Concordance Analysis in Biopharmaceutical Industry

Jason Liao, Ph.D.
Merck Research Laboratories

BASS XIV, November 5, 2007

Outline

- Literature Review
- An Interval Approach
- Three Examples
- Summaries and Recommendations

Literature Review

- Agreement problem: a broad range of data
- Where: medicine and experimental sciences
- It can happen in all phases of drug development

Scenarios

- Reliability of multiple raters (or the same rater over time) in a randomized clinical trial
  - Including and excluding of patients into a trial
- Two clinical endpoints: Surrogate vs. true, Subjective vs. objective
- Two treatments (drug A vs. drug B)
- Two formulations (bioequivalence)
- Two gene sequences (profiles)
- Two biomarkers' performance
- Test vs. re-test
- Two methods, assays, batches, devices, labs, models…

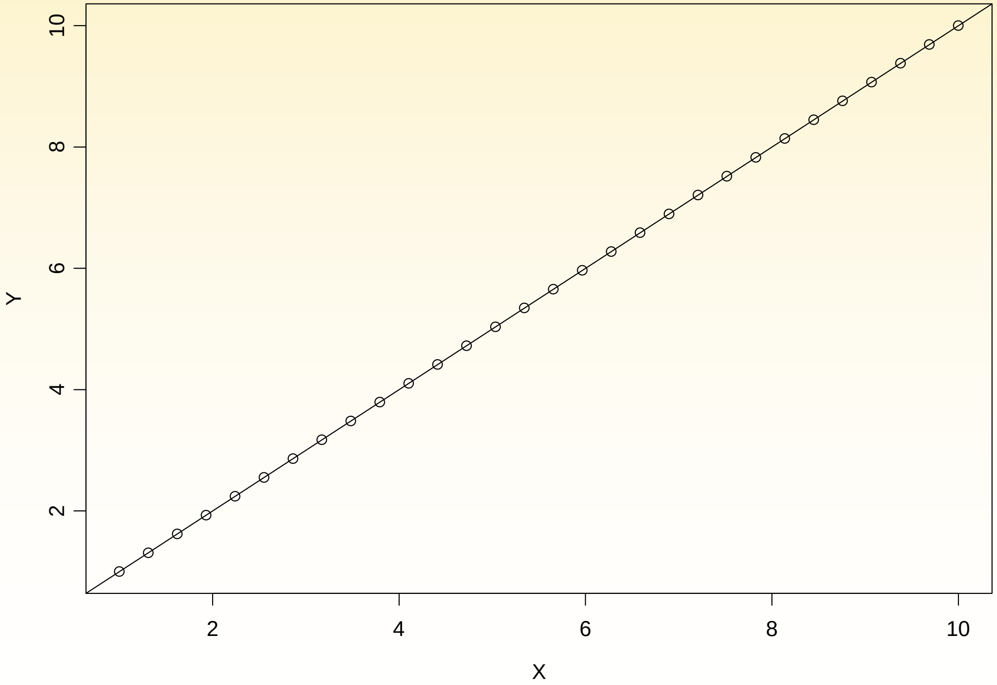
Goal of An Agreement Study

- Various questions:
  - Can the measurements from “raters” be used interchangeably?
  - How does one define and measure agreement?
  - What is the overall level of agreement?
  - How much bias and variance is there among “raters”?
- In summary:
  - Agree with each other?
  - If not, what is the bias and how to calibrate the difference?

- Recent applications:
  - Two clinical outcomes (Deyo, *et al.*, 1991)
  - Assay validation (Lin, 1992)
  - Two methods for human sperm evaluations (Coetzee, *et al.*, 1997)
  - Assay transfer (Liao, 2003)
  - Instrument validation with curved data (Liao, 2005)
  - Assay bridging (Liao, *et al.*, 2006)

How?

- If measurements X and Y are in a perfect match, i.e., agree with each other, then (X,Y) are on the 45^0 line through the origin (identity line)



Existing Approaches (1)

- Hypothesis test:
  - Paired T-test
  - Functional & structural regression approach
  - Agreement in individual means (AIIM) test
  - Mean & variance simultaneous test
  - Intersection-union test (IUT)
- Issue: Heavily depends on the residual variance
  - Reject a reasonably good agreement when the residual errors are small (good precision)
  - Accept a poor agreement when the residual errors are large (less precision)

Existing Approaches (2)

- Index approaches:
  - Correlation coefficient
  - Coefficient of variation (CV)
  - Intraclass correlation coefficient (ICC)
  - Concordance correlation coefficient (CCC) (Lin, 1989)
  - Improved CCC (Liao, 2003)
  - Random marginal agreement coefficient (RMAC) (Fay, 2005)
  - Others (JBS, 2007 special issues)

- Issues:
  - No agreement conclusion
  - A distribution with fixed mean (i.e., one level) and constant covariance
  - Only one single index not enough
  - Very sensitive to data range and sample heterogeneity
  - Not related to the actual scale of measurement
  - No bias information
  - Same value but different meanings in different experiment

Existing Approaches (3.1)

- An interval approach:
  - Limits of agreement (Bland & Altman,1986): 95% CI of sample difference

    $$(\overline{D}-2S_D, \overline{D}+2S_D)$$

    with a supplement mean-difference plot

    - a favorite of medical researcher

- Issues:
  - No agreement conclusion
  - Interpretation difficulty for a mixture of fixed, proportional bias and/or proportional error (Ludbrook, 1997)
  - Only good for additive agreement (e.g. the test-retest situation) (Rousson, *et al.*, 2002)
  - Only limited bias information
  - Metrics not valid for all situations
  - Not adjustable for covariates
  - Artifactual bias information from the mean-difference plot (Hopkins, 2004)

Existing Approaches (3.2)

- An interval approach (cont.):
  - Total deviation index (Lin, 2000): using any probability instead of 95%
  - Coverage probability (Lin, *et al.*, 2002)
  - Tolerance interval (Choudhary & Nagaraja, 2005)
- Issues:
  - Share some of the drawbacks of Bland & Altman's approach
  - Distribution with a fixed mean (i.e., one level) and variance

