Statistical tests, Bayesian analysis, or heuristic rules for demonstration of analytical biosimilarity?

Workshop on draft reflection paper on statistical methodology for the comparative assessment of quality attributes in drug development

Richard K. Burdick, Ph D
Elion Labs, a division of KBI Biopharma, Inc.

May 3, 2018

Goal of Talk

- Provide a structure for discussion and comparison of various statistical similarity and comparability approaches.
- Demonstrate the structure using four proposed comparability approaches.
- Presentation is joint work of the AAPS Biosimilar Interest Group.

Definitions

- Heuristic rule: A commonsense rule used for solving a problem.
- Statistical test: A rule used to solve a problem with definable probabilities for incorrect decisions.
- Reference product (R): Originator reference medicinal product in a test for analytical similarity or pre-process change in a comparability study.
- Test product (T): Biosimilar product candidate in a test for analytical similarity or post-process change in a comparability study.
- Objective is to compare R and T in some definable manner.

Goals for Selecting a Statistical Method to Demonstrate Comparability/Analytical Biosimilarity

1. Protect patients from consequences of concluding comparability when products are not comparable.
2. Protect sponsors from consequences of concluding lack of comparability when products are in fact comparable (the consequences include a lack of patient access to lower cost treatments)
3. Incentivize sponsors to acquire process knowledge concerning T, and perhaps R in biosimilarity.
4. Enable decision making with practical sample sizes.

Goals for Selecting a Statistical Method to Demonstrate Comparability/Analytical Biosimilarity

5. Examine entirety of the process distribution of product.
6. Statistical rigor should consider criticality and measurement scale of the attribute.
7. Demonstrate robustness to violations of assumptions.
8. Be transparent, easy to explain, and easy to compute by scientists with no formal statistical training.

Example Using the Criteria

- Four statistical procedures--two statistical tests and two heuristic rules--- are now defined for testing comparability of R and T.
- Each procedure will be assessed against the proposed criteria.
- The R population is normal with mean of μ_R=100 (known) and standard deviation of σ_R=10 (known) with specifications of LSL=70 and USL=130.
- This yields a process capability based on the out-of-specification (OOS) rate of 0.0027=0.27%.

Example Using the Criteria

- The assumption of known μ_R and σ_R may be reasonable for many comparability studies with historical data sets, but analytical similarity studies have an extra level of complexity as they are unknown and must be estimated.
- Patient will be at risk if the probability of passing when T has a shift of at least $1.5\sigma_R$ from μ_R is 0.05 or greater. (FDA criterion of practical importance)
- This shift will yield an OOS rate of at least 0.0668=6.68% in T.

Populations of T

Design	MuT	SigmaT	NT	OOST	Comparison to R
1	115	5	10	0.0013	T better than R
2	109	7	10	0.0013	T better than R
3	100	10	10	0.0027	T same as R
4	115	10	10	0.0668	T equals patient risk
5	107.5	15	10	0.0730	T exceeds patient risk
6	100	20	10	0.1336	T exceeds patient risk

$\mu_R=100$ $\sigma_R=10$

Patients at risk if Designs 4-6 “Pass” and sponsor at risk if Designs 1-3 “Fail”

Proposed Methods

- Two statistical tests for demonstrating comparability
  - Statistical equivalence test of means using a CI on difference in means (Tier 1 FDA)

$$H_0: \left|\mu_T - \mu_R\right| \geq 1.5\sigma_R = 15$$

$$H_1: \left|\mu_T - \mu_R\right| < 15 \text{ (R and T are equiv)}$$

  - Statistical noninferiority of process capability using an upper bound on the OOS rate for T

$$H_0: \ \mathrm{OOS}_T \geq 0.0668$$

$$H_1: \mathrm{OOS}_T < 0.0668 \text{ (T is not inferior to R)}$$

Proposed Methods

- Two heuristic rules for demonstrating comparability
  - 90% two-sided prediction interval (PI) computed with T data must fall within a $2.5\sigma_R$ range around μ_R.
    - 100-25=75 to 100+25=125
    - EFSPI
  - All n_T=10 individual T values must fall in a $2.15\sigma_R$ range around μ_R.
    - FDA Quality range
  - Both of these rules are calibrated to provide the same protection to patients as the two statistical tests (0.05 probability of passing in Design 4.)

