FDA U.S. FOOD & DRUG ADMINISTRATION

Statistical Issues with Aberrant IVRT/IVPT Data - FDA Perspective

Elena Rantou, PhD

Office of Biostatistics / OTS

Center for Drug Evaluation and Research, FDA

Disclaimer

1. This presentation reflects the views of the presenter and should not be construed to represent the United States Food and Drug Administration’s views or policies.
2. All data sets shown in this presentation have been de-identified

Outline

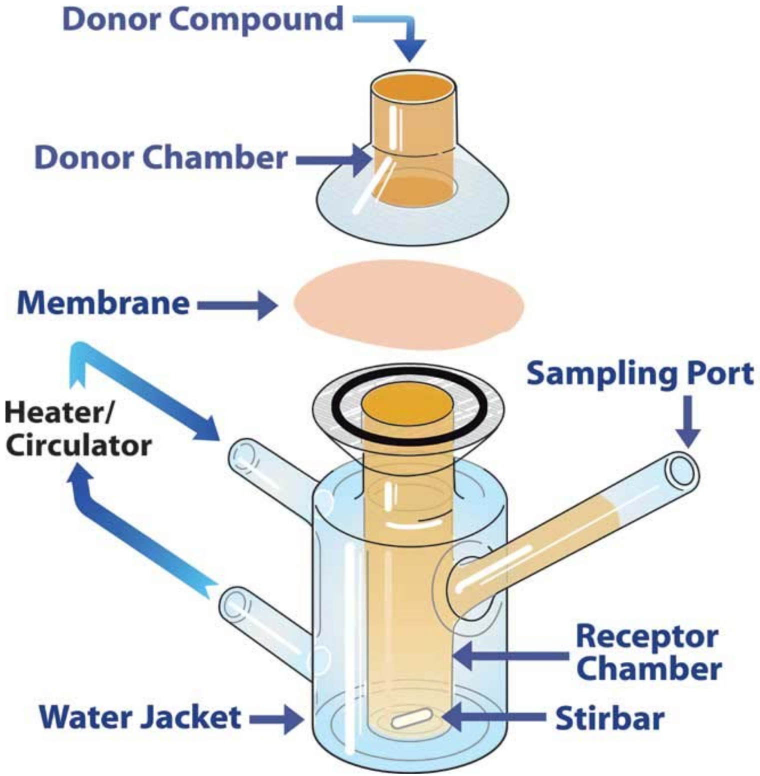
1. Overview of the IVPT and the mixed scaled criterion for assessing bioequivalence (BE)
2. Issues with IVPT
3. Issues with IVRT

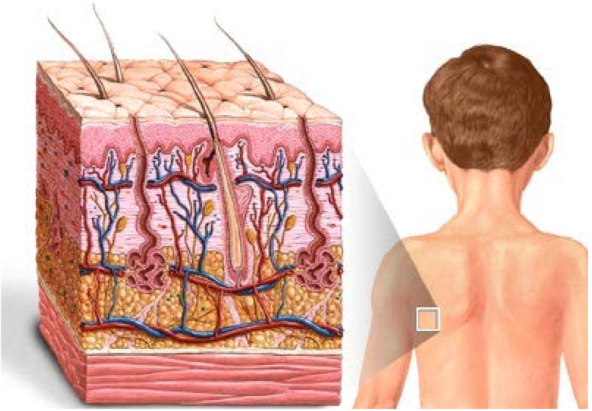
   Small sample sizes and the appropriateness of the SUPAC-SS approach

In Vitro Permeation Test (IVPT)

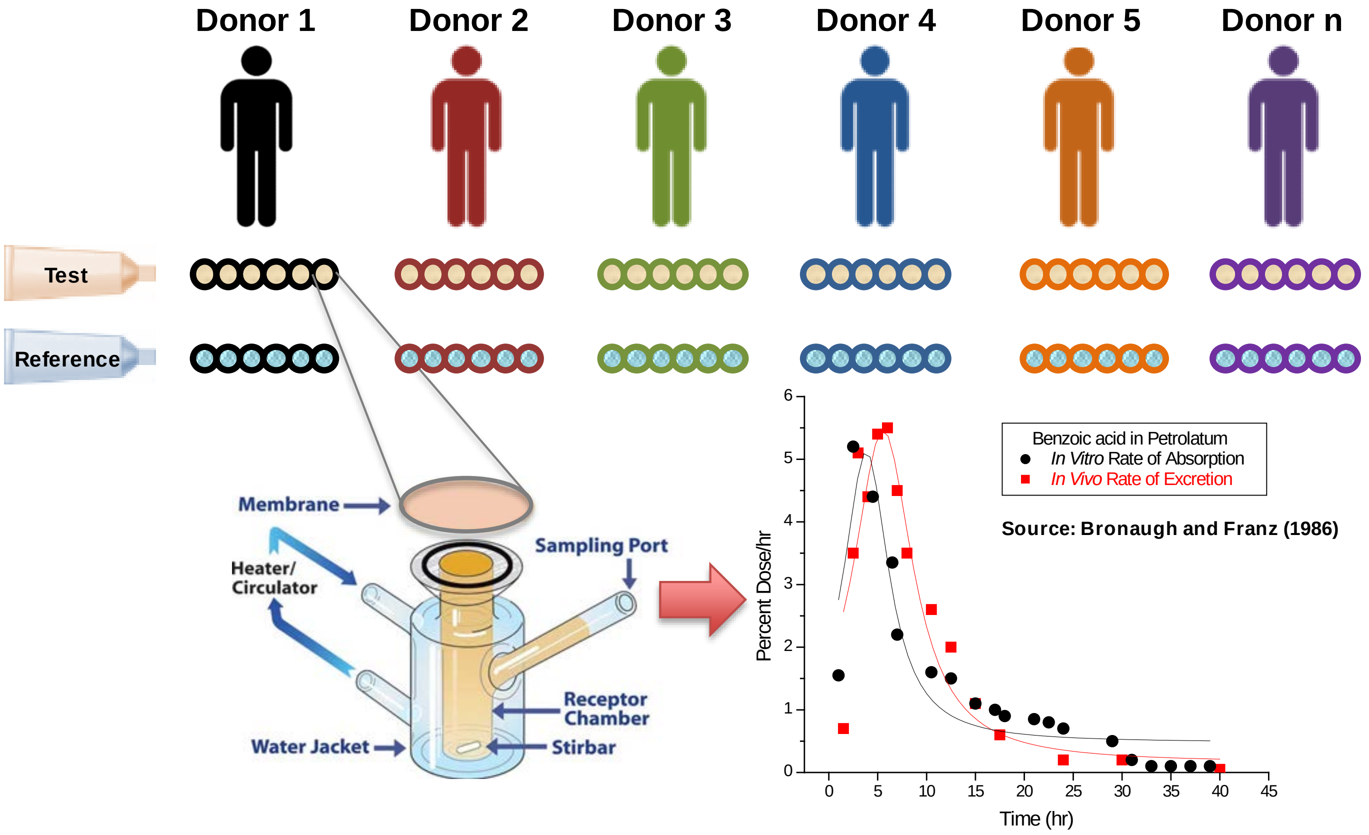
- Uses excised human skin
- Measures drug concentration
- The rate of drug delivery (flux) is measured by sampling at specific, pre-selected time-points in a way analogous to that used in blood (or plasma) concentration sampling in PK studies

