Medical Devices Quality System Case Study: 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

December 29, 2005
Taipei, Taiwan

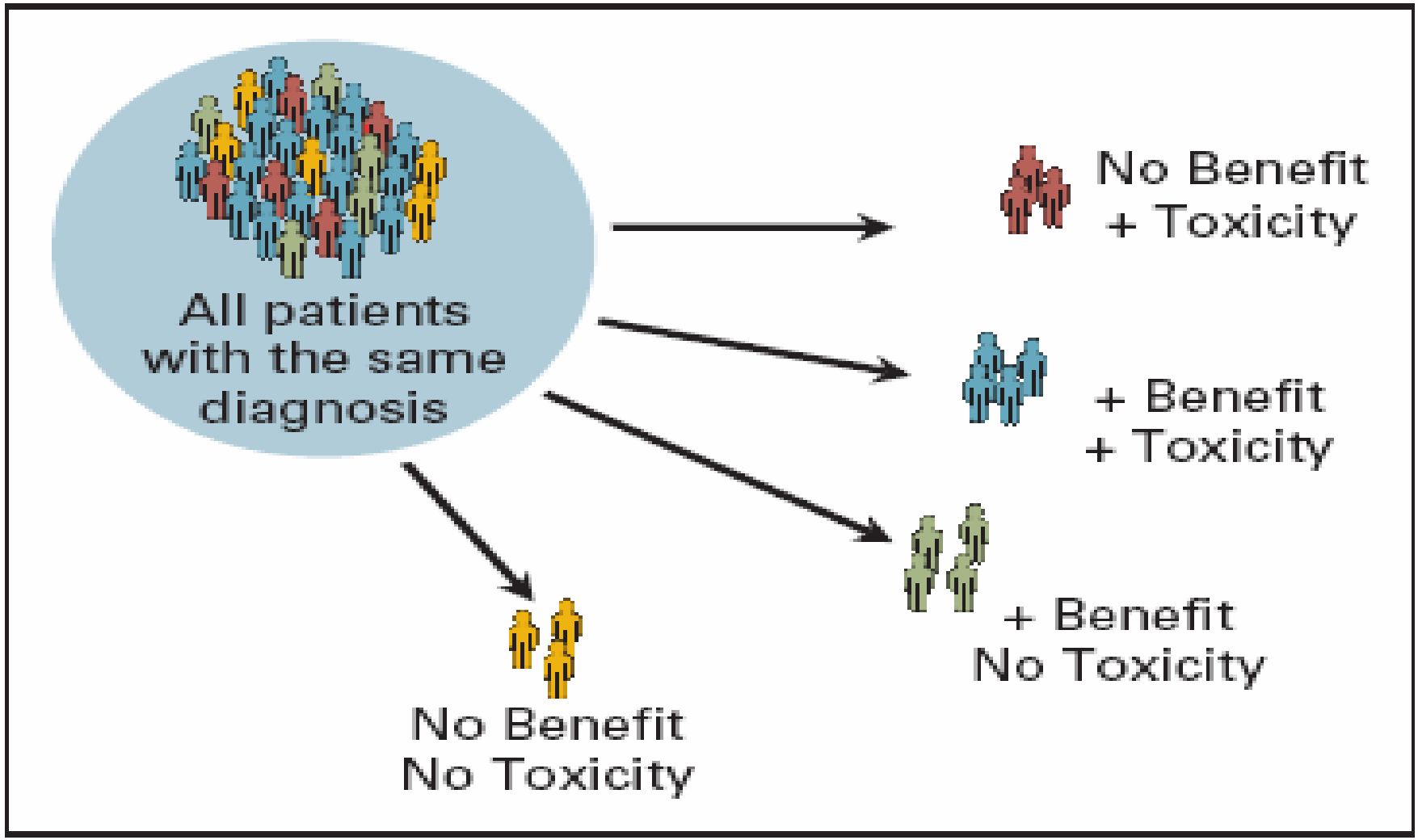
Copyright by Jen-pei Liu, PhD

Outline

- Introduction
- Overview of Biochips and Microarray Technology
- Validation and Analytical Performance
- Clinical Performance
  - Methods when clinical truth (gold standard) is present
  - Methods when clinical truth is absence
- Issues on experimental design
- Discussion and summary
- References

Copyright by Jen-pei Liu, PhD

Introduction
Biochip Diagnostic Products



Introduction
Why? Biochip Diagnostic Products

- 70% of medical devices are exported
- Pharmacogenomics and microarray for biochip diagnostic devices
- No need for high-cost and long-term clinical trials
- Ex-vivo clinical effectiveness studies if necessary
- Golden opportunity for biopharm and biotech industry in Taiwan

Copyright by Jen-pei Liu, PhD

Introduction

- The US Premarket Regulation
  - Premarket Approval (PMA)
  - Premarket Notification - 510(k)
  - *De Novo* Review – 513(f)(2)

Copyright by Jen-pei Liu, PhD

Introduction

- Classification of In Vitro Devices (IVD)
  - Intended use
    - What is tested? Presence and absence of the factor V Leiden mutation
  - Indication
    - Why a patient would be tested
  - Risk

Copyright by Jen-pei Liu, PhD

Introduction

- Class I
  - Lowest risk – immunohistochmeical reagents used as adjuncts for diagnosis
- Class II
  - Moderate risk – molecular tests for factor V Leiden
- Class III
  - Greatest risk – digital image analysis system for Pap smears

Copyright by Jen-pei Liu, PhD

Introduction

- Premarket Approval (PMA)
  - Class III devices
  - Highest level of review
  - Preclinical evaluation
    - Before the device is used in a clinical trial
  - Clinical performance
    - Samples from patients
    - Trials conducted outside clinical laboratories

Copyright by Jen-pei Liu, PhD

Introduction

- Premarket Notification – 510(k)
  - Class I or II devices
  - A 510(k) submission 90 days before offering the device for sale
  - Preclinical data and may not include clinical data if it is cleared to be substantially equivalent to the predicate (previously cleared) device for the same intended use

Copyright by Jen-pei Liu, PhD

Introduction

- *De Novo* Review – 513(f)(2)
  - No predicate device
  - Automatically classified as Class III devices
    - Submitted as *De Novo* candidates
    - Not substantially equivalent decision by FDA
    - Submit a request for *De Novo* classification
    - May be reclassified as class I or II devices
    - Roche AmpliChip CYP 450 device

Copyright by Jen-pei Liu, PhD

Introduction

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

Copyright by Jen-pei Liu, PhD

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

Copyright by Jen-pei Liu, PhD

Overview of Biochips and Microarray Technology

- **Diagnostic Tests**
  EGFR inhibitor – Candidate Gene Approach
  The Roche AmpliChip® CYP450 Microarray
  - A Genotyping Device
- **Biochip Diagnostic System**
- **Introduction to Microarray – Genome-wide Approach**
  - **cDNA microarray**
  - **In situ synthesis**
  - **Comparison of Platforms**
- **Source of Variation**
- **Review of Statistical Methods**

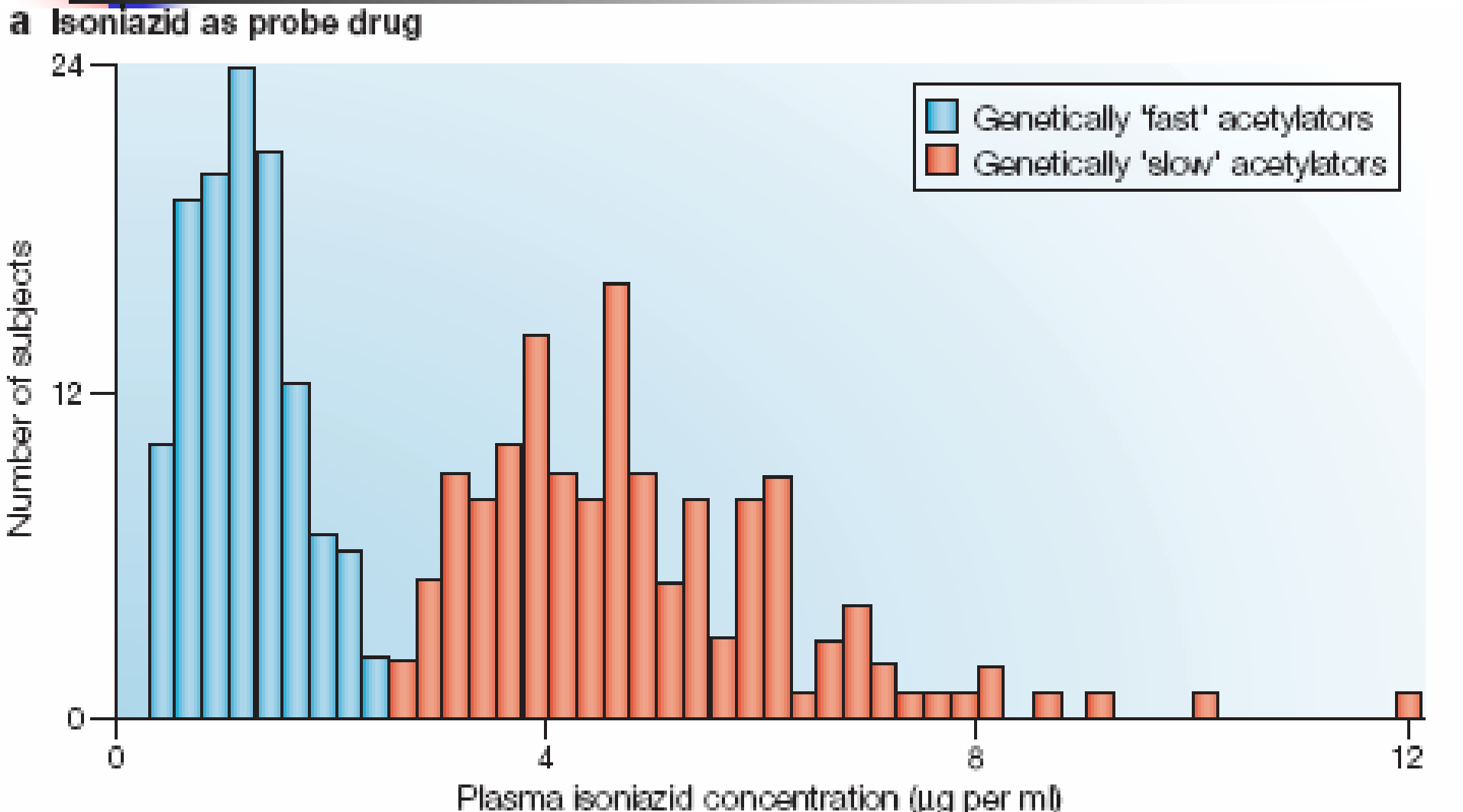
Copyright by Jen-pei Liu, PhD

Biochip Diagnostic Tests

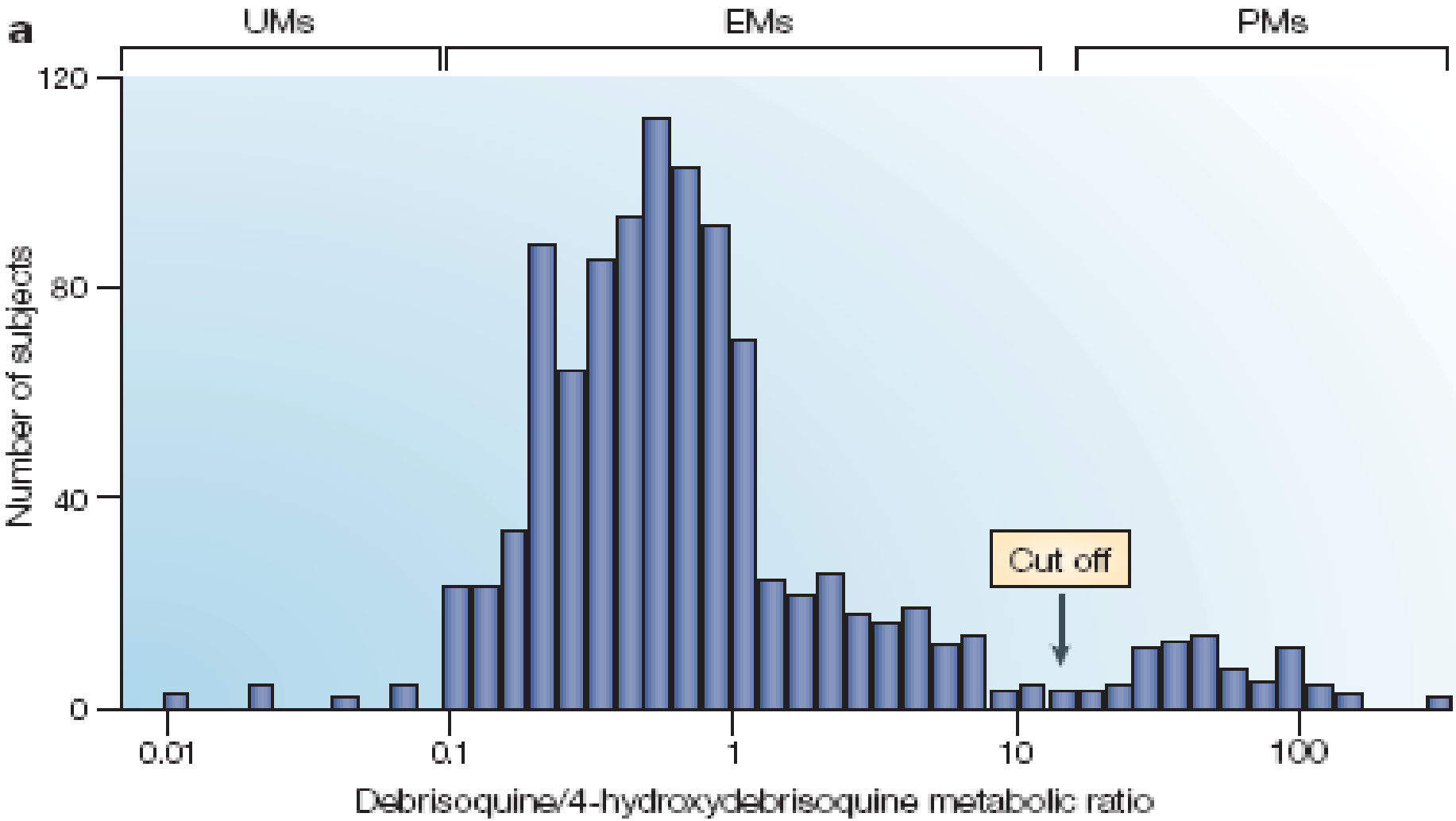
- Candidate Gene Strategies
  - Therapeutic mechanism of action studies
- Genomic-wide Strategies
  - In Vitro Discovery Approaches
  - Comparative Studies of Ex Vivo Tissues
  - Murine in Vivo Comparative Studies

