

Note: it is best to view the slides in presentation mode so the animations can be seen.
However there are presenter notes below the slides that may also be clarifying.
After the presentation there are some supplementary information that may be of interest.

Bayesian approach for similarity testing: concepts and examples

David.LeBlond@sbcglobal.net

Linas Mockus lmockus@purdue.edu

M-CERSI Workshop

In Vitro Dissolution Profiles Similarity Assessment in Support of Drug Product Quality:
What, How, and When

University of Maryland, Baltimore

May 21-22, 2019

Acknowledgements

to the colleagues who have developed these ideas over the last decade...

Yan Shen , Janssen R&D

John Peterson, GSK

Stan Altan, Janssen R&D

Hans Coppenolle, Janssen R&D

Areti Manola, Janssen R&D

Jyh-Ming Shoung, Janssen R&D

Oscar Go , Janssen R&D

Linas Mockus, Purdue University

Steve Novick, MedImmune

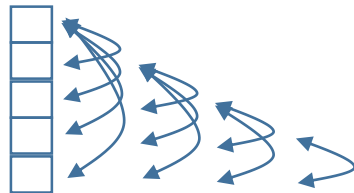
Harry Yang, MedImmune

... and to *YOU!!*

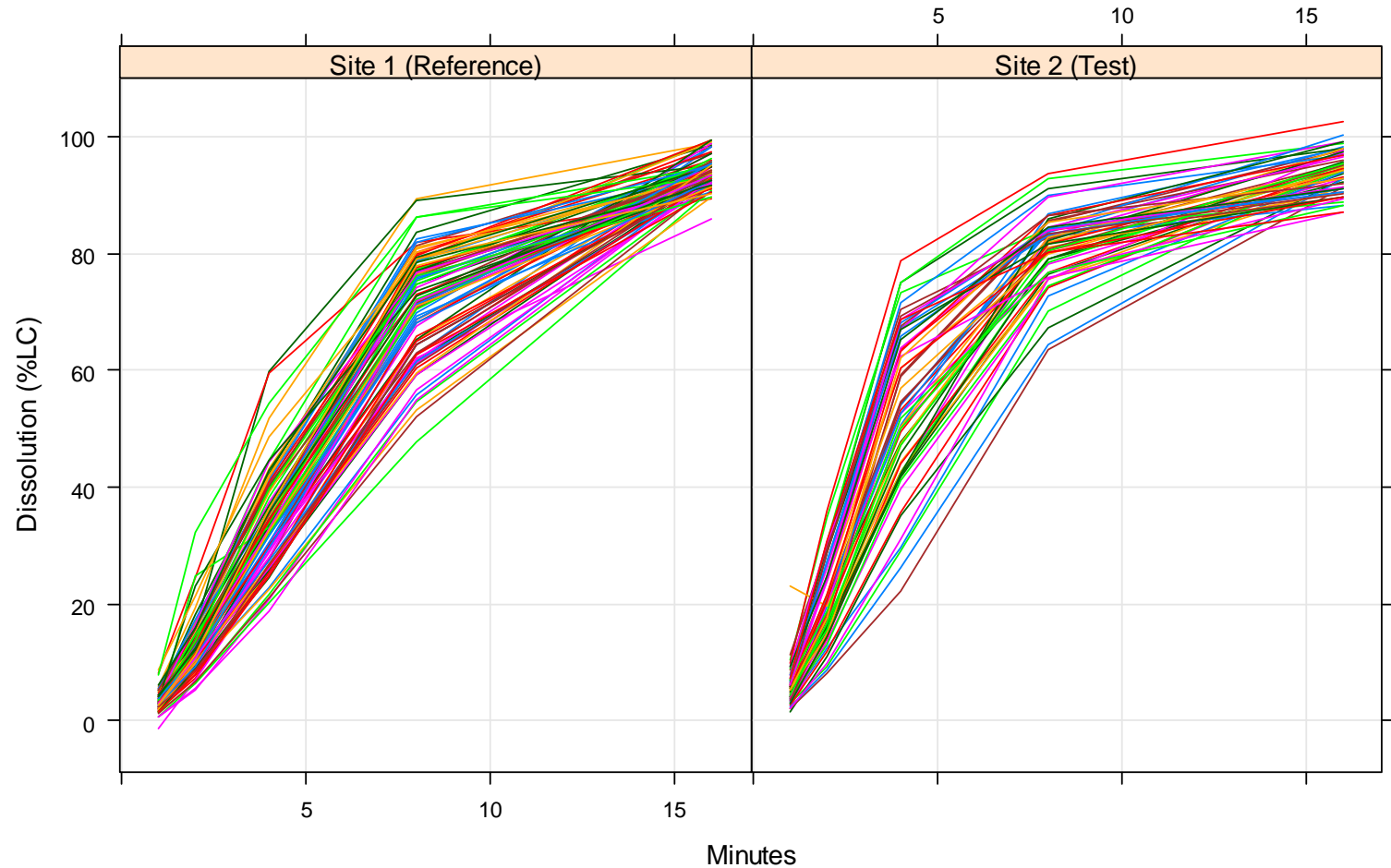
for your attention, and
for considering this approach

Some dissolution data:

- Rapidly dissolving
- IV dissolution similarity between 2 manufacturing sites
 - Site 1 (Reference): 8 lots
 - Site 2 (Test): 5 lots
- Tested in same laboratory
- 12 tablets per lot
- 5 Time points (minutes): 1, 2, 4, 8, 16
 - vector



- correlations



f_2 (non-Bayesian version)

	Site 1 (Reference)			Site 2 (Test)		
Minute	Mean	SD	%CV	Mean	SD	%CV
1	3.3	1.6	47.2	5.8	3.2	55.4
2	12.4	4.3	34.6	20.2	6.8	33.5
4	33.4	8.0	23.9	53.7	13.5	25.1
8	71.3	8.7	12.2	80.8	6.0	7.5
16	93.8	2.6	2.7	93.9	3.5	3.7

$f_2 = f(\text{Test data, Reference data})$

$f_2 = 38.0$



The Bayesian answer

Question: “My ~~diagnostic~~ test result was X. Do I have the ~~disease~~?”

- Non-Bayesian answer: “Given ~~no disease~~, the probability of X or worse is P.”
- Bayesian answer: “Given X (and other knowledge), the probability of ~~disease~~ is P.”

Which answers the question?

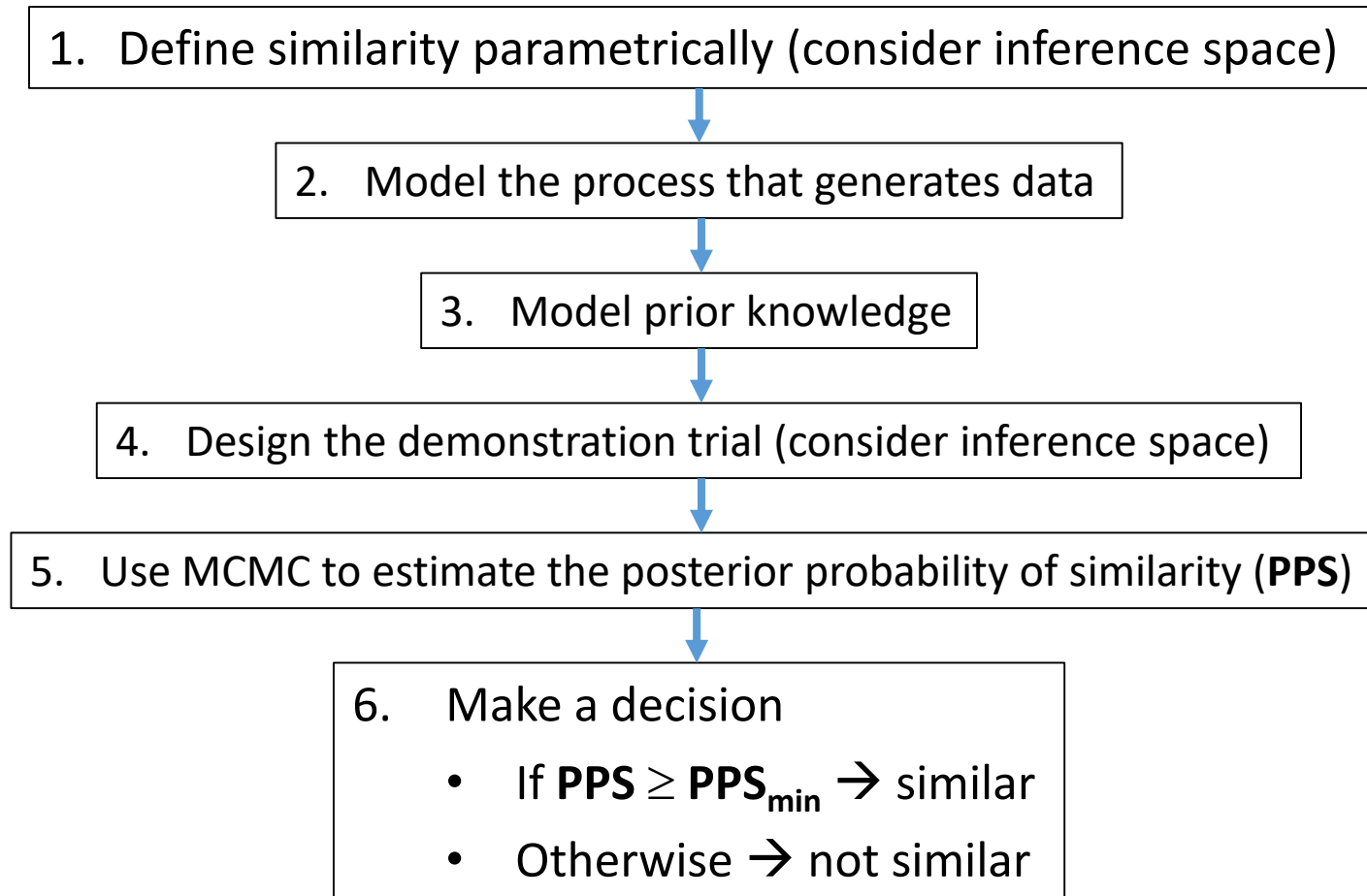
The Bayesian answer...