In Vitro Permeation Test (IVPT)





IVPT Study Design



Source: Bronaugh and Franz (1986)

Study Design

The response considered is the log-transformed

- total penetration (AUC)
- max flux rate (Jmax)

We consider a sample of

n: donors (per treatment),

r: replicate skin sections from each one of the n donors are collected for each formulation (replicates from each donor are randomly assigned to each product)

2 treatment formulations: test (generic: T) and reference (R)

Study Design

Test:

$$T_{11}, T_{12}, \dots, T_{1r}$$
$$T_{21}, T_{22}, \dots, T_{2r}$$
$$\vdots$$
$$T_{n1}, T_{n2}, \dots, T_{nr}$$

Reference:

$$R_{11}, R_{12}, \dots, R_{1r}$$
$$R_{21}, R_{22}, \dots, R_{2r}$$
$$\vdots$$
$$R_{n1}, R_{n2}, \dots, R_{nr}$$

Statistical Analysis

For each donor, we can calculate the term $I_j = \frac{1}{r}\sum_{i=1}^{r}(T_{ij} - R_{ij})$ is recorded. This leads to the derivation of the point estimate:

$$\bar{I}_{.} = \frac{1}{n}\sum_{j=1}^{n} I_j$$

estimate of the inter-donor variability:

$$S_I^2 = \frac{1}{(n-1)}\sum_{j=1}^{n}(I_j - \bar{I}_{.})^2$$

Statistical Analysis

For two different replicates R_{ij}, of the same donor, j, we define the within-reference variability as:

$$S_{WR}^{2} = \frac{\sum_{j=1}^{n} \sum_{i=1}^{r} (R_{ij} - \overline{R_{.j}})^{2}}{(r-1)n}$$

Statistical Analysis

Under normality assumptions, the following distributional results hold:

$$\bar{I}_{\cdot} \sim N(\mu_T - \mu_R, \quad \frac{\sigma_I^2}{n})$$

$$\frac{(r-1)n\, S_{WR}^2}{\sigma_{WR}^2} \sim \chi^2_{(r-1)n}$$

and the two quantities are statistically independent. Furthermore, we assume a balanced design and that no donor-by-formulation interaction exists

Assessing Bioequivalence

Mixed CDER criterion uses the intra (within) - reference variability as a cutoff point.

For $S_{WR} \leq 0.294$, the test and reference formulations are declared bioequivalent if the (1-2α) *100% confidence interval:

$$\bar{I}_{.} \pm t_{(n-1),\alpha} * \sqrt{\frac{S_I^2}{n}}$$

is contained within the limits $[\frac{1}{m}, m]$

Assessing Bioequivalence

The scaled BE methodology used in the case that $S_{WR} > 0.294$, adopts the FDA/CDER approach for the analysis of highly variable drugs, modified for the particular design.

The hypotheses to be tested are:

$$H_0: \frac{(\mu_T - \mu_R)^2}{\sigma_{WR}^2} > \ \theta$$

$$H_a: \frac{(\mu_T - \mu_R)^2}{\sigma_{WR}^2} \leq \ \theta$$

Where $\theta = \frac{(\ln(m))^2}{(0.25)^2}$

Assessing Bioequivalence

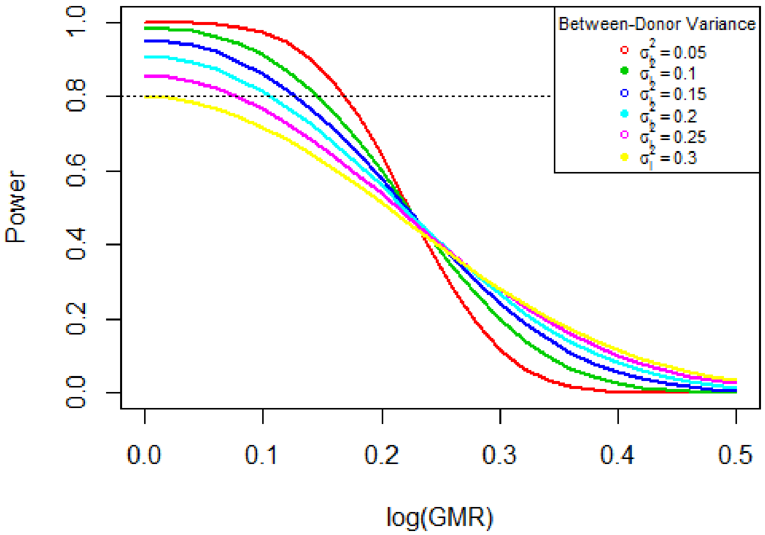
The strategy is to construct a (1-α) *100% confidence interval for the quantity $(\mu_T - \mu_R)^2 - \theta\ \sigma_{WR}^2$ and to observe its upper bound. If this is less than or equal to zero, H_0 will be rejected. Rejection of the null hypothesis, H_0, supports BE.

