In Vitro Dissolution Curve Comparisons: A Critique of Current Practice and a Proposed Bayesian Test Statistic

Stan Altan (Janssen), Dave LeBlond (Consultant), John Peterson (GSK), Yan Shen (Janssen), Harry Yang (MedImmune), Steve Novick (Mediimune)

Based on a paper published in *J. Biopharm. Stat.* **2015**, *25* (2), 351–371. "Dissolution Curve Comparisons Through the *F*2 parameter, a Bayesian Extension of the *f*2 Statistic".
Winner of the 2015 Best Nonclinical Paper Award, NCB2015 Conference Villanova, PA, October 15, 2015

Outline

- Major criticisms and concerns with f2 statistic
- Scientific vs Regulatory perspective
- Moving beyond the f2 and published multivariate approaches
- A statistically rigorous framework
  - Two examples
- Simulation study
- Nonlinear approach
- Summary

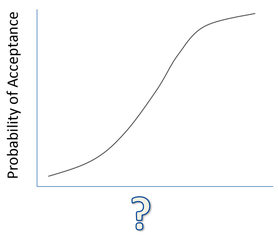
Criticisms/Questions of f2

1. What population characteristic is f2 estimating?

$$f_2 = 50\log\left(\frac{100}{\sqrt{1+\frac{D^2}{p}}}\right) > 50 \qquad D = \sqrt{\sum_{i=1}^{p}\left(T_i - R_i\right)^2}$$

- T_i and R_i are observed average dissolution of 12 units for Test and Reference at time point $i = 1, \ldots, p$
- No Guidance on underlying statistical model
- **Similarity is not defined as a function of parameters associated with the materials being compared**
- Allows a decision, but how does that decision relate to similarity?
- f2 is a biased (conservative) estimator of the corresponding population metric with $D = \sqrt{\sum_{i=1}^{p}\left(\mu_{Ti} - \mu_{Ri}\right)^2}$

2. Can an OC curve be defined when similarity is not defined in terms of model parameters?



- Probability of acceptance will depend on the measurement uncertainty which will impact decision risks.
- However, the "X axis" should not include parameters associated with measurement uncertainty.

3. Can an equivalence testing framework be possible if similarity is not defined as a function of parameters in a model of the process/ materials being compared?

- No associated "confidence level"
- If the median of the sampling distribution of f2 is 50, the Type I error = 50%
- What evidence of similarity does f2 provide?
- Bootstrapping investigated but coverage not nominal … not pursued

Other concerns

- No "standard" experimental design.
- f2 criteria becomes more liberal as the number of time points increases. Larger deviations can be accommodated.
- Test and Reference must have same time points.
- No inference about the processes.
- Variance heterogeneity not acknowledged.
- Complex sampling distribution for f2.
- f2 is a function of both material variability and analytical variability

Scientific vs. Regulatory Perspective

Fundamental difference between the scientific and regulatory perspectives.

- Scientific perspective: What is the probability that this particular change is unsafe or ineffective?
- Regulatory perspective: What is the probability (over many submissions) that we will approve a change that is unsafe or ineffective?

Statistical perspective

- Encourage use of informative decision making tools
- Statisticians calibrate these tools to understand how a metric switch impacts existing approvals. Will it raise or lower the bar, will it impact regulatory risk management? Walk the line between failure to block a bad change and failure to approve a good one.
- To improve entrenched methodology, statisticians need to wear 2 hats, make win-win arguments, and show a new tool is more informative with predictable, understandable, and consistent performance across products.

