Bioassays: big challenges yield interesting design and analysis methods

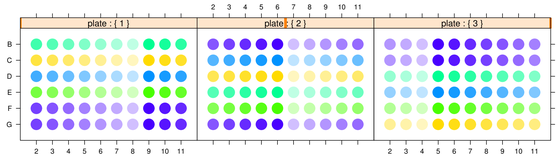
David Lansky, Ph.D. (david@precisionbioassay.com) precision bioassay inc Burlington, Vermont, USA

1. Abstract

Bioassays bring practical and statistical challenges including complex designs, non-linear responses, multiple sources of variance, non-additive effects, and more. Designs that are practical in the laboratory, are amenable to adoption of laboratory robotics, and support randomization are useful, but statistically complex. While non-linear mixed models help address many issues, they are sensitive to outliers and come with a need (particularly for complex designs) to choose among many reasonable candidate random effect models. We report some success with strategies for fixed and random effect model selection as well as ways to report results (graphically and with variance components) that help bioassay analysts monitor sources of variation that are known to be important. We will also describe several promising areas for additional research.

2. Common bioassay designs, and their challenges

- Bioassay experiment design includes groups
  - Animal groups include cages, shelf, rack, room, and facility
  - Cell assay groups: cell prep., plate, row in plate, column in plate
  - Shared preliminary dilutions created nested groupings
  - Serial dilutions create special correlation structure in concentration
- Bioassays have factorial treatment design (sample X dilution)
- Sample and dose are often applied to nested or crossed groups
- Example: 3-plate strip-unit design with 2 rows for each sample/plate

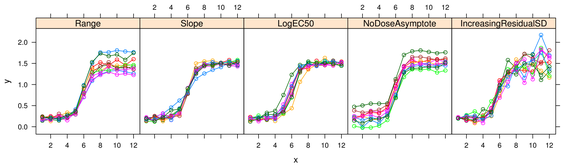


3. Outlier Management

- Regulatory likes automatic removal, statisticians not so much
- Quick investigation for cause and impact of suspected outliers
- Monitor location, analyst, etc. of candidate and removed outliers
- Periodic review for systematic causes
- Outlier Detection Considerations:
  - Transform to near-symmetric (ideally normal) before outlier detection
  - Pool across samples & blocks very helpful, but requires constant σ^2
  - Avoid outlier methods on sampleXdilution combinations (too many tests to preserve α & too few replicates to have enough power)
  - Promising outlier methods: ROUT, Rosner's, & maybe Hempel's
  - Avoid making data fit model: use non-parametric mixed models
  - Helpful to detect at multiple levels (mixed model standardized residuals)

4. Non-linearity, additivity, non-constant σ^2, & parameters

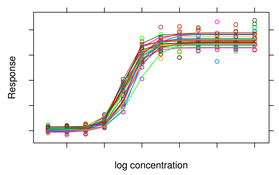
- Four parameter logistic response model a common (empirically good) choice
  - Many parameterizations, Ratkowsky & Reedy (1986) recommend a few
  - Variation associated with some parameters is not additive to response
- Full or unrestricted model: $y^* = \frac{A_i}{1+e^{-B_i(\ln(x)-C_i)}} + D_i + \epsilon$ where y^* = transformed response, A = Response Range, B = "Shape" or "Slope", C = Ln EC50, D = No-dose Asymptote and i = sample index
- Characteristics of parameter and residual variances (last 2 additive):



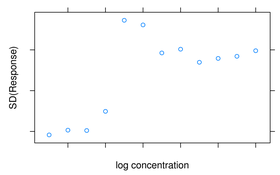
5. Approaches to Analysis

Approach	-	+
Straight lines fit to steep part after transform)	Extra data & analysis needed to assess asymptote similarity Truncation bias Narrow range Poor precision for effort	Simple Fixed dilutions in steep part (but not if $\sigma^2_{\text{Log EC50}}$ important)
Iteratively reweighted nonlinear least	Fails to address non-normality Estimating weights well is very challenging Performs poorly with multiple	Seems simple
Nonlinear mixed model	Good outlier management needed Random effects model selection challenging	Right idea Better precision and bias for similarity and potency Provides interpretable variance components for monitoring

