New paradigm to derive and assess ADA cut points

Charles Tan, PhD

2019 Regulatory-Industry Statistics Workshop

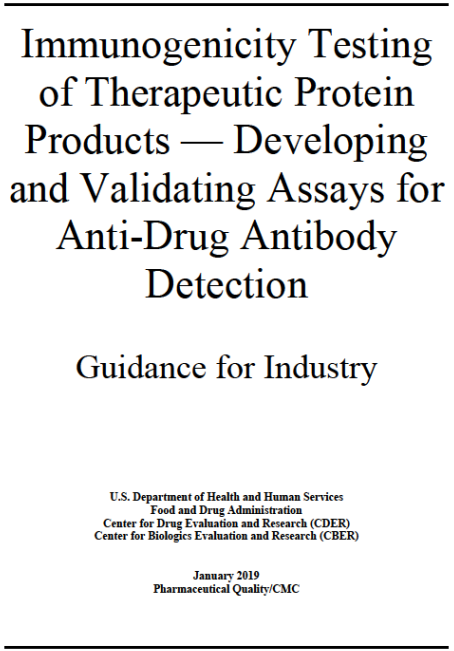
Outline

- Introduction and background
- The problem with box plot as outlier procedure
- Assay characteristic curves
- Pre-existing reactivity and Nonspecific binding
- The problem with estimating tail using middle
- Modern (and classic) nonparametric methods
- Graphs to assess cut points
- Summary of the new paradigm
- Q&A
- A few side notes

Background

- Immune responses to therapeutic protein products have the potential to affect product PK, PD, safety, and efficacy.
- The clinical effects of immune responses in subjects are highly variable, ranging from no measurable effect to extremely harmful.
- Detection and analysis of ADA formation is a helpful tool in understanding potential immune responses.
- Impact
  - During clinical trial: crucial for any therapeutic protein product development program.
  - Included in the prescribing information as a subsection of the ADVERSE REACTIONS section entitled *Immunogenicity*.

Guidance January 2019



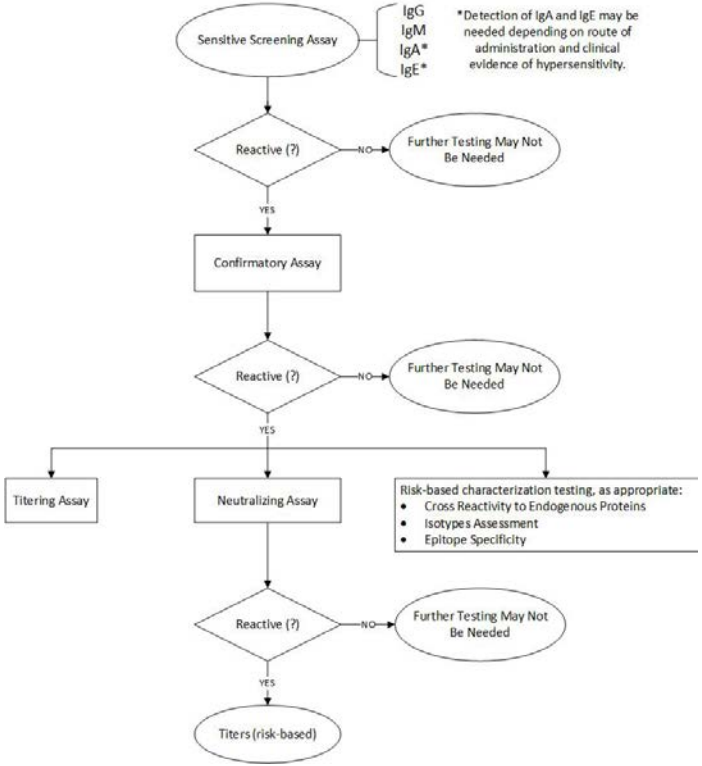
Immunogenicity Testing of Therapeutic Protein Products — Developing and Validating Assays for Anti-Drug Antibody Detection

Guidance for Industry

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)

January 2019
Pharmaceutical Quality/CMC

Multi-tiered approach



- This talk only focuses on
  - Tier 1: screening assay
  - Tier 2: confirmatory assay
  - Cut point determination
- Additional steps:
  - Titering assay: quantitation
  - Neutralizing assay: functional activity

Typical validation dataset

- Guidance: appropriate number of treatment-naïve samples
  - generally around 50, from the subject population
  - Each sample should be tested by at least two analysts on at least three different days for a total of at least six individual measurements.
- Reality: commercially procured (treatment-naïve) samples
  - Particular disease samples could be much harder to obtain than healthy volunteer samples
- Tier 1 and 2 cut points are calculated based on validation dataset
  - Tier 1 cut point: target 5% false positive rate
  - Tier 2 cut point: target 1% false positive rate

What an actual plate looks like

	PC Titer												Sample Treatment
	1	2	3	4	5	6	7	8	9	10	11	12	
A	4415	4311	147	145	142	145	145	148	142	144	152	151	No INH
B	1636	1656	145	144	139	137	144	145	138	137	148	148	INH
C	611	627	146	142	137	137	152	157	144	142	148	147	No INH
D	297	296	146	146	136	133	146	146	143	138	148	147	INH
E	188	186	2929	2965	135	141	143	139	142	138	3489	3470	No INH
F	157	148	135	132	134	135	139	133	134	137	146	141	INH
G	141	136	261	266	139	139	149	143	141	144	277	271	No INH
H	139	129	131	130	134	133	136	136	137	137	136	136	INH

	PC Titer												Sample Treatment
	1	2	3	4	5	6	7	8	9	10	11	12	
A	4363		145		144		147		143		152		No INH
B	1646				138		145		138		148		INH
C	619				137		155		143		148		No INH
D	296.5				135		146		141		148		INH
E	187		2947		138		141		140		3480		No INH
F	152.5		134		135		136		136		144		INH
G	138.5		264		139		146		143		274		No INH
H	134		131		134		136		137		136		INH

Current common practices

- Step 1: box plot to exclude outliers
  - Cut the right tail off before estimating the right tail
  - But the most informative data points for tail percentiles are in the tail
- Step 2: parametric approach to calculate cut points
  - Normal: mean + Normal constant * std
  - Lognormal: on log scale, mean + Normal constant * std
- Alternative robust: median + Normal constant * MAD
  - Still assume a symmetric distribution close to Normal
- Use the middle to predict the right tail with Normal constants

Pfizer’s goals

- Review the performance of current common practices
- Develop best practice for all Pfizer ADA assays
  - Robust enough to handle at least 80% of the assays consistently
    - Enable automation from instrument data to submission ready report
  - Convincing merit to regulatory agencies and industry peers

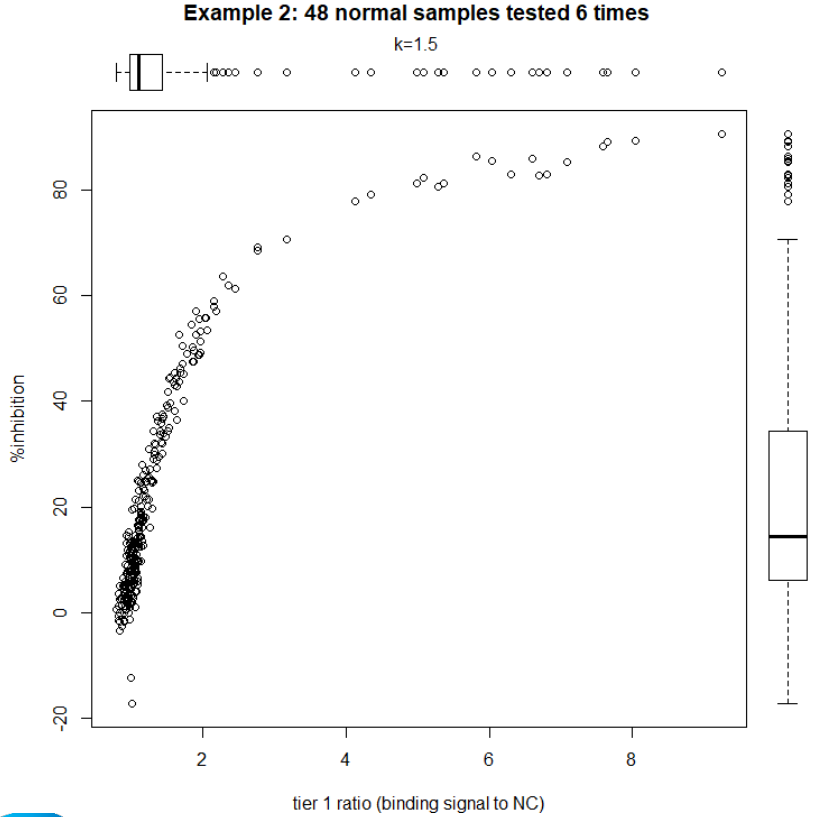
- Pfizer working group
  - Clinical Pharmacology: Daniel Baltrukonis, Sherry Cai, Puneet Gaitonde
  - Biomedicine Design: Boris Gorovits, John Kamerud
  - Statistics: Charles Tan, Gregory Steeno, Zhiping You

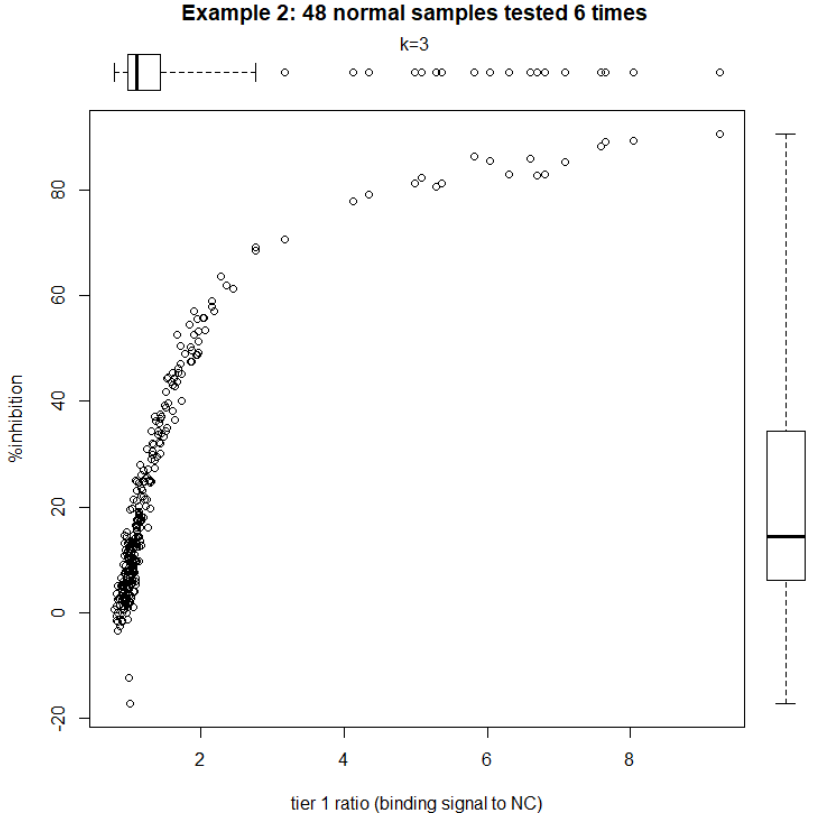
Conclusions of review

- Little differences among common methods in terms of final cut points
- Most differences are driven by differences of outlier procedures
- We don’t know whether we’re consistently good or bad
- “Colleagues across industry report problems with low cut point factors”