Copyright by Jen-pei Liu, PhD

Biochip Diagnostic Tests



Biochip Diagnostic Tests



Targeted Clinical Trials

- HER2 (the human epidermal growth factor receptor 2) gene in metastatic breast cancer - Herceptin - requirement of screening the patients with over-expressed HER2 level (Slamon, 2001).
- Estrogen receptor ploymorphism - Estrogen Replacement Atherosclerosis trial (ERA, Herrington, et al, 2002): a total of 9 SNPs were identified and interaction between treatment of HRT and some of SNPs in elevation of lipid levels is suggested
- Sample size determination: Fijal, et al. (2000) and Maitournam and Simon (2005) and Simon and Maitournam (2004)

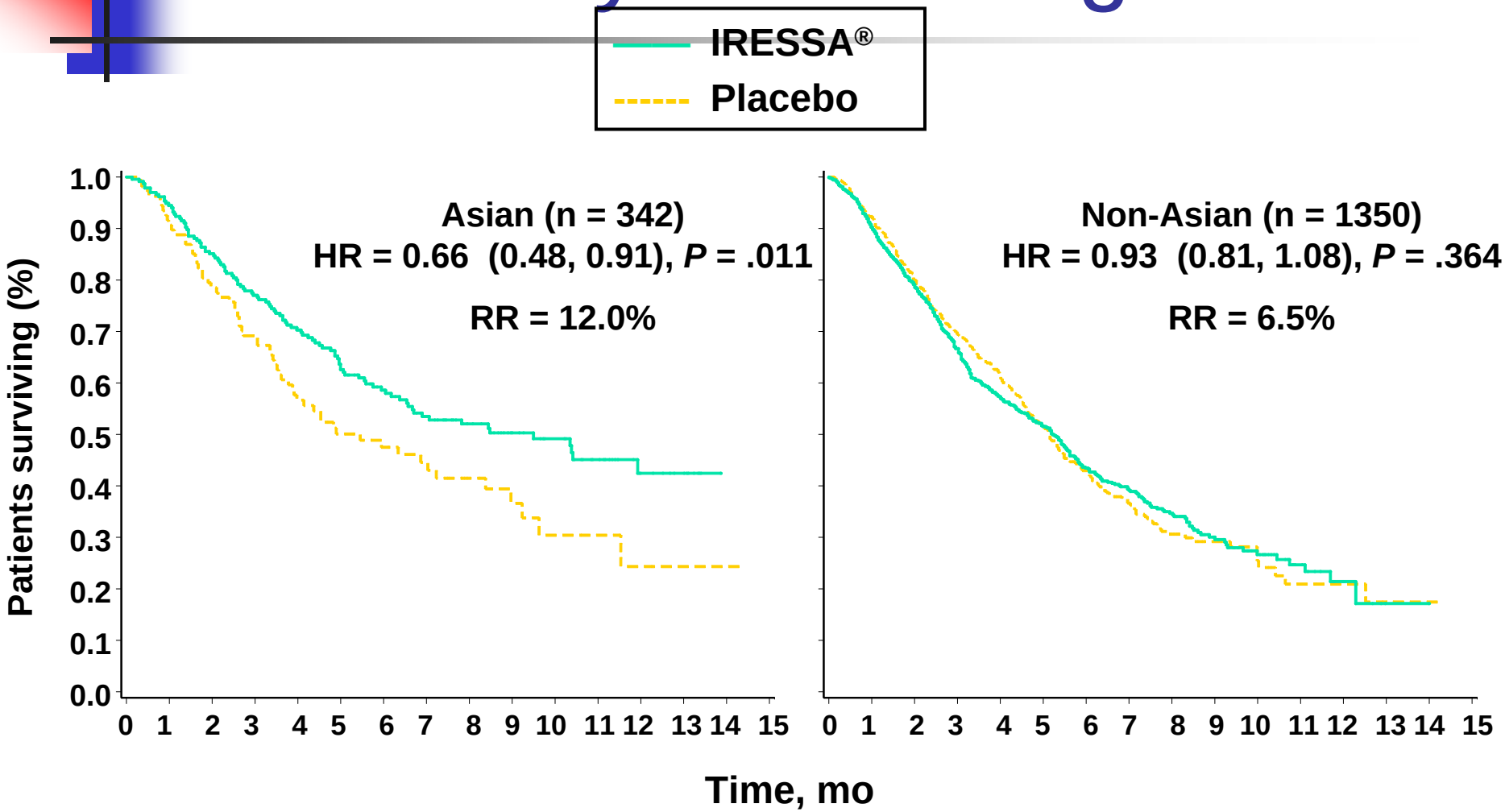
Copyright by Jen-pei Liu, PhD

Targeted Clinical Trials and EGFR

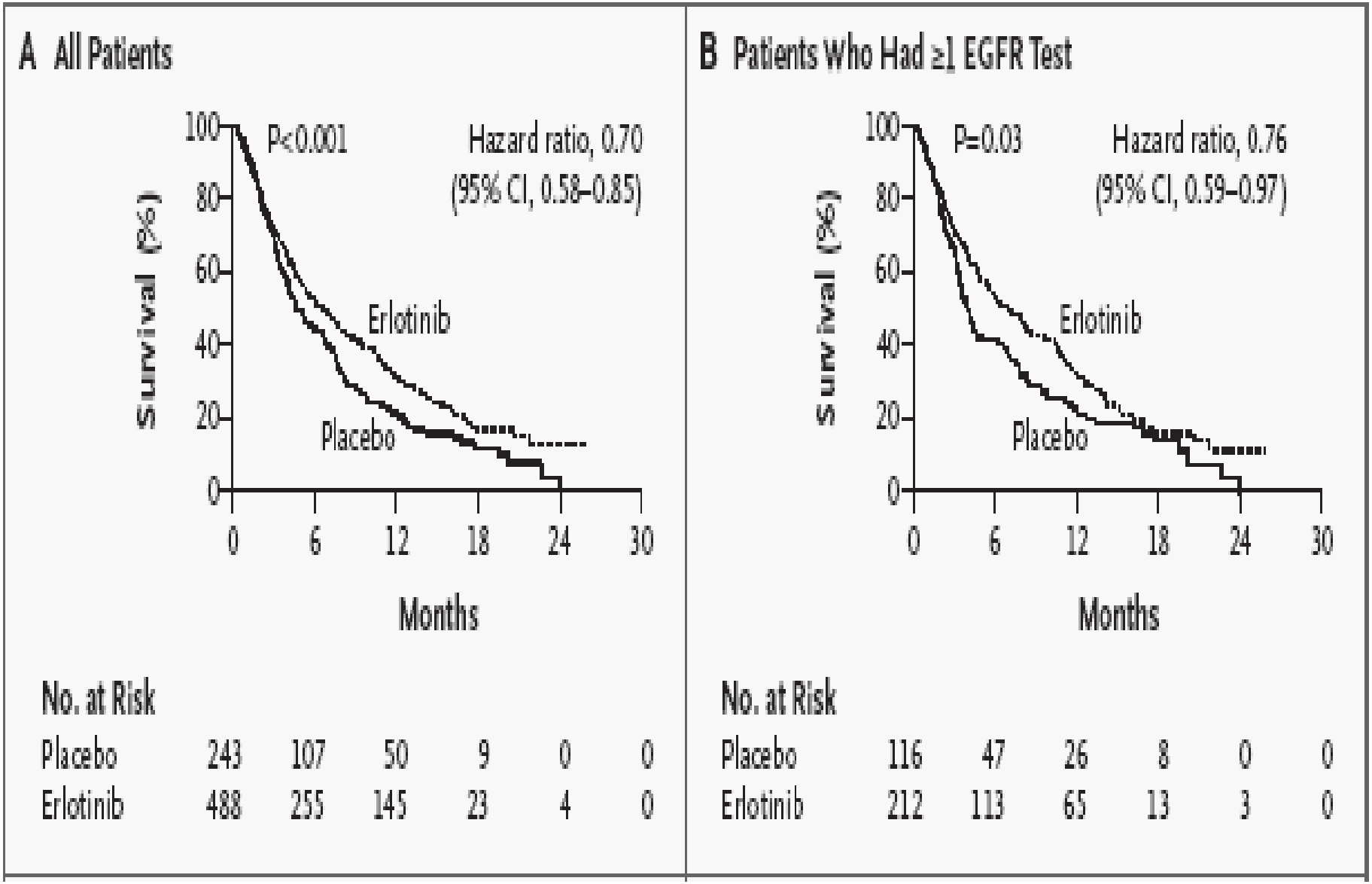
- The epidermal growth factor receptor (EGFR) inhibitor for the non-small cell lung cancer.
- Iressa (gefitnib) and Tarceva (Erlotinib) target at the EGFR pathway – EGFR tyrosine kinase inhibitor
- Efficacy is correlated to
  - race
  - number of gene copies
  - protein expression
  - EGFR mutation

Gappuzzo et al. (JNCI, 2005), Tsao, et al (NEJM, 2005)