- directly addresses the question
- quantifies the answer as a **probability**
- leverages relevant & justifiable prior knowledge
- is conditional on observed (rather than hypothetical) data



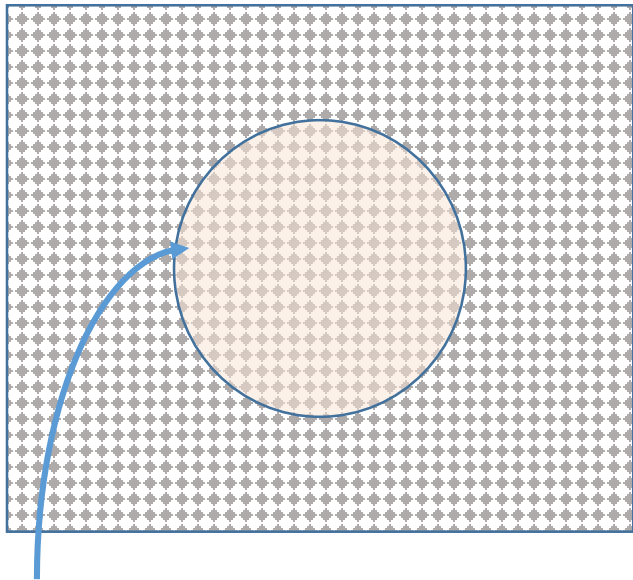
ICH Q9: “... risk is defined as the combination of the **probability** of occurrence of harm and the severity of that harm ... the protection of the patient by managing the risk to quality should be considered of prime importance.”

A Bayesian decision tree* for *in vitro* similarity



* More like a telephone pole – no branches

1. Define similarity region parametrically (what is the inference space?)



Subset we define as similar (Region of Similarity)

Set of all hypothetical:

- Dissolution profiles, or
- Profile differences, or
- Model parameters, or
- Parameter differences, or
- Univariate metrics, or
- ...

Define the comparison:

- Test vs Reference?
- Test vs some standard of performance?

What is similar?

- processes that make lots?
- lots tested?
- tablets tested?
- data results?

Define the metric of similarity

- Based on the *state of nature* we require
- Not dependent on observed data, experimental design, or analysis methodology
- Multivariate?
- Profile model parameters?
- Univariate (e.g., f2)?

1. (cont) Candidate similarity regions

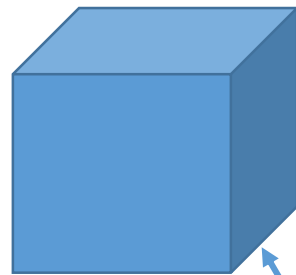
Univariate

$$F_2 = f(\text{true Test quantities, true Reference quantities}) \geq 50$$

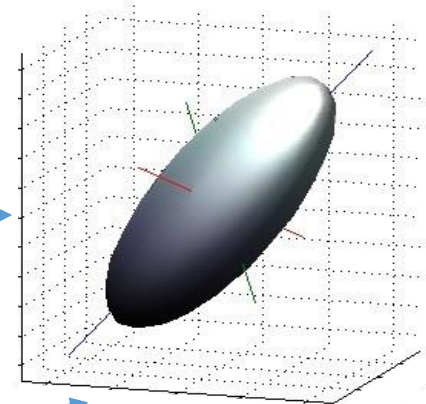
$$F_2 = f(\text{true Test quantities, fixed Standard quantities}) \geq 50$$

Multivariate

Hyper-rectangle



Hyper-ellipsoid



Allowable ranges for
Test – Reference or fixed
standard quantities of 3
time points

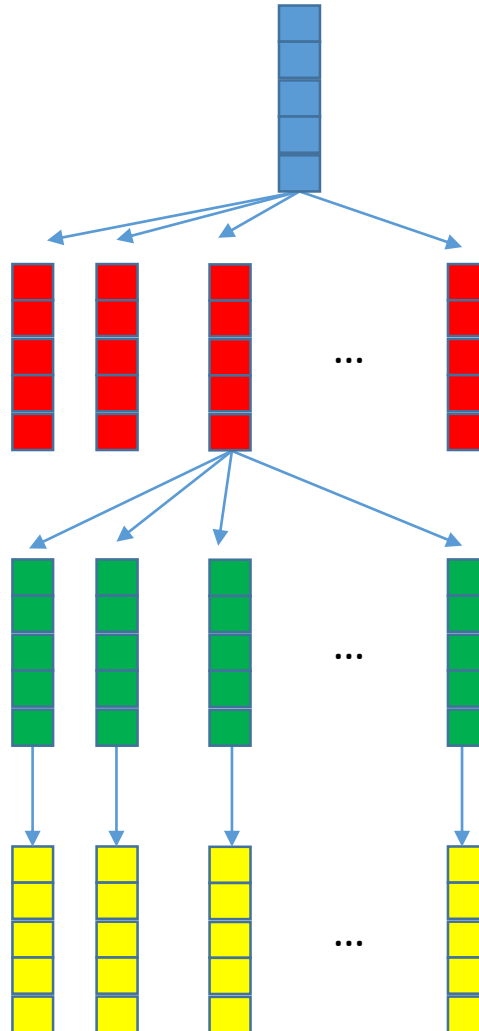
2. Model the process that generates data

Site process means
(fixed)

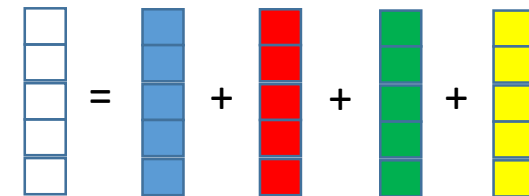
Lot to lot deviations
(random multivariate normal)

Tablet to tablet deviations
(random multivariate normal)

Analytical deviations
(random univariate normal)

