Validating, Verifying, and Evaluating Your Test Methods: It's NOT a Regulatory Exercise!

Pat Garrett, Ph.D., DABCC

Renee Howell, Ph.D., MT(ASCP)

SeraCare Life Sciences, Inc.

AACC Annual Meeting

July 29, 2008

Washington, D.C.

Forget CLIA and CAP for the moment… What are her expectations?

 IHT Sept. 17, 2007

Cancer-free, and weighing mastectomy

DNA tests provide early guide to risk

By Amy Harmon

CHICAGO: Her latest mammogram was clean. But Deborah Lindner, 33, was tired of constantly looking for the lump.

Ever since a DNA test had revealed her unusually high chance of developing breast cancer, Lindner had agonized over whether to have a mastectomy, a procedure that would reduce her risk by 90 percent.

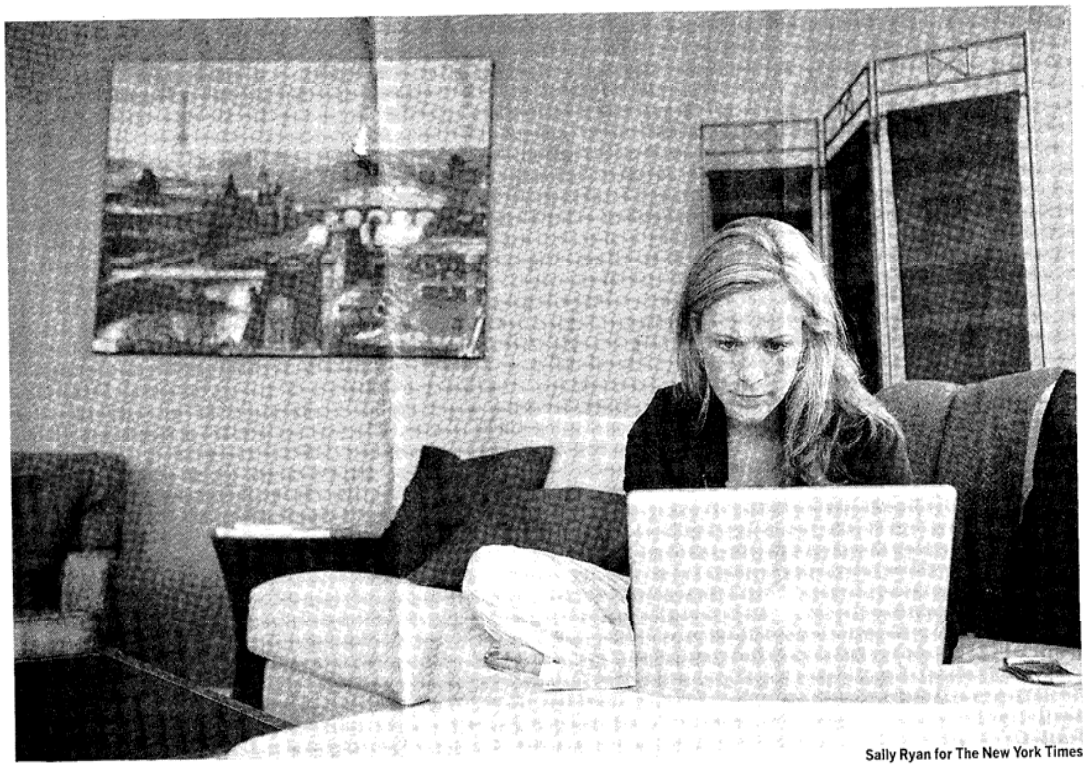
She had stared at herself in the mirror, imagining the loss of her familiar shape. She had wondered, unable to ask, how the man she had just started dating would feel about breasts that were surgically reconstructed, incapable of feeling his touch or nursing their children.

But she was sure that her own mother, who had had chemotherapy and a mastectomy after a bout with the cancer that had ravaged generations of her family, would agree it was necessary.

"It could be growing inside of me right now," she told her mother on the phone in February, pacing in her living room here. "We could find it anytime."

Waiting for an endorsement, she added, "I could schedule the surgery before the summer."

But no approval came



Sally Ryan for The New York Times

...90 percent chance of developing breast cancer.

What are your expectations for your new test?

- Accurate
- Reproducible
- Robust
- Available
- Passes proficiency testing
- Anything else?

How do you check these Performance Characteristics?

SeraCare
LIFE SCIENCES

If you're like my brother, you plunge right in…

However, a little guidance never hurts…

Begin by thinking about your expectations for the test, and about what could go wrong:

- Use your own knowledge and experience with similar methods
- Read the package insert (aka IFU – Instructions for Use)
- Start a FMEA chart
    - What can happen?
        - How bad is it?
        - How likely is it?
- Make a plan to check for and mitigate the most likely and most severe problems
- Begin considering a QA program to monitor for these

SeraCare LIFE SCIENCES

What's a FMEA?

FMEA stands for 'failure mode effects analysis'

This is industrial quality-speak to describe **brainstorming ahead of time these questions:**

- What could go wrong with a process (aka test method)?
- How bad would that be?
- How likely is it to happen?
- Should you try to prevent this? If so, how?

The FMEA 'product' is a chart that helps to guide your further work in understanding and evaluating the test method.

SeraCare LIFE SCIENCES

FMEA example adapted from real life…

Suspected pertussis epidemic at Dartmouth-Hitchcock Medical Center, 2006

- Lab-developed PCR assay used
- Hundreds of nasopharyngeal aspirates tested
- 1445 HCWs treated with antibiotics, 4524 immunized
- 134 informed they had pertussis
- Follow-up testing at CDC on samples from 116 of 134 putative cases
  - 1/116 was PCR positive but culture negative.

Conclusion: No one had pertussis.

Can you prevent this from happening to you?

SeraCare LIFE SCIENCES

FMEA for pertussis test (partial)

Steps	Failure Mode	Failure Cause	Failure Effect	Likelihood of Occurrence (1-10)	Severity (1-10)	Criticality	Actions to Reduce Failure Occurrence
1	False outbreak report	Asymptomatic individuals screened	False positive results	3	7	21	Work with infection control
2A	Positive Result	Incorrect sample matrix	False positive result	5	6	30	Confirm correct sample matrix
2B	Negative result	Wrong sample device	False negative result	7	6	42	Confirm correct sampling device
2C		Inefficient Nucleic Acid extraction	False negative result	6	6	36	Extraction control
2D		Degraded sample	False negative result	7	6	42	Internal control
2E		Incorrect Assay reagents	False negative result	2	6	12	External run control
1F		Degraded assay reagents	False negative result	3	3	9	External run control

Criticality = likelihood x severity

The values here are arbitrary, for example only.