Variation in log EC50 and increasing residual SD



SD estimated at each concentration



6. Fitting and Using Mixed Models

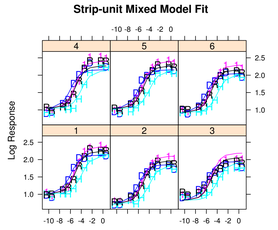
- Our strategy for random effect (RE) model selection:
  - List candidate RE models based on
    - Design: allow RE on each parameter with each unit
    - Constraint: any nested unit can only have RE of its parent unit
  - Build narrowed list of candidate RE models with good models from:
    - Fit a collection of assays together, use AIC/BIC to select good models
    - Fit dozens of assays, selecting RE in each assay; best fit by AIC/BIC
  - Routine fits: use all samples together, select from narrowed candidate set:
    - Each assay, each RE model: full, if similar: (partially) reduced
    - Choose best model by AIC/BIC among partially reduced models
  - Periodically revisit process of building list of candidate RE models
- Important benefits from this RE modeling approach
  - Better fits
  - Robust behavior of the system
  - Interpretable RE estimates help monitor process components
    - $\sigma^2_{\text{Residual}}$ Well variation (cells & dilutions): drives variance of potency
    - $\sigma^2_{\text{log EC50}_{sampleinblock}}$ Sample amount (initial dilution/block): $\sigma^2_{\text{potency}}$
    - $\sigma^2_{\text{dilution in block}}$ Within-block dilution; drives variance of potency
    - $\sigma^2_{\text{no-dose asymptote in block}}$ Not product, maybe location or cells

Reduced or constrained model for the fixed effects ($A, B, C,$ and D) and an example of a selected model for the random effects:

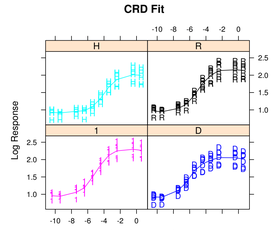
$$y^* = \frac{A + a_j}{1 + e^{-(B+b_j)(\ln(x)-C_i+c_{jkl})}} + D + d_{jk}\epsilon$$

$a, b, c,$ and d random effects, with indices: j = block, k = row in block, & l = column in block.

Strip-unit Mixed Model Fit



CRD Fit



7. Assay Acceptance or System Suitability

- Using full (unrestricted) model fit with selected RE model
- Important to use a modest number of criteria, with not-too-tight limits
- Good choices: ± 3 SD+ around long-term historical averages for:
  - Response range of standard (& control) curve
  - Log residual SD around fitted model
  - Shape or Slope parameter of standard (& control) ('B')

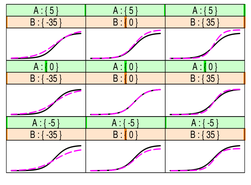
8. Sample Acceptance or Sample Suitability

- Important to demonstrate similarity
- Equivalence tests of all non-EC50 parameters
- BUT: equivalence tests will reliably fail with test EC50 outside assay range
- Strategies for setting equivalence bounds (from USP <1032>):
  - USP 2a: historical dists of standard yields bounds with known false non-similarity rate. Fails to assure adequate power. Largely undermines the goals of equivalence testing.
  - USP 2b: historical dist. of diffs between replicate standards. Like 2a.
  - USP 2c: distribution of differences between standards and known non-similar samples helpful; but, in practice, the collection of non-similar samples is very limited; hence, in practice rarely better than 2b.
  - USP 2d: sensitivity based, use knowledge of impact of various amounts of non-similarity to set equivalence bounds. Seems like the right idea, can address consumer's risk; details?
- Potency bias is sensitive to scaled shifts in non-similarity measures
  - Scaled shifts have consistent interpretation across assay systems
  - Positively correlated shifts in no-dose asymptote and range need a bound
  - Similarity equivalence bounds robustly limit (median) bias in potency
  - Important to have good power for most of the equivalence region

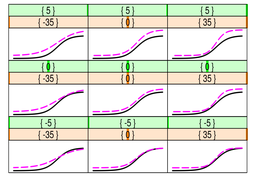
8.1 Scaled shifts have consistent meaning

