Approaches to Validating New CD4 and Viral Load (VL) Technology

The MapQuest Approach

Rebecca Gelman

Boston, MA, USA

MAPQUEST

You enter:

- where you are now
- where you want to go
- what you want to minimize:
  - miles, time, tolls, complexity

The web site returns step by step directions

(but you don’t always agree with route)

STATISTICAL CONSULT ON STUDY DESIGN

You enter:

where you are now

where you want to go

what kind of car (test material) you have

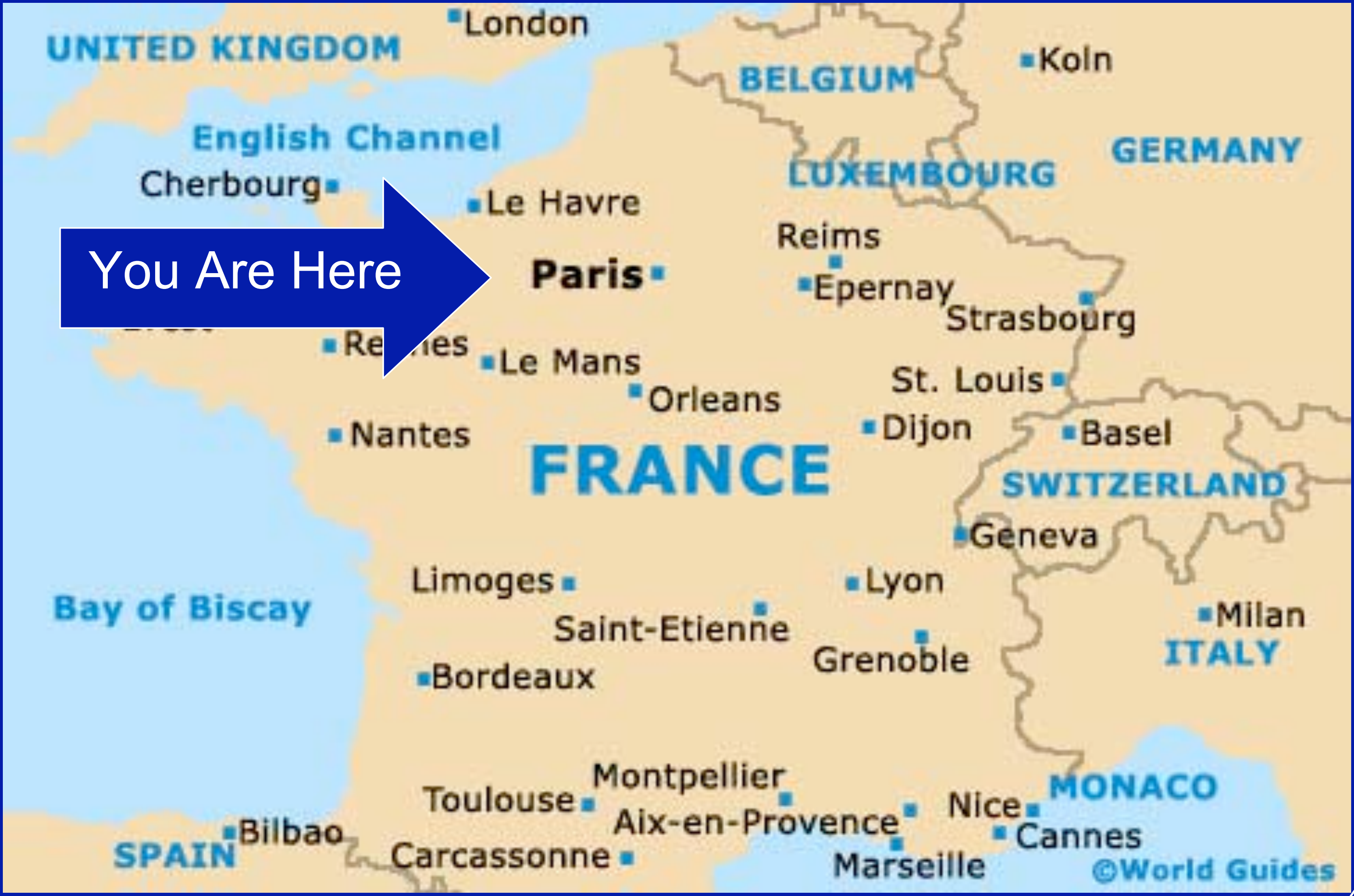
And the statistician will:

infer what you want to optimize

return step by step directions

(But you might want to discuss optimization)

Where you are



You Are Here

THE ROUTE YOU TAKE DEPENDS ON WHERE YOU ARE NOW:

multiple labs vs. single lab

> **multiple platforms/machines/reagents/**
> **storage/times/sample prep vs. not**

“clinical care” vs. “research” (group or single)