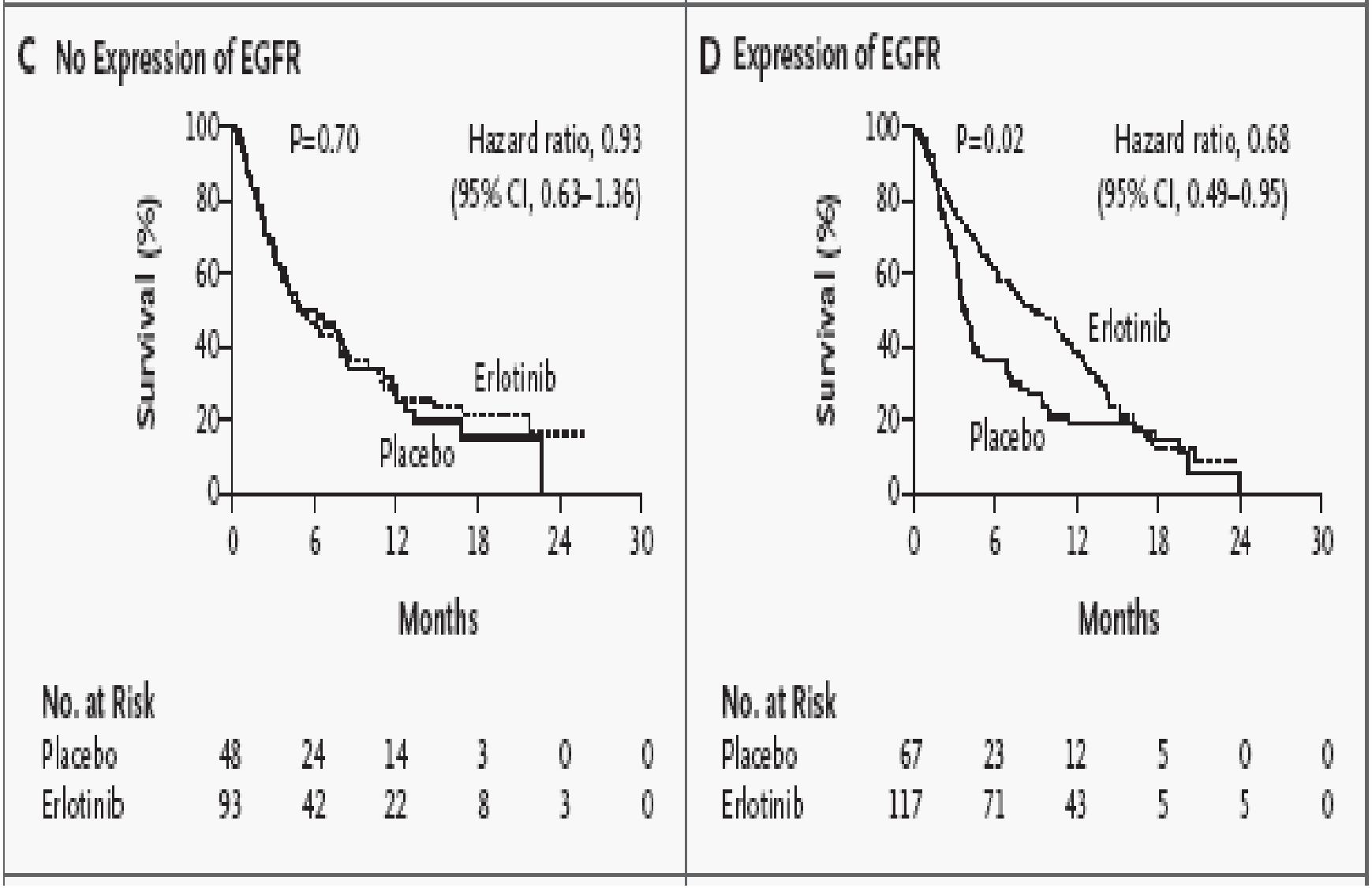
Survival by Ethnic Origin



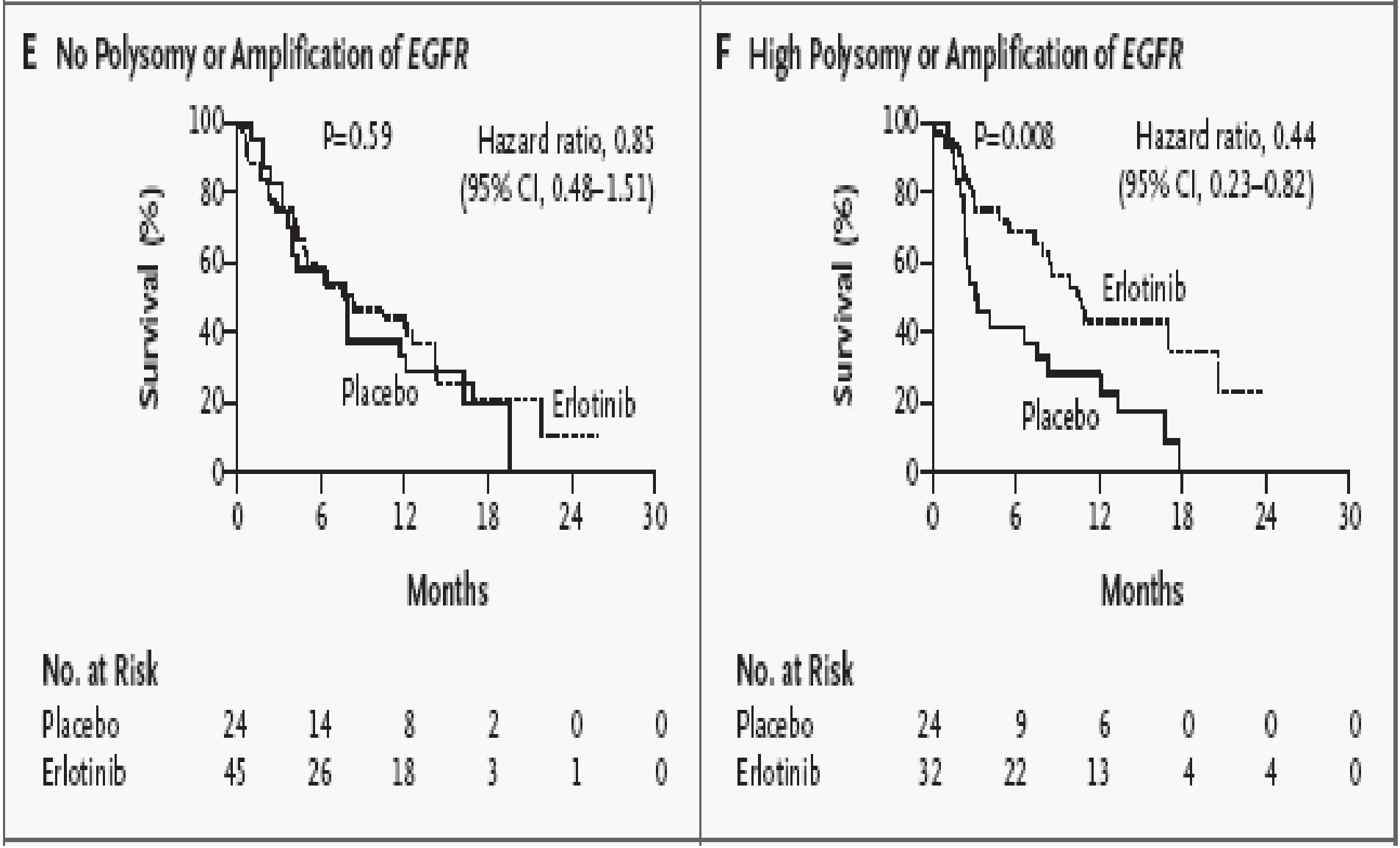
Copyright by Jen-pei Liu, PhD



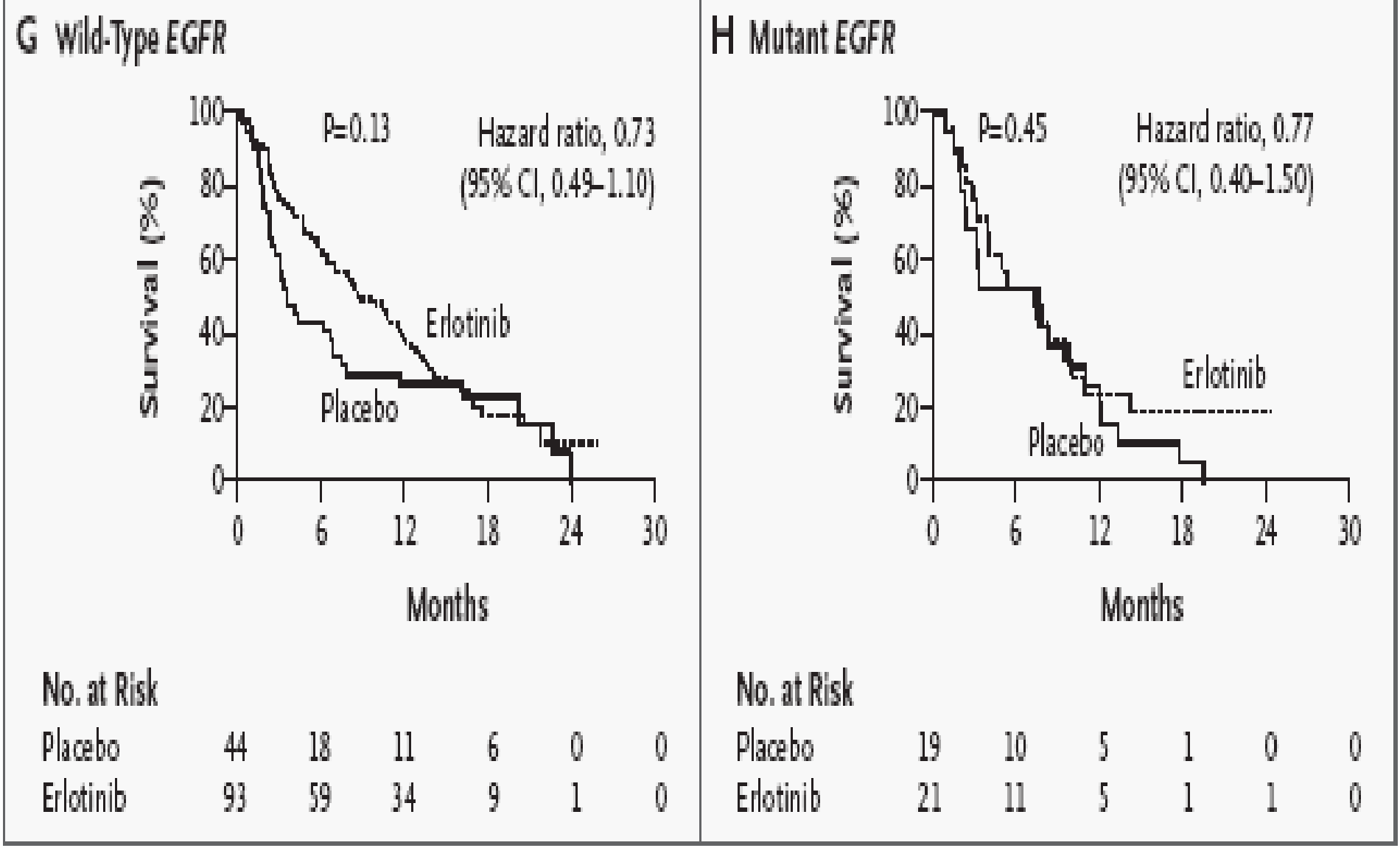
From: Tsao, et al (2005, NEJM)



From: Tsao, et al (2005, NEJM)



From: Tsao, et al (2005, NEJM)



From: Tsao, et al (2005, NEJM)

Biochip Diagnostic Tests

- Requirement of diagnostic tests for measurements of EGFR expression levels.
- Approved diagnostic devices:
  - DAKO Herceptest® (IHC assay)
  - PATHWAY™ Her 2 (Clone CB11mouse monoclonal antibody)
  - Semi-quantitative measurements (0, 1+, 2+, and 3+

Copyright by Jen-pei Liu, PhD

Roche AmpliChip CYP450 Microarray

- The Roche AmpliChip CYP 450 Test was the first DNA microarray diagnostic test to be approved by the US FDA on Dec. 23, 2004 to analyze one of the genes from a family of genes called cytochrome P450 genes which are active in the liver to metabolize drugs and other compounds such as grape fruit.

Copyright by Jen-pei Liu, PhD

Roche AmpliChip CYP450 Microarray

- Cytochrome (CYP) P450 is a family of genes in all living creatures
- CYP450 plays a primary role in metabolism
- CYP450 genes have been in existence for more than 3.5 billion years
- In humans, enzymes encoded by the CYP450 are found primarily in liver, where they metabolize drugs, toxins, and other foreign substances that enter the body
- 2C19 and 2D6 metabolize 25% of drugs

Copyright by Jen-pei Liu, PhD

CYP2C19

Proton pump inhibitors	Anti-epileptics	Others
Omeprazole	Diazepam	Amitriptyline
Lansoprazole	Phenytoin	Clomipramine
Pantoprazole	Phenobarbitone	Cyclophosphamide
		Progesterone

CYP2D6

Beta blockers	Antidepressants	Antipsychotics	Others
Carvedilol	Amitriptyline	Haloperidol	Codeine
Metoprolol	Clomipramine	Risperidone	Dextromethorphan
Propafenone	Desipramine	Thioridazine	Flecainide
Timolol	Imipramine		Mexiletine
	Paroxetine		Ondansetron
			Tamoxifen
			Tramadol
			Venlafaxine

Roche AmpliChip CYP450 Microarray

- Phenotypes of CYP2D6 can be classified
  - poor (no enzyme activity)
  - intermediate (reduced enzyme activity)
  - extensive (”normal” enzyme activity)
  - Ultra-rapid (higher than normal enzyme activity)
- Phenotypes of CYP2C19 can be classified
  - Poor (no enzyme activity)
  - extensive (”normal” enzyme activity)

Copyright by Jen-pei Liu, PhD

CYP2C19

Allele	1	2	3
1	E	E	E
2		P	P
3			P

CYP2D6

Allele	1	2	3	4	5	6	7	8	9	10	11	14A	14B	15	17	19	20	25	26	29	30	31	35	36	40	41	1XN	2XN	4XN	10XN	17XN	35XN	41XN
1	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	U	U	E	E	E	U	E
2		E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	U	U	E	E	E	U	E
3			P	P	P	P	P	P	I	I	P	P	N	P	I	P	P	N	N	I	N	N	E	I	P	I	E	E	P	I	I	E	I
4				P	P	P	P	P	I	I	P	P	N	P	I	P	P	N	N	I	N	N	E	I	P	I	E	E	P	I	I	E	I
5					P	P	P	P	I	I	P	P	N	P	I	P	P	N	N	I	N	N	E	I	P	I	E	E	P	I	I	E	I
6						P	P	P	I	I	P	P	N	P	I	P	P	N	N	I	N	N	E	I	P	I	E	E	P	I	I	E	I
7							P	P	I	I	P	P	N	P	I	P	P	N	N	I	N	N	E	I	P	I	E	E	P	I	I	E	I
8								P	I	I	P	P	N	P	I	P	P	N	N	I	N	N	E	I	P	I	E	E	P	I	I	E	I
9									I	I	I	I	N	I	I	I	I	N	N	I	N	N	E	I	I	I	E	E	I	I	I	E	I
10										I	I	I	N	I	I	I	I	N	N	I	N	N	E	I	I	I	E	E	I	I	I	E	I
11											P	P	N	P	I	P	P	N	N	I	N	N	E	I	P	I	E	E	P	I	I	E	I
14A												P	N	P	I	P	P	N	N	I	N	N	E	I	P	I	E	E	P	I	I	E	I
14B													N	N	N	N	N	N	N	N	N	N	E	N	N	N	N	N	N	N	N	N	N
15														P	I	P	P	N	N	I	N	N	E	I	P	I	E	E	P	I	I	E	I
17															I	I	I	N	N	I	N	N	E	I	I	I	E	E	I	I	I	E	I
19																P	P	N	N	I	N	N	E	I	P	I	E	E	P	I	I	E	I
20																	P	N	N	I	N	N	E	I	P	I	E	E	P	I	I	E	I
25																		N	N	N	N	N	E	N	N	N	N	N	N	N	N	N	N
26																			N	N	N	N	E	N	N	N	N	N	N	N	N	N	N
29																				I	N	N	E	I	I	I	E	E	I	I	I	E	I
30																					N	N	E	N	N	N	N	N	N	N	N	N	N
31																						N	E	N	N	N	N	N	N	N	N	N	N
35																							E	E	E	E	U	U	E	E	E	U	E
36																								I	I	I	E	E	I	I	I	E	I
40																									P	I	E	E	P	I	I	E	I
41																										I	E	E	I	I	I	E	I

Roche AmpliChip CYP450 Microarray

- Unequal genetic variations (polymorphism) of CYP2C19 and CYP2D6
- CYP2D6 has more than 80 distinct allelic variants
- Poor metabolizers of CYP2D6
  - Caucasians: 7%
  - African-American: 2-4%
  - Asian: 1-2%
    - CYP2D6*10 allele (50% allele frequency)
    - CYP2D6*17 and CYP2D6*29 alleles (~30%)

Copyright by Jen-pei Liu, PhD

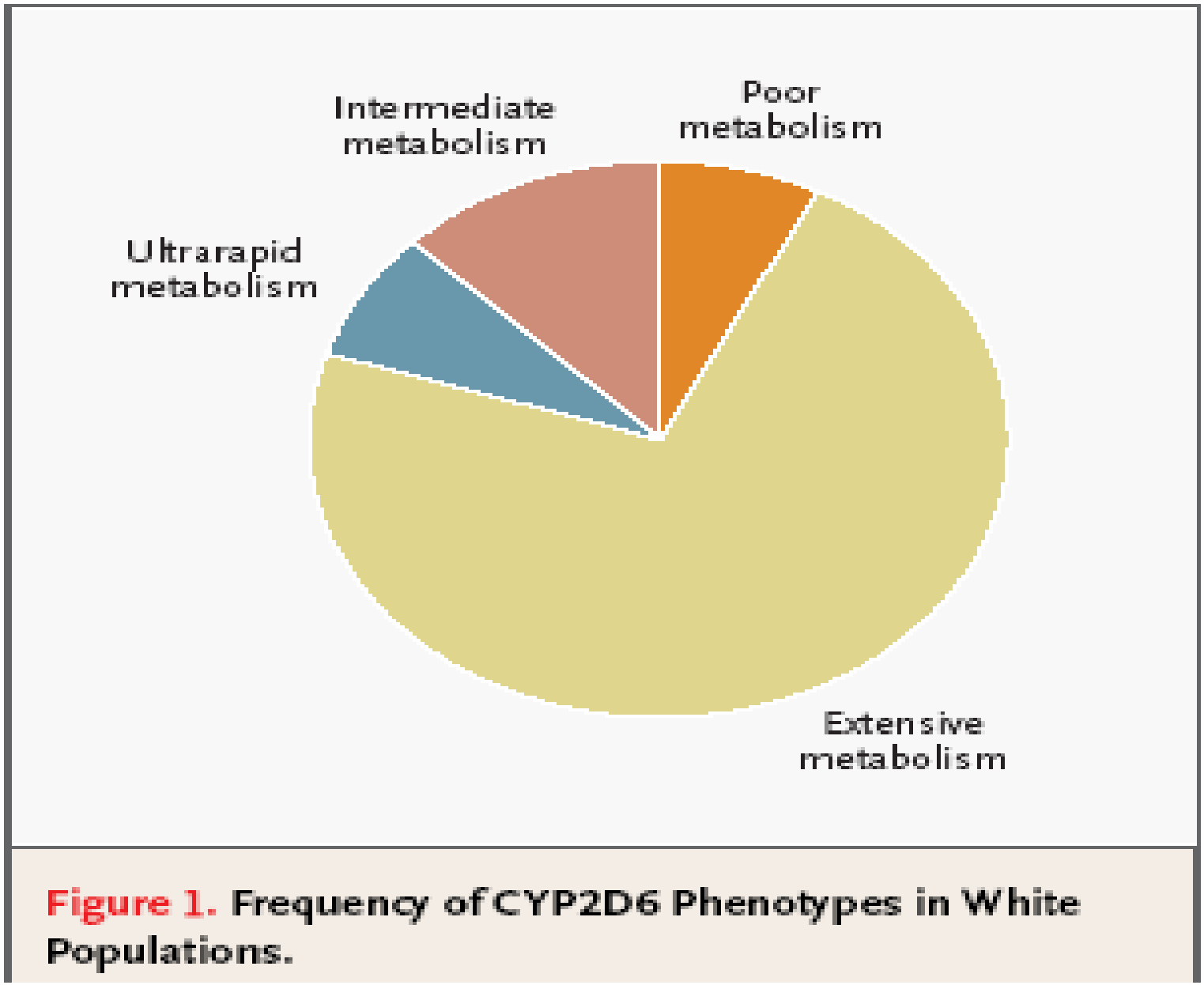


Figure 1. Frequency of CYP2D6 Phenotypes in White Populations.

