

Test Validation Strategies



Dr. Michaela Riddell
International Public Health Unit,
DEPM, Monash University

michaela.riddell@monash.edu



Outline:

- Considerations for introducing a new or alternative test
- Validation parameters for consideration
- Composition of validation panel
- Sample size of validation panel
- Indicators of test performance
- Statistical methods of comparison



Validation of standard methods

- *The validation of standard or collaboratively tested methods should not be taken for granted, no matter how impeccable the method's pedigree - the laboratory should satisfy itself that the degree of validation of a particular method is adequate for the required purpose and that the laboratory is itself able to match any stated performance data*

- (CITAC/EURACHEM, Working group. *International guide to quality in analytical chemistry: An aid to accreditation*, 2002)



Introducing a new or modified test

■ Practical Considerations

- Cost effective
- Robust/reliable
- Need for test

■ Analytical considerations

- Precision
- Validation of accuracy
- Limit of detection
- Equipment required



Introducing a new or modified test

■ Other Considerations

- Reason for introducing new method
 - New technology
 - Change from in-house to commercial (or vice - versa)
 - Development of previously unavailable test
 - Development of test to detect or diagnose infection of new pathogen
- Ethical (availability/source of panel samples)

Validation parameters under consideration

Within test	Within lab
Limit of detection	Day to day variation
Limit of quantitation	Operator variation
Linearity	Reproducibility
Range	Equipment
Precision	Reagents
Accuracy	Confirmation algorithm
Selectivity	
Robustness	
Measurement uncertainty	



Limit of detection

- The smallest amount or concentration of factor that can reliably be distinguished from zero.
- The lowest value measured by a method that is greater than the uncertainty associated with it



Limit of quantitation

- The lowest concentration that can be determined with an acceptable level of uncertainty
- Often defined as 3x LOD



Linearity

- Should use 6 or more calibration standards run in duplicate or even triplicate and evenly spaced over the range of interest
- Simple linear regression can determine the relationship (association) between the test value and the sample concentration
- Plot and inspect the fitted data and the residuals to confirm linearity and observe outliers



Range

- The working range of a method is the concentration range within which the results will have an acceptable level of uncertainty
 - = [LOQ - upper limit of linear calibration]
 - Validated range = [LOQ - highest concentration used for validation]



Precision

■ Repeatability

- same sample, same test, same operator, same equipment
- Useful indicator of method performance but underestimates spread of results expected under normal conditions

■ Reproducibility

- Same samples, same test, different operators, different equipment, different laboratories, different times
- Helpful but not mandatory to assess during internal validity, but important when comparing results to other laboratories.



Accuracy

- A reference material, with a known concentration of factor can be used to estimate the bias of the test result
 - Use certified reference material (CRM), reference material (RM), spiked samples or reference methods
 - Average estimate of bias can be determined by comparing results from different runs over different days
 - Bias can be estimated using a reference method with a known bias



Selectivity

- Accuracy of the measurement in the presence of interferences
- For example:
 - Chromatography/mass spectrophotometry may be highly selective

BUT

- Assays relying on colorimetric measurements are less selective because they are more likely affected by other things in the sample

