Statistical Methods for Clinical Evaluation of Biochip Products

Jen-pei Liu, PhD
Division of Biometry, Department of Agronomy
National Taiwan University
and
Division of Biostatistics and Bioinformatics
National Health Research Institutes

At

Workshop on Applications of Statistical Methods to Evaluation of Diagnostic Biochip Products

November 26, 2005
Taipei, Taiwan

Outline

- Introduction
- Methods when clinical truth (gold standard) is present
- Methods when clinical truth is absence
- Issues on experimental design
- Discussion and summary
- References

Introduction

- Post HGP (Human Genome Project) Era
- Pharmacogentics
- Pharmacogenomics
- Biochip Products
- Target Clinical Trials
- Personalized Medicine
- Diagnosis and Treatment

Introduction

- The US FDA guidance
  - *Multiplex Tests for Heritable DNA Markers, Mutations and Expression Pattern*
  - *Pharmacogenomic Data Submission*

Collection of samples from clinical trials in Taiwan by the global pharmaceutical company for microarray and pharmacogenomic analysis

多標的陣列平台基因診斷試劑－查驗登記審查指引(2005年3月)衛生署

Introduction

- Article 8 of the Guidance（第八條用臨床檢體進行比較研究）
- Comparison to a Reference Method – Sensitivity and Specificity for Clinical Diagnosis
- Comparison to Another Device – Percent Agreement
- Report of Discrepancy, False Positive and False Negative Results
- Evaluation of Random and Systemic Error

Introduction

- Accuracy in diagnostic performance: a measure of how faithfully the information obtained using a diagnostic device reflects truth as measured by a truth standard or gold standard.
- Comparator: An established test (device) against which a proposed test is compared to evaluate the effectiveness of the proposed test.

Methods When Presence of Clinical Truth (Gold Standard)

	("Gold Standard") Clinical Truth of Diagnosis		
Diagnosis Made from New Marker Test	**Present (+)**	**Absent (-)**	**Total**
Positive (+)	a	b	m_1
Negative (-)	c	d	m_2
Total	n_1	n_2	N

Methods When Presence of Clinical Truth (Gold Standard)

- Example 1 (FDA, 2003)

New Maker Test Result	True Positive	Diagnosis Negative	Total
Positive	44	1	45
Negative	7	168	175
Total	52	169	220

Indexes of Diagnostic Accuracy

- **Sensitivity** (True Positive rate): Capacity for making a correct diagnosis in subjects with the disease
- Estimated Sensitivity:

  100% x a/(a+c)
- **Specificity** (True Negative rate): Capacity for making a correct diagnosis in subjects without disease
- Estimated Specificity:

  100% x d/(b+d)

Indexes of Diagnostic Accuracy

Data from Example 1

- Estimated sensitivity

  = 100% x 44/51 = 86.3%

  Exact 95% confidence interval based on binomial distribution: (73.7%, 94.3%)

- Estimated specificity

  = 100% x 168/169 = 99.4%

  Exact 95% confidence interval based on binomial distribution: (96.8%, 100%)

Indexes of Diagnostic Accuracy

- **Positive Predictive Value** (Positive Predictive Accuracy): the proportion of subjects with the disease given the positive results.
  = 100% x a/(a+b)
- **Negative Predictive Value** (Negative Predictive Accuracy): the proportion of subjects without the disease given the negative results.
  = 100% x d/(c+d)
- **False positive rate**: given the positive results ,the proportion of subjects without the disease
  =1 – positive predictive value = 100% x b/(a+b)
- **False negative rate**: given the negative results, the proportion of subjects with the disease
  = 1 – negative predictive value = 100% x c/(c+d)

Methods When Presence of Clinical Truth (Gold Standard)

Example 2 (Feinstein, 2002)

New Maker Test Result	Diseased Cases	Nondiseased Control	Total
Positive	46	2	48
Negative	4	48	52
Total	50	50	100

Indexes of Diagnostic Accuracy

Data from Example 2 (Feinstein, 2002)

- Sensitivity = 100% x 46/50 = 92.0%
- Specificity = 100% x 48/50 = 96.0%
- Prevalence = 100% x 50/100 = 50.0%
- Positive Predictive Value
  = 100% x 46/48 = 95.8%
- Negative Predictive Value
  = 100% x 48/52 = 92.3%
- False Positive Rate = 100% x 2/48 = 4.2%
- False Negative Rate = 100% x 4/52 = 7.7%

Example 3 (Feinstein, 2002)

New Maker Test Result	Diseased Cases	Nondiseased Control	Total
Positive	46	38	84
Negative	4	912	916
Total	50	950	1000

Indexes of Diagnostic Accuracy

- Example 3 (Feinstein, 2002)
- Sensitivity = 100% x 46/50 = 92.0%
- Specificity = 100% x 912/950 = 96.0%
- Prevalence = 100% x 50/1000 = 5.0%
- Positive Predictive Value = 100% x 46/84 = 54.8%
- Negative Predictive Value = 100% x 912/916 = 99.6%
- False Positive Rate = 100% x 38/84 = 45.2%
- False Negative Rate = 100% x 4/916 = 0.4%

Error rates associated with screening test (Fleiss, 1981)

Prevalence	False Positive Rate	False Negative Rate
1/million	.9999	0
1/100,000	.9991	0
1/10,000	.9906	.00001
1/1000	.913	.00005
1/500	.840	.00010
1/200	.677	.00025
1/100	.510	.00051

Indexes of Diagnostic Accuracy

- False positive rate is high if prevalence of the disease is low and vice versa. False negative rate is high if prevalence of the disease is high and vice versa.

- However, sensitivity and specificity are independent of the size of the subjects used to evaluate the tests. (i.e., independent of the prevalence rate)

Indexes of Diagnostic Accuracy

Type of Diagnostic Tests (Feinstein, 1977)

- Screening or discovery tests: mammogram, fasting blood sugar-required high sensitivity => high false positive rate.
- Exclusion tests: to rule out the presence of the disease such as colonoscopic examination => require extremely high sensitivity
- Confirmation test: to verify the suspicion of the presence of the disease such as biopsy for lung cancer => require extremely high specificity with very few false positive.