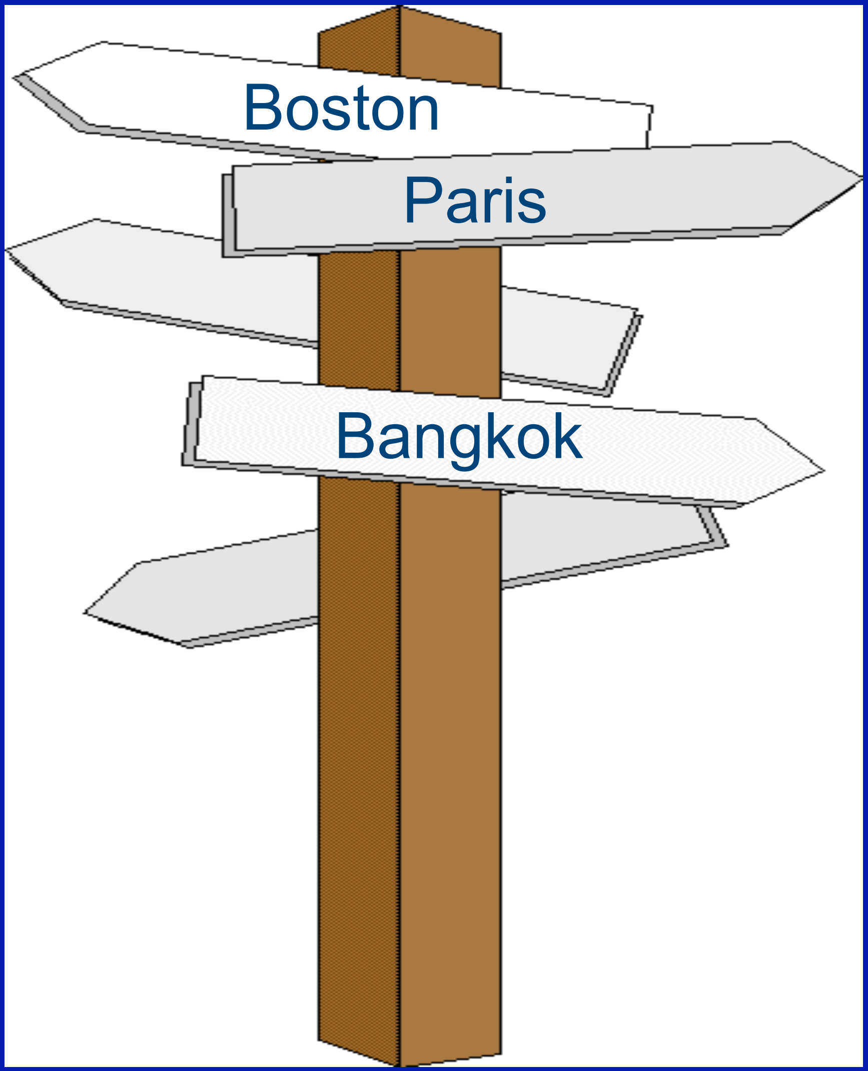
old test used for <u>ongoing</u> clinical care or research

old test considered “gold standard”

how assay is now used:

> **continuous vs. dichotomous**
>
> **diagnosis**
>
> **guidelines to start/change therapy**
>
> **evaluate response to therapy**

Where you want to go



THE ROUTE YOU TAKE DEPENDS ON WHERE YOU WANT TO GO:

add a technology vs. replace one

new gold standard vs. one of many assays

compare several technologies at once

mandate vs. suggest vs. allow switch

show new technology better vs. no worse

prediction/variability/bias/old blood/cost/

ease of use/shelf life

What kind of car you have

THE ROUTE YOU TAKE DEPENDS ON WHAT KIND OF CAR (TEST MATERIAL) YOU HAVE:

specimens with known results

via gold standard or created specimens

ability to "spike" specimens and create panel of specimens with known ratios

ability to reliably obtain donors with known "high" or "low" values

practicality of:

sending aliquots from same donor to multiple labs

obtaining enough blood from single donor for replicates

MapQuest isn’t good at telling you:

where you are

where you want to go

what kind of car you have

A statistician can at least:

ask you pointed questions

give an opinion on these matters

advise you to change your answers

Let’s go inside the MapQuest “black box” for the statistician’s view of the problem.

1. Dichotomous assay result

CD4 and VL are both continuous results, but they may be used dichotomously:

CD4 < 200 (or 350) to start ART

CD4 < 100 (or 50) to start OI prophylaxis

VL “undetectable” (< LLD) for ART response

VL “detectable” (> LLD) for diagnosis

1. Dichotomous, ctd.

However, even if assay used dichotomously, there are good reasons to study continuous differences:

how large a CD4 misclassified as <200; how large a VL misclassified as <LLD

get estimates of bias and variability
in case cut-offs change in future

difficulty getting many specimens close to cutoff value

(often) more power (fewer specimens needed) for continuous result

Dichotomous Methods

A. Two technologies

   i. Gold standard (or from studies)

      a. Random sample of specimens

      b. Not random (or stratified)

   ii. No Gold standard

      a. Specimens randomly chosen

      b. No good methods if not random

B. Extensions

   i. Tricotomous – not very relevant

   ii. More than two technologies

N specimens chosen randomly from population

	Gold Standard Result		
New Test	Good	Bad	Total
Good	a	b	a+b
Bad	c	d	c+d
Total	a+c	b+d	N

Sensitivity = a / (a+c)   Specificity = d / (b+d)

PPV = a / (a+b)   NPV = d / (c+d)

Sample size based on width of exact binomial CI (confidence intervals) for 2 of above 4 quantities (some adjustment necessary if CIs for all 4)

P values are inappropriate!!!

Increase number of old test “good” or “bad” results or stratified random sample of results

	Gold Standard Result		
New Test	Good	Bad	Total
Good	a	b	a+b
Bad	c	d	c+d
Total	a+c	b+d	N

Sensitivity = a / (a+c)   Specificity = d / (b+d)

PPV and NPV =
ftn (sens., spec., true good/bad ratio in pop.)
using Bayes Theorem; CIs are complicated

Sample size based on width of CIs

P values are inappropriate!!!

No gold standard

	Old Assay		
New Assay	Good	Bad	Total
Good	a	b	a+b
Bad	c	d	c+d
Total	a+c	b+d	N

Agreement = (a+d) / N and base N on binomial CI width

(could test if agreement > minimum acceptable level)

OR

Test if c = b vs. c > b (see if new assay puts more in good category than old assay)

McNemar test based on c + d, not on N (so total sample size N may need to be fairly large)

1. Dichotomous, ctd.

For CD4 and VL, extending to endpoints with 3 categories (good, bad, indeterminate) is not usually relevant

Don’t drop indeterminate results from analysis!