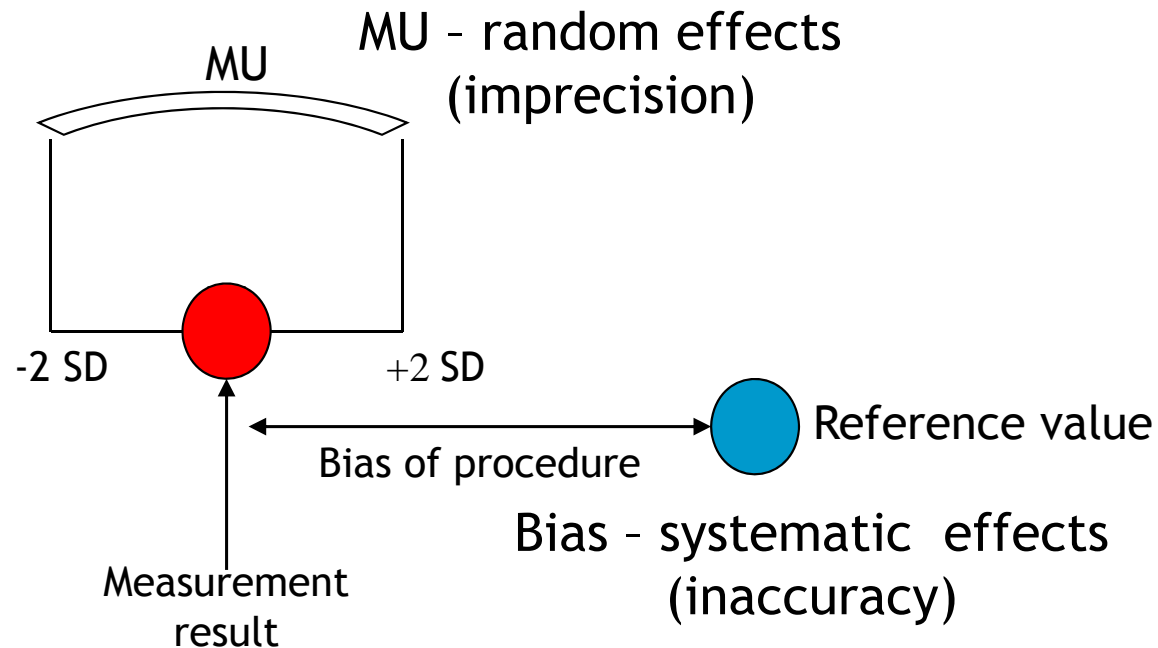


Robustness

- The degree to which results are unaffected by minor changes in the experimental method (temperature, pH, reagent concentration, incubation times)

Measurement Uncertainty

- Property of the test result - NOT the method
- Can be assessed using QC materials



REF: Requirements for the estimation of measurement uncertainty 2007 National Pathology Accreditation Advisory Council - available from

[http://www.health.gov.au/internet/main/publishing.nsf/Content/86A3CE312C612377CA257283007BC92D/\\$File/dhaeou.pdf](http://www.health.gov.au/internet/main/publishing.nsf/Content/86A3CE312C612377CA257283007BC92D/$File/dhaeou.pdf)



Validation panel composition

- Ethical considerations
- All samples must be well characterised
- Choice must be made according to purpose of testing and validation
 - Replacing an existing method
 - Introducing a new test or method
 - Representative of population
 - In accordance with manufacturers requirements (haemolysis, lipemic etc)
 - Samples should be “clean”, minimal freeze/thaw cycles
- Choose samples with adequate volume for precision and reproducibility studies
- Include appropriate proportion of “variants”



Sample size for validation

- What is “appropriate # of variants”
- Depends on prevalence of disease - changes predictive values
- Practical and economical considerations
 - Sample volume, kit & reagent cost, equipment availability
- Determine what you want to be able to say about the new or alternative assay
 - Better sensitivity/specificity, at least as good performance, quantitatively equivalent,



Sample size

- “the more the better!” - (min 30+, 200 -)
- Often “as many as possible!”
- Balance of
 - positive,
 - negative,
 - potential cross reactive samples
 - Similar clinical presentation
 - Potential cross reactive elements
 - Parasitemia
 - Rheumatoid factor
- Can reduce the sample size and achieve good validation of
 - $(1 - \beta)$ - power (usually 80%)
 - α - significance (usually 0.05)



Methods of analysis

- Sensitivity, specificity, PPV, NPV
- Kappa statistic
 - measure of concordance
- Correlation/regression
 - measures strength of association
- Non parametric methods
 - comparison of medians or means
- Method agreement analysis
 - Bland Altman method comparison analysis

Qualitative analysis

	Gold std			
		+	-	
New test				
	+	a	b	a+b
	-	c	d	c+d
		a+c	b+d	

Sensitivity = $a / (a+c)$
= proportion of positives correctly identified by the test

Specificity = $d / (b+d)$
= proportion of negatives correctly identified by the test

Qualitative analysis

	Gold std			
		+	-	
New test				
	+	a	b	a+b
	-	c	d	c+d
		a+c	b+d	

Positive Predictive value (PPV) = $a / (a+b)$

= proportion of patients with positive test who are truly positive

Negative Predictive value (NPV) = $d / (c+d)$

= proportion of patients with a negative test who are truly negative



Predictive Values

Assume: Test Sensitivity = 100% / Specificity = 99.5%

Population #1, where prevalence of infection is high (5%)

- Population: **1000 sera tested**
 - 50 sera from infected individuals
 - 950 sera from non-infected individuals
- Test Results: **50 positive**
 - 45 from the infected group
 - 5 false pos from the non-infected group
- Therefore, the positive predictive value is:
$$\text{PPV} = \frac{45}{45+5} = 90\%$$
- 9 of 10 positive results will be from infected persons

