



Evaluations of Parallelism Test Methods Using ROC Analysis

Harry Yang and Lanju Zhang
MedImmune

2009 Non-clinical Biostatistics Conference
Boston, MA

- Motivations
- Parallelism testing in bioassay
- Available methods
- Evaluation based on ROC analysis
- Conclusions

■ Parallelism testing

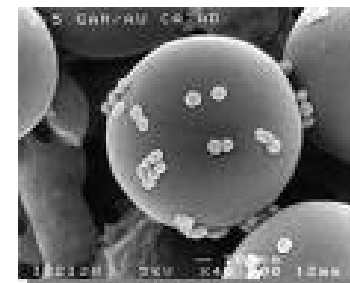
- ◆ Testing similarity between reference standard and test sample dose-response curves
- ◆ Key requirement for bioassay (Revised USP Chapter <111>)
- ◆ Several methods available
- ◆ Method comparison either biased or flawed
- ◆ No consensus on best test methods in literature

■ ROC analysis

- ◆ A common framework for objective evaluation of overall performance
- ◆ Determination of optimal cut points or equivalence bounds

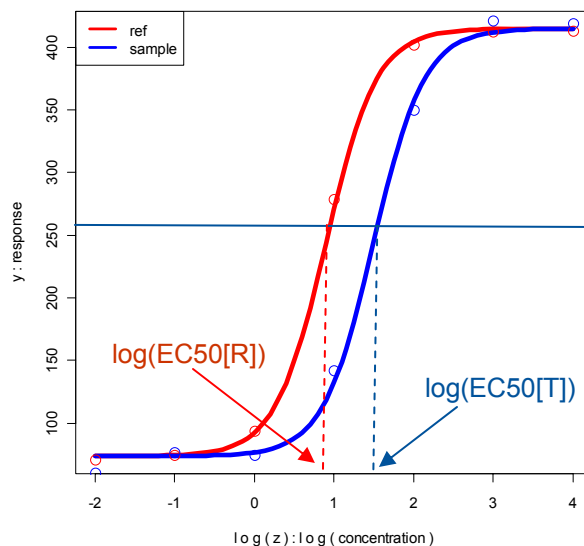
Potency Bioassay

- Measurement of effectiveness of a compound by its effect on animals or cells in comparison with a standard preparation (USP Chapter <1046>)



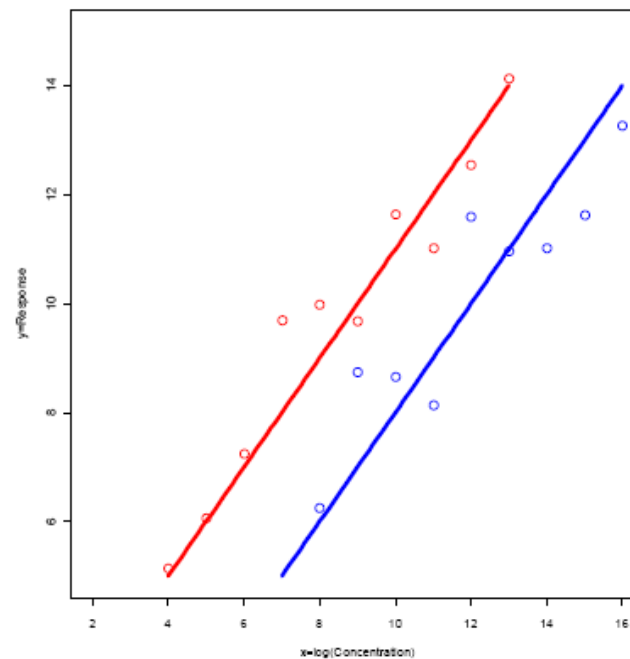
Potency of Bioassay

- Potency often determined relative to a reference standard such as ratio of EC_{50}
- Only meaningful if test sample (T) behaves as a dilution or concentration of reference standard (R)
 - $EC_n[T] = \rho \times EC_n[R]$ or $\log(EC_n[T]) = \log(\rho) + \log(EC_n[R])$, with $n = 25, 50, 75$ etc. and $\rho =$ relative potency



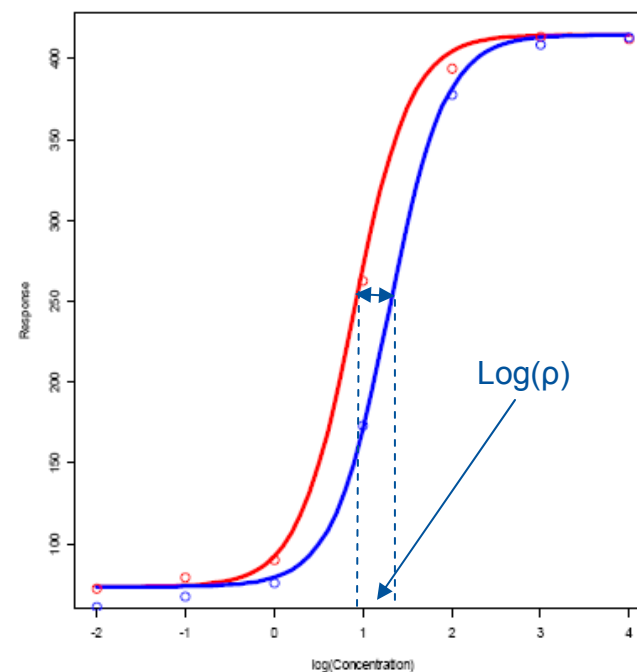
Parallelism Testing

- A procedure by which similarity between dose-response curves of test sample and reference standard is evaluated



Parallelism Testing (Cont'd)

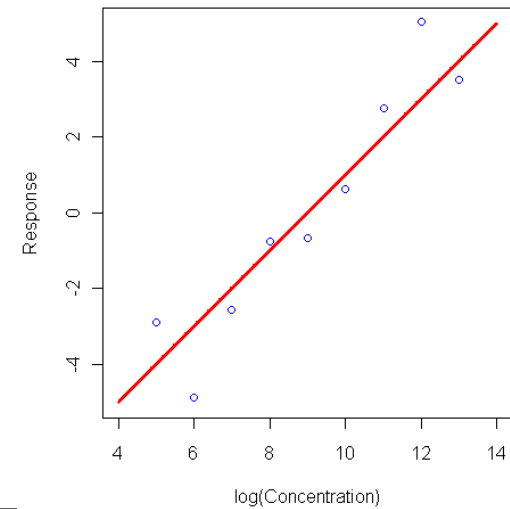
- Mathematically, parallelism implies $F_T(z) = F_R(\rho z)$ where z is concentration and ρ is the **relative potency**