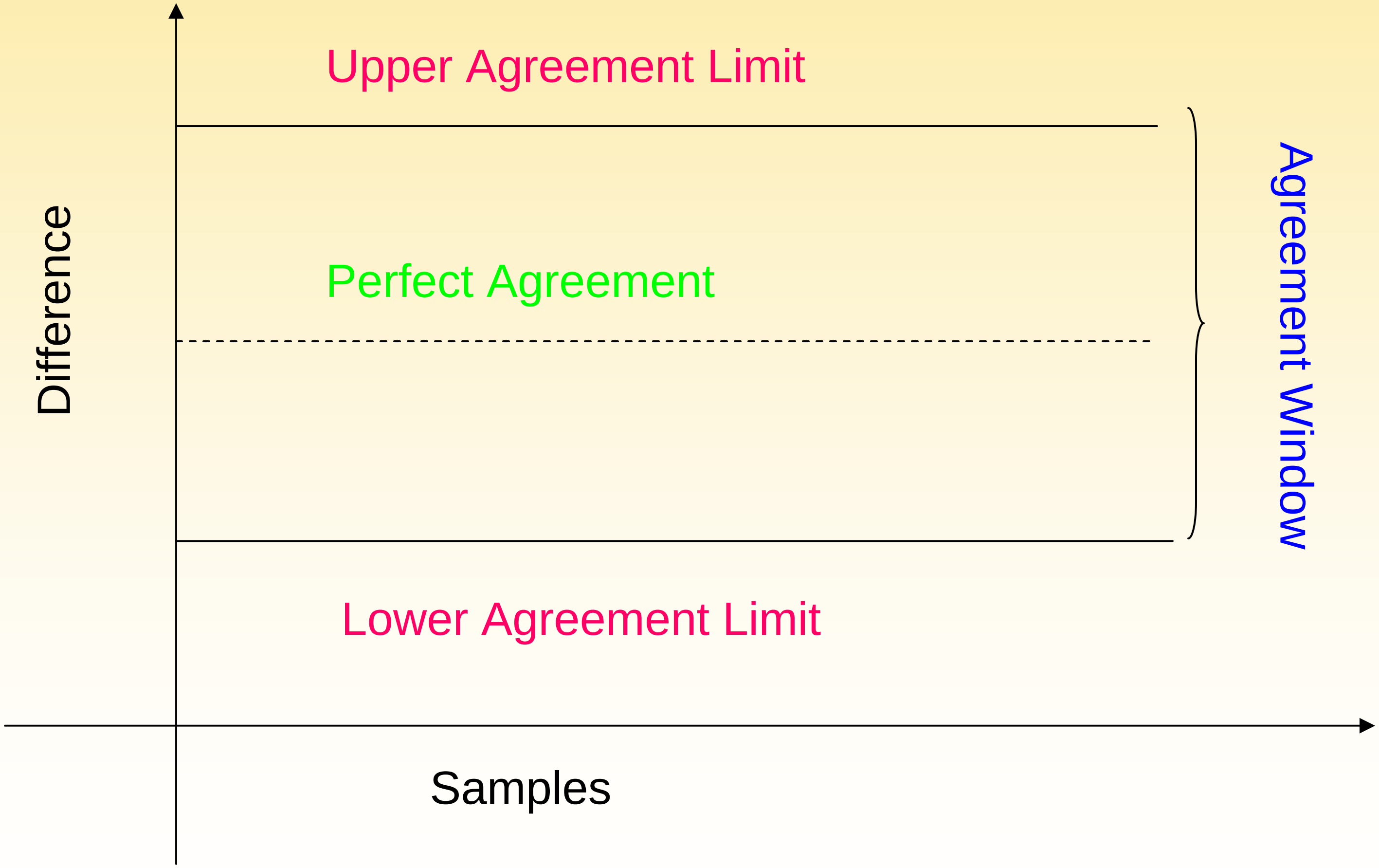
Needs for a New Method

- Practical meaningful and easy interpretation
- 1st goal of an agreement study: conclusion
- 2nd goal of an agreement study: bias information
  - fixed and/or proportional
- Easy adjustment for covariates or factors

The New Interval Approach (Liao, et al., 2006a & b)

- People judge agreement by difference
- Interval Δ: $P(y-x\in\Delta)=1-\alpha, say, 0.95$
- Accepted concordance: All paired differences fall into the agreement interval Δ
  - Informative
  - Easy bias detection
  - SPC techniques
  - The flexible acceptance criteria
    - FDA guidance (2001): 4-6-15 for accepting batches

Supplement Graphic Illustration



- Observations (X,Y):

$$Y = a + b \times X^0 + \varepsilon$$

$$X = X^0 + \delta$$

where $\varepsilon \perp \delta$ and are from $N(0, \sigma^2)$

- $\Delta = (-t_{1-\alpha/2,n-1} \times \sqrt{2}\hat{\sigma}, +t_{1-\alpha/2,n-1} \times \sqrt{2}\hat{\sigma})$
  - This agreement interval for absolute agreement
  - Same as BA's limits of agreement if no bias

- **The interval should compare to the scientifically acceptable boundary**

Advantages of New Approach

- A criterion for making a conclusion using SPC technique
- Very informative: Bias information fully available
- Covariates adjustable
- All metrics valid
- Adjustable for fixed and/or proportional bias, proportional error cases