Predictive Values

Assume: Test Sensitivity = 100% / Specificity = 99.5%

Population #2, where prevalence of infection is low (0.7%)

- Population: **1000 sera tested**
 - 7 sera from infected individuals
 - 993 sera from non-infected individuals
- Test Results: **7 positives:**
 - 2 from the infected group
 - 5 false pos from the non-infected group
- Therefore, the positive predictive value is:

$$\text{PPV} = \frac{2}{2+5} = 28.6\%$$

Chance of positive result being from a truly infected person in low prevalence population is only 28.6%

Qualitative analysis

Gold std \ New test	+	-	
+	33	4	37
-	2	61	63
	35	65	100

Measure of agreement
between the two tests
 $= (33 + 61)/(100) = 0.94$

Expect some agreement just by chance need to determine how much agreement b/w tests is there above what would be expected by chance:

To determine expected agreement:

Positive: $37 \times 35/100 = 12.95$

Negative: $63 \times 65/100 = 40.95$

Total = 53.9

Number of agreements by chance = $53.9/100 = 0.539$

Therefore agreement of these two tests is Kappa (K)

$0.94 - 0.539$

$1.00 - 0.539 = 0.86$

Qualitative analysis

Gold std \ New test	+	-	
+	33	4	37
-	2	61	63
	35	65	100

Measure of agreement
between the two tests
 $= (33 + 61)/(100) = 0.94$

Therefore agreement of these two tests is Kappa (K)

