



**Society for Clinical Trials 33<sup>rd</sup> Annual Meeting**

**Workshop P5**

**Biomarkers in Clinical Trials:**

**General Principles for Study Design and Statistical Evaluation  
with Case Studies**

**Sunday, May 20, 2012**

**8:00 AM - 12:00 Noon**

**Brickell South**

## Workshop #5

# Biomarkers in Clinical Trials: General Principles for Study Design and Statistical Evaluation with Case Studies

*Presented at Society for Clinical Trials,  
Miami, Florida, USA, May 20, 2012*

## **Workshop Organizers:**

**Li Chen (Amgen)**

**Chris Coffey (University of Iowa)**

## Session# (Faculty)

Session 1: Overview of biomarkers in drug development (Sue-Jane Wang)

Session 2: Overview of surrogate endpoint evaluation in clinical studies (Geert Molenberghs)

Session 3: Overview of biomarkers in device development (Gene Pennello)

Session 4: Biomarker trial designs: lessons from real trials (Sumithra Mandrekar)

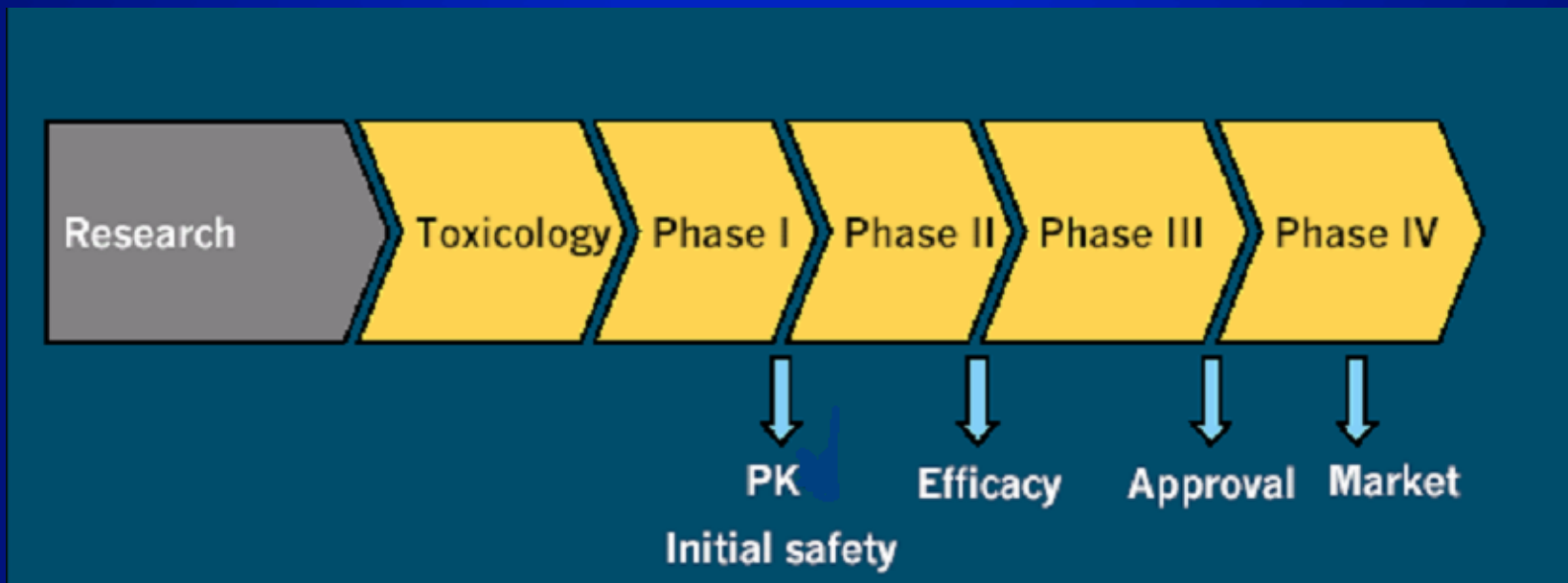
# Disclaimer

**The views expressed in this presentation  
are not necessarily of the US FDA**

# Session #1: OUTLINE

- ◆ Overview of Drug Development Process
- ◆ Movement in Pharmacotherapy
- ◆ What is Biomarker and Overview from Existing Literature
- ◆ Biomarker Translational Research – Drug Development Tools
- ◆ Design and Analysis of Pharmacogenomics Clinical Trials for Biomarker-Drug Co-development
- ◆ Summary

# Traditional paradigm Drug Research & Development



# Omic Science Evolved

- Genomics/genetics (DNA)
- Genomics (RNA, iRNA)
- Proteomics (protein)
- Methylation
- Metabolomics (systematic study of the unique chemical fingerprints that specific cellular processes leave behind; scientific study of chemical processes involving metabolites)
- Metabonomics (metabolism) (The study of metabolic responses to drugs, environmental changes and diseases)
- Next generation sequencing (wide applications, curious enthusiasm)
- Bioinformatics (computational biology vs clinical disease/therapeutics)



# Drug Discovery : New paradigm

Discovery



Powerful discovery and screening technologies



New  
Chemical  
Entity



- Combinatorial Chemistry
- Mass spectrometry
- High Throughput Screening
- Cell- and tissue- based DNA microarrays
- Proteomic technologies
- Metabonomics
- Next generation sequencing

SCT 2012 - Biomarker Short Course

8

# Drug Development: New Paradigm

Preclinical

Clinical Development

New  
Chemical  
Entity

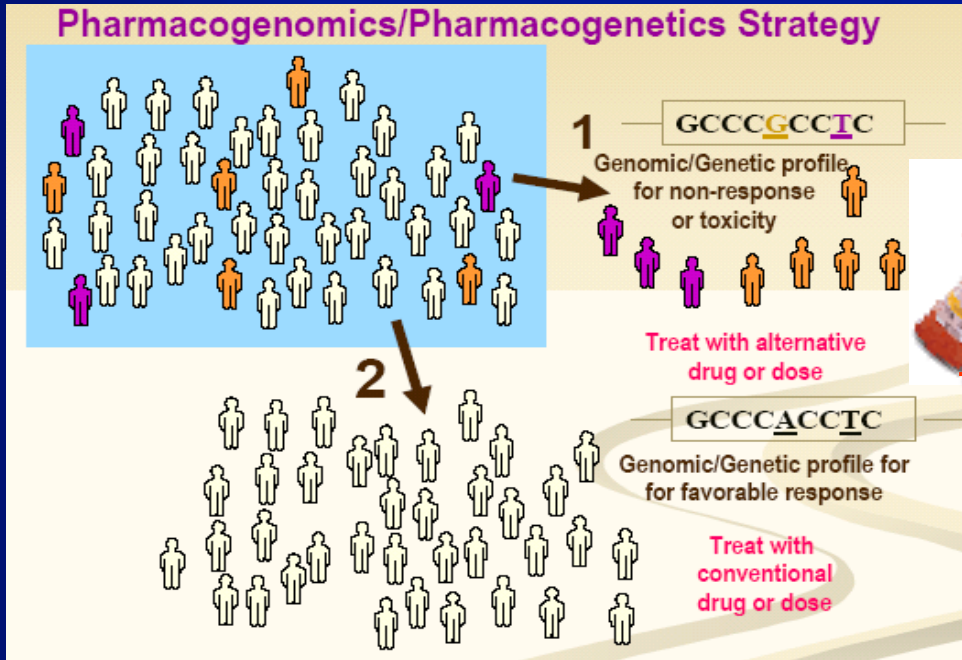


Preclinical

- Biomarkers
- Surrogate Endpoints
- precise clinical measurement ,e.g., survival, SNP
- etc.

- Outcome Endpoint
- Surrogate Endpoint
- Genotyping
- Phenotyping
- Possible Enrichment
- etc.

# Path to Individualization ?



$$Y = f(x)$$

Non-responder

Optimal response

Toxicity

Why

> Omics factors x?  
Environmental factors?

# Movements in Pharmacotherapy vs. Clinical Trial



# Definition

- **Biomarker:**

- A characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention (Biomarkers Definitions Working Group ,2001)

- ◆ A characteristic recognized as an indicator
- ◆ Disease indicator
- ◆ Therapeutic indicator
- ◆ Potential Confounder

Also cited in FDA/CDER draft guidance: qualification process for drug development tools (2010)

# Biomarker

- ◆ The biomarker may reflect biological processes closely related to the mechanism of disease or processes substantially downstream from the primary disease processes.
- ◆ Biomarkers may assess many different types of biological characteristics or parameters, including genetic composition, receptor expression patterns, radiographic or other imaging-based measurements, blood composition measurements (e.g., serum enzyme levels, prostate specific antigen), electrocardiographic parameters, or organ function (e.g., creatinine clearance, cardiac ejection fraction)

Draft guidance

SCT 2012 - Biomarker Short Course

13

# Surrogate Endpoint

A biomarker that is

intended to substitute for a clinical endpoint;

A surrogate endpoint is expected to

predict clinical benefit (or harm) based on epidemiological, therapeutic, pathophysiologic, or other scientific evidence

**Surrogate endpoints are a small subset of biomarkers**

Biomarkers Definitions Working Group, 2001

# Imaging (Surrogate) Biomarker

- Measurement of an imaging biomarker may require administration of an imaging agent drug
- Development path depends on whether the imaging agent has been approved for marketing



# Genomic Biomarker

- ◆ A measurable DNA/RNA characteristic that is objectively measured and evaluated
- ◆ Recognized as an indicator of
  - ◆ Normal biological processes
  - ◆ Pathogenic processes
  - ◆ Pharmacologic response to a therapeutic intervention
- ◆ When does it have regulatory impact ?

# Newer Genomic Biomarker Research

## DNA methylation

An epigenetic modification that changes the appearance and structure of **DNA** without altering its sequence

A potential stratification factor

# Pharmacogenomics/Pharmacogenetics

- ◆ The science of determining how the benefits and adverse effects of a drug vary among a target population of patients based on genomic features of the patient's germ line and diseased tissue
  - ◆ Simon, Wang (The Pharmacogenomics Journal, 2006)
  - ◆ Trepicchio, Essayan, Hall, Schechter, Tezak, Wang, Weinreich, Simon (The Pharmacogenomics Journal, 2006)
- ◆ The study of variations of DNA and RNA characteristics as related to drug response (draft ICH E15)

# Pharmacogenomics Clinical Studies

- ◆ Exploratory – development of a genomic biomarker
- ◆ Confirmatory – regulatory impact; labeling implication; diagnostic test for patient selection

# Genomic Biomarker Classifier or Signature

- ◆ Could be measurements of gene expression, gene function, or gene regulation
- ◆ Can consist of one or more DNA and/or RNA characteristics
- ◆ Not limited to human samples, but includes samples from viruses and infectious agents as well as animal samples

ICH E15

SCT 2012 - Biomarker Short Course

20

# Diagnostic (Multiplex) Assay

For genomic (composite) biomarker to be used along with therapeutics in medical practice, regulatory approval or clearance of genomic diagnostics depends on the class category of the diagnostics (risk based)

## Statistical issues

- ◆ Assay characterization and analytical validation
- ◆ Clinical validation

# Uses of Biomarker in Drug Development

- ◆ Tool for assessing disease as diagnostic/screening
  - ◆ disease presence
  - ◆ disease heterogeneity/subtypes
  - ◆ prognosis
- ◆ Tool for assessing drug target
  - ◆ target validation
  - ◆ target/compound interactions
- ◆ Tool for assessing therapeutics use vs predicting outcome
  - ◆ pharmacokinetics, pharmacodynamics
  - ◆ clinical (intermediate vs ultimate) endpoint as response to therapy
  - ◆ patient selection

# Decision Making vs Regulatory Approval

Internal  
Consideration

- ▶ Preclinical: target validation, interaction with targets, toxicity potential, efficacy, heterogeneity in response
- ▶ Early clinical development: pharmacokinetics, pharmacodynamics, dose selection, POC, safety signals, explore patient subsets

Regulatory  
Impacts

- ▶ Late clinical development: patient (sub)population, surrogate endpoint, biomarker qualification (Drug), companion diagnostics for drug use (Diagnostics)



# **Biomarkers Translational Research for Drug or Biologics Development**

## **Drug Development Tools**

### **Some Nomenclature of Biomarkers**

# Drug Development Tools (DDT)

- ◆ DDTs are methods, materials or measures that aid drug development
- ◆ DDTs includes **biomarker**, clinical outcome assessment, and, animal models, etc.
- ◆ Biomarker Qualification (BM): optional

Draft guidance

SCT 2012 - Biomarker Short Course

25

# Prognostic Biomarker

- ◆ A baseline characteristic that categorizes patients by degree of risk for disease occurrence or progression of a specific aspect of a disease
- ◆ Informs about the natural history of the disorder in the absence of a therapeutic intervention
- ◆ Can be used as an enrichment strategy to select patients likely to have clinical events of interest or to progress rapidly
- ◆ Biomarker-outcome relationship can change after treatment intervention

# Predictive Biomarker

- ◆ A baseline characteristic that categorizes patients by their likelihood for [of] response to a particular treatment [relative to no treatment]
- ◆ Used to identify whether a given patient is likely to respond to a treatment intervention in a particular way
- ◆ May predict a favorable response or an unfavorable response (i.e., adverse event)

FDA/CDER draft guidance: qualification process for drug development tools (2010)

# Definition of Treatment Effect

Genomic Status*	Scenario A		Scenario B		Scenario C	
	Control	Drug A	Control	Drug B	Control	Drug C
$g^-$	33%	33%	36%	46%	39%	49%
$g^+$	33%	48%	50%	60%	48%	68%

\*  $g^+$  or  $g^-$  is patient's genomic status determined from a diagnostic assay

## Predictive

Effect in  $g^+$  only  
No effect in  $g^-$

Qualitative

## Prognostic

Effect in  $g^+$  and  $g^-$  is consistent, i.e., biomarker plays a role in disease response only

## Prognostic-Predictive

Effect is larger in  $g^+$  than in  $g^-$

Quantitative

Wang et al. (2007, PS)

# Biomarker-Based Predictive Enrichment $\hat{=}$ Predictive Biomarker

- ◆ Enhanced B/R if there is toxicity (e.g., Herceptin)
- ◆ Trastuzumab (Herceptin) is cardiotoxic:

Studies in patients with metastatic cancer as well as adjuvant studies were conducted in patients with HER2/Neu positive tumors, enhancing B/R

HER2/Neu negative patients have much less response and the cardiotoxicity is unacceptable

# Pharmacodynamic (Activity) Biomarker

- ◆ A change in the biomarker shows that a biological response has occurred due to therapeutic intervention
- ◆ The magnitude of change is considered pertinent to response
- ◆ May be treatment-specific or informative of disease response
- ◆ Examples: blood pressure, cholesterol, HbA1C
- ◆ Most PD biomarkers are used to guide drug development and not as a basis for regulatory approval

# Efficacy-Response Biomarker

- ◆ Efficacy-surrogate biomarker, Surrogate endpoint
- ◆ Subset of general pharmacodynamic biomarkers
- ◆ Maybe used as basis of NDA/BLA approval decisions
- ◆ Predicts a specific clinical outcome of the patient at some later time
- ◆ May be Treatment specific



# Biomarker Uses Relate to Characteristics

- ◆ Biomarkers can have utility in more than one category
  - ◆ Depends on the specific characteristics of the particular biomarker, e.g., safety assessment to warn of toxicity vs monitor for the desired effect based on serum lipid level
- ◆ Biomarker is applied differently for utilizing the different characteristics
- ◆ For some situations, interpretation of a biomarker implies which term is being applied, e.g., disease stage

# Potential Concerns with Pharmacodynamic Biomarker

- ◆ Can mislead future development if discordant with clinical outcome
  - ◆ Falsely suggest presence or absence of benefit
  - ◆ False optimization of dose / regimen / population
  - ◆ Assumed relationship is incorrect or suboptimal
  - ◆ Inaccurate estimate of effect size or frequency of benefit
- ◆ Potential causes
  - ◆ Alternate mechanisms of action
  - ◆ Unrepresentative model of Biomarker-Clinical relationship

# Understand Surrogate Measure & Its Complexity



When the Surrogate Endpoint is in the Causal Pathway of Disease Process, e.g., a pathophysiologic process

Multiple causal pathways, e.g., multiple pathophysiologic processes may interact among themselves

# Imaging Biomarker as Predictor ?

- ➔ Predict the presence of some condition  
before vs. after drug/biologics administration
- ➔ Predict the occurrence of some future event  
before vs. after drug/biologics administration
- ➔ Value added or improvement of diagnostic ability:  
established for some aspect
  - ◆ Drug/Biologics development
  - ◆ Clinical safety
  - ◆ Clinical efficacy
- ➔ Established from supporting documentation, medical literature and clinical trials development

# Regulatory Acceptance of Biomarker Prior to DDT Qualification

- ◆ On a case by case basis
  - ◆ Within a specific IND/NDA/BLA/Labeling Update
  - ◆ For a specific drug
  - ◆ Driven by a specific drug developer's needs
- ◆ More general use accepted over extended period
  - ◆ Scientific experience accumulates through varied uses
  - ◆ Usually very extended time-frame
  - ◆ Scientific evidence collection may not be cohesively directed

# More Recent Regulatory Acceptance of Biomarker Development Paths

- ◆ Existing routes remain available
- ◆ Co-development of drug and diagnostic test
  - ◆ Companion diagnostics
  - ◆ Policy Guidance – July 2011
    - ◆ Others in development
- ◆ Biomarker Qualification Process (2010 draft)
  - ◆ Developing program within CDER
  - ◆ Outgrowth of Critical Path Initiative

# Biomarkers Qualification

- ◆ A conclusion that within a carefully and specifically stated “context of use” the biomarker has been demonstrated to reliably support a specified manner of interpretation and application in drug development
  - ◆ Utility in drug development, particularly regulatory decisions, is central to purpose of qualification
  - ◆ Particularly for biomarkers expected to have application in multiple different drug development programs
- ◆ Validation used in IOM report
- ◆ Context of Use (regulatory consideration)

# What Becomes Qualified?

- ◆ Biomarker is a measurement of a substance, analyte, anatomic image, or other describable characteristic
  - ◆ Assay methods are needed to measure the biomarker
  - ◆ Assay method is not the biomarker
- ◆ One biomarker can have multiple assays that are capable of measuring the biomarker
  - ◆ Assay method performance characteristics are important
- ◆ CDRH clears or approves commercial testing devices for clinical measurements
- ◆ **CDRH clearance does not equal CDER qualification**
  - ◆ Different purposes



# Qualification's Place in Therapeutic Development

- ◆ Qualification is not required
  - ◆ Case by case approach for accepting use in a single IND/  
NDA/BLA program remains valuable
- ◆ Qualification is voluntary
  - ◆ Holder of biomarker data can choose to pursue or not  
pursue qualification
- ◆ Qualification is intended for biomarkers that will be used in  
multiple drug development programs
  - ◆ Public knowledge and availability essential
  - ◆ Consortia or collaborative groups likely to be source of  
biomarkers for qualification

# Single Biomarker Development

## Early Stage vs Later Stage

Her2+/Neu (Herceptin)

EGFR (Tarceva, Iressa)

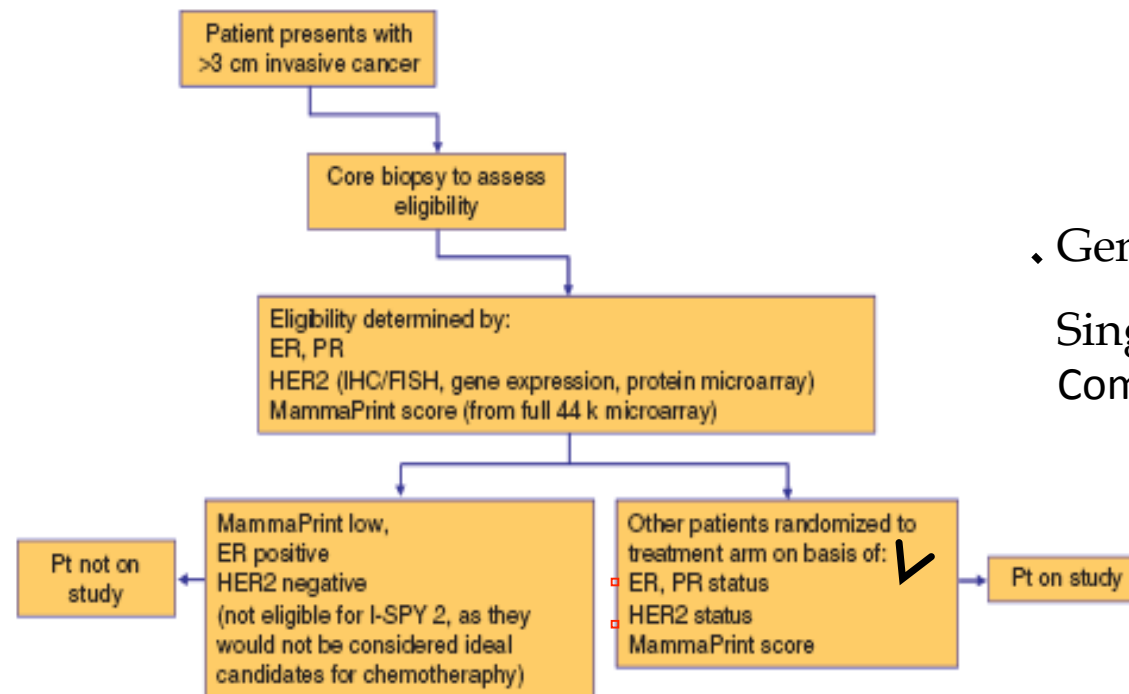
CYP2D6 variants (Strattera)

TNF- $\alpha$  238, HLA-B57 (Abacavir)

ALK+ (crizotinib)

BRAF 600E mutation (vemurafenib)

# I SPY 2 Study Design incorporate MP



- Genomic biomarker
- Single biomarker
- Composite biomarker

**Figure 1** I-SPY 2 eligibility and treatment assignment. ER, estrogen receptor; FISH, fluorescence *in situ* hybridization; HER2, human epidermal growth factor receptor 2; I-SPY 2, investigation of serial studies to predict your therapeutic response with imaging and molecular analysis 2; IHC, immunohistochemistry; PR, progesterone receptor; Pt, patient. For MammaPrint scoring, see refs. 11,12.