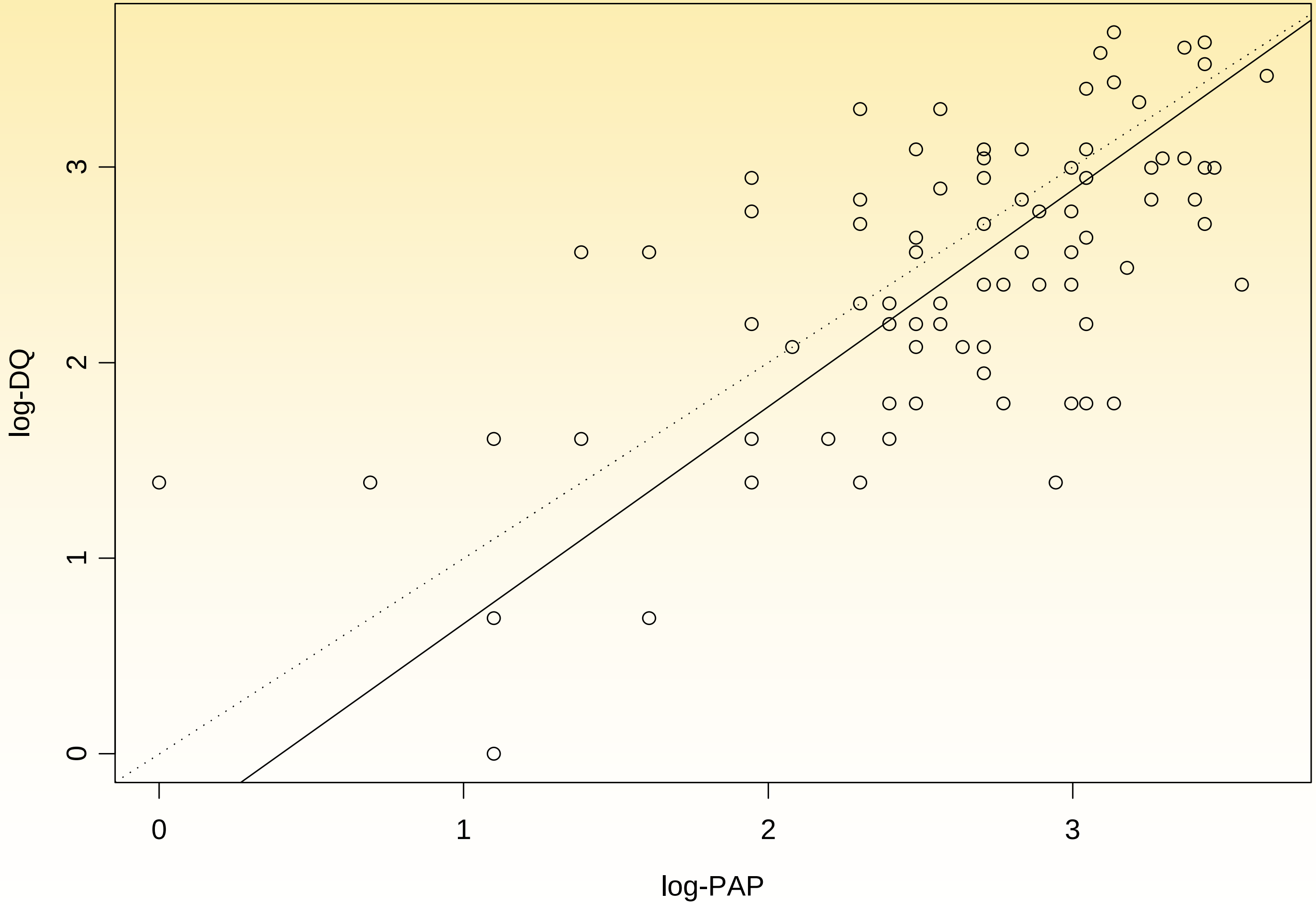
- $\Delta = (a_0 - t_{1-\alpha/2,n-1} \times \sqrt{2}\hat{\sigma}, a_0 + t_{1-\alpha/2,n-1} \times \sqrt{2}\hat{\sigma})$
  - This agreement interval for additive agreement
  - Same as BA's limits of agreement if only fixed bias
- $\Delta_i = ((b_0 - 1) \times X_i - t_{1-\alpha/2,n-1} \times \sqrt{2}\hat{\sigma}, (b_0 - 1) \times X_i + t_{1-\alpha/2,n-1} \times \sqrt{2}\hat{\sigma})$
  - This agreement interval for multiplicative agreement
- $\Delta_i = (a_0 + (b_0 - 1) \times X_i - t_{1-\alpha/2,n-1} \times \sqrt{2}\hat{\sigma}, a_0 + (b_0 - 1) \times X_i + t_{1-\alpha/2,n-1} \times \sqrt{2}\hat{\sigma})$
  - This agreement interval for linear agreement
- $\Delta = (-t_{1-\alpha/2,n-1} \times \sqrt{1+\lambda}\hat{\sigma}, +t_{1-\alpha/2,n-1} \times \sqrt{1+\lambda}\hat{\sigma})$
  - This agreement interval for proportional error case
  - Can be avoided in design stage

Example One

- Study for computerized human sperm morphology evaluations (Coetzee, *et al.*, 1997)
- The normal sperm morphology, as a diagnostic tool, has been used as an important predictor of male fertility
- Papanicolaou (PAP): to establish the standard fertility thresholds
- Diff-Quik (DQ): its simplicity

Scatter Plot

where dotted line is the identity line



Functional Regression Approach

- Intercept: -0.445

  95% CI: (-1.159, 0.270)
- Slope: 1.110

  95% CI: (0.838, 1.381)

=➔ Good agreement

Index Approaches

- ICC: 0.597

  95% CI: (0.333, 0.861)
- CCC: 0.625

  95% CI : (0.484, 0.735)
- Improved CCC: 0.629

  95% CI: (0.508, 0.725)

=➔ Moderate or substantial agreement

Interval Approaches

(Liao, et al. Approach)