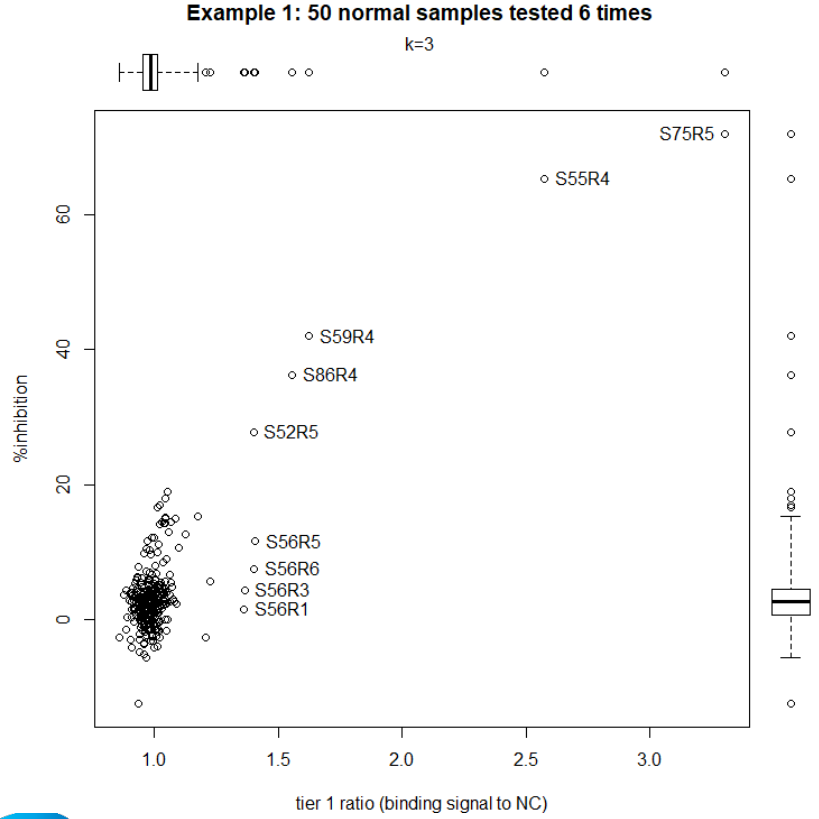
- My personal goal: graph, graph, graph
  - Bring scientific, clinical, regulatory context into the graphs to assess the cut points
  - Put the shape of the data on the table, not just the final cut point numbers

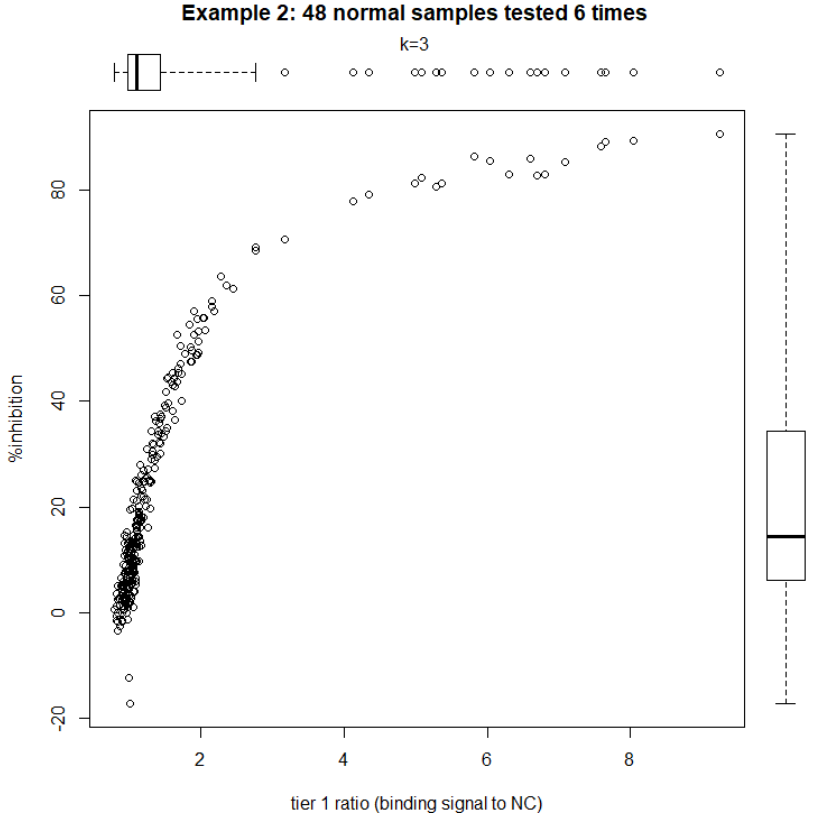
Outlier procedure drives the cut point





Clean too much and too little





The problem with using box plot to exclude outliers

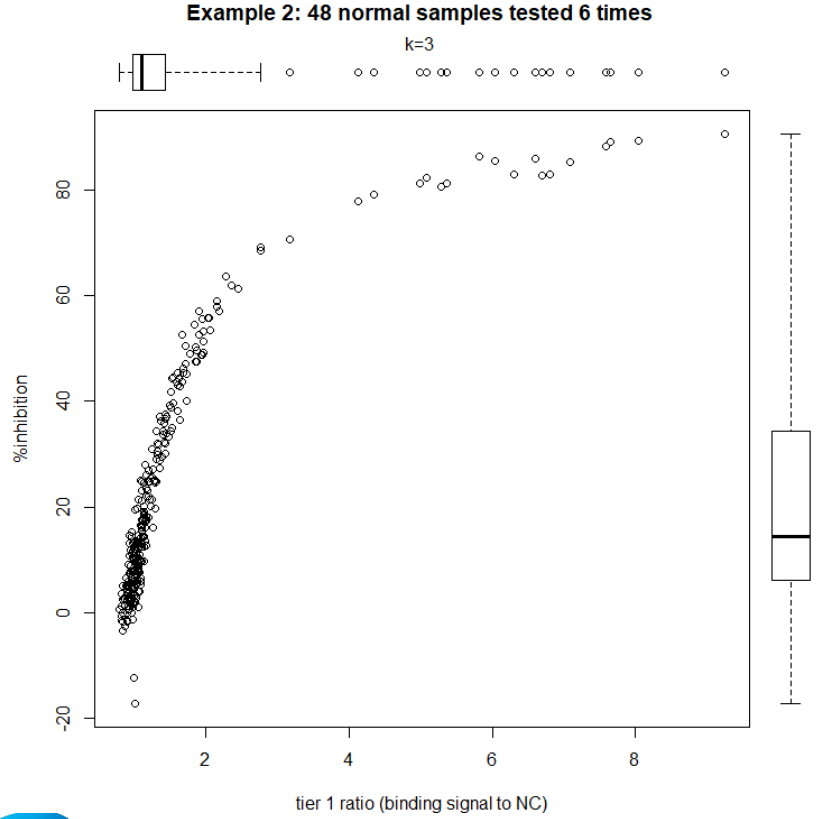
- k=1.5 vs 3 drives the cut point
- Vagueness and logic inconsistency in claiming “biological” versus “analytical” outliers
  - For some samples, every measurements are excluded
  - For others, only some of the measurements are excluded
  - The data points deemed outliers in tier 1 could be retained in tier 2, but the binding signal is part of %inhibition calculation
- Clean too much for clean dataset while clean too little for messy dataset

- Box plot is not a proper statistical outlier detection procedure. Its original purpose is to display distribution
  - “Out of the fence” doesn’t necessarily mean “outliers”
  - Skewed distributions would have many data points “out of the fence” by design

Insights that stimulated the new paradigm

- Need to use the purpose in “fit-for-purpose” to evaluate the performance of cut points
  - Clinical/regulatory purpose of the ADA assays: sensitivity given specificity constrains
  - But cannot be tighter than measurement variability can support
  - Step outside the statistics based on naïve samples alone
- Due to biological complexity of human populations, the naïve population is often not homogenous, but a mixture
- Statistically, determination of cut point is fundamentally to estimate tail percentiles
  - In contrast, the estimand of most statistical inferences in literature is the location of the distribution, or to a much lesser degree, the spread
- Nonparametric approach is a natural candidate for inference on tail percentiles
  - Responsive to the actual tail distribution
  - Robust to any distribution or mixture
  - The challenge: naïve samples with “pre-existing reactivity” that may inflate the cut points

Common pattern that was ignored



- It has long been noted that %inhibition is “correlated” with tier 1 ratio
- In fact, the relationship is much stronger than mere “correlation”
- A particular pattern, almost functional relationship, shows up repeatedly

