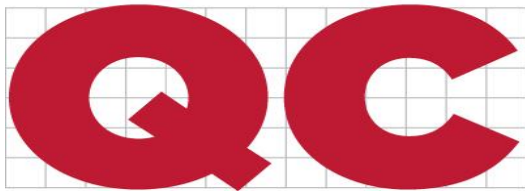


A Warning on Warning Rules

Sponsored by Thermo Fisher MAS controls



AGENDA

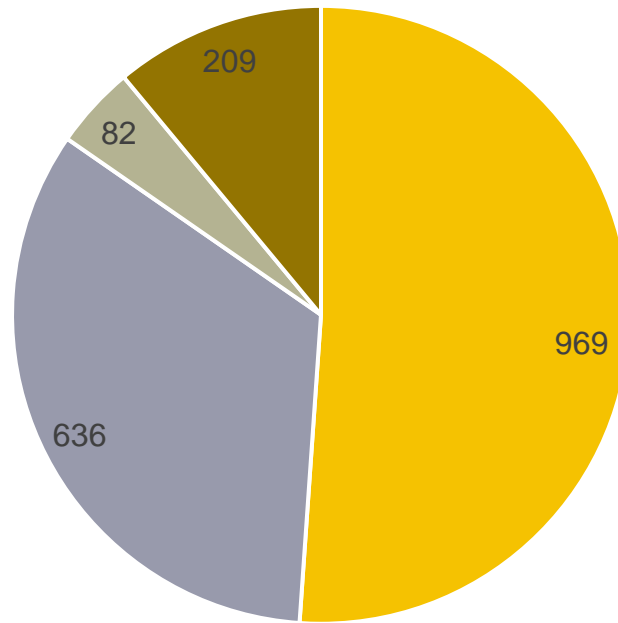
1. Ancient History of Warning Rules
2. Technical Analysis of Warning Rules by Critical-Error graphs
3. WESTGARD SIGMA RULES
4. A Warning on Widened Ranges and “Standardized” means and SDs





DO YOU USE WARNING RULES?

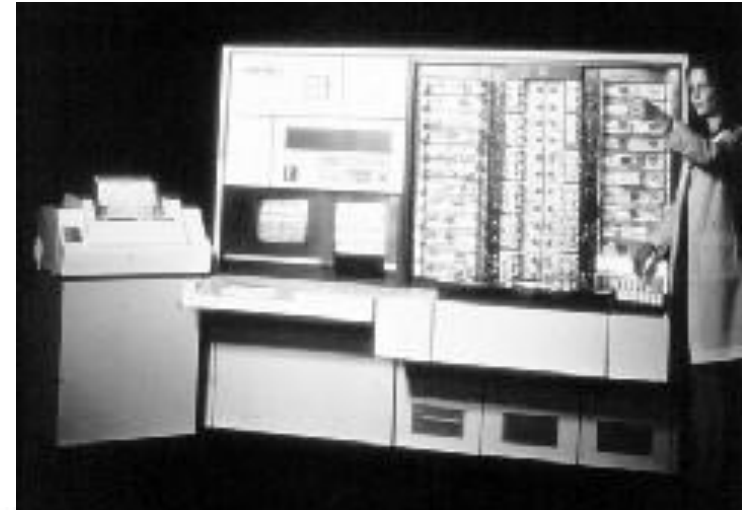
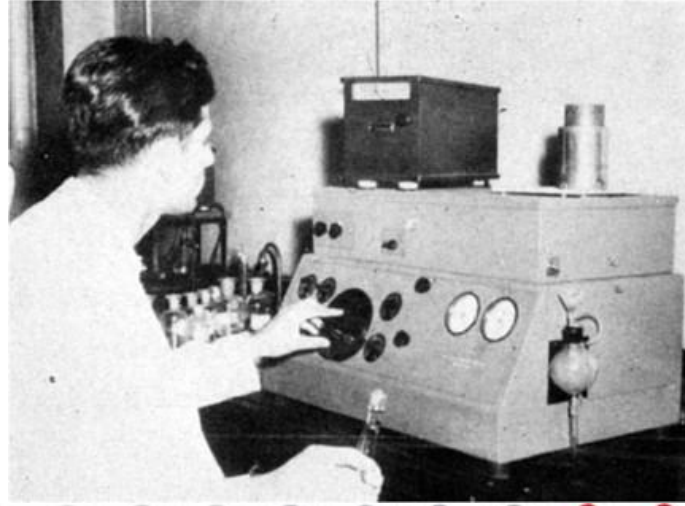
Who uses Warning Rules? 1896 Registrants



- On ALL of my tests 51%
- On SOME of my tests 33%
- On NONE of my tests 4%
- some other way 11%

LET'S GO BACK IN TIME: DO YOU REMEMBER...

- Pipetting by mouth?
- Flame photometers?
- The “SMAC” first automated instrument from Technicon?
- Using 2 SD control limits?

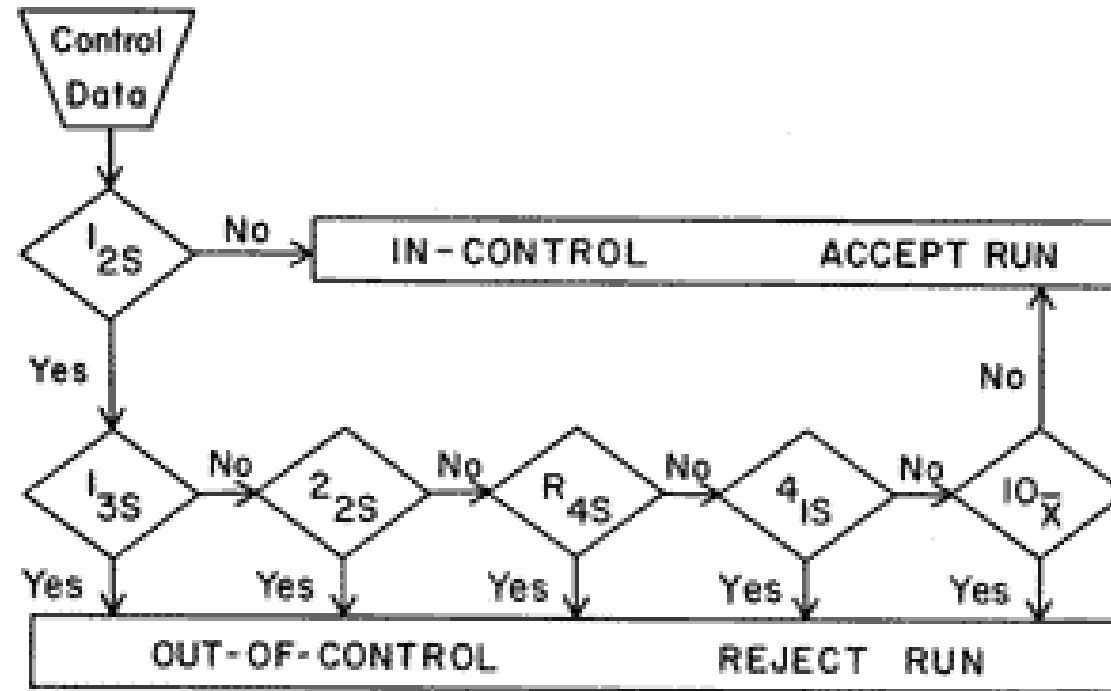


THE WARNING RULE DATES BACK TO 1981... AND WESTGARD RULES

2 SD formerly used as rejection rule

2 SD converted into a “warning rule”

Classic laboratory workaround



Westgard JO, Barry PL, Hunt MR, Groth T. A multi-rule Shewhart chart for quality control in clinical chemistry. Clin Chem 1981;27:493-501.

WHY CONVERT 2 SD INTO A WARNING RULE?

2 SD has a VERY HIGH false rejection rate

- 2 controls with 2 SD: 9% false rejection (think 1 out of 10)
- 3 controls with 3 SD: 14% false rejection (think 1 out of 7)

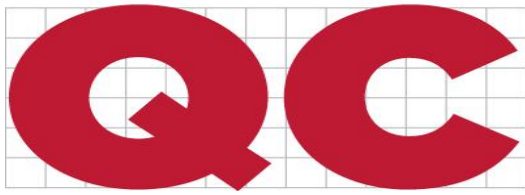
If you run 60 tests every day, that means...

- 1,971 OOCs for N=2 (think 5 outliers / day)
- 3,066 OOCs for N=3 (think 8 outliers / day)

That's a LOT of outliers – when NOTHING IS WRONG!