Indexes of Diagnostic Accuracy

Type of Diagnostic Markers

- Binary Test Results (+,-)
- Multiple Categorical Results
  - Abnormality Rating
  - Severity Rating
  - Urine test: None, trace, 1+, 2+
  - HER2 test: 0, 1+, 2+, 3+
- Continuous Test Results
  - PSA
  - Intraocular Pressure
  - Glucose tolerance test
  - Gene expression level

Example 4:

Results in Diagnostic Marker Study of Coronary Artery Disease and Level of S-T Depression in Exercise Stress Test

Patients with S-T Segment Depression of	Definitive State of Disease: Cases of Coronary Disease	Definitive State of Disease: Controls Without Coronary Disease
≥ 3.0mm.	31	0
A		
≥ 2.5mm. but < 3.0mm.	15	0
B		
≥ 2.0mm. but < 2.5mm.	27	7
C		
≥ 1.5mm. but < 2.0mm.	30	8
D		
≥ 1.0mm. but < 1.5mm.	32	39
E		
≥ 0.5mm. but < 1.0mm.	12	43
F		
< 0.5mm.	3	53
TOTAL	**150**	**150**

Source: Feinstein (2002)

Indexes of Diagnostic Accuracy

- To convert a ranking scale or a continuous measurement into a binary outcomes (+,–), we need a cutoff point or threshold.
- ***Example:***
- FBG > 126mg/dL DM (+)
- ≤ 126mg/dL DM (–)
- S-T Depression in Exercise Stress Test
- Class D < 1.5 min CAD (+)
- ≥ 1.5 min CAD (–)

Indexes of Diagnostic Accuracy

At a specific threshold, relationship of sensitivity, specificity, false positive and false negative rates can be interpreted through hypothesis testing:

H0:Absence of the disease

H1:Presence of the disease

α =Pr[Type I Error]

=Pr[test positive | no disease]

β=Pr[Type II Error]

=Pr[test negative | disease]

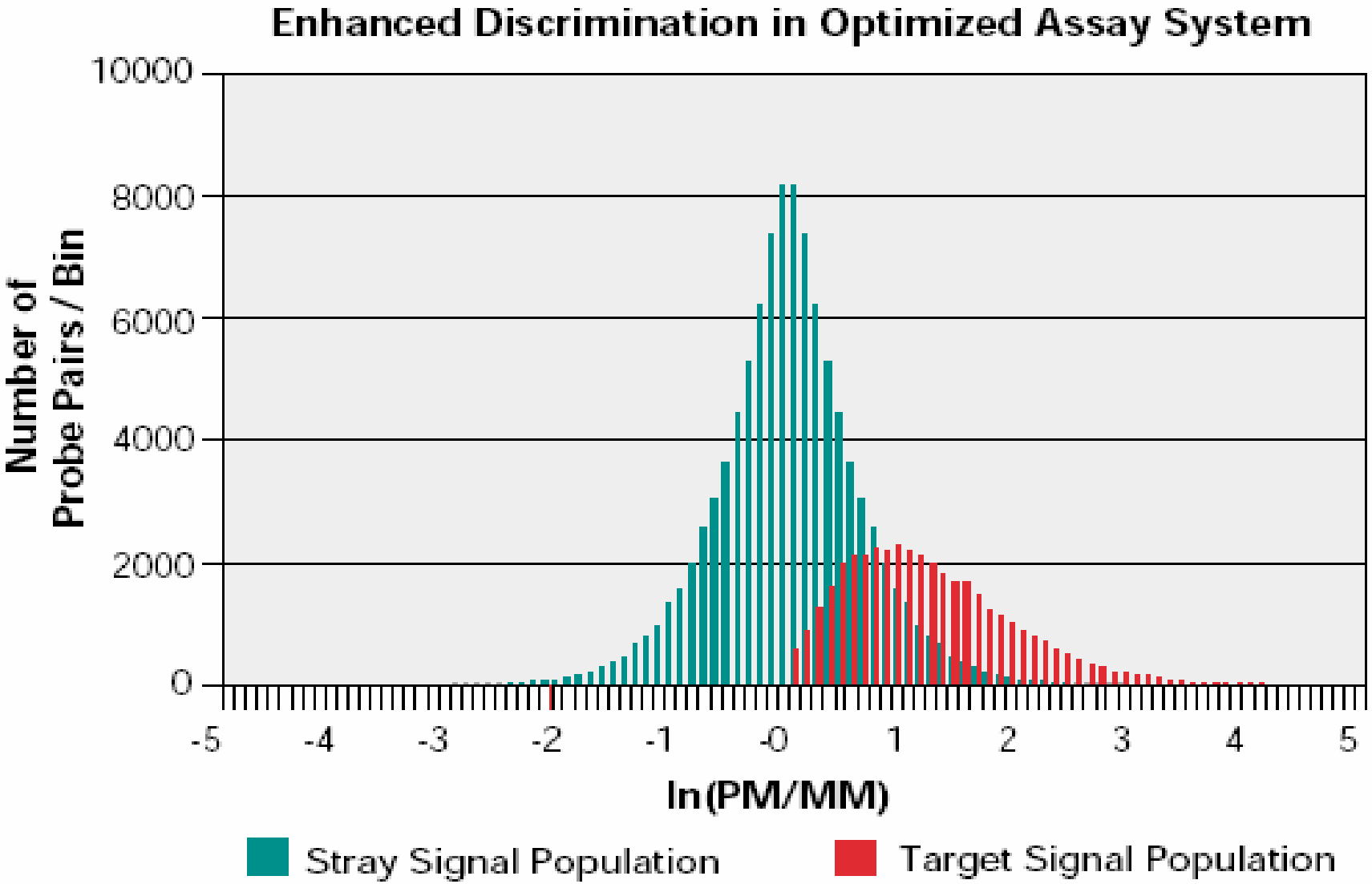
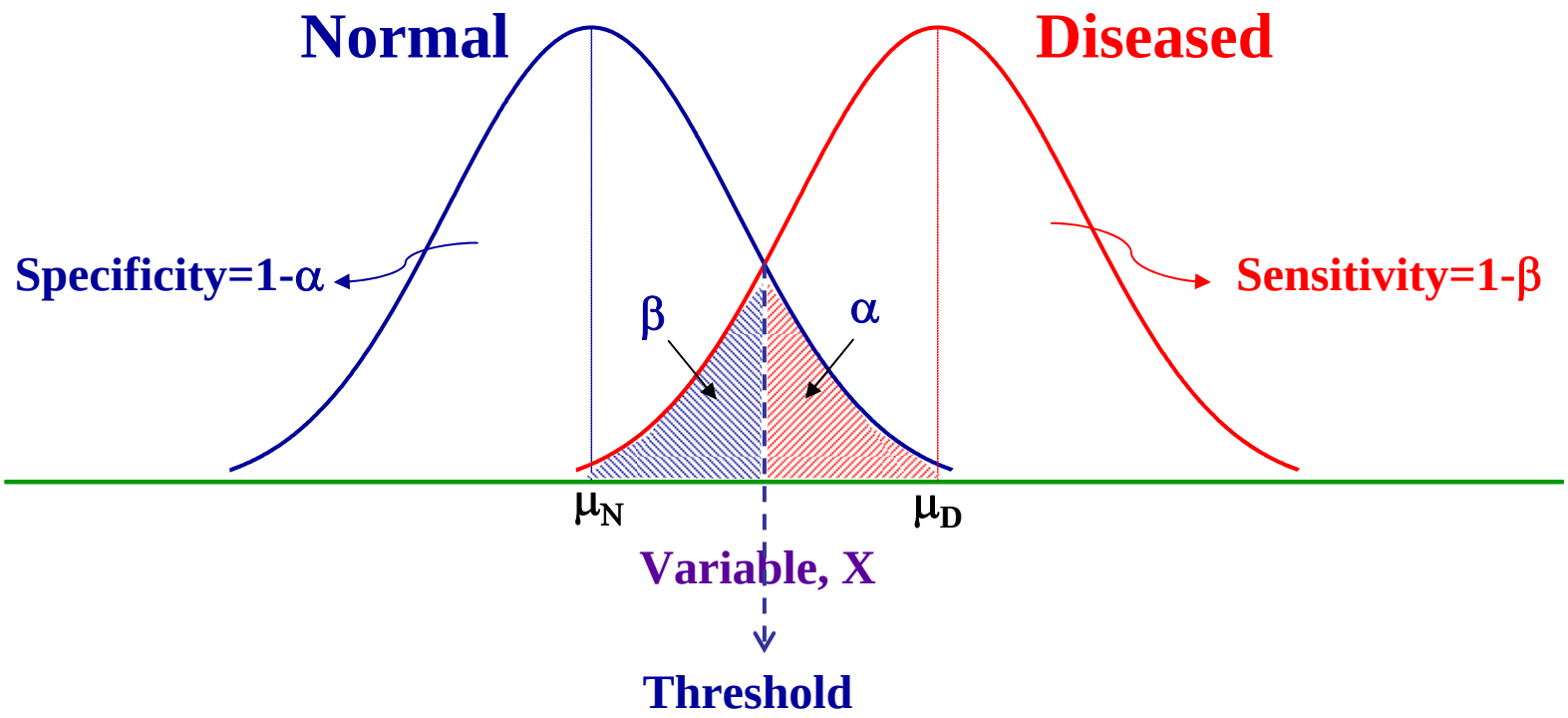