From Carco Y. NEJM 2005

Copyright by Jen-pei Liu, PhD

Roche AmpliChip CYP450 Microarray

- CYP2D6 Gene duplications (ultra rapid metabolizers)
  - Ethiopians: 29%
  - Southern Europeans: 10%
  - Northern European: 1-2%

Copyright by Jen-pei Liu, PhD

Roche AmpliChip CYP450 Microarray

- CYP2C19*2 and CYP2C19*3 alleles for poor metabolizer
- Null alleles are caused by a SNP that either causes a splice site or a stop codon
- Frequency
  - Asian: 13-23%
  - Caucasian: 3-5%
  - African-American: 3-5%

Copyright by Jen-pei Liu, PhD

Roche AmpliChip CYP450 Microarray

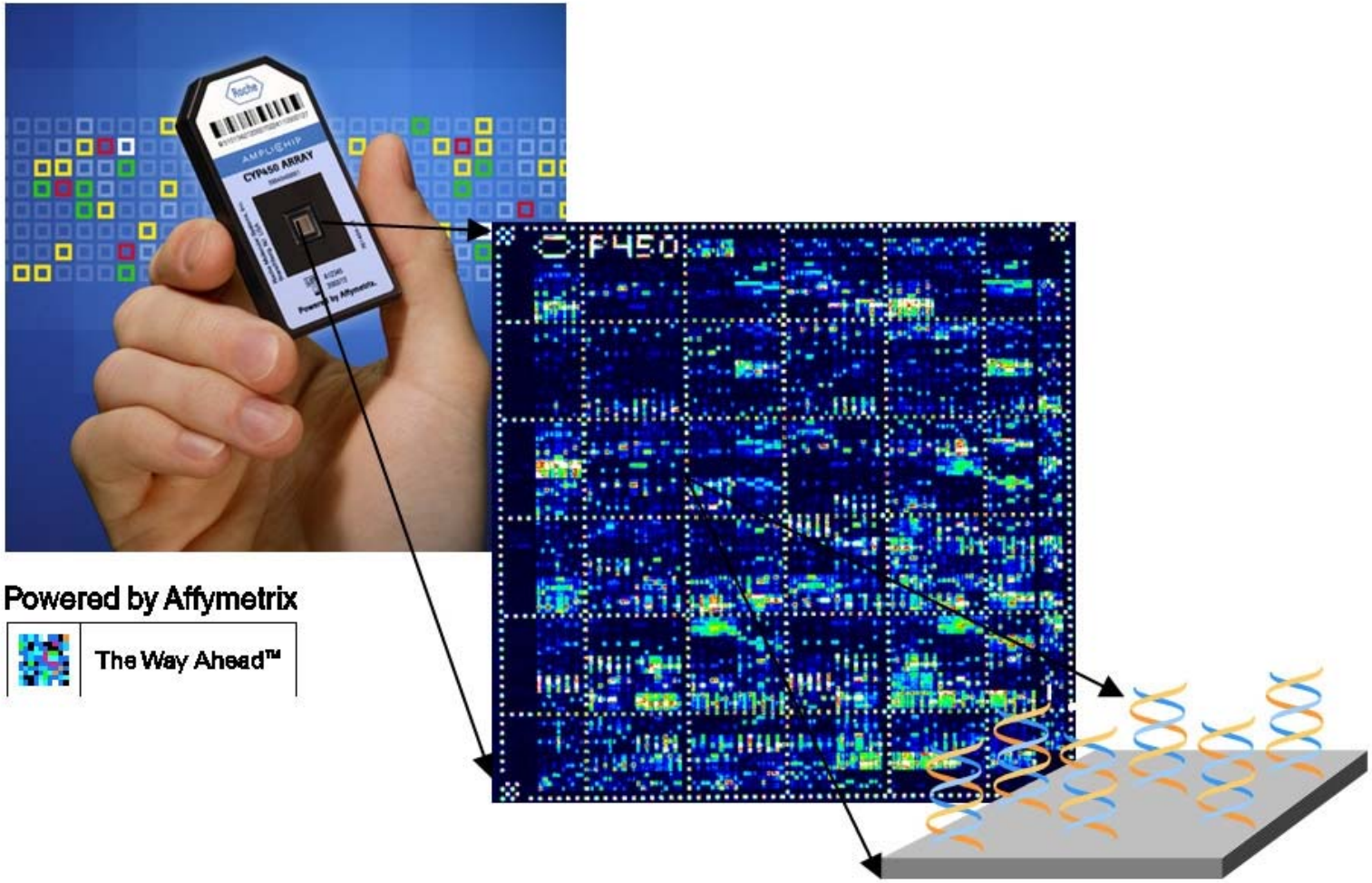
Table 10: Geographic Differences in CYP2C19 Allelic Frequencies

Allele	Predicted Enzymatic Activity	Chinese	Black	Caucasian
*1	Normal			
*2	None	30%	17%	15%
*3	None	5%	<1%	<1%

Copyright by Jen-pei Liu, PhD

Roche AmpliChip CYP450 Microarray

- Poor or immediate metabolizer: toxic adverse events (side effects) even at lower dose
- Ultra rapid metabolizer: no efficacy even at higher dose
- Necessity of the microarray-based pharmacogenomic devices to provide genotyping of the CYP2D6 and CYP2C19 and predictive phenotype of associated enzyme activities
- Prevention of harmful drug interactions and assurance of the use of the optimal dose


AMPLICHIP
CYP450 ARRAY
Powered by Affymetrix
The Way Ahead™
CYP450
Each 20 μm² cell on the array can contain
10⁷ DNA fragments, or "probes"

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

Copyright by Jen-pei Liu, PhD

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.

Quality Assurance and Quality Control

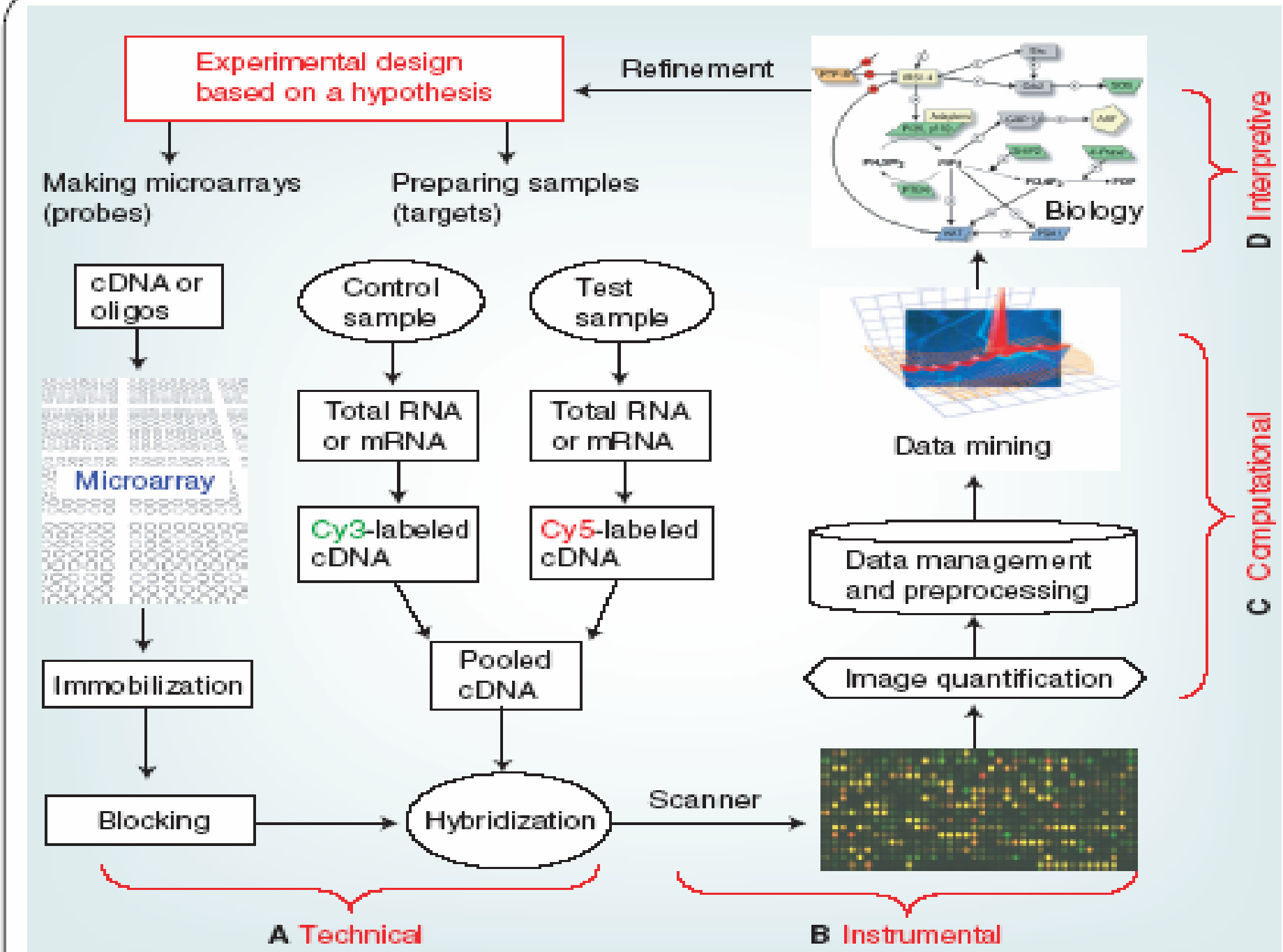
Large variability in microarray data and their interpretation is due to

- Data quality
- Standardization
- A wide diversity of different platforms
- Lack of reproducibility between labs and across platforms
- Lack of consensus QA/QC criteria for performance evaluation
- Many different approaches to analyzing the same set of data

Quality Assurance and Quality Control

- Four types of factors affecting microarray experiments
  - Technical
  - Instrumental
  - Computational
  - Interpretive

A single hidden and uncontrolled factor can completely negate a microarray experiment


Experimental design based on a hypothesis
Refinement
Making microarrays (probes)
Preparing samples (targets)
Biology
D Interpretive
cDNA or oligos
Control sample
Test sample
Microarray
Total RNA or mRNA
Total RNA or mRNA
Data mining
Cy3-labeled cDNA
Cy5-labeled cDNA
Data management and preprocessing
C Computational
Pooled cDNA
Immobilization
Image quantification
Blocking
Hybridization
Scanner
A Technical
B Instrumental

Quality Assurance and Quality Control

- Technical Factors
  - Microarray manufacturing
  - Sample collection
  - RNA extraction
  - cDNA/cRNA synthesis
  - Labeling with fluorescent dye
  - Hybridization

Copyright by Jen-pei Liu, PhD

Quality Assurance and Quality Control

- Instrumental Factors
  - Imaging acquisition
  - Quantification
- Computational Factors
  - Data processing
  - Normalization
  - Analysis
- Interpretive Factors
  - Biological reasoning