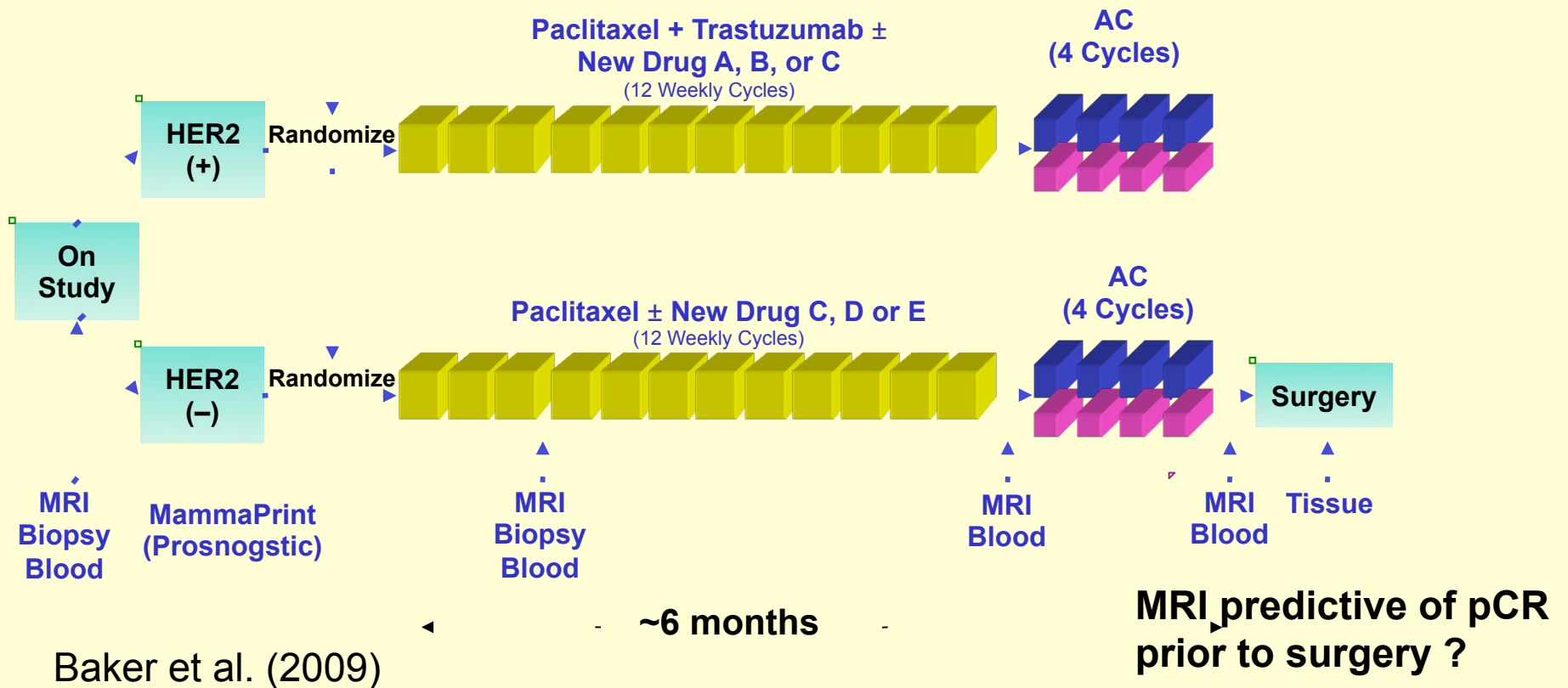
Baker et al. (2009)

SCT 2012 - Biomarker Short Course

42

# I-SPY 2 Adaptive Trial (Imaging, Genomic Single and Composite Biomarker)

ADAPT



SCT 2012 - Biomarker Short Course

43

# Performance of a diagnostic test

## Diagnostic Test Result

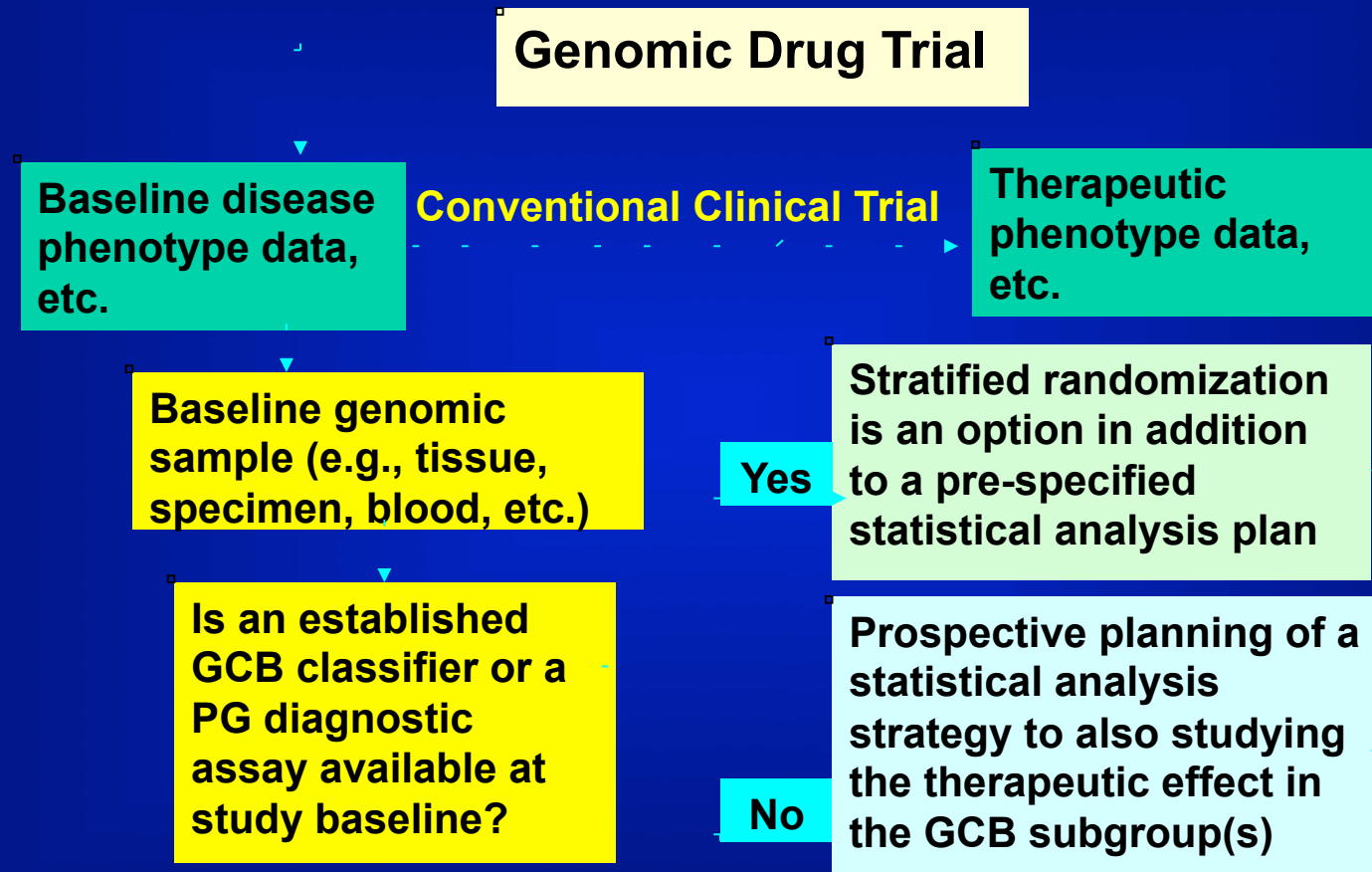
		Diagnostic Test Result	
		Negative	Positive
True State	No Event	Specificity Pr (negative   no event)	Pr (Incorrect positive diagnosis)
	Event	Pr (Incorrect negative diagnosis)	Sensitivity Pr (positive   event)

Accuracy = Pr (Correct Decision)

# Pharmacogenomics Studies

## How Convenience Genomic Samples Can Affect Interpretation of Study Finding

# Use of GCB profiling in Genomics Drug Trials



Wang SJ (2008, [Journal of the Formosan Medical Association](#))

SCT 2012 - Biomarker Short Course

46

# Imbalance Observed Within Convenience Sample

## Baseline Dissimilarity in % Males

	Placebo	Low dose	High dose
Consented sample (30%)	43%	67%	68%
Non-consented sample (70%)	69%	62%	76%

Wang, O'Neill, Hung (2010, Clinical Trials)



## Observed treatment effect - Bias as an explanation due to convenience sample (30%) Inconsistent Evidence Between Samples

Study 1	Primary Endpoint 1		Primary Endpoint 2	
	Low	High	Low	High
31% of ITT				
effect estimate	-5.0	-3.4	-7.9	-6.6
unadj. p-value	0.029	0.192	0.0141	0.135

Study 1	Primary Endpoint 1		Primary Endpoint 2	
	Low	High	Low	High
ITT				
effect estimate	-3.1	-4.5	-5.1	-8.1
unadj. p-value	0.033	0.005	0.034	0.002

Wang, O'Neill, Hung (2010, Clinical Trials)

SCT 2012 - Biomarker Short Course

48

# How Probable are Prognostic Factor Imbalances ?

- ◆ Full ITT population - factor ascertained on everyone in the RCT
  - ◆ Depends upon sample size in each treatment group within each factor (genomic + or - )
- ◆ Convenience samples - factor is only ascertained on a non-random and non-randomized subset of subjects - Depends on lack of randomization and other biases in the data

# Study Designs

- ◆ Retrospective versus Prospective
- ◆ Conventional randomized controlled
  - ◆ Stratified
  - ◆ Interaction
- ◆ Genomic biomarker guided design
- ◆ Adaptive vs Non-Adaptive Enrichment design
- ◆ One trial with separate inference between + and – subsets

# Adaptive Design Can Dramatically Improves Efficiency of Genomic Guided Design

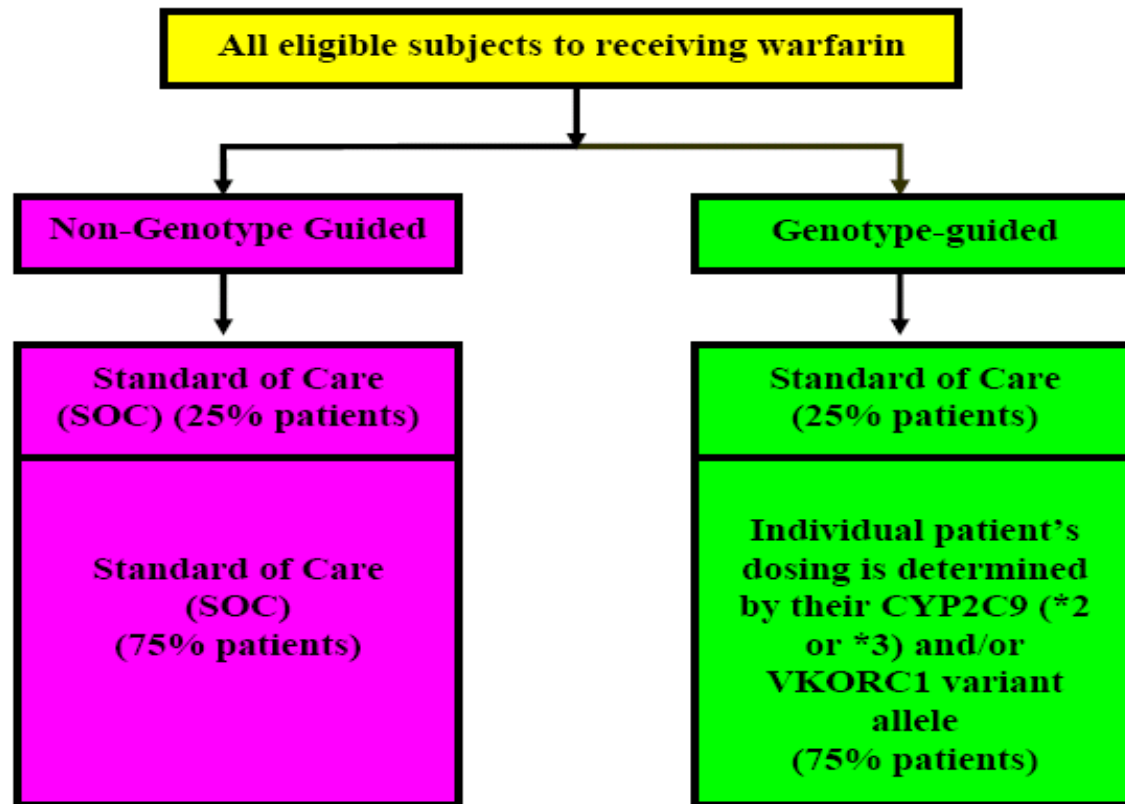
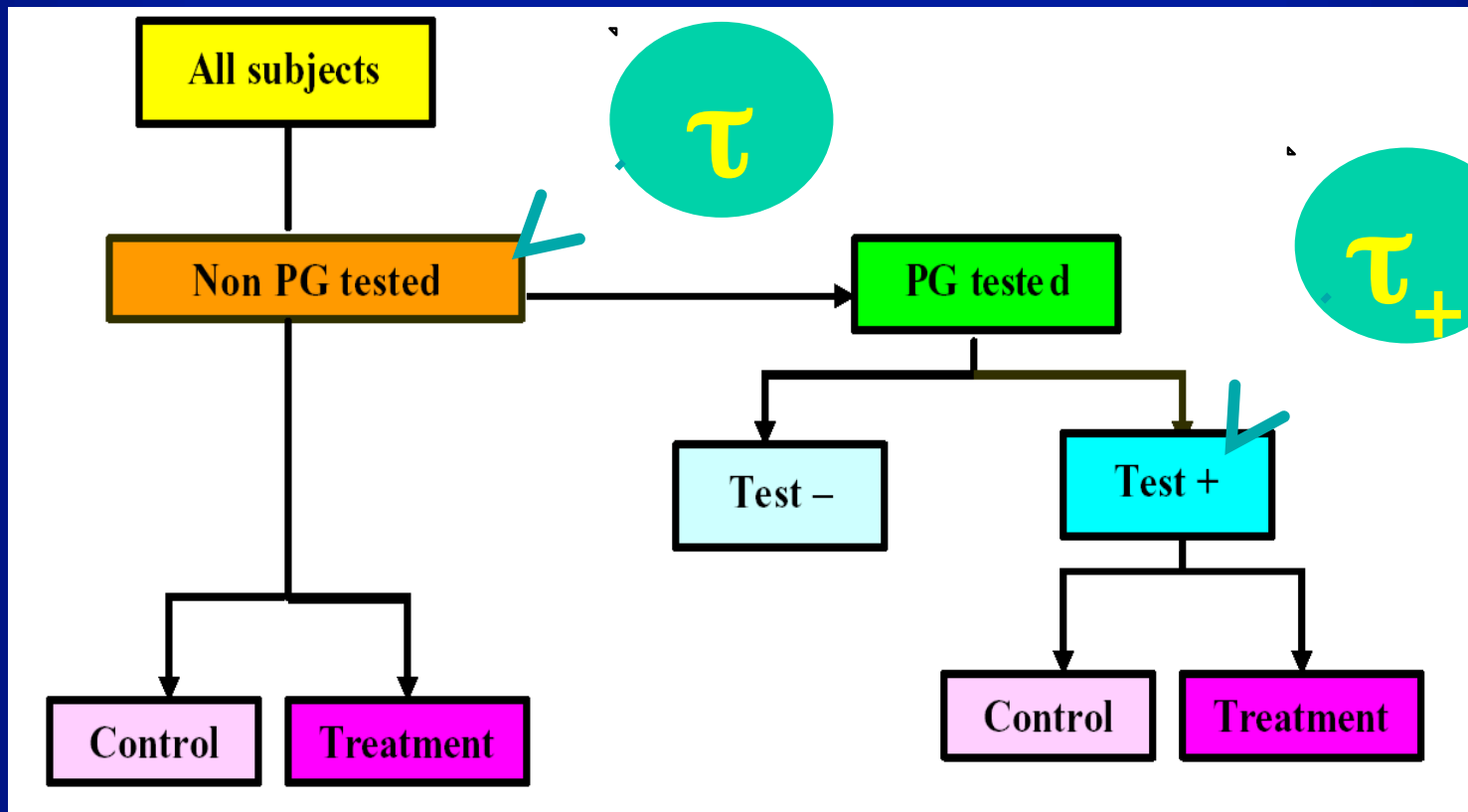


Figure. Genotype-guided (clinical + genotype) versus SOC (clinical only) Dosing Design

# Non-Adaptive with Multiplicity

- ◆ 2-arm PCT: 2-patient sets (ITT, biomarker+), 1-endpoint
- ◆ Study objectives: to evaluate treatment effect in
  - (i) ITT; (ii) subset defined by biomarker classifier (B+)
- ◆ Require showing treatment effect in ITT
- ◆ Conventional subgroup consistency problem
- ◆ Not require showing treatment effect in ITT
- ◆ Optimal balance between overall power and power in B+, e.g., search for larger effect

# Adaptive Enrichment\*



Wang et al. (2007, PS)