Modeling Dose-response Curves

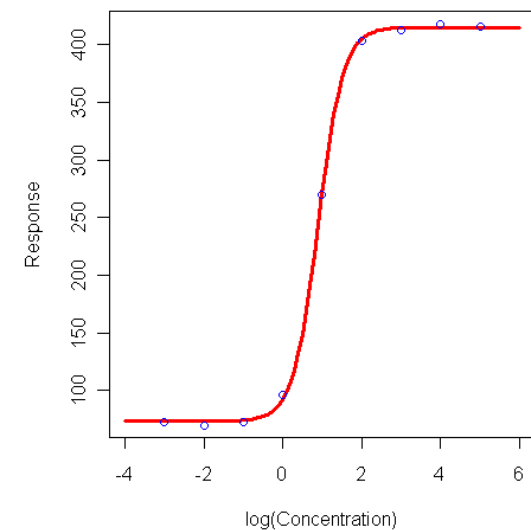
- Models for the function F

- Linear: $y = a + bz$



- Non-linear: $y = D + \frac{A - D}{1 + \exp(Bz + \log(C))}$

- D: upper asymptote
- A: lower asymptote
- C: EC_{50}
- B: slope parameter



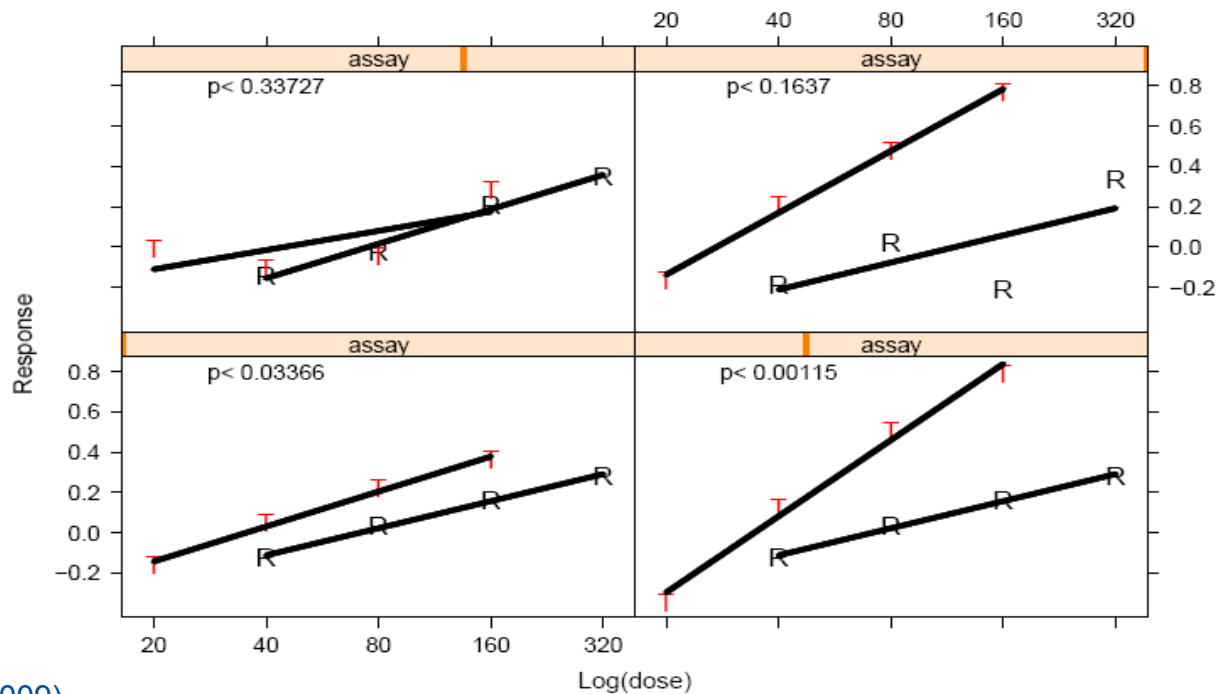
- General idea
 - A metric for non-parallelism
 - A cut off value
- Linear case: testing parallelism is the same as testing equal slopes
 - Test $H_0: \beta_T = \beta_R$ vs. $H_1: \beta_T \neq \beta_R$

$$\text{Test statistic } Z^* = \frac{\hat{\beta}_T - \hat{\beta}_R}{\sqrt{\text{var}[\hat{\beta}_T - \hat{\beta}_R]}} \sim N(0,1) \longleftarrow \text{Metric for non-parallelism}$$

Reject H_0 if $|Z^*| > z_{1-\alpha} \longleftarrow \text{Cut off value}$

- Parameter comparison
 - Compare parameters of models used to fit TS and RS data
 - Significance test vs. equivalence test (Hauck et al, 2005; Jonkman and Sidik, 2009)
- Response comparison
 - Compare fitted values between reduced model ($F_T(z) = F_R(\rho z)$) and free model
 - Chi-square test vs. F-test (Gottschalk and Dunn, 2005; revised USP Chapter <111>)

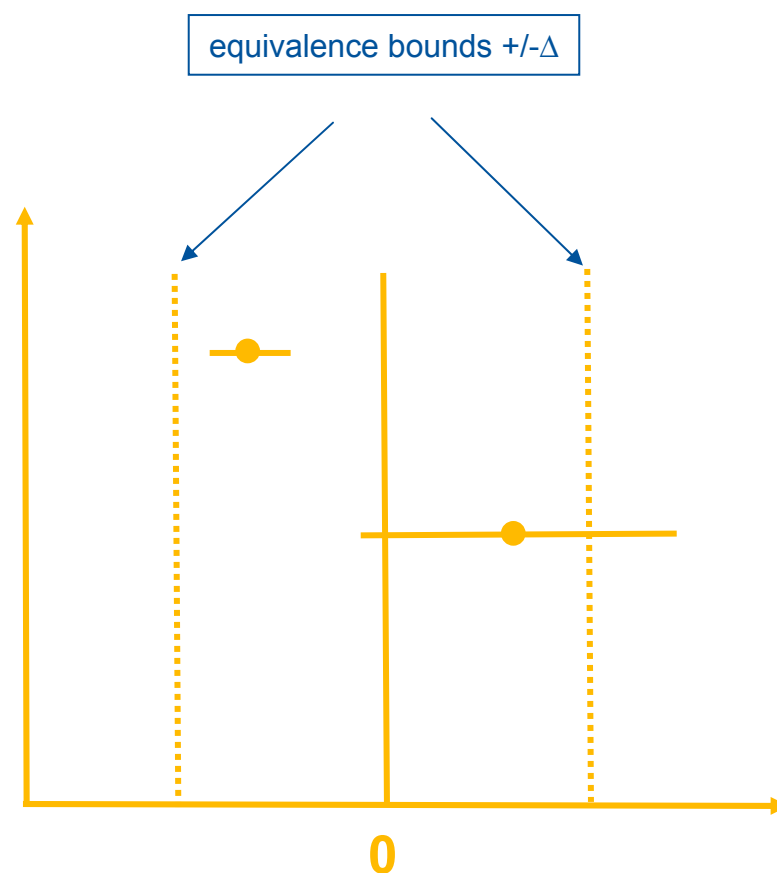
- Significance test:
 - Traditional approach for testing difference $H_0: \beta_T = \beta_R$ vs. $H_1: \beta_T \neq \beta_R$
 - Non-parallel when $p\text{-value} < 0.05$
- Penalize precise assays



* David Lansky (2009)