Copyright by Jen-pei Liu, PhD

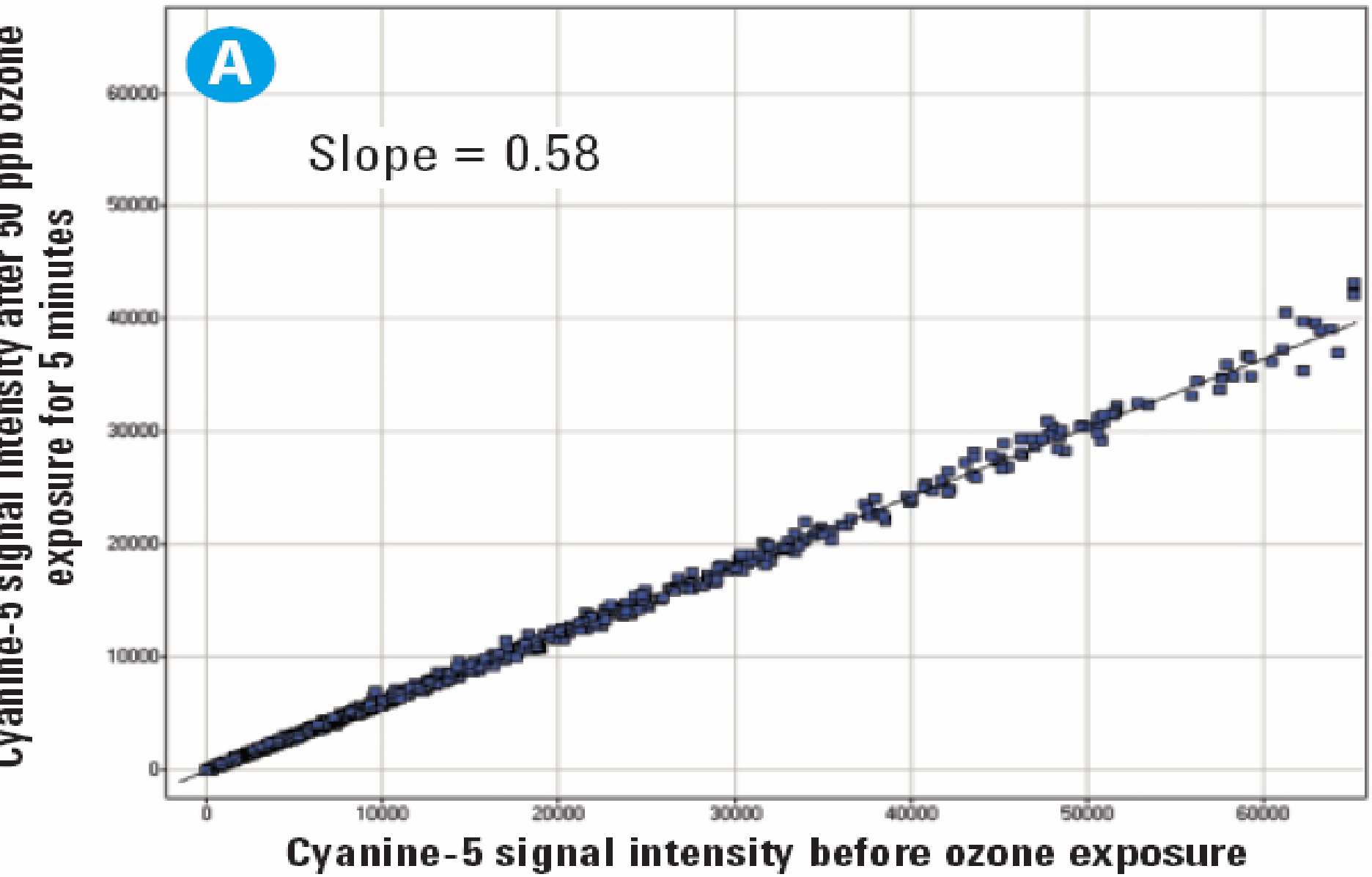


Figure 3A: Wash Protocol without Stabilization and Drying Solution.

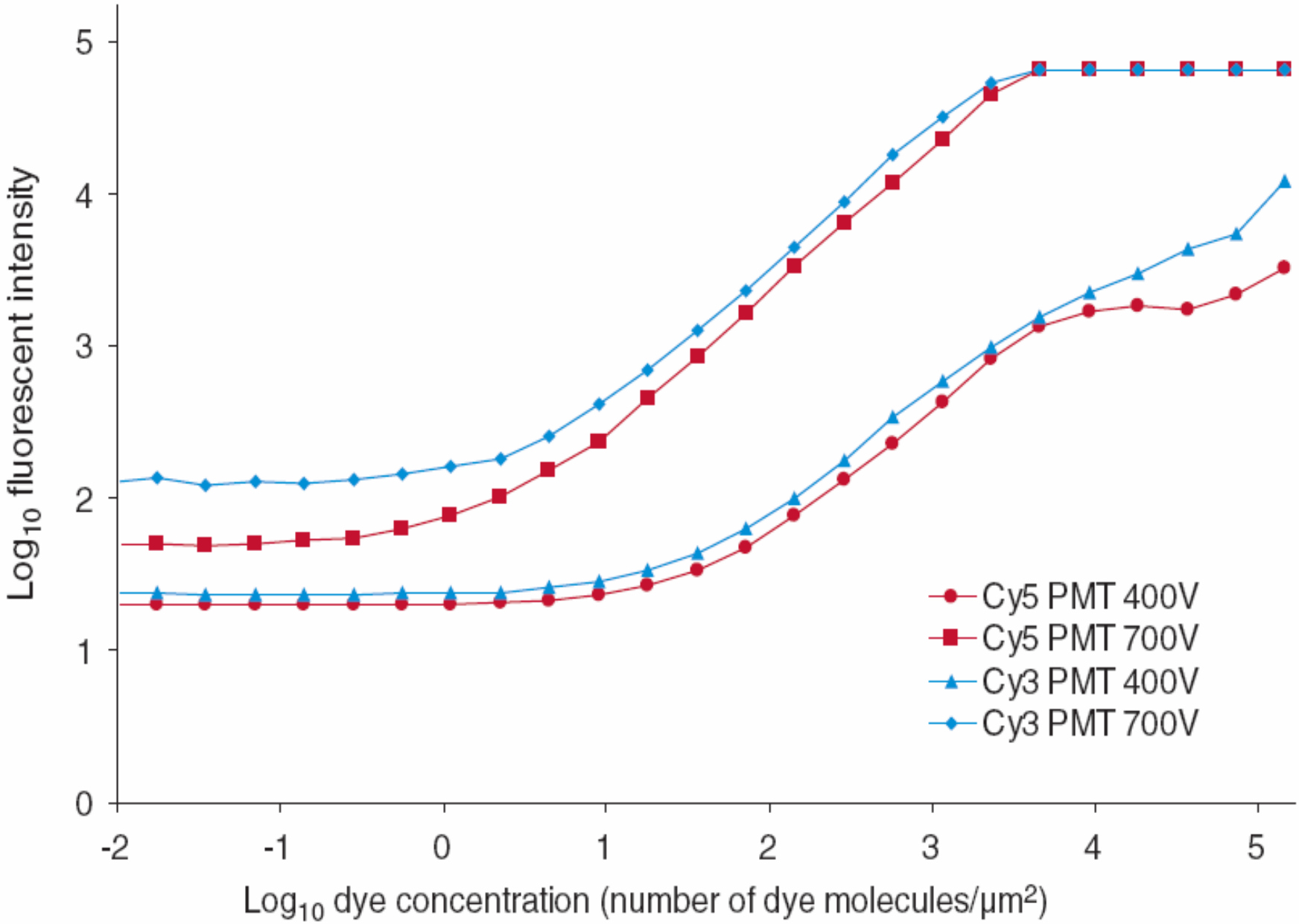


Figure 3. Calibration curves for the Cy5 and Cy3 dyes under different PMT gain settings.

Quality Assurance and Quality Control (Computational)

- Issues
  - Image quantification
  - Data management
  - Data preprocessing (e.g., quality filtering, background handling)
  - Data normalization

Copyright by Jen-pei Liu, PhD

Quality Assurance and Quality Control (Computational)

- Issues
  - Statistical methods for identifying differentially expressed genes
  - Clustering and classification methods
  - Evaluation of data quality (e.g., reproducibility and accuracy)
  - Cross-array (hybridization) reproducibility

Copyright by Jen-pei Liu, PhD

Quality Assurance and Quality Control (Computational)

- Suggestions
  - Dye-flip average after Lowess normalization for spotted cDNA arrays
  - ANOVA methods
  - Permutation tests, SAM, Bonferroni correction for false discovery rate
  - Pay attention to false non-discovery rate
  - Gene (feature) selection during the cross-validation
  - ***NO COMPUTATIONAL METHODS CAN SAVE A FAILED MICROARRAY EXPERIMENT***

Cross-platform and intra- and inter-laboratory reproducibility

- Different designs
- Different concepts
- Different referenced samples
- Different procedures for sample acquisition
- Different experimental protocols
- Different methods for data analysis

Copyright by Jen-pei Liu, PhD

Cross-platform and intra- and inter-laboratory reproducibility

- Only recognition and evaluation of reproducibility:
  - Dobbin et al. (Clinical Cancer Research, 2005)
  - Larkin et al. (Nature Methods, 2005)
  - Irizarry et al. (Nature Methods, 2005)
  - Members of the Toxicogenomic Research Consortium (Nature Methods, 2005)
  - Tan, et al. (Nucleic Acids Research, 2003)
  - Yauk, et al. (Nucleic Acids Research, 2004)

Different methods for data analysis

Copyright by Jen-pei Liu, PhD

Crossplatform and intra- and inter-laboratory reproducibility

- Conclusions from cross-laboratory and cross-platform studies are meaningless when there are fundamental problems in achieving acceptable intra-laboratory reproducibility
- Conclusions from microarray data analysis are meaningless when a minimum requirement for data quality is not met and provided

Microarray Standards and QA/QC Metrics

- Universal reference RNA samples for two-color platforms
  - The common controls for the common reference design
  - Standard test material for optimizing microarray protocols
  - Critical to manufacture reproducibly the universal reference RNA in bulky quantity

Copyright by Jen-pei Liu, PhD

Microarray Standards and QA/QC Metrics

- External RNA spike-in controls
  - A normalization set to monitor global messenger changes
  - Consensus external RNA spike-in controls independent of platforms produced by The External RNA Control Consortium for microarray and QT-RT-PCR

Copyright by Jen-pei Liu, PhD

Microarray Standards and QA/QC Metrics

- External RNA spike-in controls
  - 100 well-characterized clones of random unique sequences as determined by sequence comparison with mouse, rat, human, *Drosophila, E. coli*, mosquito, *Bacillus subtillus* and *Arabadopsis thaliana* sequence databases.
  - *Use of External RNA Spike In Controls in Gene Expression Assays* has been submitted to Clinical and Laboratory Standards Institute (formally National Committee for Clinical Laboratory Standards, NCCLS) as well as *Molecular Methods for Microarray*

Quality Assurance and Quality Control (Computational)

- A certain percent of biological samples should be measured in replicates to assess technical reproducibility
- Accuracy should be evaluated in spike-in controls and other reference materials where the test results are known
- Sample size of biological samples depends upon the level of technical variability (reproducibility)

Copyright by Jen-pei Liu, PhD

Microarray QC Metrics and Thresholds (MAQC) Project

- Sponsored by the US FDA
- Establish QC metrics and thresholds to objectively assessing the performance of microarray platforms and merits of various data analysis methods
- Two RNA samples from three species (human, rat, and mouse)

Copyright by Jen-pei Liu, PhD

The MAQC Project: MicroArray Quality Control

Microarray QC Metrics and Thresholds (MAQC) Project

- Assess the precision and cross-platform comparability
- Nature and magnitude of systematic bias assessed by QT-PCR
- Use of the calibrated RNA samples

Copyright by Jen-pei Liu, PhD

Microarray QC Metrics and Thresholds (MAQC) Project

- Four government agencies
- 10 platform providers
- 3 RNA sample provider
- 27 test sites
- 10 data analysis sites
- 200 people from more than 70 organizations

Copyright by Jen-pei Liu, PhD

Test Sites: The "official" MAQC test sites are (as of July-28-2005):

	Manufacturer (Site 1)	Site 2	Site 3
1	Affymetrix*	FDA/CDER	Ambion
2	Agilent	FDA/NCTR	Icoria
3	Applied Biosystems	EPA/NHEERL	Vanderbilt Univ.
4	Combimatrix	UCSF	Stanford
5	Eppendorf	MD Anderson	CSHL
6	GE Healthcare	UMass Boston	Genus Biosciences
7	Illumina	Duke University	Burnham Institute
8	NCI (Custom arrays)	FDA/NCTR	FDA/CBER
9	AB (TaqMan)	N/A	N/A
10	Genospectra (QuantiGene)	N/A	N/A

*Three additional sites (EPA, Novartis, and UCLA/Cedars-Sinai) will test the four RNA samples with Affymetrix platform and submit the datasets to MAQC. Ambion and Stratagene will provide RNA samples to all 29 test sites.

Microarray QC Metrics and Thresholds (MAQC) Project

- Completion of MAQC main study – Oct. 2005
- Submission of manuscript and release of MAQC datasets – Feb. 2006
- Publications – July-Sept. 2--6
- Public meeting on microarray quality control and data analysis – Dec. 2006
- Guidance on microarray quality control and data analysis – Dec. 2007

Copyright by Jen-pei Liu, PhD

Regulatory Evaluation of Biochip Products

References:

US FDA Decision Summary – P980018 (1998)

US FDA Decision Summary – P990081 (2000)

US FDA Decision Summary – K042279

US FDA Decision Summary – K043576

US FDA Decision Summary – K042259

Copyright by Jen-pei Liu, PhD

Performance Evaluation

- Analytical Performance
  - Accuracy
  - Precision
- Clinical Effectiveness
  - Diagnostic accuracy and variability
- Clinical Utility
  - Correlation with outcomes

Copyright by Jen-pei Liu, PhD

Analytical Performance

- Validation assessed functional performance
  - Precision (Reproducibility)
  - Assay Sensitivity (Limit of Detection) —ability to accurately identify positive samples
  - Assay specificity (Accuracy)
  - Interfering substances (endogenous and exogenous)
  - Validation of cut-off, reference range, or medical decision point
  - Assay range
  - Effect of excess sample and limiting sample.