1. Protect patients from consequences of concluding comparability when products are not comparable.

- This goal requires an ability to ensure a small probability of demonstrating comparability when product differences are of practical importance.
- The two statistical tests (Equiv, OOS) control this probability by defining type 1 error to be 0.05 in Design 4.
- The two heuristic tests (PI, QR) require calibration for given sample sizes.

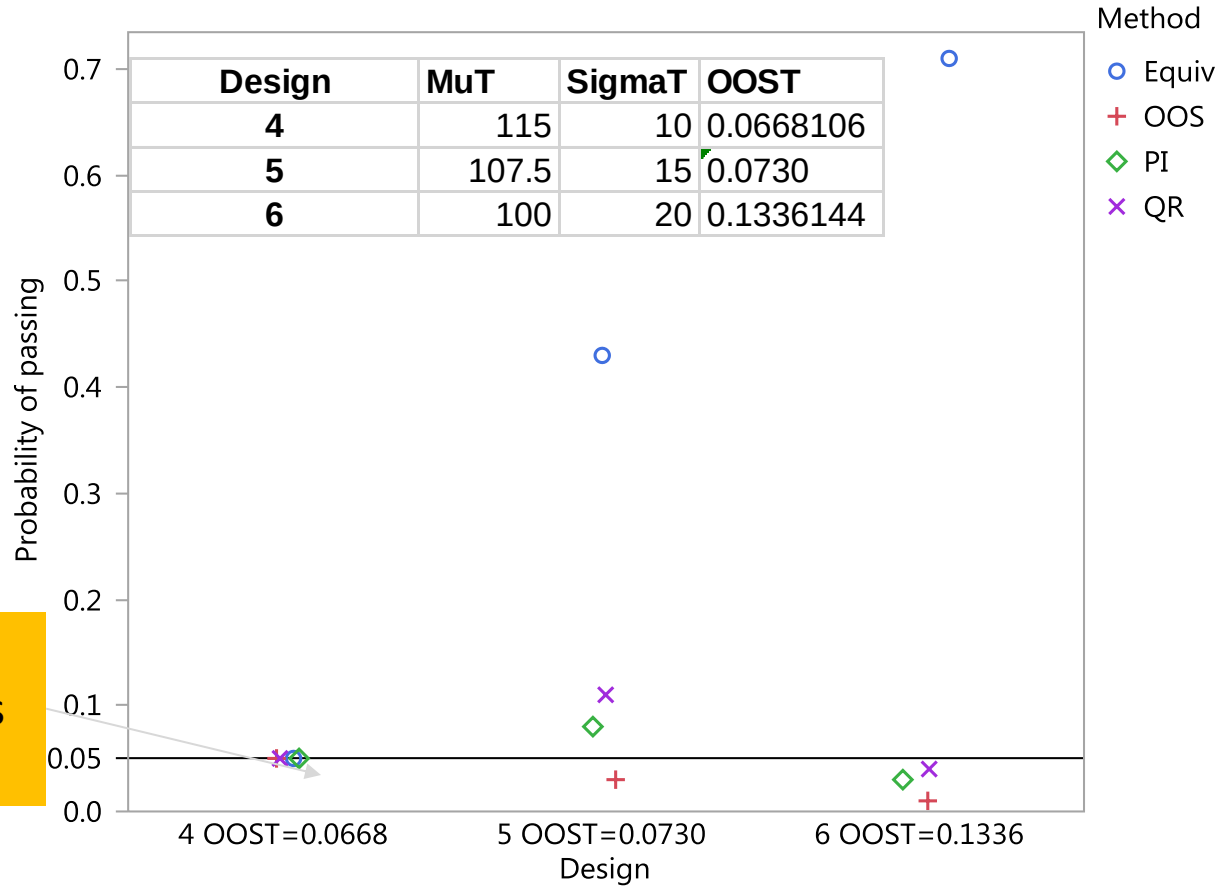
Populations of T

Design	MuT	SigmaT	NT	OOST	Comparison to R
1	115	5	10	0.0013	T better than R
2	109	7	10	0.0013	T better than R
3	100	10	10	0.0027	T same as R
4	115	10	10	0.0668	T equals patient risk
5	107.5	15	10	0.0730	T exceeds patient risk
6	100	20	10	0.1336	T exceeds patient risk