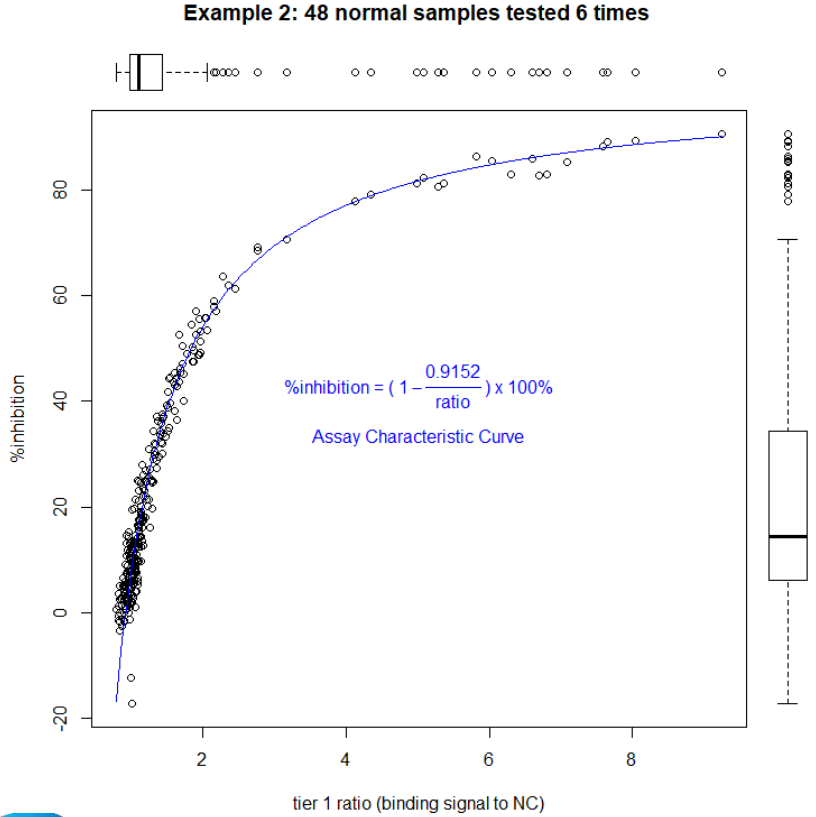
Assay Characteristic Curve

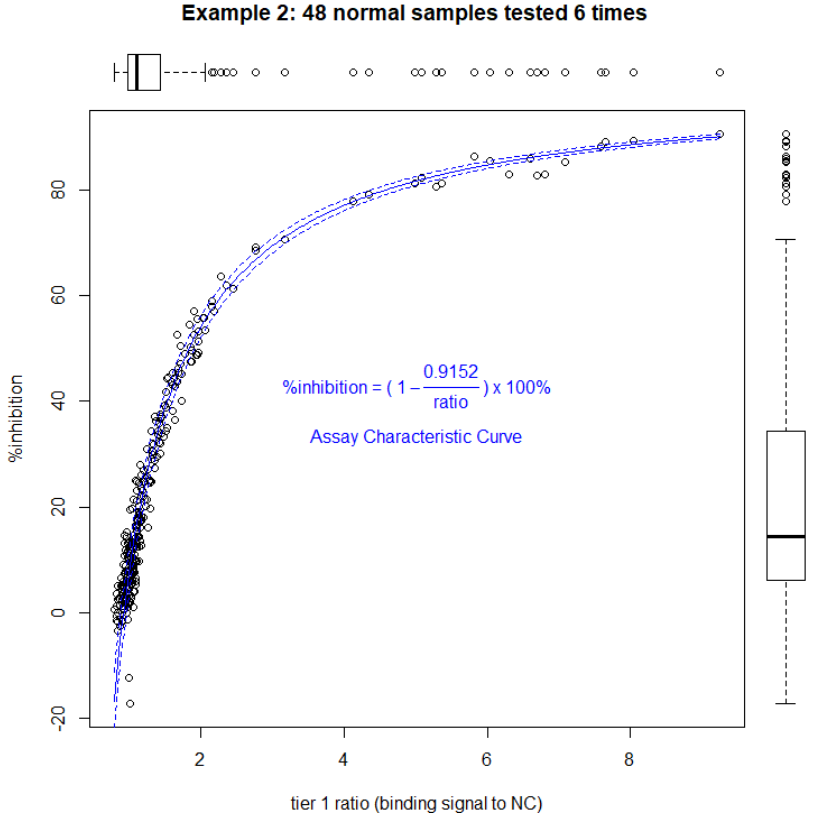
- NC: the Negative Control response on each plate
- S_1: the binding response (tier 1 raw data)
- S_2: the inhibited response (tier 2 raw data)

- Tier 1: $ratio = S_1/NC$
- Tier 2: $\%inhibition = \left(1 - \frac{S_2}{S_1}\right) \times 100\% = \left(1 - \frac{S_2/NC}{S_1/NC}\right) \times 100\% = \left(1 - \frac{S_2/NC}{ratio}\right) \times 100\%$
- Ideal: S_2/NC should be 1 regardless whether the sample is positive or negative
- Reality: S_2/NC is often tightly distributed around a constant near 1, and the constant is driven by
  - Choice of negative pool
  - Amount of inhibitor

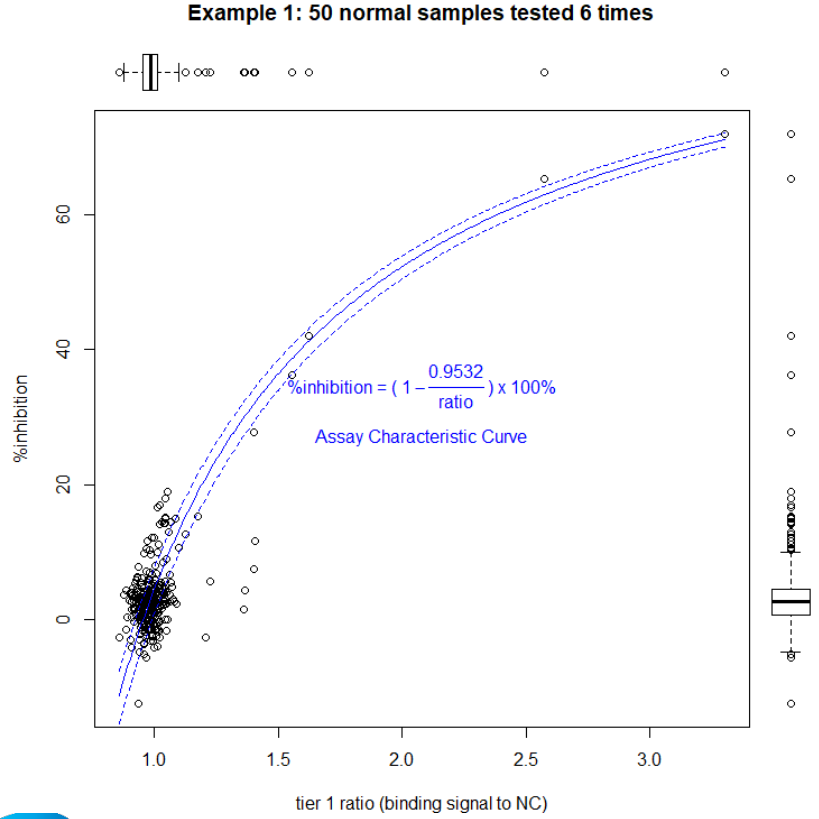
- **Assay Characteristic Curve**: $\%inhibition = \left(1 - \frac{h}{ratio}\right) \times 100\%$
  - Recognize the algebraic relationship and scientific context
- For each assay, we estimate the constant h by $median(S_2/NC)$: statistical fit