Copyright by Jen-pei Liu, PhD

Precision (Reproducibility)

- Sites should include at least two external sites, with multiple operators at each site conducting the study over multiple days.
- Should include multiple product lots, and multiple reagents and instruments
- Should use clinical samples whenever possible
- The procedures used in reproducibility studies should be the same as the procedures used in marketed assay

Copyright by Jen-pei Liu, PhD

Affymetrix GeneChip Microarray Instrumentation System

- FS450Dx Fluidics Station
  - 4 modules
  - Each containing a single GeneChip microarray
  - Perform functions required for hybridization, washing, and staining
  - Up to 9 stations communicate to a workstation
- GCS3000Dx Scanner
  - A wild-field, epifluorescent, confocal, scanning laser microscope
  - Autoloader loads arrays from array cartridges
- GCOSDx (GenChip Operating Software)

Copyright by Jen-pei Liu, PhD

Affymetrix GeneChip Microarray Instrumentation System

- GCOSDx Software
  - Interface between the user and instrument system
    - Instrument control
    - Application of processing array
    - Data collection
    - Algorithms and reporting functions to produce a clinical result

Copyright by Jen-pei Liu, PhD

Affymetrix GeneChip Microarray Instrumentation System

- GCOSDx Software
  - Alignment algorithm to superimpose a grid on the image to delineate probe cells by using a checkboard image of control probes, located at the corner of the probe array to superimpose the grid on the scanned image
  - Generate cell intensity data from the image data
  - The cell analysis algorithm analyzes the image data and computes a single intensity value for each probe cell on the array and saved as a “cel” file

Copyright by Jen-pei Liu, PhD

Affymetrix GeneChip Microarray Instrumentation System

- Operational Environment
  - Programmed in C++
  - MS Window 2000 SP3 or SP4 (will move to MS XP)
  - Internet Explorer 6.0
  - Office XP
  - MDAC 2.7 SP1

Affymetrix GeneChip Microarray Instrumentation System

- Indications for use
  - To measure fluorescence signals of labeled DNA targets hybridized to GeneChip arrays
- Special Conditions for Use Statements
  - For use with separately cleared GeneChip microarray assays

Affymetrix GeneChip Microarray Instrumentation System – Analytical Performance

- Precision/Reproducibility
  - Intensity readings generated by Scanner 3000DX were not changed, manipulated or revised after scanning was completed
  - Raw .CEL data were analyze for
    - Uniformity
    - Intra-, inter-chip, intra-, inter-scanner, intra-, inter-fluidics port reproducibility
    - Random effect of FS, FS port or scanner by ANOVA
    - No scaling nor normalization to observe variability in all system components with raw data

Affymetrix GeneChip Microarray Instrumentation System – Analytical Performance

- Array Design
  - Array to assess scanner performance without hybridization step
    - 32x32 cell array to have photobleaching stability and can be scanned over 100 times
    - A single array was scanned 4 times for each of three scanners
    - Global uniformity: CV < 10%
    - Local uniformity: CV < 1% (each 400 mM2 gridded cell with surrounding 4 cells; averaged over alternating cells of the 32x32 array)

Copyright by Jen-pei Liu, PhD

Affymetrix GeneChip Microarray Instrumentation System – Analytical Performance

- Array to assess overall system performance
  - Control probes
    - Random generated synthetic sequence
    - Located in a specific pattern across the array
    - Lengths of probes: 16-25mer
    - For grid alignment and
    - To evaluate variability introduced by hybridization
  - Discrimination controls
    - 4 control probe sets
    - 30 tiled PM/MM for each probe set
    - Measurement of discrimination of PM from MM

Copyright by Jen-pei Liu, PhD

Affymetrix GeneChip Microarray Instrumentation System – Analytical Performance

- Sample Type
  - Two target sets for control and discrimination at concentrations of 0.1, 0.5, 1 and 3 pM
- Study Design
  - Goal to generate reproducibility data for consistent system performance independent of assay type

Affymetrix GeneChip Microarray Instrumentation System – Analytical Performance

- Study Design (Continued)
  - Metric: Discrimination score (PM vs. MM)
  - Design: Replicated fully factorial fully nested
  - One target aliquot was hybridized to 12 chips on 3 Fluidics Stations
  - Each array was scanned in duplicates on each of three scanners
  - 72 scans for one target aliquot
  - 3 replicates/aliquot

Copyright by Jen-pei Liu, PhD

Affymetrix GeneChip Microarray Instrumentation System – Analytical Performance

- 643 identical features on the array
- The overall average CV was 8.0%
- Between-scan CV
  - 4 scans per day
  - Scanner 1: 4.2-5.2%
  - Scanner 2: 3.2-4.0%
  - Scanner 3: 0.3-1.9%

Copyright by Jen-pei Liu, PhD

Affymetrix GeneChip Microarray Instrumentation System – Analytical Performance

- Between-scanner CV

Time Point	CV% Range
1	4.4 – 6.8
2	5.0 – 7.4
3	5.7 – 9.2
4	8.9 – 10.9

Copyright by Jen-pei Liu, PhD

Affymetrix GeneChip Microarray Instrumentation System – Analytical Performance

- Discrimination Score: 36 average discrimination scores from all arrays fell within 95% CI (average = average of 2 scanned images/array x 36 arrays)
- P-values for variability of FS, FS ports, and scanner > 0.05
- Chip-to-chip variability
  - One PM CV > 13%
  - All MM Cvs <=15%

Roche AmpliChip CYP450 2C19 Test Analytical Performance

- 23 system errors
  - 21 scanner failures
  - 2 fluidics failures
- One miscall: a called 2*/*2 sample but later sequencing revealed a 2*/*10 sample
- 100 replicates for the failure rate of the AmpliChip CYP450 system
  - One whole system failure: 1% failure rate with 95% CI 94.55 to 99.97% - one chip failed to scan and subsequent attempts also failed

Roche AmpliChip CYP450 2C19 Test Analytical Performance Precision (Reproducibility)

CYP2C19 genotype	No. Tested	Genotype Calls	Correct Calls	Correct Call Rate (95% CI)
*1 / *1	134	134 (100.0)	133	0.99 (0.97)
*1 / *2	135	135 (100.0)	135	1.00 (0.98)
*1 / *3	135	135 (100.0)	135	1.00 (0.98)
*1 / *1	135	135 (100.0)	135	1.00 (0.98)
*1 / *2	135	134 (99.3)	134	1.00 (0.98)
*1 / *2	135	134 (99.3))	134	1.00 (0.98)
Total	809	807 (99.8)	806	1.00 (0.99)

The overall results were as follows: 806/809 samples called correctly (99.6%).

Copyright by Jen-pei Liu, PhD

Roche AmpliChip CYP450 2C19 Test Analytical Performance

- Specificity

Table 2: Specificity for the CYPC19 Gene

CYP2C196 Genotype	Number of Samples Tested	Number of Correct Calls	Number of Miscalls
*1/*1	270	270	0

Copyright by Jen-pei Liu, PhD

Roche AmpliChip CYP450 2C19 Test Analytical Performance

Table 4: Detection of CYP2C19 Alleles

CYP2C19 Allele	Number of Unique Alleles Tested	Number of Correct Calls	Number of Miscalls	Number of No Calls	Percent Agreement	Number of Replicates
*1	647	647	0	0	100%	842
*2	137	136	1	0	99.3%	176
*3	14	14	0	0	100%	32
Total	796	796	0	0	99.9%	1050

Copyright by Jen-pei Liu, PhD

Roche AmpliChip CYP450 2C19 Test Analytical Performance

Table 5: Detection of Samples by CYP2C19 Genotype

CYP2C19 Genotype	Total Unique Samples	Number of Correct calls	Number of Miscalls	Number of No Calls	Percent Agreement	Genotype Call Rate
*1/*1	270	270	0	0	100.0%	100.0%
*1/*2	101	101	0	0	100.0%	100.0%
*1/*3	6	6	0	0	100.0%	100.0%
*2/*2	15	14	1[1]	0	93.3%	93.3%
*2/*3	6	6	0	0	100.0%	100.0%
*3/*3	1	1	0	0	100.0%	100.0%
Total	399	398	1	0	99.7%	99.7%

[1] *One sample, shown by sequencing to be CYP2C19 *2/*10*, was miscalled as a CYP2C19 *2/*2 genotype.*

Copyright by Jen-pei Liu, PhD

Roche AmpliChip CYP450 2C19 Test Analytical Performance

Table 6: Reference Method for CYP2C19 Allele Identification

Method(s)	Number of Samples Tested	Number of Correct Calls	Number of Miscalls	Number of No Calls
PCR-RFLP	276	276	0	0
DNA Sequencing and PCR-RFLP	123	122	1	0

Limit of detection

Lowest and highest concentration of input sample that yields a consistent and accurate result

DNA Amount (ng)	Number of Arrays	Number of Correct Calls	Positive Rate	95% Confidence Limit
50	144	144	100%	97.5 – 100%
25	144	144	100%	97.5 – 100%
2.5	144	144	100%	97.5 – 100%

The lowest level of genomic DNA at which a ≥ 95% positive rate was obtained for correct detection of the CYP2C19 gene was 2.5 ng of input DNA.

Copyright by Jen-pei Liu, PhD

Accuracy (analytical specificity)

- Real clinical samples whenever possible
- Compare test to a reference method, e.g., bi-directional DNA sequencing
- In limited cases (i.e., very rare alleles) may use contrived samples
  - Samples should mimic the molecular composition and concentration of real clinical samples

Copyright by Jen-pei Liu, PhD

Accuracy (analytical specificity)

AmpliChip CYP450 test

Test Method: DNA sequencing as the comparator

CYP 2C19 Allele	Number of Alleles Sequenced	AmpliChip Results			
		Correct Calls	Miscalls	No Calls	Percent Agreement
*1	153	153	0	0	100.0%
*2	79	78	1	0	98.7%
*3	14	14	0	0	100.0%
Total	246	245	1	0	99.6%

The agreement between the AmpliChip CYP450 Test and sequencing for CYP2C19 alleles was 99.6%.

Copyright by Jen-pei Liu, PhD

Roche AmpliChip CYP450 2D6 Test Analytical Performance

DNA Amount (ng)	Number of Arrays	# Correct Calls	Positivity Rate	95% CI
50	144	144	100%	97.5 – 100%
25	144	144	100%	97.5 – 100%
2.5	144	134	93.1%	87.6 – 96.6%

- The lowest level of genomic DNA at which a >= 95% positivity rate was obtained for correct detection of CYP2D6 (*4DxN/*41 and *4/*5 samples) was 25ng.

Potential Interferences

- Endogenous and exogenous common substances
- Commonly prescribed drugs
- Molecules similar to the analyte

AmpliChip CYP450 Test

- 10 unique patient samples were tested with and without spiking of albumin, bilirubin and triglycerides.
- Albumin – 6000 mg/dL; Bilirubin - 60 mg/dL; Triglycerides – 3000 mg/dL (approximately 10-fold greater than normal).
- Elevated levels of lipids, bilirubin and albumin in specimens did not interfere with the performance of the AmpliChip CYP450 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

Copyright by Jen-pei Liu, PhD

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

Copyright by Jen-pei Liu, PhD

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%)

Copyright by Jen-pei Liu, PhD

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

Copyright by Jen-pei Liu, PhD

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%

Copyright by Jen-pei Liu, PhD

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%

Copyright by Jen-pei Liu, PhD

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.

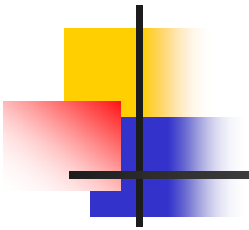
- However, sensitivity and specificity are independent of the size of the subjects used to evaluate the tests. (i.e., independent of the prevalence rate)

Copyright by Jen-pei Liu, PhD

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)

Copyright by Jen-pei Liu, PhD

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 (–)

Copyright by Jen-pei Liu, PhD

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]

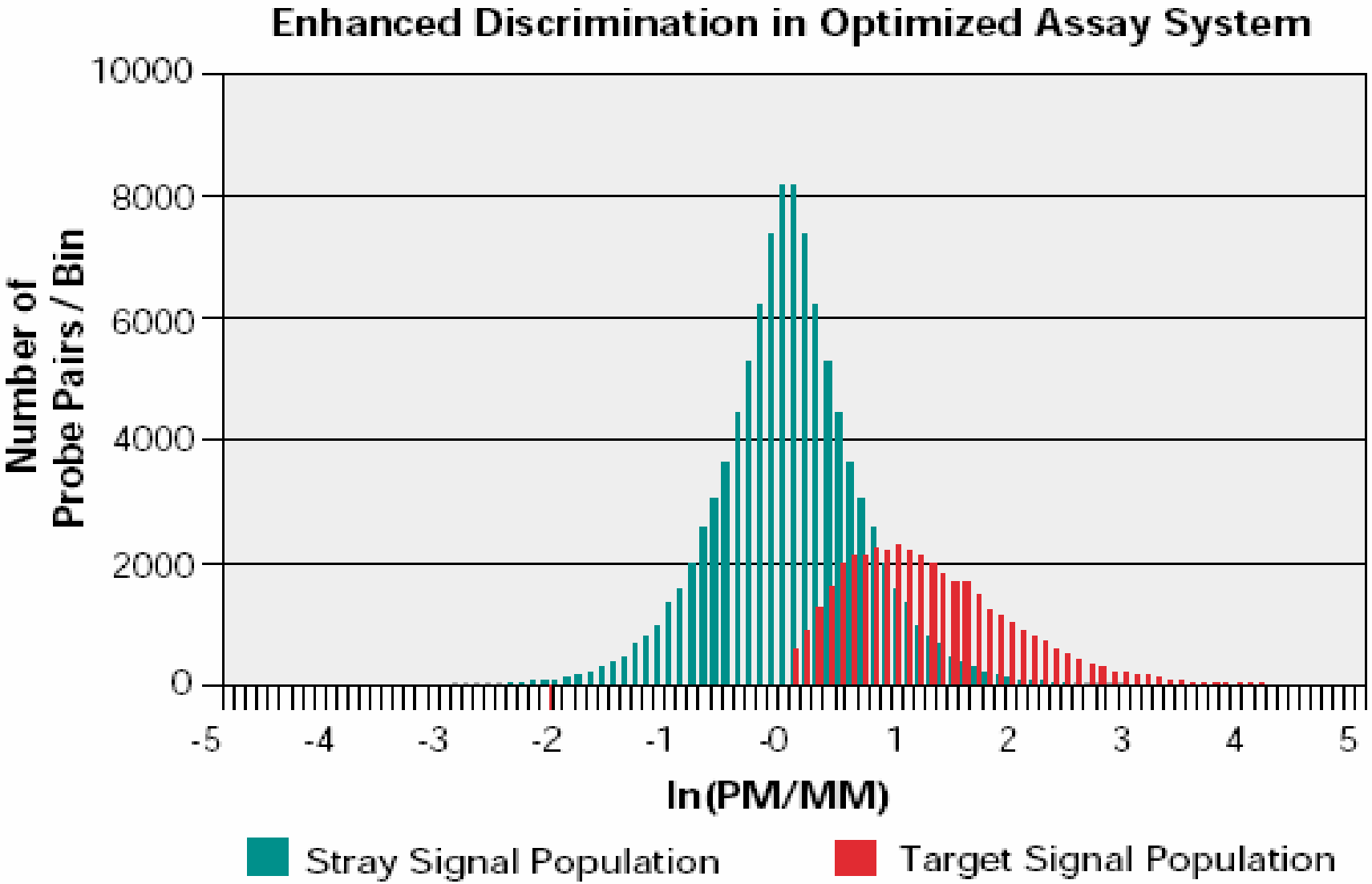
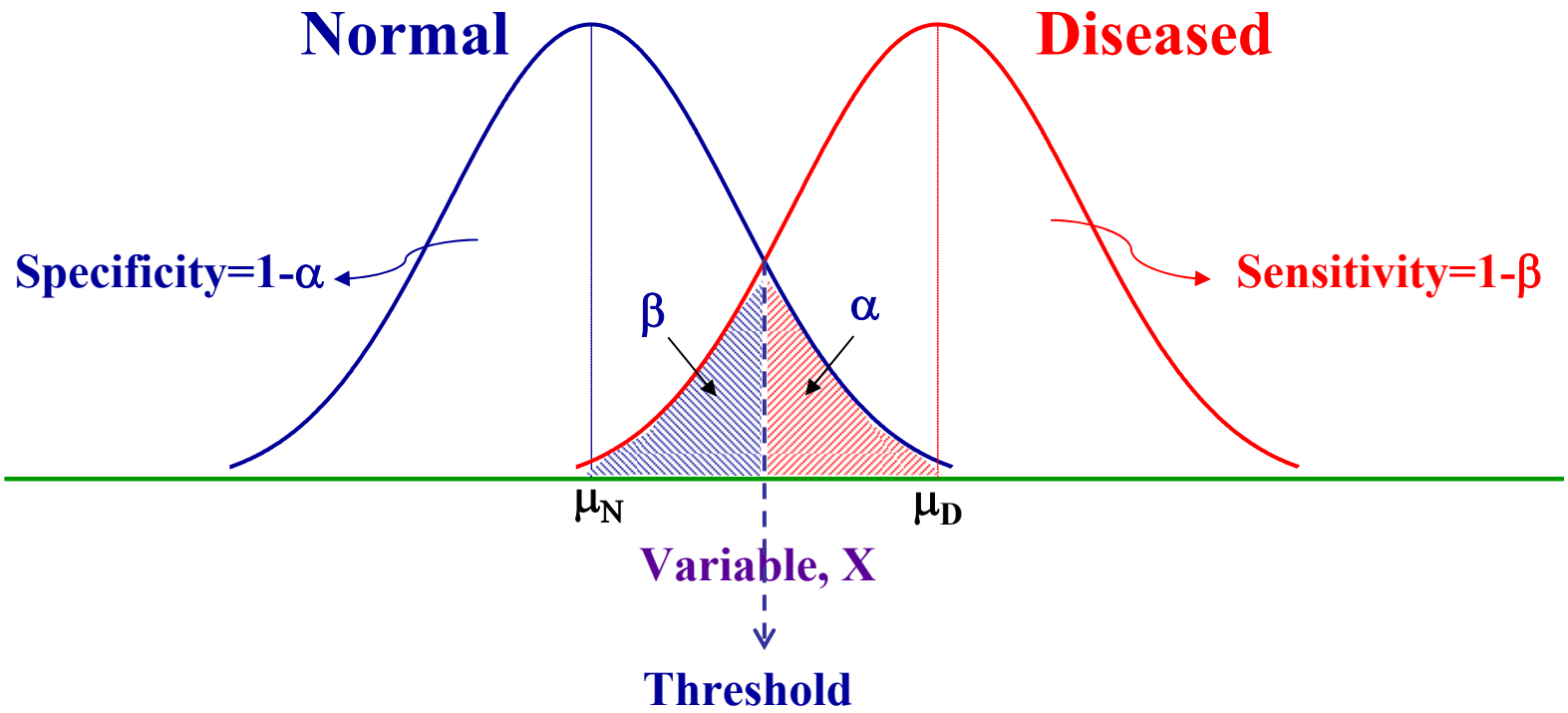