SeraCare LIFE SCIENCES

Verification vs. Validation vs. Establishment

Ongoing confusion as to precise definitions

- CLIA uses establishment and verification
- ISO has very similar definitions for validation and verification
- CLSI documents vary in their use of internationally established definitions

For purposes of our discussion:

- **Establishment:** design and development of performance characteristics
- **Validation**: determination of performance characteristics, once developed
- **Verification**: confirmation of performance characteristics previously determined by the manufacturer during validation
    - Verification studies smaller, narrower in scope, but extremely important!

This is how the test operates in your lab.

SeraCare LIFE SCIENCES

Verifying test performance in your lab

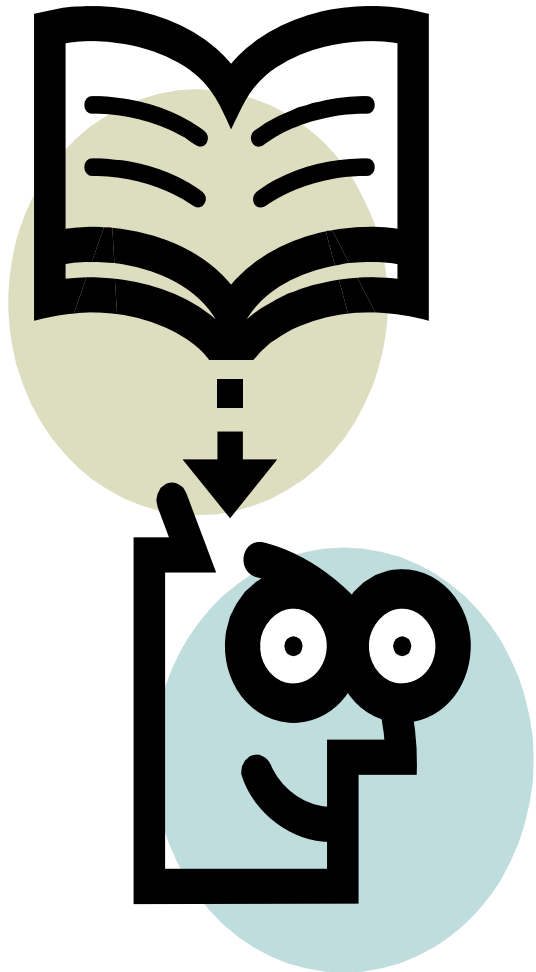
- **Accuracy**: can the test produce the correct result?
- **Reproducibility/precision**: can it do this consistently?
- **Robustness**: can all our techs run this method reliably on our schedule?
- **Reference range/reportable range**: was the manufacturer's range established with a population similar to ours?
- **Calibration plan**: what does the manufacturer recommend, and can we do this?
- **QC plan**: how do we detect errors when they occur?

Guidance (beyond this workshop)

EP15-A2

User Verification of Performance for Precision and Trueness; Approved Guideline—Second Edition

www.clsi.org



CLINICAL AND LABORATORY STANDARDS INSTITUTE ™

(Formerly NCCLS)
Providing NCCLS standards and guidelines,
ISO/TC 212 standards, and ISO/TC 76 standards

SeraCare LIFE SCIENCES

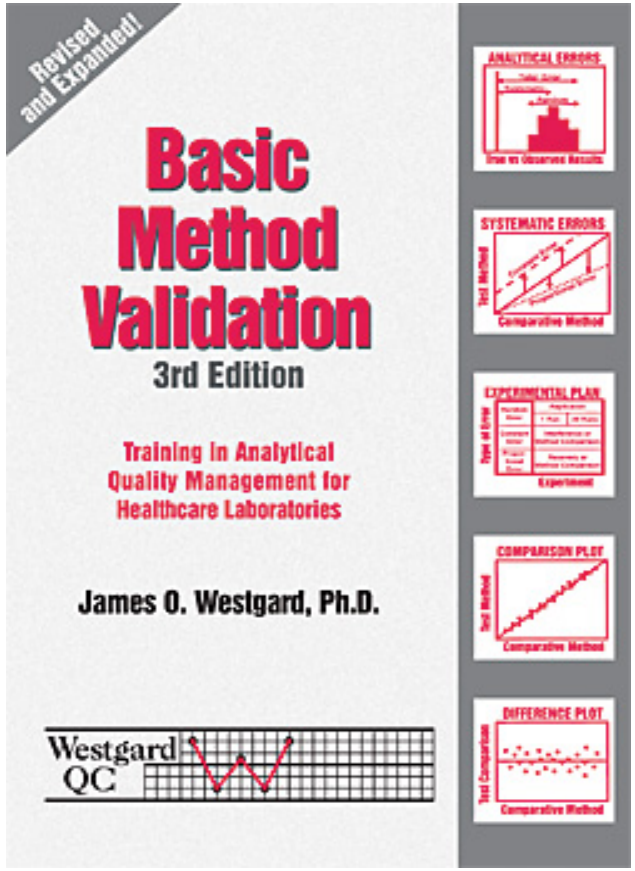
More guidance (this just in…)

Basic Method Validation, 3rd edition

Training in Analytical Quality Management for Healthcare Laboratories

James O. Westgard, Ph.D.

www.westgard.com



SeraCare LIFE SCIENCES

Verification Studies Summary

Purpose: to assess error inherent in the test, so that you can recognize unacceptable changes

Studies:

- Accuracy
  - comparison of methods experiment
    - sensitivity, specificity, PPV, NPV
  - linearity
- Precision
  - within-run (repeatability)
  - between-run (reproducibility)
- Analytical sensitivity (detection limits)
- Analytical Specificity

SeraCare LIFE SCIENCES

A quick review of error

There are only two kinds! (Well, three, actually.)

- **Random error**: an increase in the standard deviation (sd)
- **Systematic error**: a shift in the mean

The baseline settings for mean and sd are assessed in the verification studies.

The QC program monitors the test for changes in these.