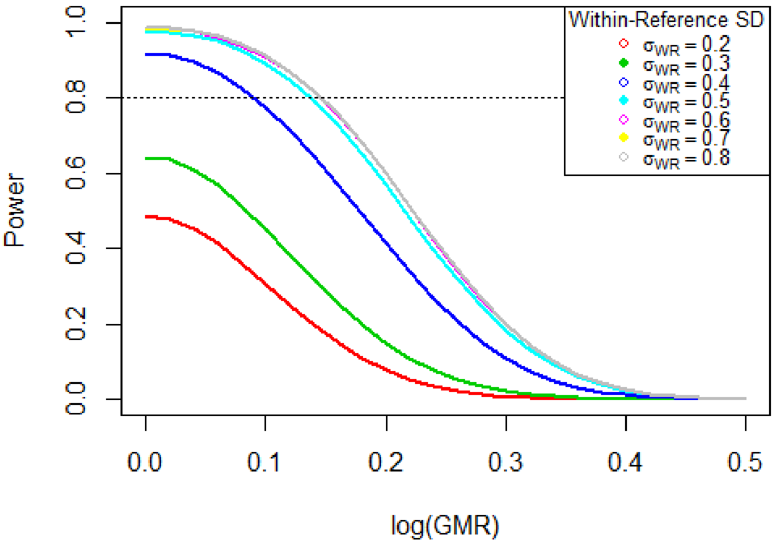
This criterion is accompanied by a ***point estimate constraint*** according to which the geometric mean ratio (point estimate of the log-transformed response has to fall within the pre-specified limits: $[\frac{1}{m}, m]$

Performance/Results

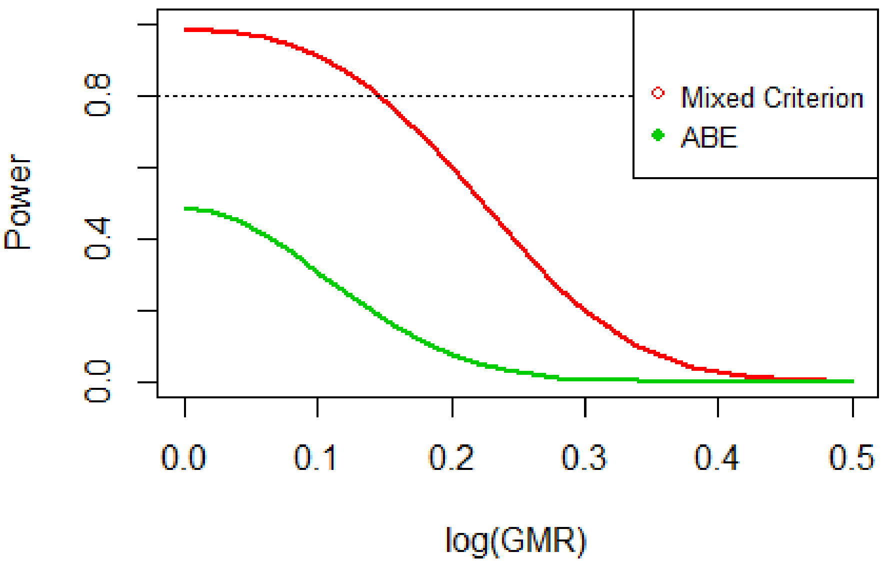
- The results obtained with IVPT and the suggested statistical analysis, were in agreement with the original results that led to regulatory approval of these products. This speaks in favor of the ***validity of this model for assessing BE***
- The test has been used for comparing two batches of the same reference product and successfully captured the similarity of these products in terms of BE. The outcomes advocate the ***model's sensitivity to meaningful differences and its resistance to the hazard of rejecting good products***

