Statistical Considerations in Biomarker Method Development & Validation

Viswanath Devanarayan, Ph.D.
Eli Lilly & Company
E-Mail: devan@lilly.com

Presented at BASS-XI, November 1, 2004

Acknowledgements

Biologists:

- Len Boggs
- **Ron Bowsher**
- Karen Cox
- Jean Lee
- Peter O'Brien
- Chad Ray
- Sitta Sittampalam

Statisticians:

- Bruno Boulanger
- Ray Carroll
- Walthere Dewe
- **Wendell Smith**

Outline

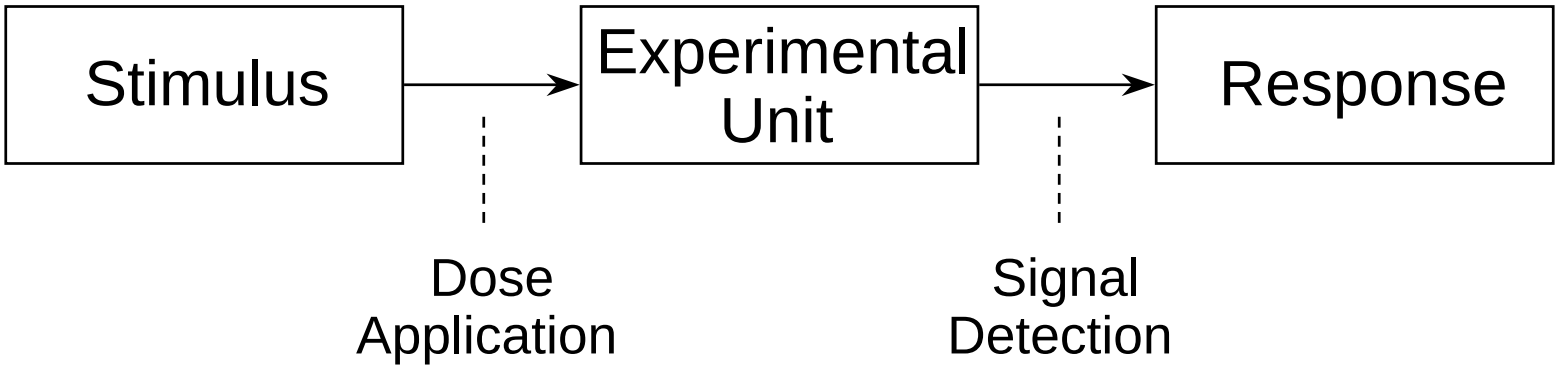
1. Assay/Method Background
2. Fundamental Validity, Similarity, Parallelism
3. Types of Biomarker Methods
4. Standard Curves, Weighting, Precision Profiles
5. Assay/Method Optimization
6. Pre-Study & In-Study Validation
7. Acceptance Criteria
8. Summary (Flow Scheme)

1. Assay/Method Background

Assay Purpose (Finney)

A (biological) assay is an experiment run to estimate the nature, constitution, or potency of a material, by means of the reaction that follows its application (to living matter).

Assay Structure



- The size of the stimulus (“dose”) is varied to obtain a dose response curve.
- “Potency” of a test sample is estimated by comparing its response to that of a **standard preparation**.

Assay Inference

Interest is primarily on "estimation" of some property of the material

Similar to methods of physical measurement, but with more complex sources of variation

Different from experimental studies that are designed to compare effects of known treatments

Nature of a Standard Preparation

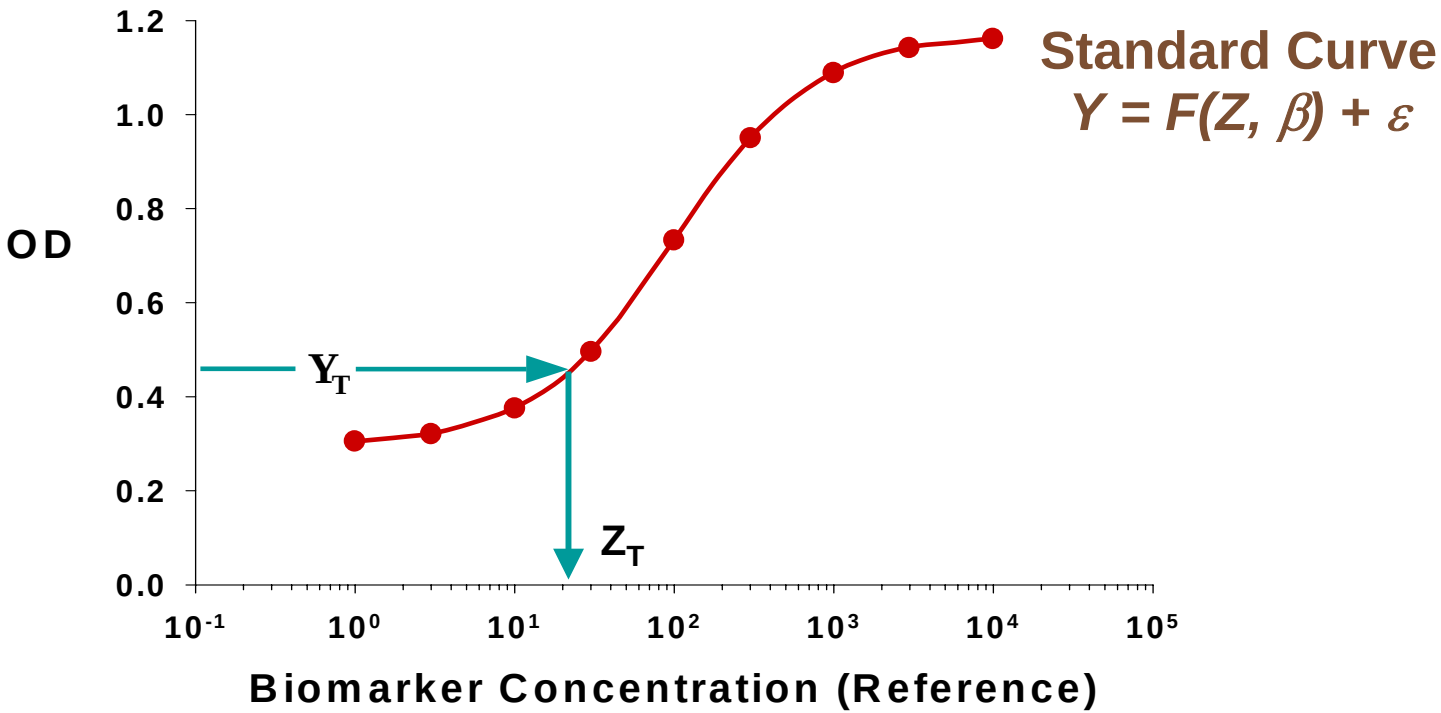
"CRASS"

- **Characterized and purified well**
- **Representative of samples to be assayed**
- **Available in large quantity**
- **Stable under well-defined conditions**
- **Accessible to participating laboratories**

Note: Samples of the standard preparation must be included in each "run" of an assay!

Biomarker Quantification

Standard Curve



Y_T = Observed Assay signal of Test (T) sample

Z_T = Calibrated Biomarker level in test sample = $F^{-1}(Y_T, \beta)$

2. Fundamental Validity, Similarity, Parallelism

Similarity Requirement

What assumptions must be satisfied for a Biomarker result to be **fundamentally valid** when calibrated from a reference (standard) material?

Similarity condition (Finney, 1978):

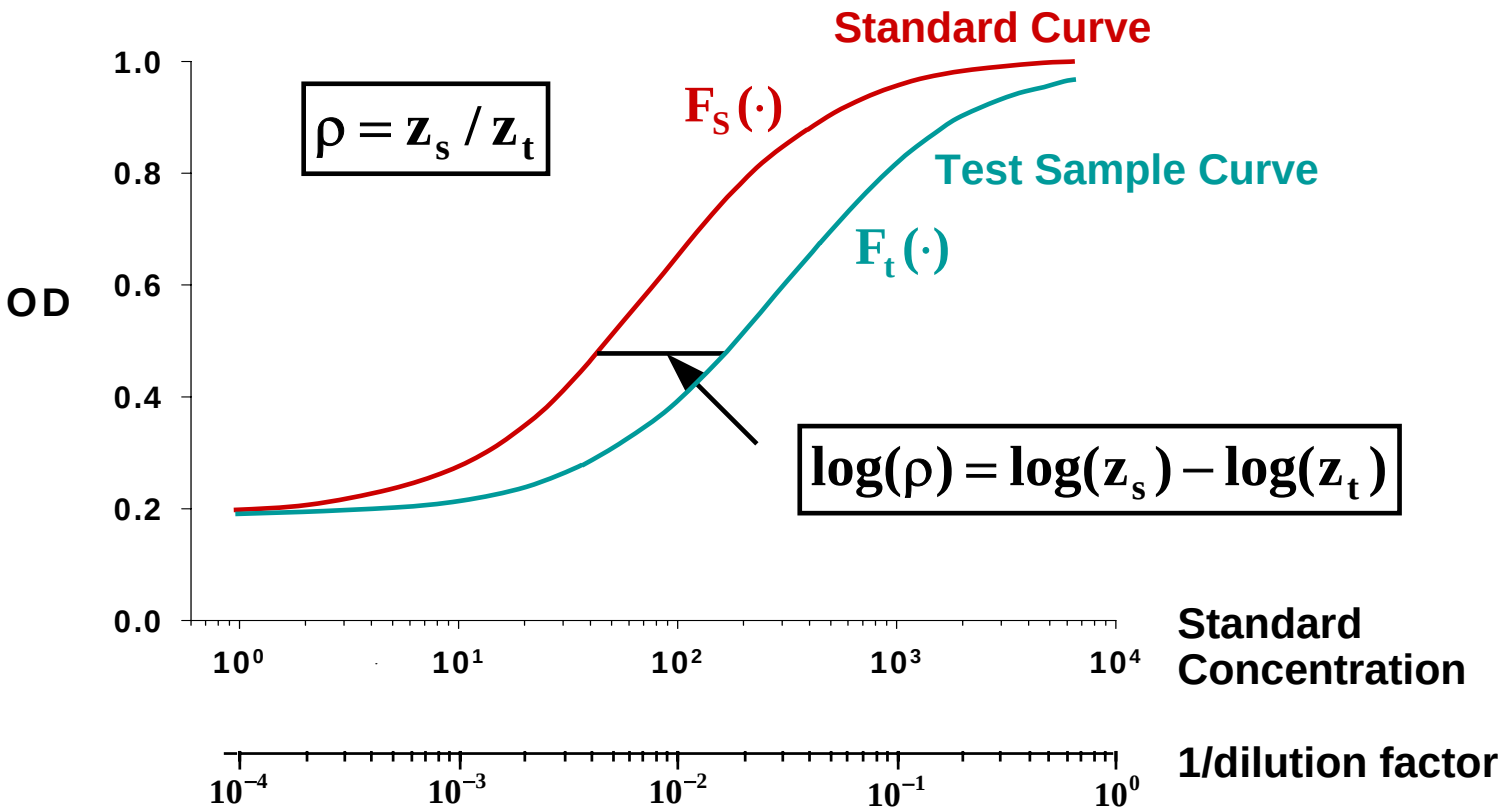
1. Dose response functions for the test (T) and standard (S) preparations must satisfy

   $$F_t(z) = F_s(\rho \bullet z) \quad \text{for all doses } z$$

2. $F_t(\bullet)$ and $F_s(\bullet)$ have the same functional form
3. ρ is a constant (defined as relative potency)