Probability of passing in Designs 4-6 should be less than or equal to 0.05 to satisfy Criterion 1.

Control of Patient Risk

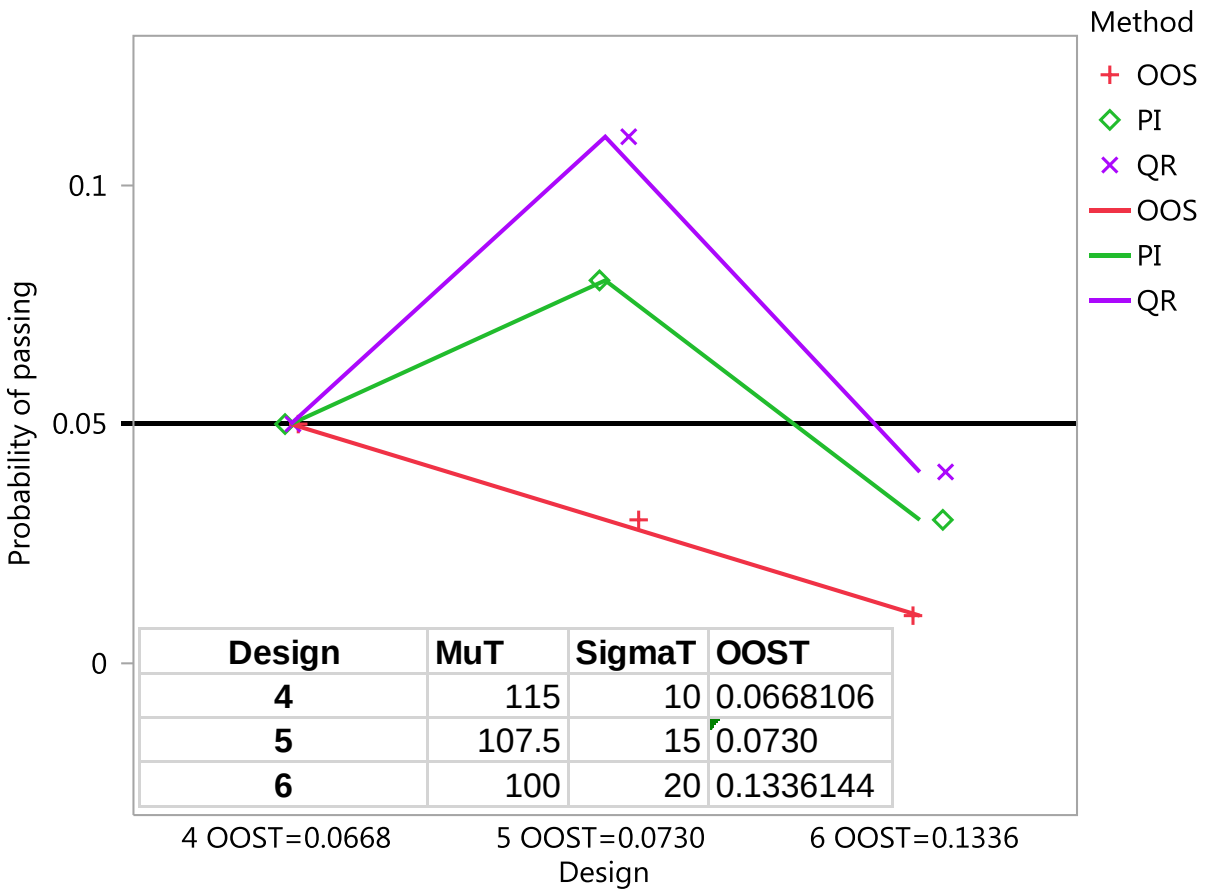
Design	MuT	SigmaT	OOST
4	115	10	0.0668106
5	107.5	15	0.0730
6	100	20	0.1336144



All methods calibrated at this point.