Equivalence Approach

- Equivalence test (*Hauck et al, 2005; Lansky, 2009; Draft USP Ch. <111>, OCT 2006*)
 - $H_0: |\beta_T - \beta_R| \geq \Delta$ vs. $H_1: |\beta_T - \beta_R| < \Delta$
 - Parallel when 90% confidence interval falls within equivalence bounds
 - Equivalent to two one-sided t-tests
 - Claim to reward precise assays



- Generalization of Hauck's equivalence test from linear case to non-linear response (Jonkman and Sidik, 2009)

$$Y = a + \frac{d - a}{1 + \exp[b(c - \log X)]} + \epsilon$$

- Equivalence test involving comparisons of lower and upper asymptotes and slopes between reference standard and test sample

$$a_1 - D_L a_2 \leq 0 \quad \text{or} \quad a_1 - D_U a_2 \geq 0 \quad \text{or}$$

$$H_0 : b_1 - D_L b_2 \leq 0 \quad \text{or} \quad b_1 - D_U b_2 \geq 0 \quad \text{or}$$

$$d_1 - D_L d_2 \leq 0 \quad \text{or} \quad d_1 - D_U d_2 \geq 0$$

$$a_1 - D_L a_2 > 0 \quad \text{and} \quad a_1 - D_U a_2 < 0 \quad \text{and}$$

$$H_1 : b_1 - D_L b_2 > 0 \quad \text{and} \quad b_1 - D_U b_2 < 0 \quad \text{and}$$

$$d_1 - D_L d_2 > 0 \quad \text{and} \quad d_1 - D_U d_2 < 0.$$

- Chi-square test (Gotschalk and Dunn, 2005)

- Based on extra-sum-of-squares

$$RSSE_{\chi^2} = SSE(\text{Reduced}) - SSE(\text{Full}) \sim \chi_{df_1}^2$$

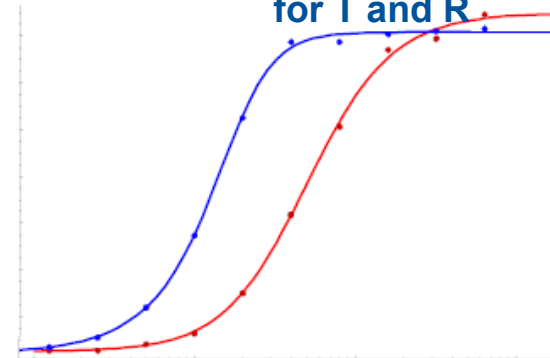
- Two curves are parallel if

$$RSSE_{\chi^2} > \chi_{df_1}^2 (1 - \alpha)$$

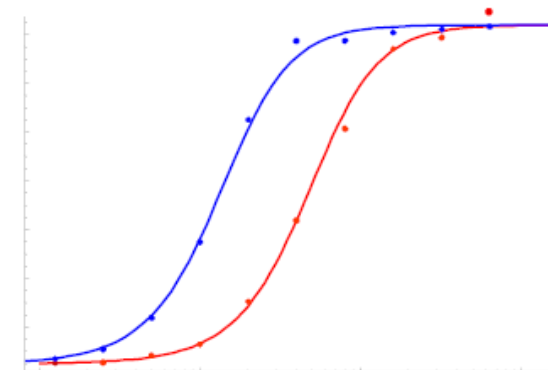
- Claim not to have the shortcomings of F-test

$$RSSE_F = \frac{[SSE(\text{Reduced}) - SSE(\text{Full})] / df_1}{SSE(\text{Full}) / df_2} \sim F_{df_1, df_2}$$

Full model: Fit separate models for T and R

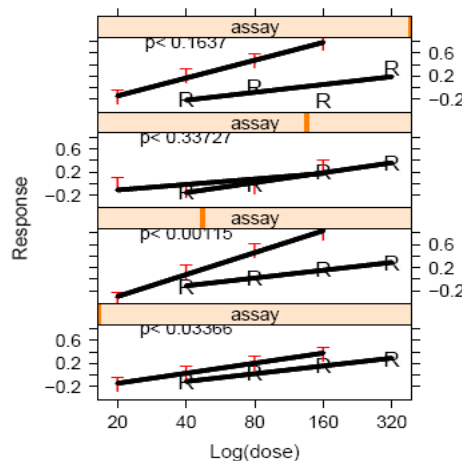
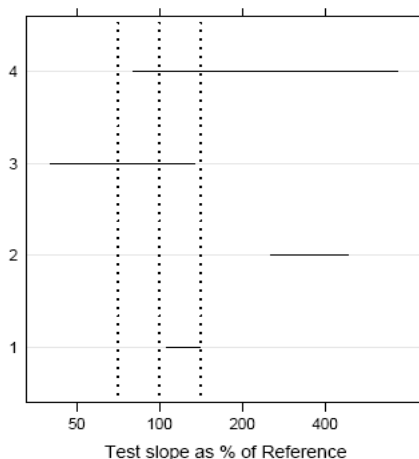


Reduced model: Fit models for T and R under constraint $F_T(z) = F_R(\rho z)$



- Significance test in USP Ch. <111> is flawed
 - ◆ Fail assays with good precision and accept faulty assays with poor precision

- Equivalence testing paradigm does not have the aforesaid shortcomings



Are two lines parallel?

| <u>P-value</u> | <u>CI</u> | <u>Assay</u> |
|----------------|-----------|--------------|
| Yes | No | Variable |
| Yes | No | Variable |
| No | No | Precise |
| No | Yes | Precise |

* David Lansky (2009)

Comparison between Apple and Orange

■ Significance test

- ◆ Test H0: $\beta_T = \beta_R$ vs. H1: $\beta_T \neq \beta_R$
- ◆ Control Type I error = Prob[Non-parallel | Parallel]

■ Equivalence test

- ◆ Test H0: $|\beta_T - \beta_R| > \Delta$ vs. H1: $|\beta_T - \beta_R| \leq \Delta$
- ◆ Control Type I error = Prob[Parallel | Non-parallel]

■ Not necessarily true that significant test fails assays with good precision and accepts faulty assays with poor precision

■ Unfair comparison

- ◆ Fixing cut point for significant test at $p=0.05$ for all assays
- ◆ Allowing equivalence test to choose acceptance criteria for each assay