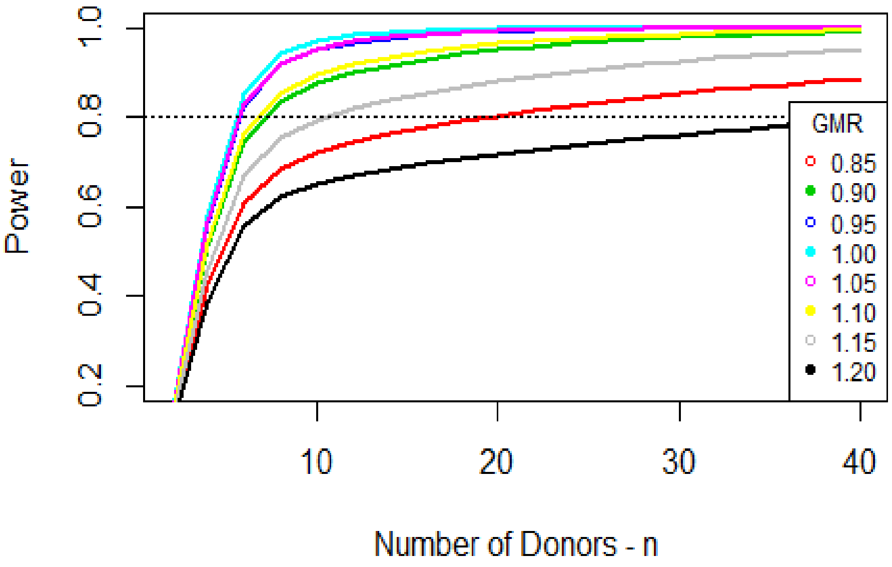
Power Analysis



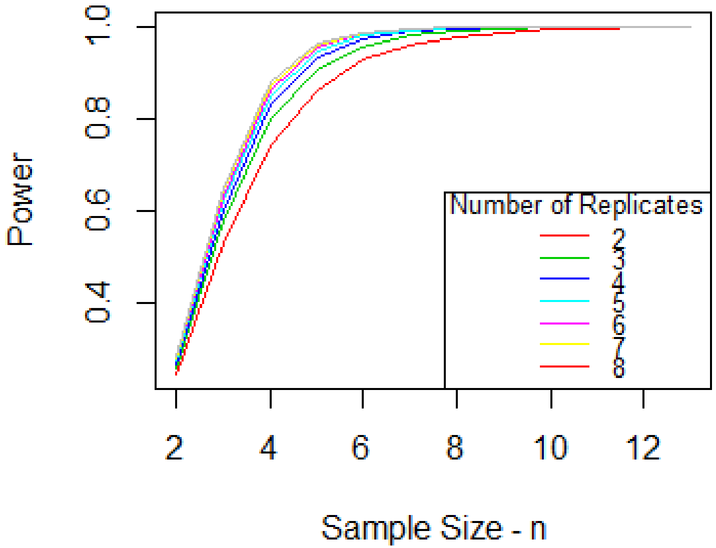


Power Analysis





Selecting the Optimal Number of Replicates



All sensitivity analyses have indicated that the number of replicates, $\boldsymbol{r}$, does not dramatically affect statistical power.

Since power appears to be stable for $\boldsymbol{r \geq 3}$, four replicates ($\boldsymbol{r = 4}$) was chosen as the suggested number in a study.

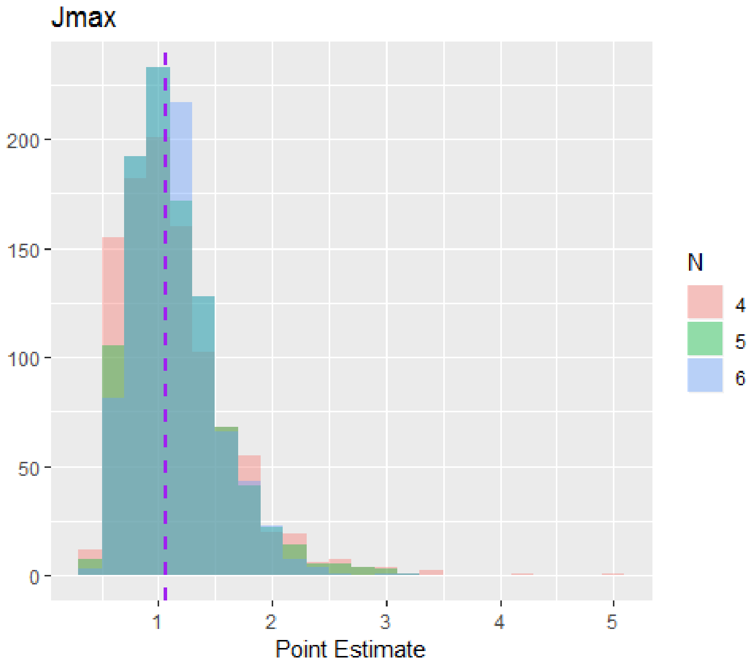
Selecting the Number of Donors for a Pilot Study

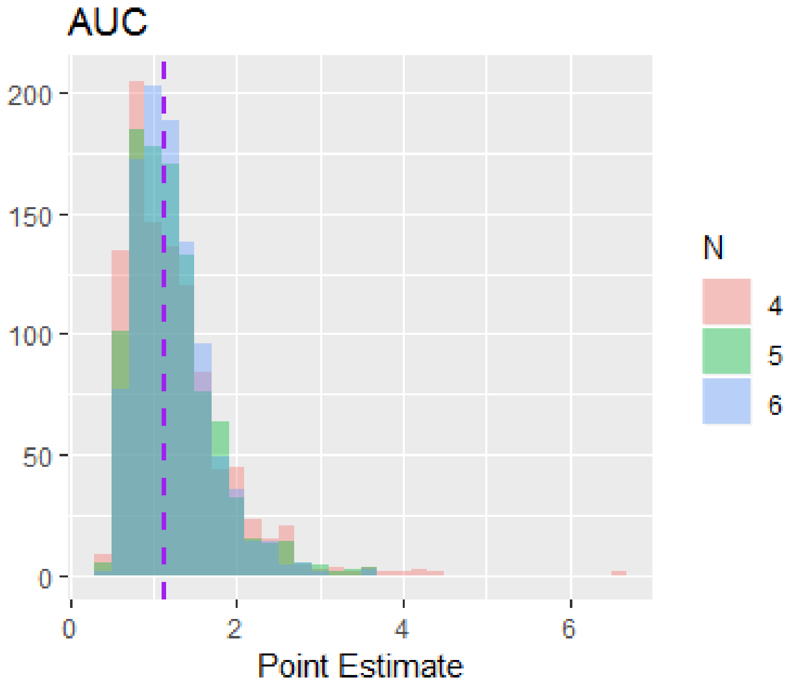
Simulation study Example 1

The distribution of the statistics $\bar{I}_{.}$ and S_{WR} was generated drawing 1000 bootstrap samples of sizes N=4, 5 and 6 from a ‘population’ of n=15 donors