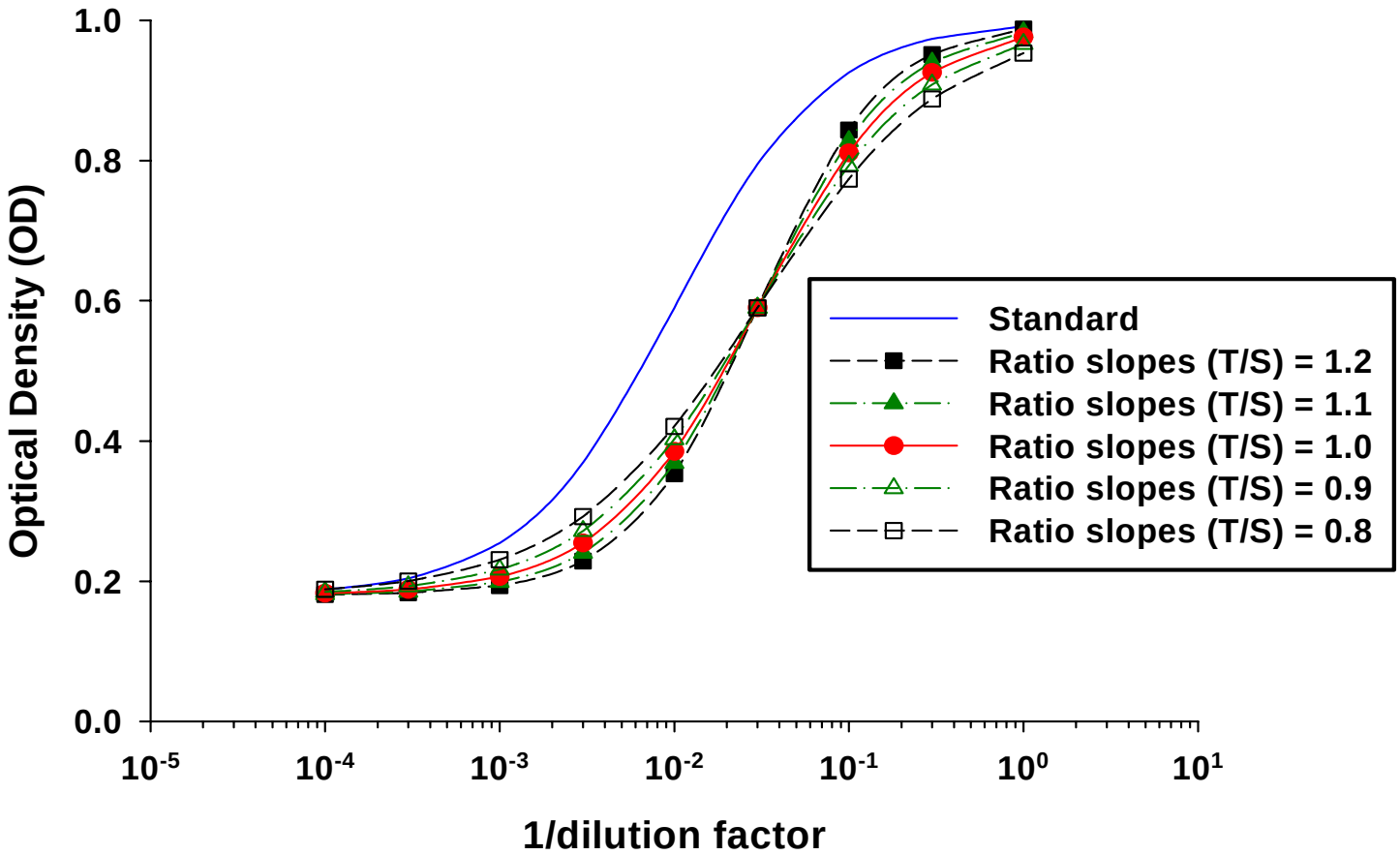
Parallelism

Illustration



Effect of Non-Parallelism

Illustration



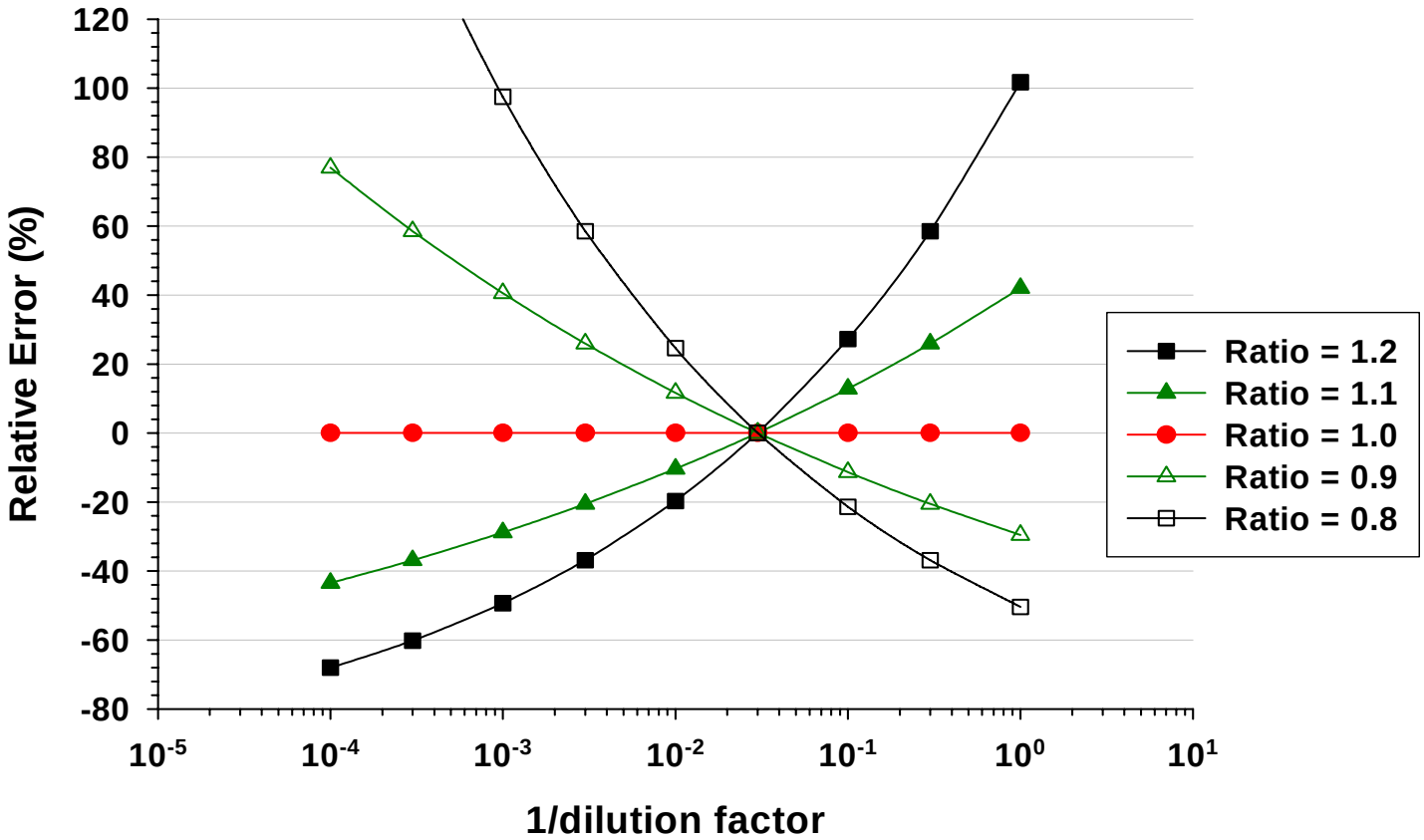
Effect of Non-Parallelism

Illustration (contd.)

μ_T: nominal value

z : interpolated result

$$\%RE = \left(\frac{z - \mu_T}{\mu_T} \right) \cdot 100$$



Parallelism

Assessment & Analysis

"Minor" differences in slopes can cause major effects in the "relative error" (bias).

Statistical significance:

- **The difference in slopes & other parameters can be tested within the framework of nonlinear models.**
- **Implementation: gnls()/nlme() function in Splus**

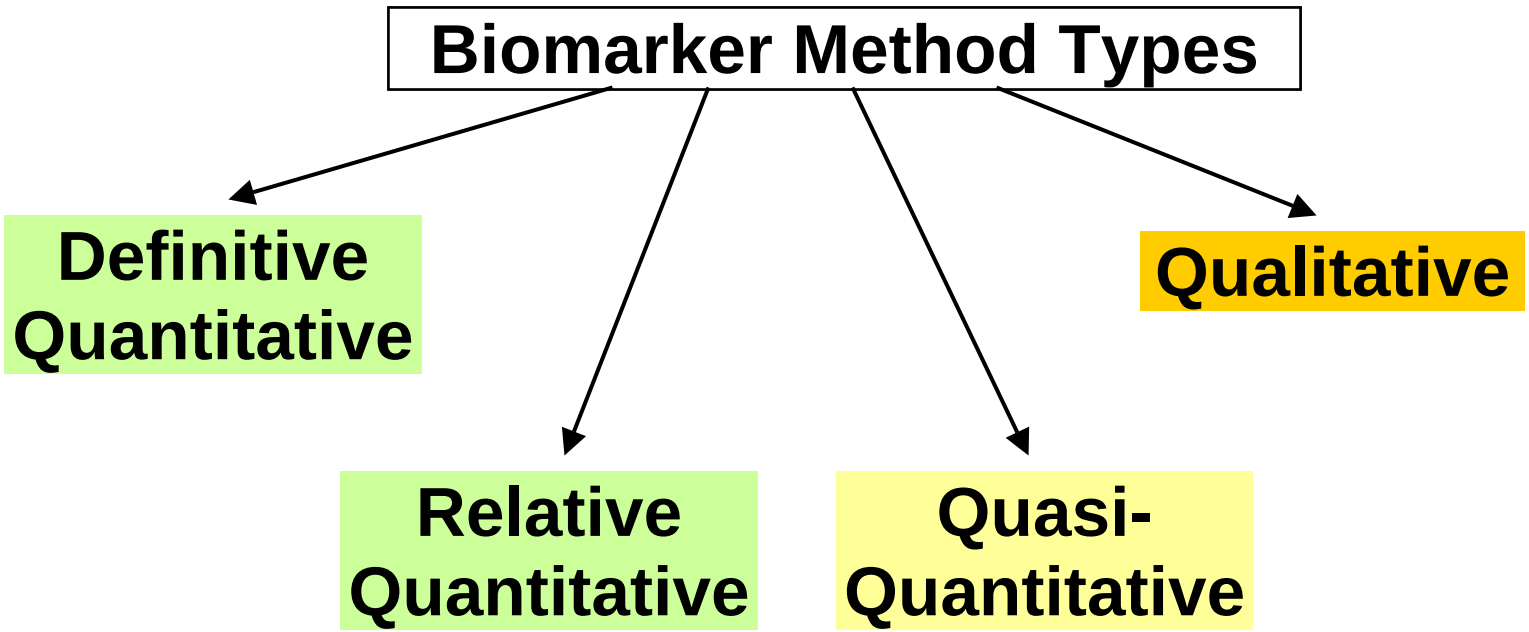
Biological significance:

- **How much relative error (bias) is clinically acceptable?**

Consider both Statistical & Biological Significance!

3. Types of Biomarker Methods/Assays

Types of Biomarker Methods



Types of Biomarker Methods

Definitive Quantitative

Reference Standard available

- Well defined,
- Fully representative of the endogenous protein.

Analytical result is expressed in continuous units of the definitive reference standard.

Examples:

- Human insulin
- Steroid Assays

Ideal situation

Types of Biomarker Methods

Relative Quantitative

Reference Standard available.

- Not well characterized,
- Not available in a purified form, or is
- Not fully representative of the endogenous protein
  - **Relative!**

Analytical result is expressed in continuous units of the relative reference standard.