- $$\hat{\sigma} = 0.412$$

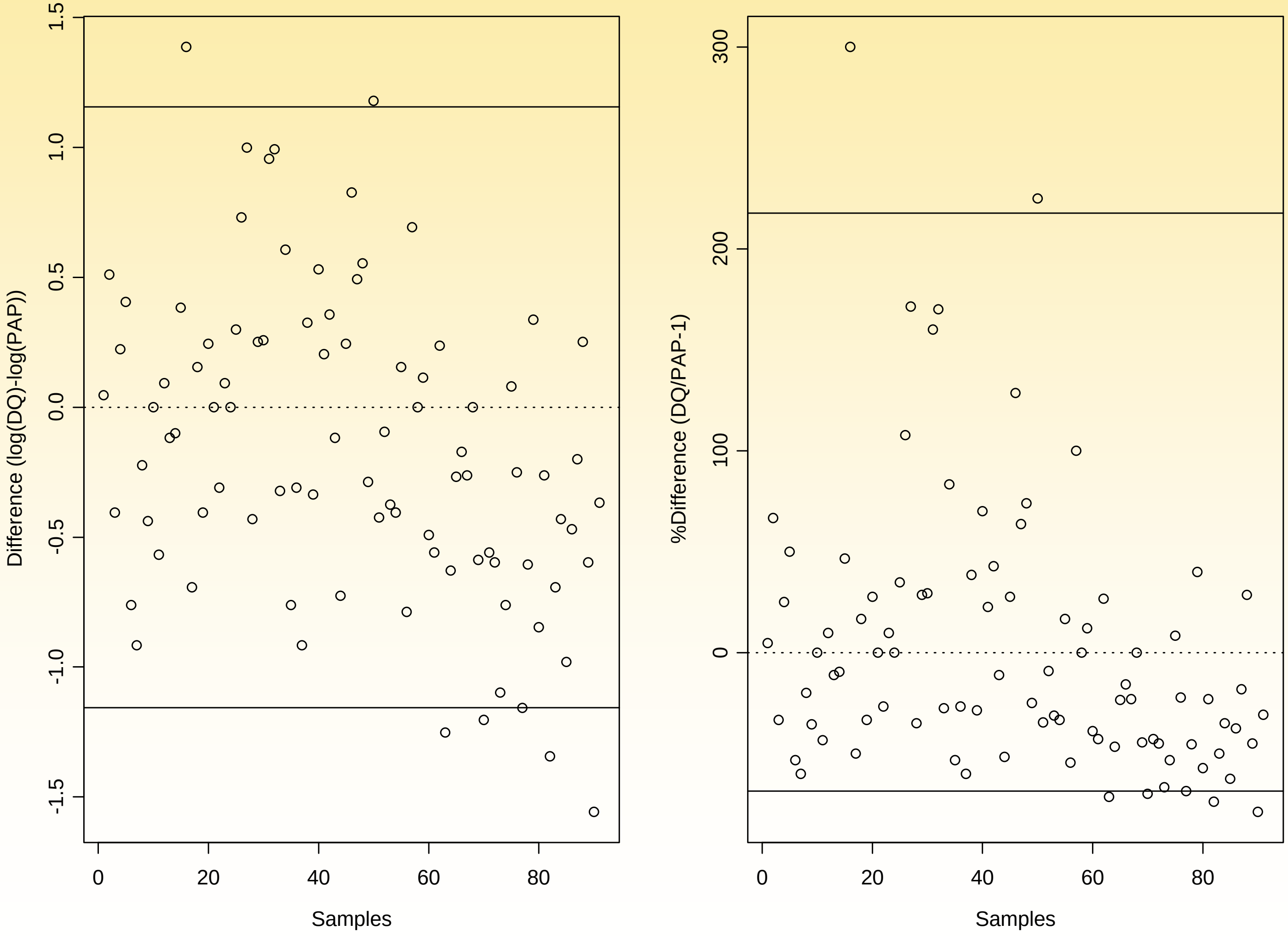
- The agreement interval

$$\Delta = (-1.156, 1.156)$$

==➔No agreement

log-bias: $-0.445 + 0.110 \times \log(PAP)$

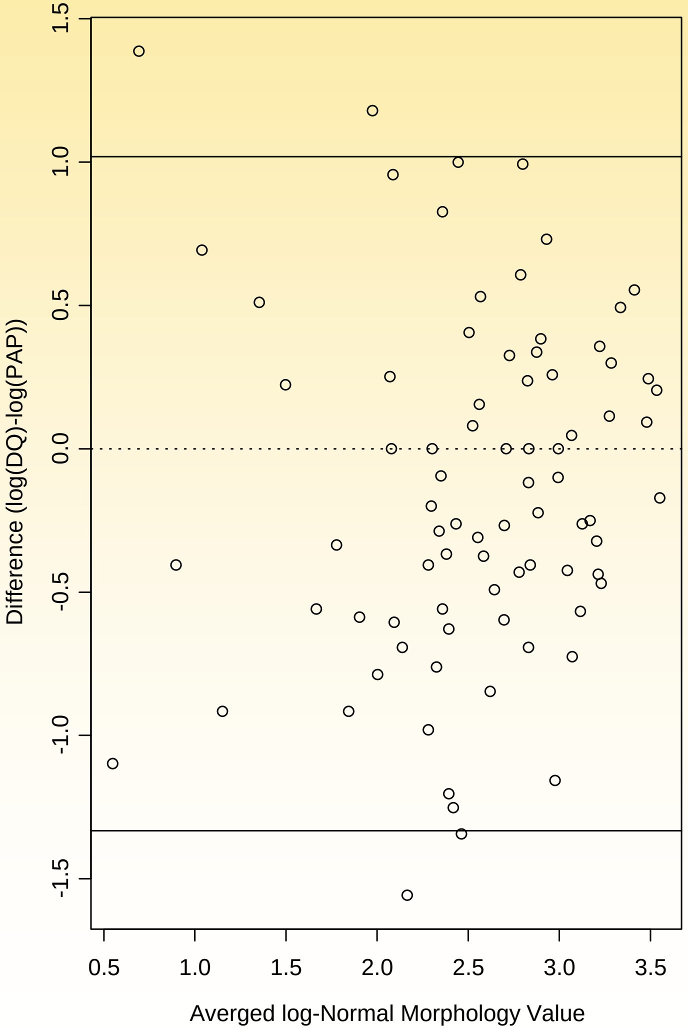
Concordance Assessment
(Liao, et al. Approach)