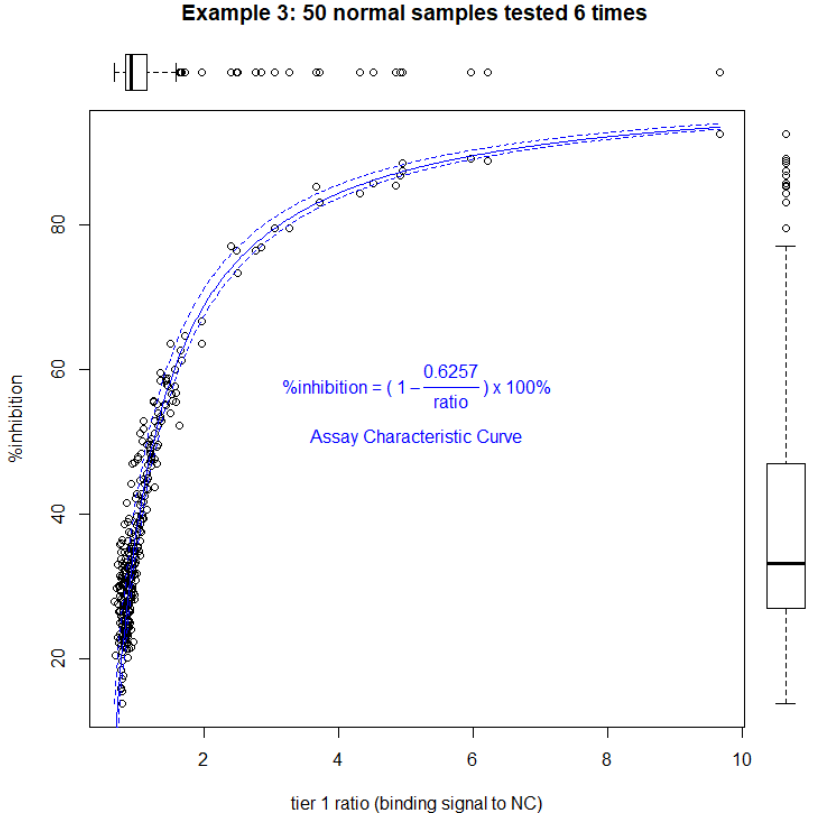
ACC and how good the “fit” is



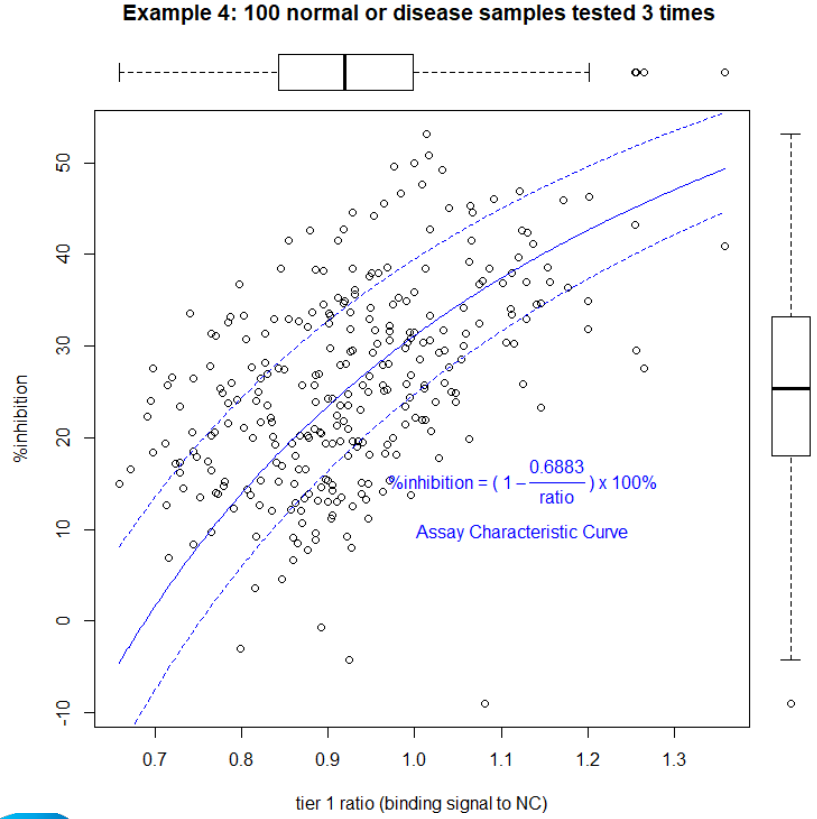


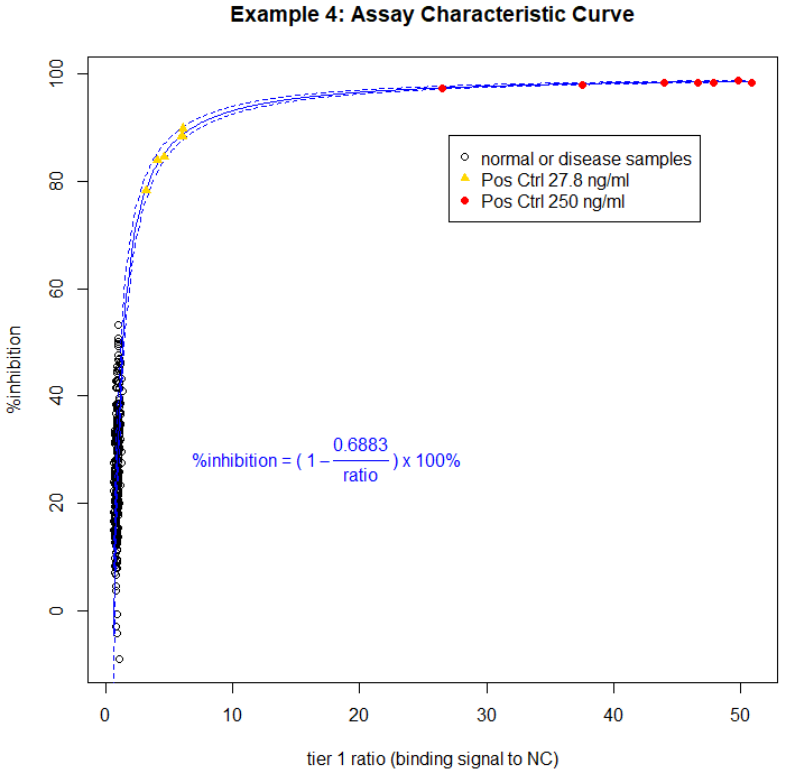
ACC: additional examples





ACC: “exception” proves the rule



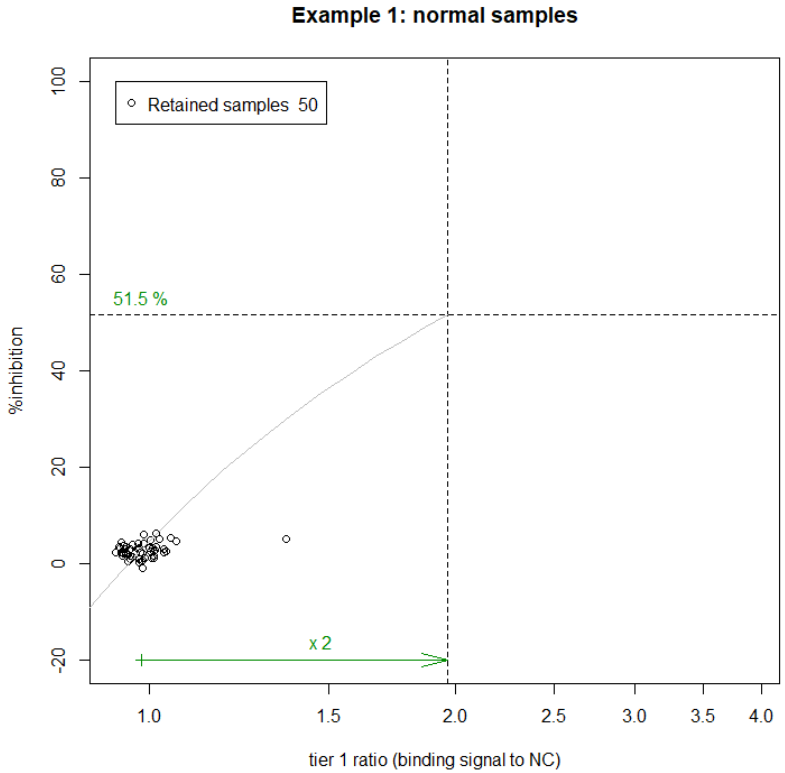


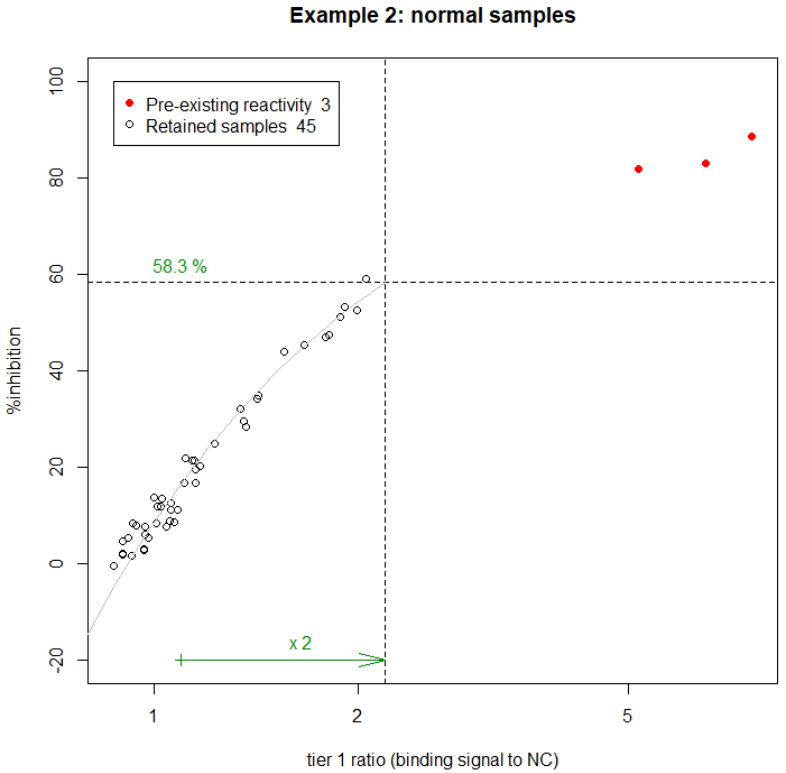
Pre-existing Reactivity

- Define at sample level:
  - The median $ratio > 2 \times median(S_1/NC)$ and
  - The median $\%inhibition \geq \left(1 - \frac{h}{2 \times median(S_1/NC)}\right) \times 100\%$
  - Every ratio and %inhibition of the sample are removed from both tier 1 and 2 datasets

- Justification:
  - High ratio and high %inhibition: indistinguishable from positives
  - Ideal case $h = 1$: $\%inhibition = \left(1 - \frac{1}{ratio}\right) \times 100\%$
    - ratio = 2 ⇔ %inhibition = 50%: a ratio of 2 could be justified as high
  - Typical naïve sample is a true negative, hence, binding signal close to NC: $median(S_1/NC) \approx 1$
    - A ratio of $2 \times median(S_1/NC)$ for naïve samples could be justified as high

Pre-existing Reactivity: Examples



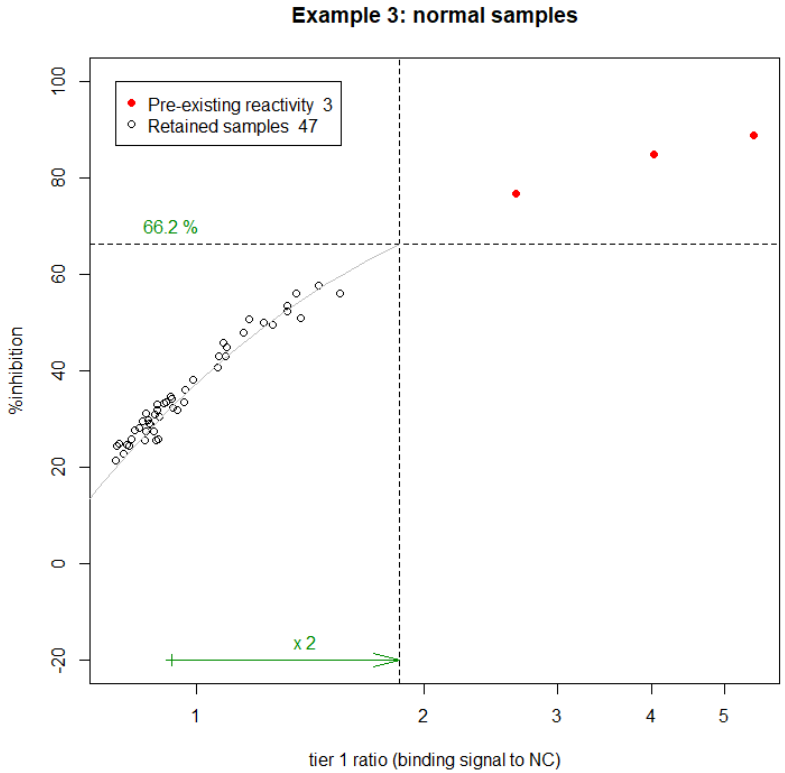


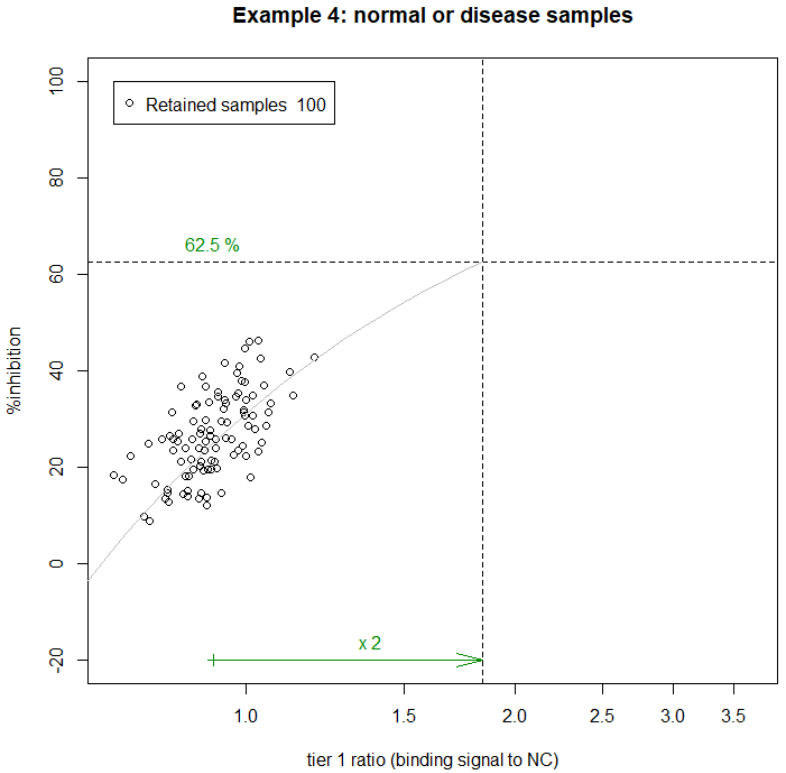
Nonspecific Binding

- Define at sample level:
  - The median $ratio > 2 \times median(S_1/NC)$ and
  - The median $\%inhibition < \left(1 - \frac{h}{2 \times median(S_1/NC)}\right) \times 100\%$
  - Every ratio and %inhibition of the sample are removed from both tier 1 and 2 datasets