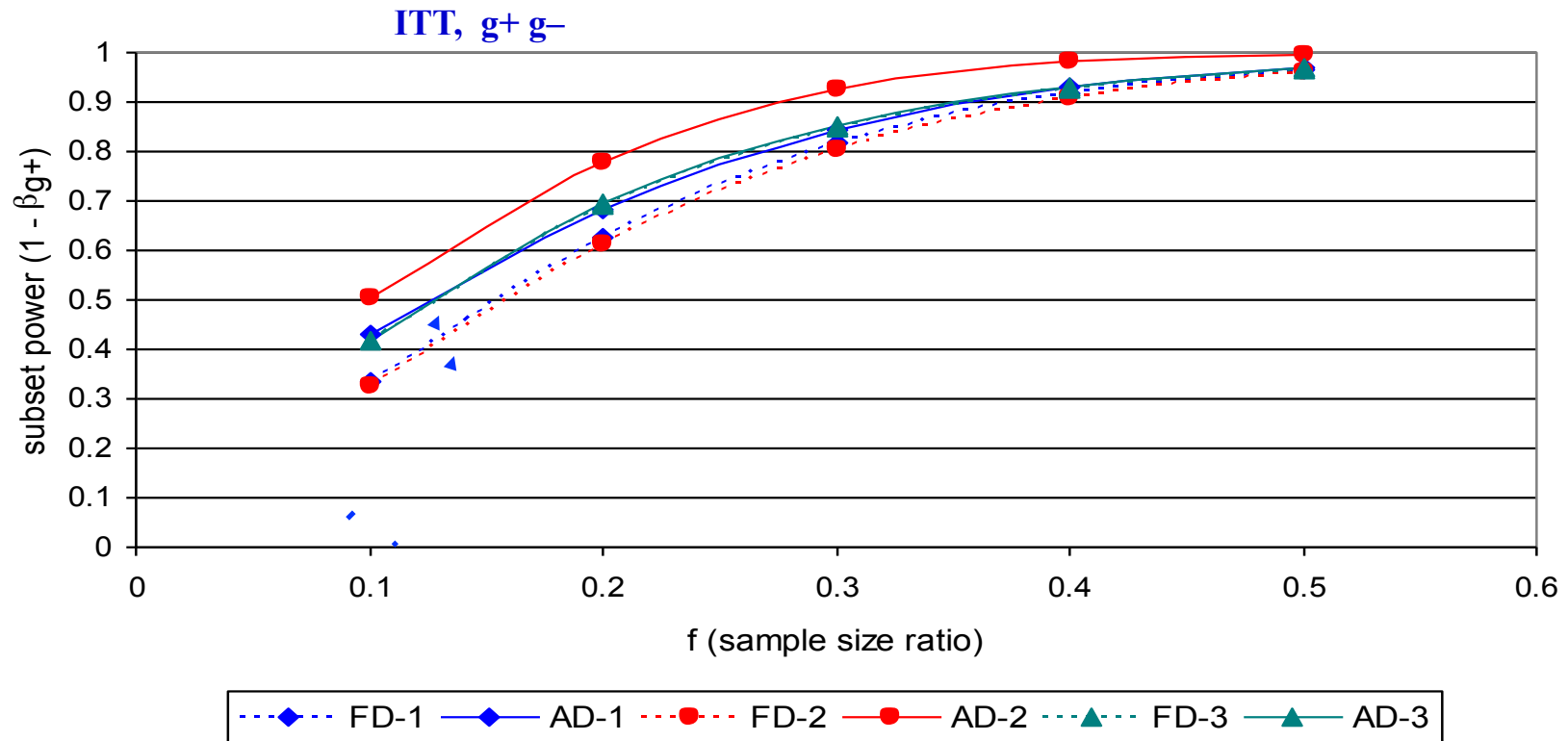
SCT 2012 - Biomarker Short Course

53

# Non-Adaptive vs. Adaptive

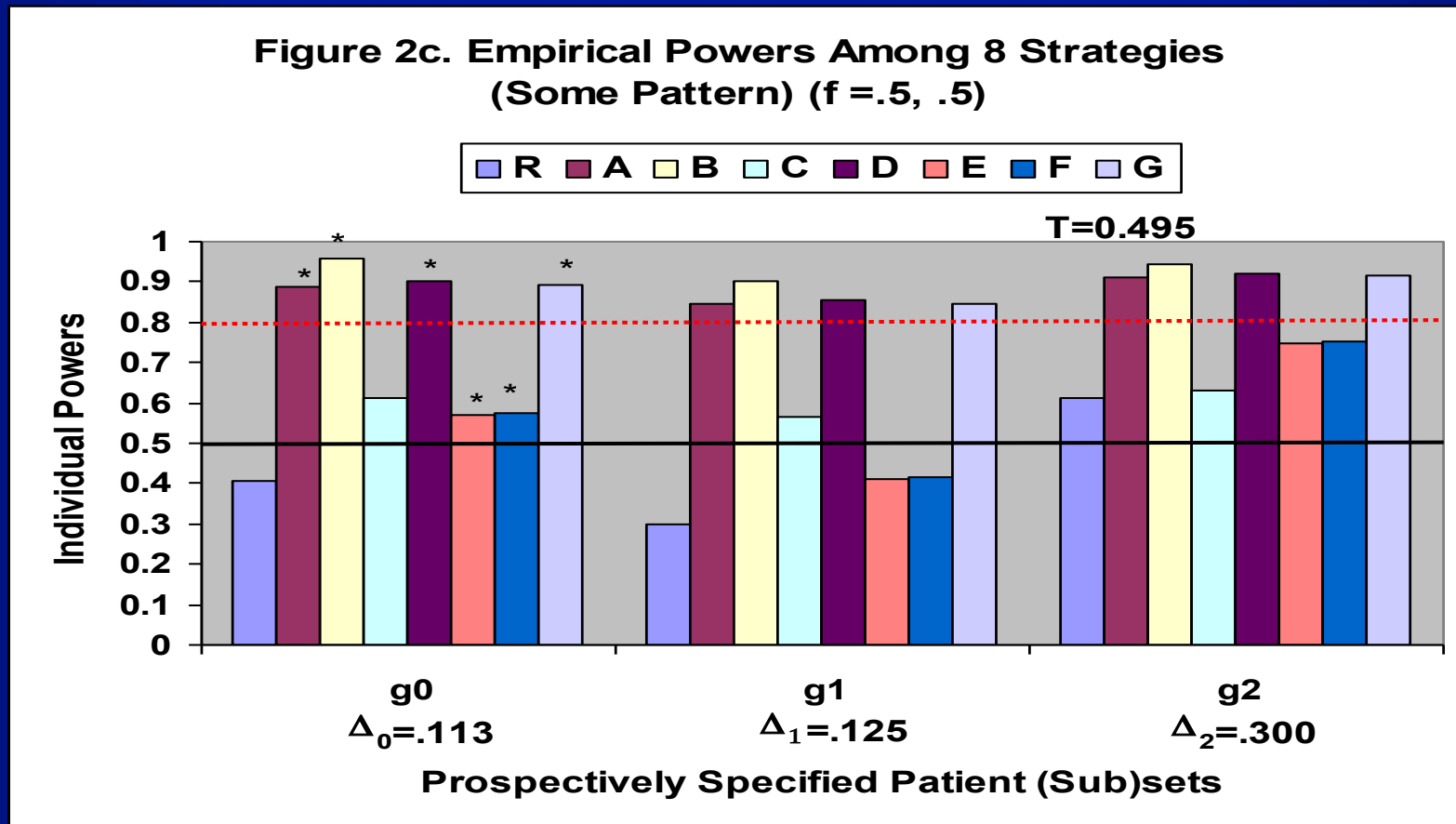
## Dashed vs. Solid

Figure 7. Power Comparison for  $\Delta_{g+}$  with Hochberg Method  
 (1=( $\Delta$ , 0.4, 0); 2=(0, 0.4,  $\Delta_{g-} < 0$ ); 3=(0.2, 0.4,  $\Delta_{g-}$ ))



# Empirical Power Comparison

## Some Nested Effect Pattern





# Design and Analysis Issues

- ◆ Impact of genomic diagnostic misclassification in pharmacogenomics clinical trials
  - ◆ Superiority
  - ◆ Non-inferiority (Wang et al., 2011, SBR)
- ◆ Bias in treatment effect estimate (Wang et al., 2010, CT)
- ◆ Strong control of studywise type I error rate with an adaptive design (Wang et al., 2009, Biometrical J)
- ◆ Replication of treatment effect (Wang et al., 2010, CT)

## Reasonable Level of Evidence that a (composite) biomarker is predictive of differential treatment effect – enrichment in Ph 3 ?

Response rate	P	T	p-value
#2 n g+ only	140 38%	140 73%	< 0.005
#1 n g+ (74%) g- (26%) ITT	131 28% 53% 31%	176 54% 47% 51%	< 0.0001 (1°) ns <0.001
#3 n g+ (79%) g- (21%) ITT	201 19% 12% 18%	298 54% 41% 51%	< 0.0001

Wang et al. (Clinical Trials 2010 to Appear)

# Summary

- ◆ Biomarker versus Pharmacogenomics
- ◆ Types of biomarker, its context of use primarily as drug development tools for qualification with regulatory bearing
- ◆ Replication of treatment effect in biomarker defined patient subset to avoid random or false positive finding
- ◆ Adaptive design can be powerful mostly when biomarker is predictive of treatment effect; requiring acceptable diagnostic performance of biomarker; interpretation problem about mixture of treatment effect if requiring overall effect shown

# Questions / Comments

# Surrogate Endpoint Evaluation in Clinical Studies

Geert Molenberghs

*SCT2012: Biomarkers in Clinical Trials, May 20, 2012*

Interuniversity Institute for Biostatistics and statistical Bioinformatics

Universiteit Hasselt, Belgium  
geert.molenberghs@uhasselt.be  
www.censtat.uhasselt.be



Katholieke Universiteit Leuven, Belgium  
geert.molenberghs@med.kuleuven.be  
www.kuleuven.ac.be/biostat/

# Motivation

---

- **Primary motivation**

- ▷ True endpoint is rare and/or distant
- ▷ Surrogate endpoint is frequent and/or close in time

- **Secondary motivation:** True endpoint is

- ▷ invasive
- ▷ uncomfortable
- ▷ costly
- ▷ confounded by secondary treatments and/or competing risks

# Definitions

---

## **Clinical Endpoint:**

A characteristic or variable that reflects how a patient feels, functions, or survives.

## **Biomarker:**

A characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention.

## **Surrogate Endpoint:**

A biomarker that is intended to substitute for a clinical endpoint. A surrogate endpoint is expected to predict clinical benefit (or harm or lack of benefit or harm).

Biomarkers Definition Working Group (Clin Pharmacol Ther 2001)

# Age-Related Macular Degeneration

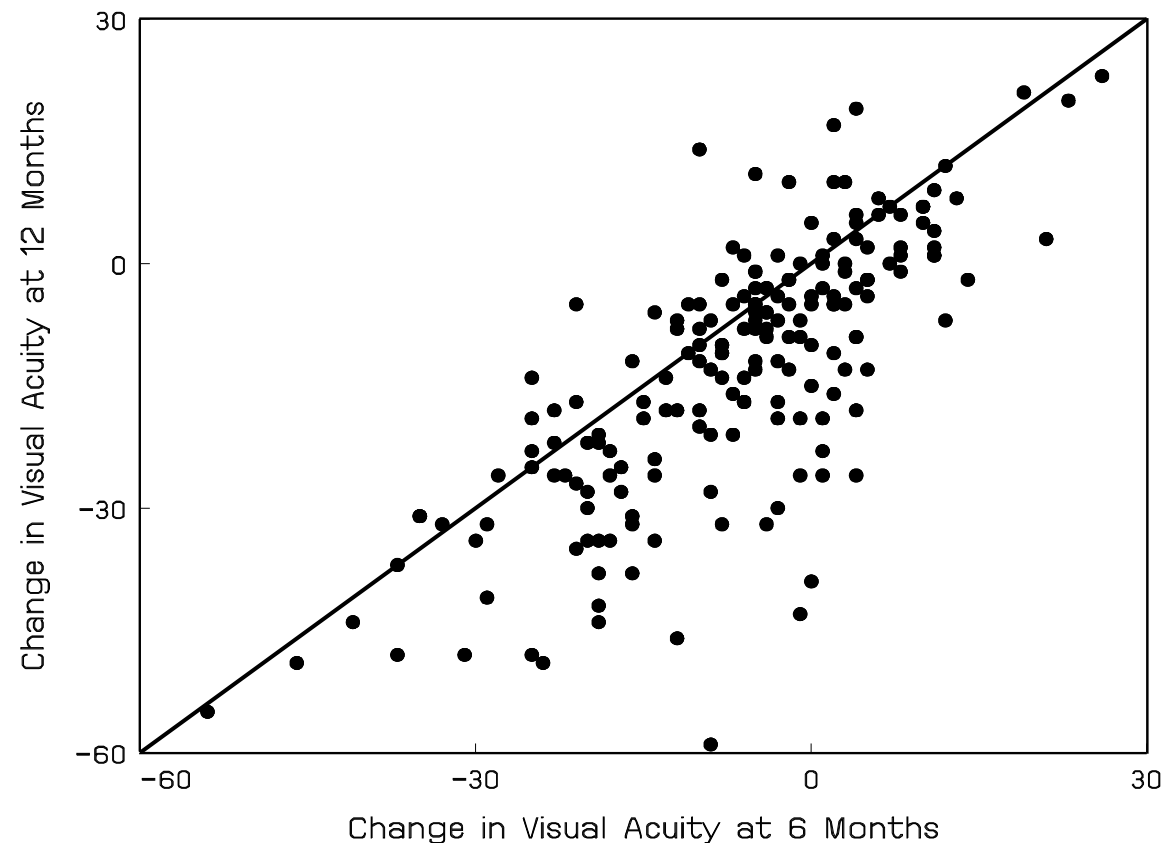
*Pharmacological Therapy for Macular Degeneration Study Group (1997)*

**Z: Interferon- $\alpha$**

**S: Visual acuity at 6 months**

**T: Visual acuity at 1 year**

**N: 190 patients in 36 centers (# patients/center  $\in [2;18]$ )**





# Definition and Single-Unit Model

---

Prentice (Bcs 1989)

“A test of  $H_0$  of no effect of treatment on surrogate is equivalent to a test of  $H_0$  of no effect of treatment on true endpoint.”

$$\begin{aligned} S_j &= \mu_S + \alpha Z_j + \varepsilon_{Sj} \\ T_j &= \mu_T + \beta Z_j + \varepsilon_{Tj} \end{aligned} \quad \Sigma = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ & \sigma_{ST} \end{pmatrix}$$

$$T_j = \mu + \gamma S_j + \varepsilon_j$$

# Prentice's Criteria and Measures

Prentice (1989), Freedman *et al* (1992)

	Quantity	Estimate	Test
1	Effect of $Z$ on $T$	$\beta$	$(T Z) \neq (T)$
2	Effect of $Z$ on $S$	$\alpha$	$(S Z) \neq (S)$
3	Effect of $S$ on $T$	$\gamma$	$(T S) \neq (T)$
4	Effect of $Z$ on $T$ , given $S$	$\beta_S$	$(T Z, S) = (T S)$



**Proportion Explained**

$$PE = \frac{\beta - \beta_S}{\beta}$$



**Relative Effect**

$$RE = \frac{\beta}{\alpha}$$



**Adjusted Association**

$$\rho_Z = \text{Corr}(S, T|Z)$$

# Prentice's Criteria and Measures

Prentice (1989), Freedman *et al* (1992)

	Quantity	Estimate	Test
1	Effect of $Z$ on $T$	$\hat{\beta} = 4.12(2.32)$	$p = 0.079$
2	Effect of $Z$ on $S$	$\hat{\alpha} = 2.83(1.86)$	$p = 0.13$
3	Effect of $S$ on $T$	$\hat{\gamma} = 0.95(0.06)$	$p < 0.0001$
4	Effect of $Z$ on $T$ , given $S$	$\hat{\beta}_S$	



## Proportion Explained

$$\widehat{PE} = 0.65 \quad [-0.22; 1.51]$$



## Relative Effect

$$\widehat{RE} = 1.45 \quad [-0.48; 3.39]$$



## Adjusted Association

$$\hat{\rho}_Z = 0.75 \quad [0.69; 0.82]$$

# Relationship and Problems

---

$$RE = \frac{\beta}{\alpha}$$

$$\rho_Z = \frac{\sigma_{ST}}{\sqrt{\sigma_{SS}\sigma_{TT}}}$$

$$PE = \lambda \cdot \rho_Z \cdot \frac{\alpha}{\beta} = \lambda \cdot \rho_Z \cdot \frac{1}{RE}$$

where

$$\lambda^2 = \frac{\sigma_{TT}}{\sigma_{SS}}$$

- Very wide confidence intervals for PE
- $PE \notin [0, 1]$

# Use of Relative Effect and Adjusted Association

---

- The two new quantities have clear meaning

- ▷ **Relative Effect:** trial-level measure of surrogacy

*Can we translate the treatment effect on the surrogate to the treatment effect on the endpoint, in a sufficiently precise way?*

- ▷ **Adjusted Association:** individual-level measure of surrogacy

After accounting for the treatment effect, is the surrogate endpoint predictive for a patient's true endpoint?

- **BUT:**

The RE is based on a single trial  $\Rightarrow$  regression through the origin, based on one point!

# Analysis Based on Several Trials...

---

- **Context:**

- ▷ multicenter trials
- ▷ meta analysis
- ▷ several meta-analyses

- **Extensions:**

- ▷ **Relative Effect** → **Trial-Level Surrogacy**

How close is the relationship between the treatment effects on the surrogate and true endpoints, based on the various trials (units)?

- ▷ **Adjusted Association** → **Individual-Level Surrogacy**

How close is the relationship between the surrogate and true outcome, after accounting for trial and treatment effects?

# ... Is Considered a Useful Idea

---

Albert *et al* (SiM 1998)

“There has been little work on alternative statistical approaches. A meta-analysis approach seems desirable to reduce variability. Nevertheless, we need to resolve basic problems in the interpretation of measures of surrogacy such as PE as well as questions about the biologic mechanisms of drug action.”