Example: Cytokine ELISAs

Types of Biomarker Methods

Quasi Quantitative

Quasi ⇨ “possesses certain attributes”

Reference Standard not available or is not ‘valid’.

Continuous response

Analytical result: characteristic of the test sample

- Assay Signal

Examples:

- Enzymatic assays (Activity Units)
- Anti-drug antibody assays (titers)
- Flow cytometry assays

Types of Biomarker Methods

Qualitative

No Reference Standard

Discrete response

Analytical result: characteristic of the test sample

- Assay Signal

Ordinal data:

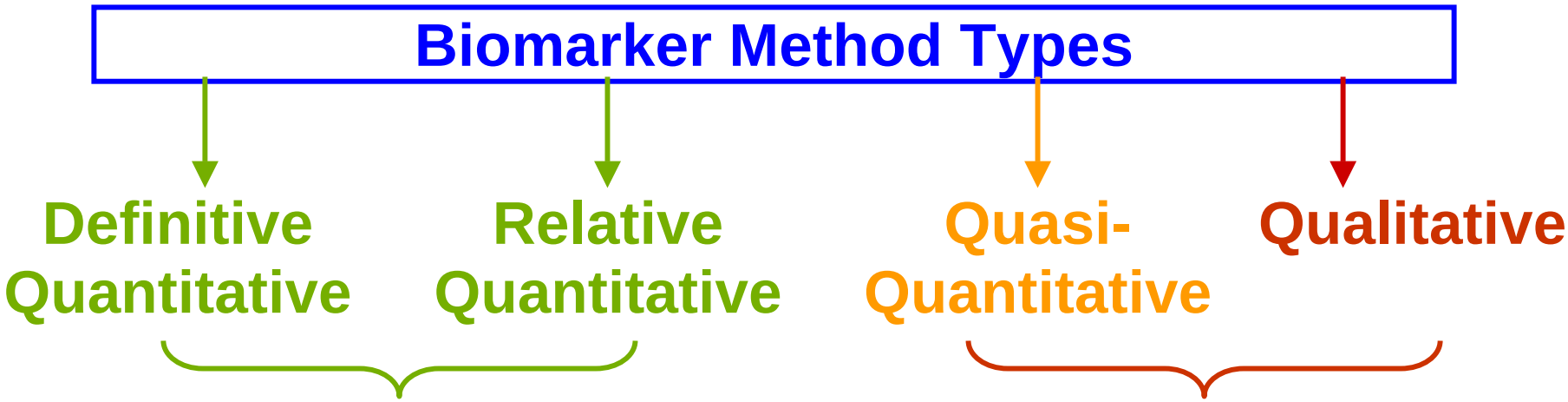
- Ordered & non-continuous responses
- Examples: +, ++, +++ or low, mid, hi

Nominal data:

- Non-ordered & non-continuous responses
- Examples: reporting results as positive (+) or negative (-)

Types of Biomarker Methods

Recap!



- Biomarker levels determined using a "reference standard".
  - Reference standard must be truly representative of test samples
    - Parallelism
    - Standard Curve

Scope of this talk!

- Biomarker levels represented only by assay signal.
  - Reference material not available or
  - Not representative of test samples.
    - Non-Parallelism
    - Standard Curve not used
  - Example: Mire-Sluis et al (2004, JIM)

4. Standard Curves, Weighting & Precision Profiles

Standard Curve

Standard Curve model includes

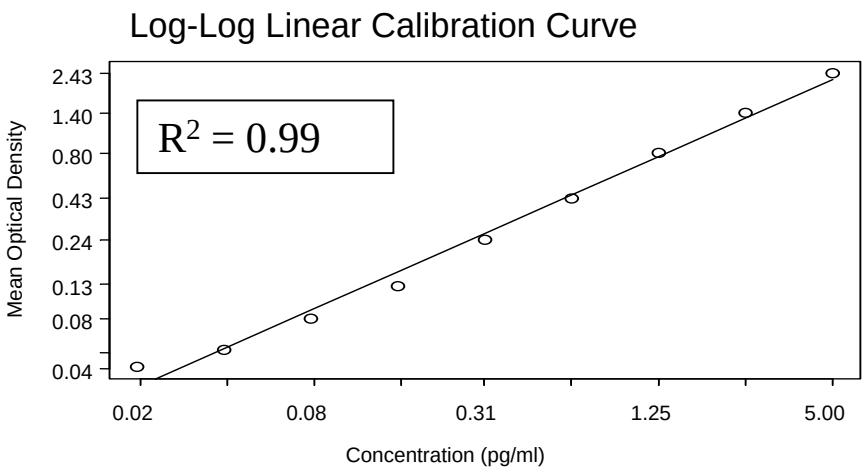
- **Mean Function**
- **Response Error Function (Weighting)**

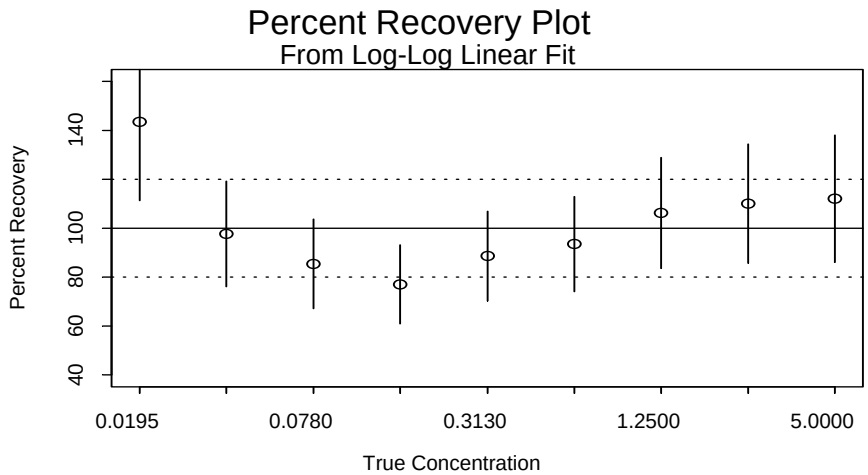
Mean Functions:

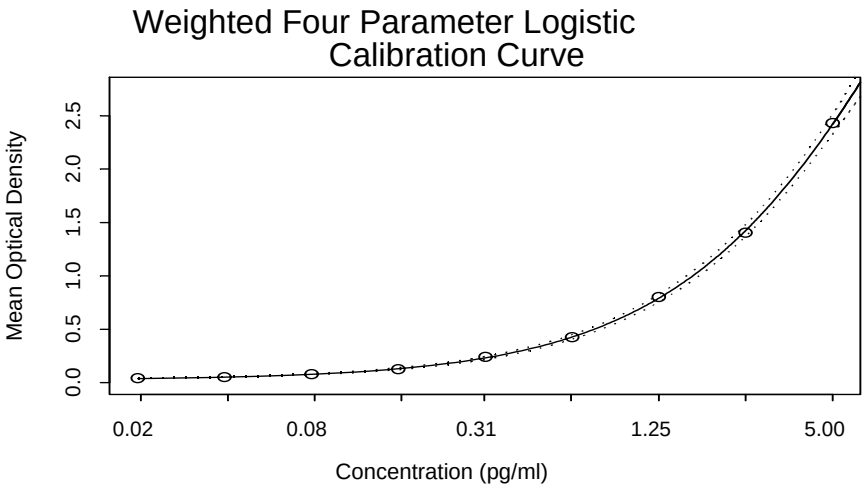
- Polynomial (linear, quadratic, linear in log-scale, etc.)
- Nonlinear (example: Logistic models)
- Spline

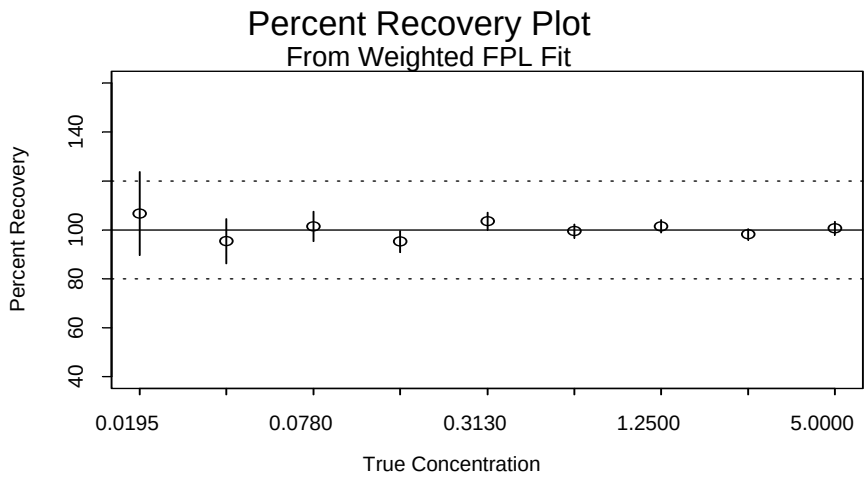
Standard Curve: Mean Function

- Linear model is far worse than 4PL, even though $R^2 = 0.99$









Percent Recovery = 100*(Estimated Conc. / True Conc.)

Standard Curve

Response Error Function (Weighting)

Most laboratory software assume constant variance.

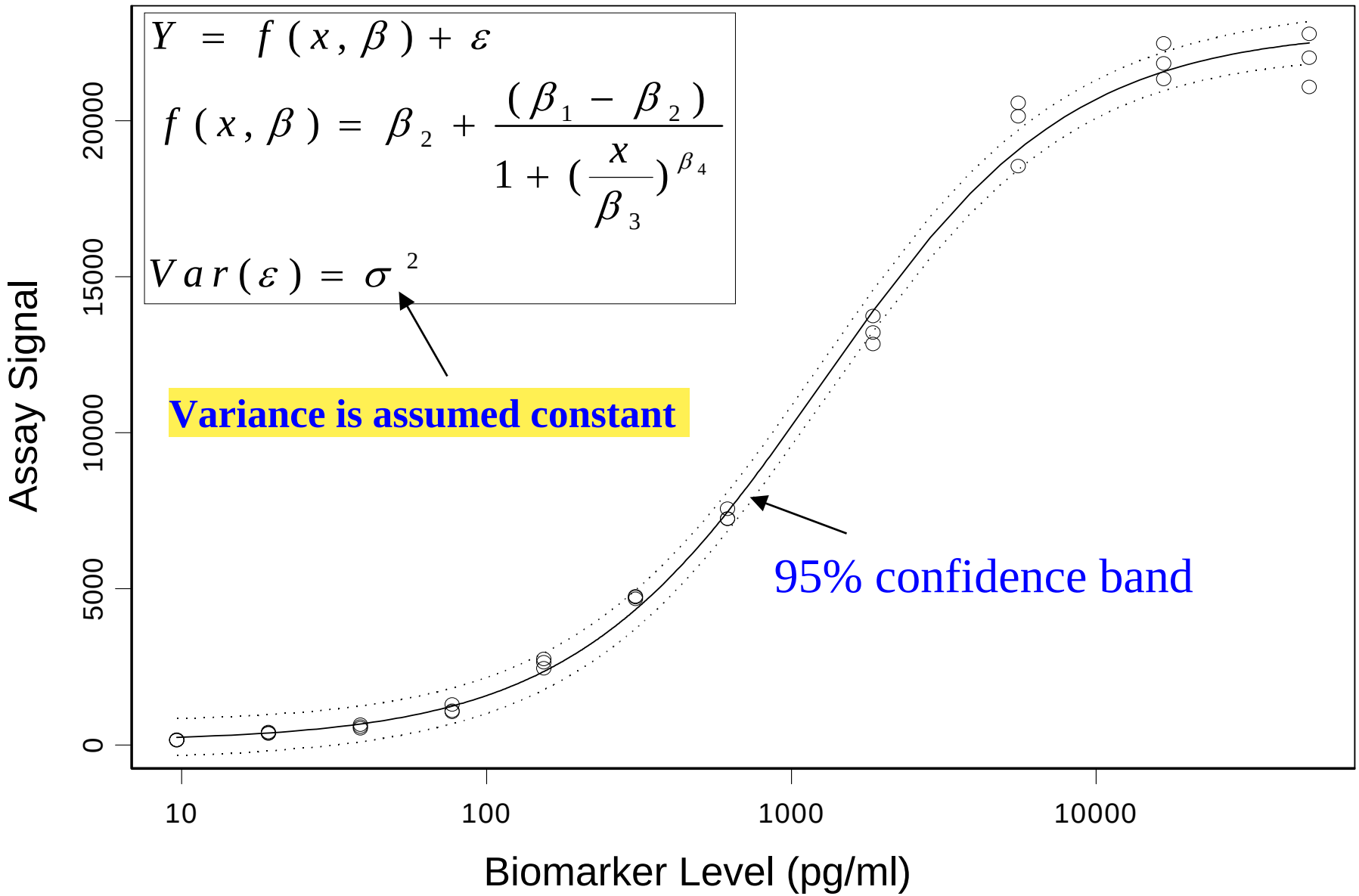
Heteroscedasticity is common with these data.