Figure 2. Global evaluation of hybridization results on a GeneChip® array.

From Affymetrix Technical Note "GeneChip® arrays provide optimal sensitivity and specificity for microarray expression analysis"

Indexes of Diagnostic Accuracy



Indexes of Diagnostic Accuracy

Sensitivity = Pr[test positive | disease]
= $1 - \beta$
= power of the statistical procedure

Specificity = Pr[test negative | no disease]
= $1 - \alpha$

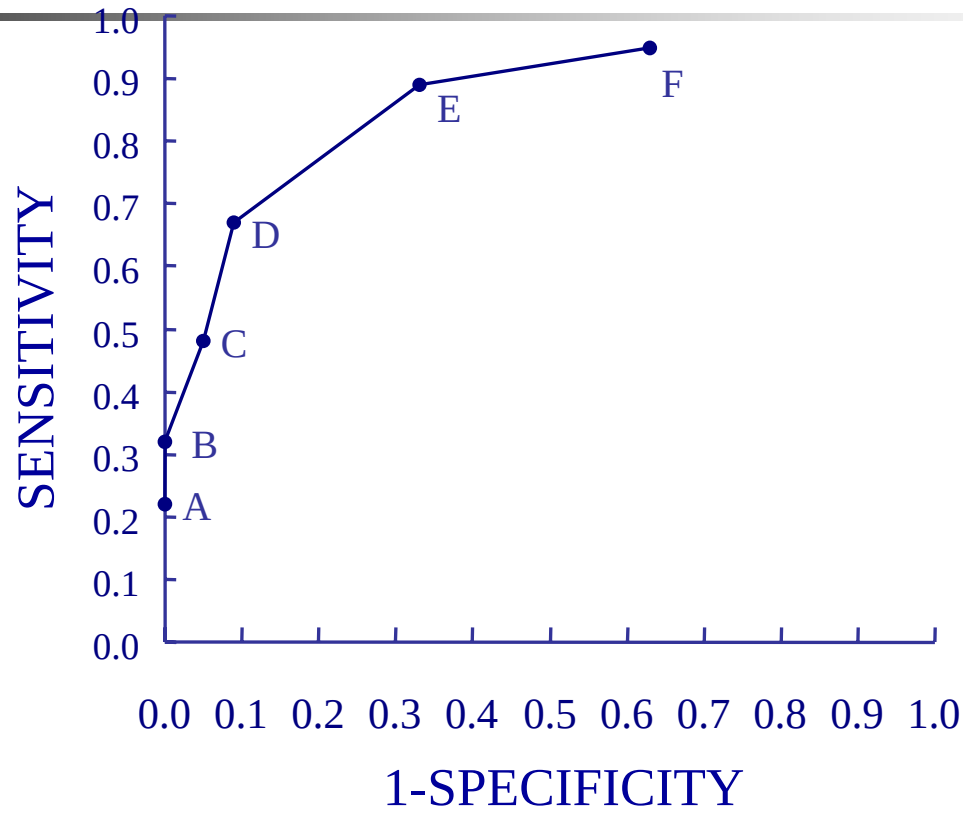
- $\alpha\uparrow \Rightarrow \beta\downarrow \Rightarrow (1-\beta)\uparrow$
- A test with a high sensitivity also has a high incorrect positive rate but a low incorrect negative rate. A test with a high specificity also has a high incorrect negative rate but a low incorrect positive rate.