# Statistical Model

---

- **Model:**

$$S_{ij} = \mu_{Si} + \alpha_i Z_{ij} + \varepsilon_{Sij}$$

$$T_{ij} = \mu_{Ti} + \beta_i Z_{ij} + \varepsilon_{Tij}$$

- **Error structure:**

$$\Sigma = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ & \sigma_{TT} \end{pmatrix}$$



# Statistical Model

- **Model:**

$$S_{ij} = \mu_{Si} + \alpha_i Z_{ij} + \varepsilon_{Sij}$$

$$T_{ij} = \mu_{Ti} + \beta_i Z_{ij} + \varepsilon_{Tij}$$

- **Trial-specific effects:**

$$\begin{pmatrix} \mu_{Si} \\ \mu_{Ti} \\ \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \mu_S \\ \mu_T \\ \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} m_{Si} \\ m_{Ti} \\ a_i \\ b_i \end{pmatrix} \quad D = \begin{pmatrix} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ & d_{TT} & d_{Ta} & d_{Tb} \\ & & d_{aa} & d_{ab} \\ & & & d_{bb} \end{pmatrix}$$

# ARMD: Trial-Level Surrogacy

- **Prediction:**

- ▷ *What do we expect ?*

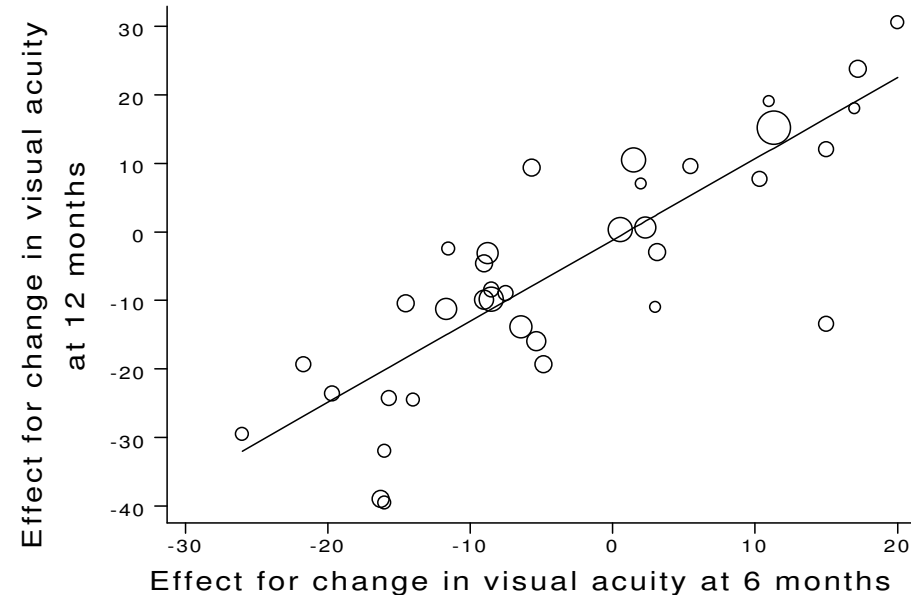
$$E(\beta + b_0 | m_{S0}, a_0)$$

- ▷ *How precisely can we estimate it ?*

$$\text{Var}(\beta + b_0 | m_{S0}, a_0)$$

- **Estimate:**

- ▷  $R^2_{\text{trial}} = 0.692$  (95% C.I. [0.52; 0.86])



# ARMD: Individual-Level Surrogacy

- **Individual-level association:**

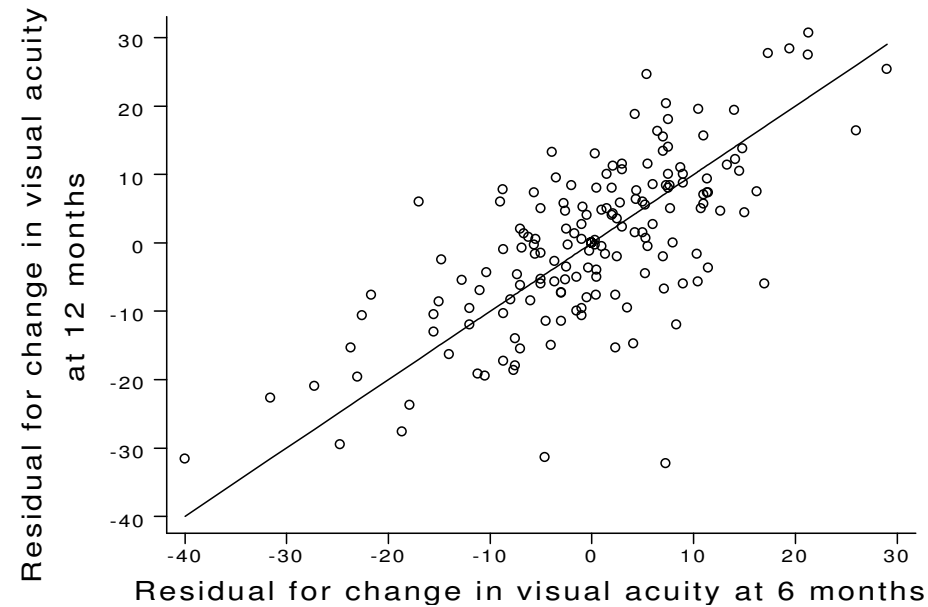
$$\rho_Z = R_{\text{indiv}} = \text{Corr}(\varepsilon_{Ti}, \varepsilon_{Si})$$

- **Estimate:**

- ▷  $R_{\text{indiv}}^2 = 0.483$  (95% C.I. [0.38; 0.59])

- ▷  $R_{\text{indiv}} = 0.69$  (95% C.I. [0.62; 0.77])

- ▷ Recall  $\rho_Z = 0.75$  (95% C.I. [0.69; 0.82])



# A Number of Case Studies

	Age-related macular degeneration	Advanced ovarian cancer	Advanced colorectal cancer
<b>Surrogate True</b>	Vis. Ac. (6 months) Vis. Ac. (1 year)	Progr.-free surv. Overall surv.	Progr.-free surv. Overall surv.
<b>Prentice Criteria 1–3 (<i>p</i> value)</b>			
<b>Association</b> ( $Z, S$ )	0.31	0.013	0.90
<b>Association</b> ( $Z, T$ )	0.22	0.08	0.86
<b>Association</b> ( $S, T$ )	< 0.001	< 0.001	< 0.001
<b>Single-Unit Validation Measures (estimate and 95% C.I.)</b>			
<b>Proportion Explained</b>	0.61[−0.19; 1.41]	1.34[0.73; 1.95]	0.51[−4.97; 5.99]
<b>Relative Effect</b>	1.51[−0.46; 3.49]	0.65[0.36; 0.95]	1.59[−15.49, 18.67]
<b>Adjusted Association</b>	0.74[0.68; 0.81]	0.94[0.94; 0.95]	0.73[0.70, 0.76]
<b>Multiple-Unit Validation Measures (estimate and 95% C.I.)</b>			
$R^2_{\text{trial}}$	0.69[0.52; 0.86]	0.94[0.91; 0.97]	0.57[0.41, 0.72]
$R^2_{\text{indiv}}$	0.48[0.38; 0.59]	0.89[0.87; 0.90]	0.57[0.52, 0.62]

# Overview: Case Studies

	Schizoph. Study I (138 units)	Schizoph. Study I (29 units)	Schizoph. Study II
<b>Surrogate True</b>	— PANSS — — CGI —		
Prentice Criteria 1–3 ( <i>p</i> value)			
<b>Association</b> ( $Z, S$ )	0.016		0.835
<b>Association</b> ( $Z, T$ )	0.007		0.792
<b>Association</b> ( $S, T$ )	< 0.001		< 0.001
<b>Single-Unit Validation Measures (estimate and 95% C.I.)</b>			
<b>Proportion Explained</b>	0.81[0.46; 1.67]		−0.94[∞]
<b>Relative Effect</b>	0.055[0.01; 0.16]		−0.03[∞]
<b>Adjusted Association</b>	0.72[0.69; 0.75]		0.74[0.69; 0.79]
<b>Multiple-Unit Validation Measures (estimate and 95% C.I.)</b>			
$R^2_{\text{trial}}$	0.56[0.43; 0.68]	0.58[0.45; 0.71]	0.70[0.44; 0.96]
$R^2_{\text{indiv}}$	0.51[0.47; 0.55]	0.52[0.48; 0.56]	0.55[0.47; 0.62]

# Two Longitudinal Endpoints

## First Stage

$$\begin{aligned} T_{ijt} &= \mu_{T_i} + \beta_i Z_{ij} + \theta_{T_i} t_{ijt} + \varepsilon_{T_{ijt}} \\ S_{ijt} &= \mu_{S_i} + \alpha_i Z_{ij} + \theta_{S_i} t_{ijt} + \varepsilon_{S_{ijt}} \end{aligned} \quad \Sigma_i = \begin{pmatrix} \sigma_{TT_i} & \sigma_{ST_i} \\ \sigma_{ST_i} & \sigma_{SS_i} \end{pmatrix} \otimes R_i$$

## Second Stage

$$\begin{pmatrix} \mu_{S_i} \\ \mu_{T_i} \\ \alpha_i \\ \beta_i \\ \theta_{S_i} \\ \theta_{T_i} \end{pmatrix} = \begin{pmatrix} \mu_S \\ \mu_T \\ \alpha \\ \beta \\ \theta_S \\ \theta_T \end{pmatrix} + \begin{pmatrix} m_{S_i} \\ m_{T_i} \\ a_i \\ b_i \\ \tau_{S_i} \\ \tau_{T_i} \end{pmatrix}$$

## Evaluation Measures?

# A Sequence of Measures

---

- **Variance Reduction Factor VRF:**

$$VRF = \frac{\sum_i \{ \text{tr}(\Sigma_{TTi}) - \text{tr}(\Sigma_{(T|S)i}) \}}{\sum_i \text{tr}(\Sigma_{TTi})}$$

- **Canonical-correlation Root-statistic Based Measure  $\theta_p$ :**

$$\theta_p = \sum_i \frac{1}{N p_i} \text{tr} \{ (\Sigma_{TTi} - \Sigma_{(T|S)i}) \Sigma_{TTi}^{-1} \}$$

- **Canonical-correlation Root-statistic Based Measure  $R_\Lambda^2$ :**

$$R_\Lambda^2 = \frac{1}{N} \sum_i (1 - \Lambda_i),$$

where

$$\Lambda_i = \frac{|\Sigma_i|}{|\Sigma_{TTi}| |\Sigma_{SSi}|}$$

# A Sequence of Measures

---

- **The Likelihood Reduction Factor LRF:**

- ▷ Consider a pair of models:

$$g_T(T_{ij}) = \mu_{T_i} + \beta_i Z_{ij}$$

$$g_T(T_{ij}) = \theta_{0i} + \theta_{1i} Z_{ij} + \theta_{2i} S_{ij}$$

- ▷  $G_i^2$  log-likelihood ratio for comparison of both models

- ▷ The proposed measure:

$$\text{LRF} = 1 - \frac{1}{N} \sum_i \exp\left(-\frac{G_i^2}{n_i}\right)$$



# An Information-theoretic Approach

---

- Can we unify all previous proposals?

- Shannon (1916–2001) defined **entropy** of a distribution:

$$h(Y) = E[-\log(f(Y))]$$

- Conditional version:

$$h(Y|X = x) = E_{Y|X}[\log f_{Y|X}(Y|X = x)] \quad \text{and} \quad I(Y|X) = E_X[h(Y|X = x)]$$

- The amount of uncertainty (entropy) that is expected to be removed if the value of  $X$  is known:

$$I(X, y) = h(Y) - h(Y|X)$$

# An Information-theoretic Approach

---

- Informational measure of association  $R_h^2$ :

$$R_h^2 = R_h^2 = \frac{EP(Y) - EP(Y|X)}{EP(Y)}$$

with

$$EP(X) = \frac{1}{(2\pi e)^n} e^{2h(X)}$$

- Version for  $N$  trials:

$$R_h^2 = \sum_{i=1}^{N_q} \alpha_i R_{hi}^2 = 1 - \sum_{i=1}^{N_q} \alpha_i e^{-2I_i(S_i, T_i)},$$

where the  $\alpha_i$  form a convex combination.

# Relationships With Previous Definitions

---

- All have desirable behavior within  $[0, 1]$  for continuous endpoints
- All can be embedded within a family
- $\theta_p$  is symmetric in  $S$  and  $T$  whereas the VRF is not
- $\theta_p$  is invariant w.r.t. linear bijective transformations; VRF only when they are orthogonal
- $R_{\Lambda}^2$  and later ones also apply to non-Gaussian settings

# Relationships With Previous Definitions

---

- Later ones specialize to earlier ones
- They all reduce to the  $R_{\text{indiv}}^2$  for cross-sectional Gaussian outcomes
- Longitudinal normal setting:

$$R_h^2 = R_\Lambda^2 \quad \text{if} \quad \alpha_i = N_q^{-1}$$

- General setting:

$$\text{LRF} \xrightarrow{P} R_h^2$$

when the number of subjects per trial approaches  $\infty$

# Other Implications

---

- Relationship with Prentice's main criterion and the Data Processing Inequality:

$$\begin{aligned} f(T|Z, S) = F(T|S) &\Rightarrow Z \rightarrow S \rightarrow T \\ &\Rightarrow I(T, Z|S) = 0 \\ &\Rightarrow I(Z, S) \geq I(Z, T) \end{aligned}$$

- PE and  $R_h^2$ :

$$\text{PE} = 1 - \frac{\beta_S}{\beta} \quad \longleftrightarrow \quad R_h^2 = 1 - \frac{\text{EP}(\beta_i|\alpha_i)}{\text{EP}(\beta_i)}$$

# Fano's Inequality

---

- Fano's Inequality:

$$E[(T - g(S))^2] \geq EP(T)(1 - R_h^2)$$

- ▷ Left hand side is prediction error
- ▷ Applies regardless of distributional form and predictor function  $g(\cdot)$
- ▷ **“How large does  $R_h^2$  have to be?”** ← The answer depend crucially on the power entropy of  $T$

# Schizophrenia Trial

- **Continuous Outcomes:**

- ▷  $VRF_{\text{ind}} = 0.39$  with 95% C.I. [0.36; 0.41]

- ▷  $R_{\text{trial}}^2 = 0.85$  with 95% C.I. [0.68; 0.95]

- **Binary Outcomes:**

Parameter	Estimate	95% C.I.
<b>Trial-level <math>R_{\text{trial}}^2</math> measures</b>		
Information-theoretic	0.49	[0.21,0.81]
Probit	0.51	[0.18,0.78]
Plackett-Dale	0.51	[0.21,0.81]
<b>Individual-level measures</b>		
$R_h^2$	0.27	[0.24,0.33]
$R_{h\text{max}}^2$	0.39	[0.35,0.48]
Probit	0.67	[0.55,0.76]
Plackett-Dale $\psi$	25.12	[14.66;43.02]
Fano's lower-bound	0.08	

# Age-related Macular Degeneration Trial

---

- Both outcomes binary:

Parameter	Estimate	[95% C.I.]
$R_{\text{trial}}^2$	0.3845	[0.1494;0.6144]
$R_h^2$	0.2648	[0.2213;0.3705]
$R_{h\text{max}}^2$	0.4955	[0.3252;0.6044]



# Advanced Colorectal Cancer

---

$S$ : Time to progression/death

$T$ : Time to death

- Models:

$$h_{ij}(t) = h_{i0}(t)\exp\{\beta_i Z_{ij}\}$$

$$h_{ij}(t) = h_{i0}(t)\exp\{\beta_{Si} Z_{ij} + \gamma_i S_{ij}(t)\}$$

# Advanced Colorectal Cancer

Parameter	Estimate (95% C.I.)	
	Dataset I	Dataset II
<b>Trial-level measures</b>		
$\hat{R}_{\text{trial}}^2$ (separate models)	0.82 [0.40;0.95]	0.85 [0.53;0.96]
$\hat{R}_{\text{trial}}^2$ (Clayton copula)	0.88 [0.59;0.98]	0.82 [0.43;0.95]
$\hat{R}_{\text{trial}}^2$ (Hougaard copula)		0.75 [0.00;1.00]
<b>Individual-level measures</b>		
$\hat{R}_h^2$	0.84 [0.82;0.85]	0.83 [0.82;0.85]
Percentage of censoring	19%	55%

# Prediction in a New Trial

---

- Consider a new trial  $i = 0$ :

$$S_{0j} = \mu_{S0} + \alpha_0 Z_{0j} + \varepsilon_{S0j}$$

- **Prediction variance:**

$$\text{Var}(\beta + b_0 | \mu_{S0}, \alpha_0, \vartheta) \approx f\{\text{Var}(\hat{\mu}_{S0}, \hat{\alpha}_0)\} + f\{\text{Var}(\hat{\vartheta})\} + (1 - R_{\text{trial}}^2)\text{Var}(b_0)$$

- where

- ▷  $f(\cdot)$  are appropriate functions of the parameters involved
- ▷  $\vartheta$  contains all fixed effects

# Prediction in a New Trial

---

- Meaning of the three terms:
  - ▷ **Estimation error in both the meta-analysis and the new trial:**  
all three terms apply
  - ▷ **Estimation error in the meta-analysis only:**

$$\text{Var}(\beta + b_0 | \mu_{S0}, \alpha_0, \vartheta) \approx f\{\text{Var}(\hat{\vartheta})\} + (1 - R_{\text{trial}}^2)\text{Var}(b_0)$$

- ▷ **No estimation error:**

$$\text{Var}(\beta + b_0 | m_{S0}, a_0) = (1 - R_{\text{trial}}^2)\text{Var}(b_0)$$

# The Surrogate Threshold Effect

---

- **STE:** The smallest treatment effect upon the surrogate that predicts a significant treatment effect on the true endpoint
- *Various versions:*
  - ▷  $STE_{N,n}$ : STE for a finite meta-analysis and a finite new trial
  - ▷  $STE_{N,\infty}$ : STE for a finite meta-analysis and an infinite new trial
  - ▷  $STE_{\infty,\infty}$ : STE when both the meta-analysis and the new trial are infinitely large

# Practical Conclusions

---

- *Are surrogate endpoints useful in practice?*
- An investigator wants to be able to predict the effect of treatment on  $T$ , based on the observed effect of treatment on  $S$ .
- $R_{\text{trial}}^2$ ,  $R_{\text{indiv}}^2$ ,  $(\psi, \tau)$ , VRF,  $\theta_p$ ,  $R_{\Lambda}^2$  LRF,  $R_h^2$ , ....: quantification of surrogacy in a meta-analytic setting
- Prediction: useful in a *new* trial

# Methodological Conclusions

---

- **Basis for new assessment strategy**

- ▷ trial-level surrogacy

- ▷ individual-level surrogacy

- **Requirements**

- ▷ Was required: joint model for surrogate and true endpoint

- ▷ Was required: acknowledgment of the hierarchical structure

- ▷ Matters simplify with information-theoretic approach

---

Society for Clinical Trials, Short Course “Biomarkers in Clinical Trials: General Principles for Study Design and Statistical Evaluation with Case Studies”, 5/20/12

---

# Assessment of Biomarker Assay Performance: When are Biomarkers Ready for Prime Time?

---

**Gene Pennello, PhD**, Team Leader,  
Diagnostic Devices Branch,  
Division of Biostatistics, FDA  
Silver Spring MD

1



# Outline

- **Biomarkers**
  - Types (Co Dx, IVDMA, etc.)
  - Validation (independent data set, “intent to diagnose”)
- **Analytical Performance**
  - Accuracy
  - Limit of Detection
  - Precision (repeatability, reproducibility)
- **Clinical Performance**
  - Prospective-Retrospective Validation
  - Missing Test Results
  - Labeling of Approved Dx Devices
  - Subgroup Misclassification
- **Concluding Remarks**

# Biomarker Intended Uses

- **Diagnosis**, in symptomatic patients
- **Early detection (screening)**, enabling intervention at an earlier and potentially more curable stage than under usual clinical diagnostic conditions
- **Monitoring of disease** response during therapy, with potential for adjusting level of intervention (e.g. dose) on a dynamic and personal basis
- **Risk assessment**, leading to preventive interventions for those at sufficient risk
- **Prognosis**, allowing for more (less) aggressive therapy for patients with worse (better) prognosis
- **Prediction**. E.g., predicts safety, efficacy (PK/PD) of a specific therapy, thereby providing guidance in selecting it for patients or tailoring its dose.

***Last three are attempts to predict the future.***

3

# Companion Diagnostic Device

- In Vitro Companion Diagnostic Devices  
*(Draft, Jul 2011)*
  - An companion in vitro diagnostic device is “one that provides information that is essential for the safe and effective use of a corresponding therapeutic product”.
  - That is, “[it] allows the therapeutic product’s benefits to exceed its risks”.
- Biomarker is used to make treatment decisions, such as treatment selection or dosing (in oncology, it is called a *predictive biomarker*).

4

# Companion Diagnostics, FDA Approved

- Safety
  - CYP2D6 genotypes' effect on metabolic rate for drugs
  - HLA allele B\*1502 as a marker for carbamazepine-induced Stevens-Johnson syndrome and toxic epidermal necrolysis
  - UGT1A1 genotype for risk of neutropenia in CRC patients taking irinotecan
  - KRAS mutation for likely absence of cetuximab, panitumumab efficacy in CRC patients.
- Effectiveness
  - HER2 +, breast cancer patient for trastuzumab.
  - EGFR +, CRC patients for cetuximab, panitumumab.
  - ALK break apart FISH +, NSCLC patients for crizotinib.
  - BRAF V600 mutation +, metastatic melanoma patients for vemurafenib (RO5185426).
- Dosing
  - VKORC1 and CYP2C9 genotype to predict warfarin dose.

# IVDMIA

- In Vitro Diagnostic Multivariate Index Assays (*Draft, Jul 2007*)
  - An IVDMIA “combines the values of multiple variables using an interpretation function to yield a single, patient-specific result (e.g., a “classification,” “score,” “index,” etc.),
  - intended for use in the diagnosis of disease or other conditions, or in the cure, mitigation, treatment or prevention of disease, and
  - provides a result whose derivation is non-transparent and cannot be independently derived or verified by the end user.”