P.S. if you run 2 SD and don't see as many outliers – you're wrong in a different way





A REAL-WORLD CASE

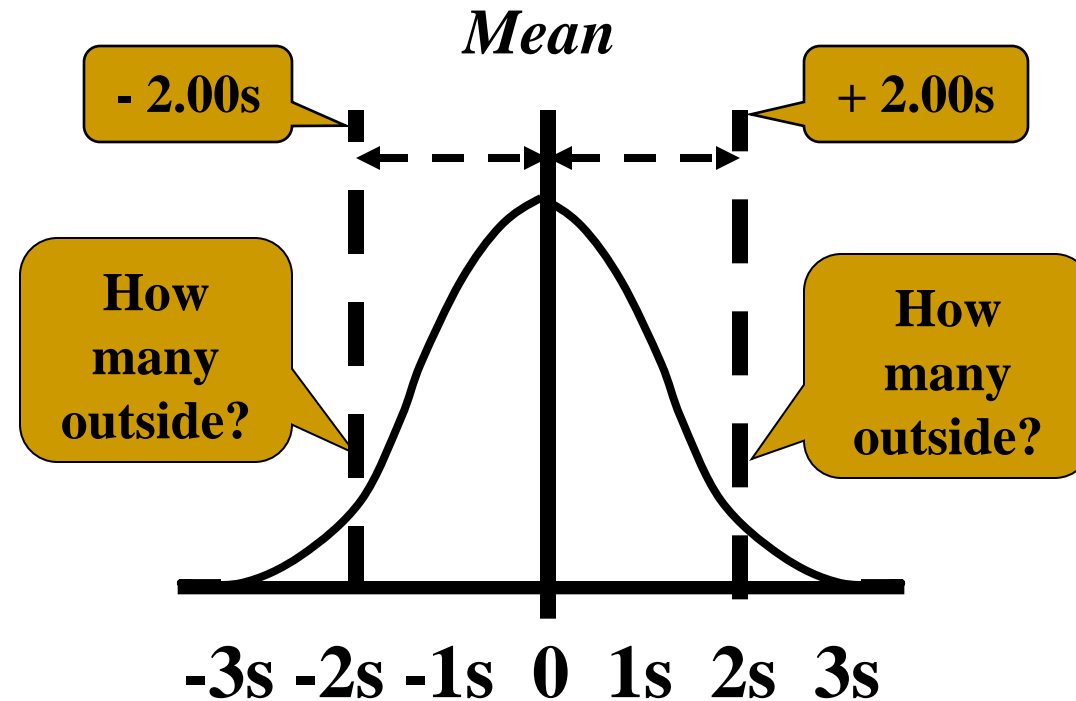
High volume laboratory in the USA:

- Almost 1,000 controls a day
- 50 flags in the morning, 50 flags at night (*as predicted*)
- 5 techs, each must check out 20 rejections every day

- 700 rejections per week
- 3,000 rejections per month
- >35,000 rejections per year
- Each tech will have to check 5,200 extra flags every year
- Even if these flags only take 30 seconds, that adds up to >43 hours of checking false flags
- *Five weeks of salary spent chasing ghosts for 5 employees*



HOW DO WE KNOW THERE'S A FALSE REJECTION PROBLEM?



Look up in "Table of area under a normal curve"

AREA UNDER A NORMAL CURVE

Z-value Area

0.00 0.500000

0.50 0.308538

1.00 0.158655

1.50 0.066807

 2.00 0.022750

2.50 0.006210

3.00 0.001350

3.50 0.000233

**First commandment
of statistics**

**Available in any
statistics textbook**

**Shows area in one
tail of a Gaussian
distribution**

- $0.02275 * 2 = 0.0455$
or 4.55% for both
tails of distribution



WESTGARD RULES REDUCED FALSE REJECTIONS

Full Westgard Rules for N=2 is about 3-4%
- HALF that of 2 SD limits with 2 controls

Full Westgard Rules for N=3 is about 3-7%
- HALF that of 2 SD with 3 controls

If fewer Westgard Rules are needed, false rejection goes even lower.



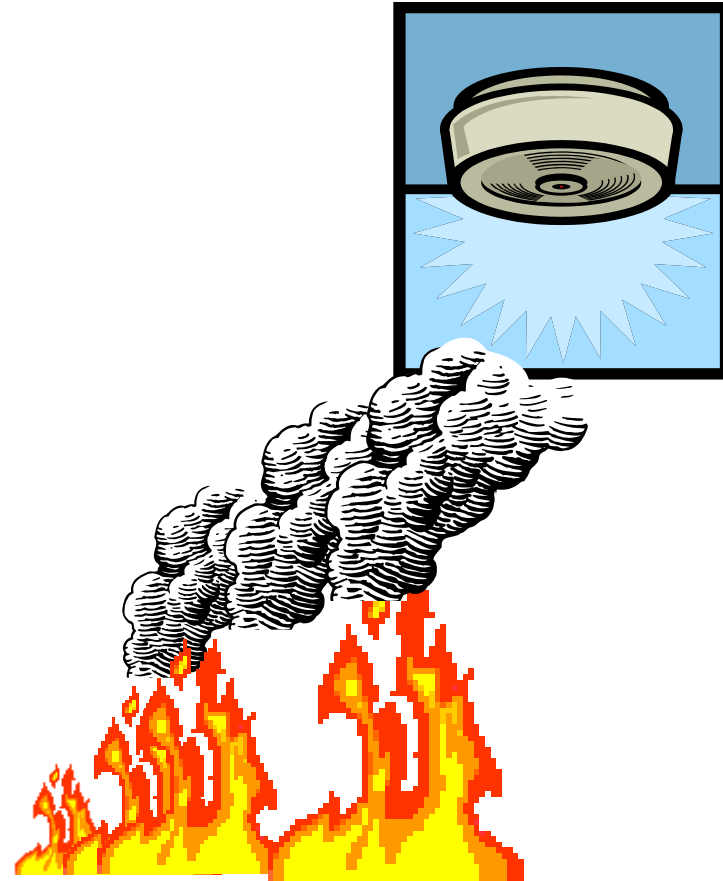
UNDERSTANDING ERROR DETECTION, FALSE REJECTION, AND POWER CURVES

False rejection:

The alarm goes off
BUT there IS NO real error

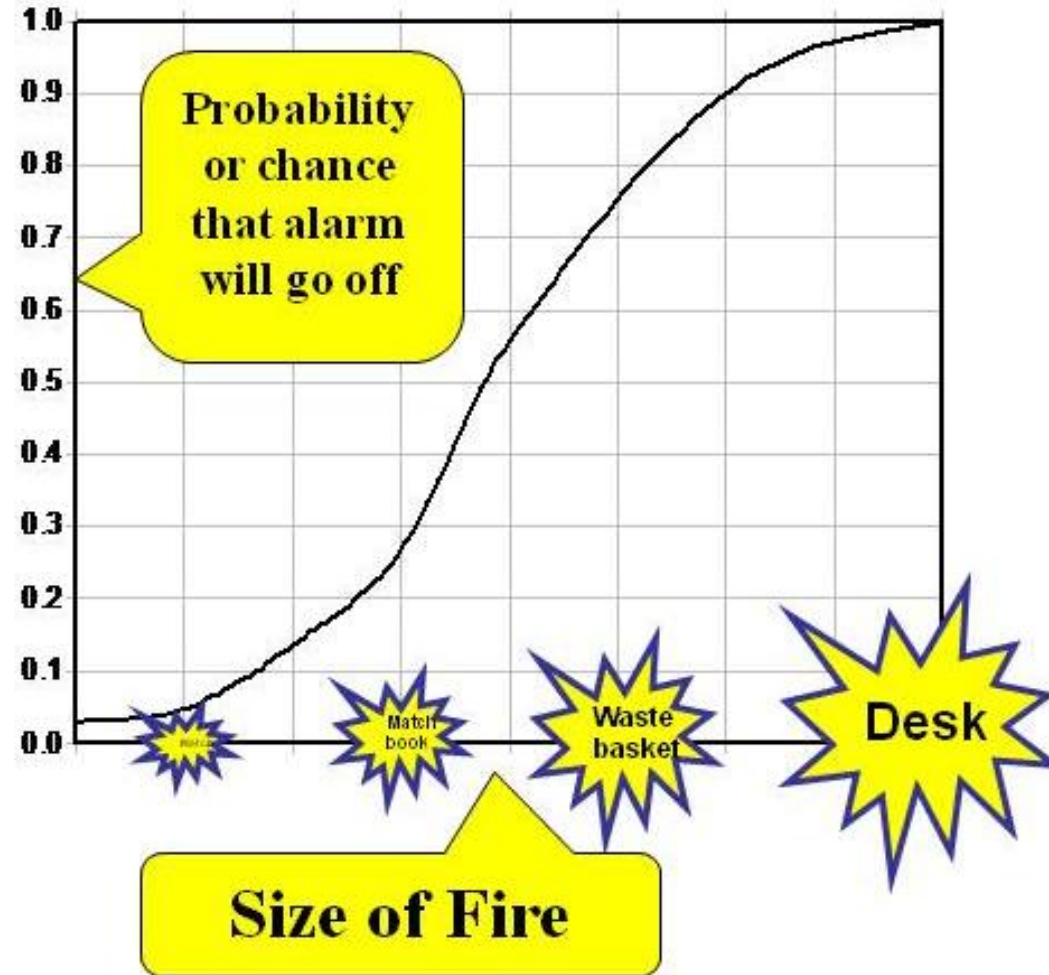
Error detection:

The alarm goes off
and there IS a real error

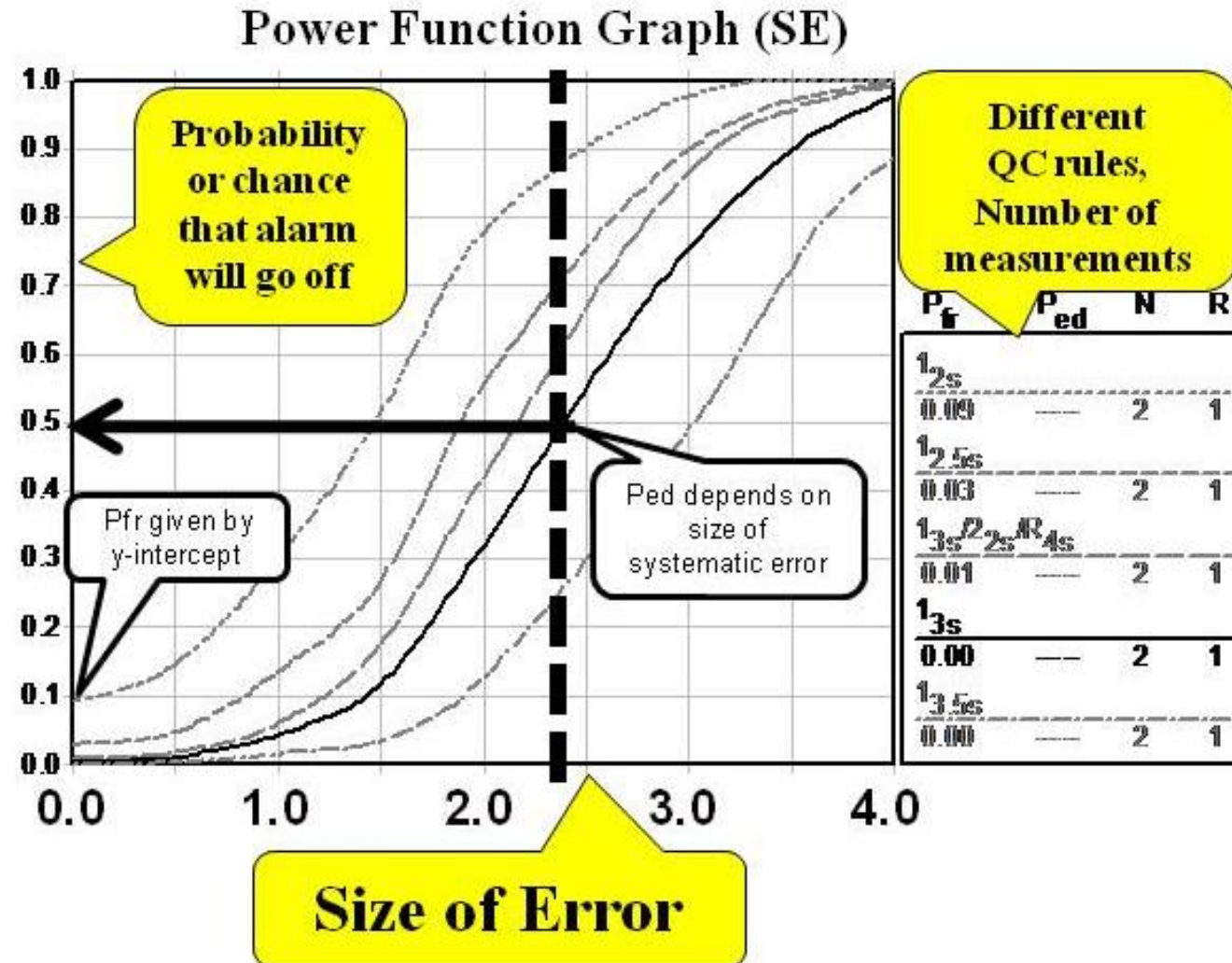


Also known as the sensitivity-specificity trade-off

“POWER” OF A “DETECTOR”



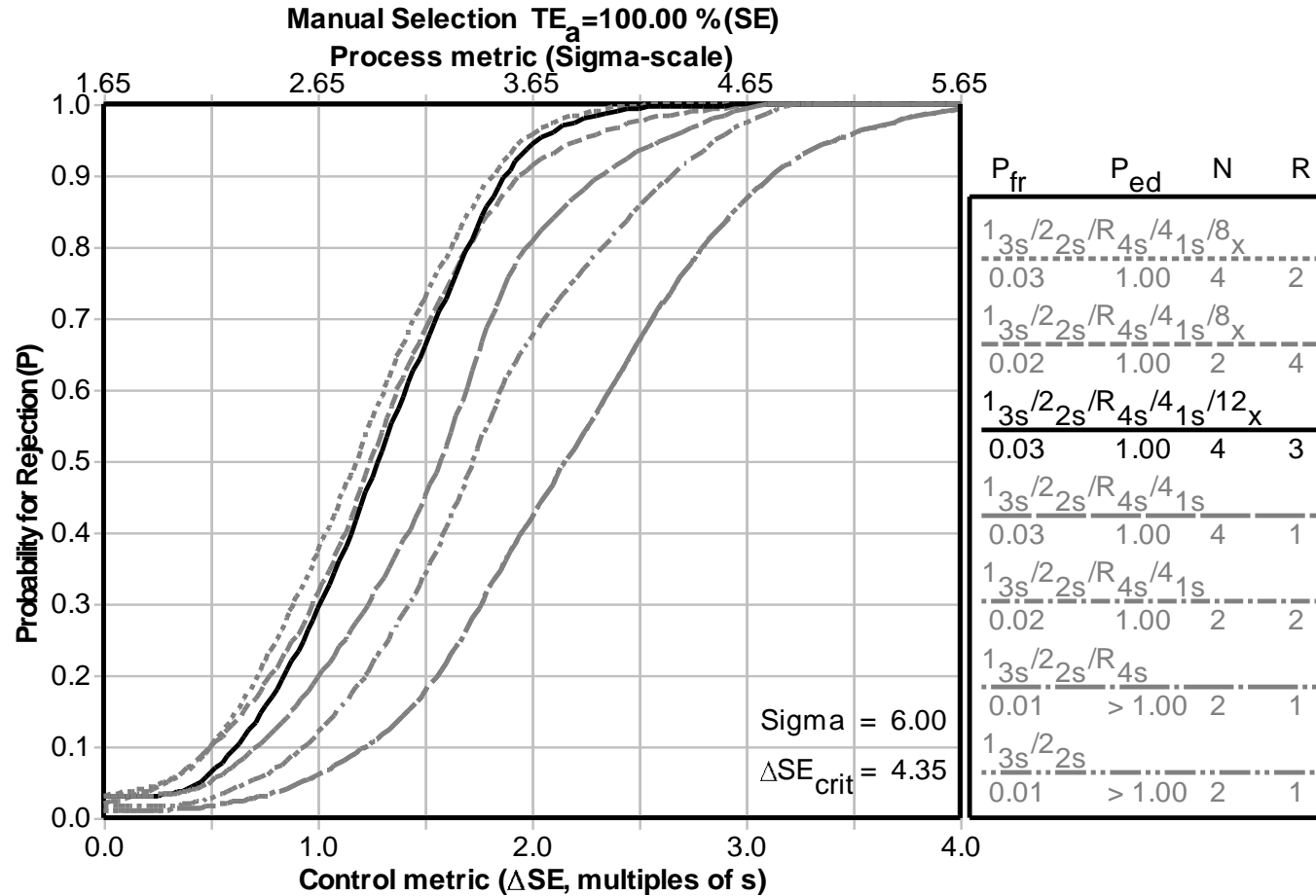
POWER OF ± 2 SD CONTROL LIMITS OR 1_{2s} CONTROL RULE



CRITICAL-ERROR GRAPH 6 SIGMA PROCESS

6 Sigma requires
 1 Westgard Rule
 1:3s

Additional warning
 rules will not add to
 error detection, but
 2:2s, 4:1s, 8:x will
 each add
 • + 1% Pfr