Indexes of Diagnostic Accuracy

- At each individual threshold (cut-off), sensitivity and specificity can be computed.
- A Receiving Operating Characteristic (ROC) curve is a graphic presentation of sensitivity against 1-specificity.
- It is a path in the unit square, from the lower left corner to the upper right corner. In fact, it can be viewed as a cumulative distribution function.
- Swets (1979), Hanley and McNeil (1982), Metz (1978, 1980)

Summary of Nosologic Sensitivity and Specificity Calculated for Demarcations of Example 4

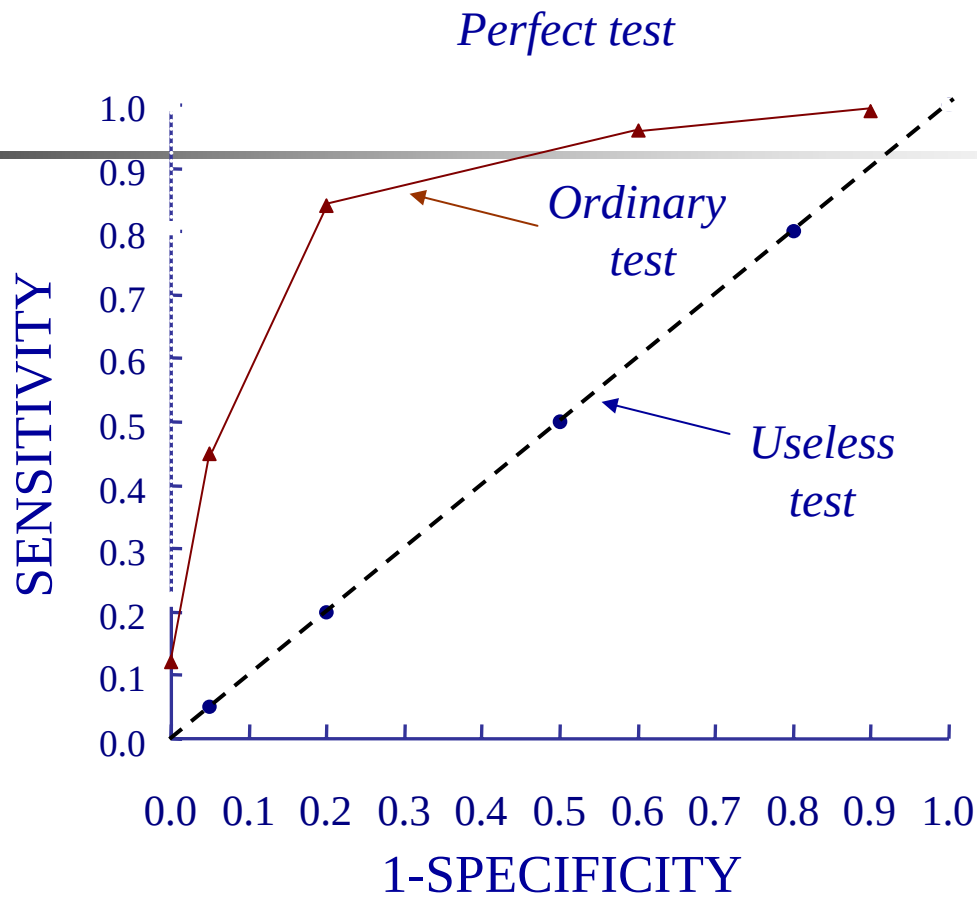
Demarcation	Location of Boundary for Abnormal	Number of Cases Included	Sensitivity	Number of Controls Included	Specificity	1 –Specificity
A	≥ 3.0mm.	31	0.21	0	1	0
B	≥ 2.5mm.	46	0.31	0	1	0
C	≥ 2.0mm.	73	0.49	7	0.95	0.05
D	≥ 1.5mm.	103	0.69	15	0.90	0.10
E	≥ 1.0mm.	135	0.90	54	0.64	0.36
F	≥ 0.5mm.	147	0.98	97	0.35	0.65
	TOTAL	150	—	150	—	—



Source: Feinstein (2002)

Indexes of Diagnostic Accuracy

- In a useless marker test, the ROC curve will be a straight line at a 45^{o} angle.
- The area under the ROC curve provides a summary index for diagnostic accuracy across over all possible values of thresholds.
- The range of the area under the ROC curve is from 0.5 (50%) to 1.0(100%)
- In a useless marker test, the area under the ROC curve is 50% which is the same as flopping a fair coin.
- For non-inferiority or equivalence test based on the paired ROC curve area, see Liu, et al. (2005, *Statistics in Medicine*)



Source: Feinstein (2002)