$$\frac{0.94 - 0.539}{1.00 - 0.539} = 0.86$$

*Value of
Kappa*

*Strength of
agreement*

<0.2

poor

0.21 - 0.4

fair

0.41 - 0.6

moderate

0.61 - 0.8

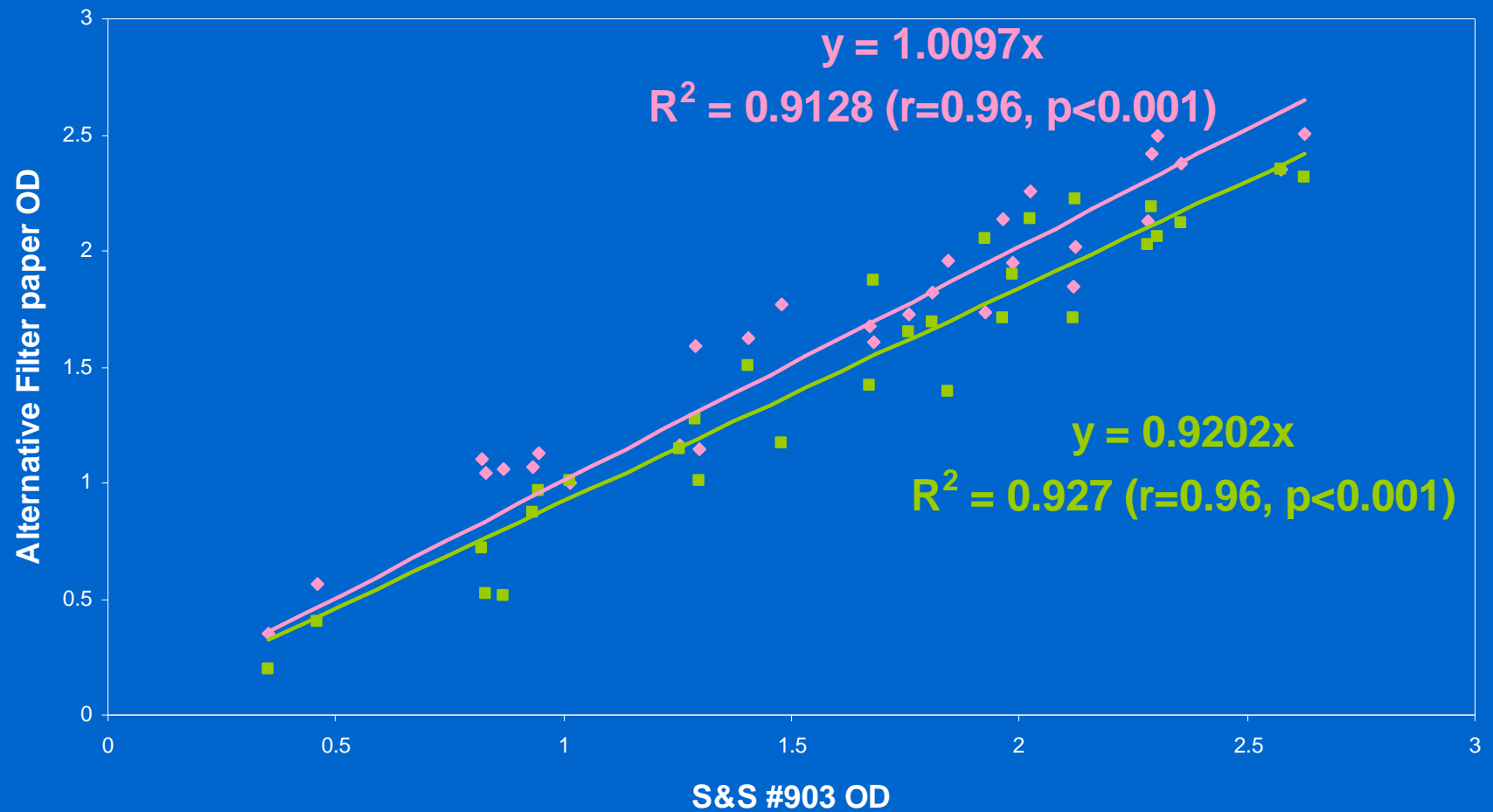
good

0.81 - 1.00

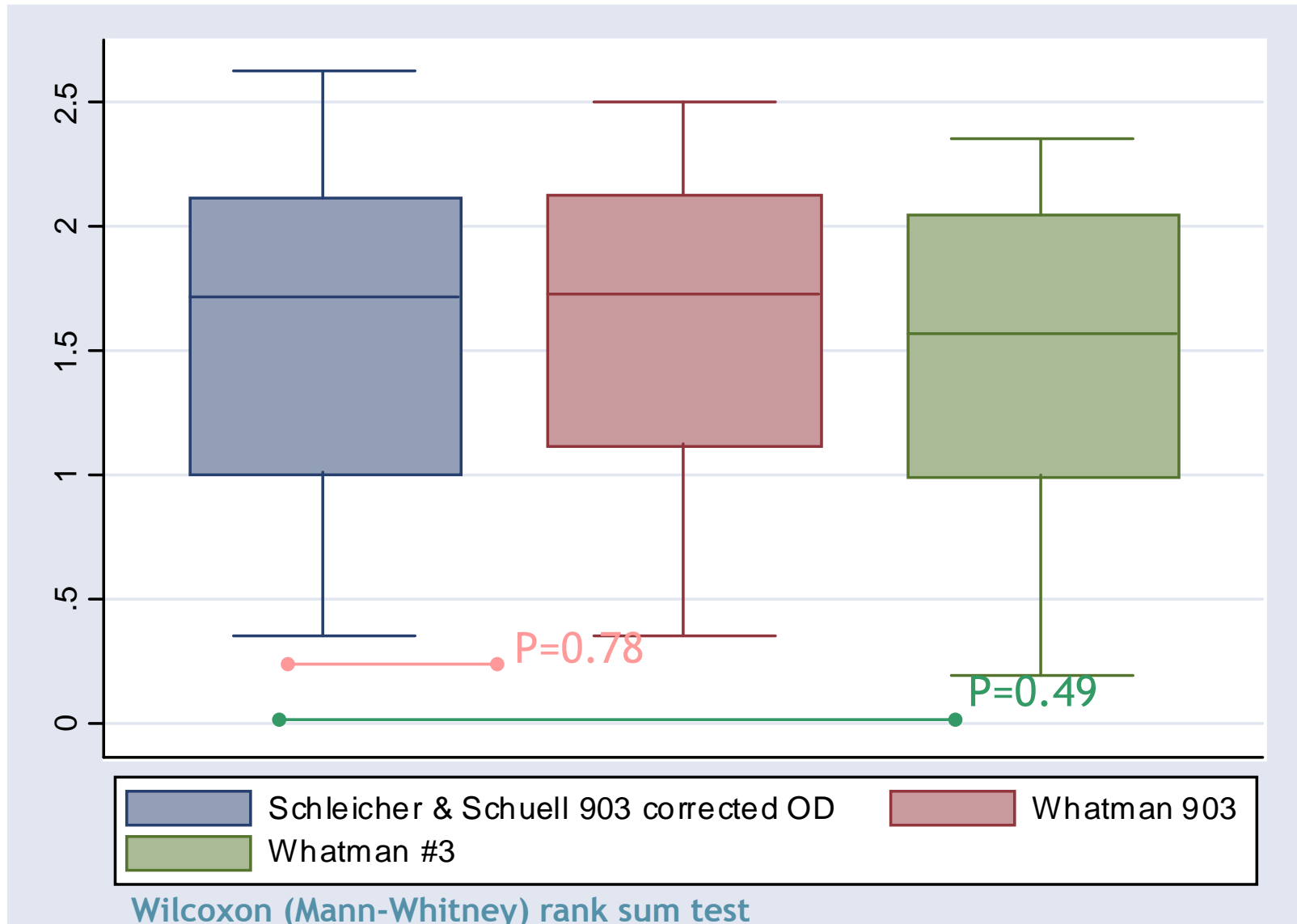
very good

Correlation & Linear regression