Ignoring this can affect method/assay performance.

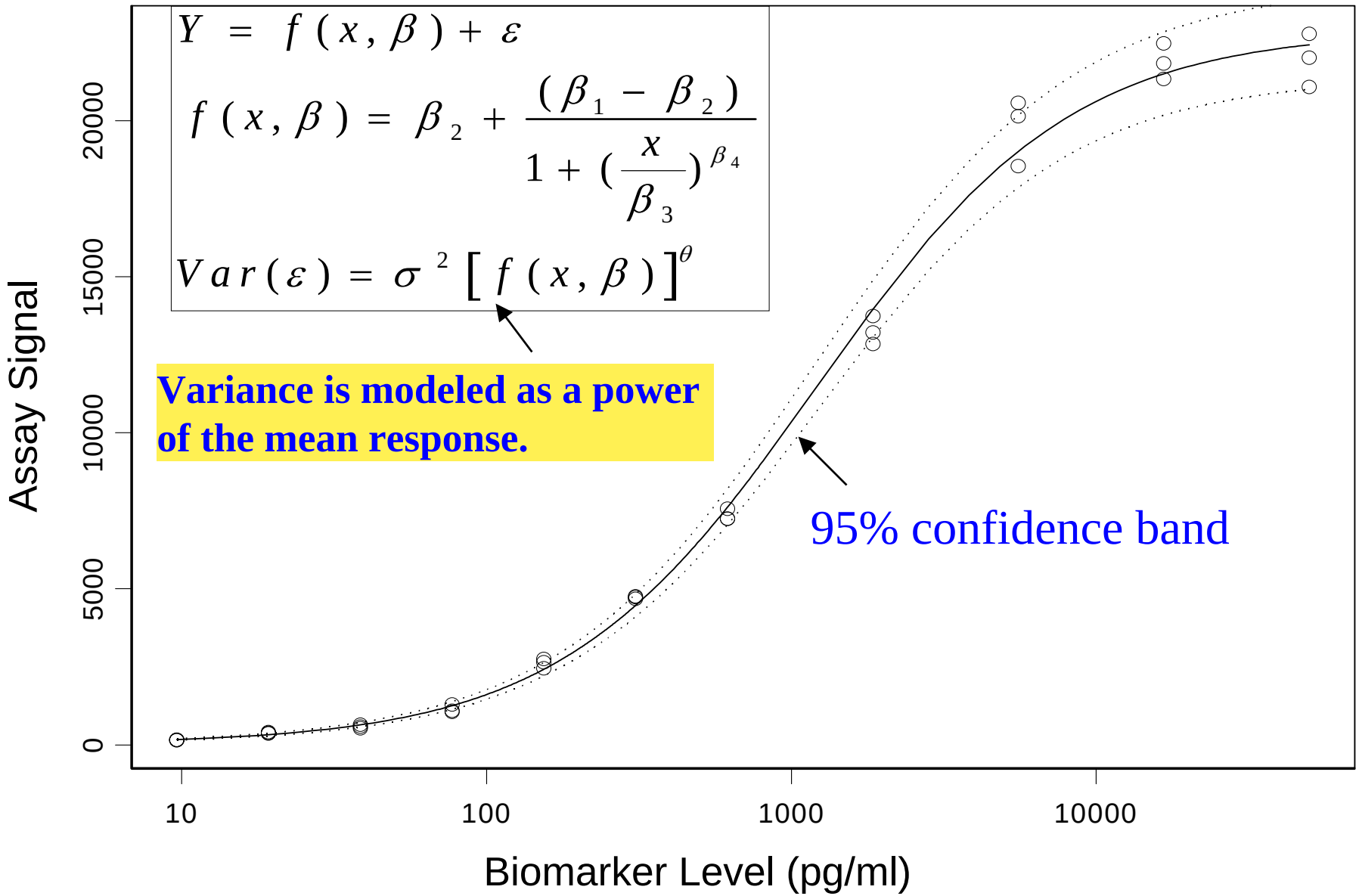
Popular methods for weighting:

- Model replicate SDs v.s. replicate means
- Pseudo-likelihood based methods (Carroll & Ruppert, 1988)

Standard Curve Without Weighting

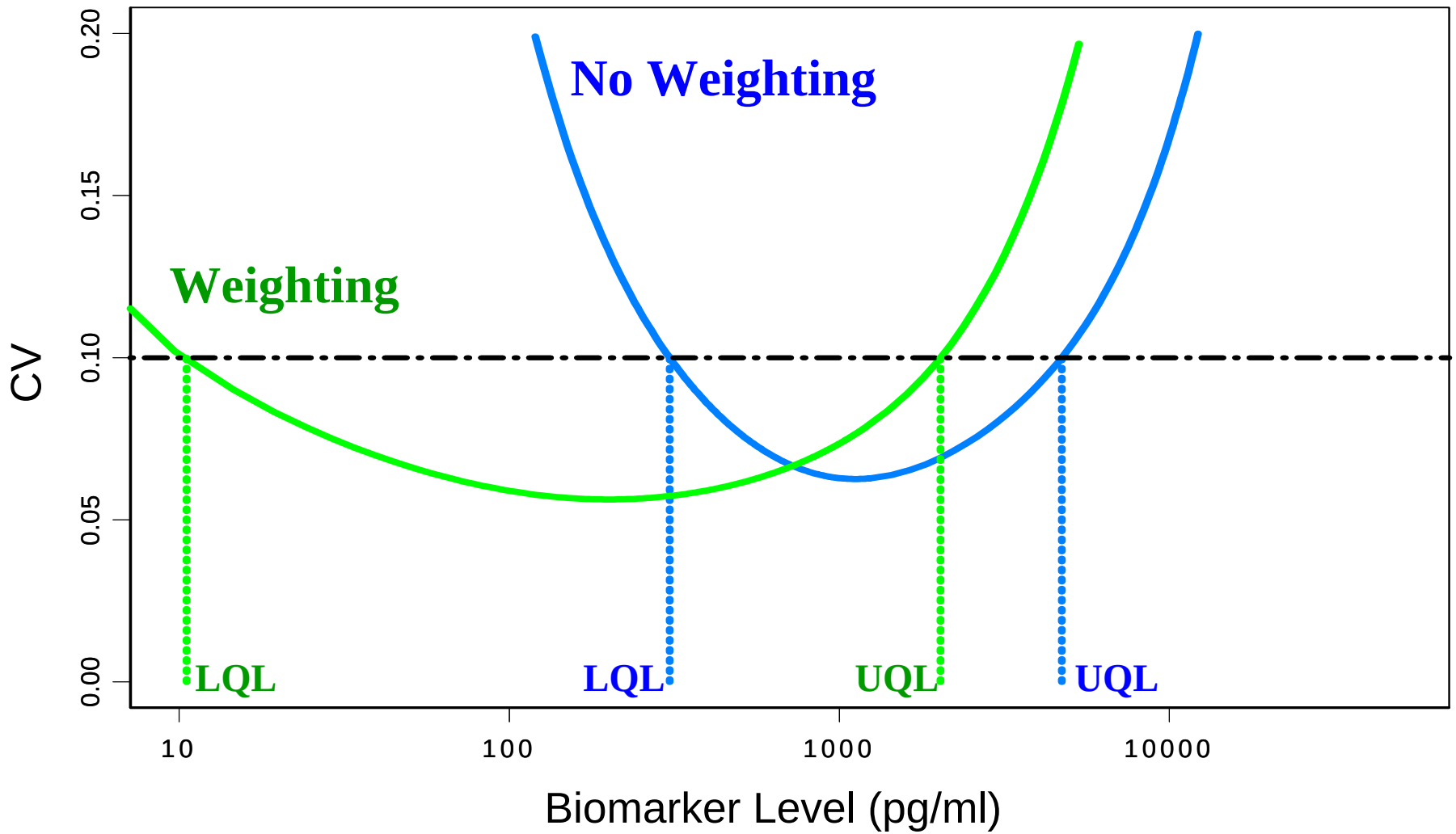


Standard Curve With Weighting



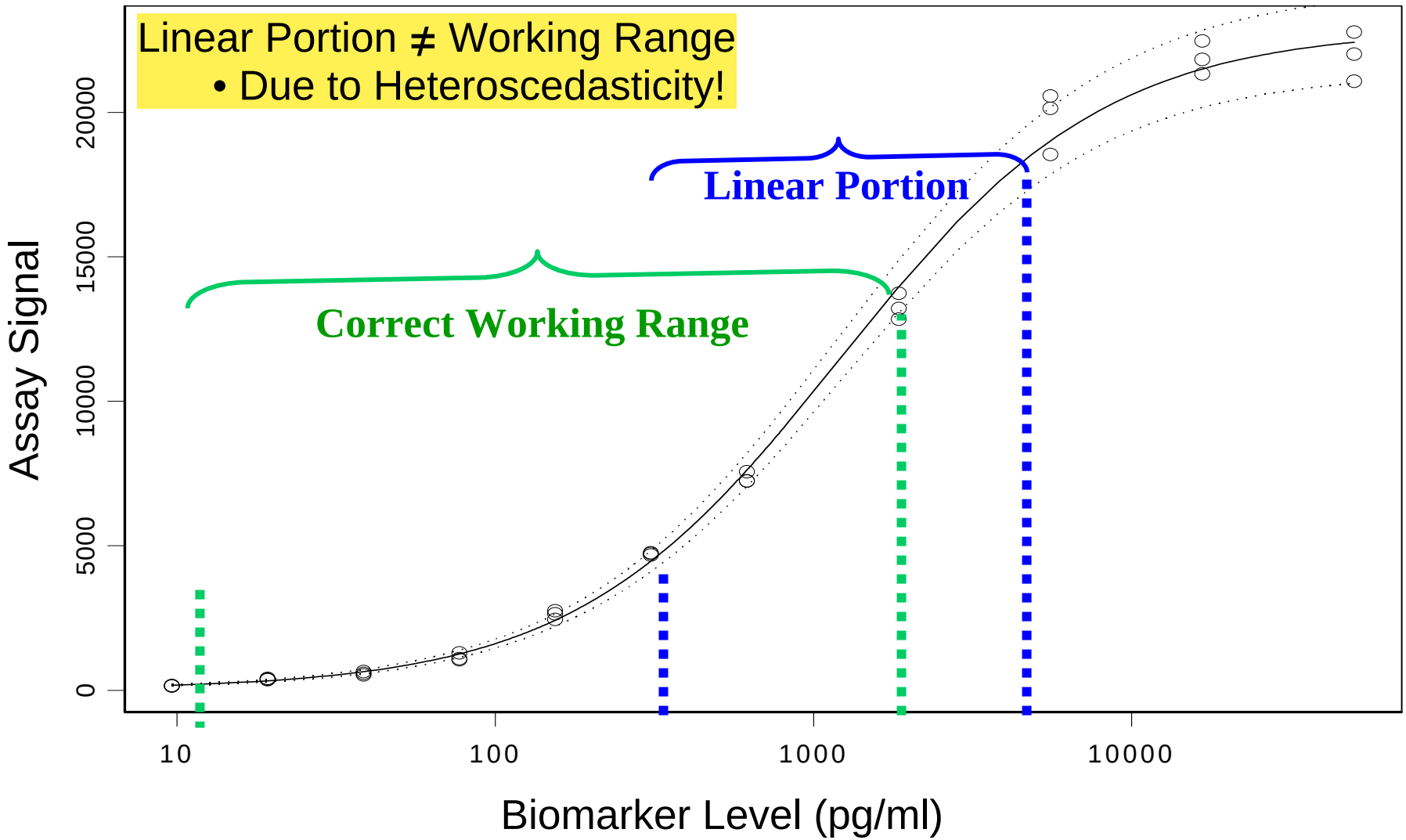
Impact of Weighting

Illustration: Precision Profiles



Standard Curve

Misconceptions About the “Linear Portion”



Note about Precision Profiles

Purpose

- Based only on standard curve, not validation samples
  - Doesn't take into account of all sources of assay variability.