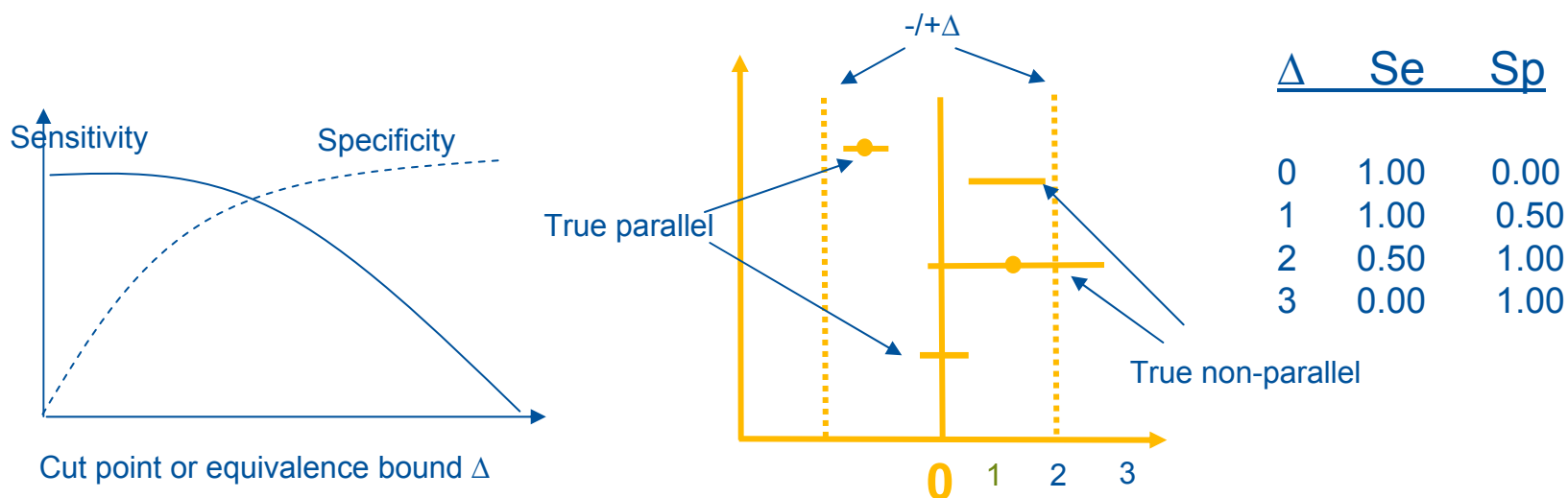
Results of A Mixed Bag

- Simulation results (Jonkman and Sidik, 2009) do not support IUT (equivalent test) is superior to F (significant test)

Table 2 Simulation results: Each value is the proportion of times out of 10,000 replicates that the specified test resulted in a declaration of parallel response curves (see Section 5)

| σ | Doses | Replicates | Case 1: Approximately parallel | | Case 2: Exactly parallel | | Case 3: Boundary Values | | Case 4: Unequal lower plateaus | | Case 5: Unequal slopes | |
|----------|-------|------------|--------------------------------------|--------|--------------------------------|--------|-------------------------------|-----|--------------------------------------|--------|------------------------------|--------|
| | | | F test | IUT | F test | IUT | F test | IUT | F test | IUT | F test | IUT |
| 0.15 | 8 | 3 | 0.2848 | 0.5681 | 0.9479 | 0.6578 | 0.0001 | 0 | 0.0006 | 0.0008 | 0 | 0.0108 |
| | 8 | 4 | 0.1464 | 0.7444 | 0.9502 | 0.8334 | 0 | 0 | 0 | 0.0005 | 0 | 0.0135 |
| | 10 | 3 | 0.0307 | 0.8806 | 0.9492 | 0.9444 | 0 | 0 | 0.0002 | 0.0022 | 0 | 0.0102 |
| | 10 | 4 | 0.0050 | 0.9559 | 0.9493 | 0.9898 | 0 | 0 | 0 | 0.0007 | 0 | 0.0090 |
| | 12 | 3 | 0.0033 | 0.9376 | 0.9507 | 0.9739 | 0 | 0 | 0.0002 | 0.0039 | 0 | 0.0079 |
| | 12 | 4 | 0.0001 | 0.9837 | 0.9491 | 0.9967 | 0 | 0 | 0 | 0.0020 | 0 | 0.0068 |
| 0.20 | 8 | 3 | 0.5474 | 0.1824 | 0.9451 | 0.2260 | 0.0093 | 0 | 0.0281 | 0.0003 | 0.0034 | 0.0058 |
| | 8 | 4 | 0.4189 | 0.3852 | 0.9500 | 0.4558 | 0.0010 | 0 | 0.0036 | 0.0004 | 0 | 0.0060 |
| | 10 | 3 | 0.2112 | 0.5692 | 0.9519 | 0.6609 | 0 | 0 | 0.0266 | 0.0019 | 0.0009 | 0.0052 |
| | 10 | 4 | 0.0925 | 0.7546 | 0.9505 | 0.8516 | 0 | 0 | 0.0035 | 0.0026 | 0 | 0.0082 |
| | 12 | 3 | 0.0658 | 0.6820 | 0.9519 | 0.7696 | 0 | 0 | 0.0222 | 0.0041 | 0.0003 | 0.0052 |
| | 12 | 4 | 0.0148 | 0.8483 | 0.9502 | 0.9091 | 0 | 0 | 0.0031 | 0.0032 | 0.0001 | 0.0060 |
| 0.30 | 8 | 3 | 0.7824 | 0.0009 | 0.9517 | 0.0017 | 0.2029 | 0 | 0.3110 | 0 | 0.1314 | 0 |
| | 8 | 4 | 0.7115 | 0.0102 | 0.9504 | 0.0136 | 0.0888 | 0 | 0.1574 | 0.0001 | 0.0439 | 0.0003 |
| | 10 | 3 | 0.5817 | 0.0641 | 0.9477 | 0.0812 | 0 | 0 | 0.2881 | 0.0004 | 0.0807 | 0.0015 |
| | 10 | 4 | 0.4547 | 0.2042 | 0.9488 | 0.2470 | 0 | 0 | 0.1564 | 0.0005 | 0.0235 | 0.0018 |
| | 12 | 3 | 0.4114 | 0.1346 | 0.9486 | 0.1583 | 0 | 0 | 0.2856 | 0.0011 | 0.0746 | 0.0015 |
| | 12 | 4 | 0.2623 | 0.3071 | 0.9520 | 0.3763 | 0 | 0 | 0.1477 | 0.0016 | 0.0213 | 0.0030 |

- Sensitivity (Se) and Specificity (Sp)
 - ◆ $Se = \Pr[\text{Test non-parallel} \mid \text{Non-parallel}]$
 - ◆ $Sp = \Pr[\text{Test parallel} \mid \text{Parallel}]$
- Need to be considered when comparing parallelism tests
- Dependent on the choice of cut point of test statistics, p-value or equivalence bound
- Higher sensitivity results in lower specificity and vice versa



Comparisons between F and χ^2 Tests by Gotschalk and Dunn

- F-test does not reflect true parallelism very accurately whenever the fit of the full (free) model is either quite good or quite bad