The third kind of error is **sporadic error**: a design flaw or implementation flaw such as when the sequence where your primer sits has an unexpected mutation and you get a false negative. QC can't help with those, but experience and a comparison of methods experiment can.

A quick review of error, continued

Total allowable error: what is it, and why do I care?

- Second question first: if the total error of your test method is bigger than the total (medically) allowable error, you're doomed!
- So what is the total allowable error?
  - The difference in results that will trigger a medical intervention
  - The %CV allowable under the CLIA regs (for some chemistry analytes)
  - EP15 assumes if precision and accuracy are ***acceptable***, the assay's total analytical error is acceptable
  - For additional info on total error, see: Krouwer JS. Setting performance goals and evaluating total analytical error for diagnostic assays. *Clin Chem.* 2002;48:919-927.

SeraCare LIFE SCIENCES

No fair looking at the outcome and deciding "Looks okay to me…"

How do you determine in advance what levels of precision and accuracy are acceptable?

It's not as hard as it sounds:

- For precision, look at your controls' performance (sd, %CV) for this or similar tests, and think about whether that's okay or you really need better.
- For accuracy, consider what the correlation coefficient (r^2) should be when you do your method comparison study.

SeraCare LIFE SCIENCES

Accuracy

Method comparison study

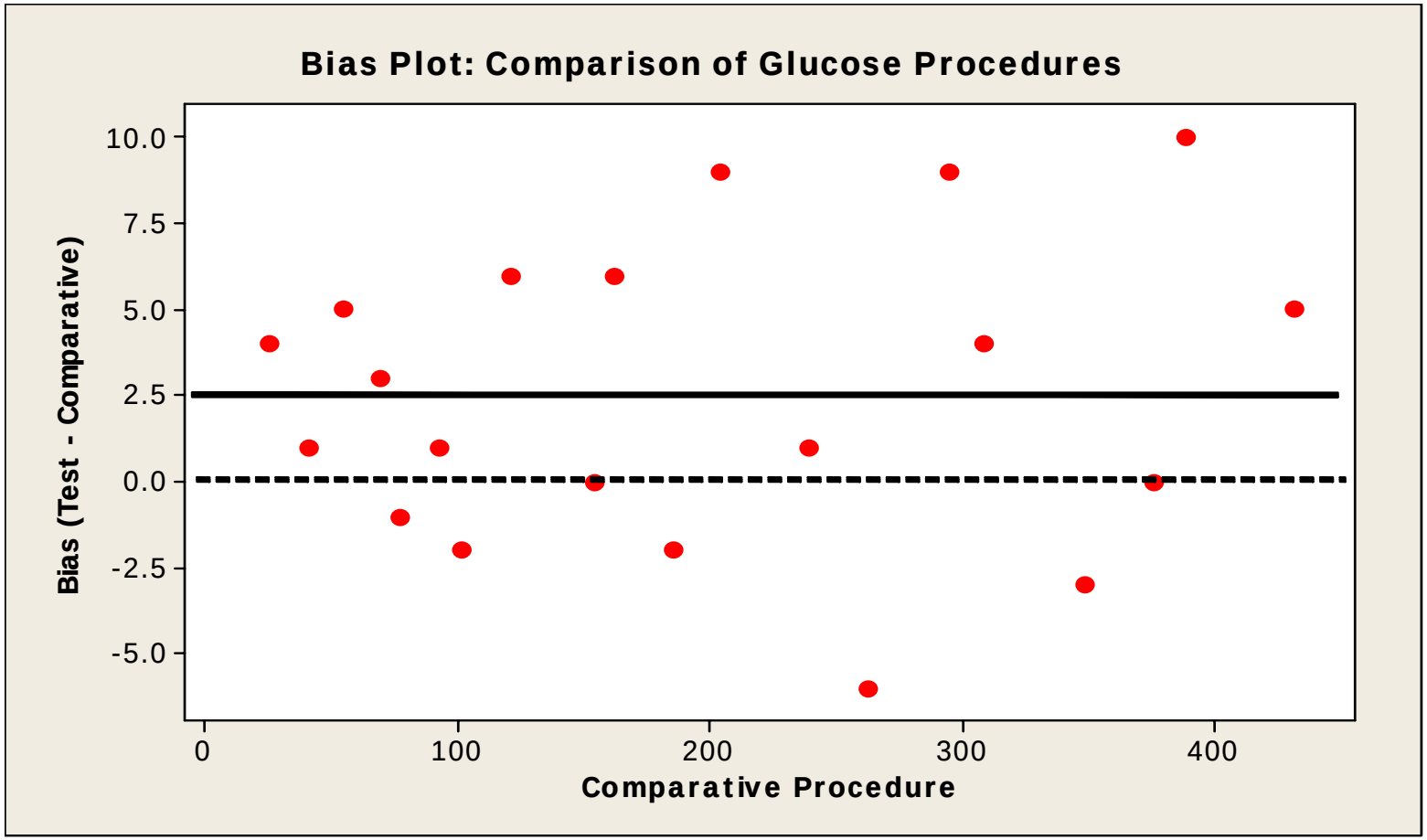
- Requires validated reference method, or the closest thing you have
- Clinical samples, standards, and/or proficiency test samples tested using both assays within narrow time frame
- Results for quantitative tests plotted as difference and scatter plots
- Results for qualitative tests typically plotted on Contingency Tables
  - 2x2 for single analyte assays
  - 4x4 for 2-plex assays
- Still unresolved: How to test a more sensitive assay (i.e. PCR) using a less sensitive reference method (i.e. DFA)

Method Comparison Study: from EP15

- Minimum of 40 different patient specimens
  - Selected to cover a wide range of target concentrations
  - Fewer samples/broader range better than more samples/narrower range – check on reportable, reference ranges
- Duplicates of each sample preferable
    - Controls for operator issues, checks for outliers
- Multiple runs over several days
- Generate a “difference plot” and linear regression plot
  - Pearson Correlation Coefficient r shows random error
  - t-test shows random and constant errors

SeraCare
LIFE SCIENCES

Difference plot borrowed from EP15



$$\%b_i = 100 \bullet \left(\frac{\text{test procedure result}_i - \text{comparison procedure result}_i}{\text{comparison procedure result}_i} \right)$$

