Sandoz
Biopharmaceuticals

Performance characteristics of quality range methods and equivalence testing in the comparative assessment of quality attributes

Thomas Stangler
RA CMC Teamlead Biosimilars
Novartis Global Drug Development

SANDOZ A Novartis Division

Agenda

1. Comparability and biosimilarity – from CMC guidances to statistics
2. Equivalence Criterion: Test population in reference population
3. Evaluating performance/operating characteristics against the equivalence criterion

Please note:

- *This presentation assumes data meeting all statistical assumptions*
  - *Case studies illustrating limitations due to real-life data were presented before*
- *Both manufacturing change comparability and biosimilarity are in scope of this presentation*
  - *differences only in sample sizes and level of prior knowledge*
- *Terminology:*
  - *Reference product: pre-change / reference biologic*
  - *Test product: post-change / biosimilar*

Comparability and biosimilarity

Comparability (ICH Q5E)

- Pre- and post-change product not necessarily identical, but **highly similar**
- Existing knowledge is sufficiently predictive to ensure that any **differences** have **no adverse impact upon safety or efficacy**

Biosimilarity (EMA/FDA)

- **Highly similar** quality profile, demonstrated by extensive comparability exercise[1]
- Any **differences** will have to be appropriately **justified** with regard to their **potential impact on safety and efficacy**[1]
- The biologic product is **highly similar** to the reference product notwithstanding **minor differences in clinically inactive components**[2]
- There are **no clinically meaningful differences** between the biologic product and the reference product in terms of safety, purity, and potency of the product[2]

1. EMA Guideline on similar biological medicinal products containing biotechnology-derived proteins as active substance : quality issues (revision 1)
2. Section 7002(b)(3) of the Affordable Care Act, adding section 351(i)(2) of the PHS Act;

Is ”Highly Similar” equivalent to ”Equivalent”?

comparable / biosimilar

↓

highly similar

↓ ?

equivalent

↓ ?

statistically equivalent

↓ ?

statistically equivalent for the means

Highly similar allows for differences if justified with respect to safety and efficacy

Merriam-Webster Dictionary
(Merriam-Webster.com, Apr 11th, 2017)
equivalent: one that is equal to another in status, achievement, or value
Equivalency: the state or fact of being exactly the same in number, amount, status, or quality

“Equivalent“ is stricter than “highly similar“

Using statistics – key considerations:

1. Relevant characteristic for comparison
2. Appropriate choice of statistical approach
3. Test parameters incl. equivalence margin / acceptance range
   - **Reference product (RP) based approach**
     - reference product defines acceptable quality
     - can be defined statistically
   - **Any other approaches feasible? No, not really**

Scientific considerations for comparability incl. biosimilarity

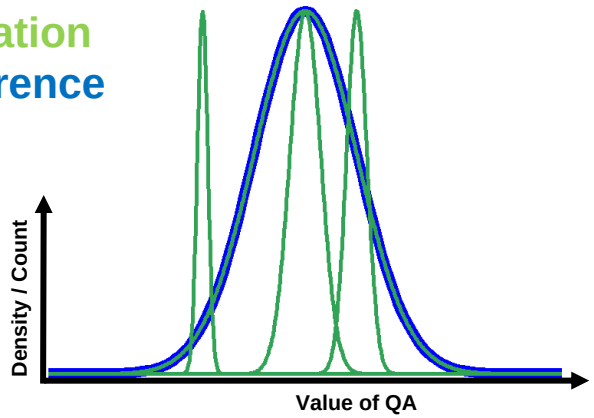
- Safety and efficacy within the reference product's variability have been demonstrated in clinical studies and by real-life experience with the reference product
- Every marketed batch from the reference product defines acceptable quality with respect to its quality characteristics
- A given quality characteristic of a reference product lot is acceptable for a test lot (e.g. biosimilar/post-change)

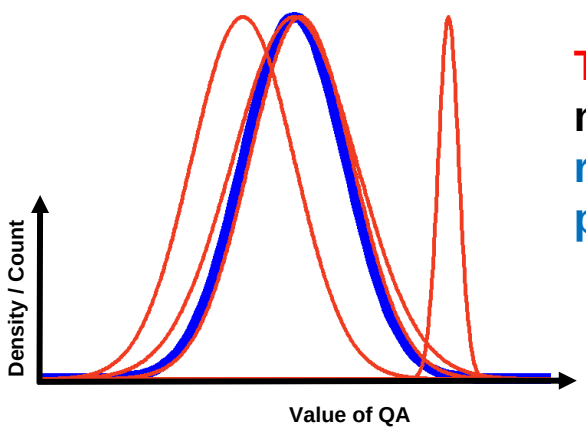
Mark McCamish & Gillian Woollett (2011) Worldwide experience with biosimilar development, mAbs, 3:2, 209-217, DOI: 10.4161/mabs.3.2.15005

Translating scientific considerations into a statistical criterion

- If the population of the test product is within the population of the reference product, all test lots are equivalent to reference lots on a batch level