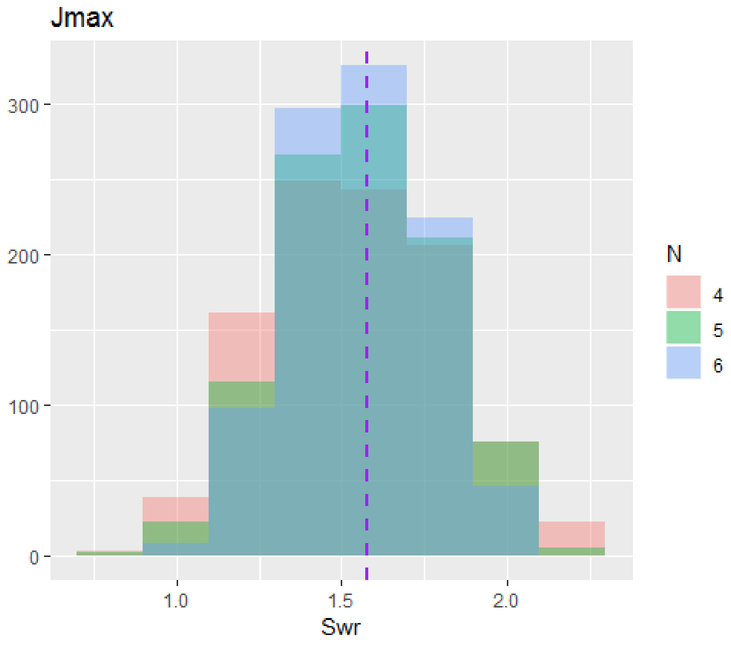
The shape and center of this distribution was then compared to the true value

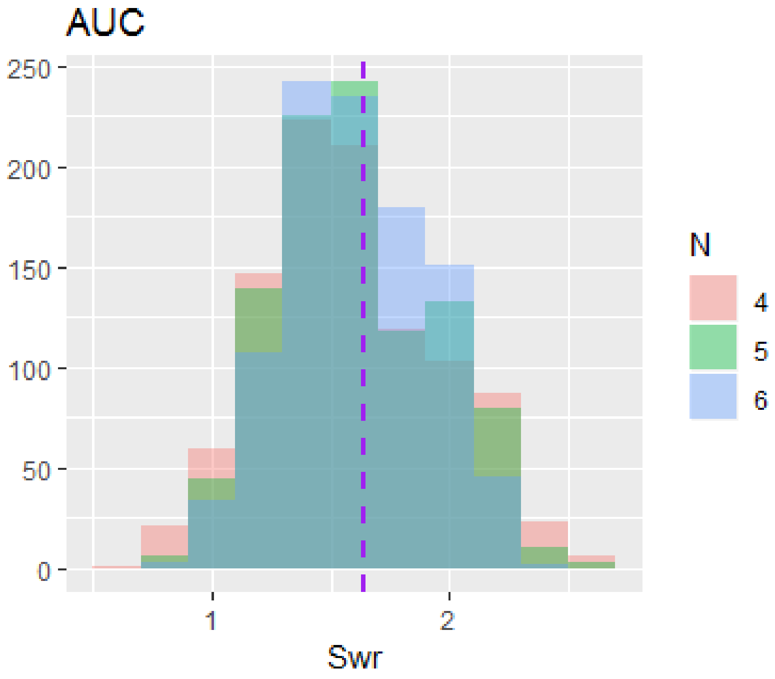
Selecting the Number of Donors for a Pilot Study for the Point Estimate, $\bar{I}$.





Selecting the Number of Donors for a Pilot Study for the Within-Reference SD, S_{WR}





Selecting the Number of Donors for a Pilot Study

Statistic	PK-metric	Point Estimate ("True Value")	Bootstrap distribution median N=4	Bootstrap distribution median N=5	Bootstrap distribution median N=6
$\overline{\boldsymbol{I}}_{.}$	AUC	1.130298	1.108010	1.142425	1.143698
	J_{max}	1.065301	1.054660	1.060841	1.090133
$\boldsymbol{s_{WR}}$	AUC	1.636378	1.534813	1.557838	1.589163
	J_{max}	1.577771	1.539472	1.563034	1.560031

Selecting the Number of Donors for a Pilot Study

- Power calculations showed no sensitivity of the power curve to the different estimates of variability
- All prior work on pilot study sample size selection indicates a constant improvement in precision, when the sample size increases
- Additionally, the choice of the sample size depends on the characteristics and variability of each data set