RSSE Nonpar

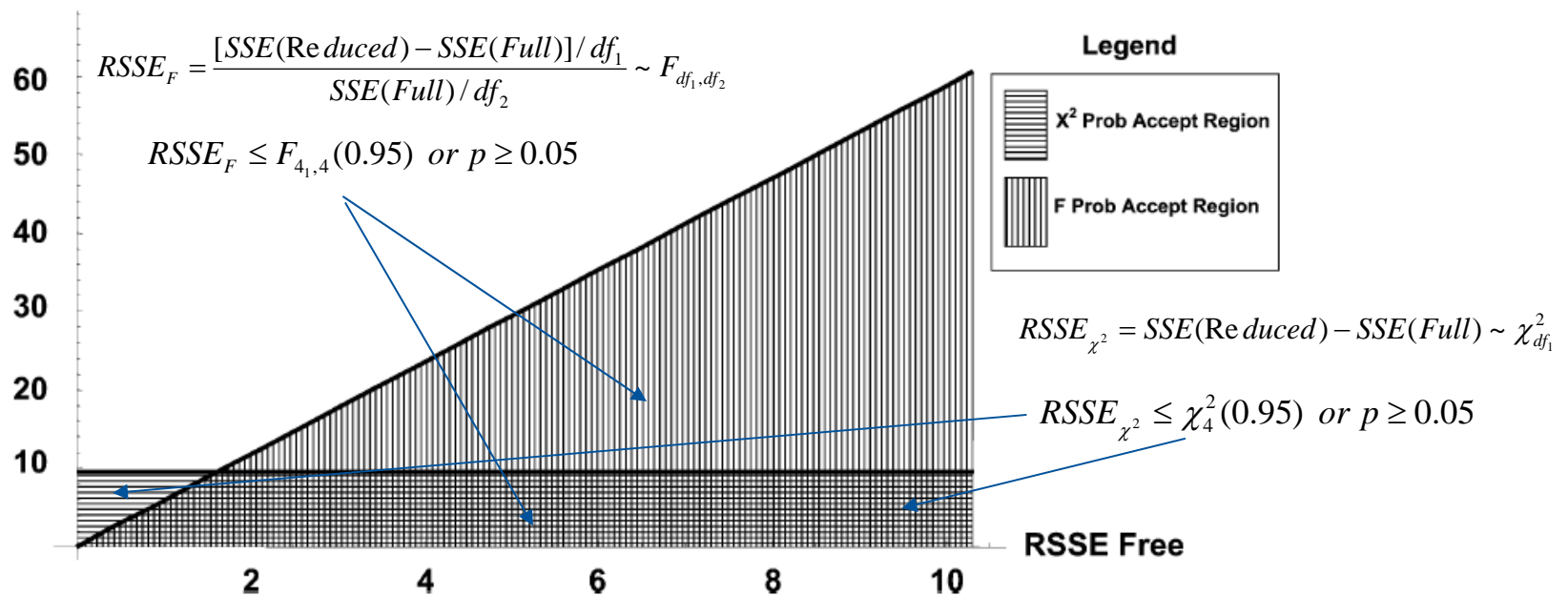


Figure 1 Regions of parallelism for the X² Prob and F Prob when the assay is parallel if the p value of the metric is greater than .05, with $df_{nonpar} = 4$ and $df_{free} = 4$.

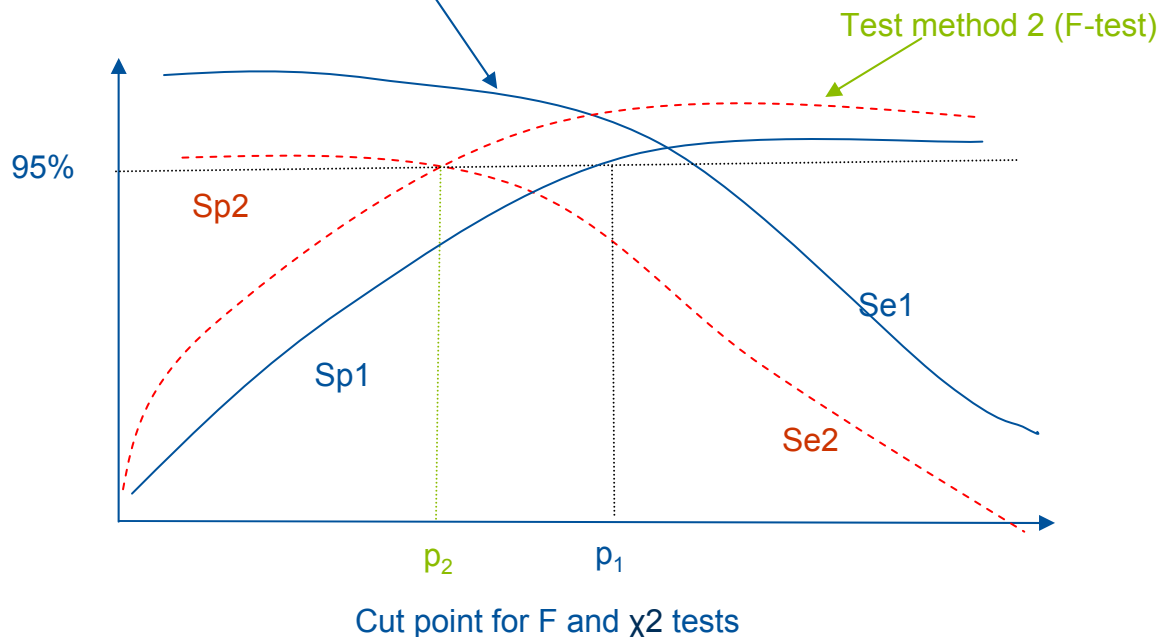
Gotschalk and Dunn, 2005

Bias in method Comparison by Gotschalk and Dunn

- Biased comparison because it is centered solely on sensitivity (=Prob[Test non-parallel | Non-parallel]), after fixing specificity (=Prob[Test parallel | Parallel])

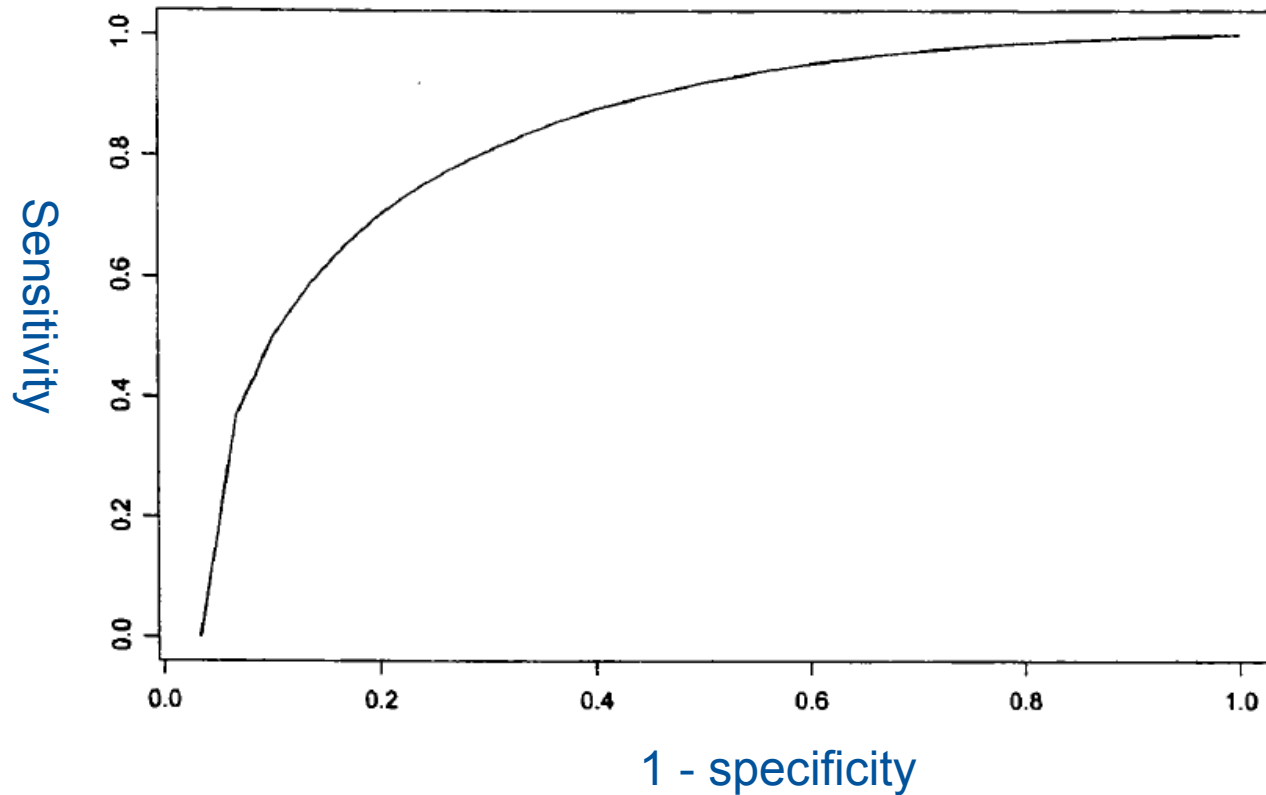
Cut points p_1 and p_2 render Test Method 1 and Test Method 2, 95% specificity, but much higher sensitivity for Test Method 1 than Test Method 2.