Test population within reference population





Test population not within reference population

- "[...] ensuring that values of the attribute being tested for the proposed biosimilar tend to fall within the reference product distribution [...]"
  One of the three criteria for the suggested form of the equivalence margin in the FDA draft guidance "Statistical Approaches to Evaluate Analytical Similarity"
- 3 standard deviations is a good estimator of the actual population width
  "three-sigma rule of thumb", Cpk/PpK=1, Statistical Process Control (Nelson rule #1), FDA's tier 2 QAs

→ **3 sigma of the test population in 3 sigma of the reference population**

SANDOZ A Novartis Division

Considered statistical approaches for the comparative assessment

Quality ranges / intervals	Assumptions	Statistical complexity	Considered implementation
Min-Max range	none	low	as is
x-sigma	normality (iid* data)**	moderate	3σ (coverage: 99.7%)
Tolerance intervals	normality (iid* data)**	moderate - high	coverage: 99% confidence: 90%
Inferential statistical methods	**allowing for a statistical quantification of uncertainty**		
Equivalence Test (for means)	normality **iid* data**	high	margin: $-1.5\sigma_R, 1.5\sigma_R$ confidence: 90%

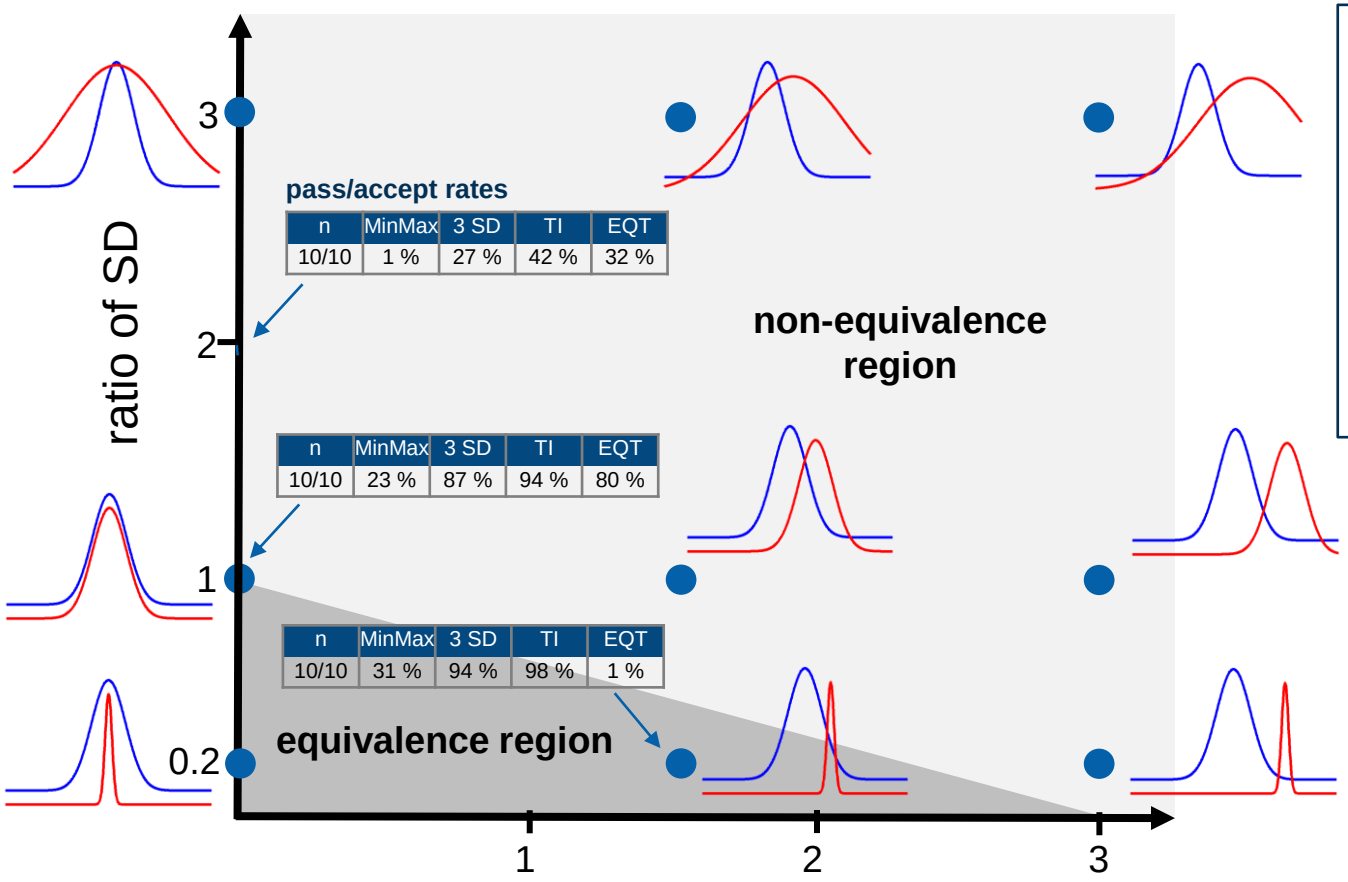
- NB: Major limitations for test interpretation may result from real-life CMC data not meeting the statistical assumptions

* independent and identically distributed data: no shifts, trends, outliers
** only necessary to draw inferential-like conclusion as drawn later in this presentation