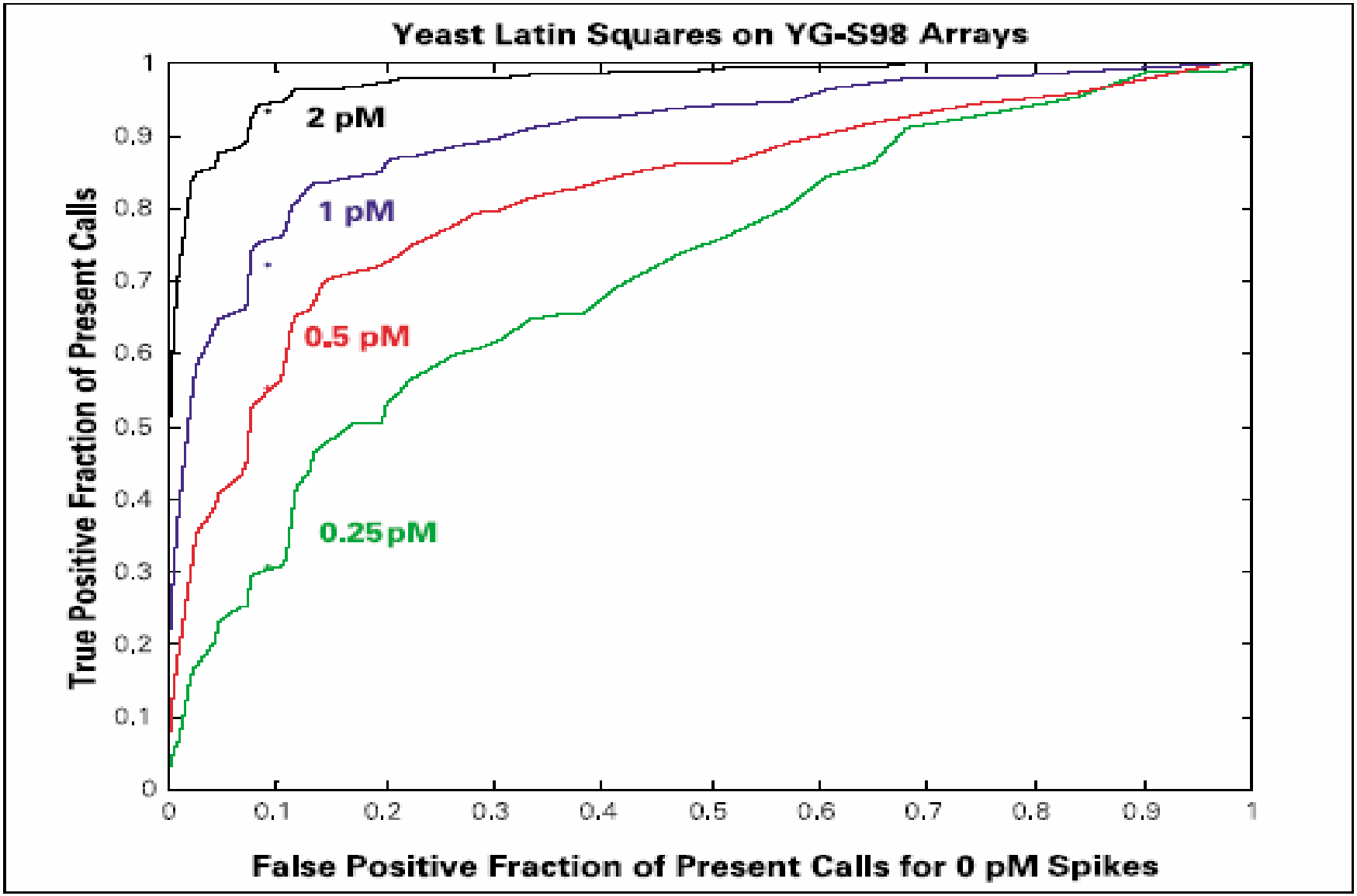


Figure 1. ***Detection Calls in Yeast Spikes: balancing sensitivity and specificity.***

From Affymetrix Technical Note 2 “Fine tuning your data analysis”

Indexes of Diagnostic Accuracy

- Other indices
  - Likelihood ratios: independent of prevalence
    - Positive test
    - Negative test
  - Odds ratio
    - Pretest
    - Posttest – positive
    - Posttest - negative

Methods When Absence of Clinical Truth

	Diagnosis Made from Another Device		
Diagnosis Made from Marker Test	**Positive (+)**	**Negative (-)**	**Total**
Positive (+)	a	b	a+b
Negative (-)	c	d	c+d
Total	a+c	b+d	N

Methods When Absence of Clinical Truth

- Overall percent agreement
  = 100% x (a+d)/N
- Agreement of new test with another device – positive = 100% x a/(a+c)
- Agreement of new test with another device – negative = 100% x b/(b+d)
- Kappa statistics (Fleiss, 1981)

Methods When Absence of Clinical Truth

	Diagnosis Made from Another Device		
Diagnosis Made from New Marker Test	Positive (+)	Negative (-)	Total
Positive (+)	40	5	45
Negative (-)	4	171	175
Total	44	176	220

Methods When Absence of Clinical Truth

- Overall percent agreement

  = 100% x (40+171)/220 = 95.9%

  95% CI = (92.4%, 98.1%)
- Agreement of new test with another device – positive = 100% x 40/44 = 90.9%
- Agreement of new test with another device – negative = 100% x 171/220 = 97.2%

Methods When Absence of Clinical Truth

Disadvantages of agreement measurements:

- "Agreement" does not mean "correct"
- Agreement changes depending on disease prevalence
- For evaluation of non-inferiority or equivalence of two diagnostic tests based on the proportions of the positive results, see Liu, et al. (2002, *Statistics in Medicine*)

HercepTest

- HER2 (the human epidermal growth factor receptor 2) is a member of the *HER* (erbB) family of transmembrane tyrosine kinase
- Enhanced level of HER2 is associated with mammary epithelial cell transformation and shorter survival in patients with breast cancer
- ≈ 25% of invasive breast cancers exhibit *HER2* gene amplification
- The rate of *HER2* gene amplification or protein in ductal carcinoma in situ (DCIS) is higher than invasive cancer ⇒ pathogenic role in the initiation of mammary carcinoma
- Treatment of Herceptin - requirement of screening the patients with over-expressed HER2 level

HercepTest

- HercepTest™ is an immunohistochemical (IHC) test intended to aid in the assessment of patients being considered for Herceptin treatment
- A Class III device – require clinical studies
- Interpret results
  - Negative for HER2 over-expression: 0 or 1+
  - Positive for HER2 over-expression: 2+ or 3+

PATHWAY™ Her 2 (Clone CB11)

- A mouse monoclonal antibody
- Semi-quantitative detection of c-erbB-2 antigen
- Binding of an antibody to an antigen of interest
- Visualization of the bound primary antibody by an indirect biotin-avidin system coupled to an enzyme
- Interpret results
  - Negative for HER2 over-expression: 0 or 1+
  - Positive for HER2 over-expression: 2+ or 3+

PATHWAY™ Her 2 (Clone CB11)

- Potential Adverse Effects
  - False Positive
    - the benefit of Herceptin to patients with normal or lower level of HER2 is unknown
    - The risks of Herceptin include infusion toxicity (chills, fever, pain, asthenia, nausea, vomiting and headache) and cardiotoxicity

Clinical Studies

- Compared to DAKO HercepTest™
- Goal: at least 75% agreement with 95% confidence

  Ho $P \geq 0.75$ vs. Ha: $P>0.75$
- One central laboratory
- Multi-center: 3 sites
- 50+ and 100- specimens by HercepTest for each site
- + > 10% of cells staining scores of 2+ or 3+

Clinical Studies

Table 4: Concordance of Ventana c-erbB-2 Primary Antibody and HercepTest™

	HercepTest™ Negative	HercepTest™ Positive	Total
c-erbB-2 Primary Antibody Negative	282	17	299
c-erbB-2 Primary Antibody Positive	17	134	151
Total:	299	151	450