6

# Pre-Market Review of IVDs

- Analytical Validation: *does my test measure the analyte I think it does? Correctly? Reliably?*
- Clinical Validation: *does my test result correlate with the expected clinical presentation? How reliably?*

# Independent Validation

- To establish the utility of a medical test, validation dataset should be completely independent of derivation dataset.
- Refinements to a test include
  - Acceptance range of control
  - Input range (e.g., of DNA)
  - Cut-off(s)
  - For IVDMIAs, the set of predictors (analytes, clinical variables, etc.)

# Intent to Diagnose (ITD)

- In statistical analysis, include all patients on whom a diagnosis could have been attempted:
  - Report number (percent) of subjects without results (invalid, unevaluable, equivocal, etc.).
  - When appropriate, consider imputation of missing test results.

FDA Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests, *Final 2007*.

<http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/default.htm>

Campbell, Pennello, and Yue, 2011, Missing Data in the Regulation of Medical Devices, *J Biopharm Stat*, 21(2), 180-195

•9



# Analytical Performance

# Analytical Validation Steps

- Accuracy (agreement with a reference)
- Precision (repeatability, reproducibility)
- Limit of Detection (sensitivity)
- Interference, Cross-reactivity (specificity)
- Matrix effects
- Sample preparation / conditions
- Performance around the cut-off
- Potential for carryover, cross-hybridization

# Analytical Validation Steps

Required Steps Vary with

- Technology
- Result Type
  - quantitative, semi-quantitative, qualitative
- Setting of use
  - e.g., marketed vs. single laboratory service
- What is reported
  - individual markers vs. composite score

# Clinical Laboratory Standards Institute (CLSI) Guidelines

- FDA formally recognizes several:
  - **EP5** Precision Performance of Quantitative Measurement Methods
  - **EP6** Linearity of Quantitative Measurement Procedures
  - **EP9** Method Comparison and Bias Estimation Using Patient Samples
  - **EP12** Qualitative Test Performance
  - **EP17** Limit of Detection
- If banking samples for later use, see also
  - **MM13** Collection, Transport, Preparation, and Storage of Specimens for Molecular Methods; Approved Guideline.

•13

# Accuracy, BRAF V600 Test

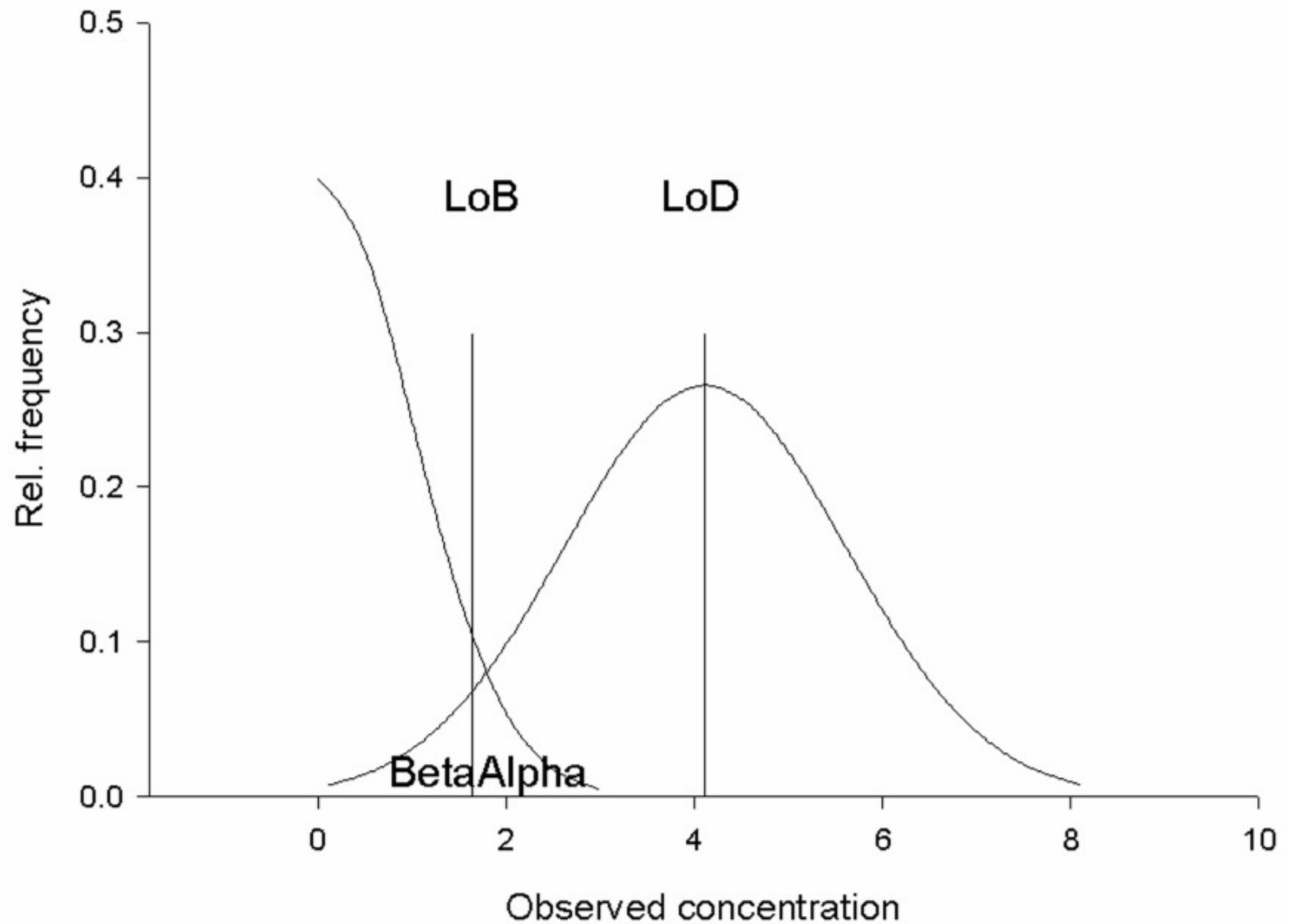
- Melanoma patients are given vemurafenib if tumor carries BRAF V600E mutation.

Cobas <sup>®</sup> 4800 BRAF V600 test <sup>†</sup>	Bi-directional sequencing*			Ttl
	V600E Not Detected	V600E Detected	Invalid	
V600 Not Detected	192	6	16	214
V600 Detected	35	216	31	282
<b>Total</b>	<b>227</b>	<b>222</b>	<b>47</b>	<b>496</b>

<sup>†</sup>Cobas test cross-reacted with V600K in 25 of 38 specimens (65.8%)

\*Bi-directional sequencing limit of detection is ~20% of mutant alleles <sup>14</sup> in FFPET specimen DNA.

# Limit of Detection



5

# LoD, BRAF V600 Test

- FFPET specimen
- Limit of Detection (LoD)
  - Genomic DNA Input Range: Recommended DNA input for the cobas@ 4800 BRAF V600 Mutation Test is  $\geq$  **125 ng**.
  - Minimum Tumor Content: **5%** BRAF V600E mutation DNA blended with BRAF wildtype DNA can be detected with probability  $\geq$  95%.
- LoD for % mutant DNA could vary with DNA input level (low, standard, high).

16

<http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfTopic/pma/pma.cfm?num=P110020>

# Precision Testing

- Intended to capture total test variability (imprecision) of repeated measurements (all steps from specimen prep to final result).
- **Repeatability:** Precision when repeated measurements are taken under the same conditions (i.e., within a run).
- **Intermediate precision:** Precision when varying some conditions (run, day, reagent lot, operator instrument,) but holding others constant (lab).
- **Reproducibility:** multi-lab precision



# Precision Experiments

<b>Factor</b>	<b>Blood</b>	<b>Tissue</b>
Labs	3	3
Days per lab	20	5
Runs per day	2	1
Replicates per run	2	2
Total	240	30

- **Tissue Sampling:** Perhaps only up to 30 serial sections may be available for precision testing to avoid biological variability in tissue.

# Intermediate Precision Study

- Repeatability imprecision  $s_{K(U,D)}$  is pooled SD of K replicates within U runs, D days.
- Intermediate imprecision is

$$s_W = \sqrt{s_D^2 + s_{U(D)}^2 + s_{K(U,D)}^2}$$

- Typically, %CV < 5-10% is considered acceptable.
- Variance components estimated by MOM<sub>19</sub>

# Vermillion OVA1™ IVD MIA

- Vermillion OVA1™, *diagnostic*
  - Combines results from five immunoassays into a score for assessing likelihood that an ovarian adnexal mass is malignant.  
[http://www.accessdata.fda.gov/cdrh\\_docs/reviews/K081754.pdf](http://www.accessdata.fda.gov/cdrh_docs/reviews/K081754.pdf)
- Immunoassays of Five Markers:
  - CA 125
  - Prealbumin
  - Transferrin
  - Apolipoprotein A-1
  - $\beta$ 2-microglobulin
- Range of numerical score 0.0 - 10.0

# OVA1™ Precision Testing

Parameters		Specimen				
		1	2	3	4	5
<b>CA125 II, U/mL</b>						
Mean		9.02	14.04	17.02	20.92	352.1
Repeatability (within run)	SD	0.354	0.210	0.839	0.525	5.131
	%CV	3.9	1.5	4.9	2.5	1.5
Between run	SD	0.176	0.590	0.679	1.054	20.53
	%CV	2.0	4.2	4.0	5.0	5.8
Between day	SD	0.140	0.176	0.386	0.000	5.054
	%CV	1.6	1.3	2.3	0.0	1.4
Between operator	SD	0.453	0.294	0.000	0.000	3.306
	%CV	5.0	2.1	0.0	0.0	0.9
Between sites	SD	0.476	0.380	0.138	0.236	5.182
	%CV	5.3	2.7	0.8	1.1	1.5
Reproducibility (total)	SD	0.708	0.766	1.146	1.187	22.22
	%CV	7.9	5.5	6.7	5.7	6.3

# OVA1™ Precision Testing

Parameters		Specimen				
		1	2	3	4	5
<b>OVA1 result</b>						
Mean		2.67	3.21	3.75	5.00	9.71
Repeatability (within run)	SD	0.069	0.094	0.157	0.364	0.157
	%CV	2.6	2.9	4.2	7.3	1.6
Between run	SD	0.034	0.087	0.091	0.000	0.129
	%CV	1.3	2.7	2.4	0.0	1.3
Between day	SD	0.000	0.000	0.146	0.032	0.045
	%CV	0.0	0.0	3.9	0.6	0.5
Between operator	SD	0.042	0.039	0.105	0.265	0.000
	%CV	1.6	1.2	2.8	5.3	0.0
Between sites	SD	0.057	0.141	0.000	0.103	0.212
	%CV	2.1	4.4	0.0	2.1	2.2
Reproducibility (total)	SD	0.098	0.176	0.250	0.447	0.271
	%CV	3.7	5.5	6.7	8.9	2.8

# Precision Testing, IVDMIAs

- Precision can be evaluated at three levels of the prediction algorithm:
  - Individual analytes (scoring algorithm inputs)
    - Evaluate with samples at low, middle, and high levels of the analyte
  - Score (given by algorithm)
    - Evaluate with samples with low, middle, and high values of the score
  - Medical decision or classification (based on cut-off(s) in the score)

# Precision Testing, IVDMIAs

- The same score can be obtained from different sets of values of the analytes.
  - A sample with a particular value of the score only represents one possible set with that value
- For  $k$  analytes,  $3^k$  possible combinations of low, middle, and high levels of each analyte.
  - Infeasible to evaluate all for  $k \gg 5$ , say.
  - Because of correlation, many combinations may never occur in clinical samples and therefore are not relevant.

•24

# Clinical Performance



# Clinical Validation

- **BGM Galectin-3 Assay.** An in vitro diagnostic device that quantitatively measures galectin-3 in serum or plasma by enzyme-linked immunosorbant assay (ELISA) on a microtiter plate platform.
- BGM Galectin-3 Assay is indicated to be used in conjunction with clinical evaluation as an aid in assessing the prognosis of patients diagnosed with chronic heart failure (HF).

•26

# Prospective-Retrospective Validation

- **Pivotal Study.** Heart Failure: A Controlled Trial Investigating Outcomes of Exercise Training (HF-ACTION).
- The HF-ACTION study involved 2,331 chronic HF patients with left ventricular dysfunction and with NYHA class II, III or IV symptoms.
- To validate the clinical effectiveness of the cut-off values for the BGM Galectin-3 assay, Galectin-3 levels were measured by the assay in 895 banked EDTA-plasma samples from chronic heart failure participants in the HF-ACTION study.

•27

# Key Conditions for Prospective-Retrospective Validation

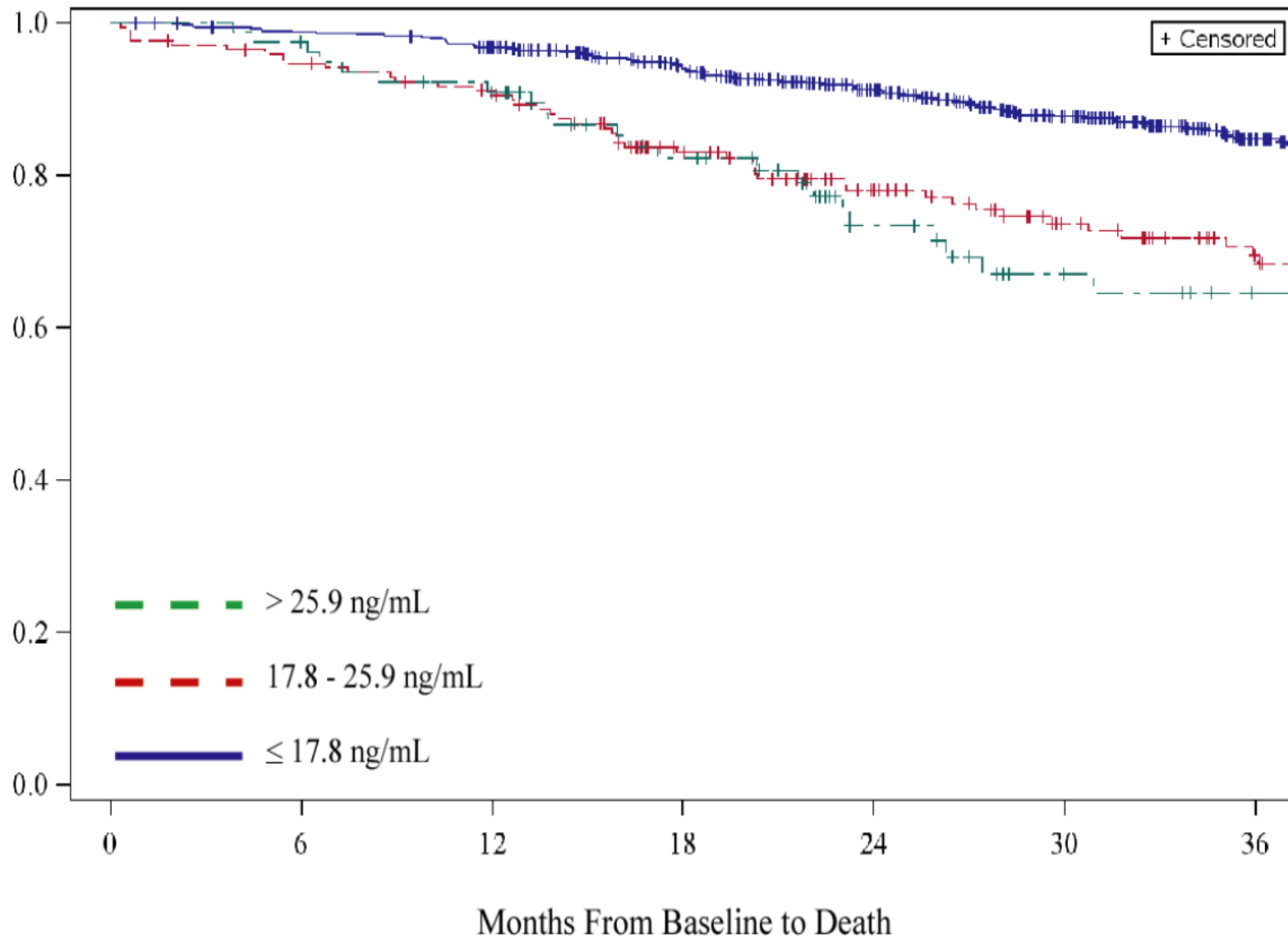
1. Adequate, well-conducted, well-controlled trial with eligibility criteria the same as the assay.
2. Specimens are available on a large predominance of subjects.
3. Analysis plan is completely pre-specified.
4. Assay demonstrates acceptable analytical performance on archived specimens.
5. Assay result is obtained on a large portion of archived specimens.
6. User of assay is masked to the clinical data.

Mack. *Nature Biotech*, 2009, 27(2), 110-2.

Subramanian, Simon. *Nat Rev Clin Onc*, 2010, 7, 327-34. .28

Simon, Paik, Hayes, *JNCI*, 2009; 101, 1446-52.

# Galectin 3 Kaplan-Meier Curves, All-Cause Mortality



# Predictive Values, All-Cause Mortality

	Cumulative Probability of All-Cause Mortality Event (95% CI) by Galectin-3 Category and Time Point (in percent)			
Galectin-3 Category	6 months	12 months	24 months	36 months
≤ 17.8 ng/mL	1.2% (0.6%-2.5%)	3.3 (2.1-5.0)	8.7 (6.7-11.3)	15.3 (12.4-18.8)
17.8-25.9 ng/mL	5.3 (2.8-10.0)	8.9 (5.5-14.4)	22.0 (16.3-29.4)	30.5 (23.4-39.1)
> 25.9 ng/mL	2.6 (0.6-9.9)	9.1 (4.4-18.1)	26.6 (17.5-39.1)	35.5 (24.5-49.5)

# Missing Data Sensitivity Analysis

- Galectin-3 values were imputed conservatively for the 1436 remaining patients in the dataset based on the probability of the assay categorizing a patient into a high or low risk group.
- The difference in survival curves for the risk groups remained statistically significant, indicating that the results on the evaluable subset (895) were robust and representative of the entire study population.

# Non-Informative Imputation

- For outcome  $Y$ , binary test result  $T$ , missing test result indicator  $V$ , assume

$$\Pr(T_+ | Y, V-) = \Pr(T_+ | V-)$$

- That is, missing test results are independent of (non-informative for)  $Y$ .
- As an ITD analysis of robustness, NI imputation is an alternative to assuming
  - test results are missing at random.
  - all missing test results “disagree” with clinical outcome  $Y$  (worse case scenario).

•32

# Robustness of Inference to Non-Informative Imputation

1. Obtain bootstrap sample of  $n$  subjects. Let  
 $n_1$  = number of subjects with test results,  
 $x_1$  = number of  $n_1$  subjects categorized as **high risk**.
2. For each missing test result in bootstrap sample,
  - (a) draw  $\text{Pr}(\text{high risk}) = p \sim \text{Beta}(x_1 + a, n_1 - x_1 + b)$ ,  
the posterior of  $p$  under prior  $\text{Beta}(a, b)$ ,
  - (b) draw  $Z \sim \text{Bernoulli}(p)$ ; impute missing result as  
**high risk** if  $Z=1$ , **not high risk** if  $Z=0$ .
3. Using completed data, compute hazard ratio between high and low risk groups.
4. Repeat 1-3 to obtain 95% bootstrap CI on hazard ratio.

Because imputed test results are non-informative for survival time, hazard ratio is conservatively estimated. See

- Efron 1994, *J Amer Stat Assoc*, 89, 463-475.
- Campbell, Pennello, Yue, 2010, *J Biopharm Stat*.

•33



# Missing Test Results

- **Types**

- Specimen not available for testing
- Specimen unevaluable
- Test result invalid
- Diagnostic testing not attempted

- **Examples**

- Retrospective analysis of available specimens
- Retest tissue specimens already tested with a reference method or a clinical trial assay.