Operating characteristics: Quantification of uncertainty

- From a pure statistical point of view
  - inferential statistics can quantify uncertainty
    - e.g. false positive rate alpha restricted to 5%, power for a give sample size & deviation from H_0
  - uncertainty cannot be quantified for range methods
    - TI‘s confidence is not an uncertainty estimation for the testing procedure
- From a combined scientific & statistical point of view
  - it‘s possible quantify the uncertainty based on a clear scientific hypothesis about acceptable quality (= equivalence criterion)
  - works for inferential methods and range methods
  - can identify false accepts (false positives) and false rejects (false negatives)

Comparing two normal populations: Test vs reference



Pass/accept rate

= n_{pass}/n_{sim}; n_{sim}=1000

Equivalence Test:

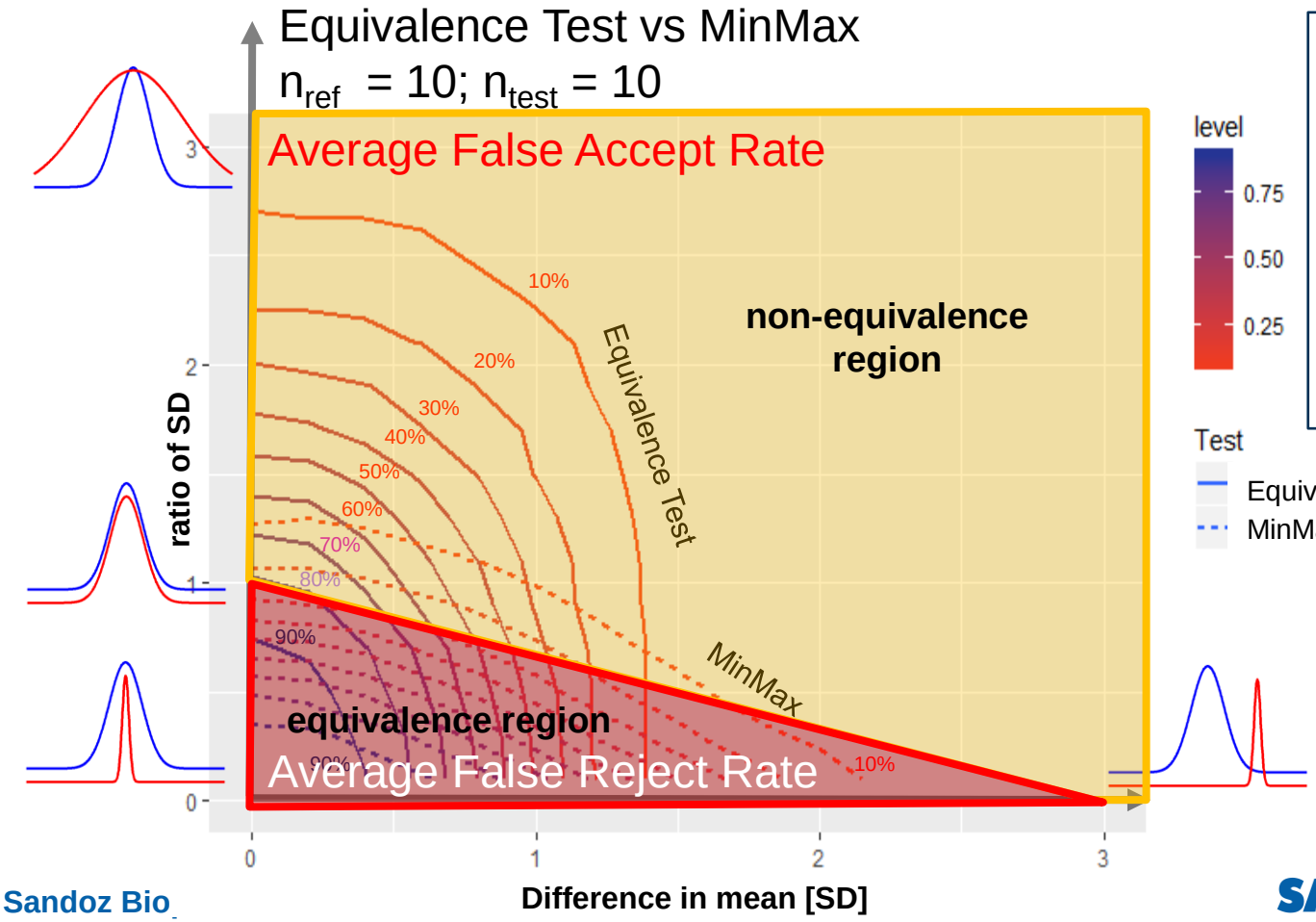
- type I error prob. 5%
- equiv. margin 1.5 σ_{ref} (sample)