CRITICAL-ERROR GRAPH 5 SIGMA PROCESS

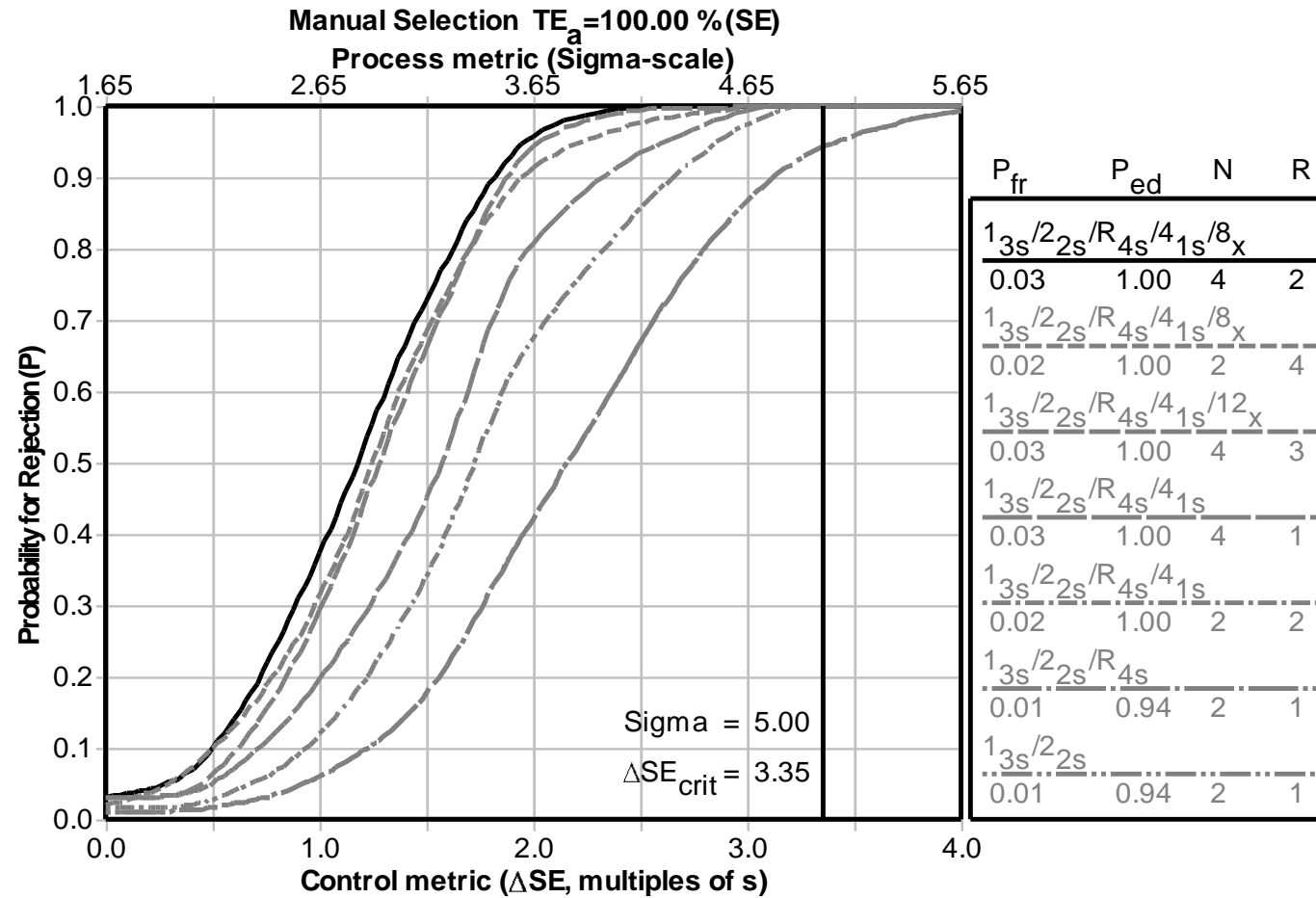
5 Sigma requires
 3 Westgard Rule
 1:3s/2:2s/R:4s

Additional warning
 rules will not add to
 error detection, but

- 4:1s will add
- + 1% Pfr
 - + 6% Ped

8:x will add

- + 1% Pfr
- No additional Ped

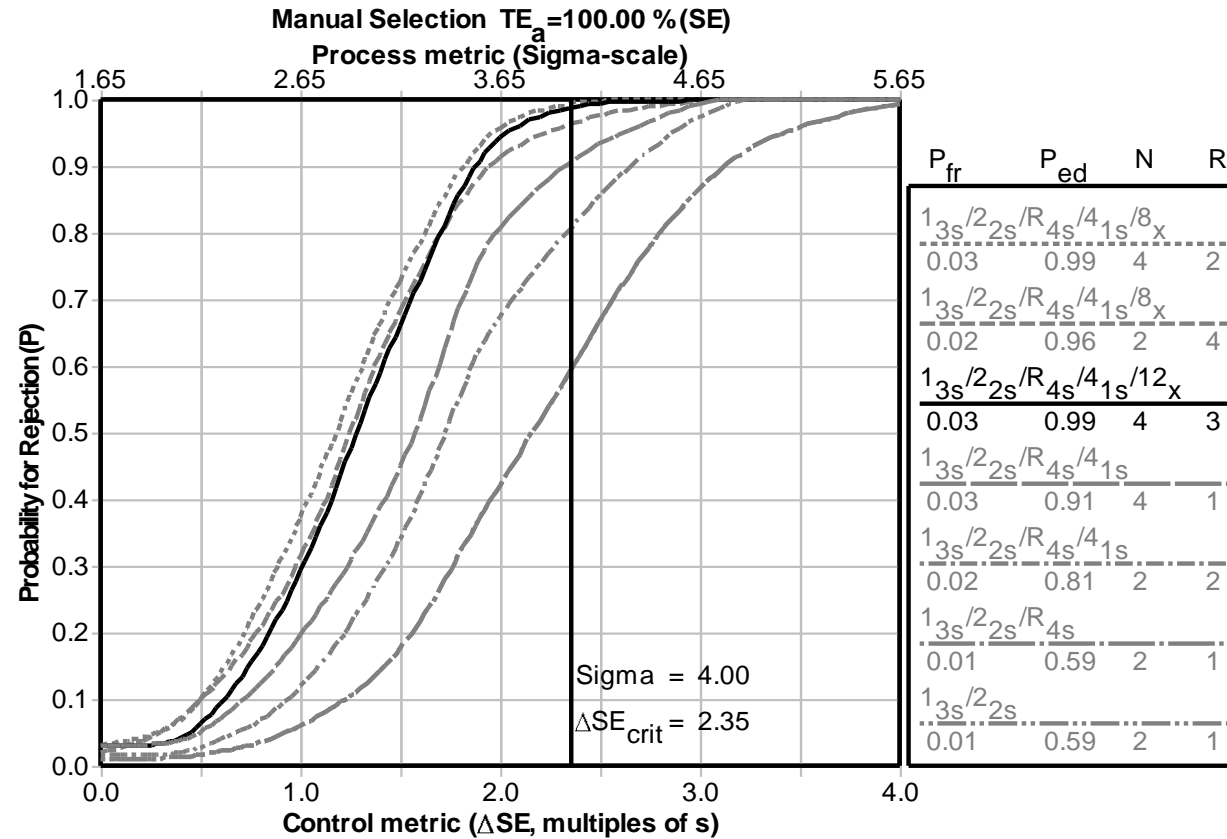


CRITICAL-ERROR GRAPH 4 SIGMA PROCESS

4 Sigma requires
 4 Westgard Rules
 1:3s/2:2s/R:4s/4:1s/

Additional 8:x will

- + 8% Ped
- + 1% Pfr



CRITICAL-ERROR GRAPH FOR 3 SIGMA PROCESS

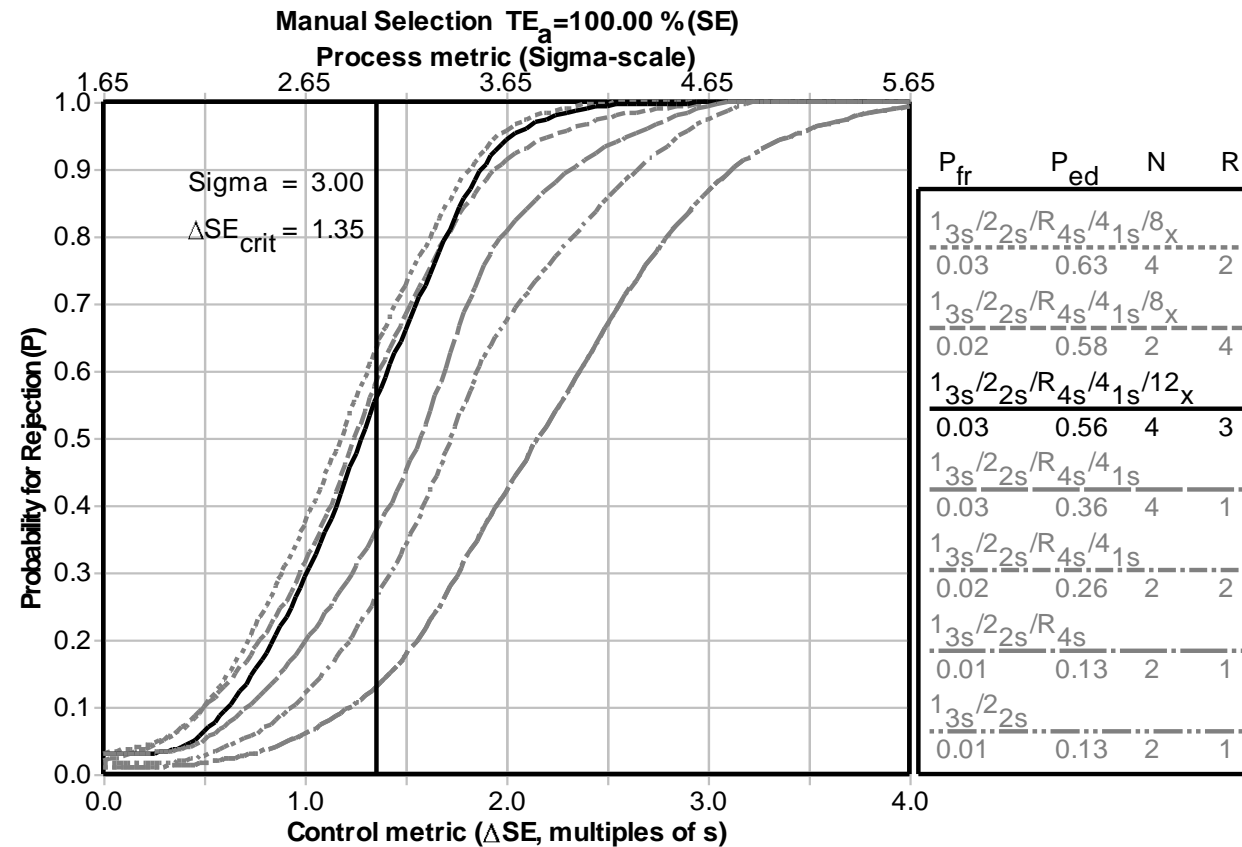
3 Sigma process
 requires ALL of
 the Westgard
 Rules