SeraCare LIFE SCIENCES

Linearity verification

Checks the useful analytical range

- EP15 recommends minimum 4 levels
- Preferable to use >5
- Make serial dilutions of a high sample, control or standard
  - panels for some analytes are commercially available
- Use clinically relevant sample matrix
- Run 3 replicates of each level
- Plot expected vs. observed (linear regression) and calculate regression line equation, correlation coefficient

Precision verification I

Repeatability: results of multiple observations (tests) under identical conditions (often in one run)

- Estimates **random error** of a test
- Use **clinically relevant analyte levels**
  - low/high clinical decision points
- Use **clinically relevant sample matrix**
  - more than one?
- Typically involves ~20 replicates
  - calculate mean, sd, CV
- QC samples customarily used

SeraCare LIFE SCIENCES

Precision verification II

Reproducibility: results of multiple observations (tests) under varied conditions

- Can allow estimation of random plus systematic error
- Multiple days, technicians, ***reagent lots***
- Same samples, analyte levels, number of observations, calculations – different variety of conditions

Repeatability standard deviations should be <0.25 of allowable total error (TE_a)

Reproducibility standard deviations should be <0.33 (TE_a)

SeraCare LIFE SCIENCES

Vendor-lab cooperation

SeraCare
LIFE SCIENCES

Analytical Studies

Analytical sensitivity: aka Limit of Detection (LOD)

- Lowest concentration of analyte that will be detected 95% of time with 95% confidence
- Minimally 20 replicates
- Narrow range of concentrations tested
- Last concentration with 19/20 positives is LOD estimate
- For multiplex assays, LOD must be determined for each target singly, as well as verified with other potential targets present.

Analytical Studies

Analytical specificity

- Ability to detect desired target in the presence of other, closely related targets
- In the presence of interfering substances
- Performed using clinically relevant sample matrix
- For genotype assays, evaluate multiple alleles
- For ID assays, closely related species
- For multiplex assays, rule out cross-reactivity of detection reagents for other target in the assay

Other test parameters

Robustness: you found some of the test's virtues and glitches while all those studies were going on. Write those down!

Passes proficiency tests: if you could get them, you tested some PT samples in your method comparison study, so you have some information about this, too.

Availability: you tried to get more than one lot of reagents for the reproducibility study, so you've been in touch with the vendor and have some sense as to whether they can deliver consistently.

Calibration plan: you've checked out the manufacturer's recommendations during your studies, and you have to follow these anyway.

SeraCare
LIFE SCIENCES

Quality control plan

Every mutation or organism or compound or enzyme that you test for is a separate analyte.

- If you don't have a control for each one, how do you know that your test is detecting each one every time you run it?
- For your quantitative viral load tests, how can you track the precision at 500 vs. the precision at 500,000? Do you think they are the same? Can you prove it?
- These are complex, sometimes manual, often first generation tests – they need MORE QC, not less.
- What about the new instruments that now have serology, autoimmune, and other formerly 'special chem' tests mixed in with the routine chems?

SeraCare LIFE SCIENCES

What QC can do, cannot do

QC can do this:

Detect **systematic error** (shift in the mean)

Detect **random error** (increase in the standard deviation)

Help you become familiar with and understand your test method (because you're consistently running a sample that you know the result of)

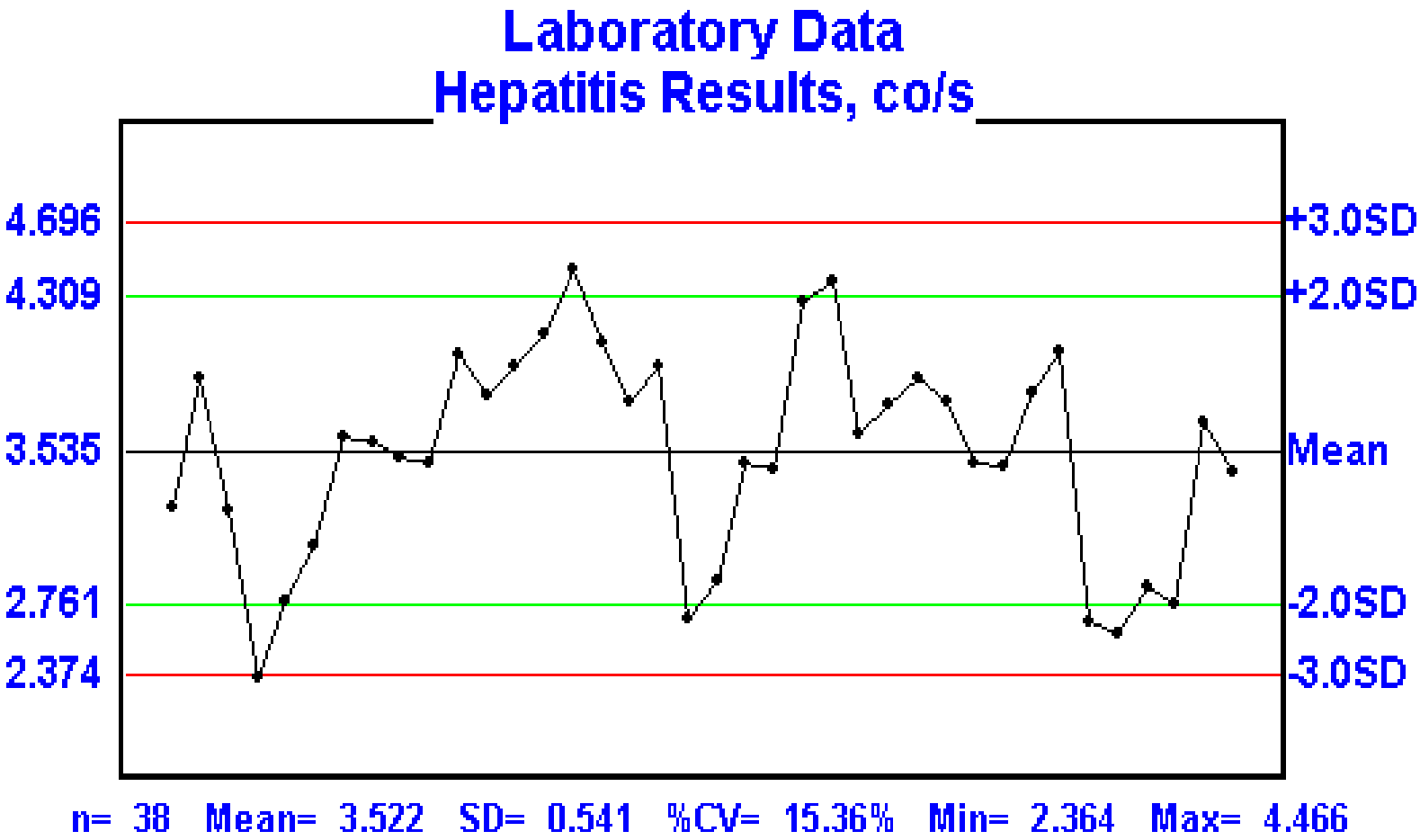
QC can't do this:

Detect sporadic error (design flaw)

Tell you exactly what went wrong (although maybe sometimes it can – if you controls different assay steps separately)

SeraCare LIFE SCIENCES

Levey-Jennings Chart

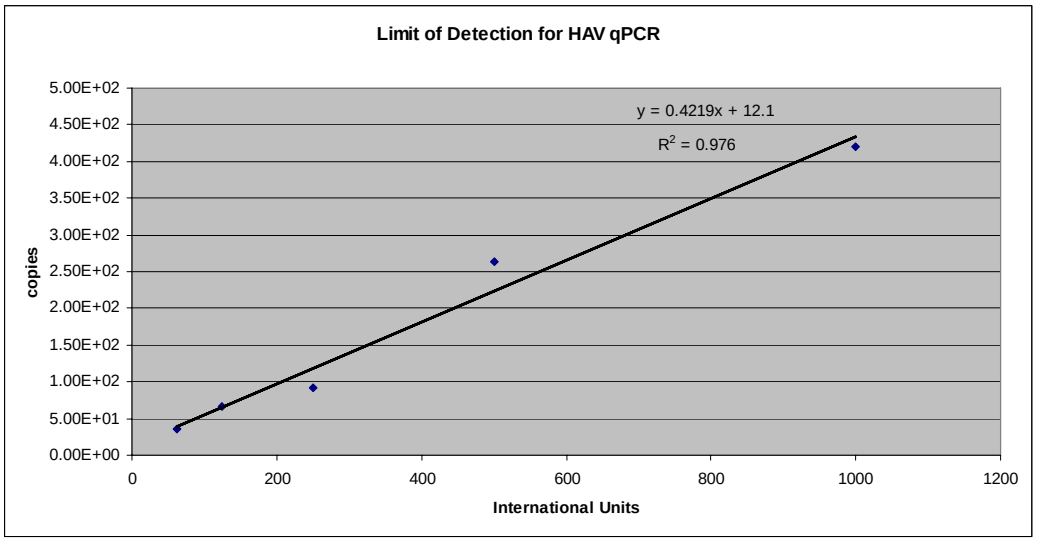


SeraCare
LIFE SCIENCES

Assay Verification

Limit of detection and linearity for qPCR method for HAV detection

- Use a well defined standard or characterized material



Input IUs	Called Concentration (copies)
1000 IU/reaction	6.01E+02
1000 IU/reaction	6.52E+02
1000 IU/reaction	5.61E+02
500 IU/ reaction	2.33E+02
500 IU/ reaction	2.85E+02
500 IU/ reaction	2.74E+02
250 IU/ reaction	1.29E+02
250 IU/ reaction	1.45E+02
250 IU/ reaction	1.29E+02
125 IU/ reaction	6.54E+01
125 IU/ reaction	6.03E+01
125 IU/ reaction	7.30E+01
62.5 IU/ reaction	[3.30E+01]
62.5 IU/ reaction	[3.62E+01]
62.5 IU/ reaction	[3.88E+01]
Run control-240 copies/rxn	1.18E+02
Basematrix negative control	0.00E+00

WHO International Standard WHO First International Standard For Hepatitis A Virus RNA Nucleic Acid Amplification (NAT) Assays

NIBSC code: 00/560

SeraCare LIFE SCIENCES

Assay verification

Lot-to-Lot variation in kit standards for qPCR viral test

	Standard Copies/rxn Lot#1	*Standard Copies/rxn Lot#2*
Kit standard 1	6.90E+05 copies/Reaction	1.28E+06 copies/Reaction
Kit standard 2	6.20E+04 copies/Reaction	1.40E+05 copies/Reaction
Kit standard 3	6.00E+03 copies/Reaction	1.28E+04 copies/Reaction
Kit standard 4	5.80E+02copies/Reaction	1.34E+03copies/Reaction
Kit standard 5	5.90E+01copies/Reaction	1.32E+02copies/Reaction

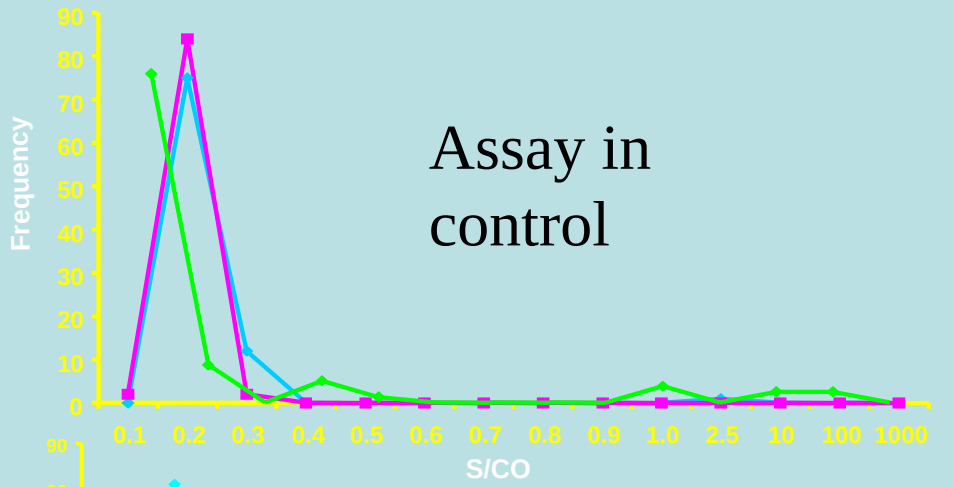
IU/Reaction	*Avg Concentration Lot#14232529*	*Avg Concentration Lot#14028820*
500	3.91E+02	3.39E+02
250	2.54E+02	2.18E+02
125	9.89E+01	[1.22E+02]
62.5	[3.47E+01]	[7.95E+01]