MinMax/3SD/TI

- complete test sample within MinMax/3SD/TI based on reference sample

Comparing two normal populations: Test vs reference

Contour plot of test's pass/accept rates



Pass/accept rate
= n_{pass}/n_{sim}; n_{sim}=1000

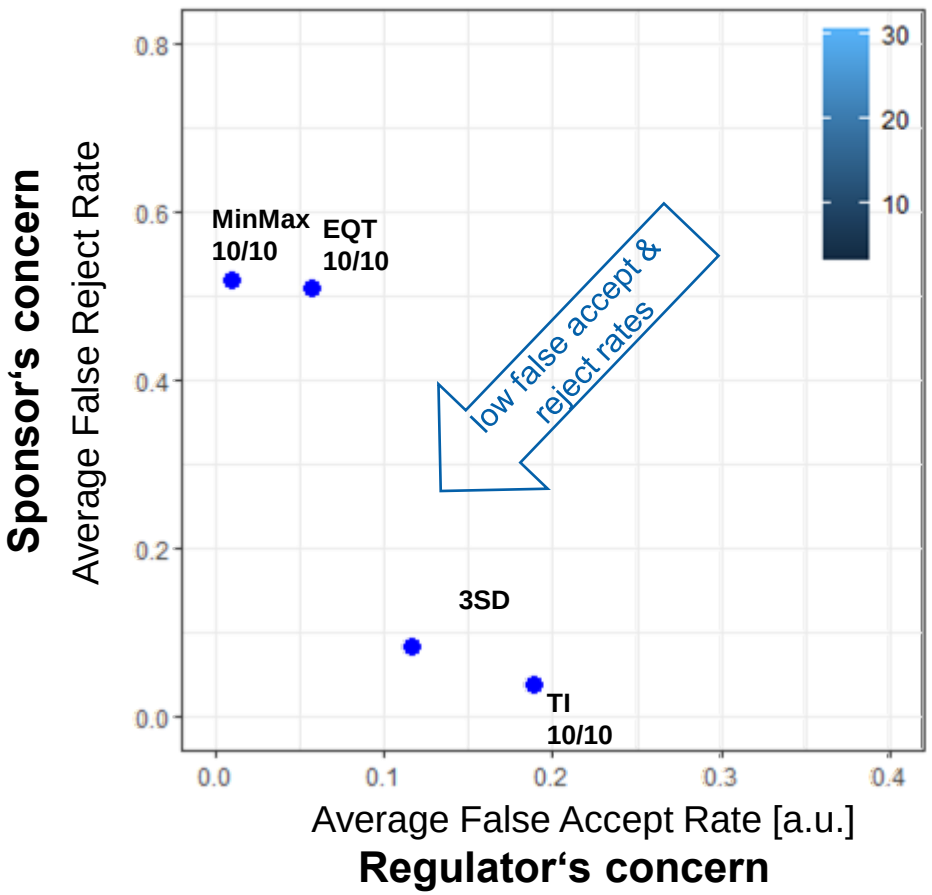
Equivalence Test:
- type I error prob. 5%
- equiv. margin 1.5 σ_{ref} (sample)

MinMax/3SD/TI
- complete test sample within MinMax/3SD/TI based on reference sample

SANDOZ A Novartis Division

Evaluating different test´s operating characteristics

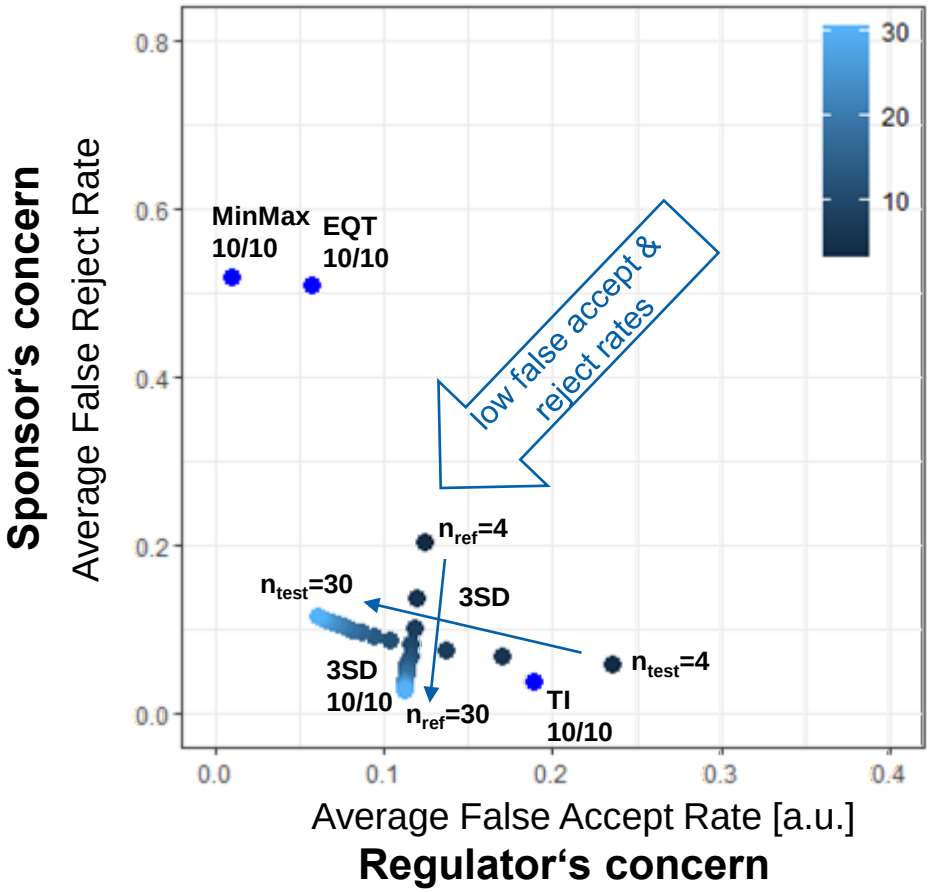
Average false accept rates & average false reject rates