Figure 2. Global evaluation of hybridization results on a GeneChip® array.

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.

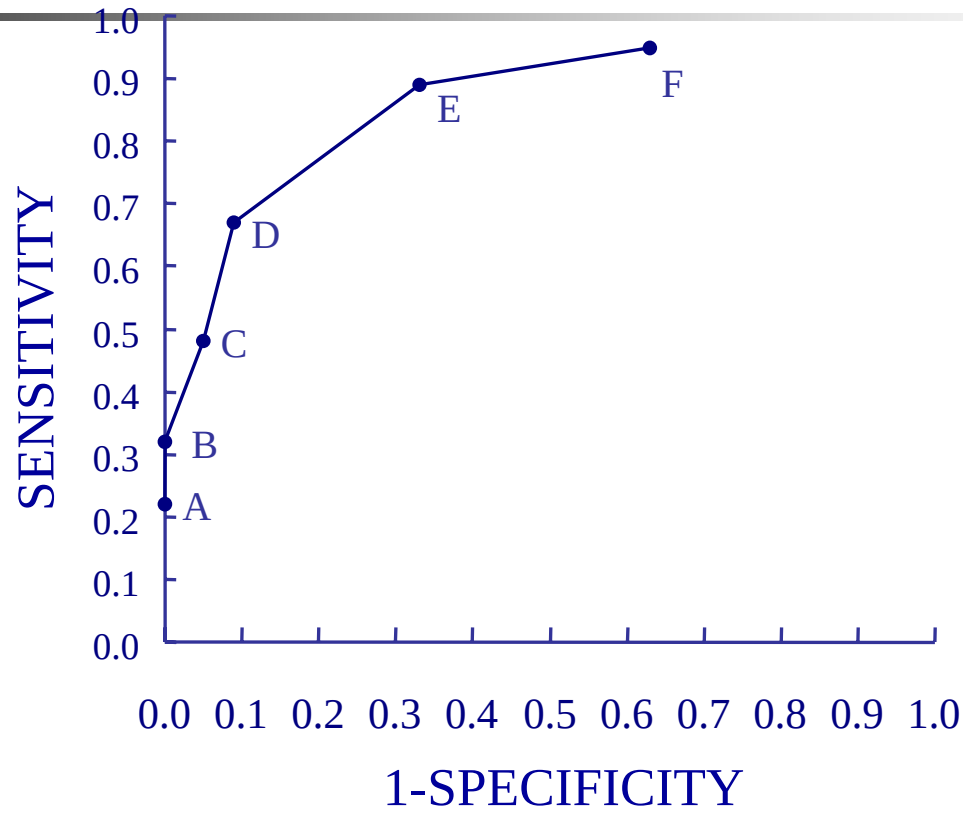
Indexes of Diagnostic Accuracy

- 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	—	—

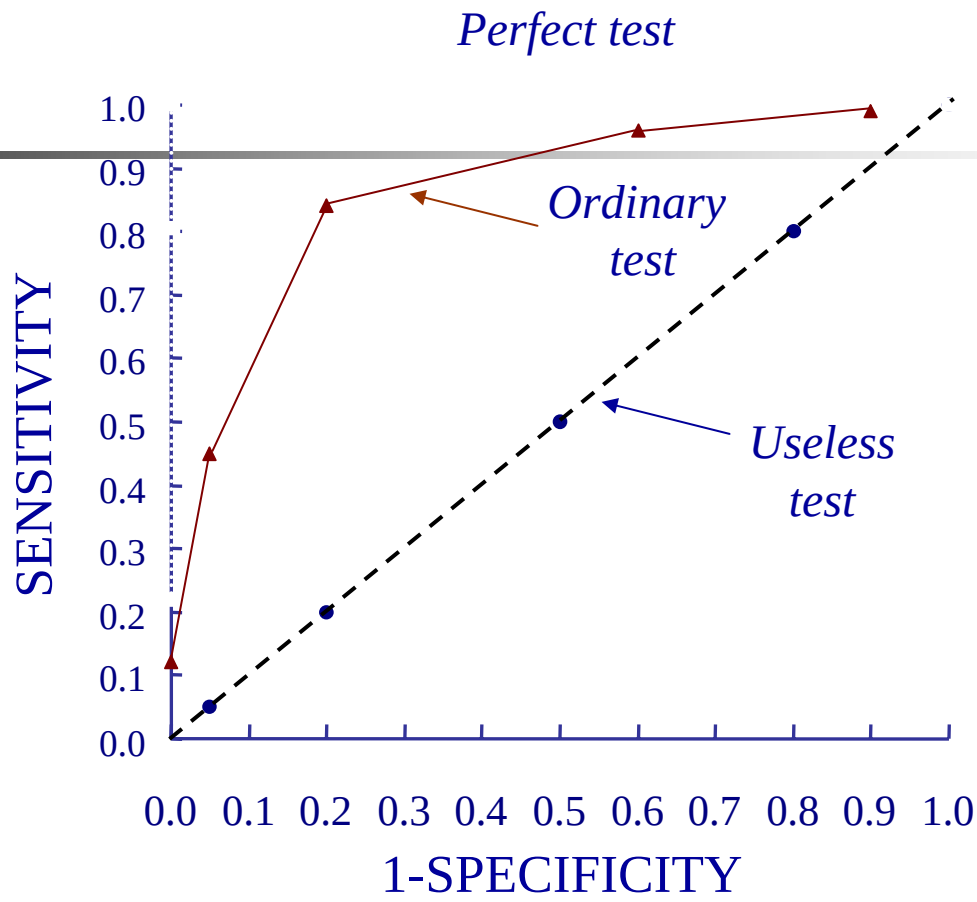
Copyright by Jen-pei Liu, PhD



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)

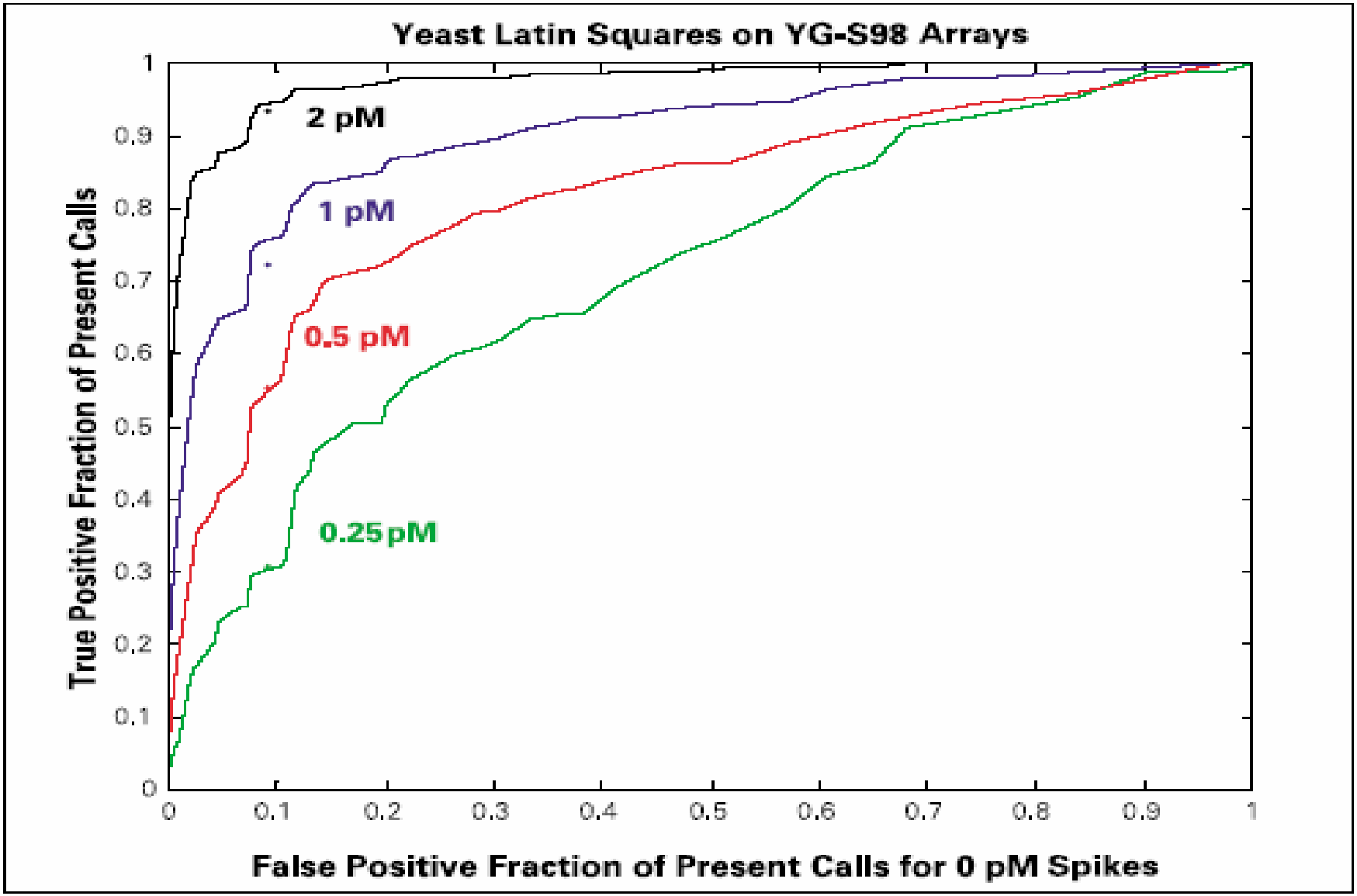


Figure 1. ***Detection Calls in Yeast Spikes: balancing sensitivity and specificity.***

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

Copyright by Jen-pei Liu, PhD

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

Copyright by Jen-pei Liu, PhD

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)

Copyright by Jen-pei Liu, PhD

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

Copyright by Jen-pei Liu, PhD

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%

Copyright by Jen-pei Liu, PhD

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*)

Copyright by Jen-pei Liu, PhD

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+

Copyright by Jen-pei Liu, PhD

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+

Copyright by Jen-pei Liu, PhD

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

Copyright by Jen-pei Liu, PhD

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+

Copyright by Jen-pei Liu, PhD

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% $p < 0.0001$

Copyright by Jen-pei Liu, PhD

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%

Copyright by Jen-pei Liu, PhD

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

Copyright by Jen-pei Liu, PhD

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

Copyright by Jen-pei Liu, PhD

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

Copyright by Jen-pei Liu, PhD

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

Copyright by Jen-pei Liu, PhD

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)