Equivalence test of means does not satisfy criterion 1.

Control of Patient Risk



Design	MuT	SigmaT	OOST
4	115	10	0.0668106
5	107.5	15	0.0730
6	100	20	0.1336144

Two heuristic rules also have increased risk above the desired 0.05 criterion in Design 5.

2. Protect sponsors from consequences of concluding lack of comparability when products are in fact comparable.

- This criterion requires an ability to ensure a large probability of demonstrating comparability when differences in products are of no practical importance.

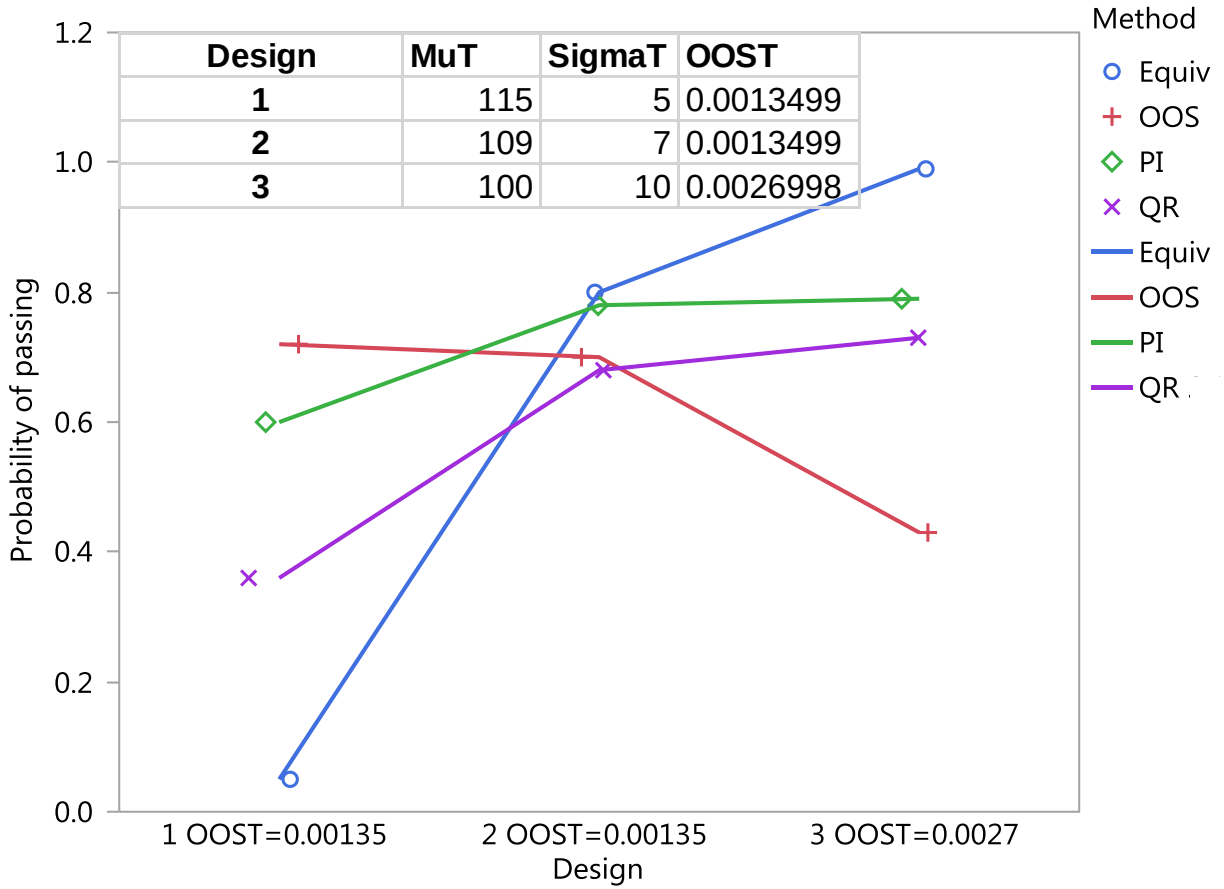
Populations of T

Design	MuT	SigmaT	NT	OOST	Comparison to R
1	115	5	10	0.0013	T better than R
2	109	7	10	0.0013	T better than R
3	100	10	10	0.0027	T same as R
4	115	10	10	0.0668	T equals patient risk
5	107.5	15	10	0.0730	T exceeds patient risk
6	100	20	10	0.1336	T exceeds patient risk