1. Educational tool
   - Illustrates the impact of weighting.
2. Method Development & Optimization
   - Objective endpoint for selecting reagents & protocol design/optimization.
3. Preliminary "Marker" of Method Performance.
   - Helps assess whether the assay is ready for Validation!

Points 2 and 3 will be evident in the next section.

5. Assay/Method Optimization

Method Optimization & Assessment

Commonly used Endpoint:

- Assay's Signal to Noise Ratio (Signal Window, Z-factor)
- Appropriate for standard screening assays (binding, functional, etc.)
- **Inappropriate** for Calibration applications.

Optimize/Assess Calibration **Precision Profiles!**

Optimization Procedure

Screening Experiments:

- Identify all potentially important assay factors/variables.
- Run 2-level fractional-factorial experiments to determine the factors that are statistically important.

Optimization Experiments:

- Run 3-level experiments on these important factors using a Central-Composite or Box-Behnken type design.

Generate standard curves for each trial of the experiment.

Estimate the optimum using response surface analysis.

Assay Optimization Endpoints

Derived from Precision Profile

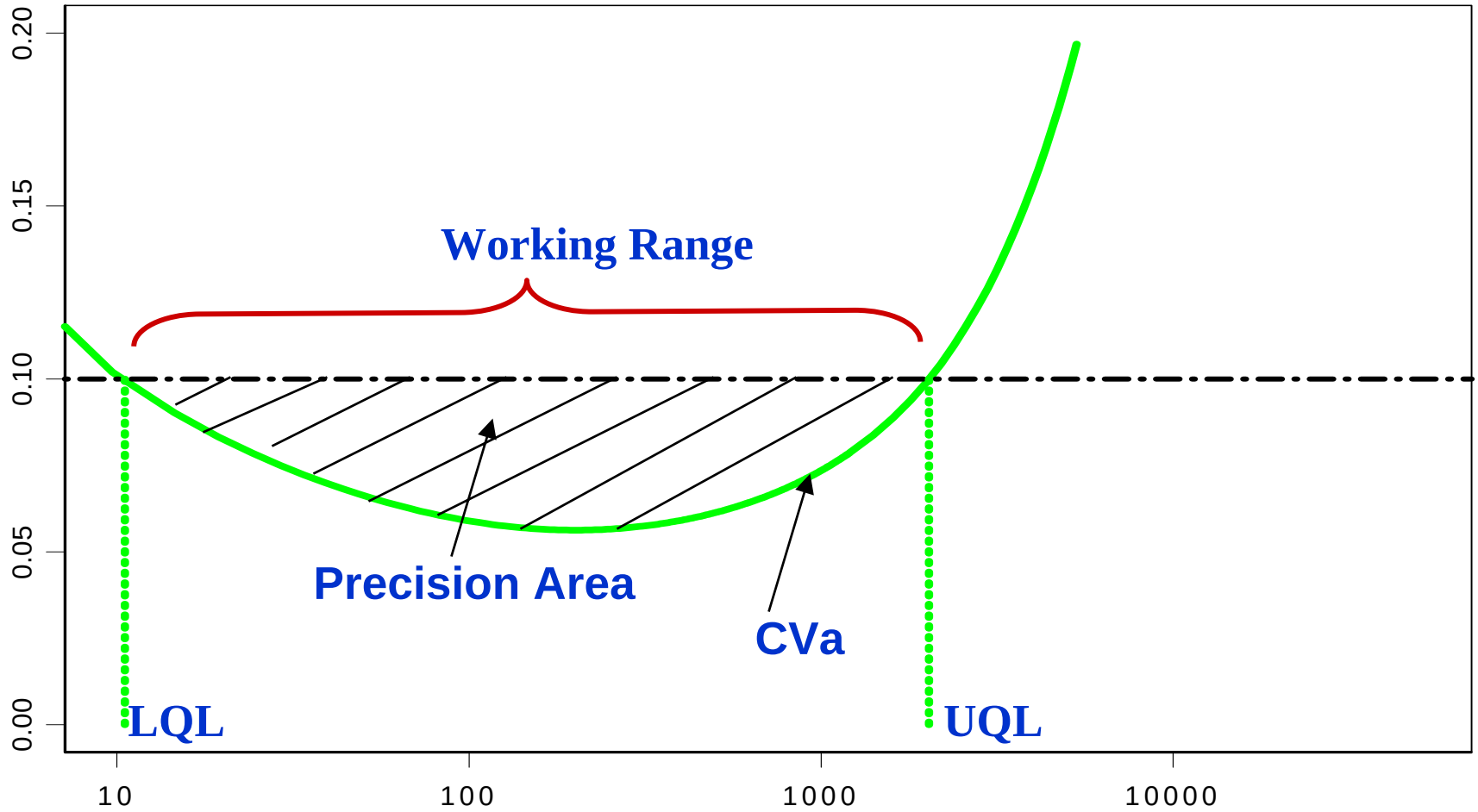
Standard Curves and hence Precision Profiles can be generated for each assay condition.

Key endpoints determined from Precision Profile.

- **Lower Quantification Limit (LQL)**
  - Lowest concentration where the precision profile intersects 20% CV.
- **Upper Quantification Limit (UQL)**
- **Working range (WR)** = Log_{10}(UQL / LQL)
- **CV_a** = Average CV within the working range.
- **Precision Area (PA)** = WR x (20% - CV_a).

Assay Optimization Endpoints

Derived from Precision Profile



Assay Optimization Endpoints

Derived from Precision Profile

Define important facets of the precision profile.

- LQL, WR, CV_a
- PA & UQL are contained in the above

Now set specification limits on these facets.

- C_{LQL}: largest acceptable value of LQL
- C_{WR}: smallest acceptable value of WR
- 15%: largest acceptable value of CV_a

Assay Optimization Endpoints

Composite Endpoint from Precision Profile

The **composite** optimization endpoint is

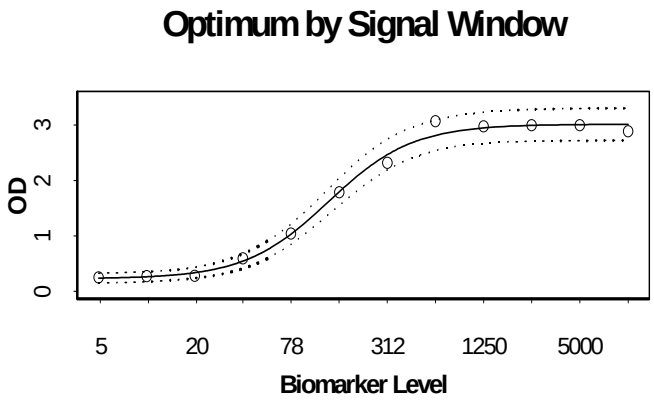
$$\left(\frac{\max(LQL)-LQL}{\max(LQL)-\min(LQL)}\right)^{w_1} I(LQL \le C_{LQL})$$

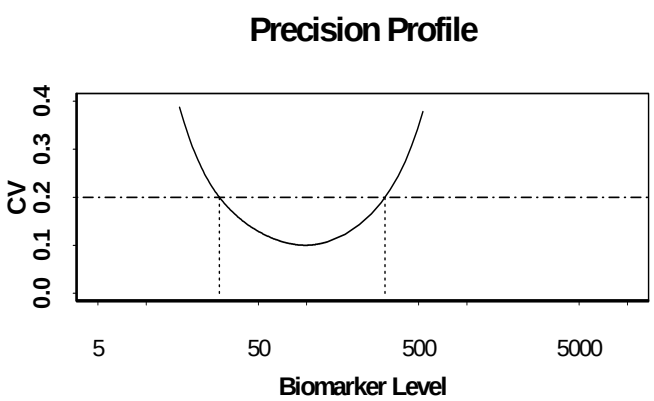
$$\times\left(\frac{WR-\min(WR)}{\max(WR)-\min(WR)}\right)^{w_2} I(WR \ge C_{WR})$$

$$\times\left(\frac{\max(CV_a)-CV_a}{\max(CV_a)-\min(CV_a)}\right)^{w_3} I(CV_a \le 15\%)$$

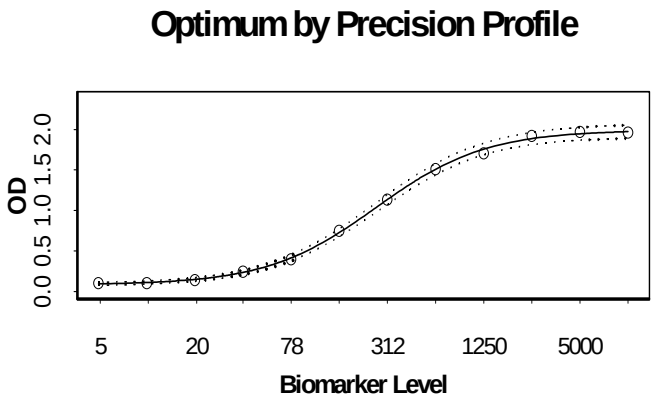
$w_1 + w_2 + w_3 = 1$ are the relative weights chosen by biochemist.

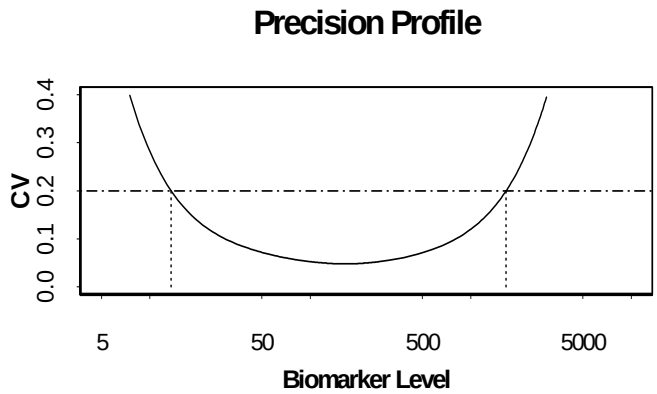
Example: Endpoints matter!





LQL = 24.2
UQL = 157





LQL = 13.6
UQL = 1662.3

Good Assay Signal does not imply good working range.

Improper optimization → Unacceptable Assay!

Focus must be on Calibration Range instead of Assay Signal.

Biomarker Methods

Are you ready for Validation?