More often – test more than two technologies

Kappa statistics NOT useful – only test that agreement is better than random

May want to know if one test (of M) is most likely to disagree with others

Designs Between Dichotomous and Continuous

For VL, If have ordered specimen panel of known values, can compare methods based on which is first specimen > LLD

e.g., WHO for diagnosis

Early in development of new continuous assay, may want to know if two technologies usually get same rank order on panel of results (not sufficient for continuous measure)

Estimate or test difference in ranks

e.g. early CD4 studies, functional assays

2. Continuous Methods

What is important to you?

a) Bias (difference in assay result between two technologies)

b) Within-laboratory variability

c) Between-laboratory variability

You can look at differences or ratios or differences of logs or some other transforms

However you measure them, which of a, b, c are likely to be approximately constant over a reasonable range of assay result values?

Continuous Methods

Other things that might be of interest I won’t talk about here:

d) Between-technician variability

e) Can result of method Y be approximated by a linear function of method X? That is,

Y - a - bX is always very small for some constants a and b (allows fudge factor)

f) Which method has smallest difference between fresh and stored (old) blood

g) Which method requires smallest amount of blood per specimen to have reasonably small within-specimen variability

2. Continuous, comments on VL and CD4 (and many other continuous lab tests)

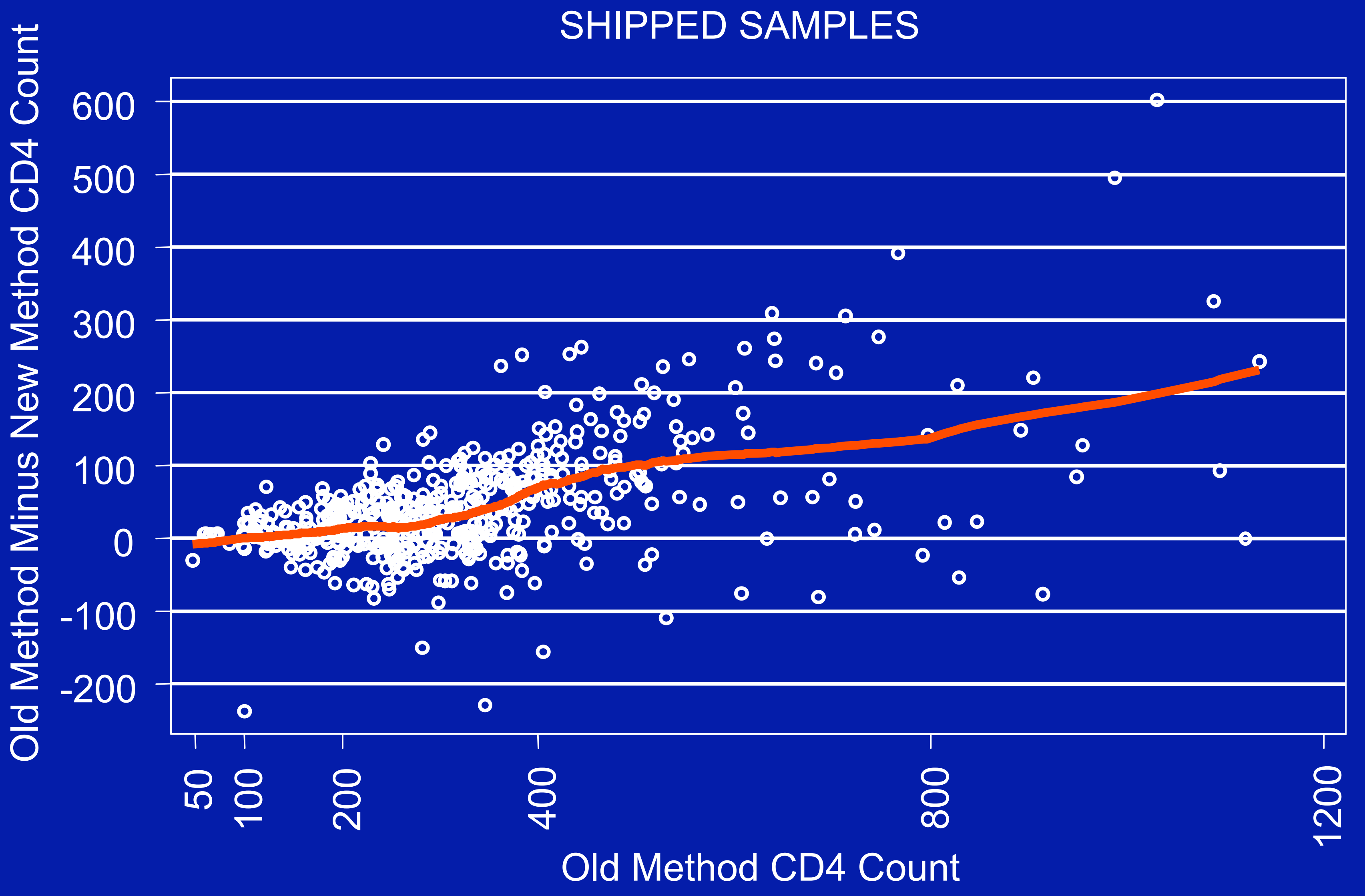
a) Difference in results (of two assay methods) tends to get bigger as results get bigger