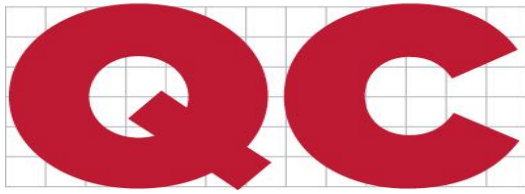
For Rejection!

No Warnings!

Even then:

- Ped 58%
- Pfr 3%





IN SUMMARY

6 Sigma: Warning Rules increase P:fr without more P:ed

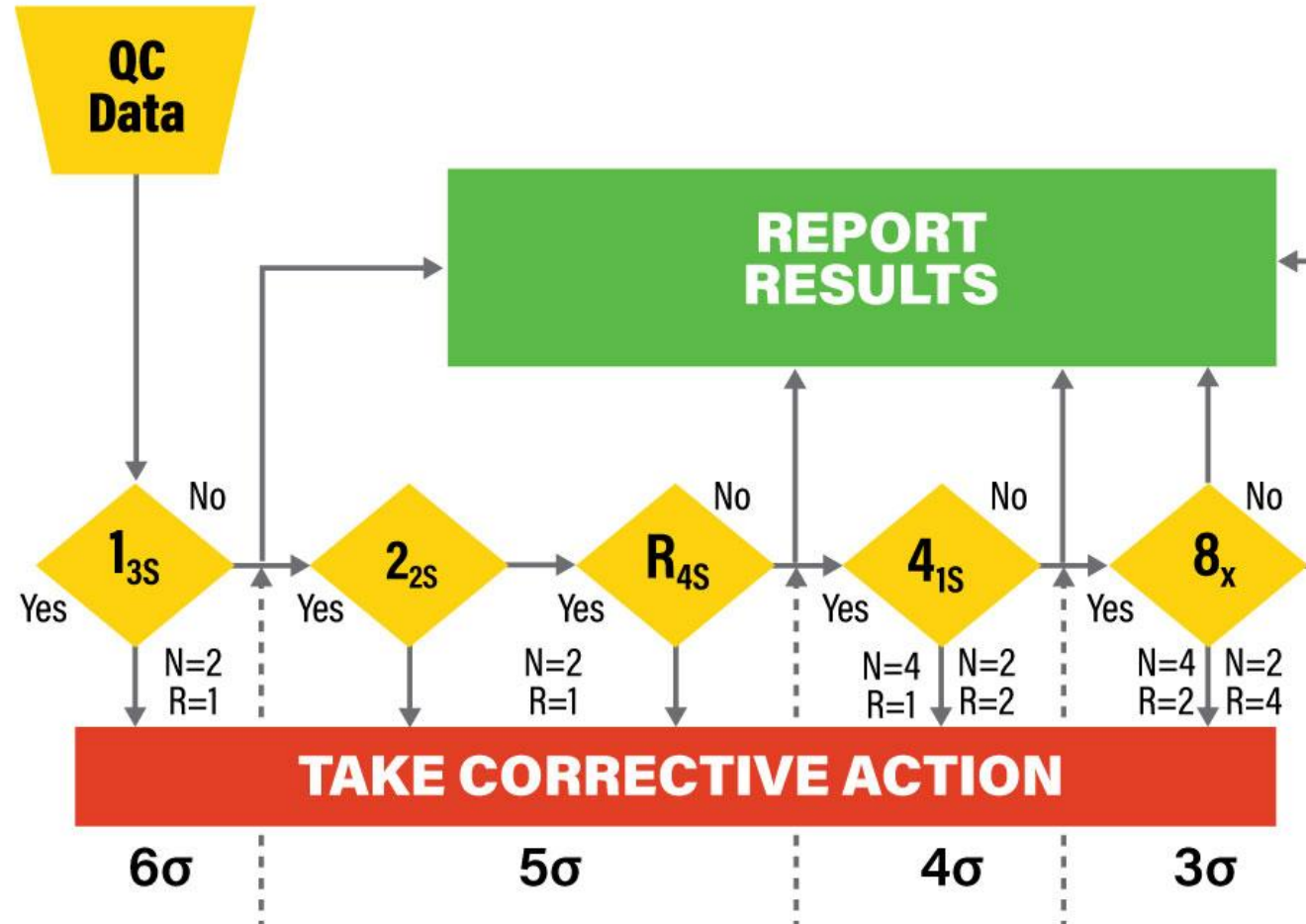
5 Sigma: Warning Rules increase P:fr without more P:ed

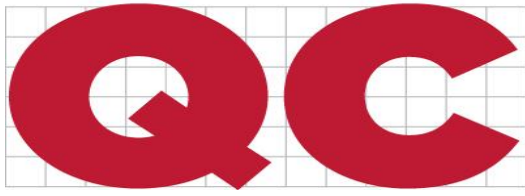
4 Sigma: Adding more rules will modestly increase in P:ed, definitely raises P:fr

**3 Sigma and below: You *can't* use Warning Rules
you need all the rules for rejection!**



WHAT DRIVES THESE DECISIONS? WESTGARD SIGMA RULES





CAN YOU COUNT TO 6? SOME SAY NO

Recent criticism in the literature has challenged the analytical Sigma metric.

1. Given an allowable total error (TEa), can you determine how many SDs fit within that space. That is, if you have a 10% TEa, how many times does a CV of 1.67% fit within it?

If you get an answer of 6, you're doing the math correctly.

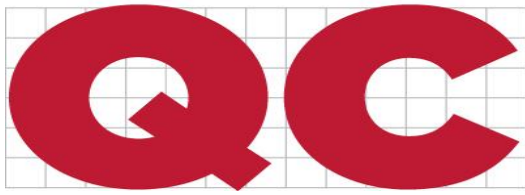
2. Do you believe that bias doesn't exist, that bias is always corrected by the manufacturer or the lab, and that bias transubstantiates into imprecision after an unknown amount of time, and that bias never appears as a shift?

If you answer no, then you accept bias as a linear shift, which can then be used in the analytical Sigma-metric.



SO WHY DO WE STILL USE WARNING RULES?





WHY

Risk Management – just in case – on critical tests

Losing sleep at night if some rules aren't in place

Tradition and Habit – it's been done this way for generations in the lab

The joy of responding to a false alarm

**But in these cases,
you're no longer doing Statistical QC,
you're doing Therapeutic QC**





WAIT? WE USE WARNING RULES BUT WE DON'T SEE LOTS OF OOCs

Using 2 SD but not seeing constant outliers?

Sorry. You haven't found an area of the universe where probability no longer applies.