# Robustness of Inference to Missing Test Results

- 1) **Identify a set of covariates which can affect test result** (e.g., use logistic regression or linear model of test result on covariates).
- 2) **Check for imbalance in the covariates** between samples in test analysis set and in non-test analysis set.
- 3) **Impute test results assuming they are**
  - missing at random
  - missing not at random by various scenarios:
    - non-informative for clinical condition
    - unfavorable relative to clinical condition (e.g., for patients surviving the longest, imputed test results confer high risk for the clinical event or high likelihood of being a non-responder to therapy).

# Variables

- Patient Characteristics
- Disease characteristics
- Handling and processing factors
- Specimen Characteristics
- Outcome

# Patient characteristics

- Gender
- Race
- Age
- Baseline ECOG PS
- Baseline weight
- Marker status by reference method

# Disease characteristics

- Months from first histological diagnosis to randomization
- Number of disease sites
- Presence of metastases (yes or no)
- Number of previous therapies
- Prior therapy (yes or no)

# Handling, processing factors

- Enrollment site
- Region (e.g., Canada, Non-Canada)
- Age of sample at testing
- Sampling method (biopsy, resection)

# Characteristics of sample

- Tumor type (primary or metastatic)
- If metastatic, then site of metastasis
- Area of tumor tissue (mm<sup>2</sup>)
- Tumor content in sample (%)
- Macro-dissection of sample (yes or no)
- Necrosis score in tumor area (0, 1, 2 or 3)
- H&E staining slide evaluable (yes or no)

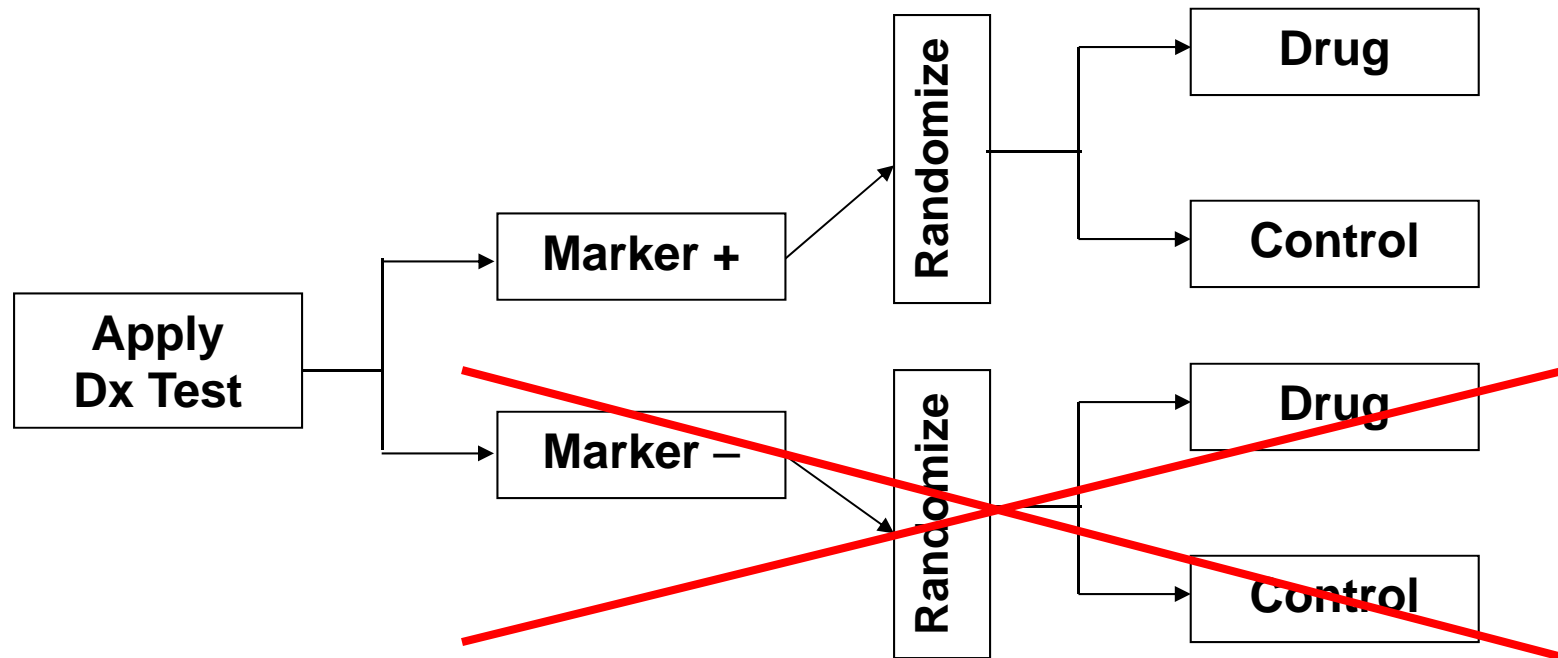
# Predictive Markers, Labeling

- Biomagene PATHIAM™ System Assisted Scoring
  - Accessory to DAKO HercepTest to aid in ... semi-quantitative measurement of HER2/neu in FFPE tissue ... of breast cancer patients for whom HERCEPTIN® (Trastuzumab) treatment is being considered.
  - HER2/neu results are indicated for use as an aid in the management, **prognosis** and **prediction** of therapy outcomes of breast cancer.
- Roche cobas® 4800 BRAF V600 Mutation Test.
  - Intended to be used as an aid in **selecting** melanoma patients whose tumors carry the BRAF V600E mutation for treatment with vemurafenib.
- Dako Egfr pharmdx IHC Kit.
  - Indicated as an aid in identifying colorectal cancer patients **eligible** for treatment with erbitux (cetuximab), or vectibix (panitumumab).

•41



# Enrichment (Targeted) Design



- Marker effectiveness (i.e., marker by treatment interaction) cannot be assessed!
- Claim is *not* that device is predictive, but can reliably identify a subset of subjects in whom drug is S & E<sup>42</sup>.

# COBAS 4800 BRAF V600 Mutation Test Label

- .....intended for the qualitative detection of the **BRAF V600E** mutation in **DNA** extracted from formalin-fixed, paraffin-embedded human melanoma tissue..... to be used as an aid in **selecting** melanoma patients whose tumors carry the **BRAF V600E** mutation for treatment with **vemurafenib**.

# Vemurafenib Label

- ... indicated for the treatment of patients with unresectable or metastatic melanoma with BRAFV600E mutation as detected by an **FDA-approved test**.
- Limitation of Use: ZELBORAF is not recommended for use in patients with wild-type BRAF melanoma.
- ...The efficacy and safety of ZELBORAF have not been studied in patients with wild-type BRAF melanoma....

# Pre-Test Screening

- A subject that is marker positive by a *laboratory developed test* (LDT +) is encouraged to enroll into the Phase II/III trial.
- In trial, drug effect is studied in subjects who are marker positive by a *market ready test* (MRT +).
- *Spectrum Effect*
  - LDT +, MRT + subjects are studied.
  - LDT –, MRT + subjects are not.



# “Get Melanoma” Tested”

(Advice of CollabRx website)

- “Based on the information you provided, testing for certain genetic mutations may help select potentially relevant treatments.....Print out this page to discuss with your doctor.”
- “Several drugs that block BRAF, such as [*Redacted*], are in clinical testing and some have shown promise in cancer patients.”

[http://therapy.collabrx.com/melanoma/view?get\\_tested\\_origin\\_skin\\*BRAF-CKIT](http://therapy.collabrx.com/melanoma/view?get_tested_origin_skin*BRAF-CKIT)


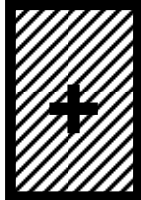

# Trial of MRT + Subjects

Study	Y	LDT	MRT
Enrolled			

---

Subjects Pre-Screened by LDT<sup>47</sup>

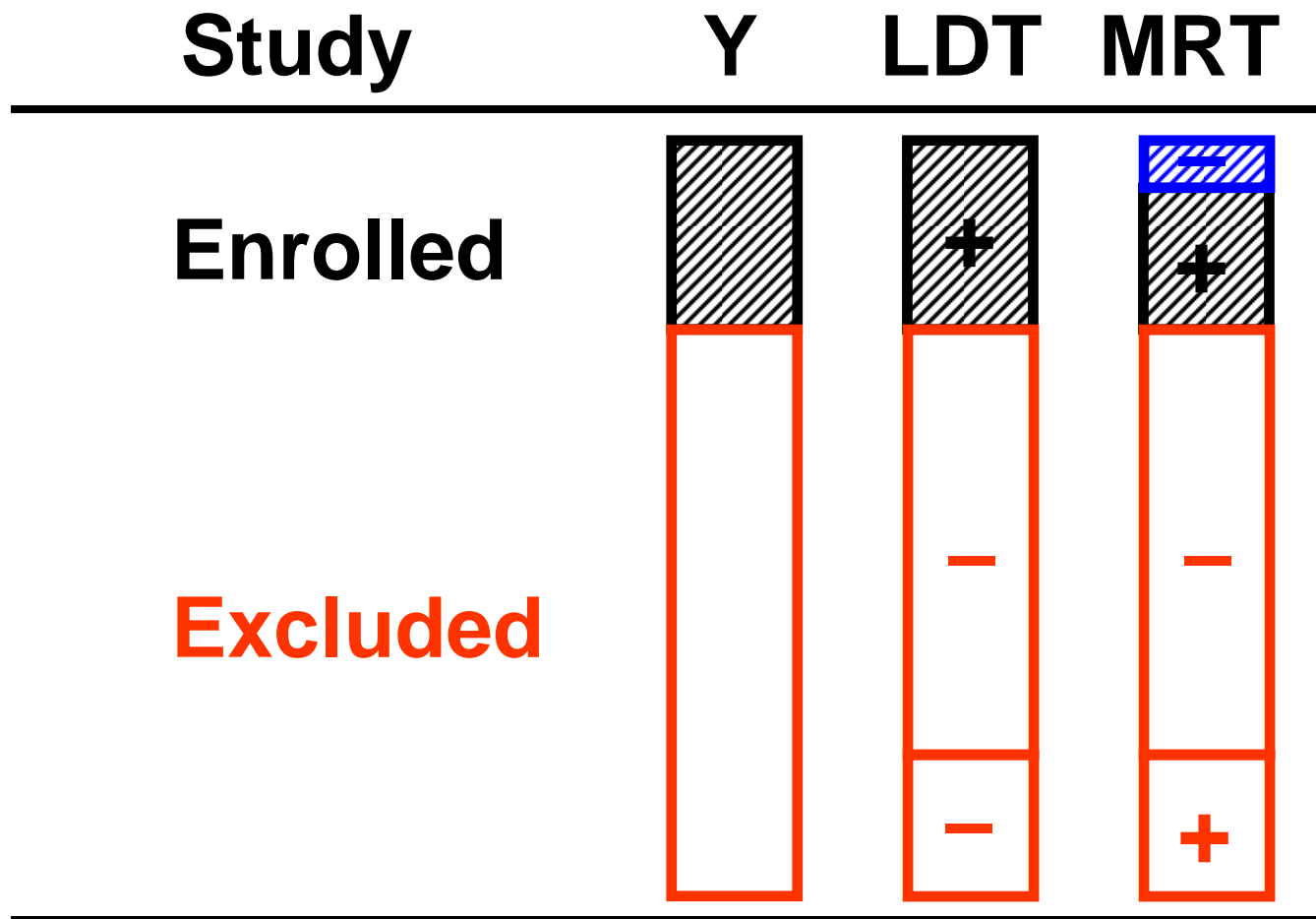
# Trial of MRT + Subjects

Study	Y	LDT	MRT
Enrolled			

---

Subjects Pre-Screened by LDT<sup>48</sup>

# Trial of MRT + Subjects



A subset of MRT + subjects were excluded from the trial.  
Study population  $\neq$  IU population for either drug or marker.

49



# Subgroup Misclassification

- Response  $R=0,1$  to treatment
- Subgroup  $S=0,1$  (reference result)
- Surrogate  $S^*=0,1$  (Dx test result)
- Assume misclassification of  $S$  by  $S^*$  is *non-differential*, that is

$$S^* | S, R = S^* | S$$

# Subgroup Misclassification

- **Attenuation Result:** Let

$$D = \Pr(R = 1 | S = 1) - \Pr(R = 1 | S = 0)$$
$$D^* = \Pr(R = 1 | S^* = 1) - \Pr(R = 1 | S^* = 0)$$

- Then  $D^* = D \times (PPV + NPV - 1)$

- where  $PPV = \Pr(S = 1 | S^* = 1)$

$$NPV = \Pr(S = 0 | S^* = 0)$$

Kuha, Skinner, Palmgren, 2005, "Misclassification Error" in *Encyc Biostat*<sup>51</sup>

# Concluding Remarks

- **Biomarker Discovery**
  - FDA has programs to assist sponsors:
    - CDRH preIDE meeting with device sponsor.
    - CDER Qualification of Drug Development Tools (DDTs), including biomarkers.
- **Analytical Performance**
  - Good performance should be demonstrated before device is applied to specimens.
- **Clinical Performance**
  - Clinical significance should be demonstrated.
  - Claims in labeling depend on studies conducted.

•52

# FDA Guidance

- In Vitro Companion Dx Devices, *Draft 2011*
- Reporting Results from Studies Evaluating Diagnostic Tests, *Final 2007*
- Design Considerations for Pivotal Clinical Investigations of Medical Devices, *Draft 2011*
- In Vitro Dx Multivariate Index Assays, *Draft 2007*
- Pharmacogenetic Tests and Genetic Tests for Heritable Markers, *Final 2007*
- Special Control – Ovarian Adnexal Mass Assessment Score Test System, *2011*
- Special Control – Cardiac Allograft Gene Expression Profiling Test Systems, *2009*

•53

# Acknowledgements

- Robert L. Becker Jr., M.D., Ph.D.  
OIVD/CDRH/FDA
- Elizabeth Mansfield, Ph.D.  
OIVD/CDRH/FDA
- Donna Roscoe, Ph.D.  
OIVD/CDRH/FDA
- Thomas Gwise, Ph.D.  
OB/CDER/FDA
- Diagnostics Devices Branch,  
Division of Biostatistics/OSB/CDRH/FDA



# SUPPLEMENTAL

# Medical Devices

- **Safety** [21CFR860.7(d)(1)] :
  - “...based upon valid scientific evidence,
  - the probable benefits ... from use of the device
  - for its intended uses and conditions of use,
  - ..... outweigh any probable risks
- **Effectiveness** [21CFR860.7(e)(1)] :
  - “... based upon valid scientific evidence,
  - .....the use of the device
  - for its intended uses and conditions of use,
  - .... will provide clinically significant results.”

# IVD Label Requirement

- 21CFR809.10(b)(12)
  - Include....such things as:
    - Accuracy
    - Precision
    - Specificity
    - Sensitivity
  - These shall be related to a generally accepted method using biological specimens from normal and abnormal populations.



# Drug Labeling

- 21CFR201.57 (2)(i)
  - If specific tests are necessary for selection ....
  - of the patients who need the drug .....,
  - [include] the identity of such tests.

# Predictive Biomarker

- Marker: Her2-neu
- Device: Pathvysion HER-2 DNA Probe Kit
- Indications: .....The PathVysion Kit is further indicated as an aid to predict disease-free and overall survival in patients with stage II, node positive breast cancer treated with adjuvant cyclophosphamide, doxorubicin, and 5-fluorouracil (CAF) chemotherapy. (PathVysion label)

The PathVysion Kit is indicated as an aid in the assessment of patients for whom HERCEPTIN® (Trastuzumab) treatment is being considered (refer to HERCEPTIN package insert).

•59

# Biomarker Trial Designs: Lessons from Real Trials

Sumithra J. Mandrekar, PhD

Director of Biostatistics

Alliance for Clinical Trials in Oncology

Associate Professor of Biostatistics

Mayo Clinic, Rochester MN

Society for Clinical Trials, Pre-conference Workshop, May 2012

# Guiding Principle

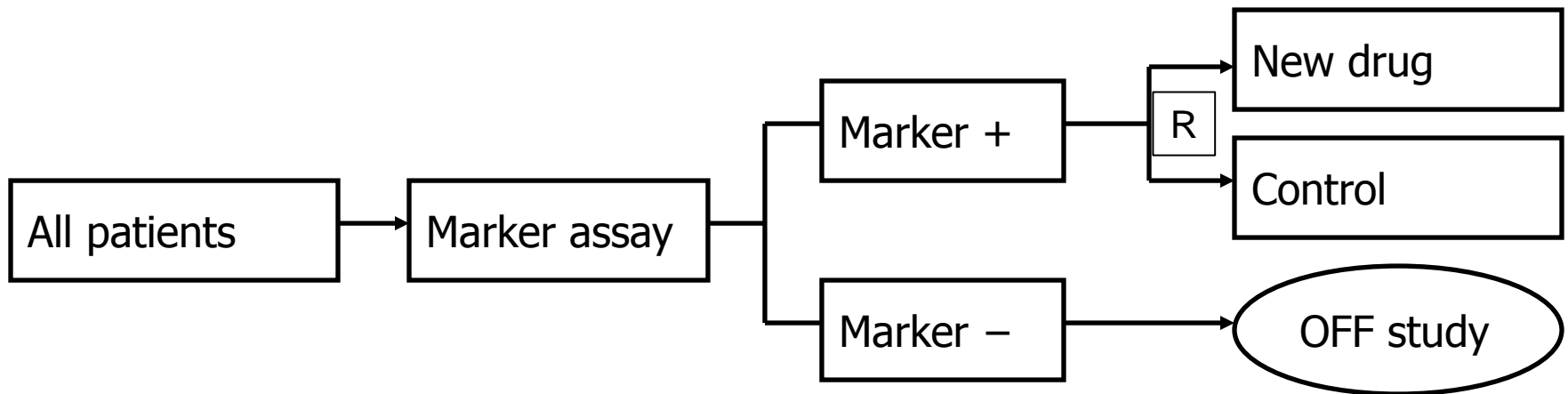
- Data used to develop the marker or classifier should be distinct from the data used to test hypotheses about marker-treatment effects
- Marker or classifier refers to:
  - Single gene / protein / other biologic variable
  - Composite score based on multiple-gene expression

# Predictive Biomarker Development

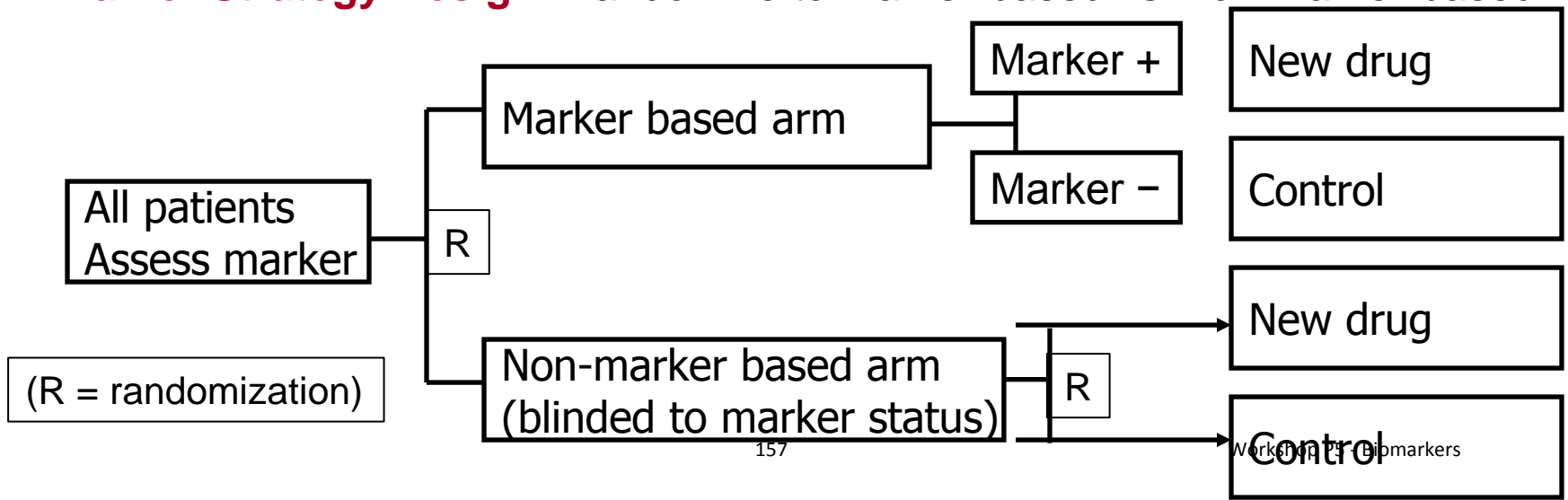
- ❖ Ideally, predictive tests (assays, signatures etc.) developed in parallel with drug development
  - ❖ Reality: biomarker and drug development not always synchronized
- ❖ A biomarker-based test might be “good enough” for the development and testing of a drug but it may not be ready for clinical use when the drug is ready

# Biomarker-Based Clinical Trial Designs

- **Enrichment or Targeted Design:** Randomize marker positive patients only



- **Marker Strategy Design:** Randomize to marker-based vs. non-marker-based.