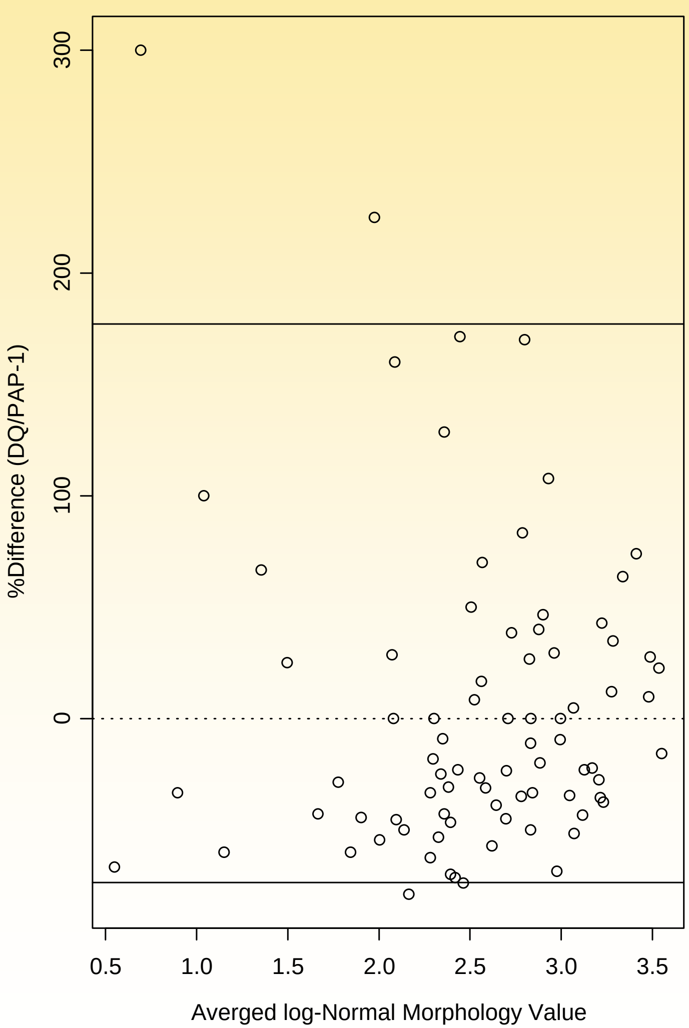


Interval Approaches
(Bland & Altman's Approach)

- $\overline{D} = -0.157$
- $S_D = 0.588$
- Limits of agreement: (-1.333, 1.019)

Concordance Assessment
(Bland & Altman Approach)





Example Two

- A modified new assay (comparator) to replace the current assay (reference) (Liao, et al., 2006)
- Current assay concentration range 10 to 800 U/mL
  - Three different sample matrices
- Issues: how many samples? how to cross-validate the new assay?

- N=32 pairs
- Three matrices with overlap in concentrations (U/mL):

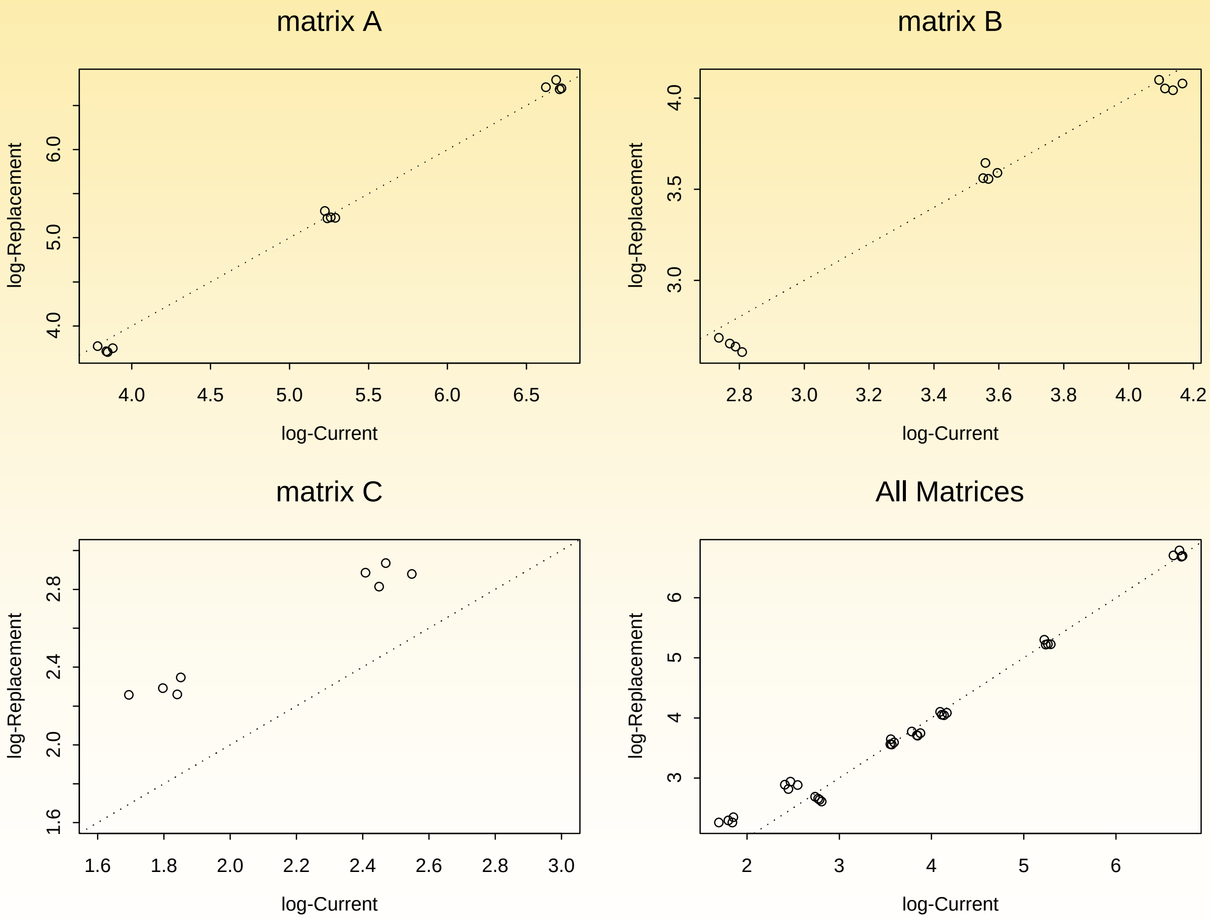
  Matrix A: 800, 200, 50

  Matrix B: 62, 35, 15

  Matrix C: 20, 10
- Four aliquots of each were prepared

Scatter Plot

where dotted line is the identity line



- The linear measurement error model:

  Matrix A: $\hat{a} = -0.284, \hat{b} = 1.049, \hat{\sigma} = 0.04$

  Matrix B: $\hat{a} = -0.292, \hat{b} = 1.067, \hat{\sigma} = 0.047$

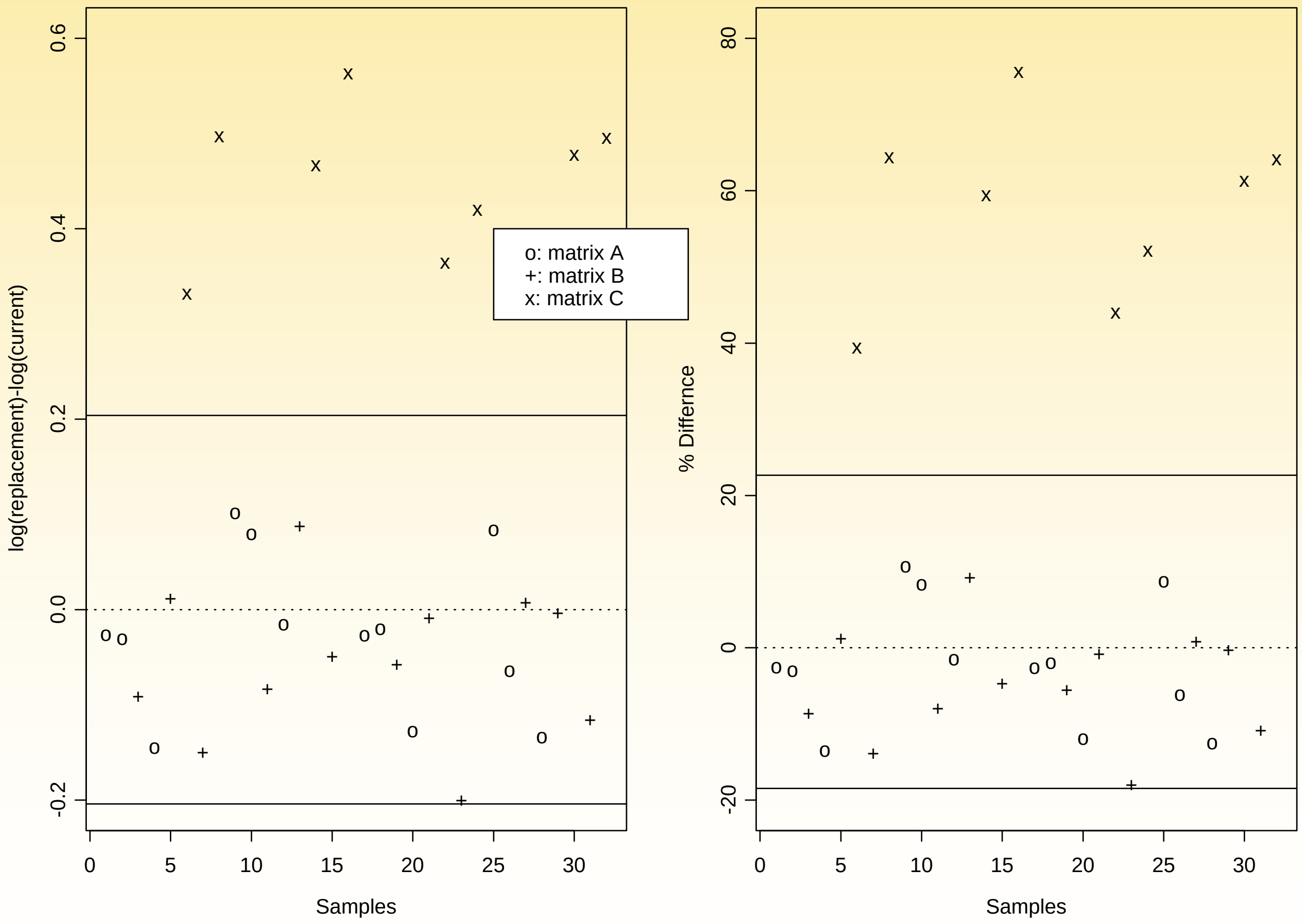
  Matrix C: $\hat{a} = 0.729, \hat{b} = 0.870, \hat{\sigma} = 0.039$

- Is there a matrix effect?

- There is no matrix effect on the variance
- There is a matrix effect on regression line
- Estimating the common variance:
  - Remove one of the four aliquots each time
  - Estimate the variance for each matrix
  - Pool the variance
- The agreement interval

$$\Delta = \left(-0.204, +0.204\right)$$

Concordance Assessment
(Liao, et al. Approach)