Test method 1 (χ^2 test)



Receiver Operating Characteristic (ROC) Curve

- A plot of sensitivity vs. 1 – specificity for various choices of cut points



Method Comparison Using Area under ROC Curve (AUC)

- AUC is the probability of correctly ranking test statistic X of a pair of parallel curves and test statistic Y of a pair of non-parallel curves
 - ◆ $AUC = \Pr[X < Y]$
- Measures overall discriminatory power of a test
 - ◆ The larger the AUC, the more accurate the test
- Can be readily calculated using parametric or non-parametric methods
- Is a useful tool for comparing parallelism test methods

- Assume dose-response curve is linear

$$F_R(z; \alpha, \beta) = \alpha + \beta \log(z)$$

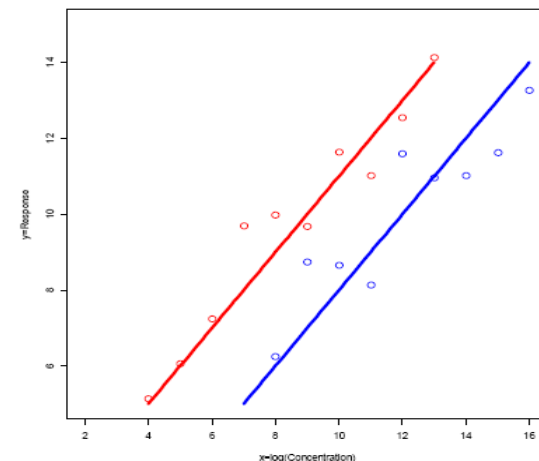
Parallelism implies

$$F_T(z) = F_R(\rho z) \Leftrightarrow \alpha_T + \beta_T \log(z) = \alpha_R + \beta_R \log(\rho z) = [\alpha_R + \beta_R \log(\rho)] + \beta_R z$$

That is

$$\alpha_T = \alpha_R + \beta_R \log(\rho), \quad \beta_T = \beta_R.$$

- Parallelism testing is equivalent to testing same slope for two dose-response lines



Step 1. (Simulate parallel curves) 1000 pairs of parallel curves were simulated using the model $y_{ij} = \alpha_i + \beta_i x + \varepsilon_{ij}$ where $i = 1, 2, j = 1, \dots, 1000$, and $\varepsilon_{ij} \sim N(0, \sigma^2)$ with $\alpha_1 = 1, \alpha_2 = 0.5, \beta_1 = \beta_2 = 1, \sigma = 0.2, 0.3$ and 0.4 , respectively;

Step 2. (Simulate non-parallel curves) 1000 pairs of non-parallel curves were simulated using the model $y_{ij} = \alpha_i + \beta_i x + \varepsilon_{ij}$ where $i = 1, 2, j = 1, \dots, 1000$, and $\varepsilon_{ij} \sim N(0, \sigma^2)$ with $\alpha_1 = 1, \alpha_2 = 0.5, \beta_1 = 1, \beta_2 = 1.5, \sigma = 0.2, 0.3$ and 0.4 , respectively;

Step 3. (Estimate sensitivity and specificity) Perform regression analysis for each pair of curves generated in steps 1 and 2, and construct F, χ^2 statistics and 90% confidence interval based on a given positive number Δ . Vary cut point c , and Δ , and calculate sensitivity (Se) or specificity (Sp) using the following formula:

$$Se = \frac{\# \text{ Pairs from Step 2 with test statistic } > c \text{ or } 90\%CI \notin \pm \Delta}{1000}$$

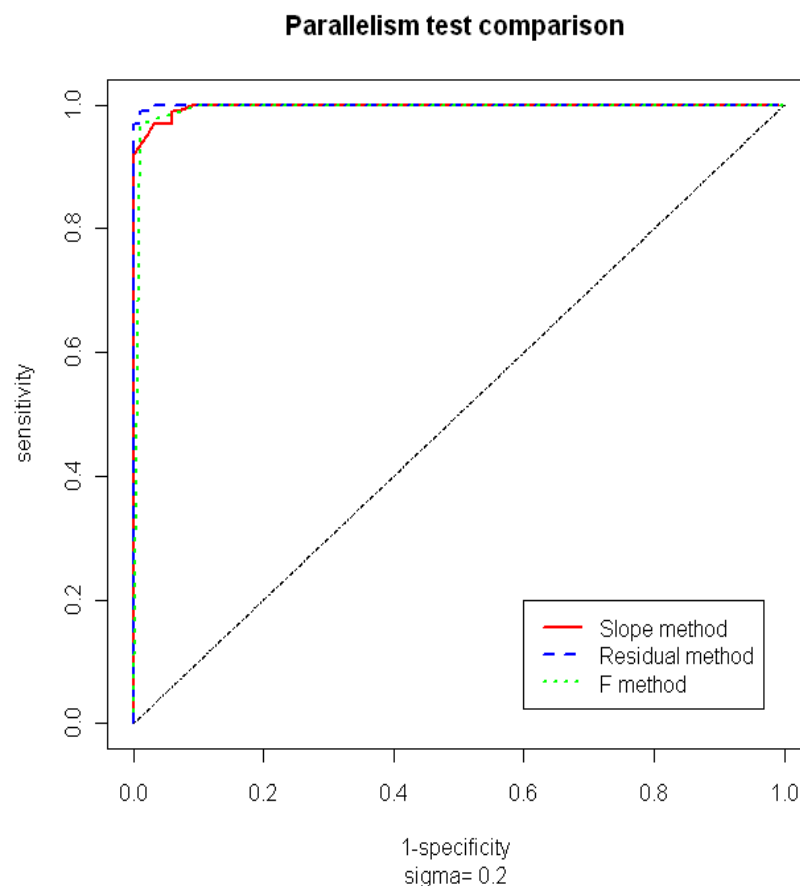
$$Sp = \frac{\# \text{ Pairs from Step 1 with test statistic } \leq c \text{ or } 90\%CI \in \pm \Delta}{1000}$$

Step 4. (Plot ROC curves and calculate AUCs) Plot Se and $1 - Sp$ obtained from *Step 3*, and calculate the AUC under the ROC curve;

Step 5. (Estimate mean AUC and its SD) Repeat *Steps 1-4* for n times, and calculate the mean AUC and its SD.