Concordance = 92.4% 95% Confidence Interval = 89.6% - 94.7% $p < 0.0001$

Clinical Studies

- Observed overall agreement = 92.4 (416/450) with an exact 95% CI (89.6% to 94.7%)
- P-value for testing Ho: $P \leq 0.75$ is < 0.0001
- The observe kappa statistic = 0.83 with a p-value for testing no agreement < 0.0001
- P-value for McNemar test for equal proportion of clinically + is 1.00
- Assume that HercepTest is gold standard
  - Sensitivity: 88.7% (134/151); 95%CI: 82.6% - 93.3%
  - Specificity: 94.3% (282/299); 95%CI: 91.1% - 96.7%

Clinical Studies

Table 5: 3 X 3 Concordance of Ventana c-erbB-2 Primary Antibody and HercepTest™

	HercepTest™			
c-erbB-2 Primary Antibody ↓	Negatives	2+	3+	Total:
Negatives	282	14	3	299
2+	13	24	13	50
3+	4	14	83	101
Total:	299	52	99	450

Concordance = 86.4% 95% Confidence Interval = 82.9% - 89.5% $p < 0.0001$

Clinical Studies

- Observed overall agreement = 86.4 (389/450) with an exact 95% CI (82.9% to 89.5%)
- P-value for testing Ho: $P \leq 0.75$ is < 0.0001
- The observe kappa statistic = 0.73 with a p-value for testing no agreement < 0.0001
- P-value for test for marginal homogeneity is 1.00
- Assume that HercepTest is gold standard
  - Sensitivity for intermediate category: 46.2% (24/52); 95%CI: 32.3% - 60.5%
  - Sensitivity (+): 83.8% (83/99); 95%CI: 75.1% - 90.5%
  - Specificity (-) : 94.3% (282/299); 95%CI: 91.1% - 96.7%

Kappa Statistic

- $\kappa = (p_0 - p_e)/(1 - p_e)$
  - Where $p_0 = \Sigma p_{ii}$ and $p_e = \Sigma p_{i.}\, p_{.i}$
  - p_{ij} is the proportion of (i,j) entry of a rxr contingency table
  - $P_{i.}$ ($p_{.j}$) is the sum of p_{ij} over category i, i=1,...,r
  - SE = {1/[1- p_e]sqrt(n)}{sqrt[C]}
  - $C = p_e - p^2_e - \Sigma p_{i.}\, p_{.i}(p_{i.}+p_{.i})$

Kappa Statistic

Example

	HercepTest		
PATHWAY	+	+	Margin
+	282 (0.6267)	17 (0.0378)	299 (0.6644)
-	17 (0.0378)	134 (0.2978)	151 (0.3356)
Margin	299 (0.6644)	151 (0.3356)	450 (1.0000)

$p_0 = 0.6267 + 0.2978 = 0.9245$

$p_e = 0.6644*0.6644 + 0.3356*0.3356 = 0.5541$

$\kappa = (0.9245 - 0.5541)/(1 - 0.5541) = 0.8307$

Methods When Absence of Clinical Truth

	Diagnosis Made from Another Device		
Diagnosis Made from New Marker Test	Positive (+)	Negative (-)	Total
Positive (+)	40	5	45
Negative (-)	4	171	175
Total	44	176	220

Methods When Absence of Clinical Truth

Table A

New Marker	Another Device	Total Patients	True Diagnosis +	True Diagnosis --
+	+	40	39	1
+	--	5	5	0
--	+	4	1	3
--	--	171	6	165
Total		220	51	169

New and another device agree for 211 patients
They agree and are both wrong for 6+1 = 7 patients

Methods When Absence of Clinical Truth

	Diagnosis Made from Another Device		
Diagnosis Made from New Marker Test	**Positive (+)**	**Negative (-)**	
Positive (+)	40	5	← retest
Negative (-)	4	171	

↑

Retest

Methods When Absence of Clinical Truth

Table C

New Marker	Another Device	Total Patients	True Diagnosis	
			+	--
+	+	40	N/A	N/A
+	--	5	5	0
--	+	4	1	3
--	--	171	N/A	N/A
Total		220	N/A	N/A

Methods When Absence of Clinical Truth

- New marker agrees 8 specimens with the resolver
- Another device agrees 1 specimen with the resolver
- Impossible to estimate the relative magnitude of this difference unless we know the true state for all specimens (Table A) or the disease prevalence in the target population

Methods When Absence of Clinical Truth

Common Mistakes:

- When original results agree, assume that they both correct and do not make any change to the table
- When original results disagree, and another device disagrees with the resolver, change the result of another device to the resolve result.

Methods When Absence of Clinical Truth

Table D

New Marker	Another Device	Total Patients	True Diagnosis +	True Diagnosis --	Revised Total
+	+	40	40*		45
+	--	5	↑ 5	0	0
--	+	4	1	3↓	1
--	--	171		171*	174
Total		220			220

*All specimen results incorrectly assumed to be correct

Methods When Absence of Clinical Truth

	Revised Results (Incorrect) Diagnosis Made from Another Device		
Diagnosis Made from New Marker Test	**Positive (+)**	**Negative (-)**	Total
Positive (+)	45	0	45
Negative (-)	1	174	175
Total	46	174	220

Percent Agreement
= 99.5% (219/220)

Issues on Experimental Design

- Objective: To achieve the maximal accuracy with the best precision at the minimal cost
- Use “least burdensome approaches”
- Collection of specimens (samples)
- Assays of specimens (samples)
- Pre-plan in the protocol

Issues on Experimental Design

Collection of specimens (samples)

Characteristics of diagnostic trials:

- A number of diagnostic tests can simultaneously can be applied to the same person without need of washout periods
- Two diagnostic tests are usually positively correlated

Issues on Experimental Design

Characteristics of diagnostic trials:

- Subjects serve their own control
- Between-subject variation is greater than within-subject variation
- Paired design to increase the power and efficiency
- Sometimes order of diagnostic tests can be randomly assigned
- Prefer prospective studies

Issues on Experimental Design

Assays of Specimens:

- Blindness is utmost important
- Fully blinded evaluation - blinded to patient status, clinical information, and results of other tests or diagnosis by gold standard, etc.
- Separate and unpaired evaluation: order of assays should be randomly arranged
- Consideration of day, analyst, and site

Discussion and Summary

- Indexes of diagnostic accuracy
- Presence vs. absence of clinical truth
- Quantitative method comparison in absence of a true standard
- Systemic error (y=x): Deming regression
- Random error: Bland-Altman Difference plot