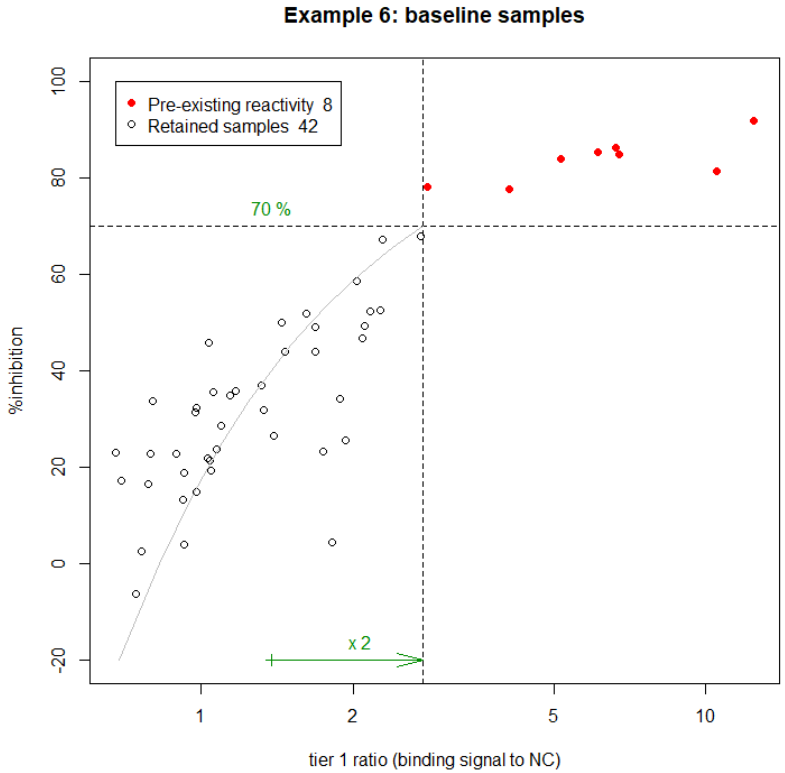
- Justification:
  - High ratio and low %inhibition: nonspecific binding
  - Low %inhibition is caused by high $S_2/NC \gg 1$
    - The binding signal is insufficiently inhibited by the absorber

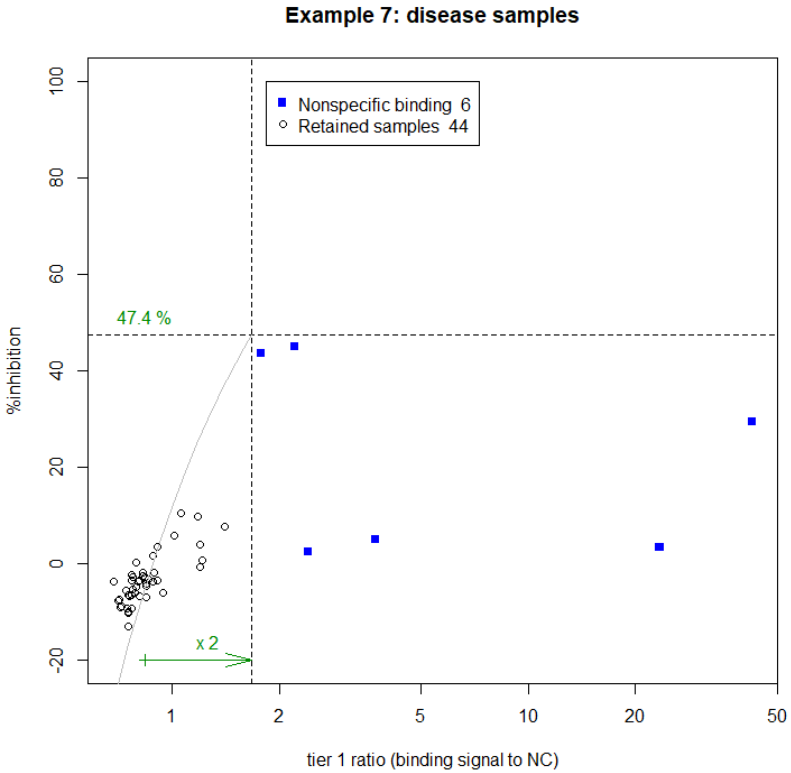
Crisp Cases


Example 3: normal samples
Pre-existing reactivity 3
Retained samples 47
66.2 %
x 2
%inhibition
tier 1 ratio (binding signal to NC)


Example 4: normal or disease samples
Retained samples 100
62.5 %
x 2
%inhibition
tier 1 ratio (binding signal to NC)

Messier Cases





Why parametric approach could be problem

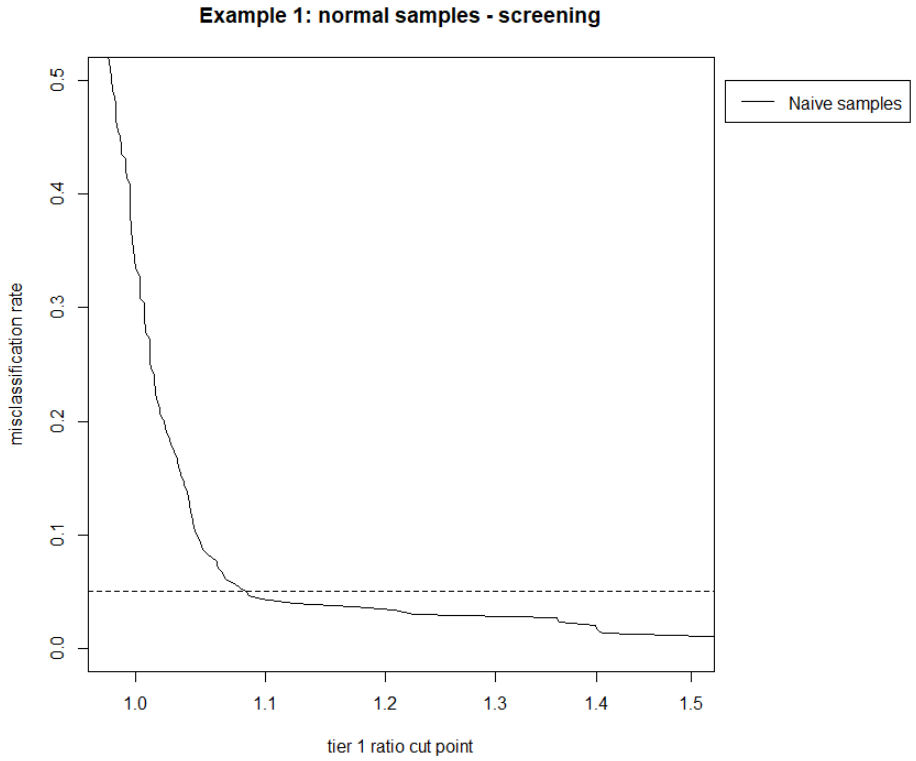
- Parametric approaches are driven by bulk of the data
- Insensitivity to small mixture at the tail
- Naïve population is often not homogenous, but a mixture biologically

- Only two parametric models (normal and lognormal) are in real contention
  - Sometimes both p-values are very small
- Lack of consistency: different data sets from the same population could end up with different models
- The “robust” approach still relies on Normal constants
- If the naïve samples are truly “clean”, there is theoretical base for the tier 1 ratio to be lognormal
  - When the variability is small enough, the lognormal is hard to distinguish from normal
- But there is no theoretical base for %inhibition to be either normal or lognormal
- Excessive cleaning would force data into normality

Modern nonparametric method for cut point calculation

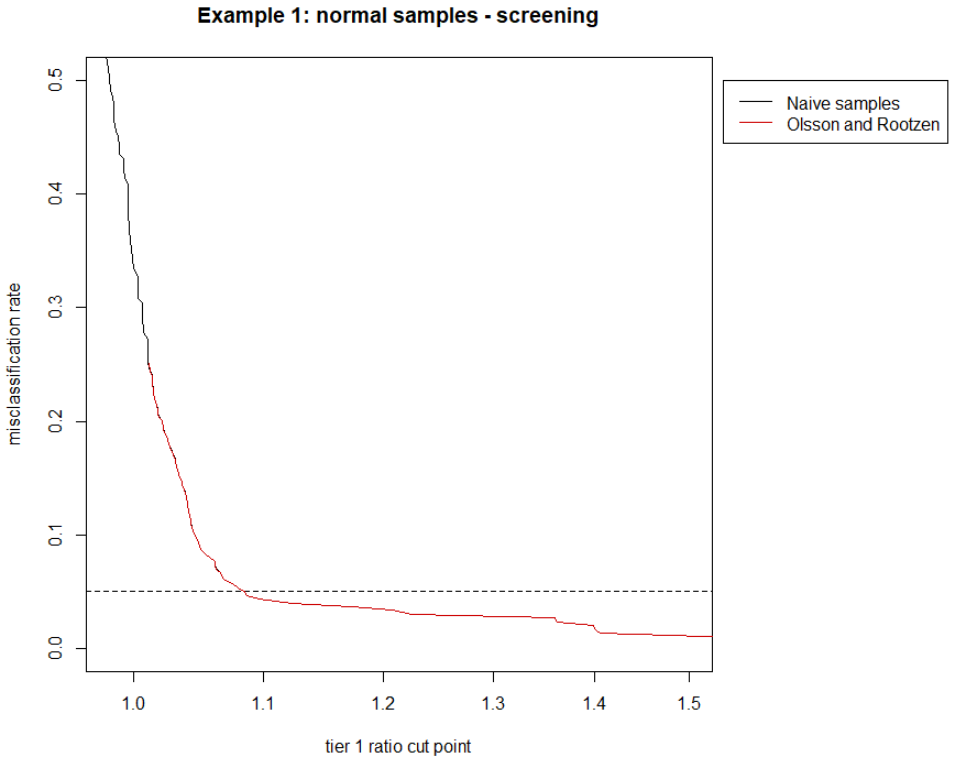
- Olsson, J. and H. Rootzen (1996). "Quantile Estimation From Repeated Measurements." JASA **91(436):** 1560-1565
  - Recognize the fact that 50 samples were tested 6 times each, not 300 independent samples
  - Classic nonparametric method is based on order statistics which ignores the fact that 50 samples were tested 6 times each, and treats the data as if 300 independent samples
  - Numerical method: 2nd order correction to the classic nonparametric method estimate based on local correlation
  - Provides confidence bounds for percentile estimates
  - The actual point estimates turn out very similar to those from the classic nonparametric method, almost identical after rounding