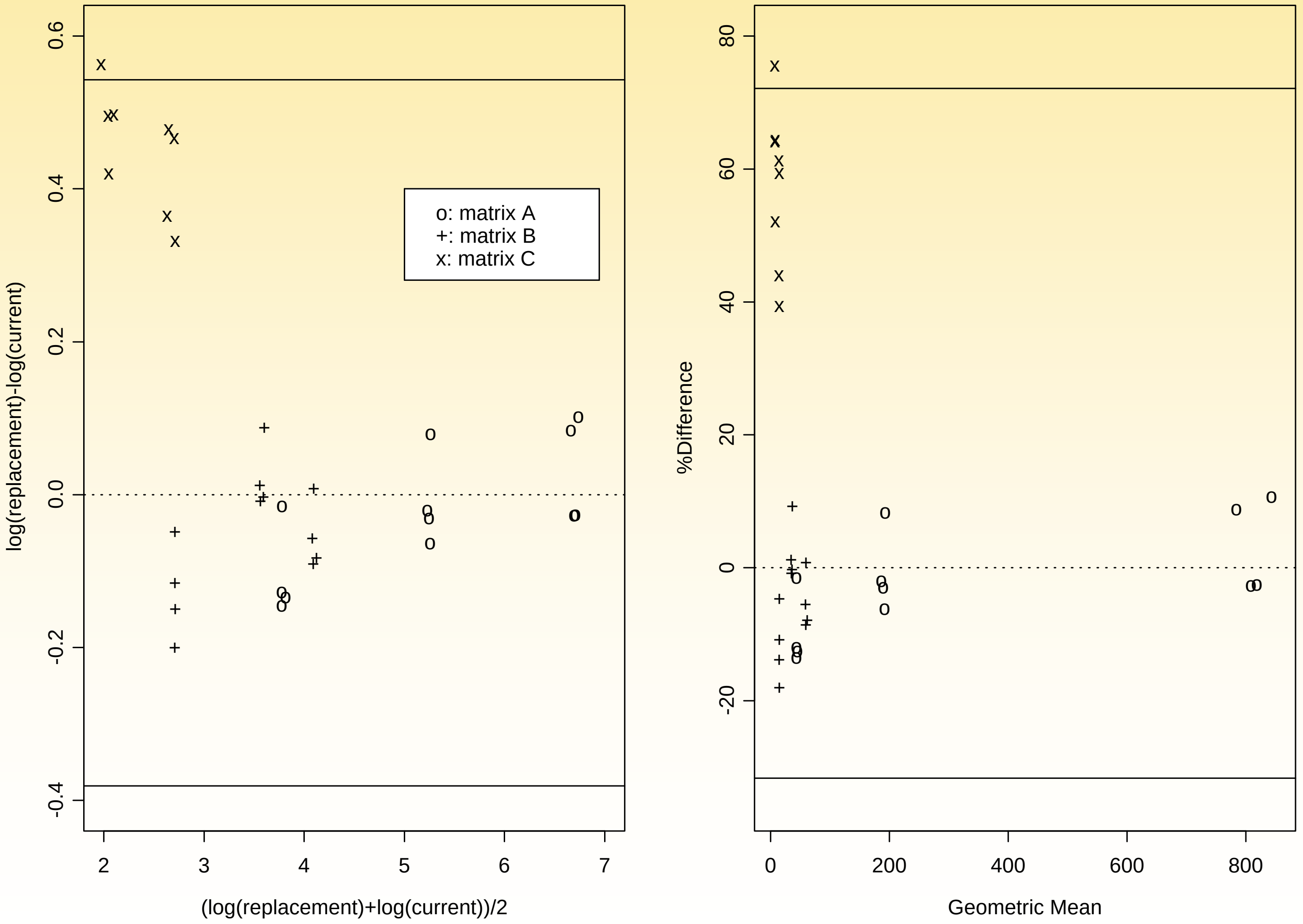
- All 24 paired observations fell within the agreement interval for matrices A and B
- All eight paired observations fell outside of the agreement interval for matrix C
- Large bias $(0.729-0.13*\ln(current))$ in matrix C
- Two assays do not agree with each other

Bland & Altman's Approach

- $\overline{D} = 0.081$
- $S_D = 0.231$
- Limits of agreement: (-0.381, 0.543)

Concordance Assessment

(Bland & Altman Approach)