SeraCare LIFE SCIENCES

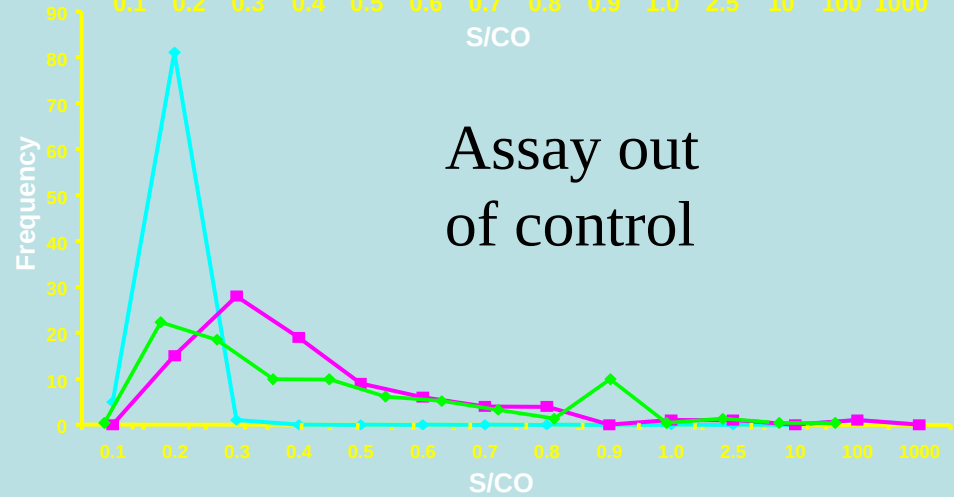
Role of Independent Controls in routine QC

- Are not biased to one specific test method.
- Are not biased to a particular lot of a specific test - manufacturers pair controls with specific lots.
- Act similar to patient samples – can demonstrate matrix effect.
- Can control the whole assay process – an example could be an internal control for PCR that is used post purification.
- Can allow continual comparison of site performance – multiple labs within an organization or across multiple sites of unrelated labs offering the same tests.

Examples of Issues with Kit Controls



Assay in control



Assay out of control

Kit control

Population #1

Population #2

Shift in performance of a lot

SeraCare LIFE SCIENCES

Statistics for Verification & Validation

The usual suspects

- Mean, standard deviation, %CV
- Sensitivity, specificity

- Agreement – kappa and McNemar
- confidence intervals
- ROC curves for cutoff determination

SeraCare
LIFE SCIENCES

Confidence Intervals

Confidence Intervals

- **the confidence interval is the likely range of the true value**
- the confidence interval is NOT the variability of the true value or of any other value between subjects. It is nothing like a standard deviation.
- The confidence interval is dependent on sample size and assay precision.