Copyright by Jen-pei Liu, PhD

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

Copyright by Jen-pei Liu, PhD

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

Copyright by Jen-pei Liu, PhD

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

Copyright by Jen-pei Liu, PhD

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

Copyright by Jen-pei Liu, PhD

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

Copyright by Jen-pei Liu, PhD

Figure 1. Difference Graph for Lab 615

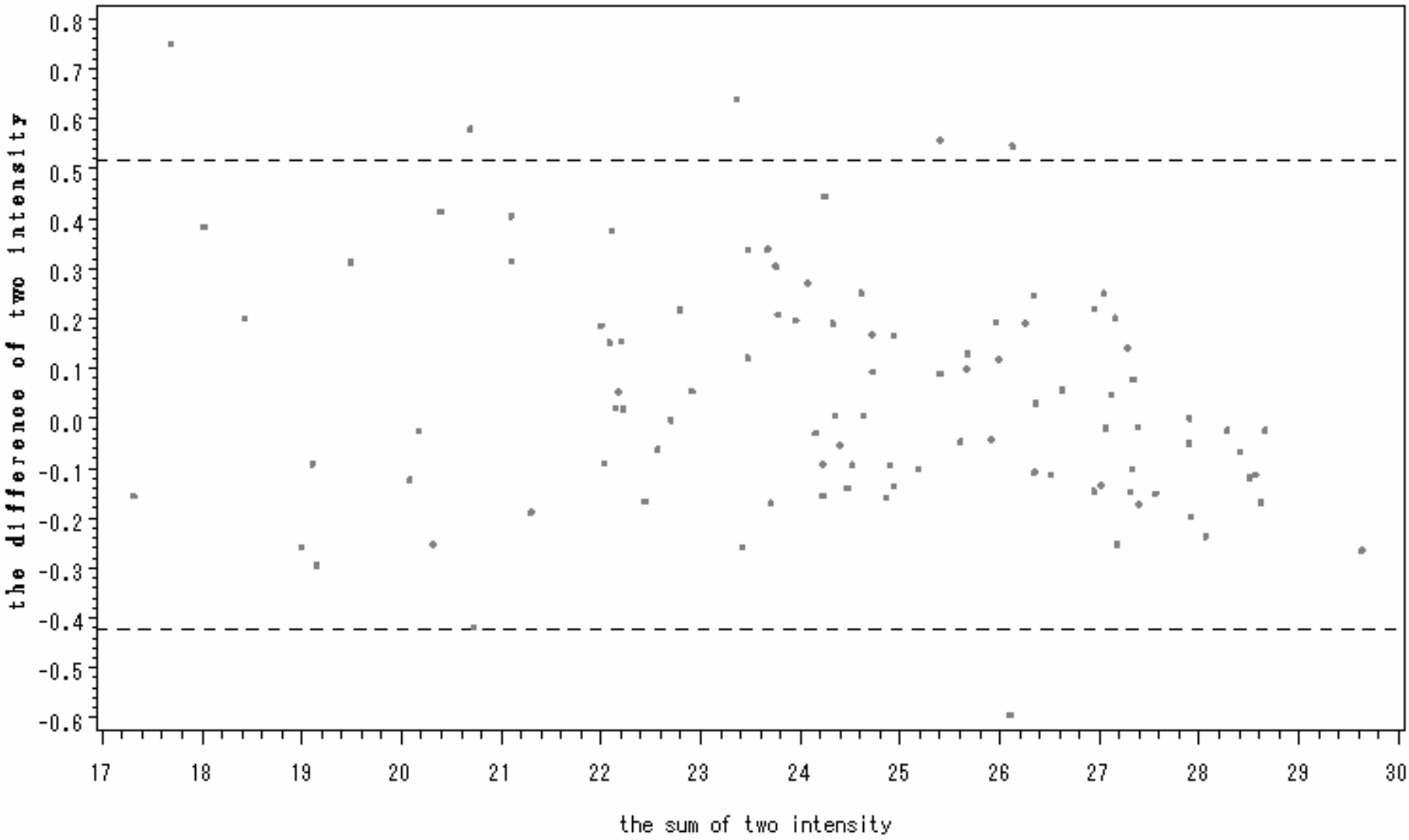
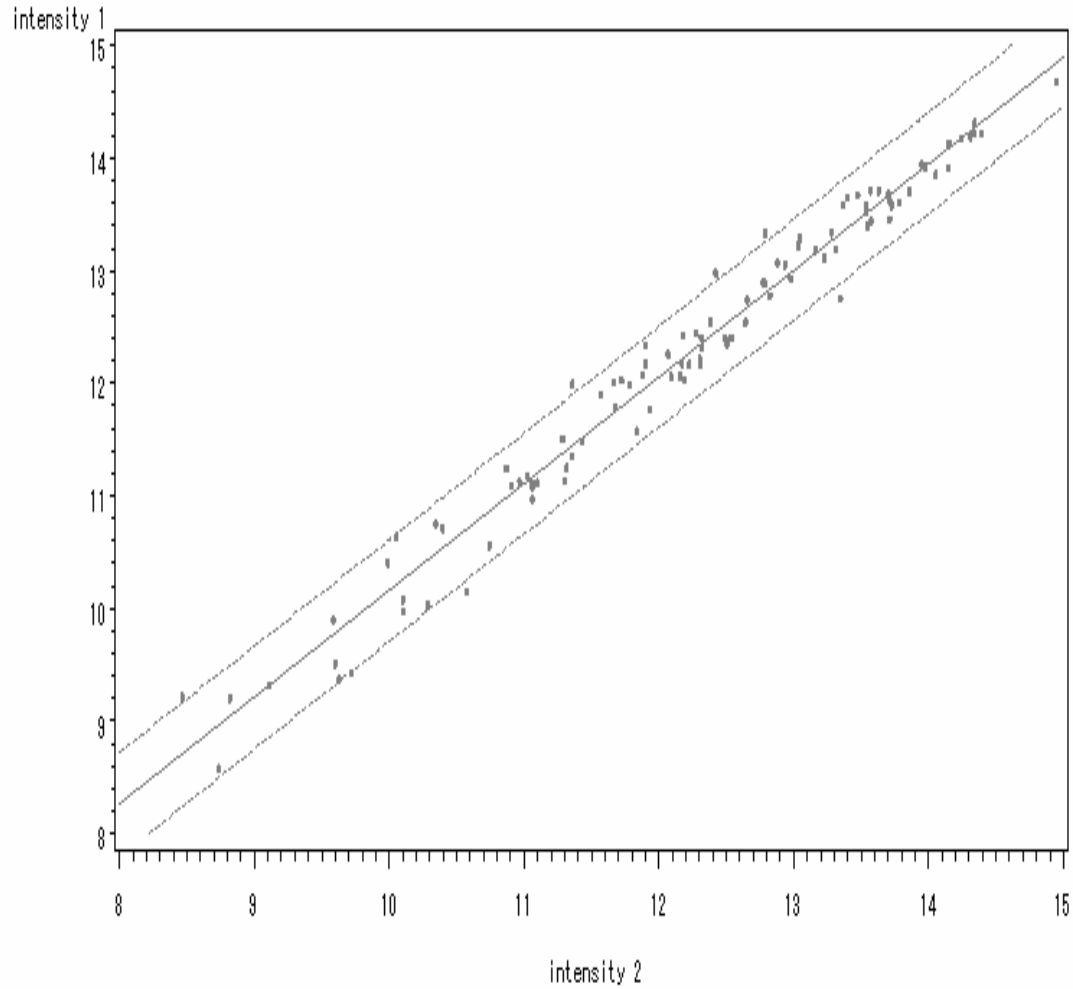


Figure 5. Scatter Plot of Replicate 1 vs. Replicate 2 for Lab 615



Regression Equation:
log2A1 = 0.654606 + 0.950092*log2A2

Five-Year View

- Calibration methods to systematically correct ratio under-estimation
- Minimization of technical variation to reproducibly detect subtler changes (e.g. 1.4-fold)
- High-throughput tools for a large number of samples (instead of large number of genes) to identify a small set of biomarkers for drug discovery and development and patient screening

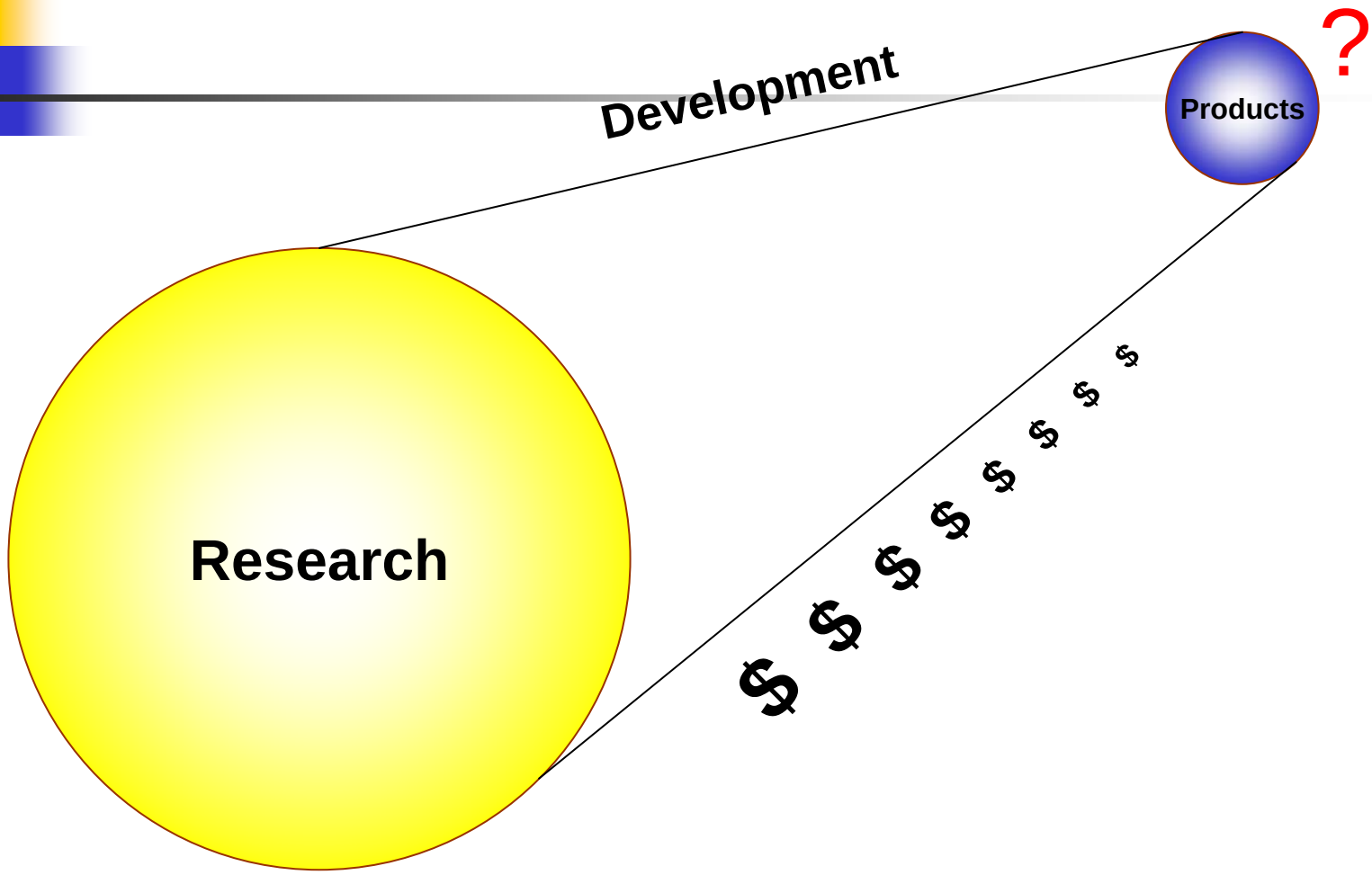
Copyright by Jen-pei Liu, PhD

Five-Year View

- A technology platform for assay a small to medium number (dozens to hundreds) of established biomarker genes in a high-throughput fashion in terms of samples is needed for diagnostic purposes (Genomic Composite Biomarker Classifier)
- A balance between high-density microarrays and QT RT-PCR

Copyright by Jen-pei Liu, PhD

Realty or Dreams



Copyright by Jen-pei Liu, PhD

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.

McNemar配對樣品檢定法

適合度檢定，獨立性檢定與同質性檢定中，每一個觀測值均來自不同的個體，觀測值是互相獨立。

選民在看電視辯論兩次投票行為結果皆來自同一位選民，故此觀測值是具有相關而種配對二項隨機變數。

例：立委候選人電視辯論對選民投票之影響

	辯論後		
辯論前	欲投甲	欲投乙	和
欲投甲	491	9	500
欲投乙	1	499	500
和	492	508	1000

Copyright by Jen-pei Liu, PhD

$$\begin{cases} H_0 : \text{電視辯論無影響} \\ H_a : \text{電視辯論有影響} \end{cases}$$

$$\begin{cases} H_0 : P_{1.} = P_{.1} \\ H_a : P_{1.} \neq P_{.1} \end{cases}$$

$$P_{1.} = P_{11} + P_{12}$$

$$P_{.1} = P_{11} + P_{21}$$

$$\begin{cases} H_0 : P_{12} = P_{21} \\ H_a : P_{12} \neq P_{21} \end{cases}$$

	辯論後		
辯論前	甲	乙	和
甲	n_{11} (p_{11})	n_{12} (p_{12})	$n_{1.}$ ($p_{1.}$)
乙	n_{21} (p_{21})	n_{22} (p_{22})	$n_{2.}$ ($p_{2.}$)
和	$n_{.1}$ ($p_{.1}$)	$n_{.2}$ ($p_{.2}$)	n (p)

$$P_{ij} = n_{ij}/n, \quad P_{i.} = n_{i.}/n, \quad P_{.j} = n_{.j}/n$$

- 僅需考慮n_{12}及n_{21}
- 令 $S = n_{12}+n_{21}$
- 若H_0為真n_{12}為二項 $Bin(n_1,1/2)$
- 其期望數為：$(n_{12} + n_{21})/2$

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{[n_{11} - (n_{12} + n_{21})/2]^2}{(n_{12} + n_{21})/2} + \frac{[n_{21} - (n_{12} + n_{21})/2]^2}{(n_{12} + n_{21})/2}$$

$$= \frac{(n_{11} - n_{21})^2}{n_{12} + n_{21}}$$

決策分法　$X^2 > \chi^2_{\alpha,1}$，拒絕H_0

電視辯論

$$X^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

$$= \frac{(9-1)^2}{9+1} = \frac{8^2}{10}$$

$$= 6.4 \quad > \quad \chi^2_{0.05,1} = 3.84$$

拒絕Ho，電視辯論對選民投票行為有影響

HercepTest vs. Pathway

$$X^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

$$= \frac{(17 - 17)^2}{17 + 17} = \frac{0^2}{34}$$

$$= 0 \quad < \quad \chi^2_{0.05,1} = 3.84$$

Reject Ho，the proportion of positive samples by Pathway is not different from that of HercepTest

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)

$p_0 = 0.6267 + 0.2978 = 0.9245$

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

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

Copyright by Jen-pei Liu, PhD