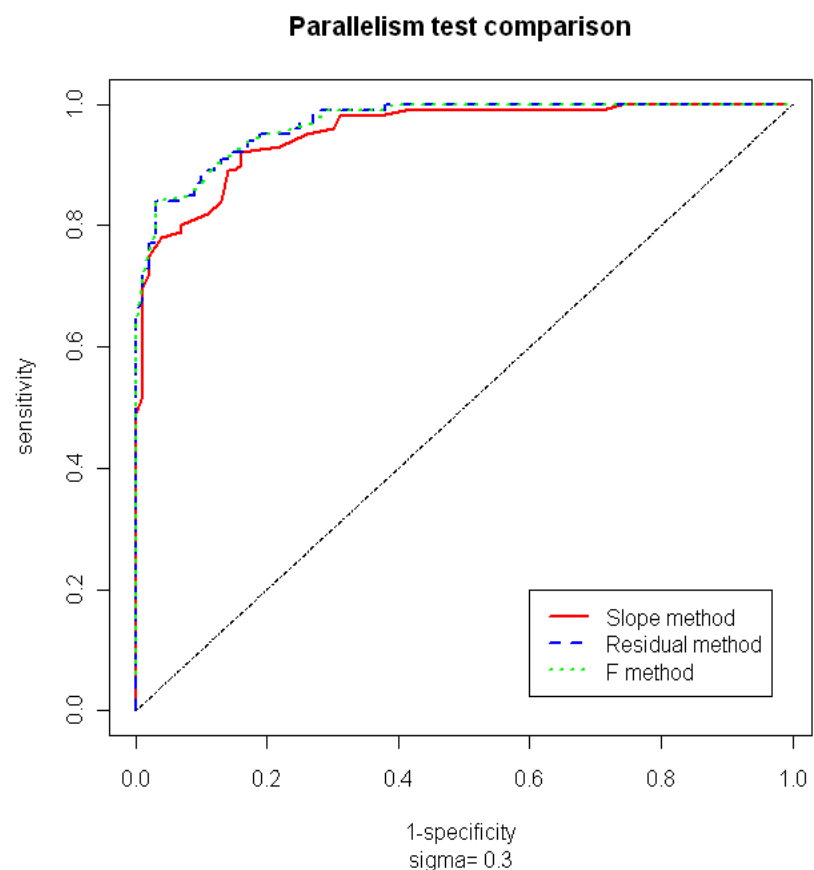
Results of Simulation Study

- Assay variability = 0.2
- Similar performance of F, χ^2 and equivalence methods
 - ◆ $AUC_F = 0.9608$ (0.00096)
 - ◆ $AUC_{\chi^2} = 0.9974$ (0.0005)
 - ◆ $AUC_{\text{Equiv}} = 0.9945$ (0.0433)
- χ^2 and F-tests ranked the best and worst, respectively.



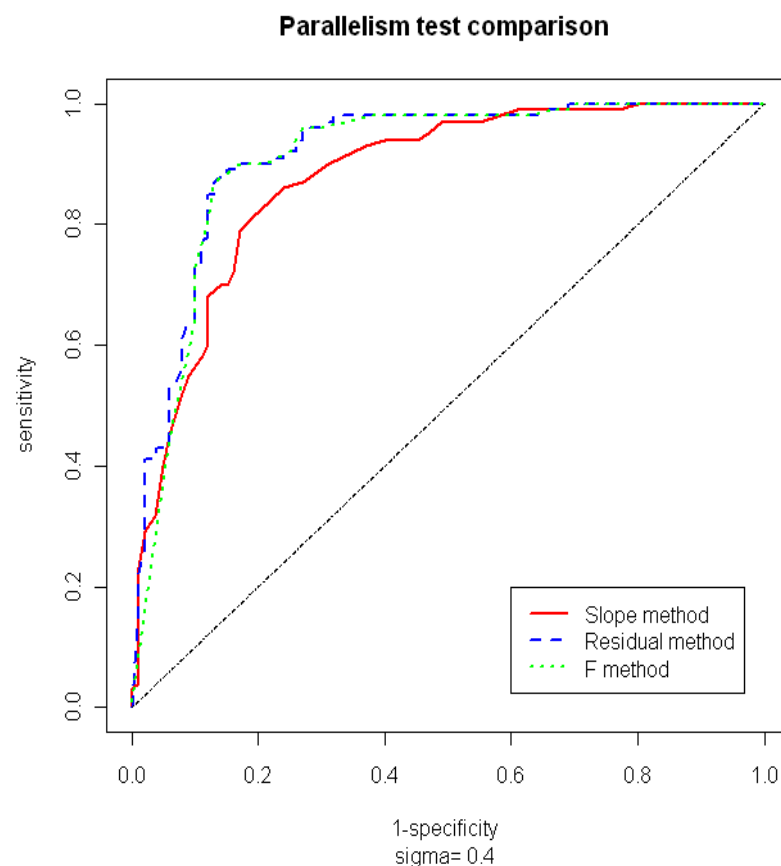
Results of Simulation Study (Cont'd)

- Assay variability = 0.3
- Similar performance of F, χ^2 and equivalence methods
 - ◆ $AUC_F = 0.9505$ (0.0131)
 - ◆ $AUC_{\chi^2} = 0.9635$ (0.0032)
 - ◆ $AUC_{\text{Equiv}} = 0.9439$ (0.0047)
- χ^2 -test and equivalence methods ranked the best and worst, respectively