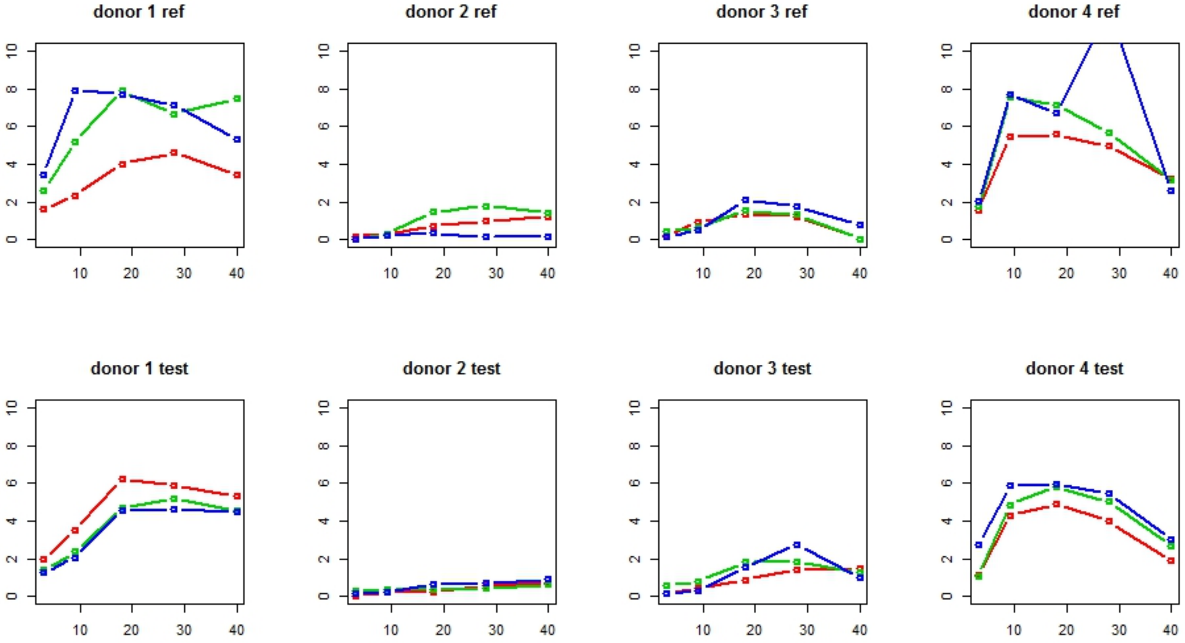
Outliers

- The nature of an outlying observation in this setup

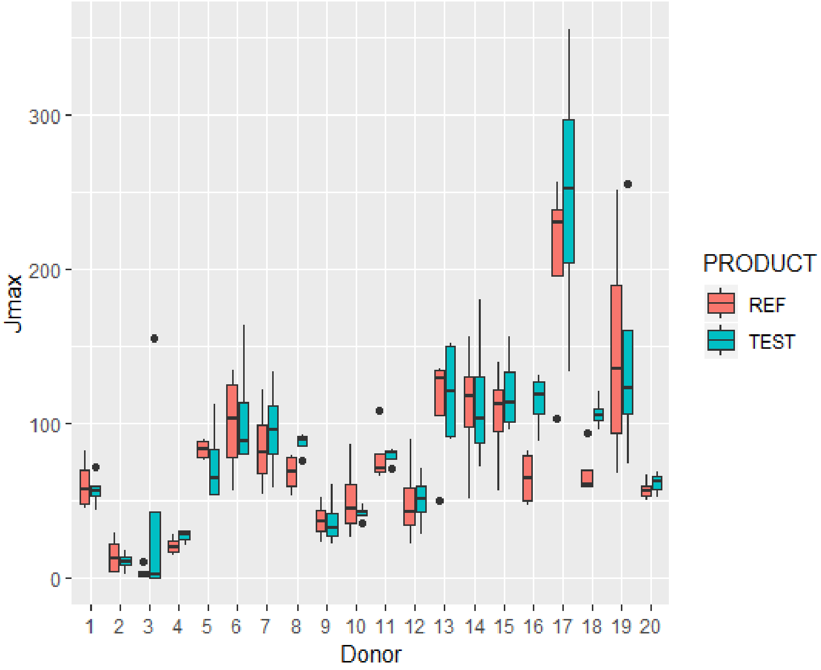
  Within-donor, extreme replicate values

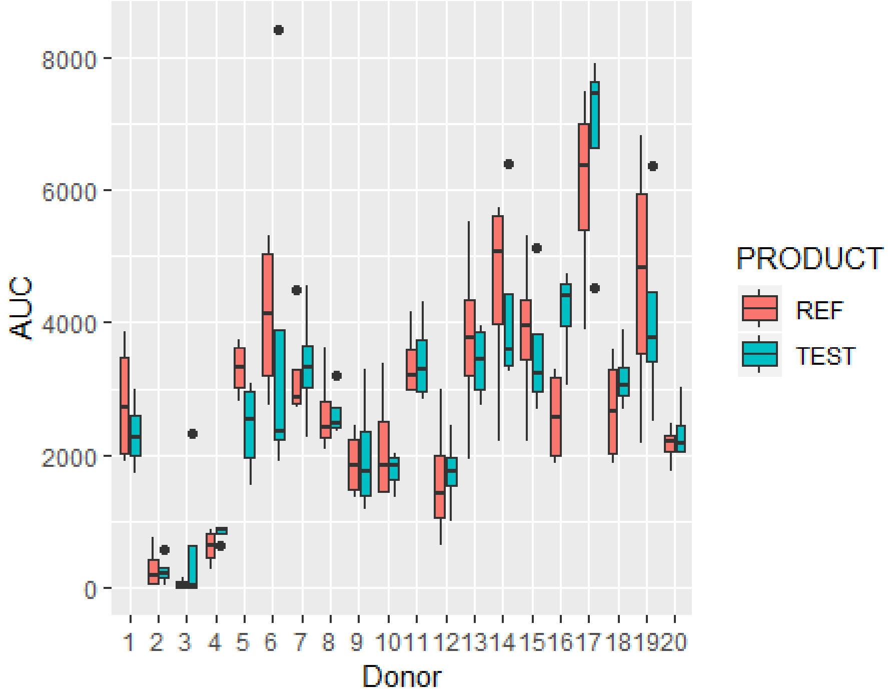
- Is it meaningful to consider ‘outlying donors’?