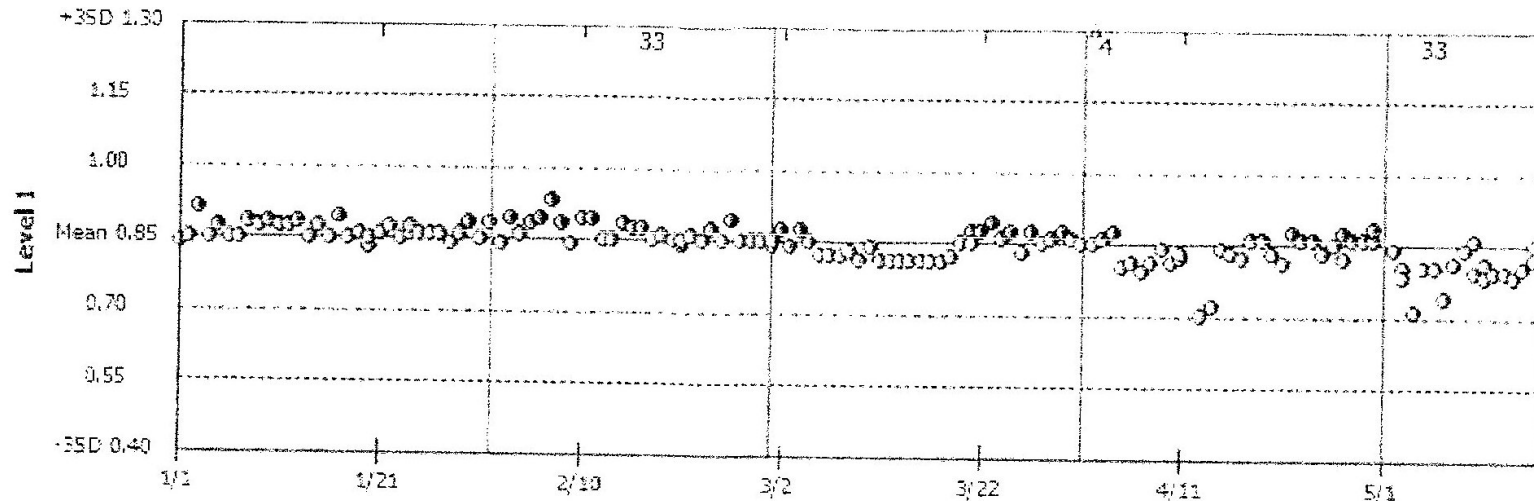
Instead, you've probably widened your limits artificially

- **Straight out "adjusting" of the SD / range**
- **Adoption of a manufacturer's SD / range**
- **Adoption of a peer group SD / range**



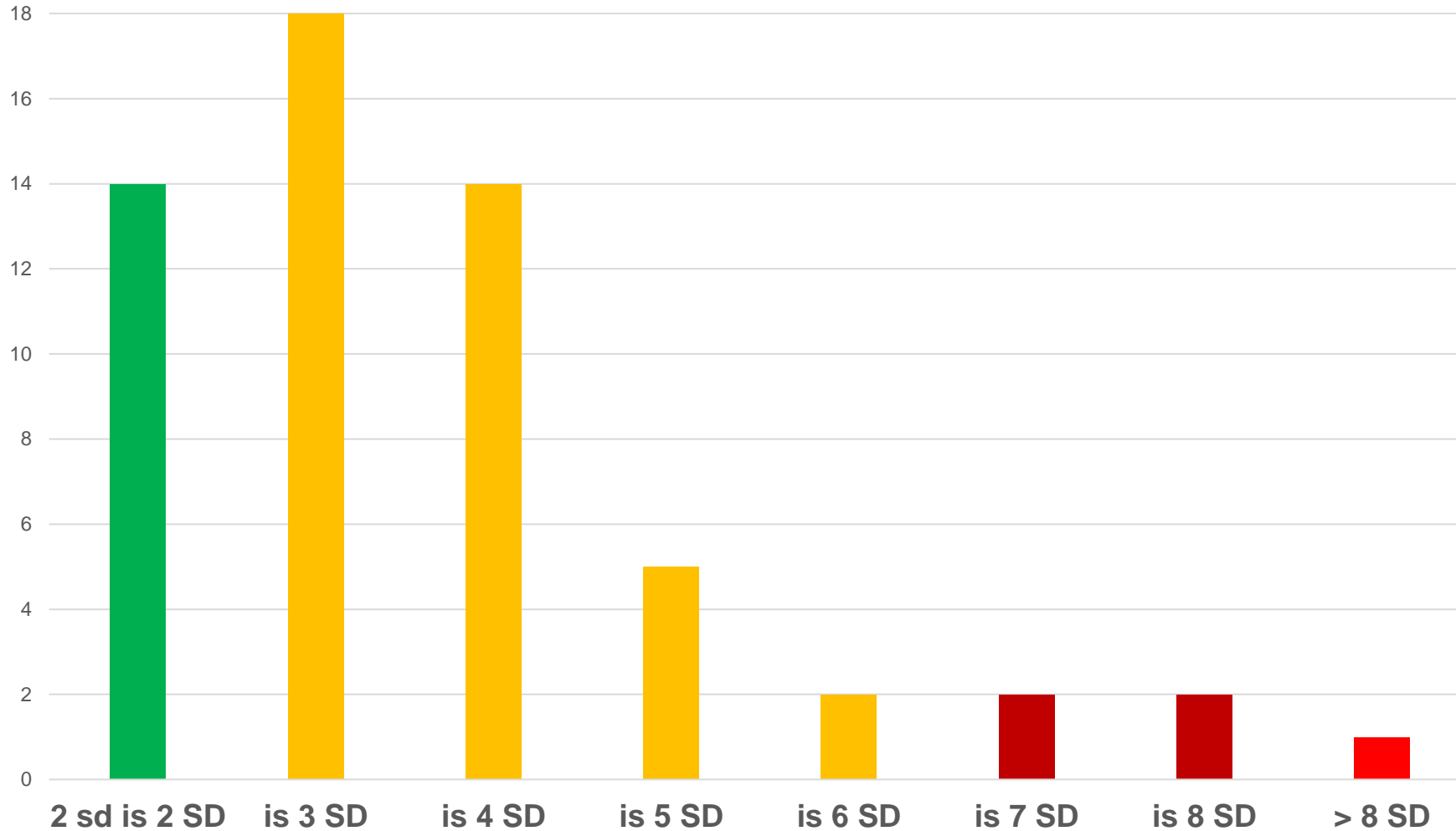
WHAT'S WRONG HERE? (ANONYMOUS US LAB)

- LJ Chart SD set to 0.15 mg/dL
- Actual cumulative SD is 0.04 mg/dL
- 1 SD limit is actually 3.75 SD limit!



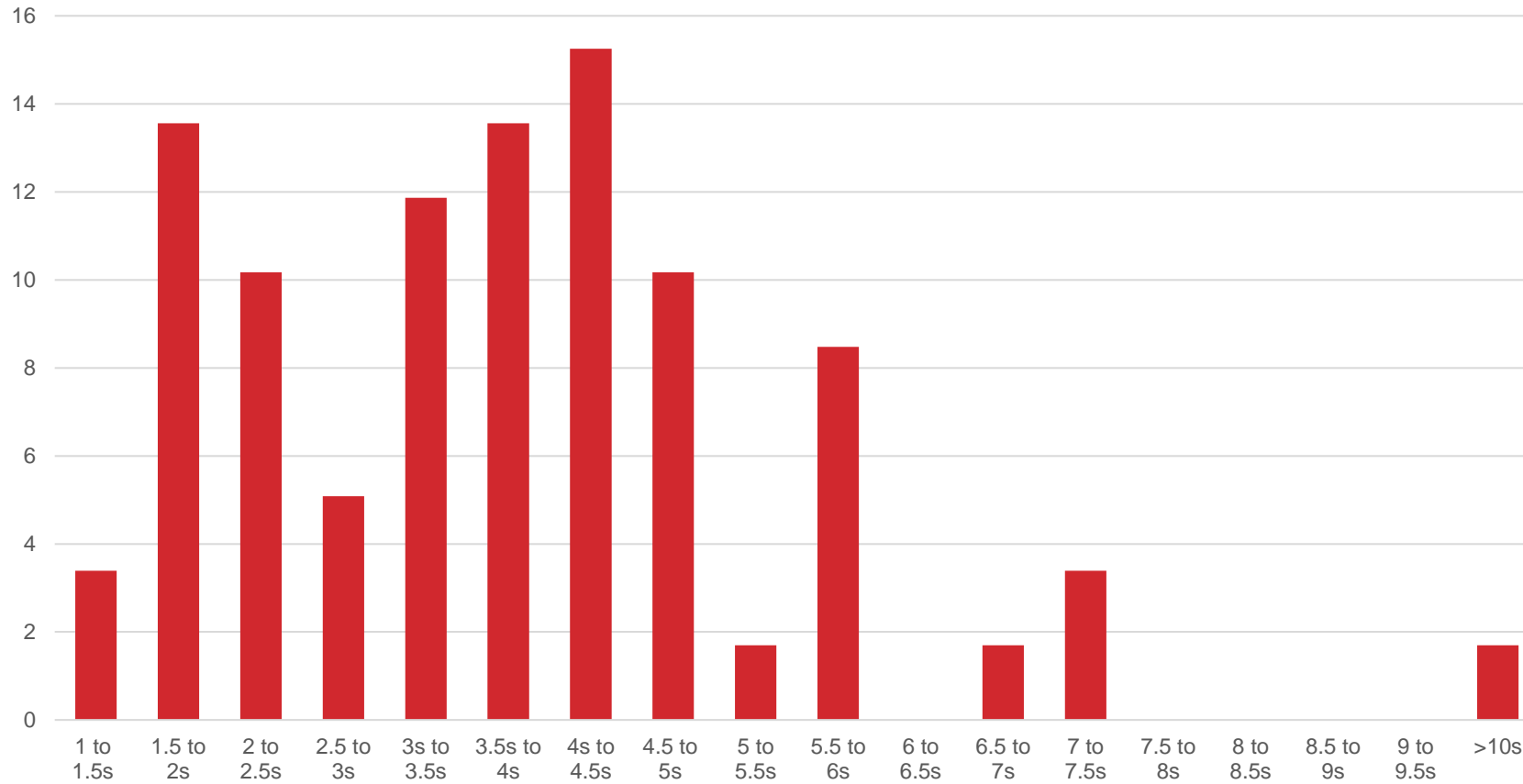
EXTENT OF BLIND MAN QC (MANUFACTURER RANGES)

Manufacturer SD recommendations for 2 SD are...
(58 levels, 29 biochemistry methods)



BLIND MAN QC IS STANDARD IN HEMATOLOGY

Hemoglobin Limits Recommended in the UK (59 labs)



WHEN YOUR “SD” IS NOT YOUR ACTUAL SD...