Predict the Sensitivity & Range (QLs) using precision profiles.

- Does not take into account of other assay problems
  - Cross-reactivity, Interference, Operational factors, etc.
- So the predicted QLs are **optimistic**!

If the predicted QLs are not within the target range:

- **Not ready** for Validation!
  - Re-optimize some assay conditions using precision profile.

If the predicted QLs are within the desired range:

- **Ready** for Validation!

6. Pre-Study & In-Study Method Validation, Acceptance Criteria

Biomarker Analytical Validation

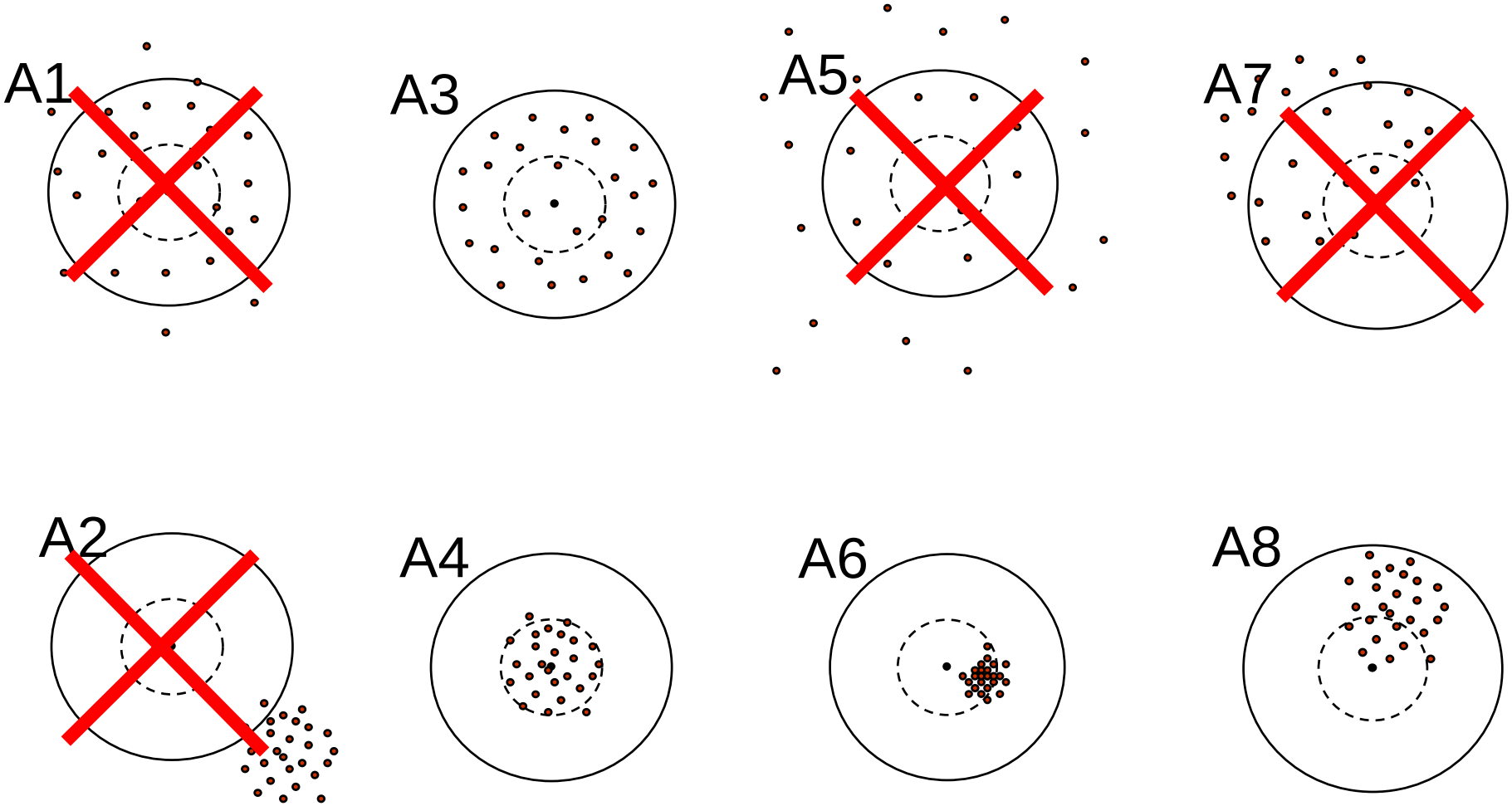
Objective

To demonstrate that an analytical procedure is acceptable (suitable, reliable, ...) for its intended application.

Implicit assumption: Acceptance criteria are defined prior to the initiation of development.

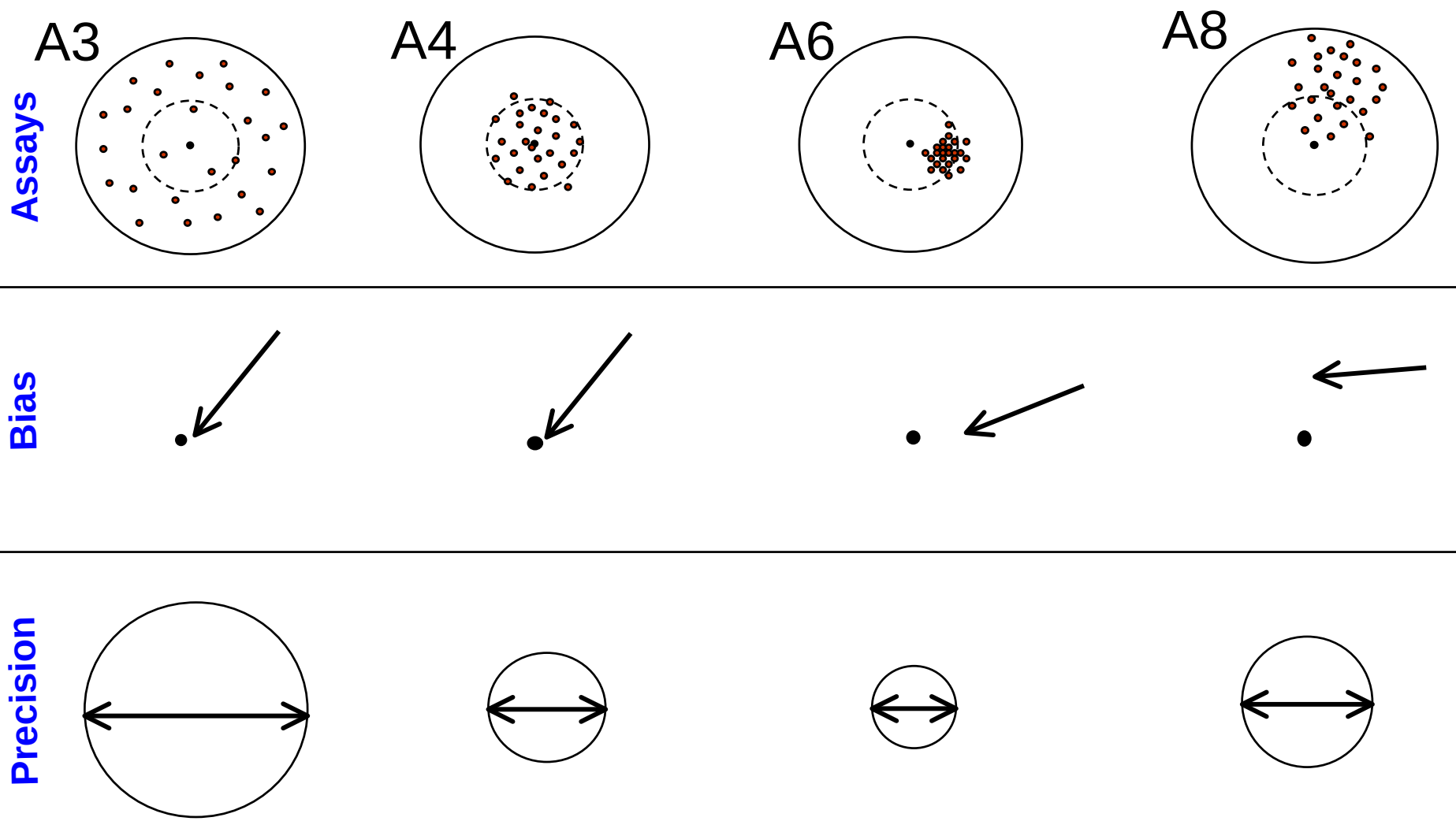
Which Assay is on Target?

Dartboard Analogy!



Which Assay is on Target?

Bias + Precision



Biomarker Analytical Validation

Parameters

Primary parameters

- Trueness (systematic error → bias)
- Overall Precision (random errors → variance)
  - **Intra-Run, Inter-Run, Analyst, Equipment, Plate, …**

✓ Total Error = |Bias| + Overall Precision

Derived parameters (from primary parameters)

- Sensitivity (LQL)
- Assay range (LQL, UQL)

Diagnostic parameters (provide insight into possible sources of systematic and random errors)

- Specificity
- Dilution linearity
- Parallelism
- Stability

Pre-Study Validation

Expectations for Biomarker Method Types

	Biomarker Method Types		
	Definitive & Relative Quantitative	**Quasi-Quantitative**	**Qualitative**
Trueness (Bias)	✓		
Precision	✓	✓	
Sensitivity	✓ LLOQ	✓	✓
Specificity	✓	✓	✓
Dilution Linearity	✓		
Parallelism	✓		
Assay Range	✓ LLOQ / ULOQ	✓	
Standard & Reagent Stability	✓		
Matrix Stability	✓	✓	✓

Reminder: Focus of this talk is on Definite & Relative Quantitative Methods

Pre-Study Validation

Experiment

Generate these data from each of 6 independent runs.

- Standard Curve: 8-12 pt, triplicates
- Validation Samples: 6-8 concentrations, > 2 replicates
  - Independent samples spiked with nominal amount of analyte.
  - 2 conc near desired LQL, 2 near desired UQL, and 2-4 within the range.

Consider other sources of variation in the design.

- Analyst, Equipment, Plate, Vendor, etc.

Pre-Study Validation

Data Analysis

Perform Variance Component Analysis

- JMP, Proc Mixed in SAS, LME in S-plus, etc.

Estimate Bias, Overall Precision, Total Error.

- Investigate each component of systematic and random errors.

Determine Sensitivity and Assay Range.

- LQL, UQL

Confirm/Finalize the model for Standard Curve.