- Compare tests e.g. for given sample sizes (n_{ref} & n_{test})
- Most desirable: low false rejects and low false accepts
- Evaluate the impact of sample size (n_{ref} & n_{test})
  - Examples:
    - n_{test} 4,6,8,..,30 for n_{ref}=10
    - n_{ref} 4,6,8,...,30 for n_{test}=10

Evaluating different test´s operating characteristics

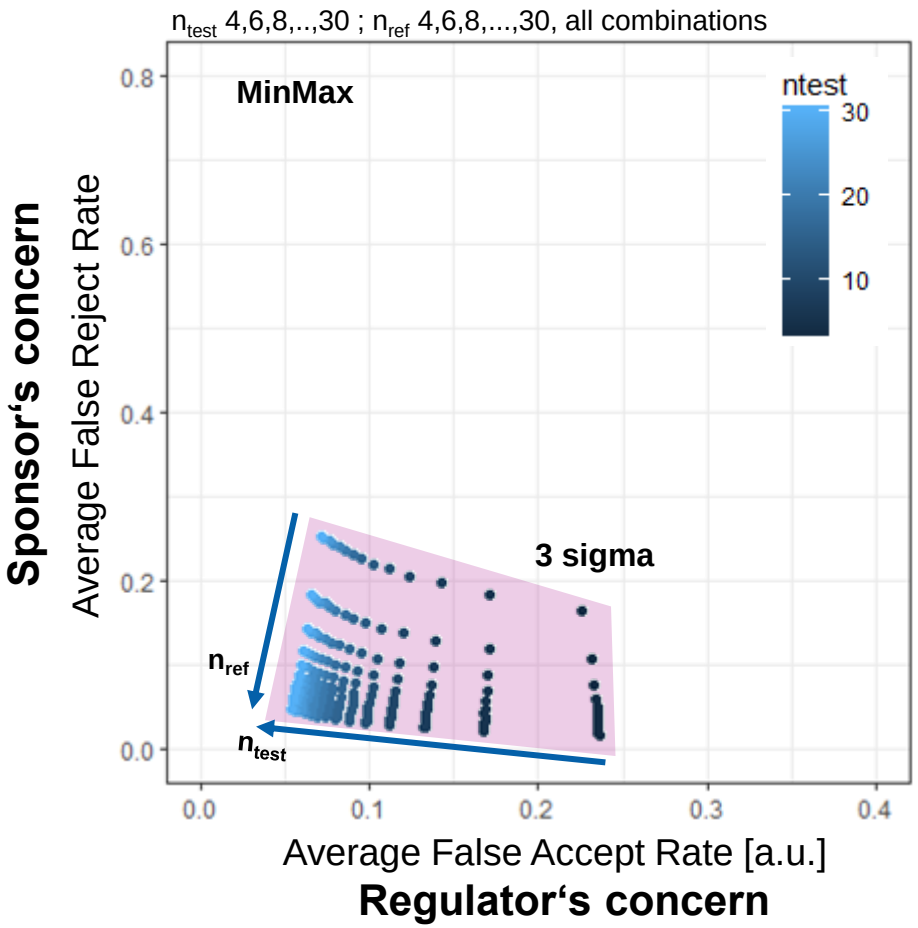
Average false accept rates & average false reject rates



- Compare tests e.g. for given sample sizes (n_{ref} & n_{test})
- Most desirable: low false rejects and low false accepts
- Evaluate the impact of sample size (n_{ref} & n_{test})
  - Examples:
    - n_{test} 4,6,8,..,30 for n_{ref}=10
    - n_{ref} 4,6,8,...,30 for n_{test}=10

