

$\bar{x} \pm k^*s$

the sample average of the N pre-change lots

sample standard deviation of the pre-change lots

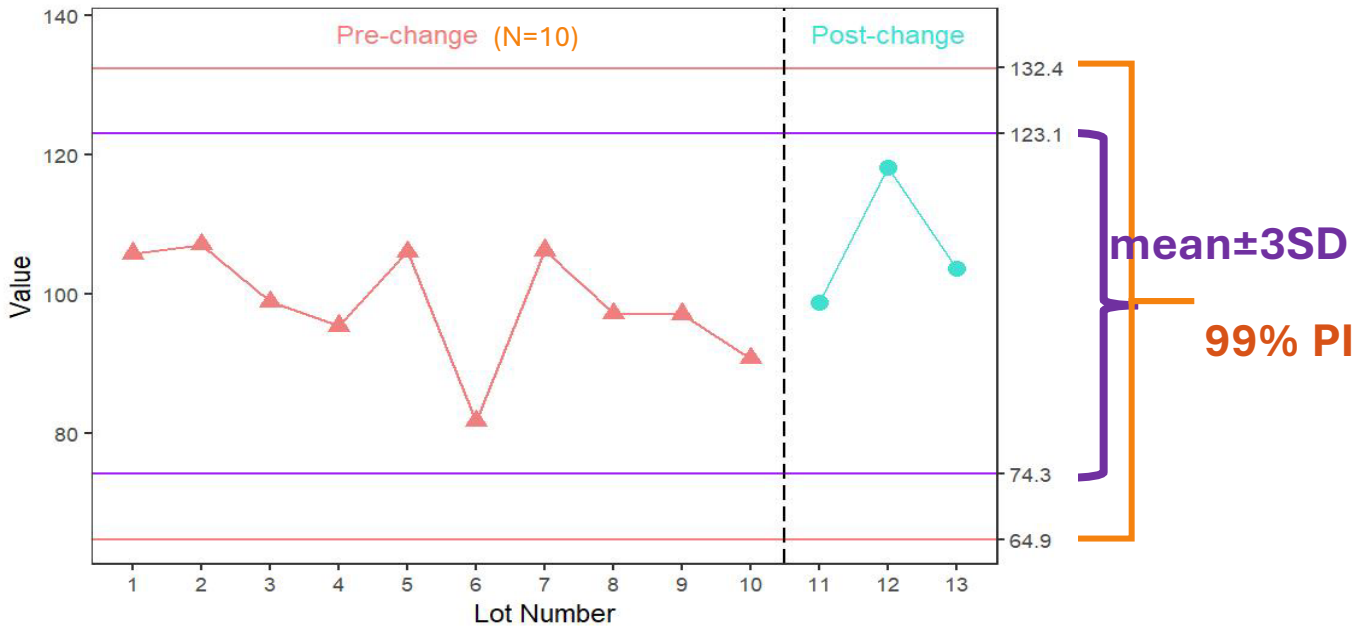
K=2 or 3 are commonly used for quality range for biosimilars. However, we would recommend simultaneous prediction intervals when there is reasonable number of pre-change batches.

$$k = t_{1-\frac{\alpha}{2m},(N-1)} * \sqrt{1+\frac{1}{N}} \text{ for two-sided;}$$

$$k = t_{1-\frac{\alpha}{m},(N-1)} * \sqrt{1+\frac{1}{N}} \text{ for one-sided.}$$

Ref: Hahn, Meeker, 1991, p62-63

Small Dataset Example

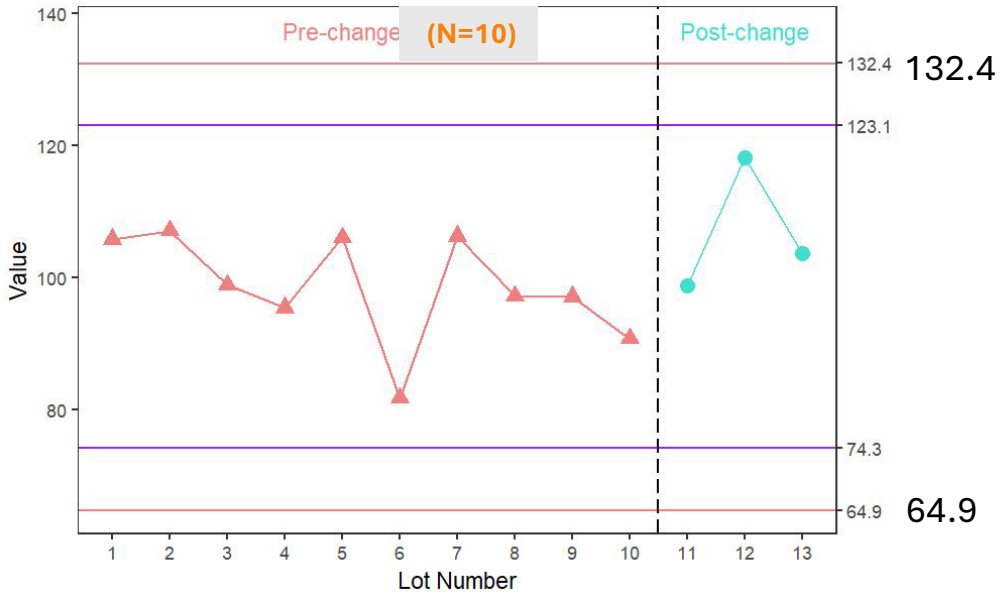


- For small pre-change data set, reference interval mean±3SD can be narrower than 99% PI
- Usually works well when the number of pre-change lots is much larger than number of post-change lots