- $\%\Delta_A = 100 \times (A_{\text{Test}} - A_{\text{Ref}}) / \overline{A^*_{\text{Ref}}}$
- $\%\Delta_D = 100 \times (D_{\text{Test}} - D_{\text{Ref}}) / \overline{A^*_{\text{Ref}}}$ (Not a typo)
- $\%\Delta_B = 100 \times (B_{\text{Test}} - B_{\text{Ref}}) / \overline{B^*_{\text{Ref}}}$ (* Long term averages)

No-dose Asymptote, 0% scaled shift

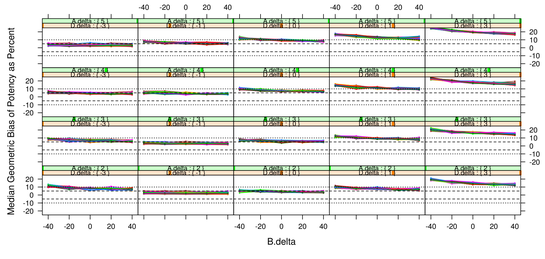


No-dose Asymptote, +5% scaled shift



8.2 Scaled shifts a robust source of potency bias

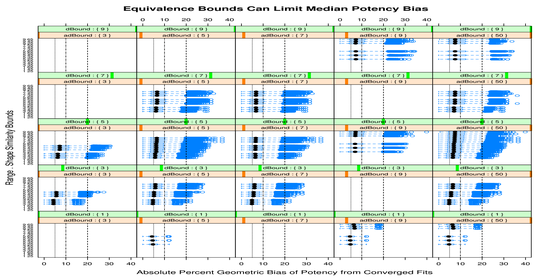
Potency bias due to non-similarity is robust to nDoses, dose range, SD



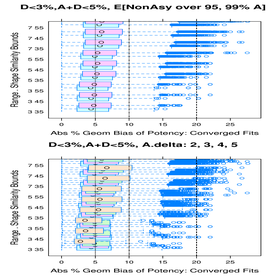
- D (no-dose asymptote) -6% to 6% (cols & lower header)
- A (range) 0% to 7.5% (rows & upper header)
- B (shape) -70% to 70% (x axis)

8.3 Equivalence bounds for similarity limit median potency bias

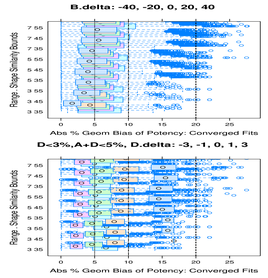
Equivalence Bounds Can Limit Median Potency Bias



D<3%,A+D<5%, E[NonAsy over 95, 99% A]



B.delta: -40, -20, 0, 20, 40

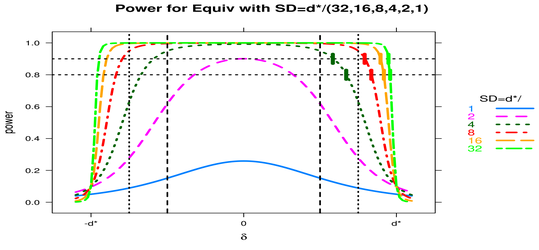


D<3%,A+D<5%, A.delta: 2, 3, 4, 5

D<3%,A+D<5%, D.delta: -3, -1, 0, 1, 3

8.4 Consider power for equivalence tests of similarity

Power for Equiv with SD=d*/(32,16,8,4,2,1)



- Minimal power at $\delta = 0$ risky because similarity decisions are very sensitive to assay precision and n
- Similarity equivalence bounds to prevent potency bias are demanding
- Decent power for most of similarity region: very demanding
- Suggestion: combine results across assays before assessing similarity

9. Recommendations

- Collaborate with labs to understand the designs they are actually using
- With labs: help them adopt randomization, good blocking, and robotics
- Potency bias limit for assay purpose drives similarity bounds
- Ensure good power for most of (better 3/4) the equivalence region
- Discourage plate specific similarity assessment

10. Acknowledgements

- Consulting clients
- USP and USP bioassay panel members
- Carrie Wager
- Ramiro Barrantes
- NSF EPSCoR
- NIH SBIR 3R44RR02198-03S1