Build a cut point graph: step 1



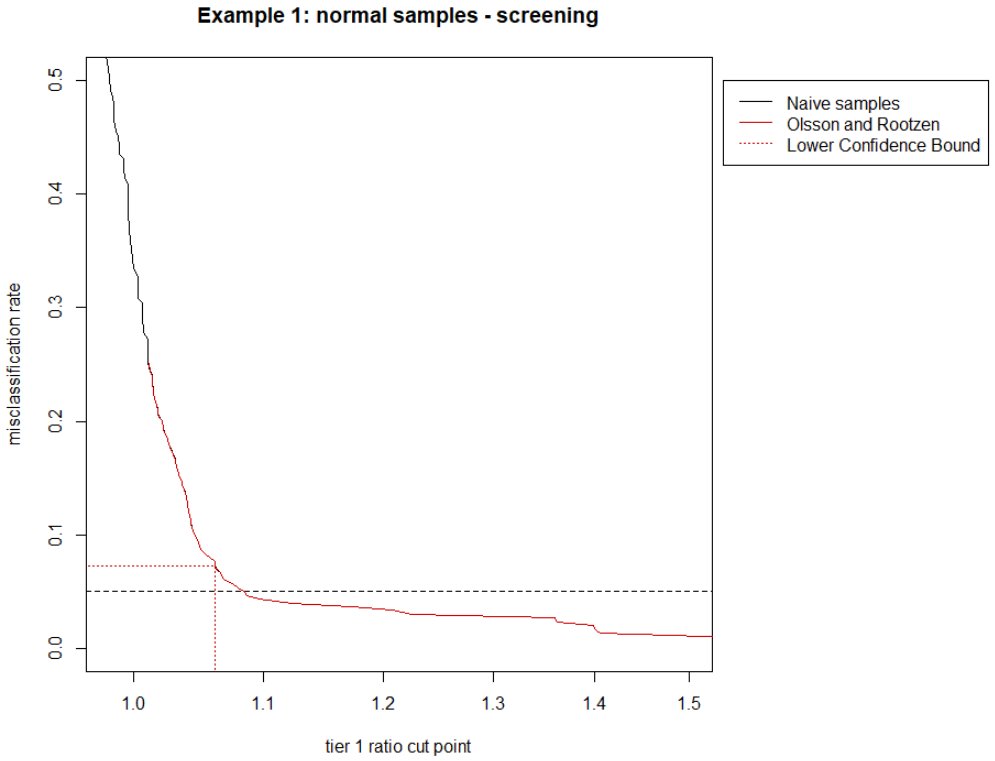
- False positive rate for all possible cut points
- Basically 1-F, where F is the cumulative distribution of the naïve sample data
- Only the right half is plotted
- The cut point for 5% FPR is where the solid line intersects the 5% misclassification line
- This is a way to display the distribution of actual data, as well as a nonparametric analysis (classic)

Build a cut point graph: step 2



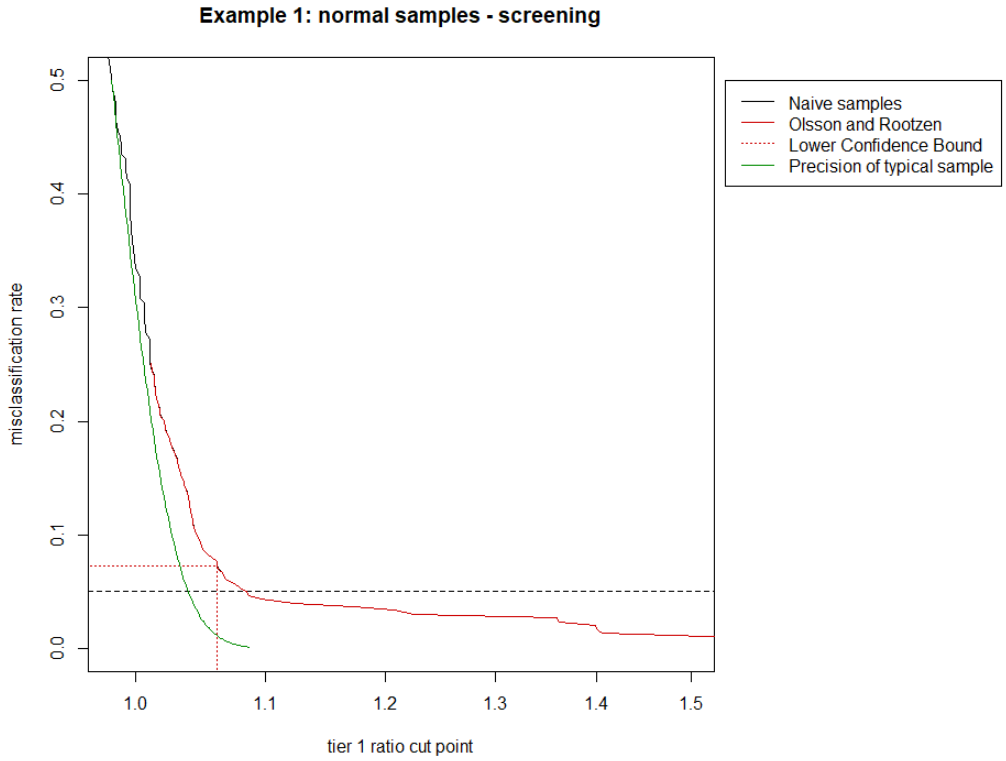
- The Olsson and Rootzen (local) correction for correlation among measurements for the same samples
  - This 2nd order correction turns out to be very minor
- Only part of the right tail of Olsson and Rootzen is plotted here
- The cut point by Olsson and Rootzen is almost identical to the cut point by classic nonparametric method, particularly after rounding
- If you don't have statistician support, just use the classic nonparametric method

Build a cut point graph: step 2.5



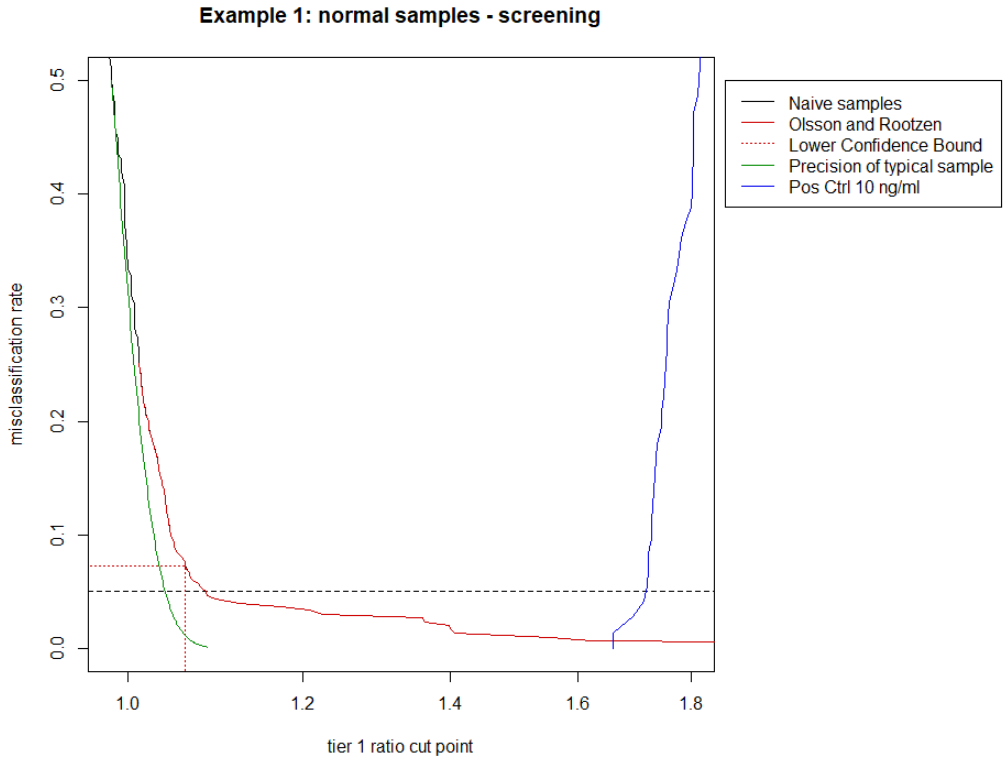
- 90% lower confidence bound for the 95th percentile based on Olsson and Rootzen is plotted
- The false positive rate is inflated a bit
- The Olsson and Rootzen steps need statistician

Build a cut point graph: step 3



- The right tail of a normal distribution for repeated measurement on a typical naïve sample
  - Centered at $median(S_1/NC)$
  - Variability pooled from within samples variability
- This represents the lower limit for cut point that can be supported by assay precision
- The difference between the green line and black/red line represents the sample-to-sample variability beyond the measurement variability