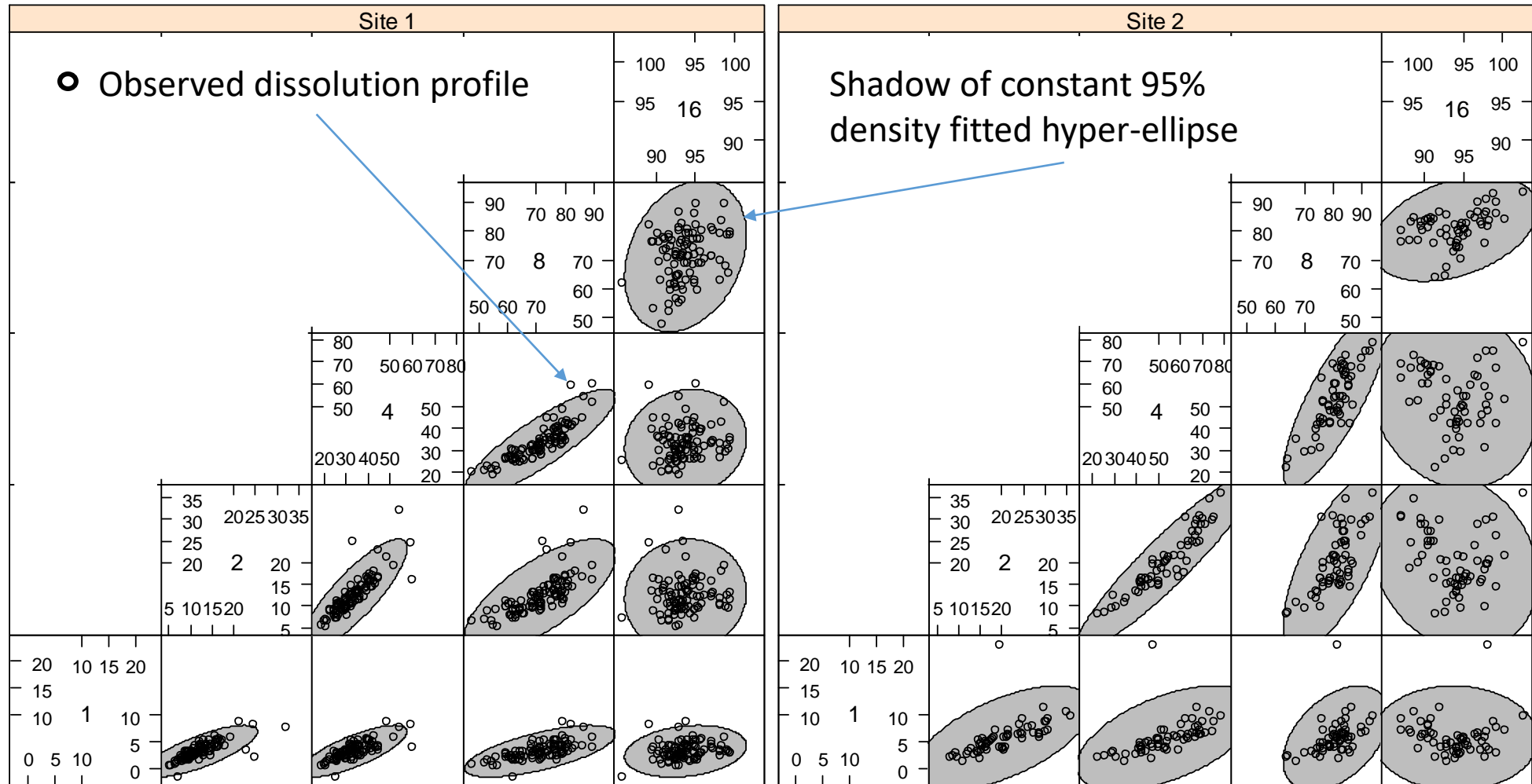


Observed dissolution profile =



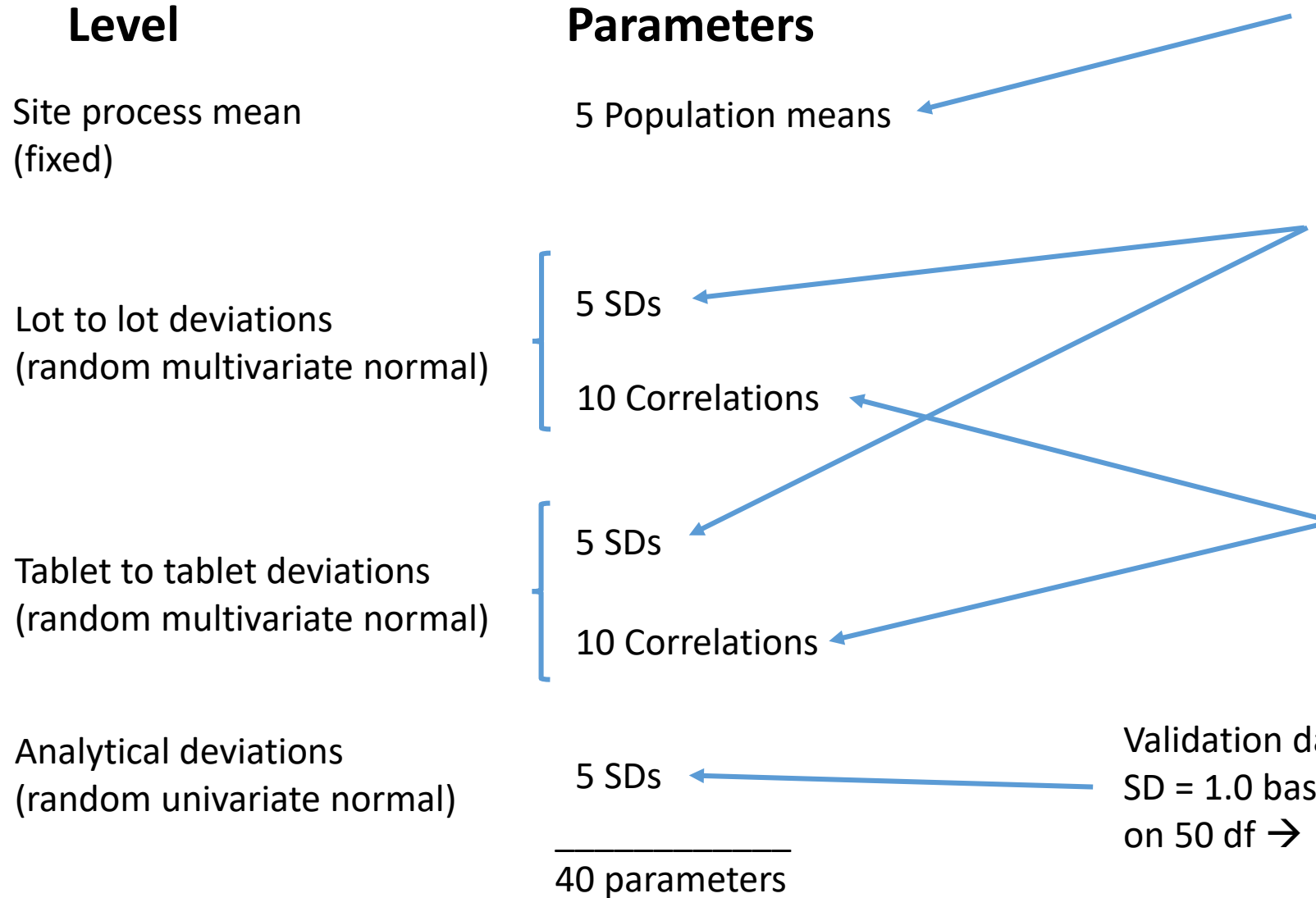
- For illustration, we will focus on the process mean level
- Inferences at other levels are equally possible

2. (cont) Modeling correlations among 5 time points

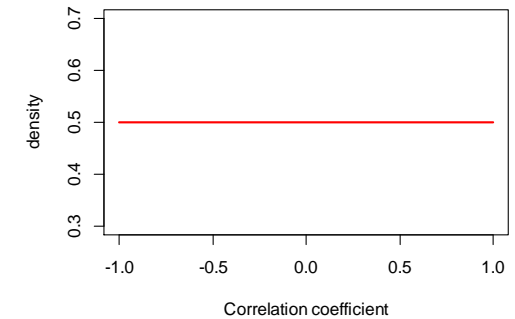
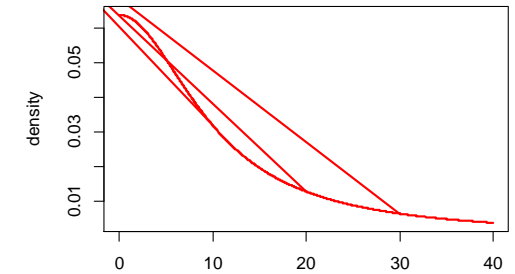