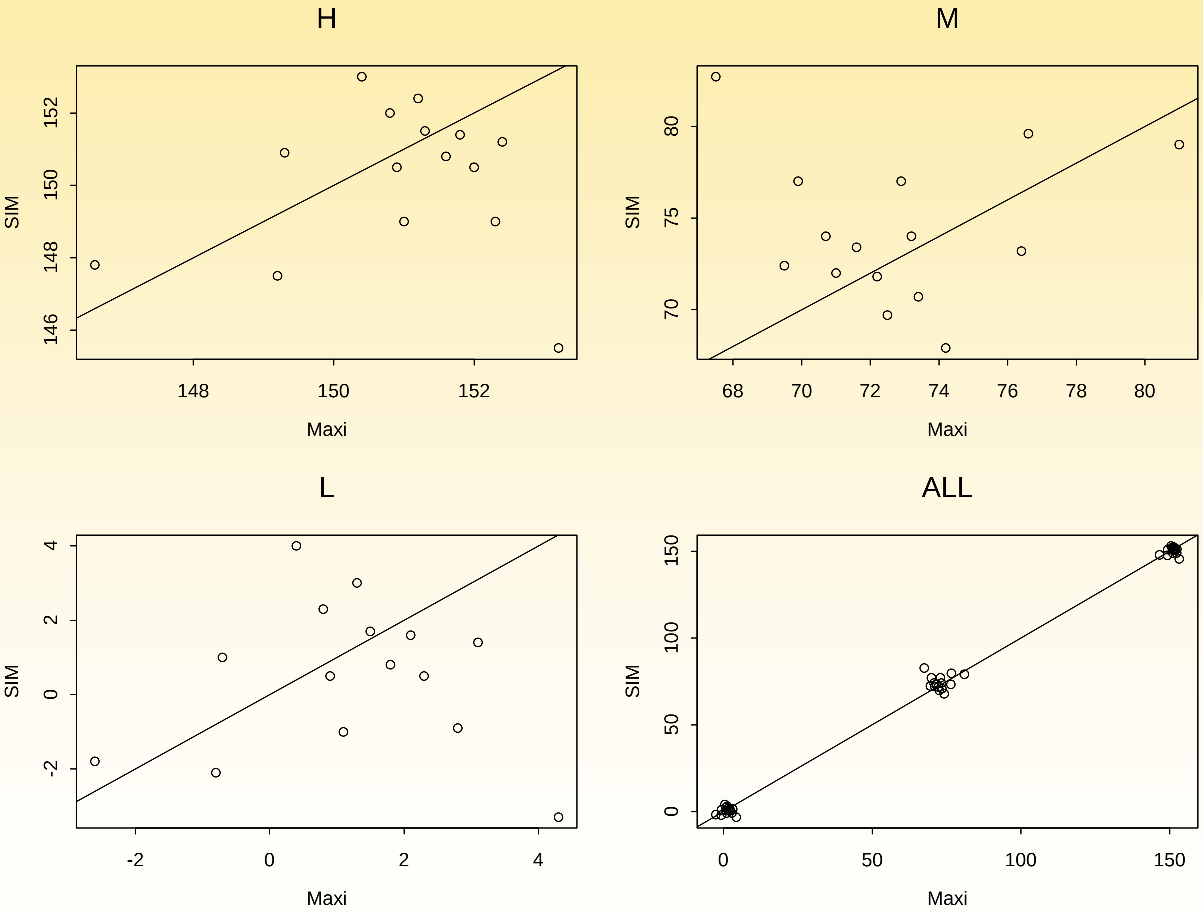
Example Three

- Phase II clinical dental study of a protein
- The bone density: at three different cross-sectional areas, called ``L", ``M" and ``H", using CT scan at visits 1, 3 and 8
- MAXI: used for visits 1 and 3
- SIM: future visit 8
- Validate SIM: how many samples? how to evaluate concordance?

- N=45 pairs
- Each patient has three measurements: “L”, “M” and “H”
- Therefore, 15 patients were randomly selected

Scatter Plot

where slide line is “S=M”



- There was one outlier in all locations “L”,”M” and “H”
- It was the same patient: No.21 whose scan was degraded by spray artifact

- The linear measurement error model:

$$S = a + b \times M^0 + \varepsilon$$

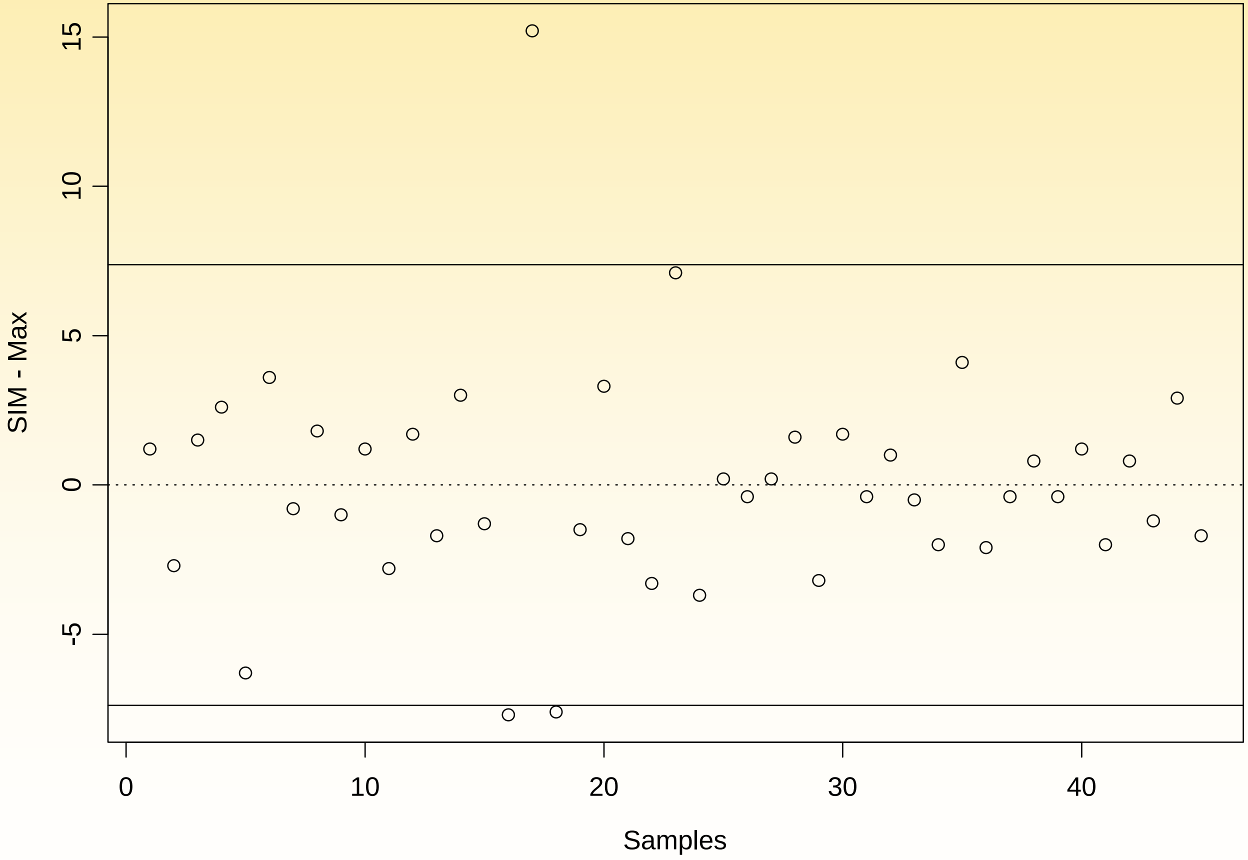
$$M = M^0 + \delta$$

- The agreement interval

Δ =(-7.38, 7.38).

Concordance Assessment

(Liao et al. Approach)



- The differences from the remaining 42 pairs of 14 patients were within the agreement interval (-7.38, 7.38)
- The two programs (Maxi vs. SIM) agreed with each other

Summaries and Recommendations

- A very informative method was suggested for assessing the concordance of two measurement methods
  - Detect any bias easily
  - Can be subject specific in defining acceptance criteria
- This approach handles the measurement range, bias, etc.
- The concordance can be adjusted for covariates, factors such as the matrix effect
- A non-zero k can be used to make an agreement conclusion but this k should be chosen before the data are available
  - FDA 4-6-15 rule for batches acceptance
- Suggested sample size: 32 or 45

Thank you!

- Any Questions?