e.g., if CD4 < 250, average bias = 10

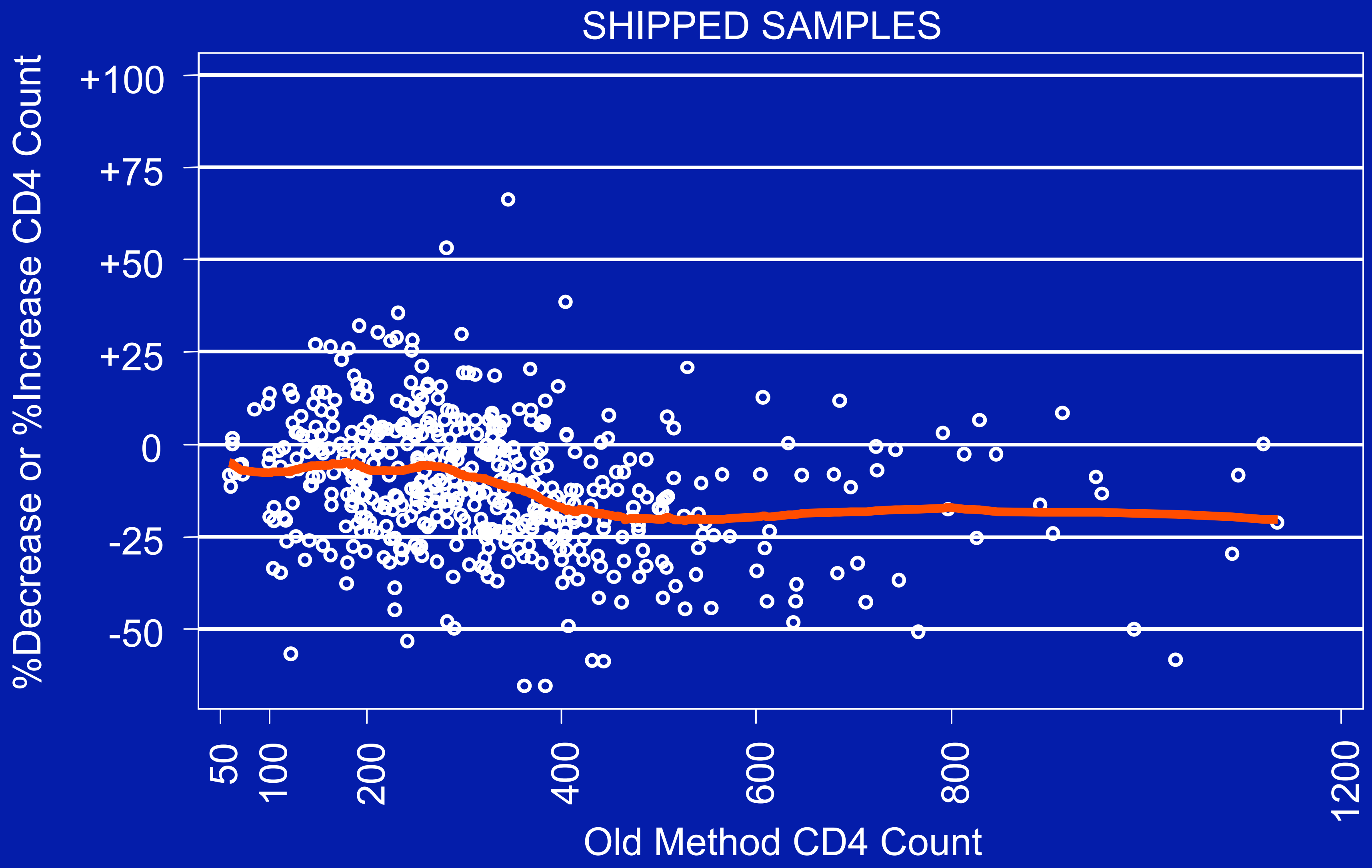
if CD4 > 250, average bias = 60

Sometimes ratio of values (or difference in log values) is more constant

Old Method Minus New Method CD4 Count



New Method Divided by Old Method CD4 Count



2. Continuous, comments on VL and CD4 (and many other continuous lab tests)

b) For a single assay method, standard deviations (between or within lab) are usually larger for bigger VL or bigger CD4

But sometimes standard deviation is also large for small VL (near LLD) or small CD4 (near 0).

c) Differences between standard deviations of two technologies are sometimes (not always) larger for bigger VL or bigger CD4.