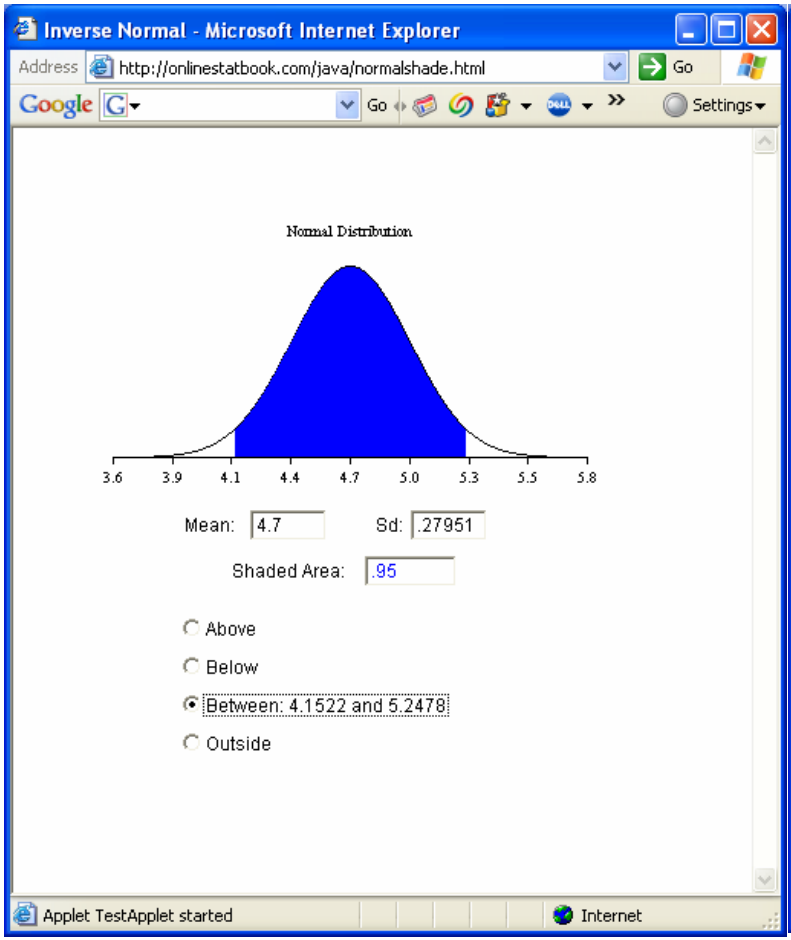
Example

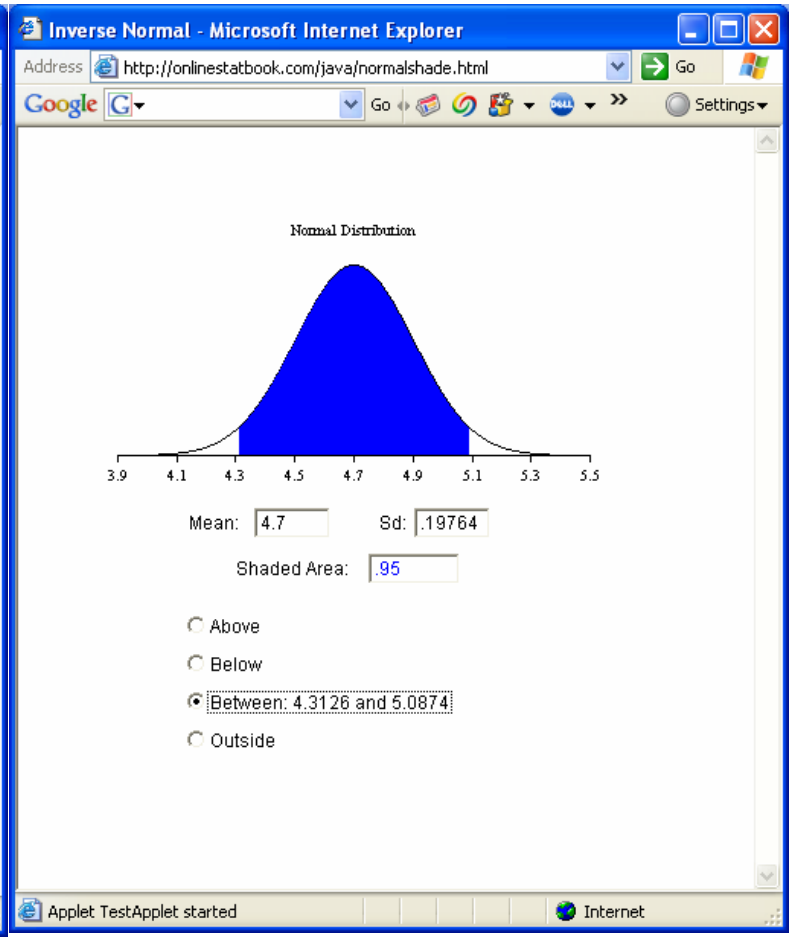
Calculate the 95% confidence interval for the following assay mean:

- the average S/N of 20 samples low positive samples was 4.7 with a standard deviation of 1.25
  - $\mu \times \sigma/\sqrt{N} = \sigma_m$
  - $4.7 \pm 1.96(\sigma_m)$

SeraCare LIFE SCIENCES

Effect of sample size on estimation of confidence interval





SeraCare LIFE SCIENCES

http://onlinestatbook.com/chapter8/mean.html

Evaluation of Other Potential Interferents

Potential interference from HAMA and rheumatoid factor (RF) in the ARCHITECT Anti-TPO assay is designed to be ≤ 15%. In a study, the ARCHITECT Anti-TPO assay was evaluated by testing specimens with HAMA and RF to further assess the clinical specificity. Specimens positive for HAMA and specimens positive for RF were evaluated for % interference with anti-TPO levels spiked between 163.0 and 184.3 IU/mL. Mean absolute % interference is summarized in the following table.*

Other Potential Interferents	Number of Specimens	Mean Absolute % Interference
HAMA Positive	10	2.1
RF Positive	10	1.8

* Representative data; results in individual laboratories may vary from these data.

Clinical Sensitivity

In two studies, clinical sensitivity was evaluated by testing 139 clinically defined Hashimoto's thyroiditis specimens and 125 Graves' disease specimens. The clinical diagnosis was based on the criteria of the respective laboratory and the presence of autoantibodies against thyroglobulin and/or TPO was not necessarily a diagnostic criterion. Results were compared with a commercially available immunoassay. Anti-TPO concentrations ≥ 5.61 IU/mL were considered positive for ARCHITECT Anti-TPO and values ≥ 12 IU/mL were considered positive for the Comparison Assay. In these populations of specimens from patients with Hashimoto's thyroiditis, a 20.9% difference (95% CI of 11.4 to 30.4%) was observed in the number of positive specimens detected by the ARCHITECT Anti-TPO assay relative to the Comparison Assay. Data from these studies are summarized in the following table.*

ARCHITECT Anti-TPO

		Hashimoto's Thyroiditis			Graves' Disease	
Study	n	Number of Positives	% Pos (95% CI)	n	Number of Positives	% Pos (95% CI)
1	89	57	64.0 (53.2 to 73.9)	75	69	92.0 (83.4 to 97.0)
2	50	37	74.0 (59.7 to 85.4)	50	50	100.0 (92.9 to 100.0)
Total	139	94	67.6 (59.2 to 75.3)	125	119	95.2 (89.8 to 98.2)

Comparison Assay

		Hashimoto's Thyroiditis			Graves' Disease	
Study	n	Number of Positives	% Pos (95% CI)	n	Number of Positives	% Pos (95% CI)
1	89	76	85.4 (76.3 to 92.0)	75	71	94.7 (88.0 to 98.5)
2	50	47	94.0 (83.5 to 98.7)	50	50	100.0 (92.9 to 100.0)
Total	139	123	88.5 (82.0 to 93.3)	125	121	96.8 (92.0 to 99.1)

* Representative data; results in individual laboratories may vary from these data.

High Dose Hook

High dose hook is a phenomenon whereby very high level specimens may falsely read within the dynamic range of the assay. For the ARCHITECT Anti-TPO, no high dose hook effect was observed when samples containing up to approximately 17000 IU/mL of Anti-TPO antibody were assayed.

Method Comparison

The performance of the ARCHITECT Anti-TPO assay was compared to a commercially available immunoassay for the determination of anti-TPO. A total of 500 specimens were evaluated in a study, encompassing a population of apparently healthy individuals and patients with autoimmune thyroid disease (Graves' disease and Hashimoto's thyroiditis). Specimens were tested in replicates of one using the ARCHITECT Anti-TPO assay with three reagent lots on three instruments and compared with a commercially available immunoassay (Comparison Assay). Data from this study are summarized in the following table.*

ARCHITECT Anti-TPO	Comparison Assay Negative	Comparison Assay Positive
Negative	242	32
Positive	5	221

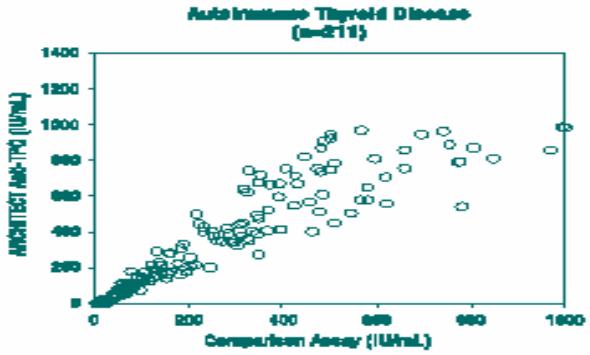
Negative Agreement = 98.0% (242/247) with 95% CI: 95.3% to 99.3%
Positive Agreement = 87.4% (221/253) with 95% CI: 82.6% to 91.2%

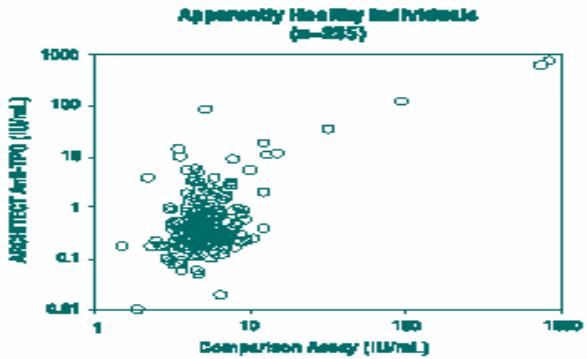
Total Agreement = 92.6%

Sample Range (ARCHITECT) = 0.0 to 27430.8 IU/mL
Sample Range (Comparison Assay) = 1.5 to 50390.0 IU/mL

* Representative data; results in individual laboratories may vary from these data.