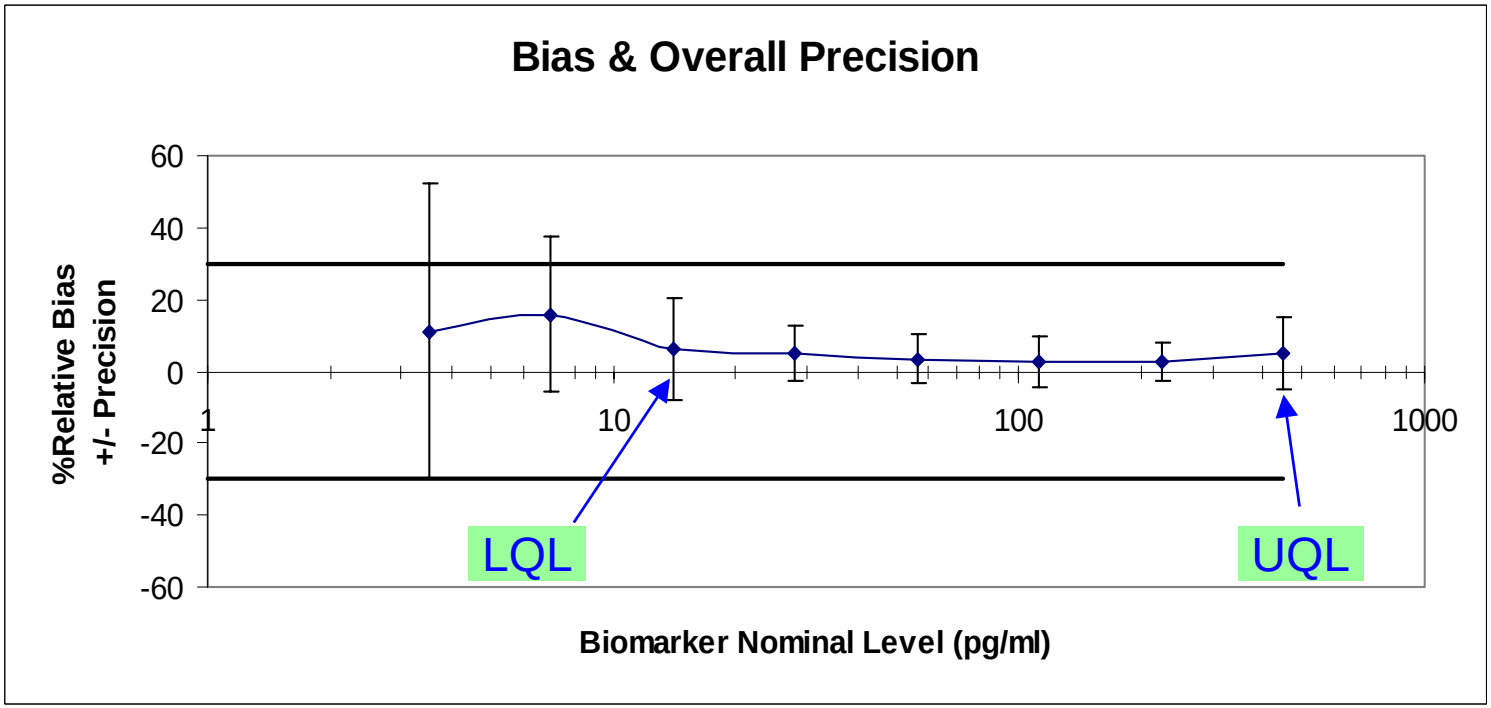
- Compare popular models, weighting methods, etc.
- Select/Confirm the optimal model based on the Total Error, Sensitivity, etc.

Assess Dilution Linearity

- Determine Maximum Tolerable Dilution.

Pre-Study Validation: Illustration

Bias, Precision, Total Error, Sensitivity



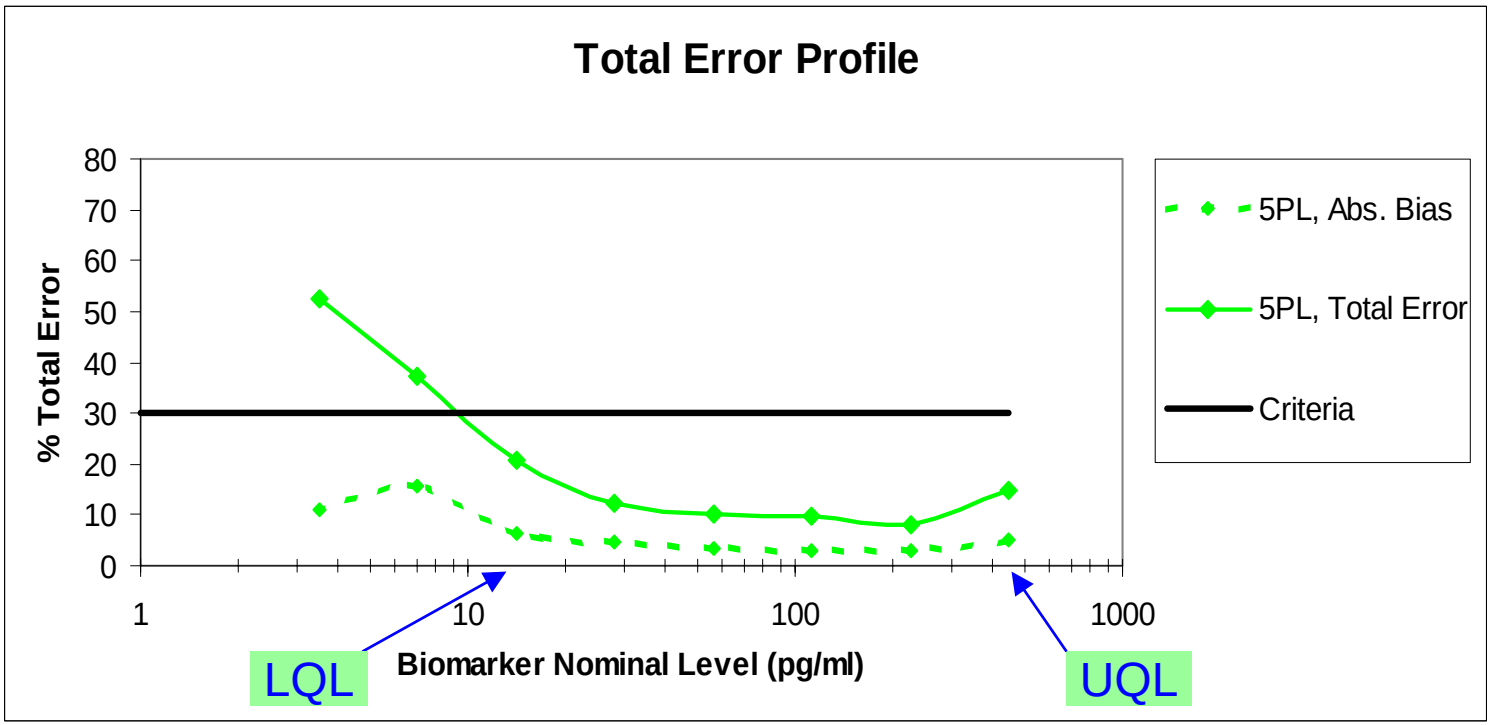
Data from a pre-study validation experiment – Biomarker (IL-6) ELISA.

Each data point represents the Mean %Bias across 6 runs.

Error Bars represent the Overall Precision at each nominal level.

Pre-Study Validation: Illustration

Bias, Precision, Total Error, Sensitivity

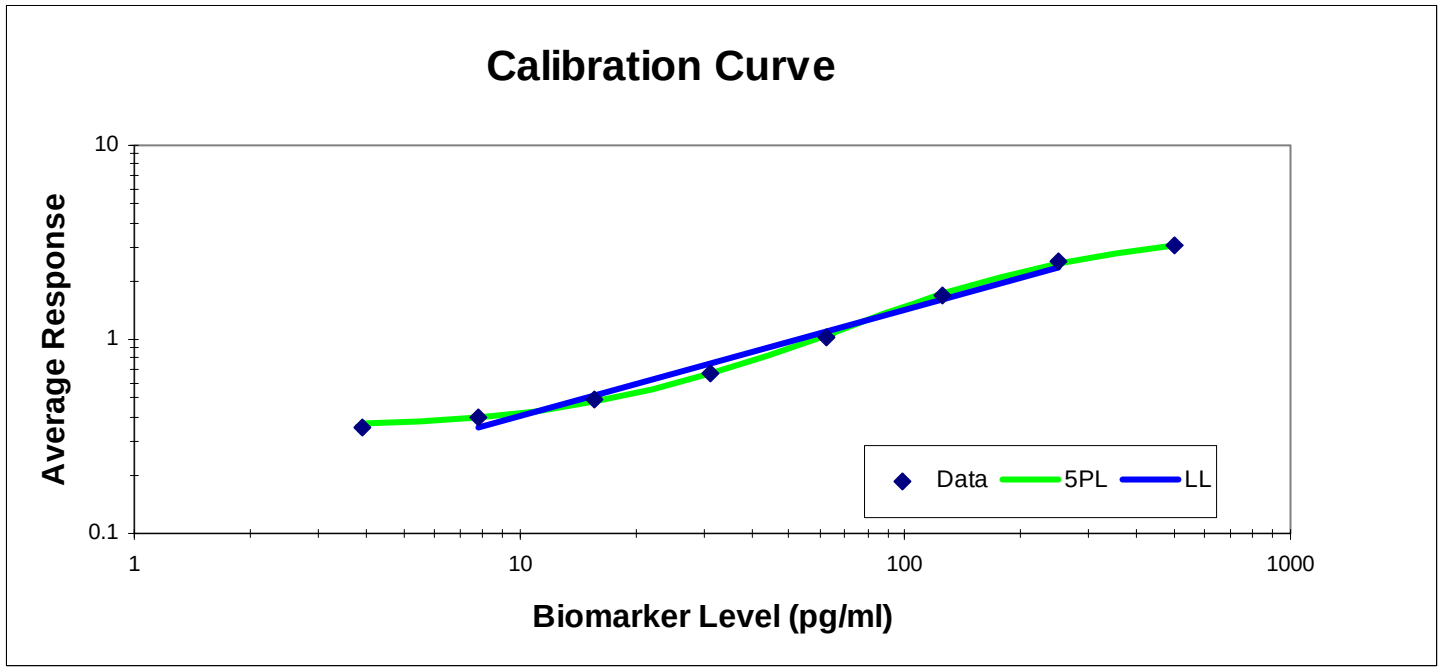


Data from the same pre-study validation experiment (previous slide).

Plotted differently to represent the Total Error, with |Bias| & Precision

Pre-Study Validation: Illustration

Finalize/Confirm Model Selection



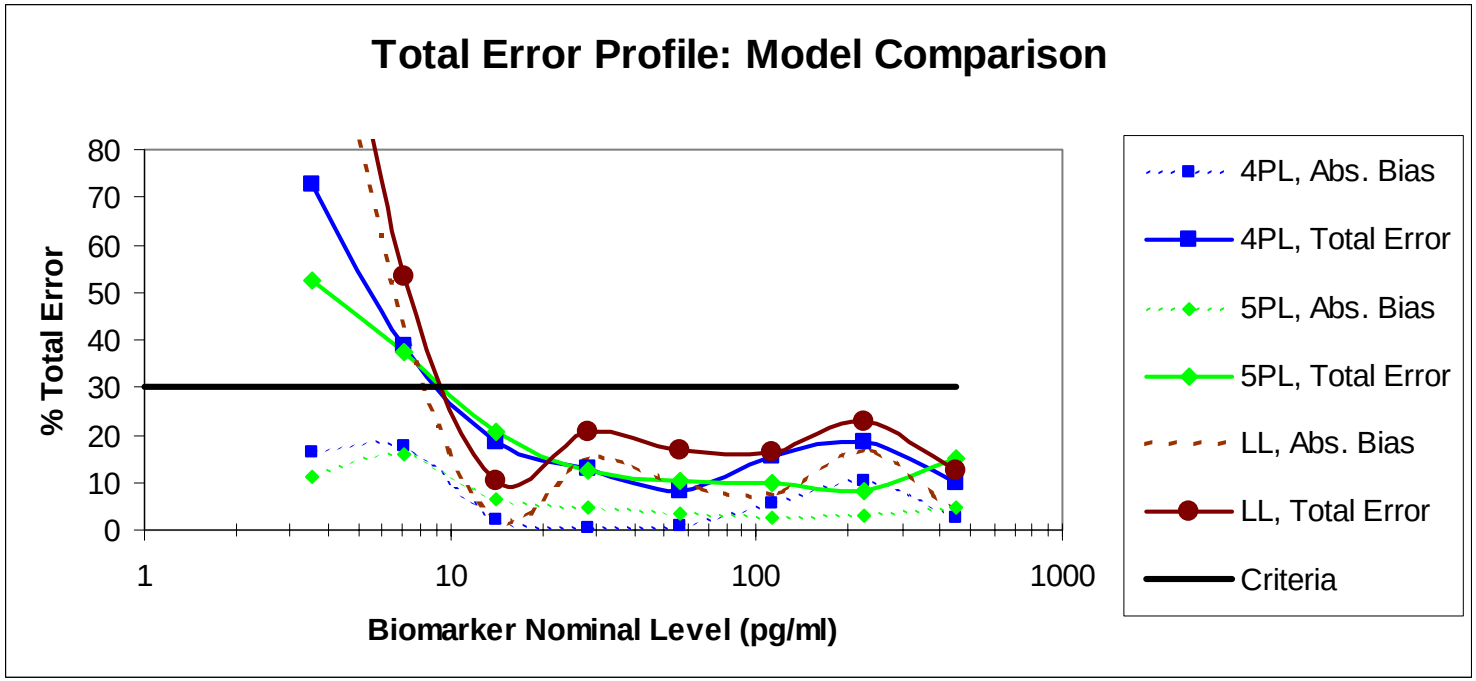
LL: Linear Model in log-scale (R^2 = 98.5%)