2. Continuous, comments on VL and CD4

Define CV (coefficient of variation) as mean/std. dev. or median/IQR

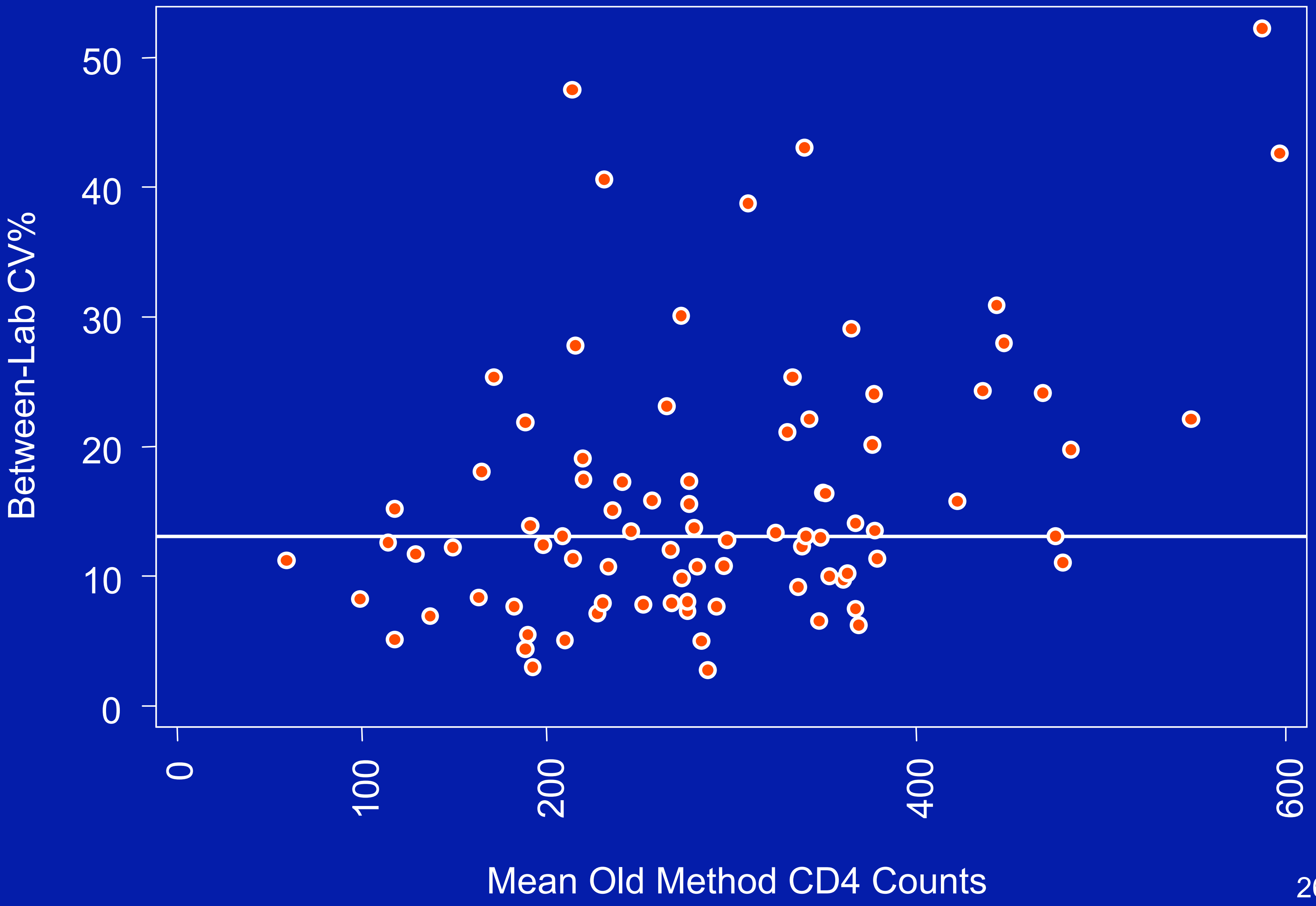
d) For a single assay method, CVs (between or within lab) are often (not always) more constant than standard deviations

Sometimes CV is also large for small VL or small CD4

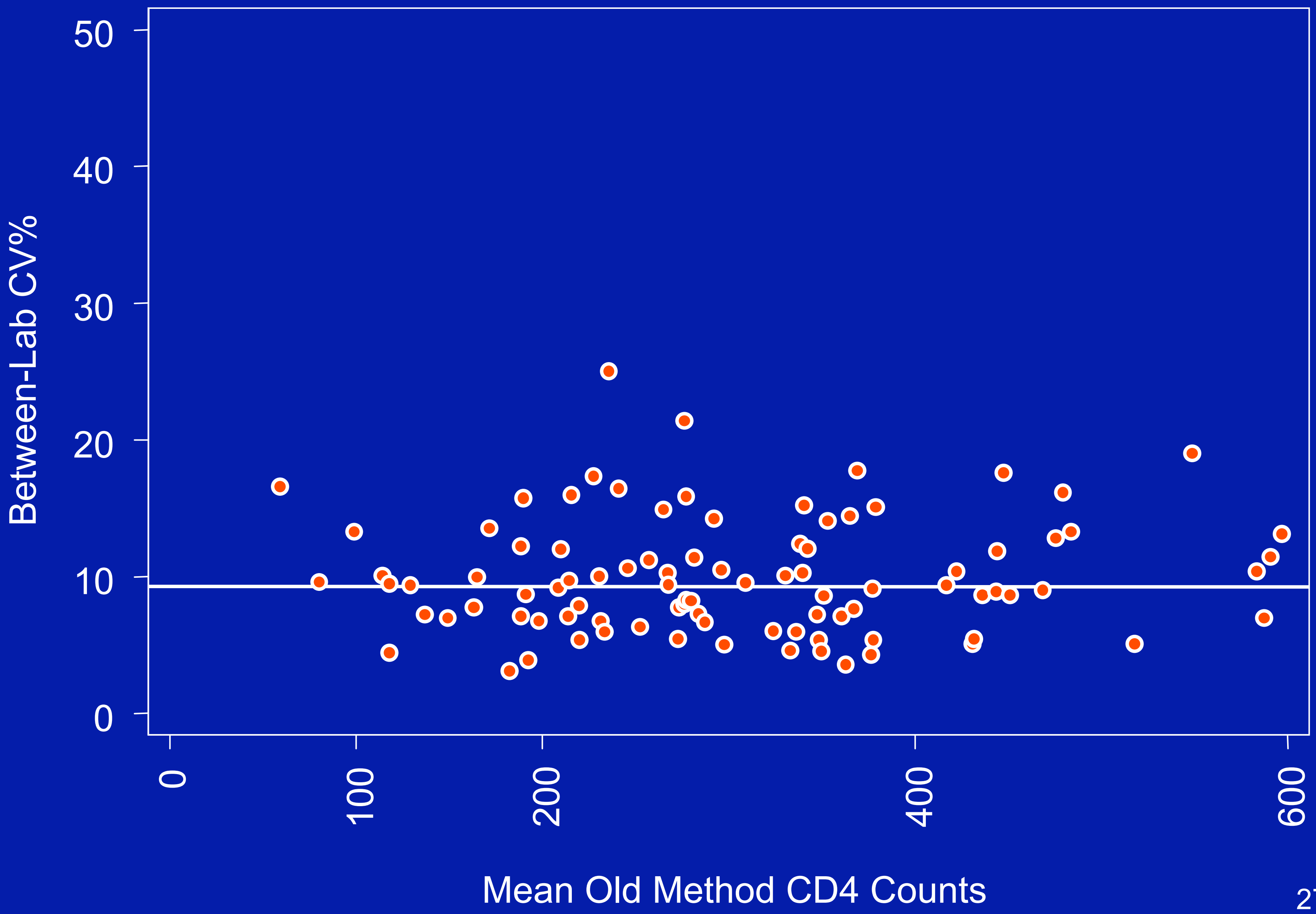
e) Differences between CVs of two technologies are frequently more constant than differences of standard deviations.