# Biomarker-Based Clinical Trial Designs

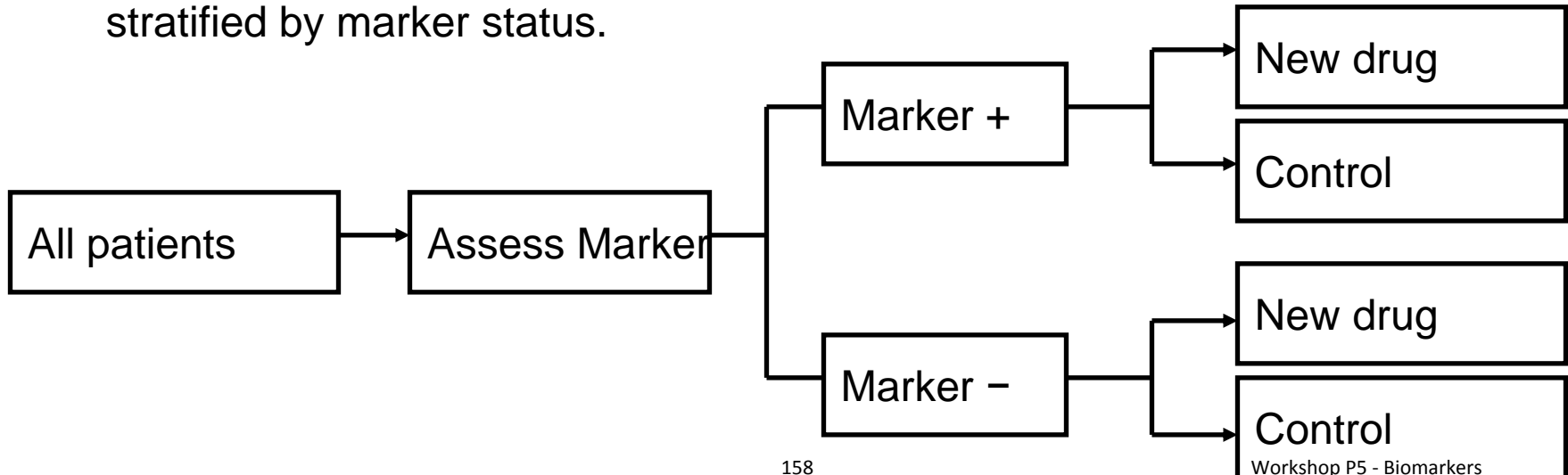
- **All-comers Design:** Randomize all patients, measure marker.



Biomarker tested on all patients, but treatment assignment/randomization is not based on marker results.

M+: marker positive pts.  ; M-: marker negative pts.  ; T1 : Treatment 1; T2: Treatment 2.

- **Marker by treatment interaction Design:** Randomize all patients, stratified by marker status.



# Sequential Testing Strategy Designs

- Test treatment effect in the overall population first and then in a prospectively planned subset if overall effect is not significant, or
- Test effect in the marker-defined subgroup first, and then in the entire population if the subgroup analysis is significant (closed testing procedure)
  
- Subset Analyses
- Adaptive Threshold Design
- Adaptive Signature Designs

Mandrekar and Sargent, JCO 2009;  
Freidlin et al., CCR 2005; 2010;  
Jiang et al., JNCI 2007



# ENRICHMENT DESIGNS

# Schema: N9831

## HER2+ Breast cancer patients

R  
A  
N  
D  
O  
M  
I  
Z  
E

Arm A: AC q 3w x 4 → Paclitaxel qw x 12

Arm B: AC q 3w x 4 → Paclitaxel qw x 12 → H qw x 52

Arm C: AC q 3w x 4 → Paclitaxel qw x 12 + H qw x 12 → H qw x 40

↑  
RT and/or hormonal  
therapy as indicated

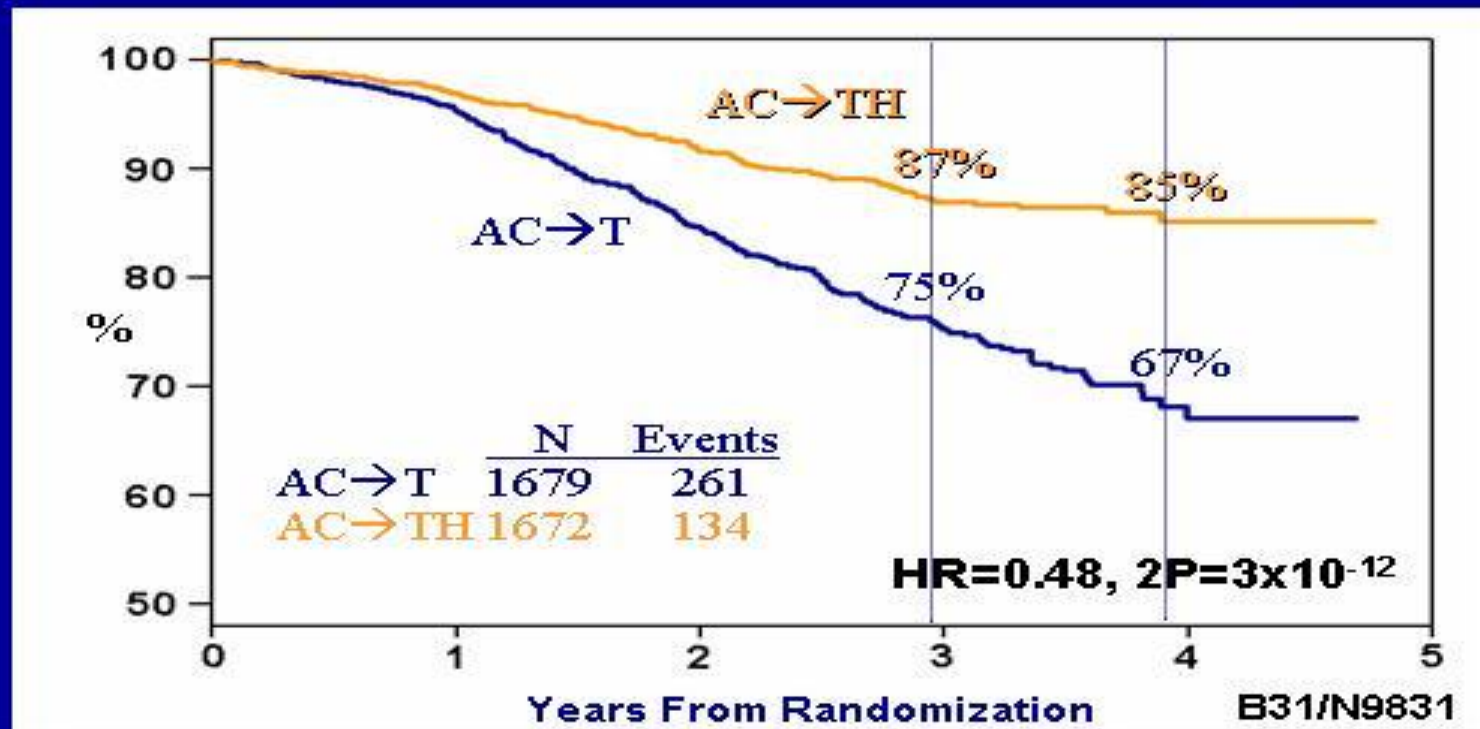
n=3,505

Perez EA

H=trastuzumab (4 mg/kg loading dose, followed by 2 mg/kg); A=doxorubicin dose 60 mg/m<sup>2</sup>; C=cyclophosphamide, 600 mg/m<sup>2</sup>; paclitaxel, 80 mg/m<sup>2</sup>; q 3w=every 3 weeks, qw=weekly

# Using markers to restrict trial eligibility: Success – Her 2+ Breast Cancer

## Disease-Free Survival: Joint Analysis



Romond et al, NEJM 2005

# Herceptin in Her2- breast cancer?

- High discordance between local and central testing for HER 2 status
- Herceptin therapy may benefit a potentially larger group than the approximately 20% of patients defined as HER2 positive by central testing in these two trials

Paik et al., NEJM 2008;  
Perez et al., JCO 2006

# Using markers to restrict trial eligibility: beware

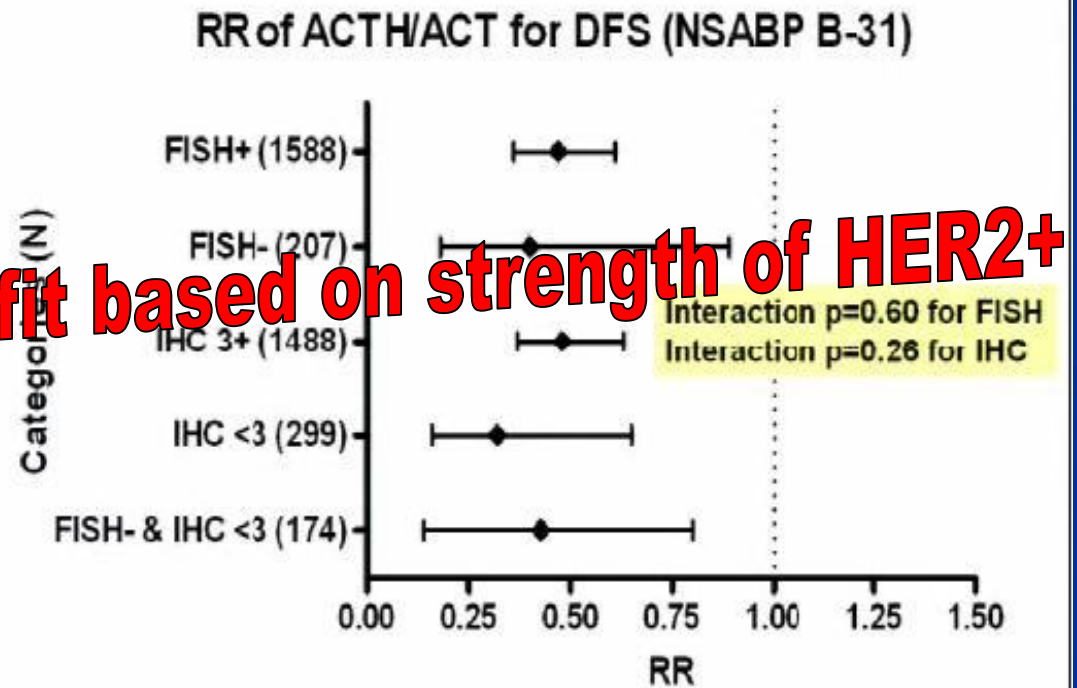
**Table 1.** Relative Risks of Disease Progression and Death among Patients in the ACT with the ACT Group.\*

End Point and Central HER2 Assay†	ACT no. of events/total no. of events	ACTH	Relative Risk (95% CI)
<b>Disease progression</b>			
HER2-positive	163/875	85/804	0.47 (0.37–0.62)
HER2-negative	20/92	7/82	0.34 (0.14–0.80)
<b>Death</b>			
HER2-positive	55/875	38/804	0.66 (0.43–0.99)
HER2-negative	10/92	1/82	0.08 (0.01–0.64)

\* The 95% confidence intervals (CI) and P values were adjusted according to the number of events from the univariate Cox proportional-hazards model for each subgroup in the Adjuvant Breast and Bowel Project B-31 trial. ACT denotes doxorubicin, cyclophosphamide, and epirubicin plus trastuzumab.

† Central HER2 assay results were defined as negative if they were negative by both fluorescence in situ hybridization (PathVysion, Vysis) and immunohistochemical analysis (Herceptest, Dako) and were positive otherwise.

**No difference in benefit based on strength of HER2+**



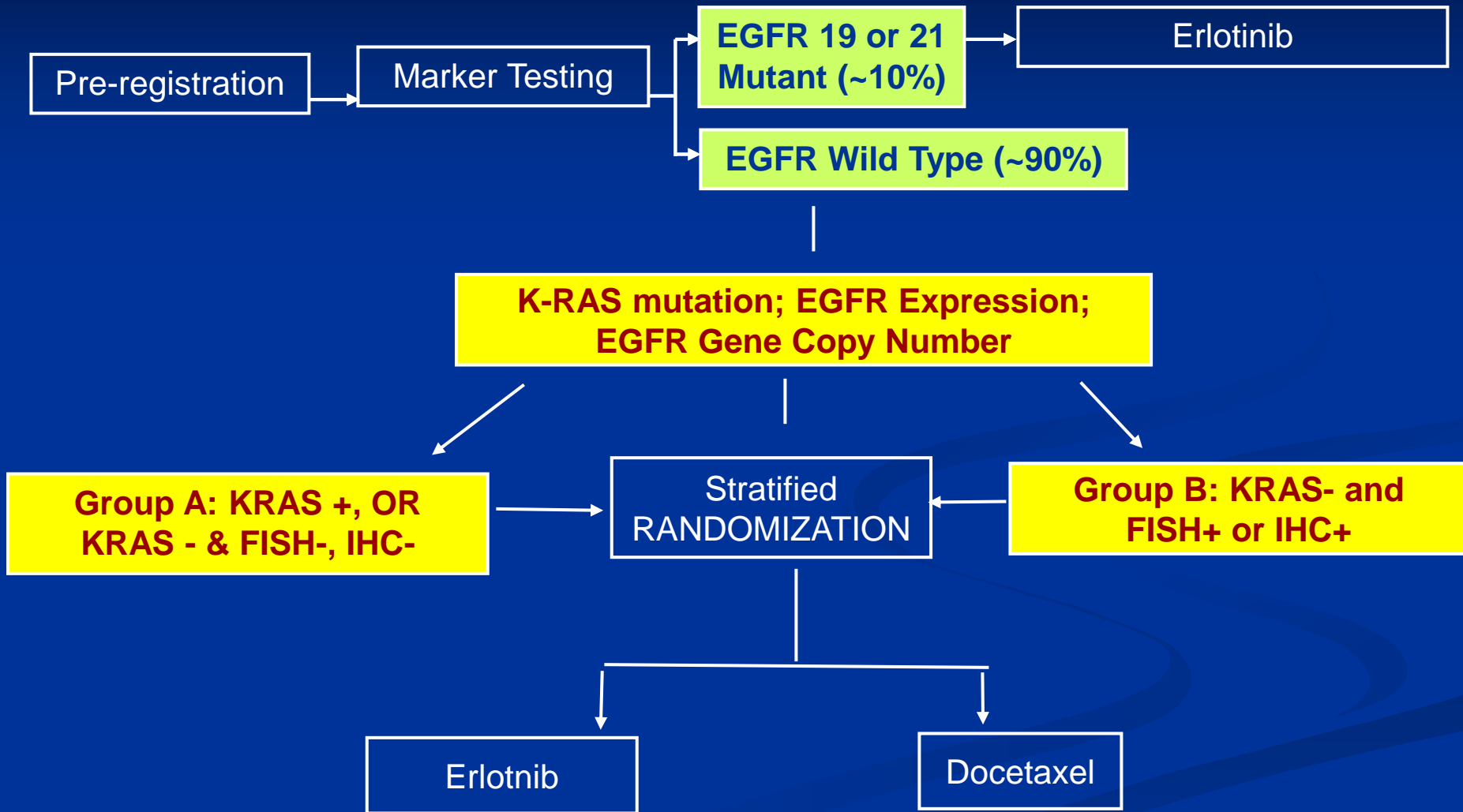
**Ongoing study of Herceptin in patients with low (1+ or 2+) HER2-positive BC.**

Paik et al, NEJM 2008

Hayes et al., NEJM 2011

# MARKER BY TREATMENT INTERACTION DESIGNS

# TAILOR: Phase III Second line NSCLC Marker by Treatment Interaction Design



(Farina et al., Clinical Lung Cancer 2011)

# TAILOR: Primary Hypothesis

- Endpoint: OS
  - Erlotinib (E) and Docetaxel (D) have similar OS in the unselected population; median OS ~ 7 months
- Primary Hypothesis:
  - D better than E in Group A: 30% improvement in OS , for a HR of 1.43 in favor of D
  - E better than D in Group B: 21% improvement in OS, for a HR of 0.79 in favor of E
- Equal allocation of patients to groups A and B
- N= 650 (325/arm); Interaction Test
  - Overall alpha=0.05 (two-sided); Power=95%

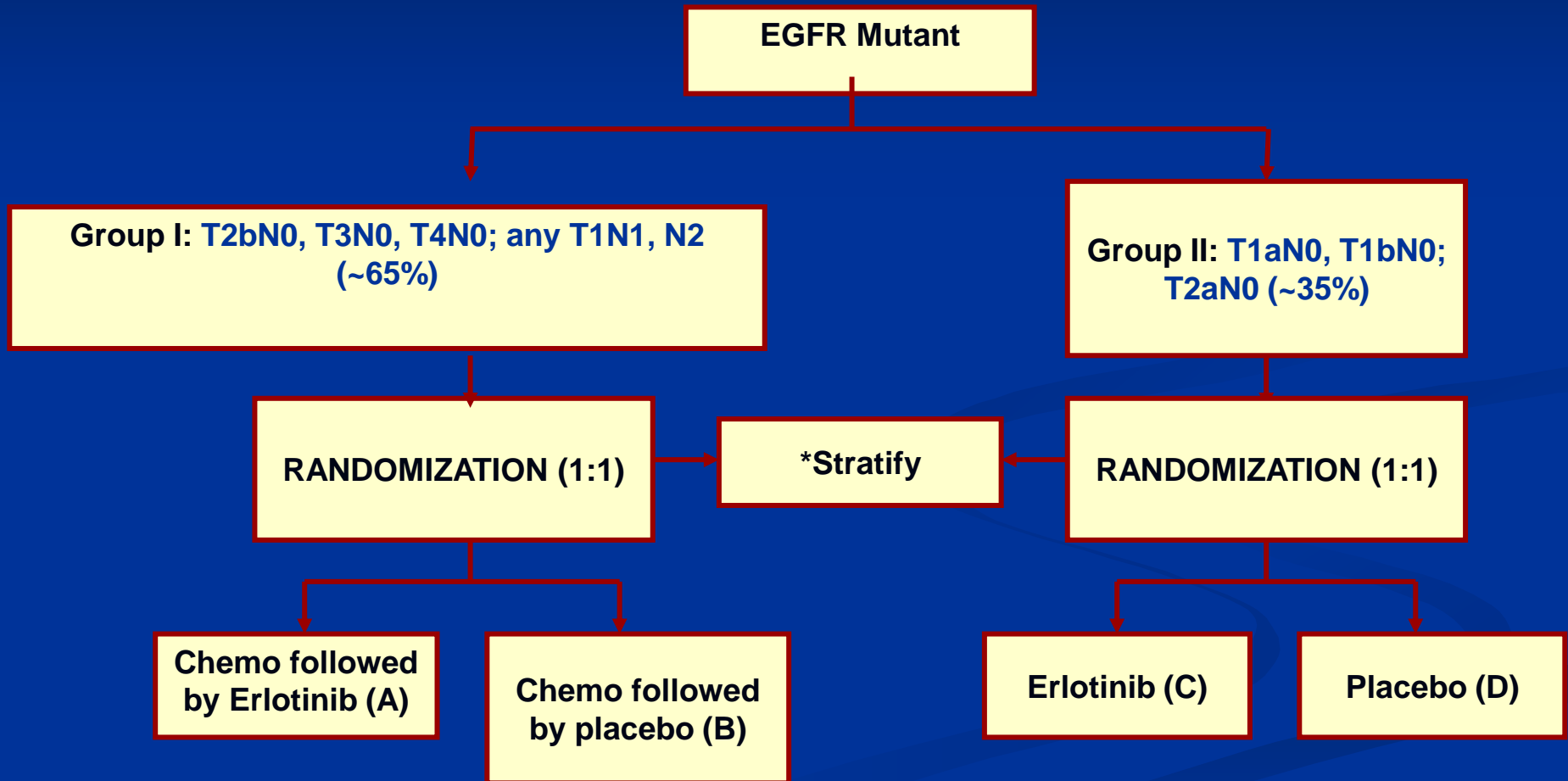


# TAILOR: Secondary Hypothesis

- Within Group Comparisons (not adequately powered to detect clinically relevant differences?)
  - Group A: 325 patients
  - D better than E in Group A: 30% improvement in OS , for a HR of 1.43 in favor of D
  - Two-sided  $\alpha=0.05$ , power=86%
  