Scatter Plot Matrix

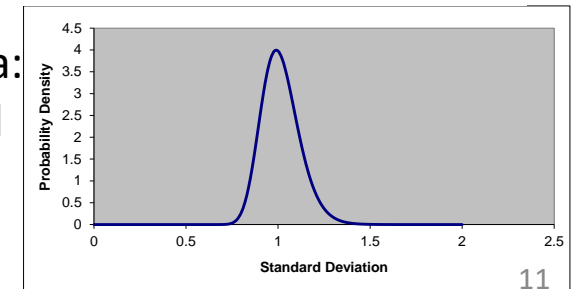
3. Model prior knowledge



Prior distributions



Validation data:
SD = 1.0 based
on 50 df →

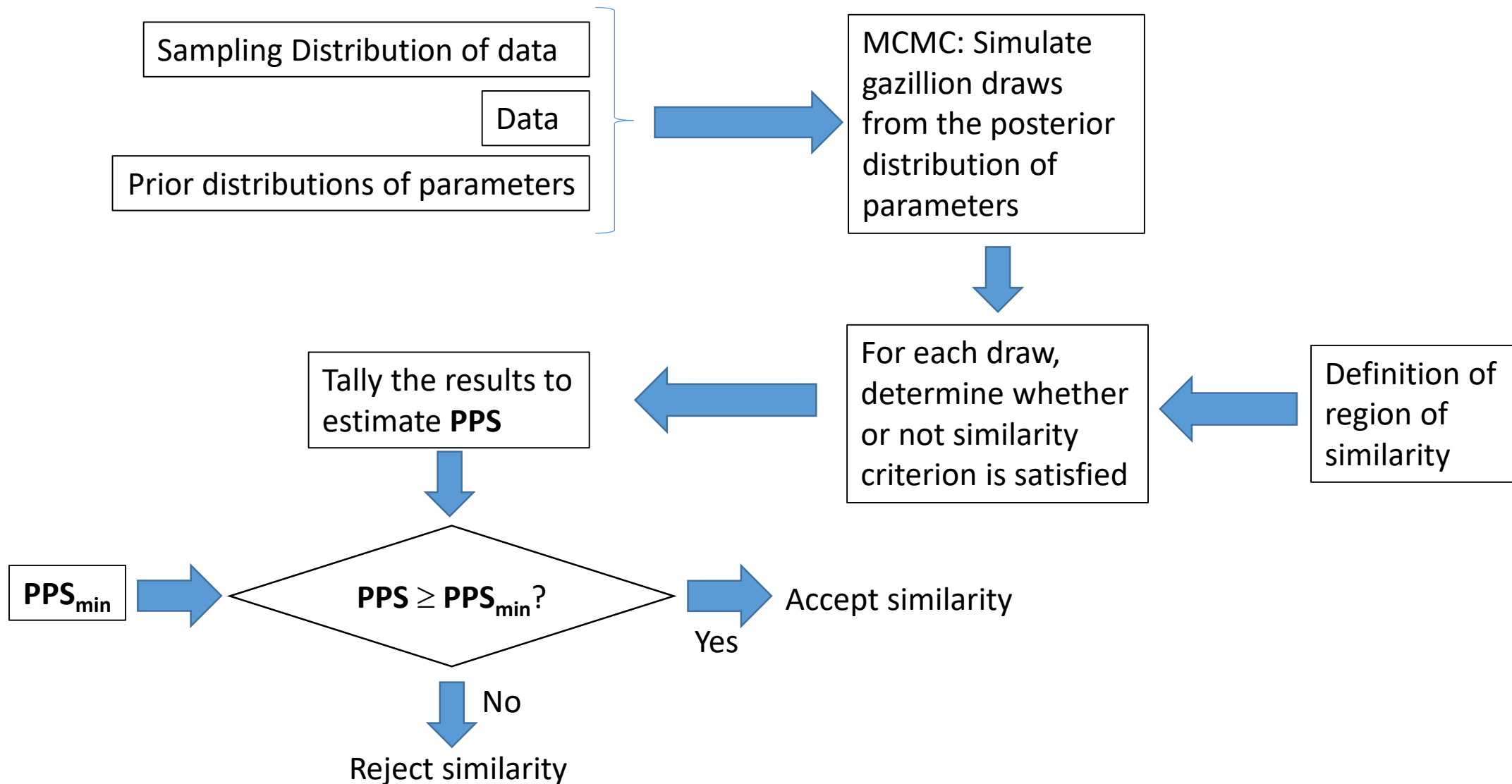


4. Design the demonstration trial

(consider inference space)

- Time points?
- Burden of proof?
 - H0: Assume similarity unless contradicted by the trial (“difference test”)?
 - H0: Assume non-similarity unless contradicted by the trial (“equivalence test”)?
- Required statistical confidence (probability of incorrectly rejecting H0) and power (probability of correctly rejecting H0)?
- Sources and magnitude of variances?
 - Inter-lot?, Inter-tablet within lot?, analytical?
- Number of lots from Test and Reference (unless comparison is to a fixed standard)?
- Sampling plan for lots?
- Number of tablets from each lot?
- Decision metric and its acceptance criterion? (Bayesian: **PPS** and **PPS_{min}**)

5. Use MCMC to estimate PPS



Three illustrations

1. F_2 (Bayesian version), univariate similarity region
2. Hyper-rectangular multivariate similarity region
3. Hyper-ellipsoid multivariate similarity region

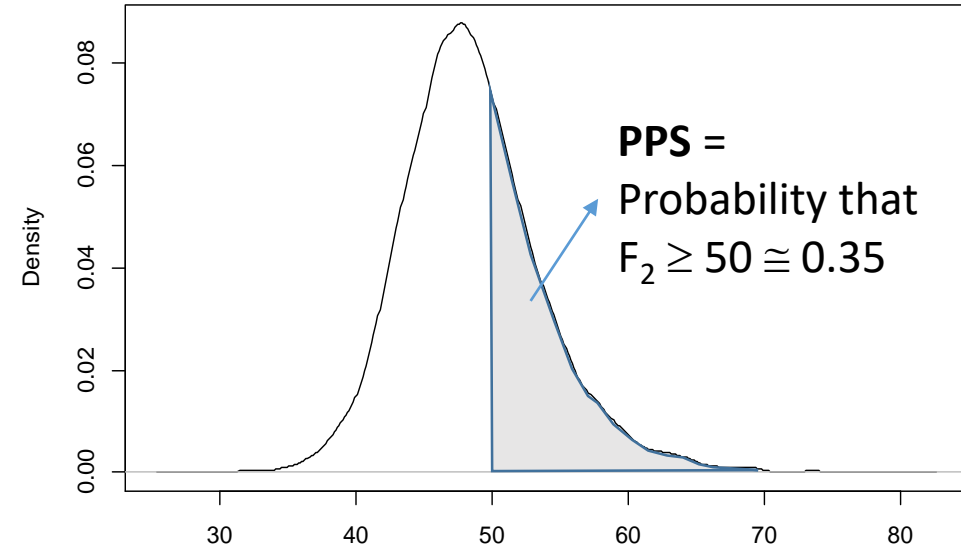
F_2 (Bayesian version): Estimating PPS

MCMC draws from joint posterior distribution of process means



Draw	Site 2 process mean (Test)					Site 1 process mean (Reference)					Difference (Site2 - Site1)					F_2
	1	2	4	8	16	1	2	4	8	16	1	2	4	8	16	
1	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
2	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
...
15999	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
16000	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•