Evaluating different test´s operating characteristics

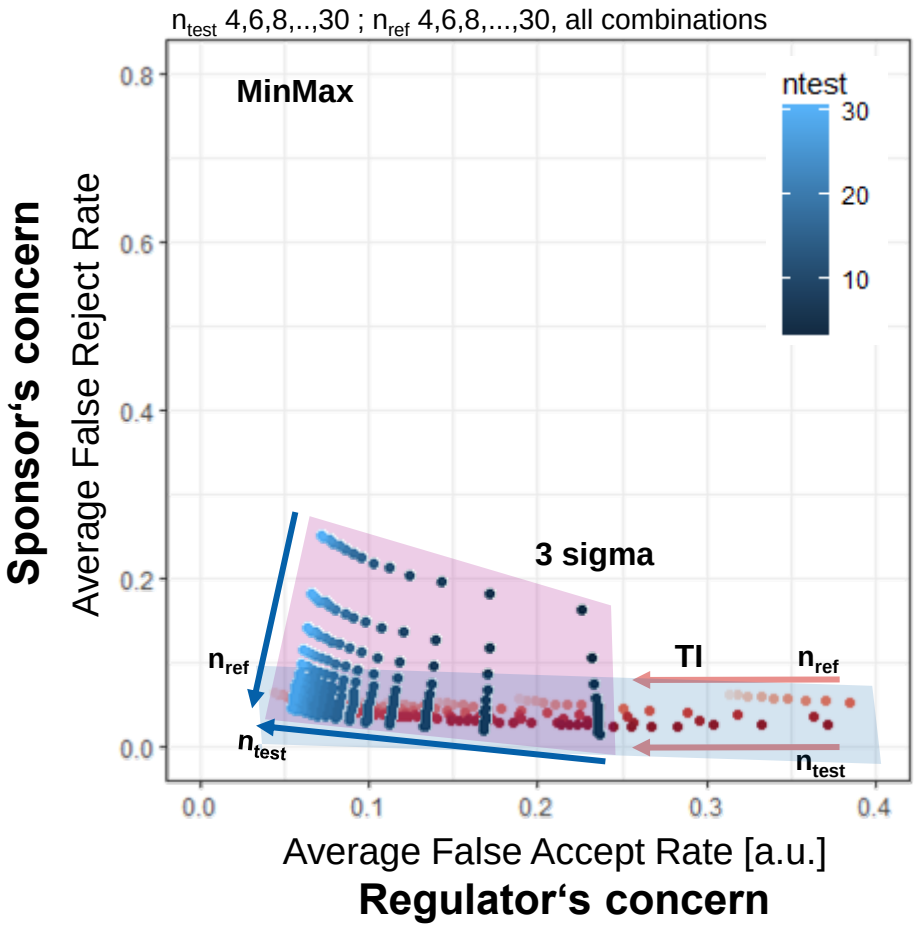
Average false accept rates & average false reject rates



- 3 sigma
  - relatively low av. false reject rates
  - increasing sample sizes decrease error rates
- Tolerance Intervals (TI)
  - low samples (test & ref) increase only av. false accept rates (but not av. false reject rates)
- MinMax
  - lowest average (av.) false accept rates but high av. false reject rates
- Equivalence Test (EQT)
  - high av. false reject rates
  - av. false accept rates increase with sample size
- Significant av. false reject rates for all approaches (& aggravated by multiplicity)
- For samples n ≥ 10, all quality range methods exhibit av. false accept rates not higher that those seen for the EQT

Evaluating different test´s operating characteristics

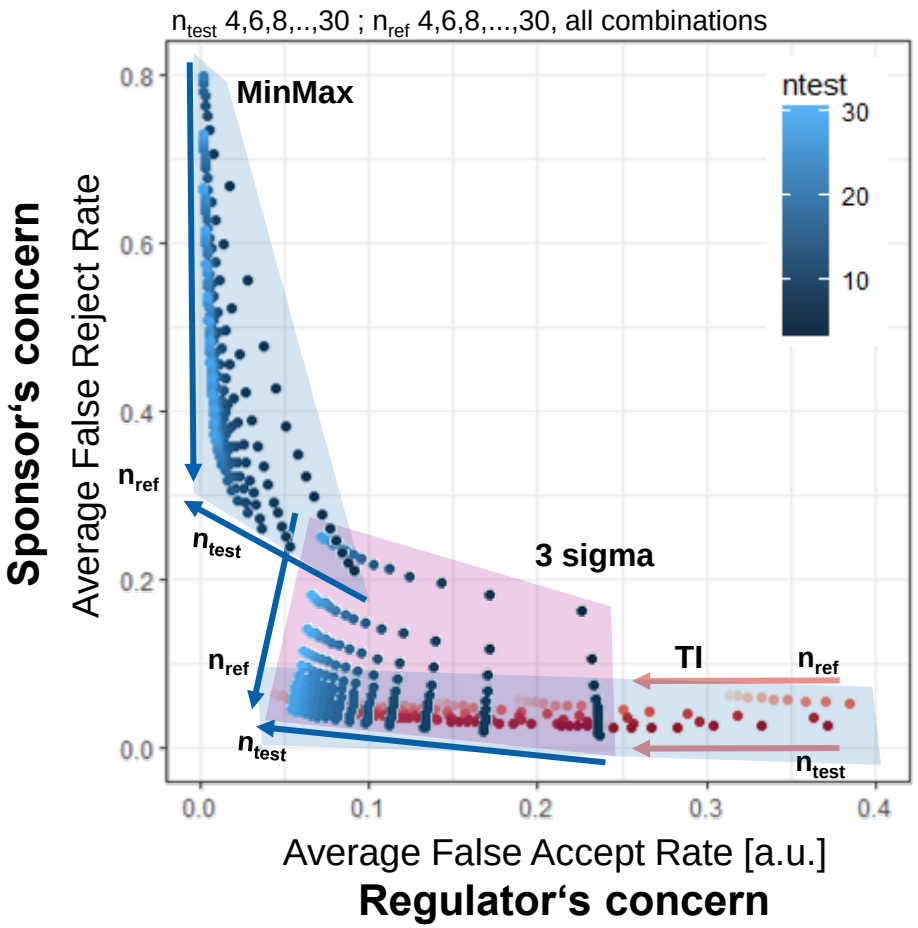
Average false accept rates & average false reject rates