Figure 1. Difference Graph for Lab 615

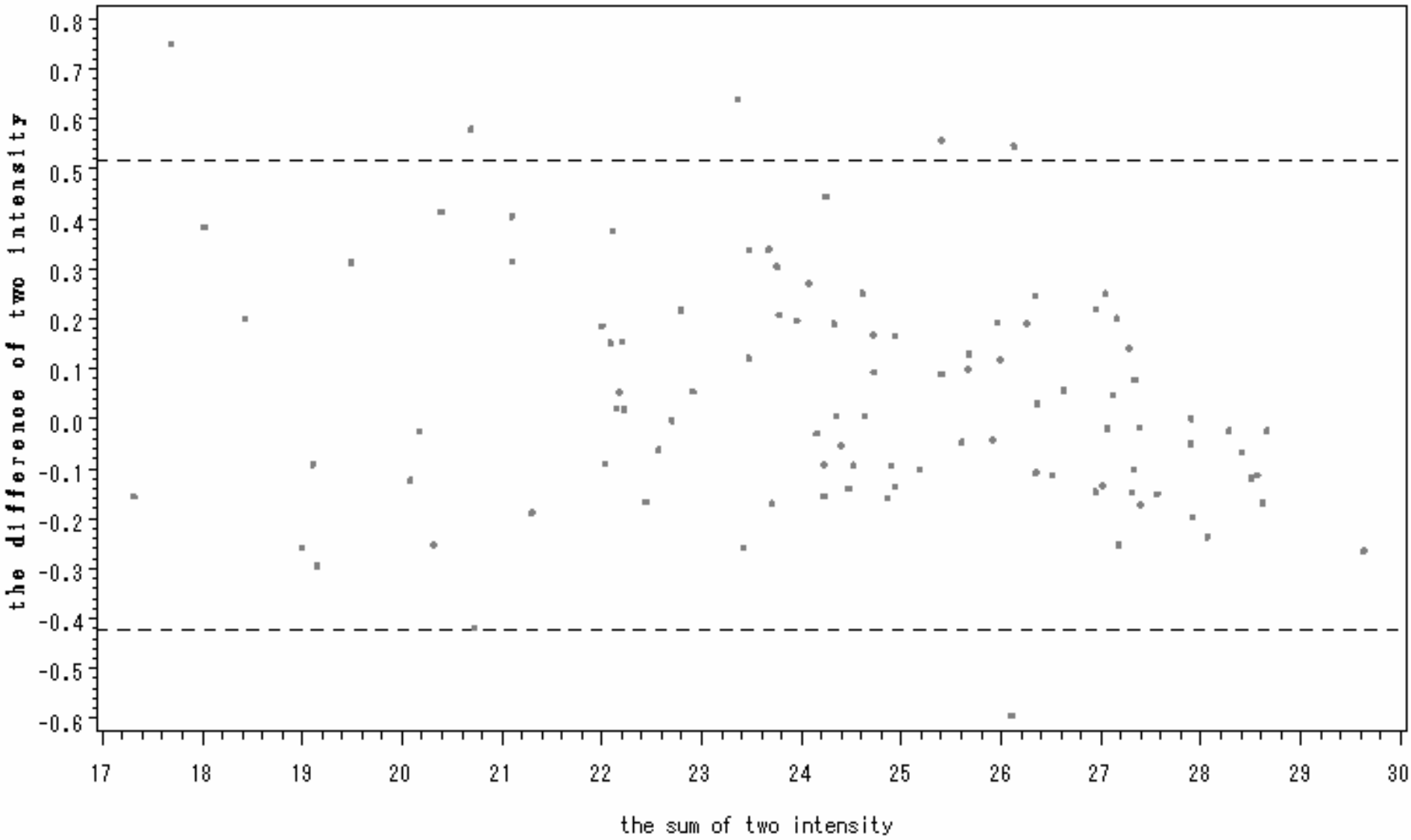
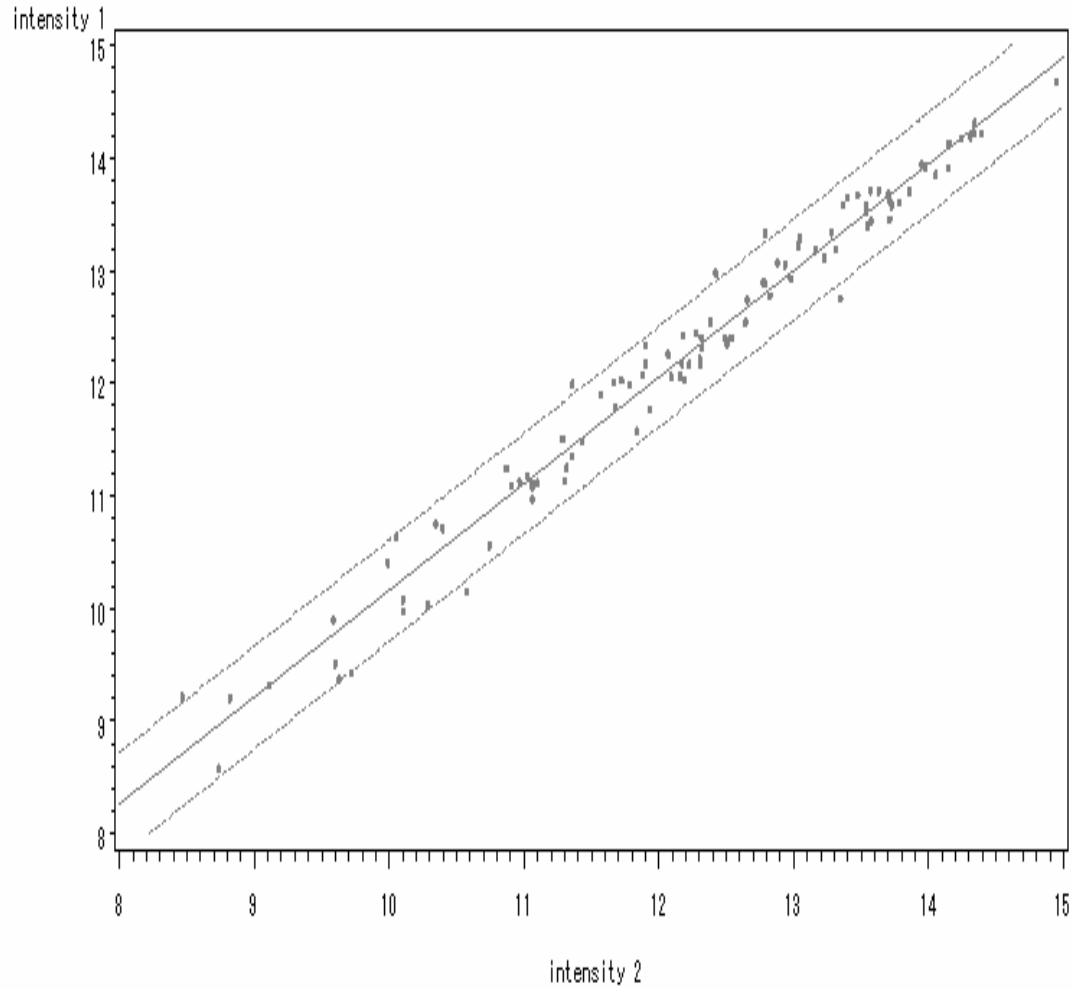
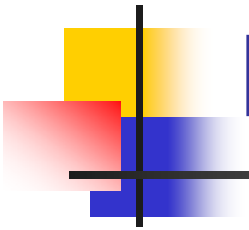