But often ratios of CVs are more constant over range of VL or CD4 than are differences of CVs

Old Method Between-Lab CV%



New Method Between-Lab CV%



Continuous Methods

A. Two technologies

Gold standard or no gold standard same

Spiked panel of specimens, random sample, and stratified random sample all same

But CV and bias usually not constant over whole range, so studies with different distributions of CD4 or VL may not be comparable

B. Extensions to more than two

Continuous Methods

Sample size and analysis can be based on:

a) CI for bias or test if bias = 0 or if bias < 2

do both assays on same specimens, so paired test

Wilcoxon sign rank better than t test (more robust)

If do k replicates of assay 1 and k replicates of assay 2 on same donor, randomly pair results for k pairs, not k^2 pairs

b) CI for difference in CVs or test if one CV larger

CVs should be paired, so paired test (Wilcoxon)

c) To test if comparison of bias (or CV) in two methods differs in different groups of specimens, can do Wilcoxon rank sum (unpaired Wilcoxon)

Continuous Methods

If more than two technologies, say M > 2:

Can do pairwise comparison of each to old method or to gold standard. Can’t do all M(M-1)/2 comparisons even if adjust for multiple comparisons (because they are dependent)

There aren’t well-known extensions of Wilcoxon sign rank (paired) test (but there are some)

The extension of the Wilcoxon rank sum test (to more than 2 groups of specimens) is the Kruskal-Wallis test

Special Size Considerations for CV

True CV is estimated by $\hat{C}$ = std. dev. / mean or $\hat{C}$ = IQR / median.

The bias of $\hat{C}$, i.e., CV – $\hat{C}$, is a function of 1/N and so is large for N < 10 or so.

As sample size increases, $\hat{C}$ increases (so 50 lab study has bigger $\hat{C}$ than one based on 10 labs).

IQR is more robust (less affected by outliers) than std. dev. if N > 7.

I don’t practice what I preach (have done studies with 8 replicates per donor and 5 labs).

The studies of new CD4 technologies I have worked on recently have had sample sizes based on testing if new method has better between-lab CV (and to a lesser extent, better within-lab CV).

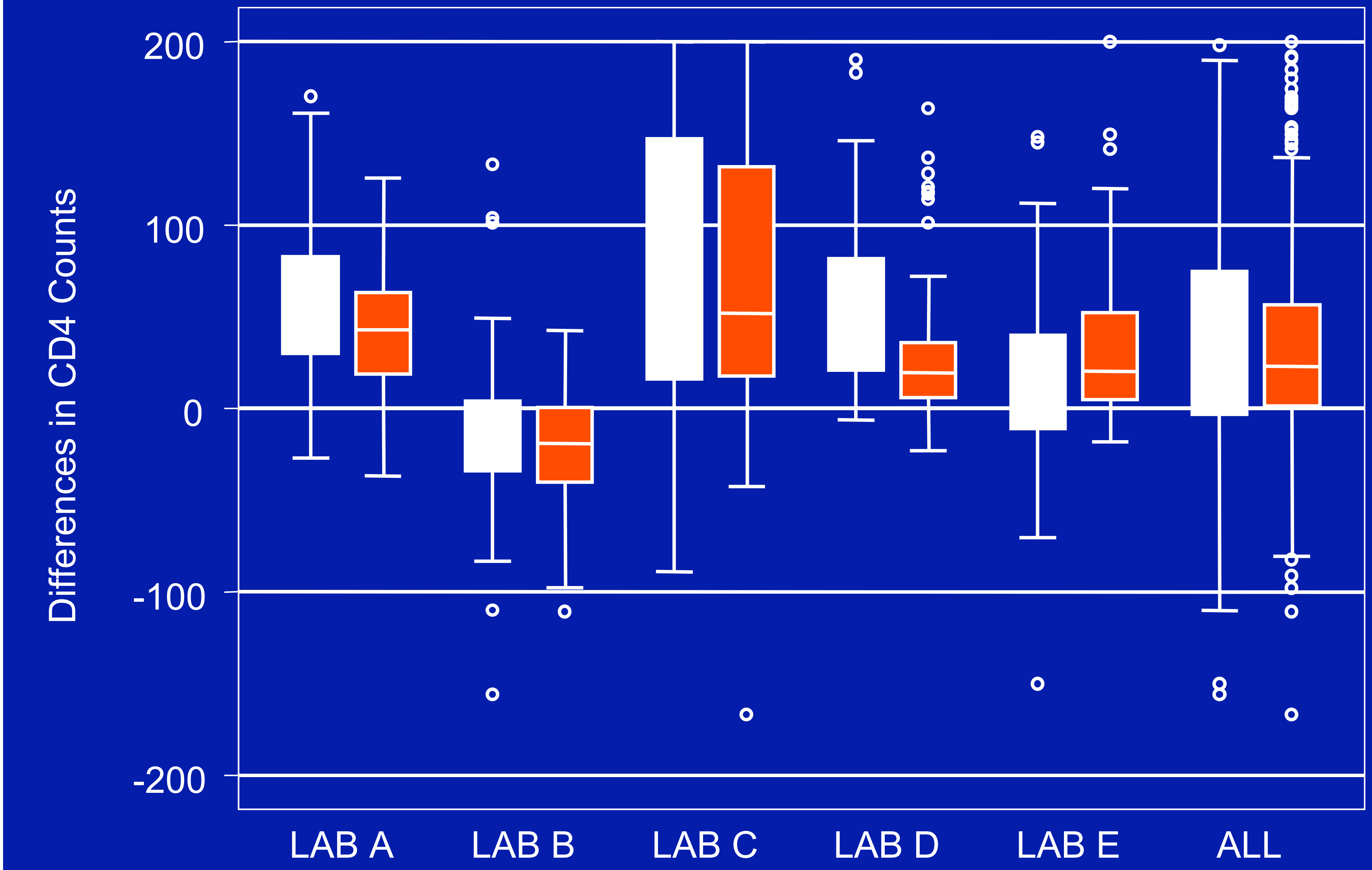
For 4 studies in a row, the average difference in CD4 count was nearly 0 (overall and within most labs).