The greater the probability of passing in Designs 1-3, the better the procedure relative to Criterion 2.

Control of Sponsor Risk

Design	MuT	SigmaT	OOST
1	115	5	0.0013499
2	109	7	0.0013499
3	100	10	0.0026998



- Only OOS uniformly increases probability of passing as OOST decreases and satisfies Criterion 2.
- Large differences in all but OOS when T is most capable.

3. Incentivize sponsors to acquire process knowledge concerning T.

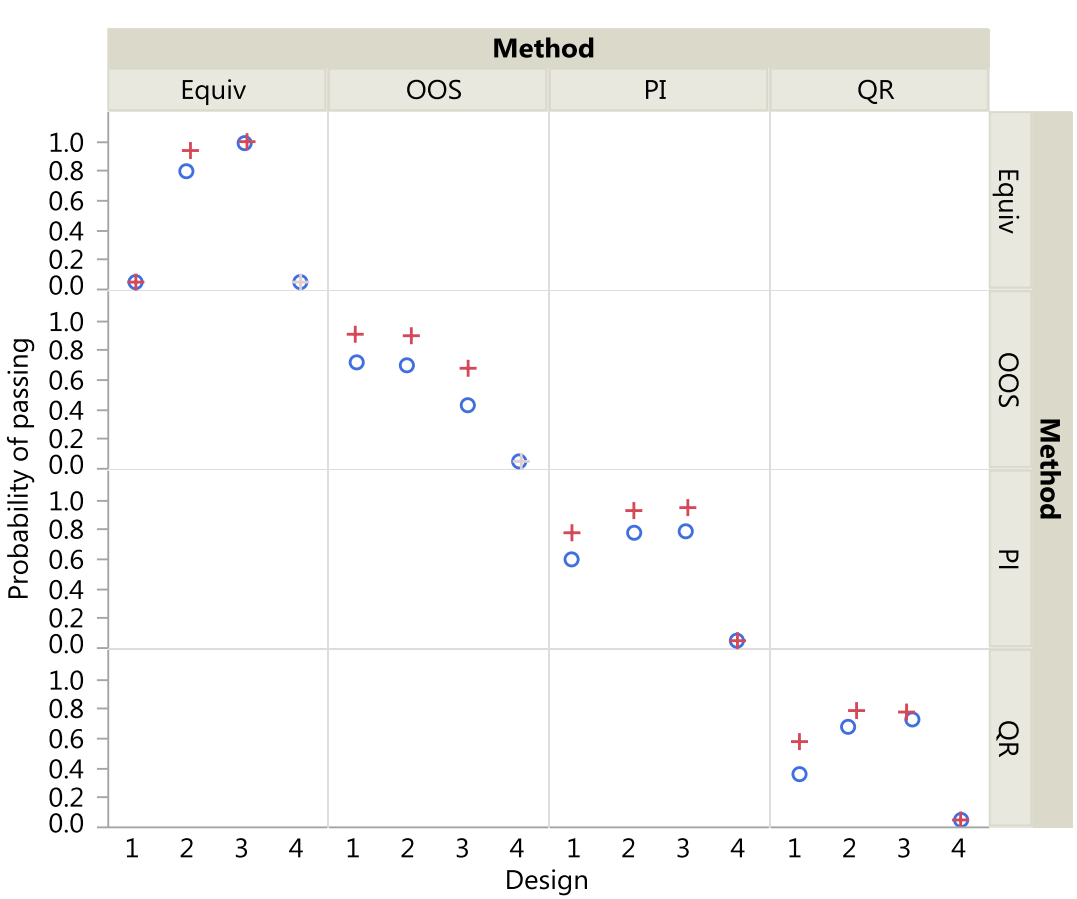
- Increase probability of passing for a given type 1 error and acceptance criterion by increasing sample sizes of T.
- To demonstrate, T sample size increased to 15.
- QR recalibrated from range of a $2.15\sigma_R$ around μ_R to a range of $2.4\sigma_R$ around μ_R to maintain 0.05 risk to patient.
- PI recalibrated from 90% to 88% to maintain 0.05 risk to patient.