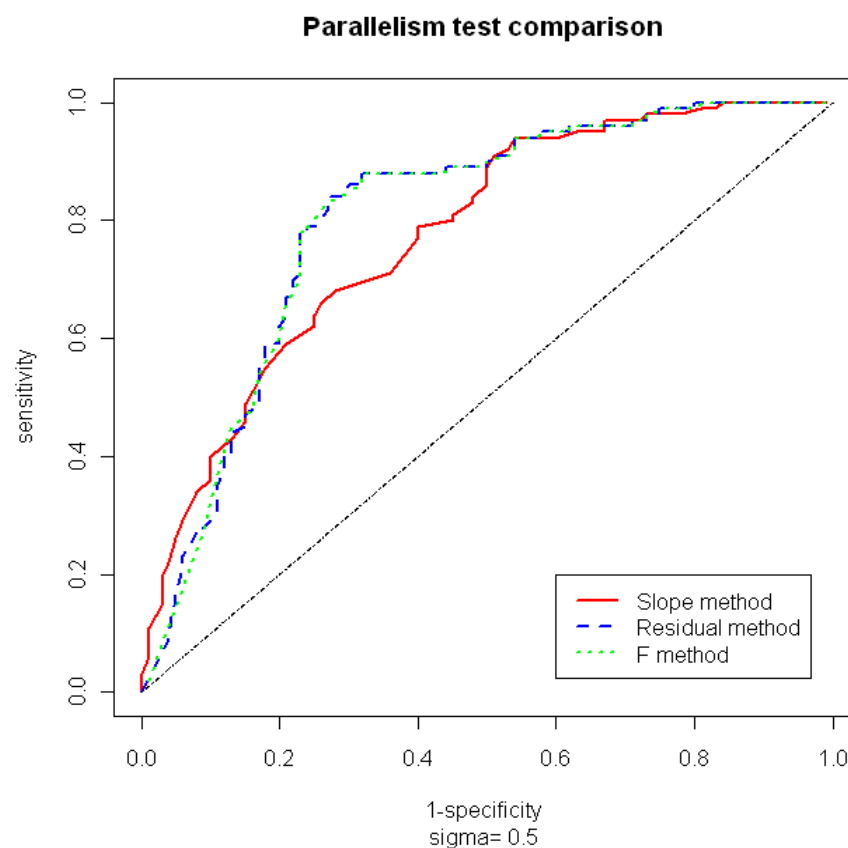
Results of Simulation Study (Cont'd)

- Assay variability = 0.4
- Similar performance of F, χ^2 and equivalence methods
 - ◆ $AUC_F = 0.8831$ (0.0097)
 - ◆ $AUC_{\chi^2} = 0.8895$ (0.0071)
 - ◆ $AUC_{\text{Equiv}} = 0.8597$ (0.0080)
- χ^2 -test and equivalence methods ranked the best and worst, respectively



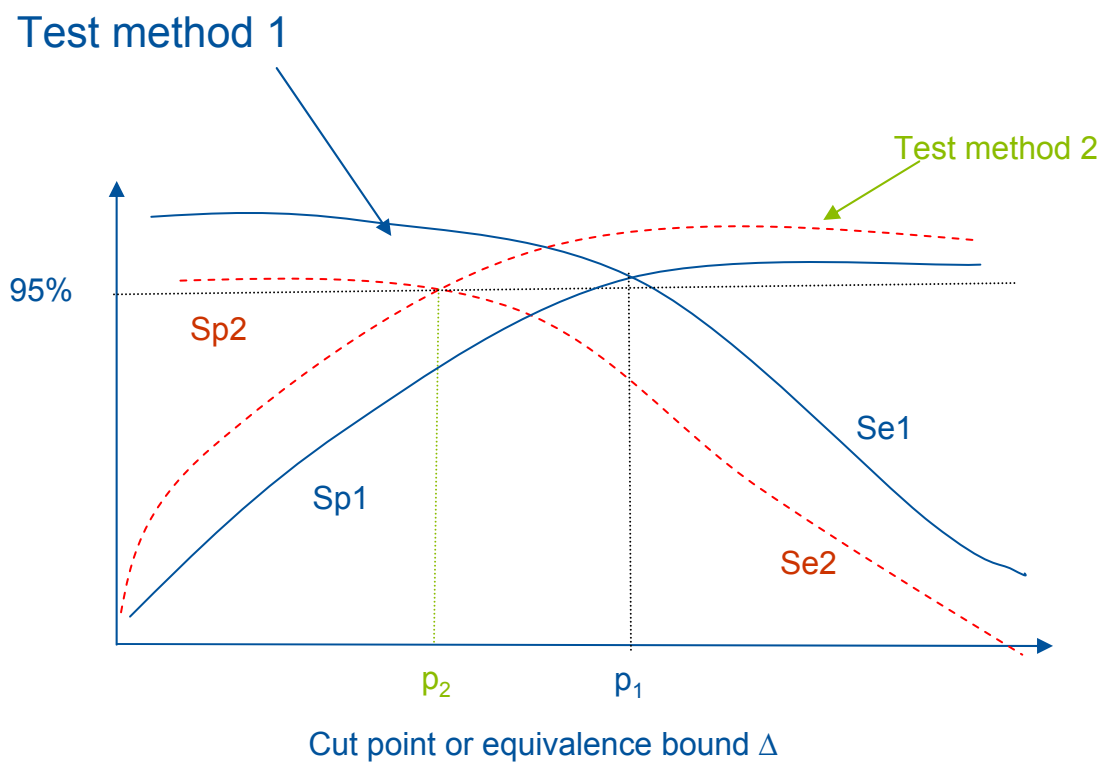
Results of Simulation Study (Cont'd)

- Assay variability = 0.5
- Similar performance of F, χ^2 and equivalence methods
 - ◆ $AUC_F = 0.8106$ (0.0085)
 - ◆ $AUC_{\chi^2} = 0.8126$ (0.0089)
 - ◆ $AUC_{\text{Equiv}} = 0.7810$ (0.0108)
- χ^2 -test and equivalence methods ranked the best and worst, respectively



Cut Point Determination

- For each test method, an optimal cut point, p-value or equivalence bound can be obtained
- The cut point is chosen to make best tradeoff between sensitivity and specificity



An Alternative Method Based on Risk Analysis

- Decision theory can be used to allow for different costs to be assigned to various test outcomes
 - ◆ For example, a high cost can be given to Type I error as opposed to Type II error
- Choose cut point or equivalence bound to minimize mean risk

An Alternative Method Based on Risk Analysis (Cont'd)

| Two curves are | Parallel | Non-parallel |
|----------------|----------|--------------|
| Accept | L_0 | L_1 |
| Reject | L_2 | L_3 |

Choose cut point, c , to minimize the mean risk:

$$R(c) = pL_0Sp(c) + (1-p)L_1[1-se(c)] + pL_2[1-sp(c)] + (1-p)L_3Se(c)$$

where p is the prevalence of the two dose response curves of test sample and reference standard being parallel.

- Parallelism is critical to potency bioassays
- Current method comparisons are biased
- A framework for method comparison based on ROC analysis is proposed
- Simulation studies showed performance of F, χ^2 and equivalence test varies, pending on noise level in dose-response curves
- An optimal cut off value, in terms of test statistic, p-value or equivalence bound can be chosen to make best trade-off between sensitivity and specificity
- Further evaluations using non-linear model are ongoing

- Gottschalk PG, Dunn JR (2005). Measuring parallelism, linearity, and relative potency in bioassay and immunoassay data. *Journal of Biopharmaceutical Statistics*, 15, 237-463.
- Hauck WW, Capen RC, Callahan JD, De Muth JE, Hsu H, Lansky D, Sajjadi NC, Seaver SS, Singer RR, Weisman D (2005). Assessing parallelism prior to determining relative potency. *PDA Journal of Pharmaceutical Science and Technology*, 59: 127-137.
- Jonkman and Sidik K. Equivalence testing for parallelism in the four-parameter logistic model. *Journal of Biopharmaceutical Statistics*, 2009
- Lansky D. (2009). Equivalence testing in nonlinear model bioassay. Presented at 2nd Annual Pharmaceutical Statistical Conference, Arlington, VA.
- The draft USP chapter <111>, October, 2006