Then 1 study with big bias (10-20%).

Would still base sample sizes on CVs (because bigger than for CD4 counts)

but may need to worry about bias in subsets of range and within labs

Old Method Minus New Method CD4 Counts



Statements about Bias with which I Disagree

(from study where new method has CD4 counts 10-20% lower than old method)

1)We don’t need to worry about bias because it is in the good direction (means more patients will be treated because they have CD4 < 200)

Bad economics, bad way to change guideline

2)A bias of 10-20% isn’t so important. After all, within-lab CV on the old method is 5-10% corresponding to a difference of 10-20%

new method still has non-zero CV%

p% bias can be worse (for misclassification) than p% CV

Statements about Bias with which I Disagree

(from study where new method has CD4 counts 10-20% lower than old method)

3) We can deal with bias by adjusting for it. Variability (CV%) is much harder to deal with.

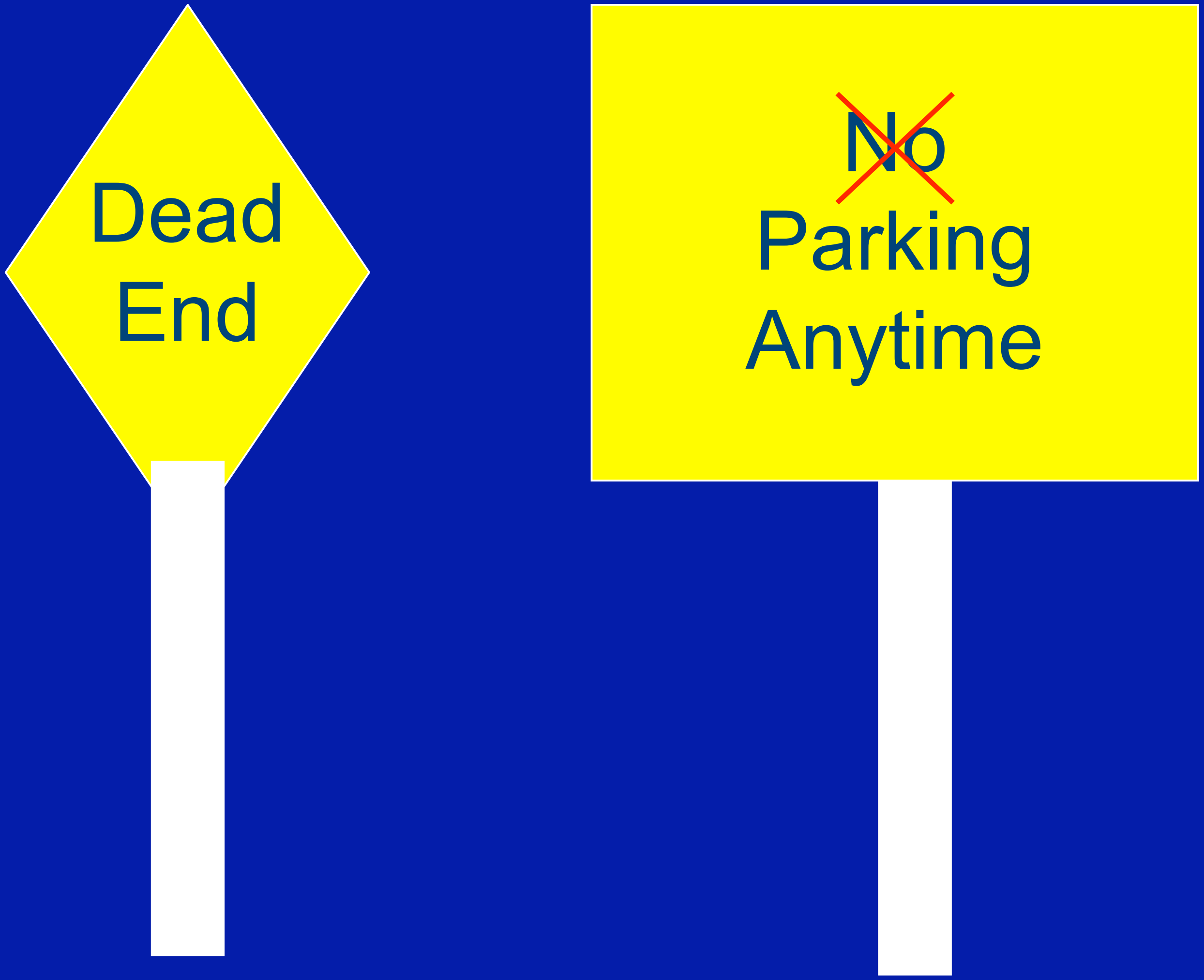
True, but:

bias estimate has variability

bias may not be even approximately constant over time or CD4 count

few labs willing to do post-processing adjustment (ok if company does it in black box)