- 3 sigma
  - relatively low av. false reject rates
  - increasing sample sizes decrease error rates
- Tolerance Intervals (TI)
  - low samples (test & ref) increase only av. false accept rates (but not av. false reject rates)
- MinMax
  - lowest average (av.) false accept rates but high av. false reject rates
- Equivalence Test (EQT)
  - high av. false reject rates
  - av. false accept rates increase with sample size
- Significant av. false reject rates for all approaches (& aggravated by multiplicity)
- For samples n ≥ 10, all quality range methods exhibit av. false accept rates not higher that those seen for the EQT

Evaluating different test´s operating characteristics

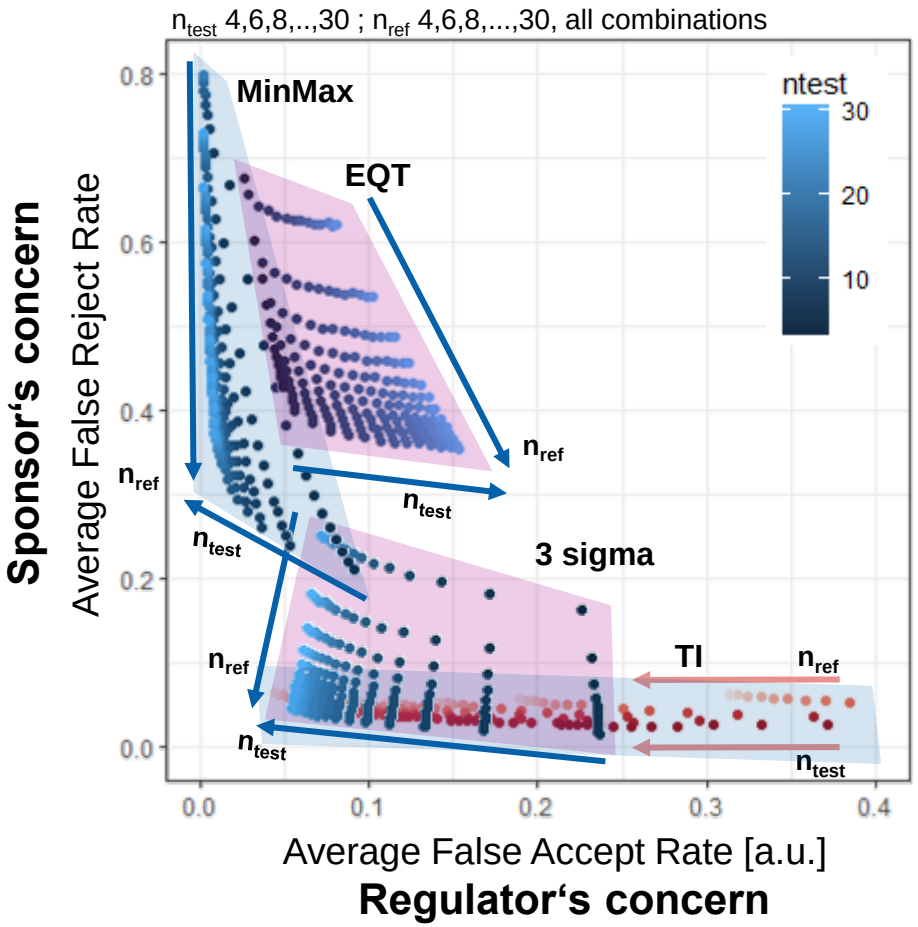
Average false accept rates & average false reject rates



- 3 sigma
  - relatively low av. false reject rates
  - increasing sample sizes decrease error rates
- Tolerance Intervals (TI)
  - low samples (test & ref) increase only av. false accept rates (but not av. false reject rates)
- MinMax
  - lowest average (av.) false accept rates but high av. false reject rates
- Equivalence Test (EQT)
  - high av. false reject rates
  - av. false accept rates increase with sample size
- Significant av. false reject rates for all approaches (& aggravated by multiplicity)
- For samples n ≥ 10, all quality range methods exhibit av. false accept rates not higher that those seen for the EQT