Posterior Distribution of F_2



F_2



Count the fraction of F_2 posterior draws that are ≥ 50

Hyper-rectangle: Defining similarity

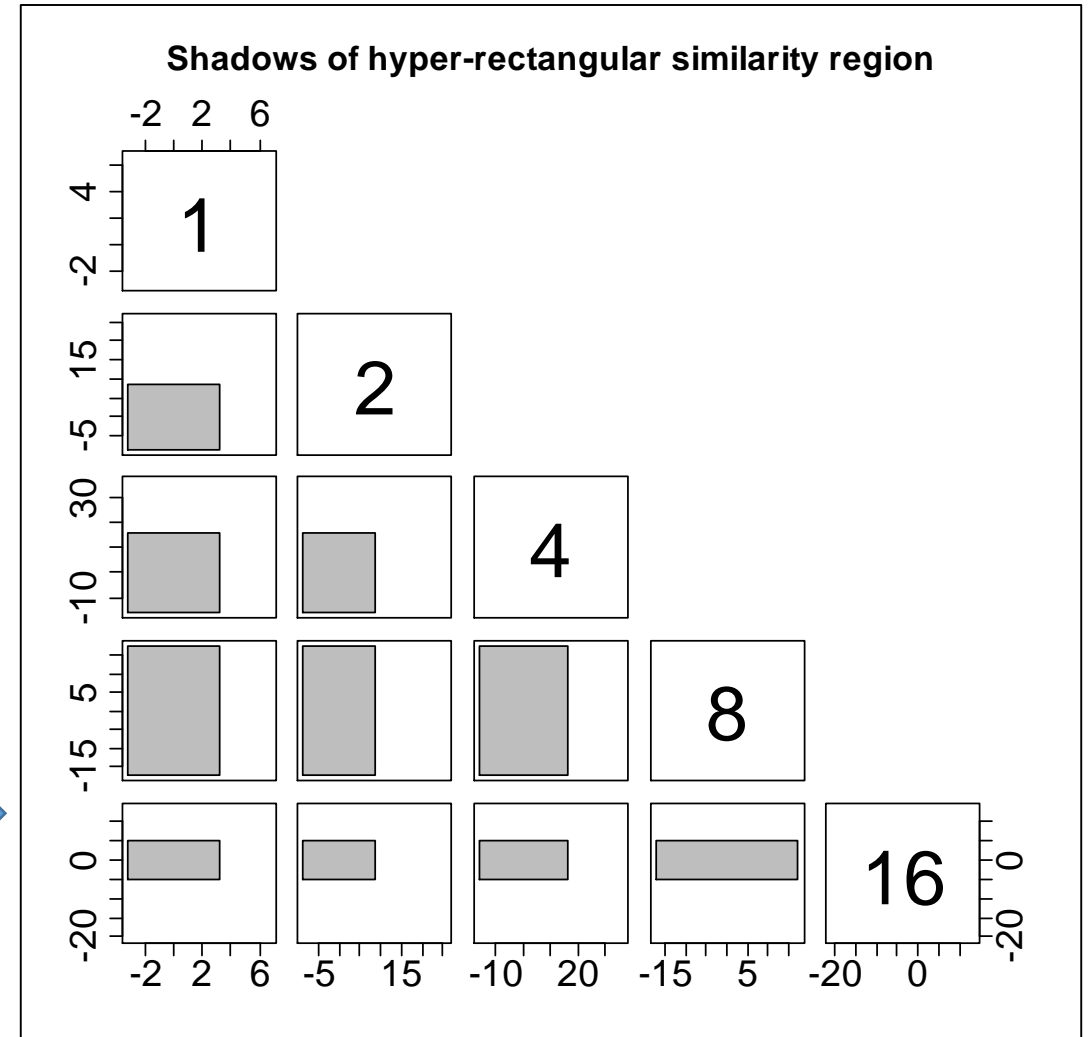
Set a *fixed* similarity range for each time point based on ...

- Deviations from mean profiles of site 1 clinical lots?
- Efficacy/safety considerations (if any)?
- Process capability? Tolerance intervals? n-sigma?
- Negotiation with regulators?
- ...

e.g.,

Minute	Mean	Range
1	3.3	± 3.2
2	12.4	± 8.6
4	33.4	± 16.0
8	71.3	± 17.3
16	93.8	± 5.1

- Ranges define a *fixed* hyper-rectangular similarity region
- Applied to *all future* (Test – Reference) similarity tests.



Hyper-Rectangle: Estimating PPS

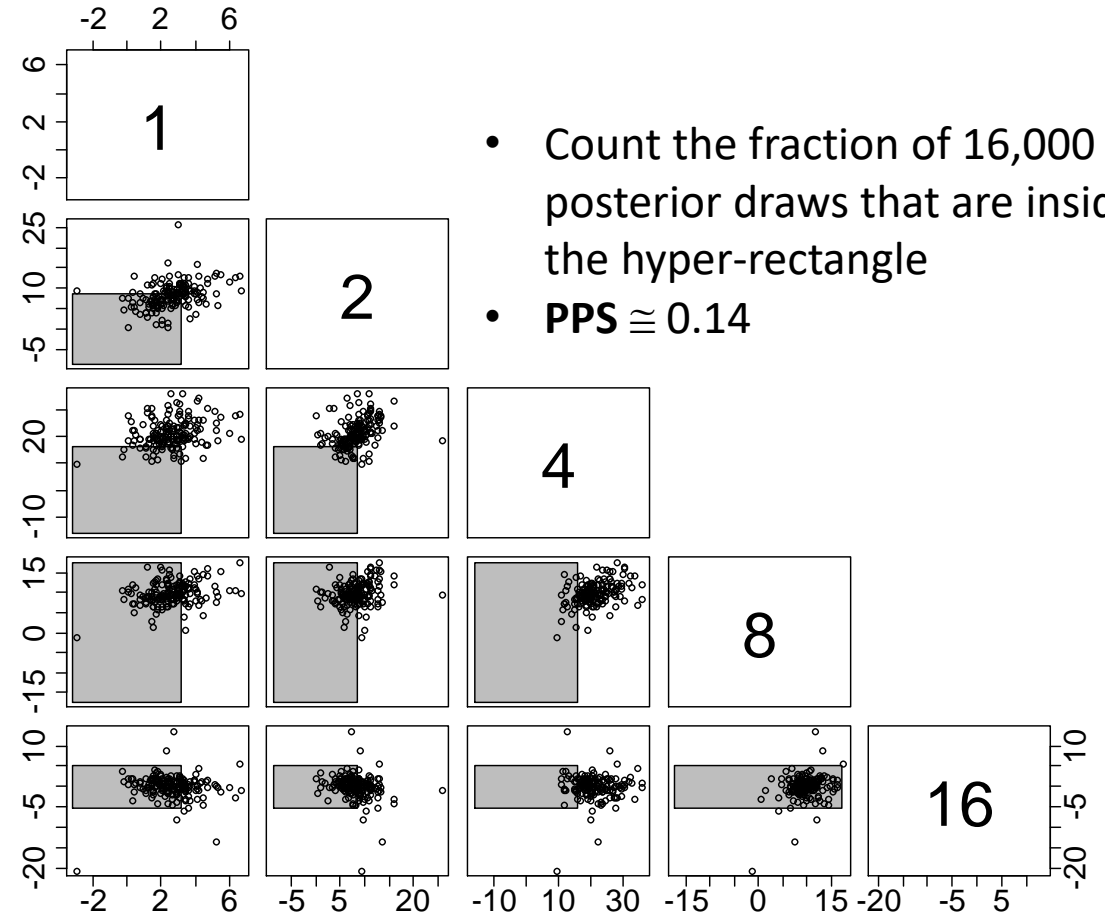
MCMC draws from joint posterior distribution of process means



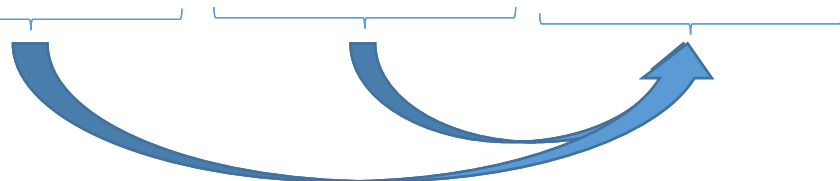
Draw	Site 2 process mean (Test)					Site 1 process mean (Reference)					Difference (Site2 - Site1)				
	1	2	4	8	16	1	2	4	8	16	1	2	4	8	16
1	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
2	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
...
15999	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
16000	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•



Posterior Sample of Process Mean Differences (Site 1 - Site 2)
1% of 16,000 draws shown