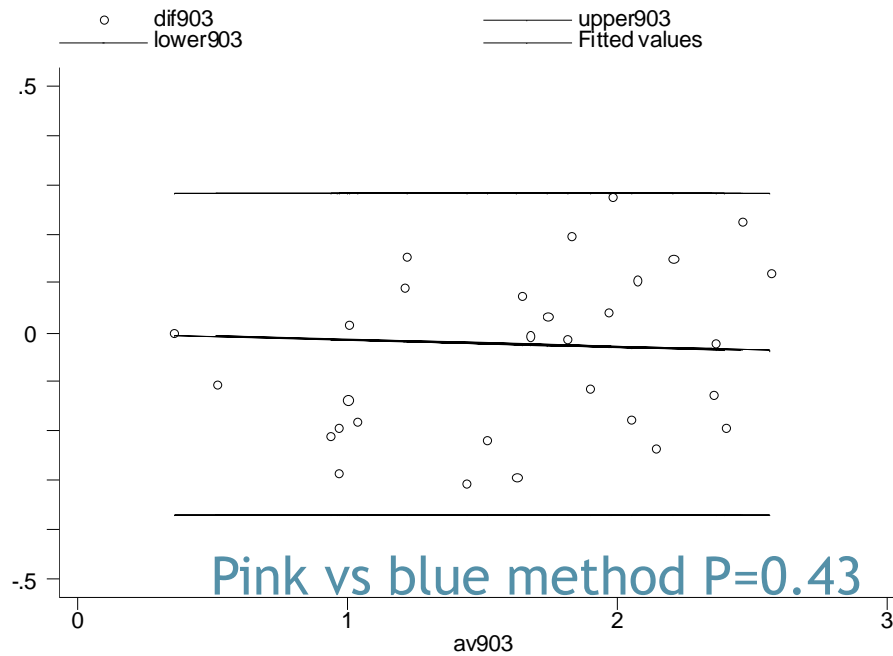
Linear regression of SS903_W903 and W3 Optical density - Measles IgG results



Box plot comparing median OD

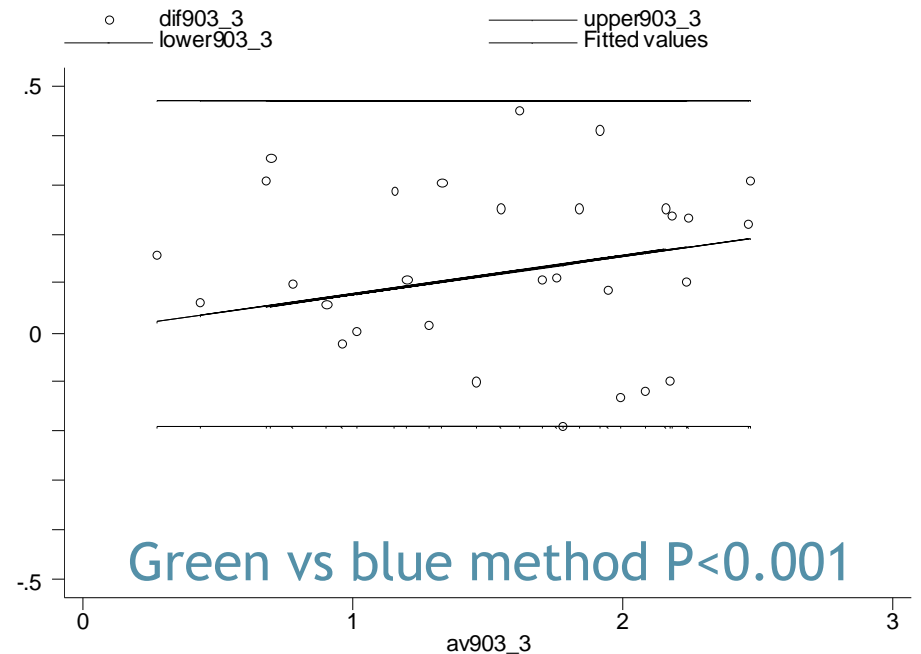


Bland Altman plot difference against average



Regress dif on average,
if methods are
equivalent then slope
of regression line = 0,