Evaluating different test´s operating characteristics

Average false accept rates & average false reject rates

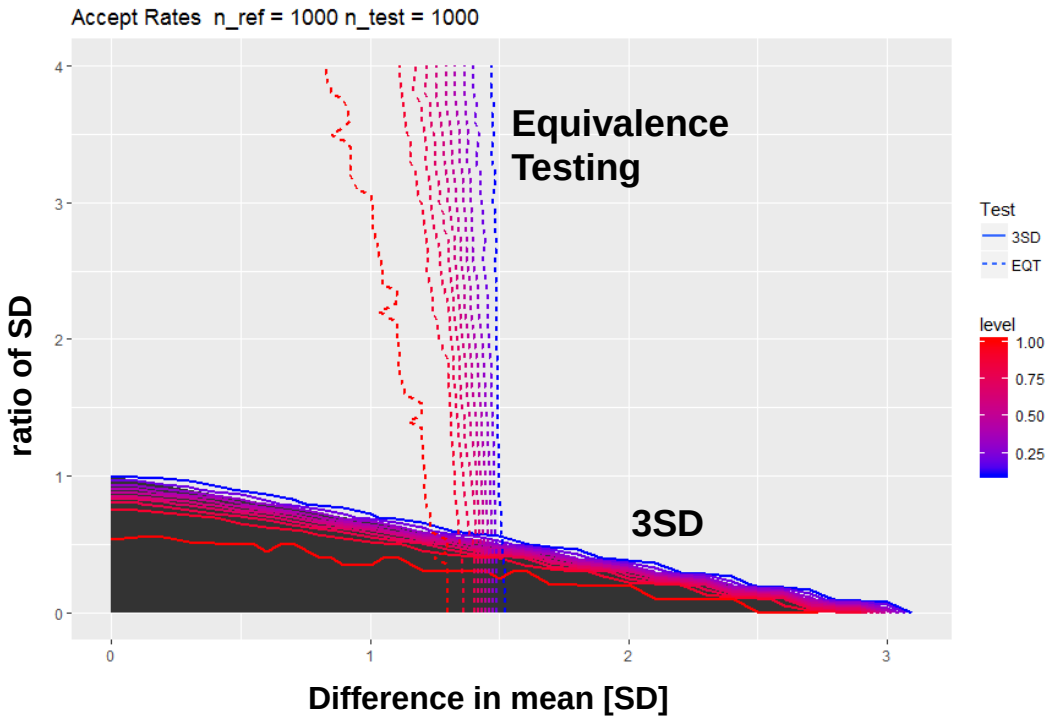


- 3 sigma
  - relatively low av. false reject rates
  - increasing sample sizes decrease error rates
- Tolerance Intervals (TI)
  - low samples (test & ref) increase only av. false accept rates (but not av. false reject rates)
- MinMax
  - lowest average (av.) false accept rates but high av. false reject rates
- Equivalence Test (EQT)
  - high av. false reject rates
  - av. false accept rates increase with sample size
- Significant av. false reject rates for all approaches (& aggravated by multiplicity)
- For samples n ≥ 10, all quality range methods exhibit av. false accept rates not higher that those seen for the EQT

Different scientific hypotheses for quality ranges vs equivalence testing

illustrated by large test and reference sample sizes

- The average false accept rate of the equivalence test increases with sample size
- Equivalence testing is the wrong tool to control a population in a population
  - EQT controls the mean to be within the equivalence margin
  - EQT does not control the variance (ratio of SD)
    - variance is a minor matter for equivalence testing for the mean
    - done decreasingly well for larger sample sizes



Multiplicity implications for overall average success rates

Testing more than one quality attribute: Overall success rates for truly equivalent products

$n_{ref} = 10$, $n_{test} = 10$

# of QA	Min Max	3SD	TI	EQT
1	48.0%	92.0%	96.0%	49.0%
3	11.1%	77.9%	88.5%	11.8%
10	0.1%	43.4%	66.5%	0.1%
20	0.0%	18.9%	44.2%	0.0%

$n_{ref} = 30$, $n_{test} = 10$

# of QA	Min Max	3SD	TI	EQT
1	71.8%	97.1%	95.8%	62.2%
3	37.0%	91.4%	87.9%	24.1%
10	3.6%	74.1%	65.1%	0.9%
20	0.1%	54.9%	42.4%	0.0%

Success rates < 50% colored red for illustration purposes only. 50 % should not be considered a reasonable success rate.

- Significant multiplicity issues due to high statistical uncertainty
  - MinMax and EQT have already for a single QA very low average success rates
- From the evaluated approaches, 3 sigma is certainly not perfect but the test of choice for any larger number of quality attributes
- In any case, false alarms are very likely and should not be overrated

Statistical conclusions

- Low sample sizes in comparability / biosimilar settings create considerable uncertainty (aggravated by multiplicity)
- Increasing sample size can have surprising and undesirable consequences
  - e.g. increase in false accept rate with test sample size for equivalence testing
- Test performance depends on scientific hypothesis
  - range methods better suited than EQT to test for „population in population“
- Typically trade-off between false accepts and false reject
  - exception EQT – which is just worse since not aligned with scientific hypothesis
- Sample sizes are of key importance
  - Scientific expectation: larger sample sizes should primarily improve the conclusion
  - for Biosimilars, consider to include representative small scale studies, where possible, to have more lots (e.g. at least 10)

Conclusions

- The presented framework allows to evaluate operating characteristics of statistical approaches
  - against a clear scientific hypothesis of equivalency (population in population)
  - other test proposals can be easily evaluated
  - equally applicable for manufacturing change comparability and biosimilarity
- Any remaining benefit from inferential vs non-inferential methods?
  - with a clear scientific hypothesis, uncertainty can be equally well estimated for non-inferential and inferential methods
- Statistics cannot be a pass/fail criterion due to
  - very limited sample size which leads to a high uncertainty
  - „Comparability" (highly similar) is less strict than statistical equivalence
  - the fulfillment of the assumptions for statistical inference is unclear
- How to find the right balance between false accept and false reject error rates?
  - depends on risk profile (e.g. QA risk in tiered approach, prior knowledge in context of a manufacturing change); multiplicity (testing of more than one quality attribute)
- Unless a complex test has clear benefits – go for simplicity (KISS*: keep it simple, stupid)