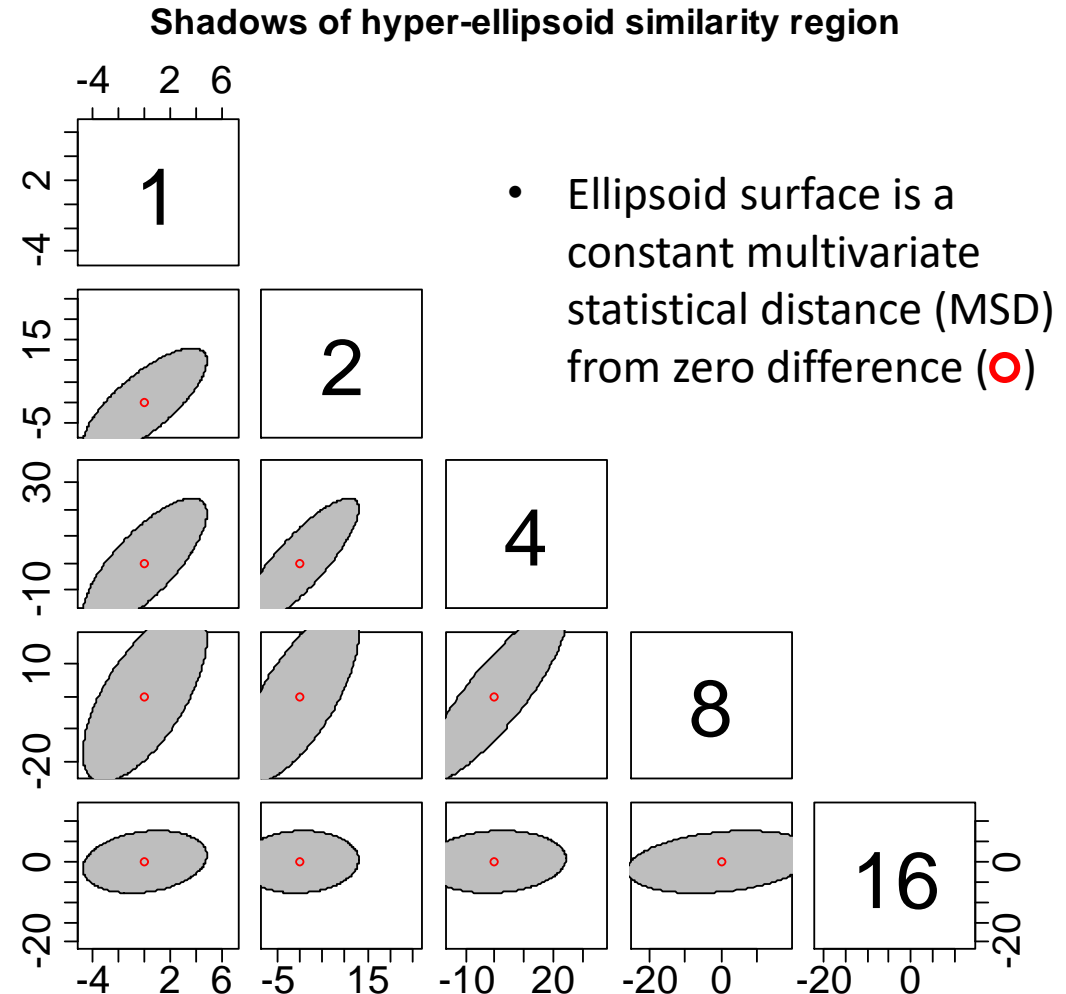
- Count the fraction of 16,000 posterior draws that are inside the hyper-rectangle
- **PPS** $\cong 0.14$



Hyper-ellipse: Defining similarity

Set a *fixed* similarity region based on ...

- Multivariate normal distribution model of dissolution profiles from clinical lots (site 1)
- Constant density (95% ?) ellipsoid
- Ellipsoid is centered about zero difference
- Ellipsoid similarity region defined by a multivariate covariance matrix
- Hyper-ellipse region is *fixed* and applied to *all future* (Test – Reference) similarity tests

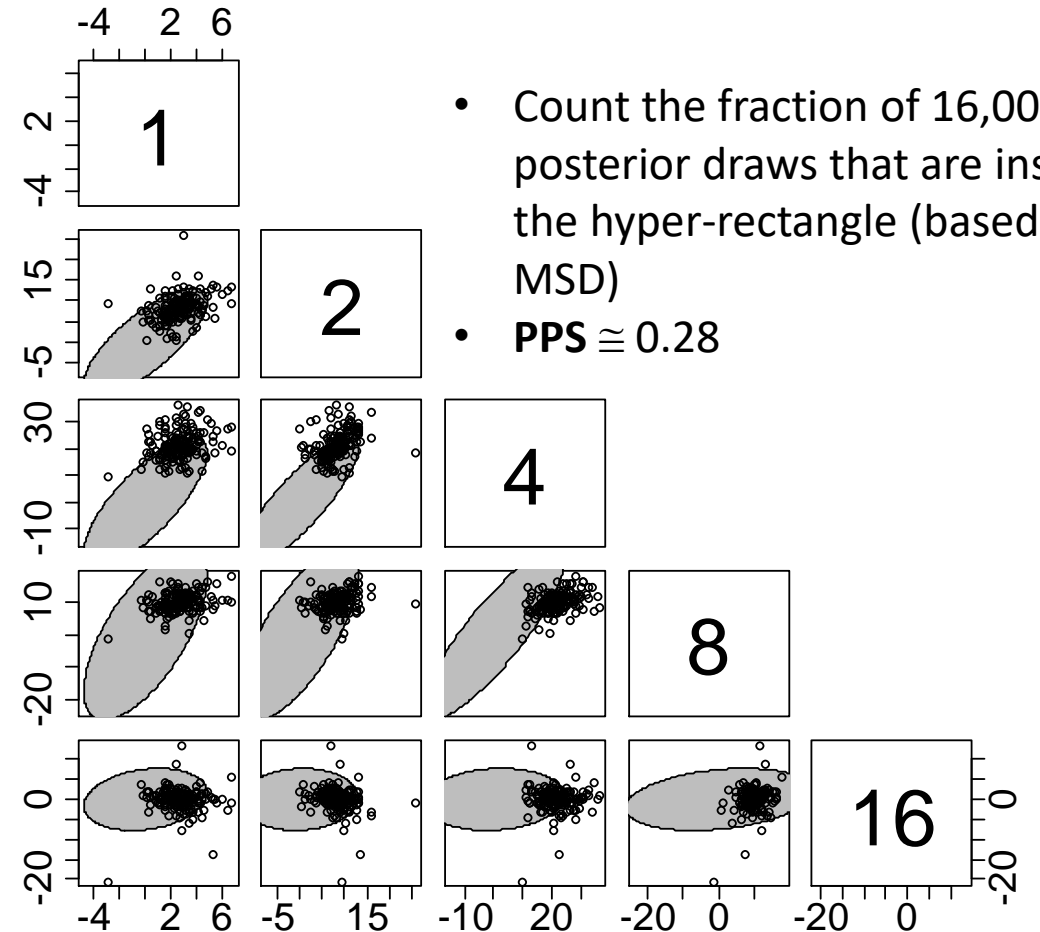


Hyper-ellipse: estimating PPS

MCMC draws from joint posterior distribution of process means

Draw	Site 2 process mean (Test)					Site 1 process mean (Reference)					Difference (Site2 - Site1)				
	1	2	4	8	16	1	2	4	8	16	1	2	4	8	16
1	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
2	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
...
15999	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
16000	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•

Posterior Sample of Process Mean Differences (Site 1 - Site 2)
1% of 16,000 draws shown



- Count the fraction of 16,000 posterior draws that are inside the hyper-rectangle (based on MSD)
- **PPS** \cong 0.28

Not seen here ...

- Test vs Fixed-standard similarity
- Similarity at other hierarchical levels
 - e.g., probability that *future lots* made by the Test process will fall within a similarity region
- More complex models (e.g., non-normal, additional sources of variance)
- Profile models (e.g., Weibull)
- Dissolution instability

... all eminently amenable to Bayesian modeling tools

Bayesian approach: pros and cons

- Pros**
- Probability metric (**PPS**) supports risk assessment
 - A single coherent approach (univariate, multivariate, \pm profile model,...)
 - Based on simple counting exercise (MCMC)
 - Leverages prior information as appropriate
 - Equivalence format rewards good experimental design & high data information content
 - Software (BUGS, JAGS, Stan, SAS) widely available
- Cons**
- Software novel/unfamiliar
 - Forces difficult (but critical) communication
 - Coverage properties require calibration studies
 - Regulatory acceptance?

Recommendations

Regulators

1. Revise global guidance
2. Not proscriptive. Stress “best statistical practices”. Cite literature/ textbook references.
3. Explicitly include Bayesian options.
4. Explicitly include a “Test vs Fixed-standard” similarity option.
5. General guidance about specifics*
6. Outline conditions for use of prior knowledge

Sponsors

1. Consider a Bayesian approach. If appropriate, ...
2. Collaborate with a statistician familiar with Bayesian tools and good experimental design.
3. Recommend specifics*, negotiate with regulators in consideration of risk.