Wider limits centered on your **ACTUAL** mean

REDUCE false rejection (good)

REDUCE error detection (bad)

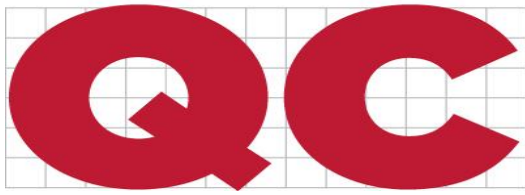
MISS medically important errors (worse)

Wider limits centered on a **DIFFERENT** mean

COULD BE MUCH WORSE...

DOUBLE-BLIND IN A BAD WAY





THE NEW TEMPTATION: “STANDARDIZING” SD & MEAN

High volume labs with multiple instruments running same tests

Networks/Systems of laboratories running same instruments and same tests

Why not treat them all the same? So the patient can flow through the system with a “standard” mean and range?



“STANDARDIZATION” IS A TEMPTING LIE

LOOKS simple – all tests have single mean, single SD, or both

Under the hood, reality persists: different instruments have different means and different SDs

When/How do you find out when one instrument is no longer “standardized”?

Either the lab monitors the REAL mean and SD, or it pretends reality doesn’t exist.



“STANDARDIZATION” REQUIRES KEEPING TWO SETS OF BOOKS

- 1. The “Standard” mean and SD of the lab / network**
- 2. The REAL mean and SD of each instrument**

At some point, the REAL must be compared against the “Standard” to see how close they are

**Small deviations or differences (how small?) may be acceptable.
*You must specify test-by-test.***

Larger deviations or differences will force the instrument out of the “Standard” herd. Recalibration or correction will be necessary.





IN ORDER TO HAVE A “STANDARD”, YOU STILL NEED TO CALCULATE THE ACTUAL MEAN AND SD

This actually isn't simpler – it's more complicated.

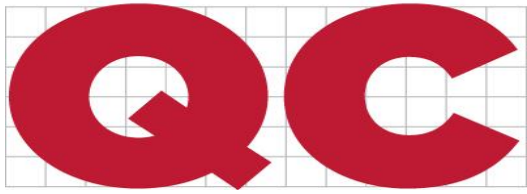
It not only requires *more* work. It *doubles* the work.

If the lab doesn't check “standardization”, *the patient is the check.*

Discrepant patient values are being used to flag “non-standard” methods and instruments.

Lag-Time Patient-Based QC (LTPBQC)





Letter to the Editor

Tony Badrick* and Jean-Marc Giannoli

Managing the Quality Control of multiple instruments

<https://doi.org/10.1515/cclm-2023-0592>

Received June 6, 2023; accepted August 16, 2023;
published online August 28, 2023

Keywords: Quality Control; network of multiple instruments;
clinical significance; error

To the Editor,

Despite very high engineering standards for analysers, there is always some variation in the outcome of the measurement of a particular sample between instruments, both in terms of

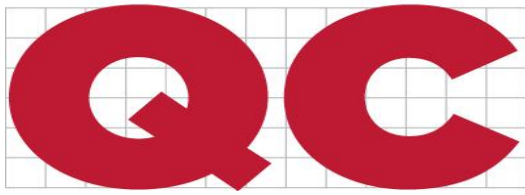
targets for all instruments (reference mean and SD). In fact, there are four possible models.

- a. The QC rule uses a mean, and an SD centered on the individual instrument mean and SD, that is, monitors each instrument as if it was stand-alone. The focus is on detecting statistically significant out-of-control error conditions.
- b. The QC rule uses a fixed mean for all instruments, but an SD based on the individual instruments' SD. That is, monitor each instrument with one mean but still use individual instrument SD as if it was stand alone.
- c. The QC rule uses an individual instrument mean and an

<https://www.degruyter.com/document/doi/10.1515/cclm-2023-0592/html>

“[T]he optimal theoretical approach in the situation of controlling multiple instruments is to use a common mean and SD, however, this assumes commutable QC material and non-significant bias between instruments.”





THERE'S A LOT OF FINE PRINT...

“A basic prerequisite is that there is **no significant bias between instruments** at the time of stable performance.

“This assumption must be verified.

“If the analytical systems are the same, then non-commutable QC material can be used to detect introduced bias. If valid comparisons are to be between instruments, then **instruments and reagents (same lot number) must be calibrated with the same calibrator using the same calibration procedure with a very small-time difference.**

“Of course, where changes in an assay occur that could lead to the introduction of bias such as a reagent lot change, the use of stored patient samples covering the measuring range should be used to ensure there is no impact on patients (lot to lot changes). Recalibration of an instrument or group of instruments can also introduce a bias.”



CAUTION IN THE LITERATURE

Impact of combining data from multiple instruments on performance of patient-based real-time quality control

Qianqian Zhou¹, Tze Ping Loh^{*2}, Tony Badrick³, Chun Yee Lim¹

¹Engineering Cluster, Singapore Institute of Technology, Singapore, Singapore

²Department of Laboratory Medicine, National University Hospital, Singapore, Singapore

³Royal College of Pathologists of Australasia Quality Assurance Programs, Sydney, Australia

*Corresponding author: tploh@hotmail.com

Abstract

Introduction: It is unclear what is the best strategy for applying patient-based real-time quality control (PBRTQC) algorithm in the presence of multiple instruments. This simulation study compared the error detection capability of applying PBRTQC algorithms for instruments individually and in combination using serum sodium as an example.

Materials and methods: Four sets of random serum sodium measurements were generated with differing means and standard deviations to represent four simulated instruments. Moving median with winsorization was selected as the PBRTQC algorithm. The PBRTQC parameters (block size and control limits) were optimized and applied to the four simulated laboratory data sets individually and in combination.

Results: When the PBRTQC algorithm were individually optimized and applied to the data of the individual simulated instruments, it was able to detect bias several folds faster than when they were combined. Similarly, the individually applied algorithms had perfect error detection rates across different magnitudes of bias, whereas the error detection rates of the algorithm applied on the combined data missed smaller biases. The performance of the individually applied PBRTQC algorithm performed more consistently among the simulated instruments compared to when the data were combined.

Discussion: While combining data from different instruments can increase the data stream and hence, increase the speed of error detection, it may widen the control limits and compromising the probability of error detection. The presence of multiple instruments in the data stream may dilute the effect of the error when it only affects a selected instrument.

Keywords: quality control; laboratory management; moving median; moving average; average of normal

Submitted: November 21, 2020

Accepted: February 28, 2021

<https://pubmed.ncbi.nlm.nih.gov/33927555/>

CONCLUSION: DON'T DO IT FOR PBRTQC

Results: “When the PBRTQC algorithm were **individually optimized** and applied to the data of the individual simulated instruments, it was able to **detect bias several folds faster** than when they were combined....”

Discussion: “While combining data from different instruments can increase the data stream and hence, increase the speed of error detection, it may **widen the control limits and compromising[sic] the probability of error detection**. The presence of multiple instruments in the data stream may dilute the effect of the error when it only affects a selected instrument.”



THE MATH ON “STANDARDIZED” MEANS IS DISCOURAGING

<https://journals.sagepub.com/doi/10.1177/00045632241226916>



The Association for
Clinical Biochemistry &
Laboratory Medicine
Better Science, Better Testing, Better Care

Zoom Out

Annals of Clinical Biochemistry
2024, Vol. 0(0) 1–7
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00045632241226916
journals.sagepub.com/home/acb

S Sage

Assay error detection when using common quality control targets across multiple instruments: An analysis using simulated and real-world data

Eric S Kilpatrick^{1,2}

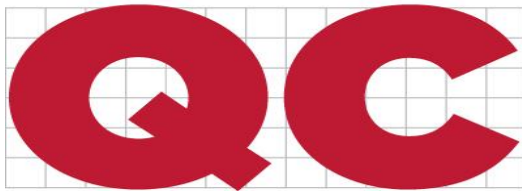
Abstract

Background: Clinical laboratories frequently implement the same tests and internal quality control (QC) rules on identical instruments. It is unclear whether individual QC targets for each analyser or ones that are common to all instruments are preferable. This study modelled how common QC targets influence assay error detection before examining their effect on real-world data.