- Note that R^2 is commonly reported by Lab software

4/5PL: Four/Five-Parameter Logistic Model

Pre-Study Validation: Illustration

Finalize/Confirm Model Selection



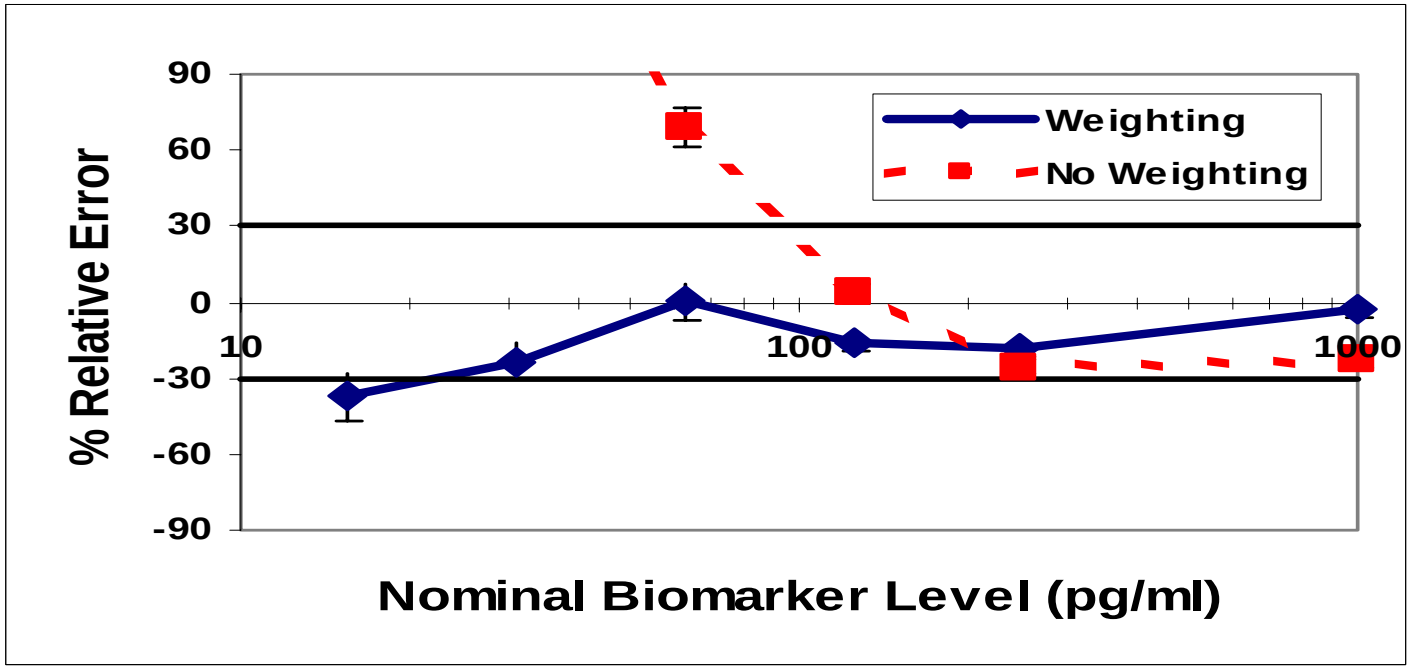
5PL vastly better than LL, and slightly better than 4PL.

- ***With respect to Total Error, Bias and Precision***

This confirms that for this particular assay, 5PL is the optimal choice for the in-study (production) phase.

Pre-Study Validation: Illustration

Finalize/Confirm Weighting



Data from a pre-study validation experiment.

Assay characterization is greatly impacted by weighting.

The optimal weighting factor estimated from the pre-study validation phase can be used to "fix" weights for the in-study (production) phase.

In-Study Validation/QC

"4-6-x Rule"

QC Samples in each run:

- 3 levels (typically low, mid, high) in 2-3 replicates.

4-6-x Rule:

- 2/3rd of all the samples must be within x% of the nominal.
- Half the samples at each level must be within x% of the nominal.

The choice of "x" varies across applications and formats.

- Typically, for biomarker immunoassays, x = 30%.

Parallelism of the test samples must be assessed at different points during the production phase.

7. Acceptance Criteria

Biomarker Analytical Validation

Acceptance Criteria (Immunoassays)

Characteristic	Pre-study Validation	In-study Validation
Trueness (%Relative Bias)	**≤ ± 20 (± 25 at LQL)**	**-**
Overall Precision (%CV)	**≤ 20 (25 at LQL)**	**-**
Total Error	**≤ 30%**	**"4-6-30" rule**

Total Error = |%Relative Bias| + Overall Precision (%CV)

DeSilva, et al: *Pharm Res* 20(11): 1885-1900, 2003.

Lee, et al.: AAPS Biomarker Method White Paper*, in preparation*

Comments on the Acceptance Criteria

Pre-Study & In-Study criteria "appear" to be consistent because

$$\text{Total Error} = \left|X_i - \mu\right| < 30\%$$
$$\Rightarrow \left|(X_i - \overline{X}) + (\overline{X} - \mu)\right| < 30\%$$
$$\Rightarrow \left|\text{Bias} + \text{Precision}\ (\sigma)\ \right| < 30\%$$
$$\Rightarrow 2/3\text{rd of the results are within } 30\%$$
$$\rightarrow 4-6-30 \text{ rule, as n} \rightarrow \infty$$

This doesn't hold as n is small and Total Error is just an estimate!

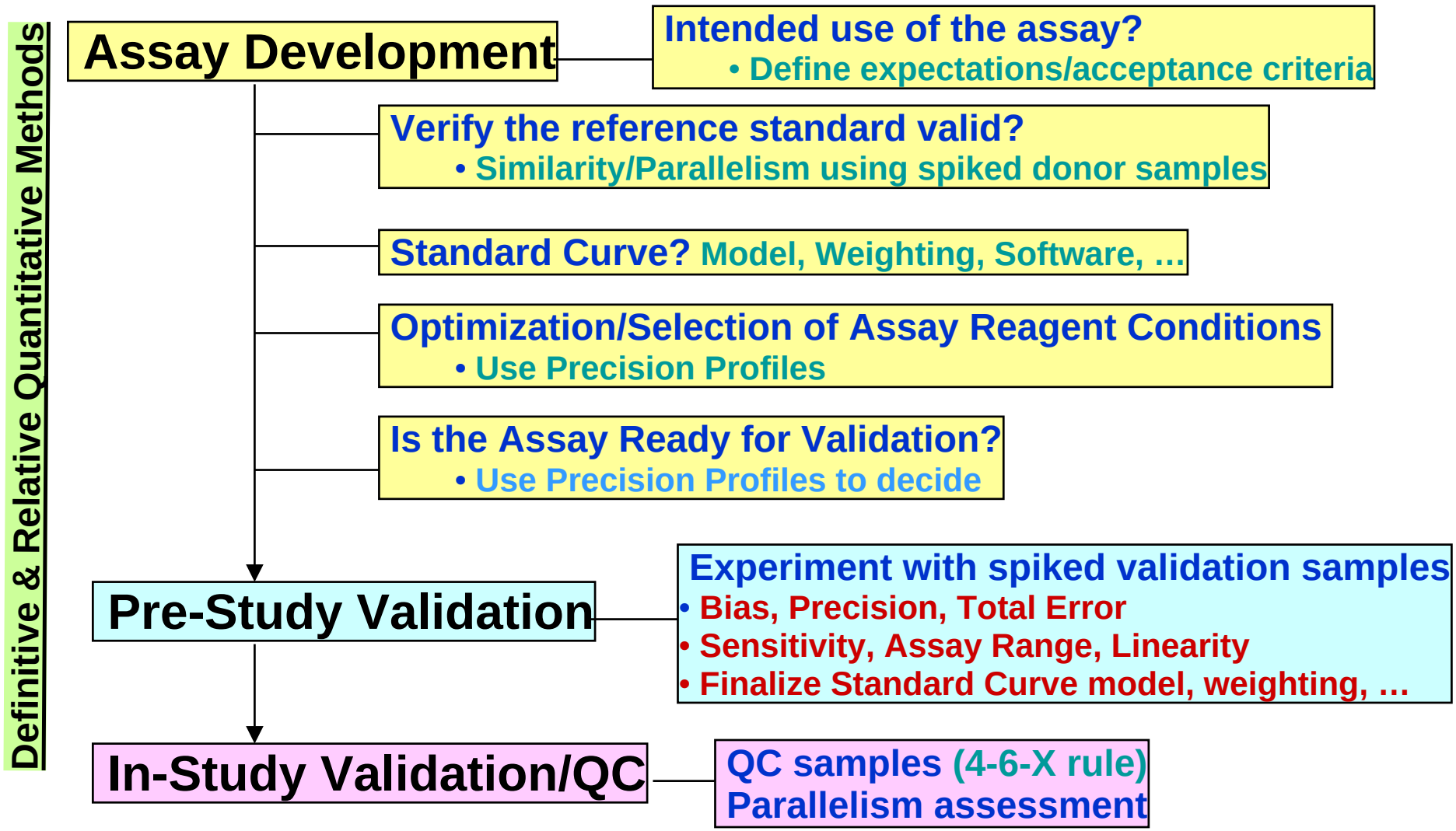
So for true consistency with the 4-6-x rule, we need to have Total Error < 30% - γ% as the pre-study criteria!

- Choice of γ depends on level of uncertainty in the estimate of total error.

8. Summary

Summary Flow Scheme

Statistical Thinking all the way!



Use statistics to provide estimates of errors. Statistics do not directly tell you whether the method is acceptable.

- Westgard, 1998

Key issues in Biomarker Method Development & Validation are governed by a combination of

Statistical + Practical + Biological Thinking!

Suggested Reading

- Belanger, B.A., et al.: D.M., *Biometrics*, 1995
- Callahan JD and Sajjadi NC: *Bioprocessing J (Apr/May)*: 1-6, 2003.
- Carroll, RJ & Ruppert, D: *Chapman & Hall, New York*, 1988
- DeSilva, et al: *Pharm Res* 20(11): 1885-1900, 2003.
- Hartmann C, et al: *Anal Chem* 67: 4491-4499, 1995.
- Hubert H, et al: *Analytica Chemica Acta* 391: 135-148, 1999.
- Karnes HT and March C: *J Pharm Biomed Anal* 10-12: 911-918, 1991.
- Kringle R, et al: *Drug Information Journal*, Vol. 35, 1261-1270, 2000.
- Mire-Sluis et al.: *J Immunological Methods*, 2004.
- O'Connell, M.A., et al.: *Chemo Intel Lab Systems*, 20, 97-114, 1993
- Plikaytis BD, et al: *J Clin Microbiol* 32(10): 2441-2447, 1994.
- **AAPS Biomarker Method White Paper*, will be ready by late 2004***