Populations of T

Design	MuT	SigmaT	NT	OOST	Comparison to R
1	115	5	10	0.0013	T better than R
2	109	7	10	0.0013	T better than R
3	100	10	10	0.0027	T same as R
4	115	10	10	0.0668	T equals patient risk
5	107.5	15	10	0.0730	T exceeds patient risk
6	100	20	10	0.1336	T exceeds patient risk

To satisfy Criterion 3, probability of passing in Designs 1-3 should increase as n_T increases (with probability of passing Design 4 equal to 0.05).

Incentivize Sponsors



Design	MuT	SigmaT	OOST
1	115	5	0.0013499
2	109	7	0.0013499
3	100	10	0.0026998
4	115	10	0.0668106

All methods satisfy Criterion 3.

Summary of Demonstration for First Three Criteria

Criterion	Equiv	OOS	PI	QR
1-Patient	No	Yes	OK	OK
2-Sponsor	No	Yes	No	No
3-Incentivize	Yes	Yes	Yes	Yes

4. Enable decision making with practical sample sizes.

- Practicality of the manufacturing process and T sample sizes need to be considered.
- If power is too low for practical sample sizes, acceptance criterion must be loosened or type 1 error rate increased.
- Regulatory agencies could play a role with establishing these standards.

5. Examine entirety of the process distribution of product.

- Individual assessment of means or variances ignores their interrelationship in impacting process capability.
- A T process with a different mean than the R process may still produce acceptable product if it has lesser variance.
- Equivalence test of means does not meet this criterion.

6. Statistical rigor should consider criticality and measurement scale of the attribute.

- Can be controlled by defined type1 error rate and acceptance criterion.
- Scientific relevance of acceptance criterion (if possible) is always desired.
- It is important to consider the measurement scale (e.g., nominal, ordinal, continuous) and interrelationships of attributes to determine how conflicting results might affect the totality of evidence.

7. Demonstrate robustness to violations of assumptions.

- Normality of data has been assumed in many of the applications proposed to date.
- Properties of heuristic rules and statistical tests may be impacted by violation of assumptions.

8. Be transparent, easy to explain, and easy to compute by scientists with no formal statistical training.

- Spreadsheet solutions would be useful, but should not be limiting if procedures can be performed with user friendly statistical software.
- Statistical elegance may need to be sacrificed in order to provide a uniform streamlined assessment strategy.
- Meaningful visual displays aligned with the numerical conclusions should always be provided.

Conclusions

- Objective of talk is to provide a structure for comparing approaches.
- Criteria can be used for evaluation of both statistical tests and heuristic rules.
- Bayesian intervals and other procedures that incorporate both location and spread of the distributions should be considered.
  - e.g., distribution overlap as discussed in Inman and Bradley (1989) and proportion of similar response as discussed in Giacoletti and Heyse (2011) could be used to form a statistical test.

References

- Henry F. Inman & Edwin L. Bradley Jr (1989) “The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities”, *Communications in Statistics - Theory and Methods*, 18:10, 3851-3874.
- Katherine ED Giacoletti and Joseph Heyse (2011) “Using proportion of similar response to evaluate correlates of protection for vaccine efficacy”, *Statistical Methods in Medical Research*, DOI: 10.1177/0962280211416299, published online August 2011.