Build a cut point graph: step 4

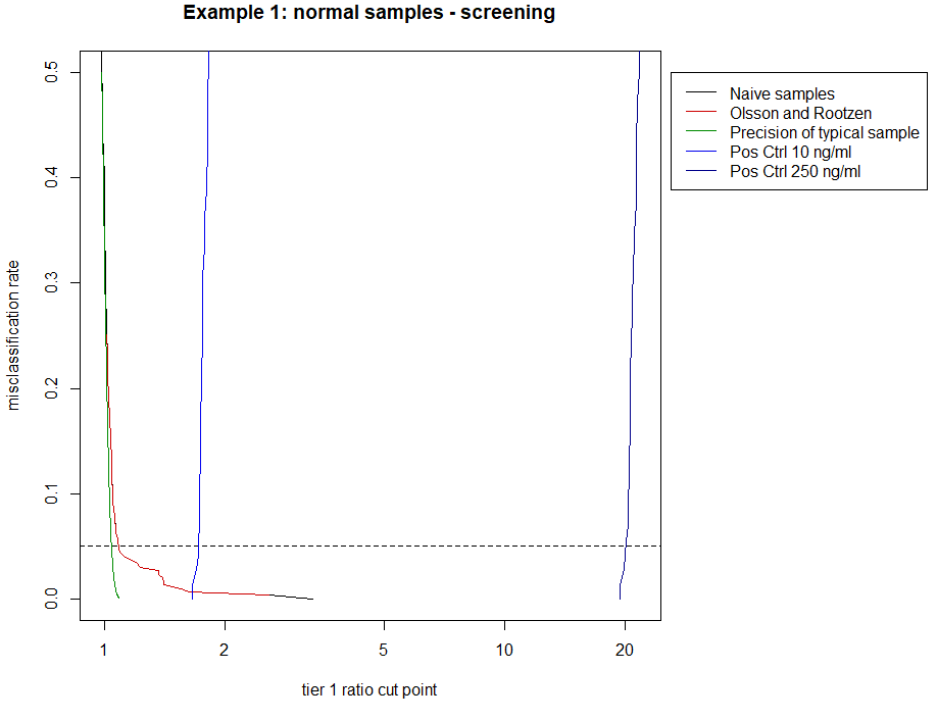


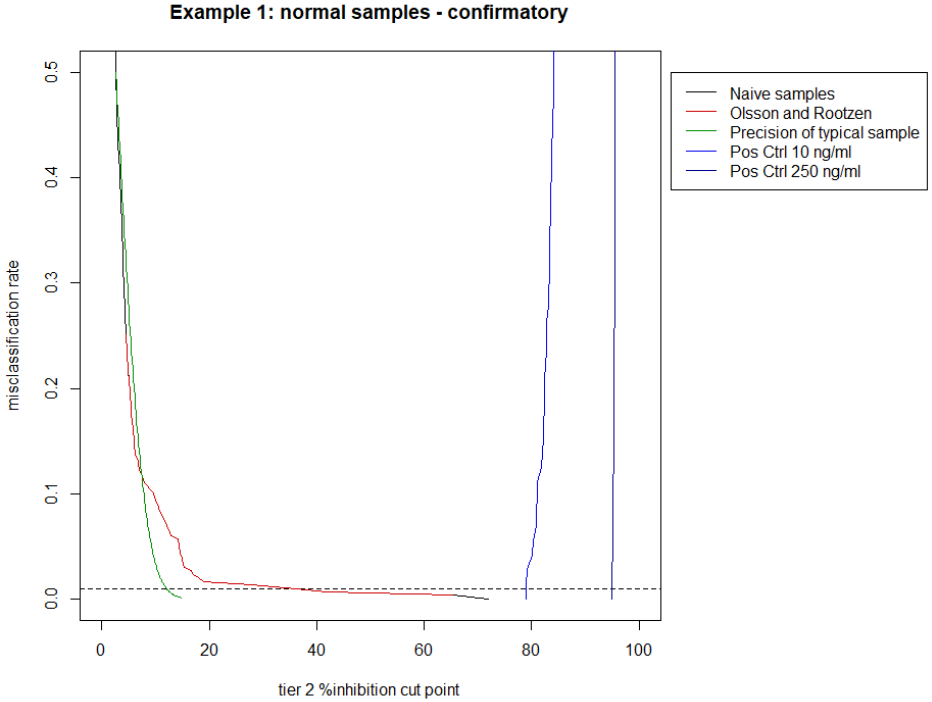
- Misclassification rate for positive control is false negative rate
- The left half of the cumulative distribution F of the positive control is plotted
- The nonparametric cut points are so far below the 10 ng/ml positive control that there is little reason to use lower confidence bound

Regulatory and Clinical requirements on sensitivity

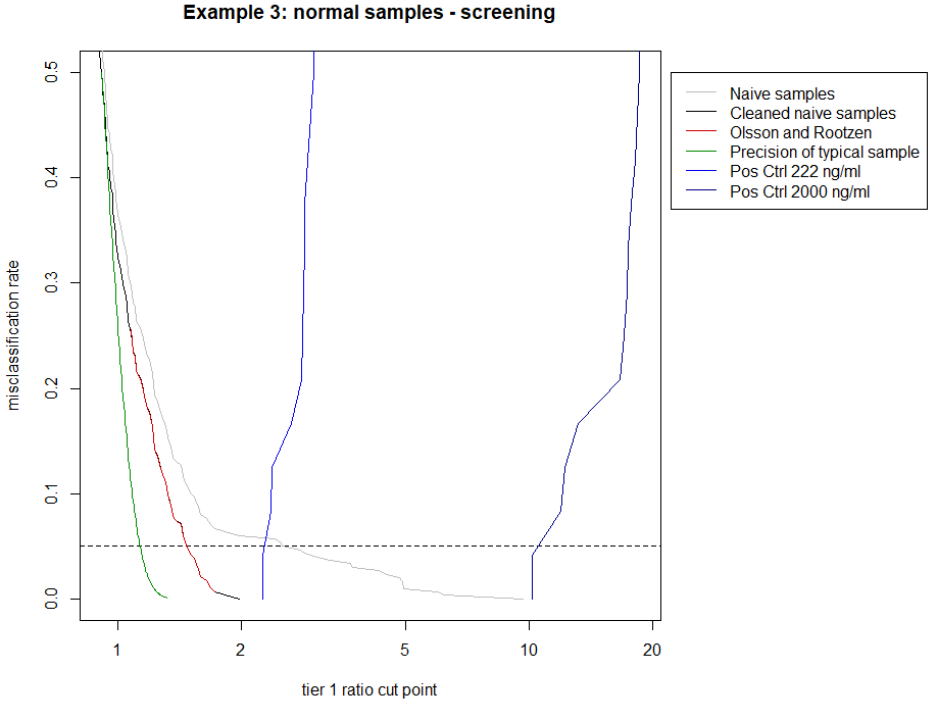
section IV.C.2).[20] FDA recommends that screening and confirmatory IgG and IgM ADA assays achieve a sensitivity of at least 100 nanograms per milliliter (ng/mL) although a limit of sensitivity greater than 100 ng/mL may be acceptable depending on risk and prior knowledge. Traditionally, FDA has recommended sensitivity of at least 250 to 500 ng/mL. However, recent data suggest that concentrations as low as 100 ng/mL may be associated with clinical events (Plotkin 2010; Zhou et al. 2013). It is understood that neutralization assays may not achieve that level of sensitivity. Assays developed to assess IgE ADA should have sensitivity in the high picograms per milliliter (pg/mL) to low ng/mL range.

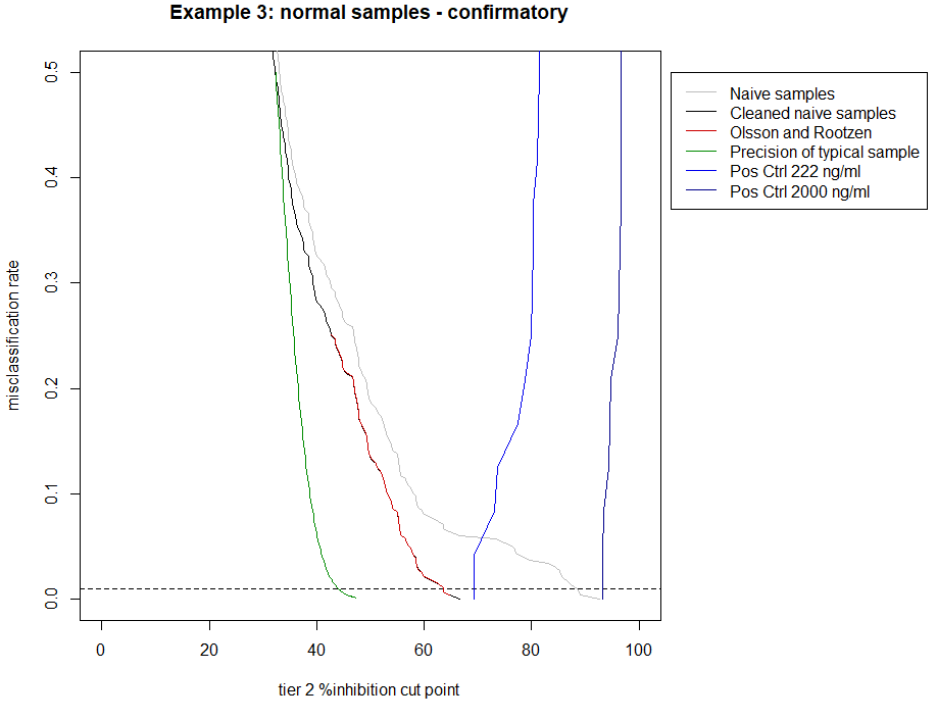
Step 5 for both tier 1 and 2



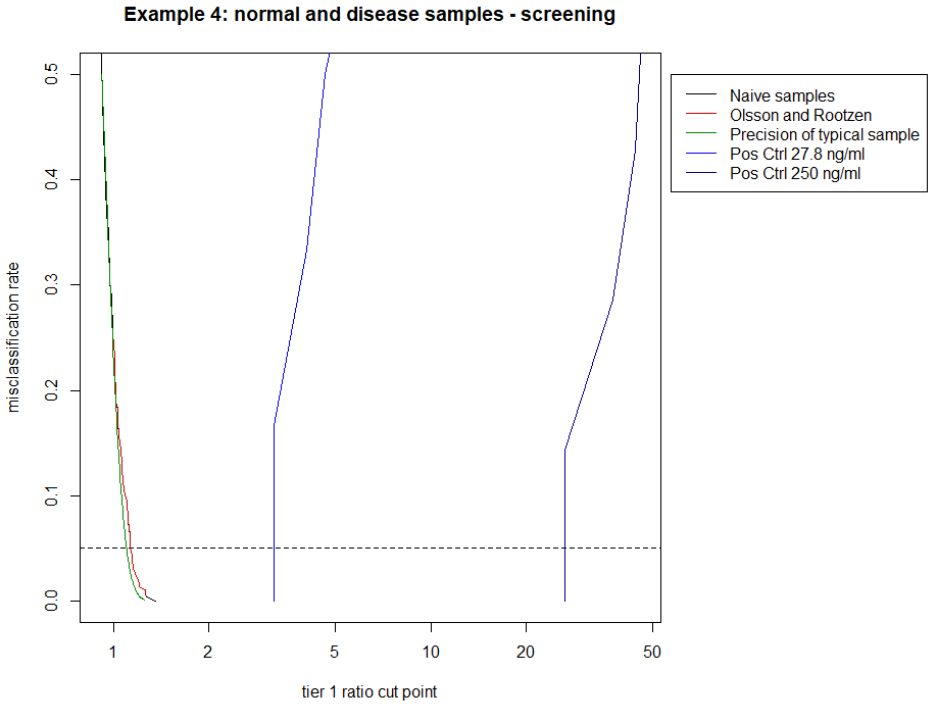


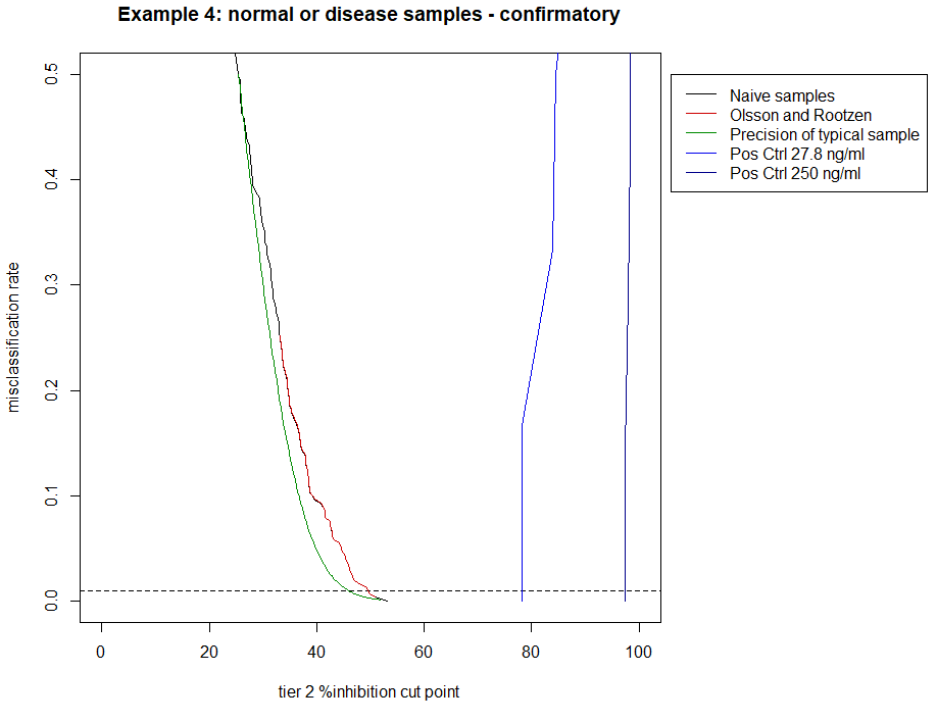
When cleanup made difference





No pre-existing reactivity, low sample-to-sample variability





Thoughts on lower confidence bound

- ONLY use when clinical/regulatory context demands conservative cut points
  - Still should not be below what precision can support
- The point estimator should be sufficient if it is “fit-for-purpose”
  - Deliver the sensitivity the program needs
- LCB is usually not far below the point estimator
  - Raise the false positive rate by a few percentage points