Outliers

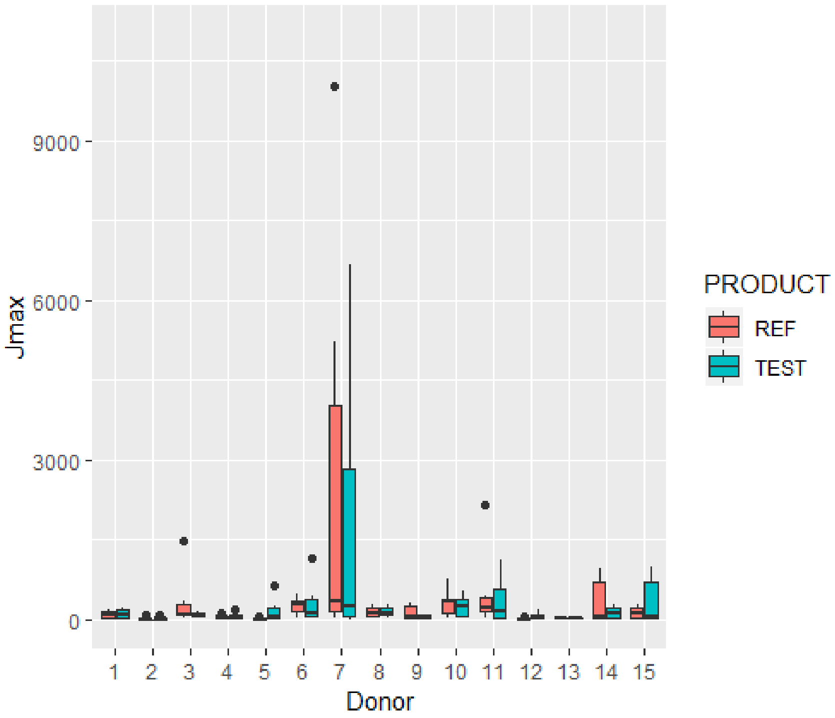

donor 1 ref
donor 2 ref
donor 3 ref
donor 4 ref
donor 1 test
donor 2 test
donor 3 test
donor 4 test
0
2
4
6
8
10
10
20
30
40

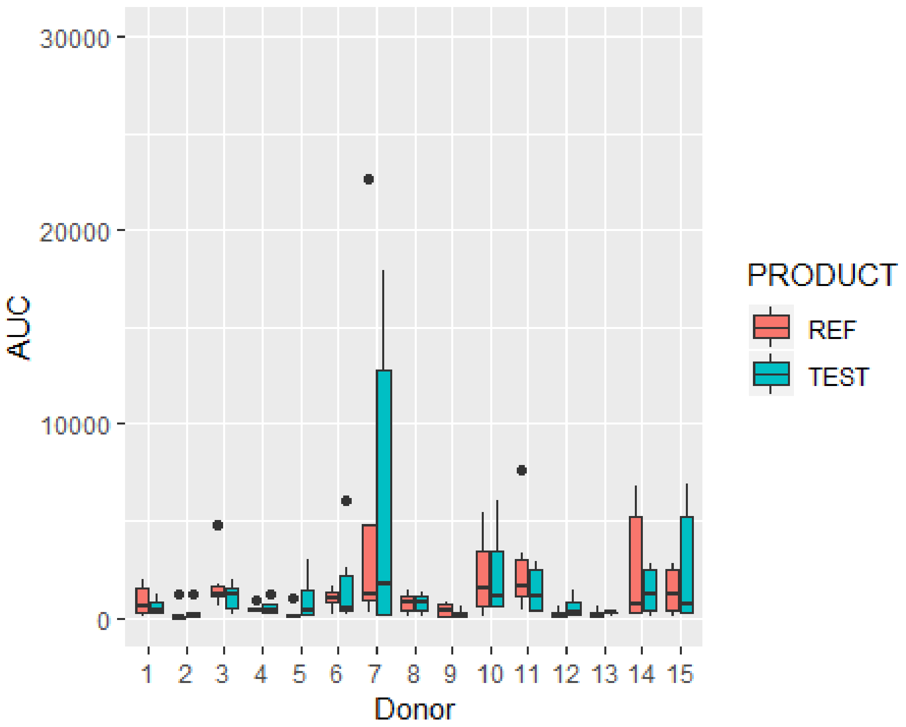
Example (20 donors)