Assess the 95% limits
of agreement to
determine if the range
is clinically acceptable



Conclusion

Within test
Limit of detection
Limit of quantitation
Linearity
Range
Precision
Accuracy
Selectivity
Robustness
Measurement uncertainty

Each of these parameters need to be assessed and addressed during validation processes

Comprehensive documentation of all validation processes is required



Conclusion

■ Sensitivity/ Specificity

- Assesses performance only in relation to the comparator test (gold std/alternative test)

■ Predictive values

- Helpful to assess clinical performance of test,
- Should be assessed in relation to prevalence of disease

■ Kappa statistic

- Appropriate for determining concordance but may not assess agreement

■ Correlation, regression analysis and comparison of means or medians

- inappropriate and misleading methods for assessing quantitative agreement



Conclusion

- Bland Altman graphical approach to QUANTITATIVE method comparison
 - Simple to perform and interpret
 - Doesn't require sophisticated software (can use MS excel)
 - Is a rigorous interrogation of the agreement between two methods
 - Reveals systematic bias/ measurement error
 - Facilitates a true assessment of the ability of new to replace old by quantifying agreement
 - Assessment of clinical significance of limits of agreement

Free-ware to assist with statistics

<http://statpages.org/javasta2.html#Biostatistics>



DAG Stat



DAG_Stat provides a comprehensive range of statistics calculable from 2 by 2 tables that are useful in evaluating diagnostic tests and interrater agreement. Statistics for the evaluation of diagnostic tests include **sensitivity**, sensitivity of a random test given the observed prevalence and test level, sensitivity quality index, **specificity**, specificity of a random test, specificity quality index, efficiency (the correct classification rate), efficiency of a random test, quality index, Youden's index, the **predictive value of positive test**, **predictive value of a positive random test**, predictive value of negative test, predictive value of a random test, likelihood ratio of a positive and negative tests, the odds ratio. Also included are the false positive and false negative rates, prevalence observed in the sample and test level (proportion of subjects classified as 'positive.' For investigating interrater agreement DAG_Stat calculates **Cohen's Kappa**, observed agreement, chance agreement, agreement about positive and negative cases, Byrt's bias index, prevalence asymmetry index, bias adjusted Kappa, prevalence & bias adjusted Kappa. DAG_Stat also calculates Dice's index, Yule's Q (Gamma), Phi, Scott's agreement index, the tetrachoric correlation coefficient, Goodman & Kruskal's tau, Lambda, the Uncertainty Coefficient, Pearson's **Chi Square** (with and without Yates' correction), the likelihood ratio Chi Square, **McNemar's Test** (with and without Yates' correction).



[How to use DAG_Stat](#)



[Learn more about DAG_Stat](#)



[Download DAG_Stat](#)



[Register as a DAG_Stat User](#)



[About Dags](#)



[Download DAG_Stat](#)



[Download DAG_Stat in Microsoft Excel 5/95 and 7/97 compatible format. \(Also suitable for Excel 2000.\)](#)



[Download DAG_Stat optimized for Microsoft Excel 98 for Macintosh format.](#)



[Download DAG_Stat in Microsoft Excel 4 format. \(Available soon: \[contact me\]\(#\) for more details.\)](#)

Notes on Downloading and Opening DAG_Stat Because browsers have varied and unpredictable responses when downloading Excel spreadsheets, DAG_Stat has been encoded as both Zip and Stuffit formats to suit PC and Mac users. Depending on your configuration, the download may be automatically unzipped and saved as an Excel file, or you may have to UnZip or UnStuff the file manually.

http://www.mhri.edu.au/biostats/DAG_Stat/ -- calculates an enormous number of quantities from a 2-by-2 table

http://www.macorr.com/ss_calculator.htm -- free sample size calculator

<http://www.r-project.org/> - free software environment for statistical computing and graphics.