*inference space, similarity region, confidence, power, PPS_{\min}

*Thank You for
your attention !!*

References

1. [LeBlond D \(2009\) Dissolution stability of a modified release product, 32nd MBSW, May 19, 2009,](#)
2. Shen Y, LeBlond D, J Peterson, S Altan, H Coppenolle, A Manola, J-M Shoung (2011) A Bayesian Approach to Equivalence Testing in a Non-linear Mixed Model Context, Non-Clinical Biostatistics Conference Boston, MA Oct 19, 2011
3. [LeBlond D, Peterson J, & Altan S \(2011\) The posterior probability of dissolution equivalence, Midwest Biopharmaceutical Statistical Workshop, Muncie IN, May 25, 2011](#)
4. [Novick S, Shen Y, Yang, Peterson J, LeBlond D & Altan J \(2015\), Dissolution Curve Comparisons through the F2 parameter, a Bayesian extension of the f2 statistic, Journal of Biopharmaceutical Statistics 25:2 351-371](#)
5. [LeBlond D, Altan S, Novick S, Peterson J, Shen Y, & Yang H \(2016\) In Vitro Dissolution Curve Comparisons: A Critique of Current Practice, Dissolution Technologies, February, 2016](#)
6. [Altan A, Coppenolle H, LeBlond D, Manola A, Mockus L, Novick S, Peterson J, Shen Y, Shoung J-M, Yang H, MedImmune \(2018\) The posterior probability of in vitro dissolution equivalence, ASA Biopharmaceutical Section Regulatory-Industry Statistics Workshop, Washington DC, Sept 12-14, 2018](#)
7. Mockus L & LeBlond D (expected, 2019) Bayesian methods for in vitro dissolution drug testing and similarity comparisons, in Bayesian Methods in Pharmaceutical Research (editors: Lesaffre E, Baio G, & Boulanger B), CRC Press

The goal of this presentation was to communicate (through a worked example) the following statistical perspectives:

1. The shape of an n-time point dissolution profile is uniquely captured as a point in n-dimensional space where the n orthogonal axes are the % dissolution observed (or expected) at each time point.
2. This n-dimensional representation provides a useful way to represent a region of similarity such as a dissolution safe space or an analogous process control space that defines regions within which the dissolution profiles obtained from bioequivalent or in-control batches, respectively, are expected to lie.
3. Similarly differences between 2 dissolution profiles (e.g., TEST – REF) are uniquely captured as a point in n-dimensional space where the n orthogonal axes are the difference in % dissolution observed (or expected) at each time point.
4. While it is impossible to visualize such n-dimensional spaces and the regions within them when $n > 3$, they can be conveniently displayed as a matrix of bivariate projections. They can also be easily manipulated by computer for purposes of statistical analysis.
5. This representation provides a revealing comparison between the popular metrics of similarity such as f_2 , Mahalanobis distance (MD), and safe (or process control) space. It can be easily shown that in this representation:
 - a. The f_2 metric defines an n-dimensional hypersphere centered at zero with a radius equal to $\sqrt{99n}$.
 - b. The MD metric defines an n-dimensional hyper-ellipse centered at zero whose axes lengths and rotations are governed by the SDs of the differences at each time point, their mutual correlations, and other statistical constants.
 - c. A safe (or process control) space defines an n-dimensional hyper-rectangle centered at the median dissolution profile whose edges have the length of the allowable range at each time point.
 - d. A requirement such as “no more than X difference at any of the n time points” defines an n-dimensional hypercube centered at zero with edges of length X.
6. Dissolution profile comparison is essentially a multivariate problem that requires a multivariate decision metric. Reducing the decision metric to a univariate metric such as f_2 or MD, leads to many fundamental difficulties:
 - a. The connection to profile shape, and thus to bio-relevance, is masked. There is no longer a 1:1 and onto mapping between a given profile shape/difference and a single point in n-dimensional space. Profile differences are averaged across time points.
 - b. Such reduced metrics prove overly sensitive to the choice of the number and location of time points.
 - c. The sampling distribution of f_2 is complex, requiring adjusted bootstrapping or asymptotic methodologies to obtain approximate statistical tests.
 - d. The MD metric uses a covariance weighting among time points. In this way time points are weighted statistically but not necessarily with respect to bio- or process control- relevance. The need to estimate the covariance matrix adds uncertainty to the similarity decision.

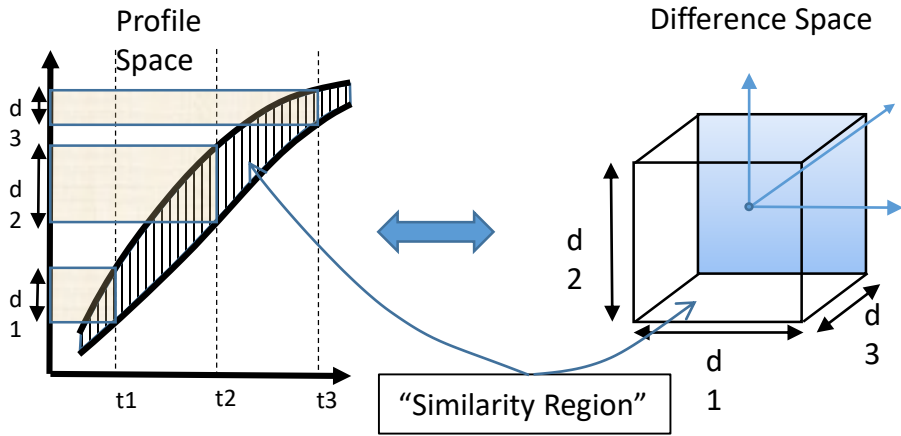
7. Traditional statistical approaches have difficulty with multivariate decision metrics because of the mismatch in shape between the estimator and the region of similarity.
8. Bayesian concepts and methodology provide a coherent path forward for dissolution similarity problem.
 - a. It can accommodate any arbitrary n-dimensional similarity region, most particularly a simple hyper-rectangle derived from safe- or process control space considerations.
 - b. Markov chain Monte-Carlo (MCMC) posterior simulation allows the similarity decision to be made by a simple counting exercise. The fraction of MCMC posterior dissolution differences that fall within the pre-defined similarity region approximates the posterior probability of similarity (PPS). If the estimated PPS is above some lower limit (e.g., 95%), similarity can be accepted.
 - c. The use of a probability metric (PPS) is consistent with risk-based decision making advocated by ICH Q9.
 - d. Bayesian approaches readily accommodates complex models (e.g., multiple batches, non-linear profile models) without complex analytical derivations or approximations.
 - e. The ability to leverage relevant and justifiable prior knowledge can potentially reduce required sample sizes and permit separation of analytical and process related sources of variation.

It is recognized that Bayesian concepts and software will be unfamiliar to many practitioners and decision makers. However, because Bayesian thinking provides a coherent pathway that preserves the connection to bio- and process control relevance and profile shape, it seems inevitable that Bayesian approaches will, in future, become important tools in dissolution similarity decision making. Therefore it is recommended that

- Regulatory guidance in the area of dissolution similarity acknowledge the applicability of Bayesian methodology in this field.
- Sponsors engaging in dissolution similarity comparisons consider collaborating with statisticians familiar with Bayesian methodology.

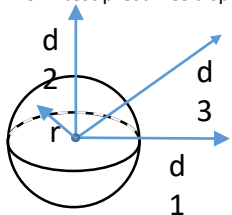
Additional Information provided at
the workshop by Dave LeBlond and Thomas Hoffelder

Thinking about the definition of similarity



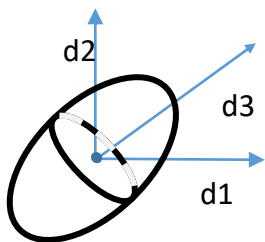
Key Points:

- The above rectangular “similarity region” may be helpful in visualizing a “safe space” (or even just a “control space” when BE is not available).
 - The difference space origin is the “target” profile.
 - All dissolution profiles contained within the profile space are “similar”
 - All dissolution profiles contained within the rectangular difference space are “similar”
- The profile space similarity region is easily visualized, but cannot always be used
- The difference space similarity region is difficult to visualize with >3 time points. Requires imagination.
- A “definition of similarity” and a “test for similarity” are not the same thing
- It is important to state whether the similarity region is based on observed or hypothetical profiles
- Similarity tests are often based on metrics (e.g., Euclidian or Mahalanobis distance) that can be challenging to visualize.
- Similarity tests presume some “similarity region” that can always be visualized in the difference space.
- The f_2 test presumes a spheroidal “similarity region” based on observed profiles:



- $d_1, d_2, d_3 = \text{TEST} - \text{REF}$ differences
- Origin $\rightarrow d_1=d_2=d_3=0$
- $r = \text{radius} = \sqrt{99 \times \text{number of time points}}$
- Note how limits at one time point depend on the differences at other time points
- More points in less deviant times can compensate for more deviant times (see next page)

- Some similarity tests (e.g., Mahalanobis distance) may presume an ellipsoid “similarity region”:



- $d_1, d_2, d_3 = \text{TEST} - \text{REF}$ differences
- Origin $\rightarrow d_1=d_2=d_3=0$
- Ellipse axes lengths/angles depend on the variance at each time point, the correlations among time points, and possibly other test “statistical constants”
- Note how limits at one time point depend on the differences at other time points
- Note that $f_2 \rightarrow \text{MD}$ changes the similarity region

10% difference as the similarity standard

- The similarity factor f_2 is a transformation of the quadratic mean (over time) of the differences between reference and test mean profile (QMD):

$$QMD := \sqrt{\frac{1}{n} \sum_{t=1}^n (R_t - T_t)^2} \Rightarrow f_2 = 50 \log_{10} \left(\frac{100}{\sqrt{1 + QMD^2}} \right)$$

- The acceptance criterion “ $f_2 > 50$ ” is identical to the criterion “ $QMD < 9.95 \approx 10$ ”. This means that dissolution profiles should be assessed as similar if the average difference between the profile means is below 10%.

Example:

$n=3$ time points. Differences between reference and test mean profiles:

- 11% at first time point
- 11% at second time point
- 3% at third time point

$$QMD = \sqrt{\frac{1}{3} (11^2 + 11^2 + 3^2)} = 9.15 \Rightarrow f_2 = 51.81$$

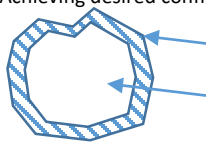
- Even though the difference between the profile means is above 10% in two of three time points, the QMD between the profile means is lower than 10%. This results in $f_2 > 50$ and in a decision in favor of profile similarity.
- The example shows that using f_2 , a difference above 10% at some time points can in some cases be compensated by means of a sufficiently low difference at the other time points.
- A regulatory recommendation of f_2 as standard approach for comparing dissolution profiles is connected with the focus on the average difference between the profile means.
- The acceptance criterion of some Mahalanobis distance based approaches as T2EQ or ACLMD can as well be interpreted as “average difference between the profile means < 10%”.
- Other approaches focus on the maximum, not on the average difference between the profile means (e.g. the TOST procedure separately applied at all dissolution time points with equivalence interval [-10% ; 10%]). The corresponding criterion “maximum difference < 10%” is more strict than the criterion “average difference < 10%”.

Thinking about statistical tests for similarity

Similarity test decision	Similar	Type I error	OK
	Not Similar	OK	Type II error
		Not Similar	Similar
Hypothetical state of nature			

Key Points:

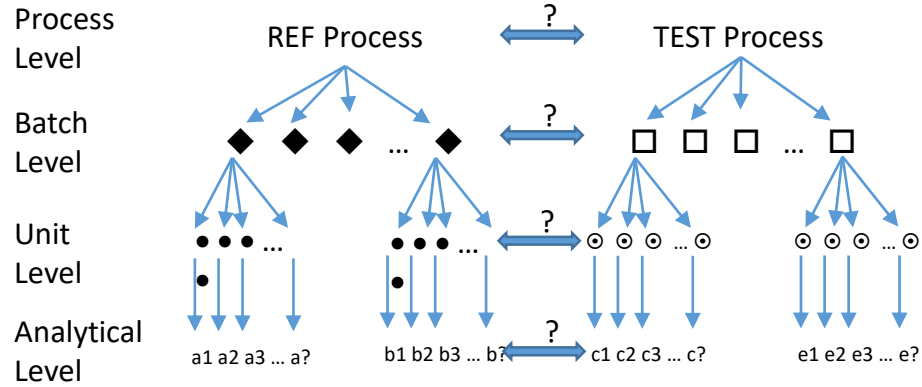
- We employ a “statistical test of equivalence”
 - Assumes “not similar” unless this is contradicted by evidence
 - Demands a greater burden of proof (i.e., more data) than a “statistical test of equality” → manufacturer benefits from higher sample size (→ higher power)
- The statistical % confidence level of the test = $100 \times (1 - \text{Type I error probability})$
 - The confidence level of the test should ideally be known prior to performing the test
 - A customary confidence level is 95%
 - Be aware... the confidence level IS NOT the “probability of similarity”
- The statistical % power of the test = $100 \times (1 - \text{Type II error probability})$
 - A customary power level is 80% (e.g., clinical trials)
- Achieving desired confidence and power levels effectively “shrinks” the similarity region


 - Set of hypothetical differences considered similar
 - Set of hypothetical differences for which similarity can be concluded
 - The amount of shrinkage drops as data variance ↓ and sample size ↑
- The traditional f_2 is not a statistical test
 - The spheroidal region of similarity is based on observed rather than hypothetical profiles
 - For borderline profiles, % confidence and % power $\cong 50\%$ (coin flip) regardless of variance or sample size.
- Some tests are “exact” (confidence level is known from theory), others are only approximate
- Only Bayesian tests can provide an estimate for the “probability of similarity”. But they do not claim to provide a pre-defined confidence or power level.
- The confidence and power of approximate and Bayesian tests are best determined by computer simulation
- Traditional similarity tests focus on profile shape difference, not differences in variability.

Some of the more common similarity tests

Test	Description	Decision Metric	Difference space Similarity region
f_2	Most common. Simple but not a statistical test. Regulatory provisions about RSD and number of time points beyond 15-85%	Euclidian	Spherical
f_2 bootstrap (Shah)	Commonly used when f_2 disallowed. Approximate statistical equivalence test. Requires corrections due to bias.	Euclidian	Spherical
T^2 (Wellek)	Exact (assuming multivariate normality) statistical equivalence test	Mahalanobis	Ellipsoid
ACLMD (aka MSD) (Tsong,1996)	Ellipsoid confidence region must fit within similarity region. Conservative.	Mahalanobis	Cuboid (± 10 to 15%)
T2EQ (Hoffelder)	Approximate equivalence test.	Mahalanobis	Ellipsoid
Bayesian F_2	Estimates posterior probability of similarity	Euclidian	Spherical (may also include cuboid)
Bayesian General	Estimates posterior probability of similarity	Euclidian or Mahalanobis	Arbitrary
SK (Saranadasa & Krishnamoorthy)	Assumes parallel profiles. Requires < 10% difference	Euclidian	Cuboid ($\pm 10\%$)
MD bootstrap (BCA δ^2)	Approximate (BCA correction) statistical equivalence test. Some computational challenges.	Mahalanobis	Ellipsoid
IUT (Berger&Hsu)	Multiple exact equivalence (TOST) tests. Very conservative.	Euclidian	Rectangular
g_2 (Shen)	Based on the average absolute deviations. Less conservative than f_2	Euclidian	Hyper-diamond
Weibull profile model	Decision based on either Euclidian or profile model parameter space. Dimensional reduction & Interpolation of points feasible.	Varies	Varies

Thinking about the “multiple batch” situation



Key points:

- Processes that produce dosage units are hierarchical
 - Each hierarchical level contributes its own sources of bias and variability
- At which hierarchical level do we wish to test for similarity?
 - This is an important part of choosing the “inference space” for the similarity test
 - Future batches and dosage units will be made by the TEST process
 - Making inferences at the analytical level confounds the manufacturing and analytical processes
- To make similarity claims specifically about the TEST manufacturing process, we must ...
 - Perform testing on units from a representative sample of TEST and REF batches
 - Challenging to randomly sample batches (e.g., campaigns)
 - Separately model both inter- and intra-batch variation
 - Modern statistical software can do this (mixed/hierarchical modeling)
 - Allows prediction of dissolution similarity for future batches and/or tablets
 - Consider prior knowledge about analytical bias/precision to exclude its contribution
 - Bayesian modeling offers a possible solution
- Pairwise comparisons creates difficulties
 - Results may be contradictory
 - Requiring all pairs to pass raises the overall Type II error, reducing power
 - Hierarchical models avoid these difficulties, but require a sufficient number of batches
 - Computer simulation can help estimate the required number of batches, given...
 - A similarity definition
 - An appropriate hierarchical model
 - Estimates of model parameter values
 - The specific statistical test to be used for decision making
 - The required confidence and power