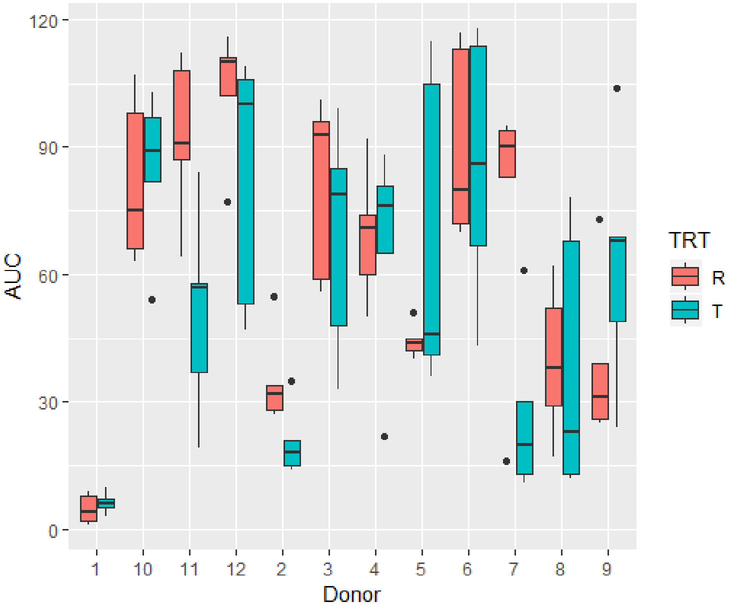


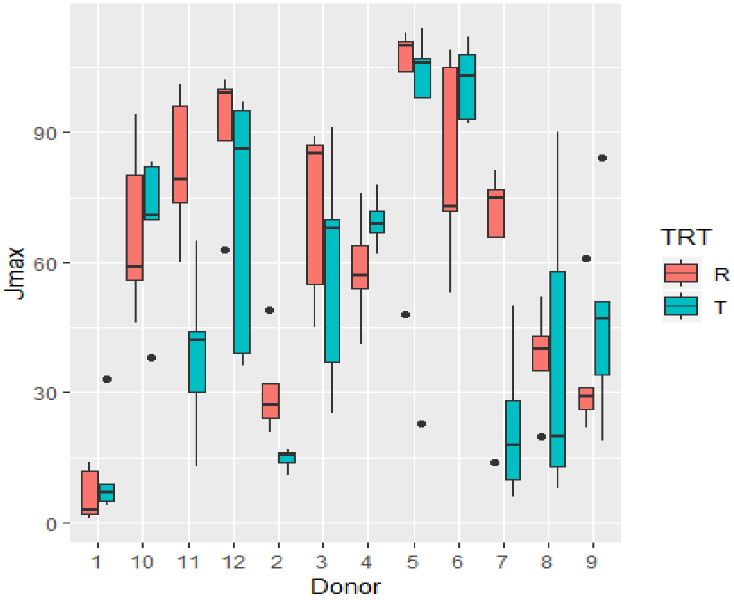
Example (15 donors)





Example (12 donors - 2 outlying cases)





Example (12 donors – 2 outlying cases)

Two donors were considered outlying due to their shape being distinctively different from the rest of PK-profiles

Excluding these donors from the data *did not affect* the BE outcome in both cases of AUC and Jmax

	PK-metric	**BE-outcome**
12 Donors	AUC	✓
	Jmax	✓
10 Donors	AUC	✓
	Jmax	✓

Outlier Detection

- Standard practices used in PK-studies (standardized residuals) do not apply here because of the small sample size of replicate values within one donor
- Is the Dean-Dixon test for outlier detection appropriate for small n in cases of experimental conduct anomalies that are detected once the sample analysis is completed?

Unbalanced Data

Replicate skin sections are withdrawn when

- Samples from the diffusion cell are destroyed (anticipated experimental event)
- If there is reason to believe that skin is damaged during the course of the experiment

In such cases, replicate values can be replaced so that there is no informational loss

Unbalanced Data

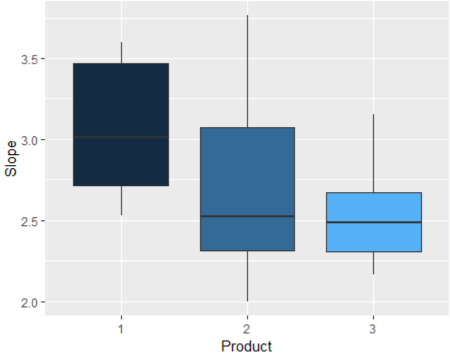
In situations that are

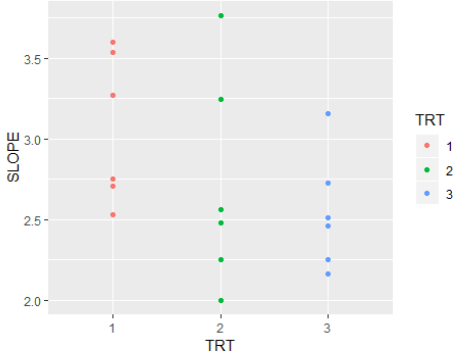
- Not pre-specified in the protocol or
- We are unable to replace the diffusion cell

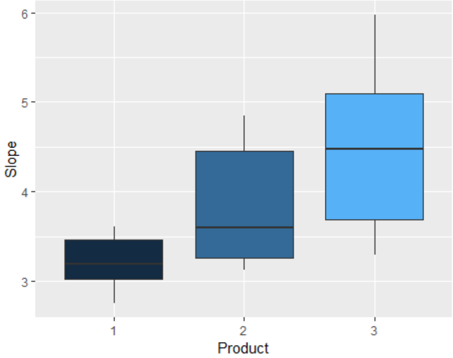
Replicate values are dropped but not uniformly for all donors. This violates the assumption of a balanced data set and can be potentially addressed by

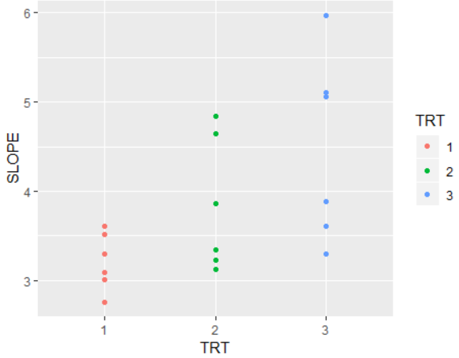
- Taking r: closest to the median replicates
- Randomly selecting r: replicates
- Imputing the missing replicate value with the average of the rest of replicate values for the same donor

IVRT









Two examples of 3 repeated runs of reference products, each having 6 replicate values of release slopes (cum. penetration on $\sqrt{sampling\ point}$)

IVRT

Is the Wilcoxon Rank-Sum test (as suggested by the SUPAC-SS guidance) still the most appropriate statistical analysis for IVRT data?

In particular

- What are the considerations for inflating type-I error under the two-stage structure of the test?
- What are the consequences of low power when the coefficient of variation (CV) is high?

References

Grosser, S., Park, M., Raney, S. G., & Rantou, E. (2015). Determining equivalence for generic locally acting drug products. *Statistics in Biopharmaceutical Research*, *7*(4), 337-345.

Grosser, S., Park, M., Raney, S. G., & Rantou, E. (2013). Clinical Endpoint Studies: Assessing Equivalence of Generic Locally Acting Drug Products. In *Encyclopedia of Biopharmaceutical Statistics* (pp. 1-7). CRC Press.

Acknowledgments

Stella Grosser

Don Schuirmann

Sam Raney

Priyanka Ghosh

Tannaz Ramezanli

Thank you!

FDA U.S. FOOD & DRUG

ADMINISTRATION