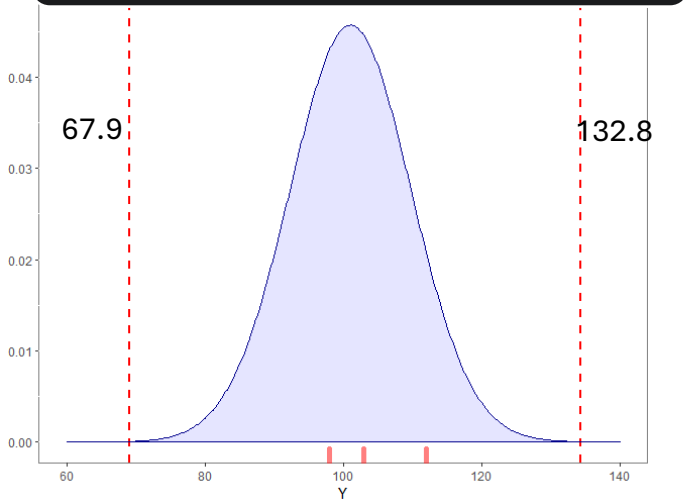
Data was generated by simulation

Frequentist vs. Bayesian for Small Dataset

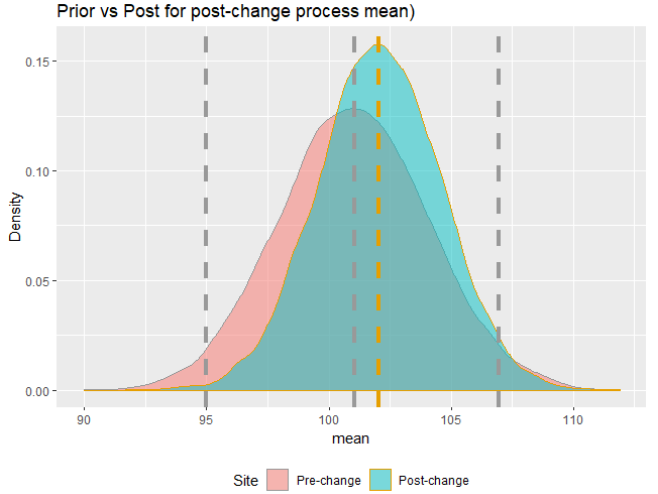
Frequentist: 99% PI



Bayesian: 99% HDI



Bayesian: Posterior distribution for the mean



- Both approaches result in the same conclusion
- Bayesian approach makes it easier to monitor the post-change posterior mean shifts

Reference:

- Cheng, A. and Wang, K., 2019 Midwest Biopharmaceutical Statistics Workshop, Carmel, IN
- Case Studies in Bayesian Methods for Biopharmaceutical CMC, Chapter 11, 2022
- Development of Gene Therapies Strategic, Scientific, Regulatory, and Access Considerations, Chapter 12, 2024

Challenges in Applying Interval Approach

Selection of Appropriate Value for k

- Various intervals serve distinct purposes.
- The FDA (2019) did **not** endorse tolerance intervals for similarity acceptance criteria.
- Significant sample size needed for meaningful intervals.

Concerns of using reference interval Mean±3SD

- A special case of tolerance interval with coverage that depends on sample size.
  - For 95% confidence, the coverage is poor for n < 15.
- Scientists often use it without clear understanding of confidence and coverage levels.

Recommendation Should be Based on Available Sample Size, but for small sample sizes interval will likely be too wide.

Reference: FDA draft guidance, 2019, Development of Therapeutic Protein Biosimilars: Comparative Analytical Assessment and Other Quality-Related Considerations. Line 812-815

Comparison between Frequentist and Bayesian

Method	Pros	Cons
Frequentist Approach	Widely used, straightforward	Equivalence test and statistical intervals assume data can be approximated by a normal distribution, less flexible
Bayesian Approach	Handles different distributions, flexible	Complex, requires prior knowledge; specialized software.

Factors to Consider for Method Choices

Choice of Method Depends on Various Factors

- Development stage of the process.
- Type of process change.
- Prior knowledge and data available.
- Goal of the assessment.
- Data distribution.

Assumption of Normal Data

- Equivalence test and statistical intervals assume data can be approximated by a normal distribution.
- If not, a transformation or different distribution is needed.
- Bayesian approach handles different distributions better.
- Frequentist approach has limitations with non-normal data.

Collaboration between scientists, engineers, and statisticians is key.