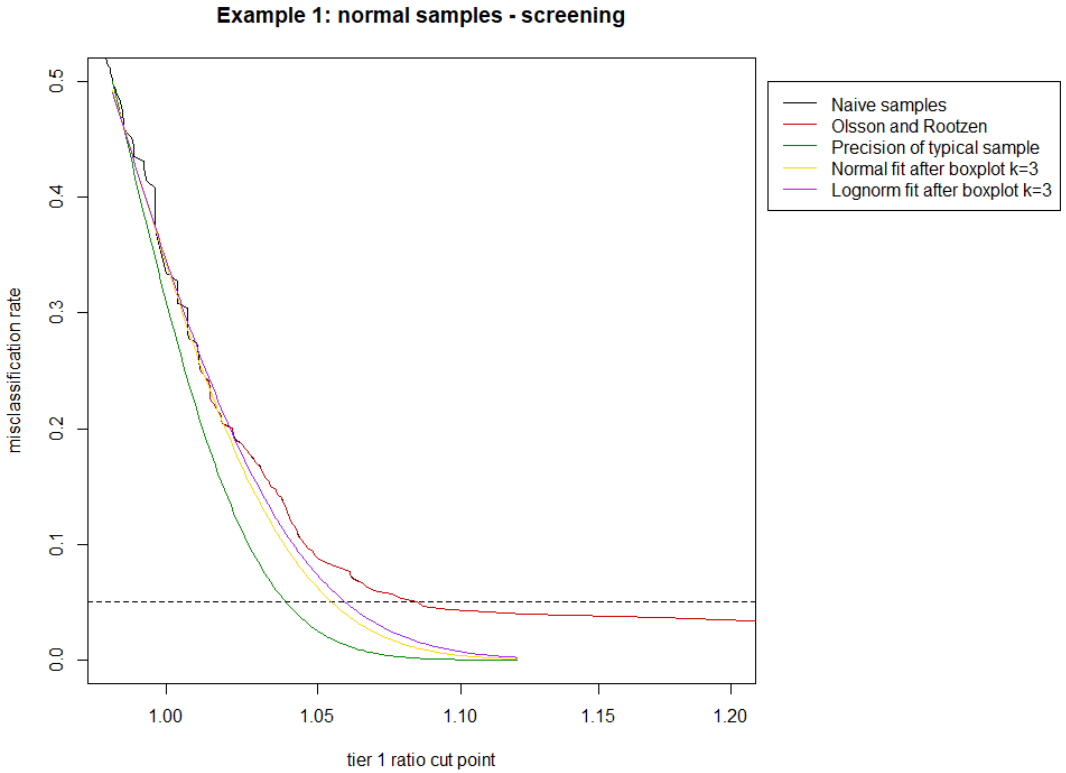
- Things we do that raise actual false positive rate above nominal level
  - Exclude naïve samples with pre-existing reactivity
  - Exclude naïve samples with nonspecific binding
  - Use lower confidence bound, instead of point estimator
  - Trade specificity for sensitivity

Summary

- **Step 1: data preparation**
  - Remove analytical failures: exclude the pair of wells with CV > 20%
  - Remove samples with pre-existing reactivity and nonspecific binding
    - Based on assay characteristics (2 graphs)
- **Step 2: calculate cut points**
  - Nonparametric method
    - Olsson and Rootzen
- **Step 3: assess the cut points**
  - Use 2 graphs to represent the right half of tier 1 and 2 data, and compare them to
  - Minimum cut point implied by the precision of the assay, and
  - Maximum cut point implied by the positive controls around desired sensitivity level
  - Discuss with clinical/management, negotiate with regulatory agencies based on these 4 graphs

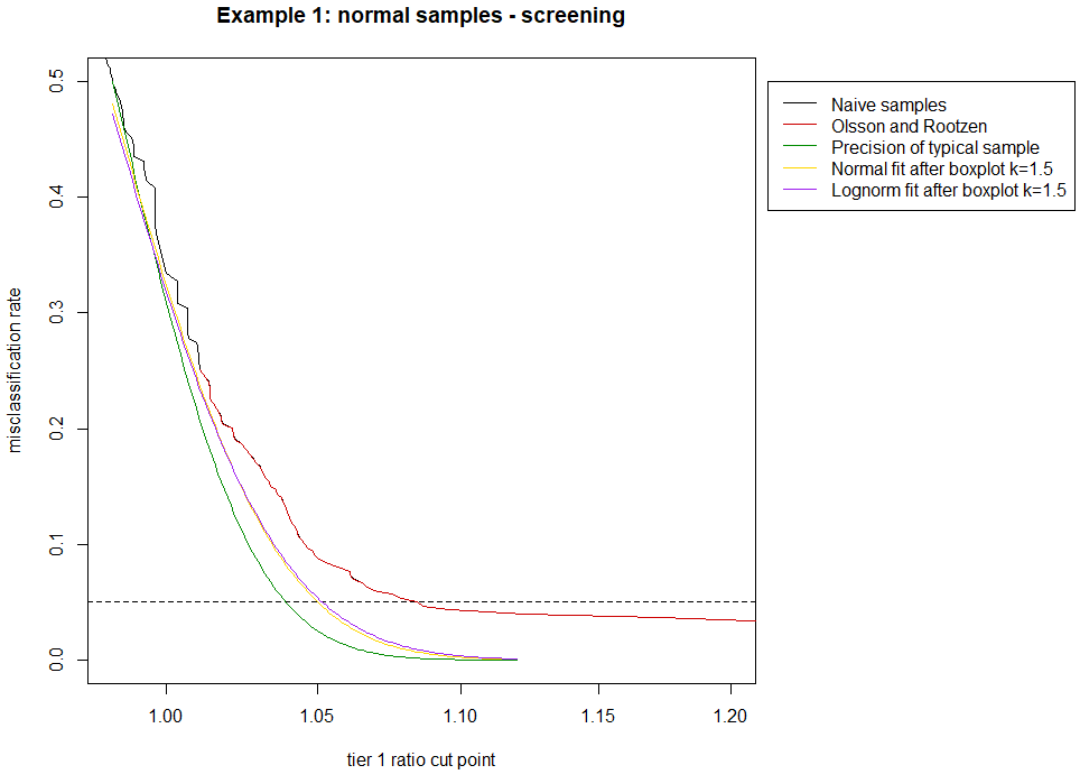
Questions?

Parametric models fail to capture the tail



- Box plot with k=3 applied on ratio scale and log ratio scale
- Fit Normal and Lognormal
- Both parametric models underestimate the right tail
- The difference between Normal and Lognormal fits are very small

Parametric models fail to capture the tail



- Box plot with k=1.5 applied on ratio scale and log ratio scale
- Fit Normal and Lognormal
- After excessive cleaning, little difference remains between Normal and Lognormal fits
- The difference between k=1.5 vs 3 is more dominant

Sources of “2 to 11%”

- Devanarayan, V., et al. (2017). "Recommendations for Systematic Statistical Computation of Immunogenicity Cut Points." The AAPS Journal.
  - “A frequently asked question, especially for phase II and phase III clinical studies or when there is considerable time lag between the validation and in-study testing, is whether the cut point derived during assay validation is relevant for the study samples from a clinical trial. Before answering this question, it is important to recognize that the SCP during prestudy validation that was set to yield 5% false positives is just an estimate. Therefore, the observed FPER of the clinical study baseline samples, after excluding the samples with preexisting ADA, is expected to fall within a certain range of this 5% target value. To assess this range, a simulation study was carried out (33) for typical screening cut point evaluations from the balanced design format (1). This assessment showed that the FPER values can range from **2 to 11%** for a SCP targeted to yield 5% FPER on the average.”
- Amaravadi, L., et al. (2015). "2015 White Paper on recent issues in bioanalysis: focus on new technologies and biomarkers (Part 3--LBA, biomarkers and immunogenicity)." Bioanalysis 7(24): 3107-3124.
  - “The suitability of using the prestudy validation cut-point factor for testing clinical study samples was discussed. Based on the simulation studies done for typical screening cut-point evaluations for a balanced design described in the validation white-paper [29], the false positive rate for a screening cut-point that is targeted around a 5% false positive rate is expected to vary between **2 and 11%**.”

Proper yard stick for 95th percentile

Size	Acceptable Range
50	0-10%
100	2-9%
150	2-8%
200	2.5-7.5%
250	2.8-7.2%
300	**3-7%**
400	3.25-6.75%
500	3.4-6.6%

- Distribution-free test of a 95th percentile claim:
  - At α=0.05 level
  - Size = 100
  - If 10% or more are outside the claimed 95th percentile, the claim could be rejected at α level (p-value = 0.0282)
- The criteria are sample size dependent
  - Impose a healthy “reluctance” to change without sufficient data

Proper yard stick for 99th percentile

Size	Acceptable Range
50	0-4%
100	0-3%
150	0-2.67%
200	0-2.5%
250	0-2%
300	**0.33-2%**
400	0.25-2%
500	0.4-1.8%

- Distribution-free test of 99th percentile claim:
  - At α=0.05 level
  - Size = 100
  - If 4% or more are outside the claimed 99th percentile, the claim could be rejected at α level (p-value = 0.0138)
- Sample size implication:
  - For sample size less than 300, it is entirely probable that the true 99th percentile is not within observed data range