Figure 5. Scatter Plot of Replicate 1 vs. Replicate 2 for Lab 615



Regression Equation:
log2A1 = 0.654606 + 0.950092*log2A2



Discussion and Summary

- Determination of cut-off (thresholds)
- Evaluation of classification error rate: leave-one-out method, bootstrap method
- Feature selection of differentially expressed genes for each leave-one-out training set: correct error rate (Simon, et al., 2003)
- Sensitivity, specificity, positive and negative predictive value, percent agreement, and ROC curve

Discussion and Summary

Four factors in QA/QC of diagnostic tests based on microarray or other pharmacogenomic technology

- Technical: manufacturing of microarrays, sample collection, RNA extraction, cDNA/cRNA synthesis, labeling with fluorescent dye and hybridization.
- Instrumental: image acquisition, and quantification
- Computational: data preprocessing, normalization and analysis
- Interpretative: biological reasoning
- Microarray QC Metric and Thresholds (MAQC, US FDA)

References

多標的陣列平台基因診斷試劑 － 查驗登記審查指引（2005年3月）衛生署

Draft preliminary Concept paper – Drug –Diagnostic Co-Development Concept Paper (April, 2005) The US FDA

Draft Guidance on Multiplex tests for Heritable DNA Markers, Mutations, and Expression Pattern (Feb,, 2003) The US FDA

Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests (March, 2003) The US FDA

Campbell, G. (2004) Some statistical and regulatory issues in the evaluation of genetic and genomic tests, *Journal of Biopharmaceutical Statistics* 14:539-552.

Shi, L., et al. (2004) QA/QC: challenges and pitfalls facing the microarray community and regulatory agencies, Expert Rev. Mol. Diagn. 4(6):761-777.

References

Yerushalmy, J (1947) Statistical problems in assessing methods on medical diagnosis, with special reference to X-ray technique, Pub. Health Research 62:1432-1449.

Feinstein, AR(1977) Clinical Biostatistics, Mosby, St Louis.

Feinstein, AR (2002) Principles of Medical Statistics, Chapman and Hall/CPC, Boca Raton, FL.

Fleiss, JL (1981) Statistical Methods for Rates and Proportions, Wiley, New York.

Armitage, P and Berry, G (1987) Statistical Methods in Medical Research, Blackwell, Oxford.

Zhou, XH, Obuchowski, NA, McClish, DK (2002) Statistical Methods in Diagnostic Medicine, Wiley, New York.

Pepe, MS (2003) The Statistical Evaluation of Medical Tests for Classification and Prediction, Oxford University Press, New York.

.

References

Swets, JA(1979) ROC analysis applied to the evaluation of medical imaging techniques, Investigative Radiology, 14:109-121.

Hanley, JA and McNeil, BJ(1982) The meaning and use of the area under a receiver operating characteristic(ROC) curve, Diagnostic Radiology, 142:29-36.

Hanley, JA and McNeil, BJ(1982) A method of comparing the area under a receiver operating characteristic curves derived from the same cases, Radiology, 148:839-843.

Begg, C.B. (1987) Bias in the assessment of diagnostic test, Statistics in Medicine 6: 411-423

Hujoel, PP, Moulton, LH, and Loesche(1990) Estimation of sensitivity and specificity of site-specific diagnostic tests, Journal of Periodontal Research, 25:193-196.

Smith PJ, and Hadgu, A(1992) Sensitivity and specificity for correlated observations, Statistics in Medicine, 11:1503-1509.

References

Hui, SL, and Walter, SD(1980) Estimating the error rates of diagnostic tests, Biometrics, 36:167-171.

Thibodeau, LA(1981) Evaluating diagnostic tests, Biometrics, 37:801-804.

Lachenbruch PA(1988) Multiple reading procedures: the performance of diagnostic tests, Statistics in medicine, 7:549-557.

Lachenbruch PA(1992) On the sample size for studies based upon McNemar's Test, Statistics in medicine,11:1521-1523

Connor, RJ(1978) Sample size for testing differences in proportions for the paired-sample design, Biometrics, 43:629-638.

Feuer, EJ, and Kessler, LG(1989) Test statistics and sample size for a two-sample McNemar test, Biometrics, 45:629-638.

Metz CE(1979) Basic principles of ROC analysis, Seminar Nuclear medicine, 8:283-298.

References

Metz, CE, and Kronman, HB(1980) Statistical significance tests for binomial ROC curve, Journal of Mathematical psychology, 22:234-245.

Metz, CE(1989) Some practical issues of experimental design and data analysis in radiological ROC studies, Investigative Radiology; 24:234-245.

Begg, CB, and McNeil, BJ(1988) Assessment of radiological tests: control of bias and other design considerations; Radiology:167:565-569.

Hsueh, H.M., Liu, J.P., Chen, J.J. (2001) Unconditional exact tests for equivalence or non-inferiority for paired binary data", Biometrics:Vol.57(2), 478-483.

Liu, J.P., H.M. Hsueh, E. Hsieh, J.J. Chen(2002) "Tests for equivalence or non-inferiority for paired binary data", Statistics in Medicine, 21:231-245.

Liu, J.P., Ma, M.C., Wu, C.Y., Tai, J.Y. (2005) "Tests of equivalence or non-inferiority for diagnostic accuracy based on the paired areas under ROC curves, Statistics in medicine, in press.