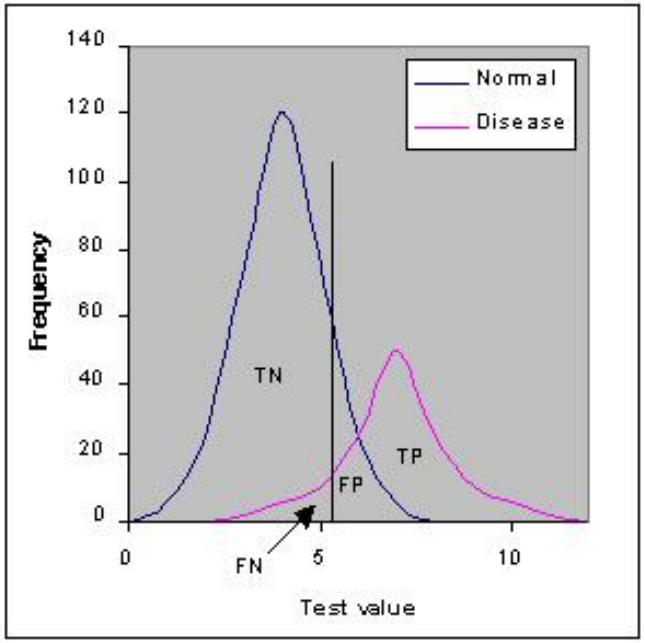
The data from this study was also analyzed by linear regression and it is summarized in the following charts and table.* Samples with anti-TPO levels greater than 1000 IU/mL (n=54) were excluded from this study.

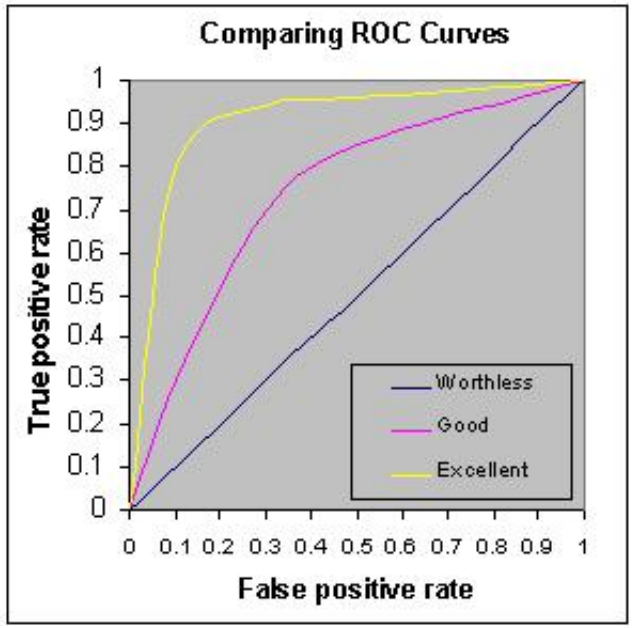




LIFE SCIENCES

ROC (Receiver Operating Curve)





ROC curves are used to look at the relationship between sensitivity and specificity – it's always a balance.

SeraCare LIFE SCIENCES

Agreement

kappa - a measure of agreement between two observations taking into account agreement that could occur by chance (expected agreement).

kappa = Observed agreement - Chance agreement /Total observed - Chance agreement

or

kappa = Observed agreement - Expected agreement/100% - Expected agreement

Tells us nothing about validity of the agreement

Agreement

Interpretation of kappa values*	
kappa	Interpretation
<0	No agreement
0.0-0.19	Poor agreement
0.20-0.39	Fair agreement
0.40-0.59	Moderate agreement
0.60-0.79	Substantial agreement
0.80-1.00	Almost perfect agreement

SeraCare LIFE SCIENCES

Agreement

McNemar

Is the agreement significant?

Assume both observations are from same sample.

	REF POS	REF NEG	
NEW POS	a	b	a+b
NEW NEG	c	d	c+d
	a+c	b+d	

a+b=a+c
c+d=b+d
Then b=c

$X^2=(b-c)^2/b+c$

If X^2 has 1 degree of freedom and $p<0.5$ result is significant

SeraCare LIFE SCIENCES

Does overall agreement tell the whole picture?

Method 1

	REF POS	REF NEG	
NEW POS	5	5	10
NEW NEG	5	85	90
	10	90	100

Method 2

	REF POS	REF NEG	
NEW POS	35	5	40
NEW NEG	5	55	60
	40	60	100

Overall pct. Agreement is 90.0% for both methods, but are they equivalent?

PPA=50.0% (5/10)
[95% CI= 18.7%,81.3%]
NPA=94.4% (85/90)
[95% CI= 87.5%,98.2 %]

PPA=87.5% (35/40)
[95%CI=73.2%,95.8%]
NPA=91.7% (55/60)
[95% CI=81.6%,97.2%]

SeraCare LIFE SCIENCES

Before you carry out the verification or validation – consider the goals and the assay capability

- Is the goal to “know” agreement or to replace an imperfect standard with a new method?
- What precision will be acceptable?
- Are you verifying existing product claims?
- Are you validating a home brew assay or a new assay using individual ASR reagents? Is your sample size appropriate, references appropriate?
- What statistical methods will help you achieve your goals?
- Are your data normally distributed?

SeraCare
LIFE SCIENCES

Thank you! And we hope we can answer your questions…

Pat Garrett, Ph.D., DABCC

Renee Howell, Ph.D., MT(ASCP)

SeraCare Life Sciences, Inc.

E-mail:

pgarrett@seracare.com

rhowell@seracare.com

SeraCare LIFE SCIENCES