Moving beyond f2 and MV

A good alternative method to f2/MV would include:

- A model based definition of "dissolution similarity"
  similarity metric should be defined in terms of parameters of the model that describes material properties, not data, and not parameters of the model of analytical measurement.
- Independent of statistical methodology used
- Clarification of the proper inference space (conclusions apply only to lots in hand or to future lots,…?)
- Proper modeling of lot to lot variance
- Consideration of "confidence level"
- Computer simulations to address the operating characteristics
- Strong experimental design recommendations
- Recommendations on how to implement the new approach (even when statistical support is lacking)
- Same general approach regardless of %CV
  - Avoid culture shock

Statistically rigorous testing

- $F_2 = 50\log_{10}\left(\frac{100}{\sqrt{1+\Delta^2}}\right)$, $\Delta^2 = \left(\frac{1}{p}\right)\sum_i^p\left(\mu_{Ref,t_i} - \mu_{Test,t_i}\right)^2$
- H_0: $F_2 \leq 50$ vs H_a: $F_2 > 50$
- Declare equivalence if Pr($F_2 > 50$ | data) = Pr($\Delta^2 < 100$ | data) ≥ 0.95
- Let $\delta(p) = \max_{t=1,\ldots,p}\left|\mu_{Ref,t} - \mu_{Test,t}\right|$

Should test individual mean differences too:
$H_0: \delta(p) \geq 15$ OR $F_2 \leq 50$
$H_a: \delta(p) < 15$ AND $F_2 \leq 50$

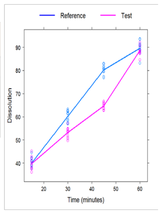
Examples/Simulation Study

Example 1: Large difference @ Time = 45 min

Means of 12 vessels

Time (minutes)	Reference	Test	Difference
15	40	40	0
30	60	52	8
45	80	64	16
60	90	90	0

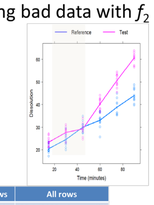
f_2	53.1 (PASS)
Pr($F_2 > 50$)	1.00 (PASS)
Pr($F_2 > 50$, $\delta(p) < 15$)	0.2 (FAIL)



Example 2: Rescuing bad data with f_2

Means of 12 vessels

Time	Ref	Test	Diff
15	20	22	-2
30	25	27	-2
45	30	30	0
60	35	40.5	-5.5
75	40	51	-11
90	45	61.5	-16.5



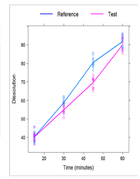
	Last 4 rows	All rows
f_2	48.2 (FAIL)	52.2 (PASS = Rescued!)
Pr($F_2 > 50$)	0.12	0.92 (Almost rescued!)
Pr($F_2 > 50$, $\delta(p) < 15$)	0.06	0.07 (FAIL)

Simulation Study

Means of 12 vessels

Time	Ref	Test	Diff
15	40	40	0
30	60	60 - η	η
45	80	80 - 2η	2η
60	90	90	0

Scenarios:
η = 0, 5, 6, 7.5, 9, 10
σ = 4



Example data set with η=5

Summary of 5,000 Monte Carlo runs

η	True F_2	True δ(p)	$f_2 > 50$	Pr($F_2 > 50$) ≥ 0.95	Pr($F_2 > 50$, δ(p) < 15) ≥ 0.95
0	100	0	1.00	1.00	1.00
5	62	10	1.00	1.00	0.88
6	58	12	1.00	0.96	0.49
7.5	54	15	0.96	0.39	0.03
9	50	18	0.45	0.02	0.00
10	47.5	20	0.05	0.00	0.00

SUMMARY/Future Research

- f2 has become entrenched as a similarity metric and is unlikely to be displaced.
- f2 and Multivariate approaches, as currently mandated, have statistical issues.
- Bayesian paradigms and methodology have the potential of overcoming many of the issues with f2 and MV, while maintaining a link to the established metric/criterion.
- **Statisticians can add value to the discourse by** :
  - ***taking the lead in communicating these issues***
  - ***identifying opportunities to improve decision making***
  - ***wearing both scientific and regulatory hats***
  - ***working toward "win-win" solutions.***
- Modeling with continuous nonlinear function opens up new possibilities
- Sharper focus on connections between hierarchical modeling and claims of equivalence