Bad Routes and Misdirection



Beware of Bad Routes or Misdirection

Correlation Coefficient

- Most common method in medical and lab literature, but virtually worthless for this problem

ANOVA

- Suggested in email sent to conference attendees – but almost never appropriate for this problem

Beware of Bad Routes or Misdirection

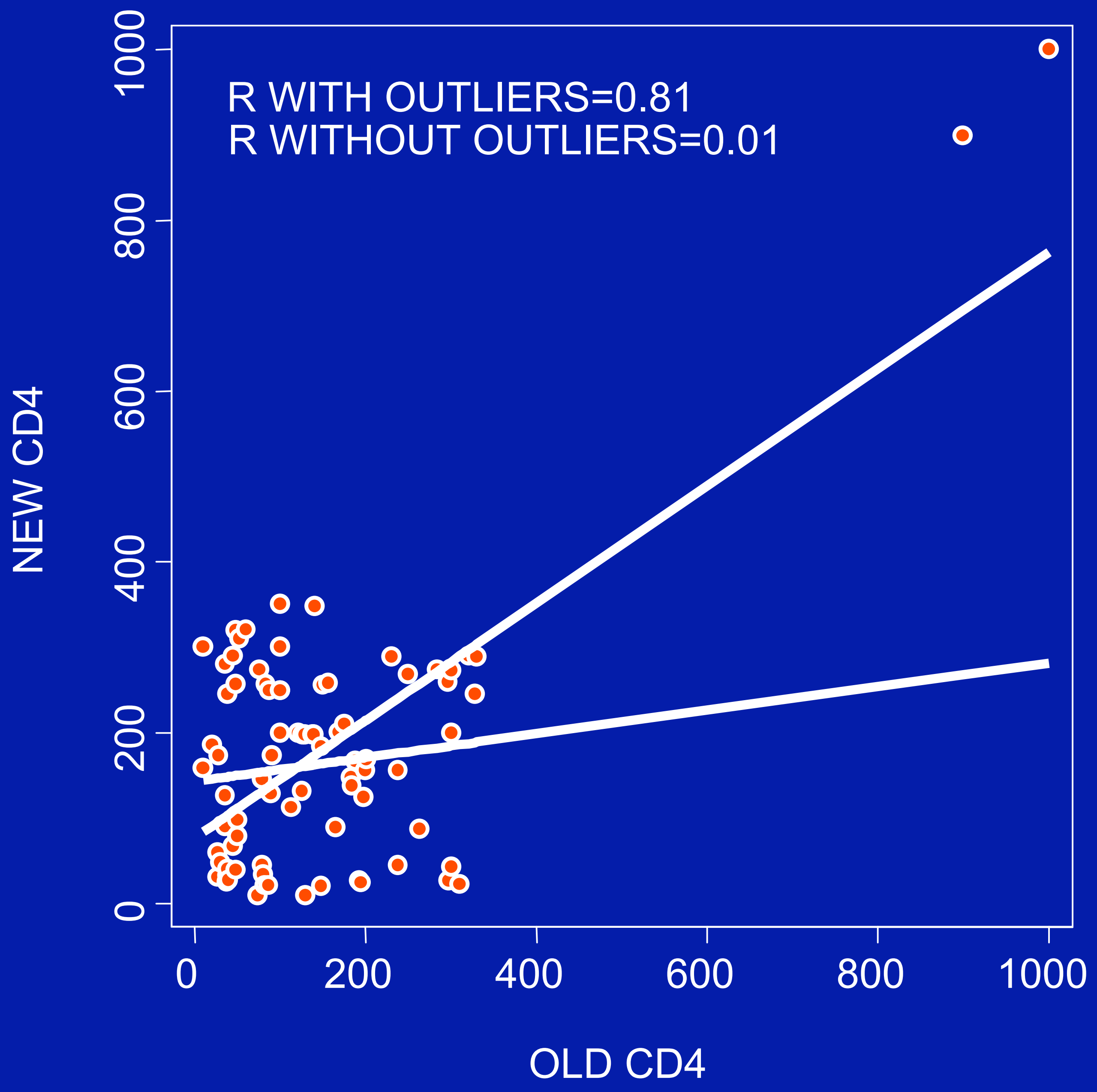
Correlation Coefficient R

Should be near 1.0, but that is not sufficient; can be .999 and two methods still far apart. Y = 500 + .5X but std. dev. on 500 is 400 and std. dev. on .5 is .3

Much affected by outliers; dropping a few outliers can change R from .999 to .001

Much affected by range of values; combining HIV+ and HIV- samples can change R from .999 to .001

For other topics we use rank correlation, but guidelines aren’t based on ranks.


R WITH OUTLIERS=0.81
R WITHOUT OUTLIERS=0.01
NEW CD4
OLD CD4
0
200
400
600
800
1000
0
200
400
600
800
1000

Truth (or Falsehood) in Advertising

ANOVA (analysis of variance) is the very worst-named method in all of statistics.

Should be ANWVA (analysis without variance) because it is based on <u>knowing</u> all variances are <u>equal</u> (so if calculated variances are different this must imply means are different).

I've occasionally seen two technologies with zero mean difference. But I've never seen two technologies with the same variance.

I don't see how ANOVA is useful for comparing two technologies.

Princess detecting the pea or throwing out the baby with bath water

Back to MapQuest

All I’ve done is vaguely describe a few routes for a few specific

starting locations

ending locations

options for optimizing

because the routes themselves are boring unless you are traveling them.

For more (and more explicit) routes, go see a statistician!!!