Methods: The effect of variable bias and imprecision on error detection and false rejection rates when using common or individual QC targets on two instruments was simulated. QC data from tests run on two identical Beckman instruments (6-month period, same QC lot, $n > 100$ points for each instrument) determined likely real-world consequences.

Results: Compared to individual QC targets, common targets had an asymmetrical effect on systematic error detection, with one instrument assay losing detection power more than the other gained. If individual in-control assay standard deviations (SDs) differed, then common targets led to one assay failing QC more frequently. Applied to two analysers (95 QC levels and 45 tests), common targets reduced one instrument's error detection by ≥ 0.4 sigma on 15/45 (33%) of tests. Such targets also meant 14/45 (31%) of assays on one in-control instrument would fail over twice as frequently as the other (median ratio 1.62, IQR 1.20–2.39) using a 2SD rule.

Conclusions: Compared to instrument-specific QC targets, common targets can reduce the probability of detecting changes in individual assay performance and cause one in-control assay to fail QC more frequently than another. Any impact on clinical care requires further investigation.



WHAT DOES A COMMON MEAN... MEAN?

<https://journals.sagepub.com/doi/10.1177/00045632241226916>

“Common targets **reduced one instrument’s error detection by ≥ 0.4 Sigma on 15/45 (33%) tests**. Such targets also meant 14/45 (31%) of assays on one in-control instrument would **fail over twice as frequently as the other... using a 2 SD rule.**”

“Compared to instrument-specific QC targets, **common targets can reduce the probability of detecting changes in individual assay performance and cause one in-control assay to fail QC more frequently than another.**”



Abstract

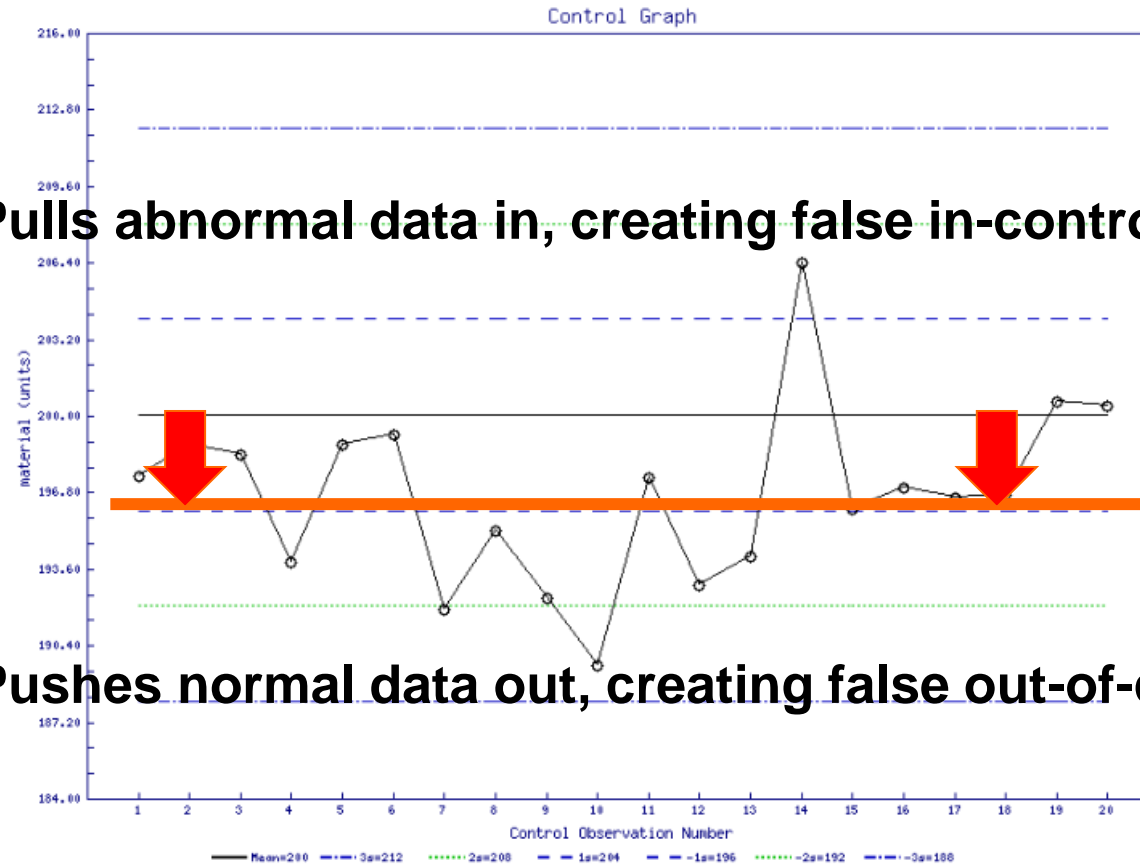
Background: Clinical laboratories frequently implement the same tests and internal quality control (QC) rules on identical instruments. It is unclear whether individual QC targets for each analyser or ones that are common to all instruments are preferable. This study modelled how common QC targets influence assay error detection before examining their effect on real-world data.
Methods: The effect of variable bias and imprecision on error detection and false rejection rates when using common or individual QC targets on two instruments was simulated. QC data from tests run on two identical Beckman instruments (6-month period, same QC lot, n = 100 points for each instrument) determined study real-world consequences.
Results: Compared to individual QC targets, common targets had an asymmetrical effect on systematic error detection, with one instrument assay losing detection power more than the other gained. If individual in-control assay standard deviations (SDs) differed, then common targets led to one assay failing QC more frequently. Applied to two analyzers (95 QC levels and 45 tests), common targets reduced one instrument’s error detection by ≥ 0.4 sigma on 15/45 (33%) of tests. Such targets also meant 14/45 (31%) of assays on one in-control instrument would fail over twice as frequently as the other (median ratio 1.62, IQR 1.20-2.39) using a 2SD rule.
Conclusions: Compared to instrument-specific QC targets, common targets can reduce the probability of detecting changes in individual assay performance and cause one in-control assay to fail QC more frequently than another. Any impact on clinical care requires further investigation.

“STANDARDIZED” MEAN AND SD HURTS TWICE!

Chart Group
mean=200
Lab SD=4.0
Observed Lab
mean=196.7
Lab SD=3.7

Pulls abnormal data in, creating false in-control points

Pushes normal data out, creating false out-of-control points



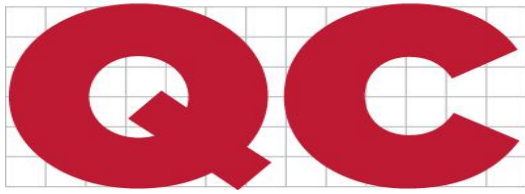
ONE FINAL FLAW...

Most discussions of these “manufacturer range” “standardization” techniques make a **HUGE** assumption, that **ALL TESTS PERFORM THE SAME.**

Of course we know that *different* tests have *different* performance.

Question: Are poor performing tests more or less eligible to use manufacturer ranges?





THE ROLE OF RULES: LEARN *HOW* TO USE THEM, NOT JUST WHAT THEY ARE

Rejection Rule: If it's out, we stop the run, trouble-shoot, fix something, start up again

“Warning Rule” – CLASSIC: A “Heads-up” to *start checking* all the rejection rules

“Warning Rule” – MODERN: A “Heads-up” to *anticipate* a developing problem

“Trouble-shooting Rule”: Using multirules *after* a rejection rule has been triggered – to figure out what might be wrong



MEAN AND SD “STANDARDIZATION”

Apparently simple, but actually
more than twice the work

Constant monitoring of bias required to maintain any standardization

“Standard” means will cause *more* false rejection on one side as was as
less error detection on the other side.





WHY WE NEED QC (AND SIX SIGMA)

"I spent many long nights independently doing training on QC systems and Westgard Rules when I worked at Theranos.

"Before reporting Theranos... to CMS, I tried to collect and present lots of evidence on how our QC systems were severely failing and it wasn't just my opinion, but violated basic QC procedures, Westgard Rules, and was far from Six Sigma laboratory principles. I was fortunate to be a young scientist who stumbled upon all the content you developed and was able to leverage it to understand how a company was endangering patients..."

- Erika Cheung

FROM THE DIRECTOR OF GOING CLEAR





YOUR GREATEST OBSTACLE? PARALYSIS

Your current QC practices have been in place for years, possibly decades.

The effort to make a change seems too hard, too much

Continuing to do the wrong thing because it's convenient is a long-term death spiral, waste of time, resources, and money (more accurately: money, money, and money), plus it is a drain on morale.

Moral of this webinar: It's time to change your QC.





**THANK YOU FOR YOUR KIND
ATTENTION!**



sten@westgard.com