  - Group B: 325 patients
  - E better than D in Group B: 21% improvement in OS, for a HR of 0.79 in favor of E
  - Two-sided  $\alpha=0.05$ , power=56%

# Z41102: Personalized Adjuvant Treatment in completely resected NSCLC

## Double Blind Placebo controlled trial



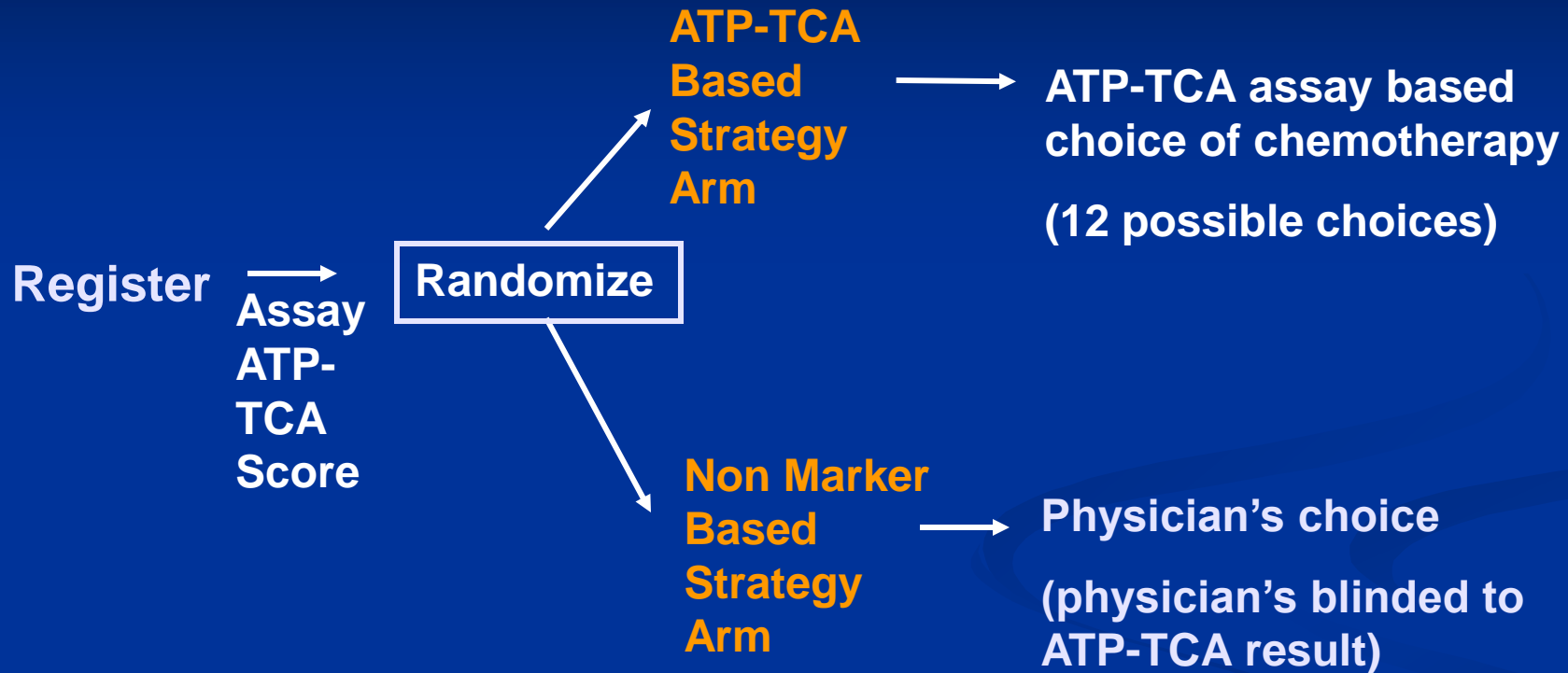
\* Strat factors: PS, smoking, histology, EXON 19 deletion

# Z41102: Design Details

- Primary Endpoint: OS
- Primary comparison: Compare PAT to SOC
  - Compare OS between Arms A and C versus B and D
  - Detect a hazard ratio (HR) of at least 0.67 in favor of erlotinib
  - 50% improvement, or 7.5 years versus 5.0 years in median OS
- Target sample size: 410
  - 1-sided  $\alpha=0.05$ ; power=86%
  - Stratified log rank test

# MARKER BASED STRATEGY DESIGNS

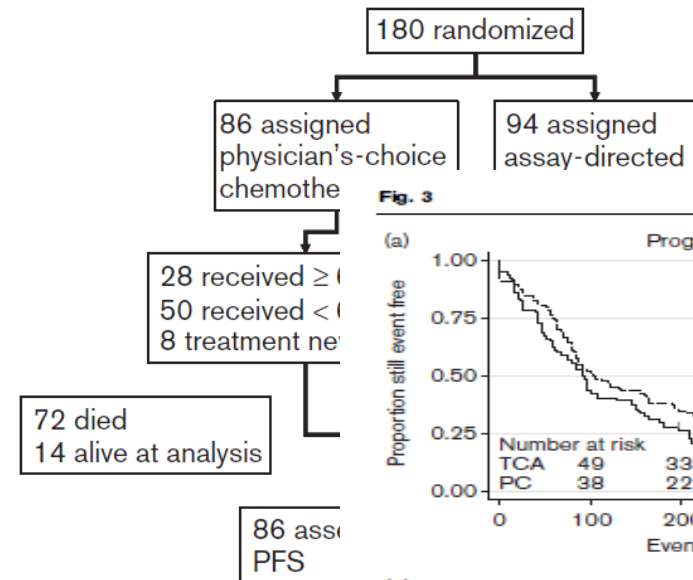
# Tumor Chemosensitivity Assay in recurrent platinum resistant Ovarian Cancer Marker Strategy Design



Primary endpoint: compare response rates between the ATP-TCA based arm to that of the non-marker based arm

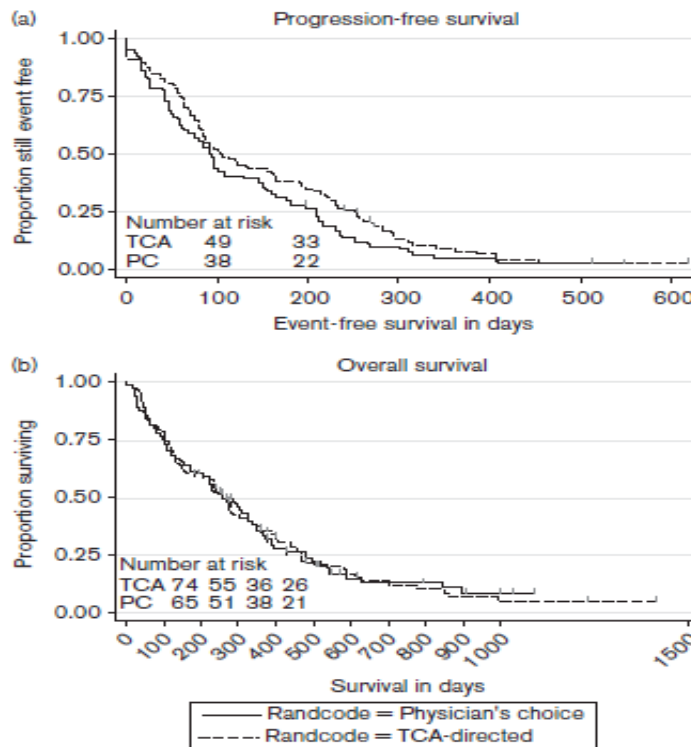
Design: 90 patients/arm; alpha=10%; power=80%; RR of 30% versus 50% (ATP-TCA arm)

Fig. 1



Trial profile, showing the completing treatment. PFS

Fig. 3



Kaplan–Meier survival curves for (a) progression-free survival and (b) overall survival, showing a trend towards improved progression-free survival (hazard ratio 0.80, 95% confidence interval 0.59–1.10) with no difference between the groups in overall survival (hazard ratio 1.01, 95% confidence interval 0.7–1.3). PC, physician's choice; TCA, tumour chemosensitivity assay.

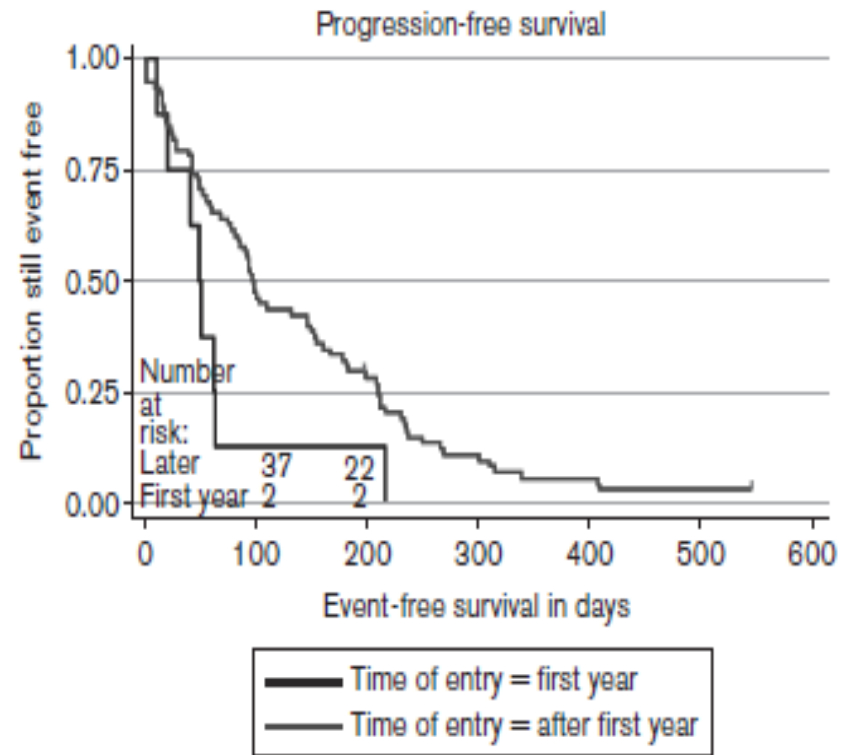
Chemotherapy regimens used during the trial, with the number of patients used

Regimen	Physician's choice	Assay directed
1	9	5
2	3	0
3	1	2
4	5	0
5	9	1
6	1	1
7	1	0
8	5	5
9	10	10
10	13	23
11	3	3
12	18	31
<b>Total</b>	<b>50/78</b>	<b>72/81</b>

# Learning Curve?

- Physician's choice arm:
  - Oncologists switched to the use of similar combinations in the non-marker based arm as the ATP-TCA directed arm.
  - Late randomization – better PFS!
- ~ 70% Overlap in treatments on both arms – dilutes the ability to distinguish treatment from marker effect!

Fig. 4



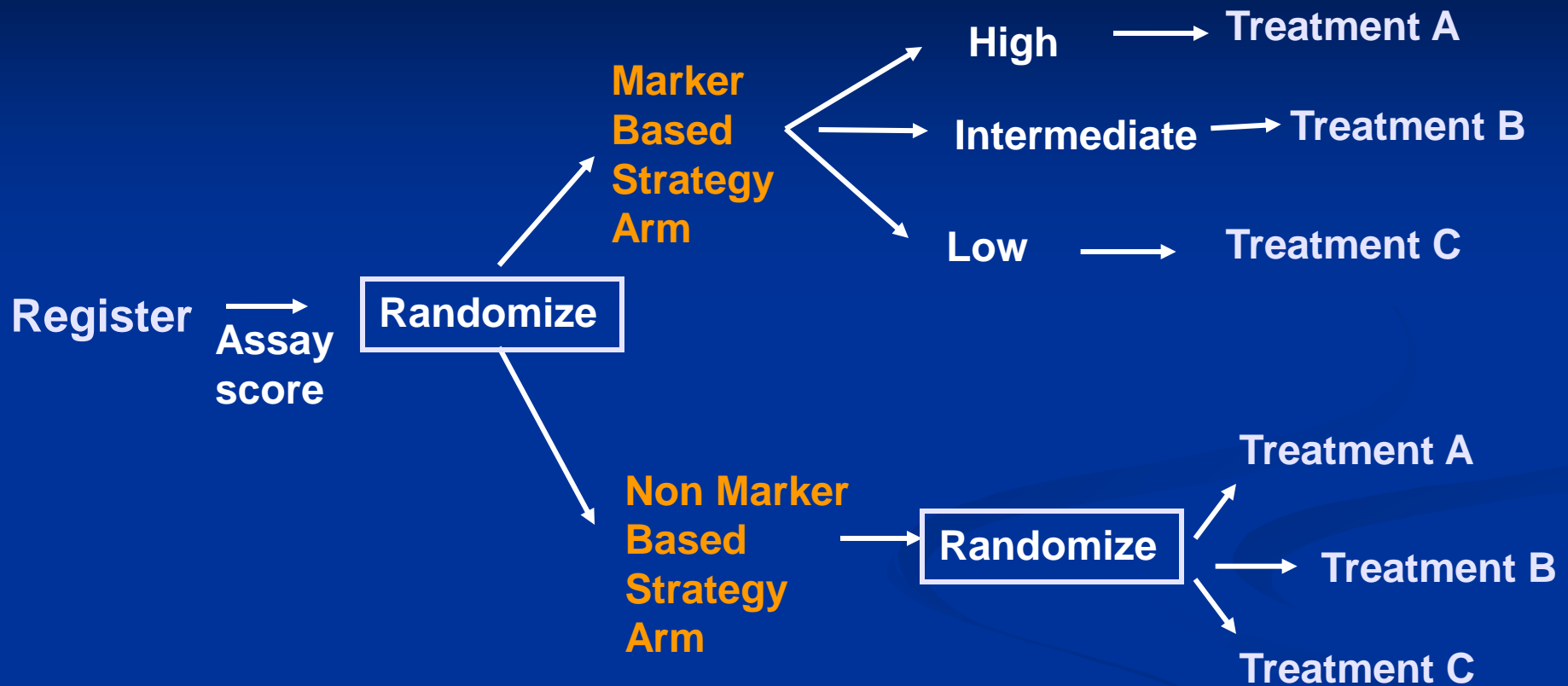
Patients entering this arm during the first year had a significantly reduced progression-free survival in comparison with later entrants (hazard ratio 0.44, 95% confidence interval 0.2–0.9,  $P < 0.03$ , log-rank analysis).

# Design Limitations

- Significant overlap of pts (depending on prevalence) receiving the same regimen in both arms
  - Dilutes the treatment effect, thus lowers power
- Independent comparisons of each regimen not possible
  - All marker subgroups do not receive all treatments
- Ethical issues - cannot give a certain treatment to a certain subgroup
- Logistically challenging (long time to accrue, large trial etc.)



# Marker Strategy Design (Version 1.1)



Independent comparisons of each regimen now possible

Is the efficacy of the marker directed approach due to the effect of the marker, or due to a better treatment regardless of marker?

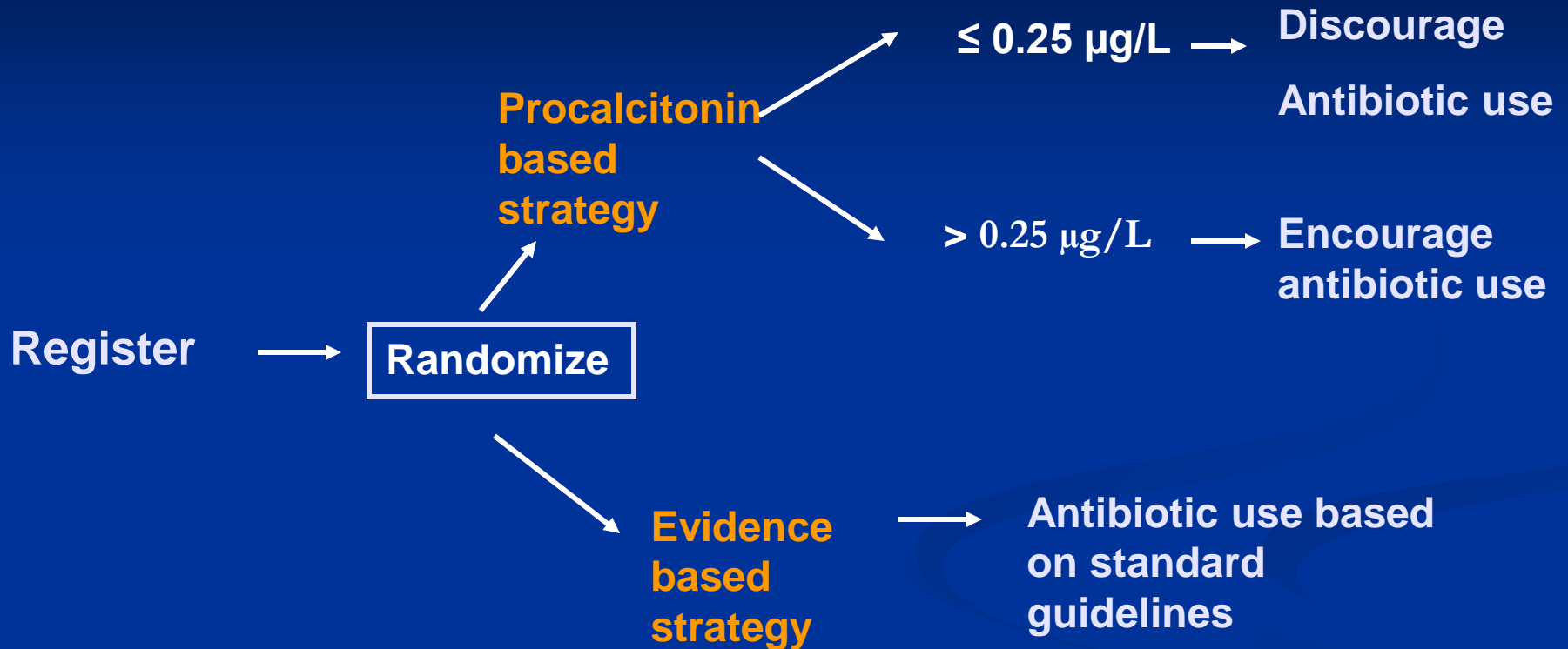
# Antibiotic use for respiratory tract infections: Procalcitonin (PCT)-based vs. Standard Guidelines

Schuetz et al., JAMA, 2009

Two Key issues -

- Who gets treated with antibiotics?
- What is the duration for antibiotic therapy?

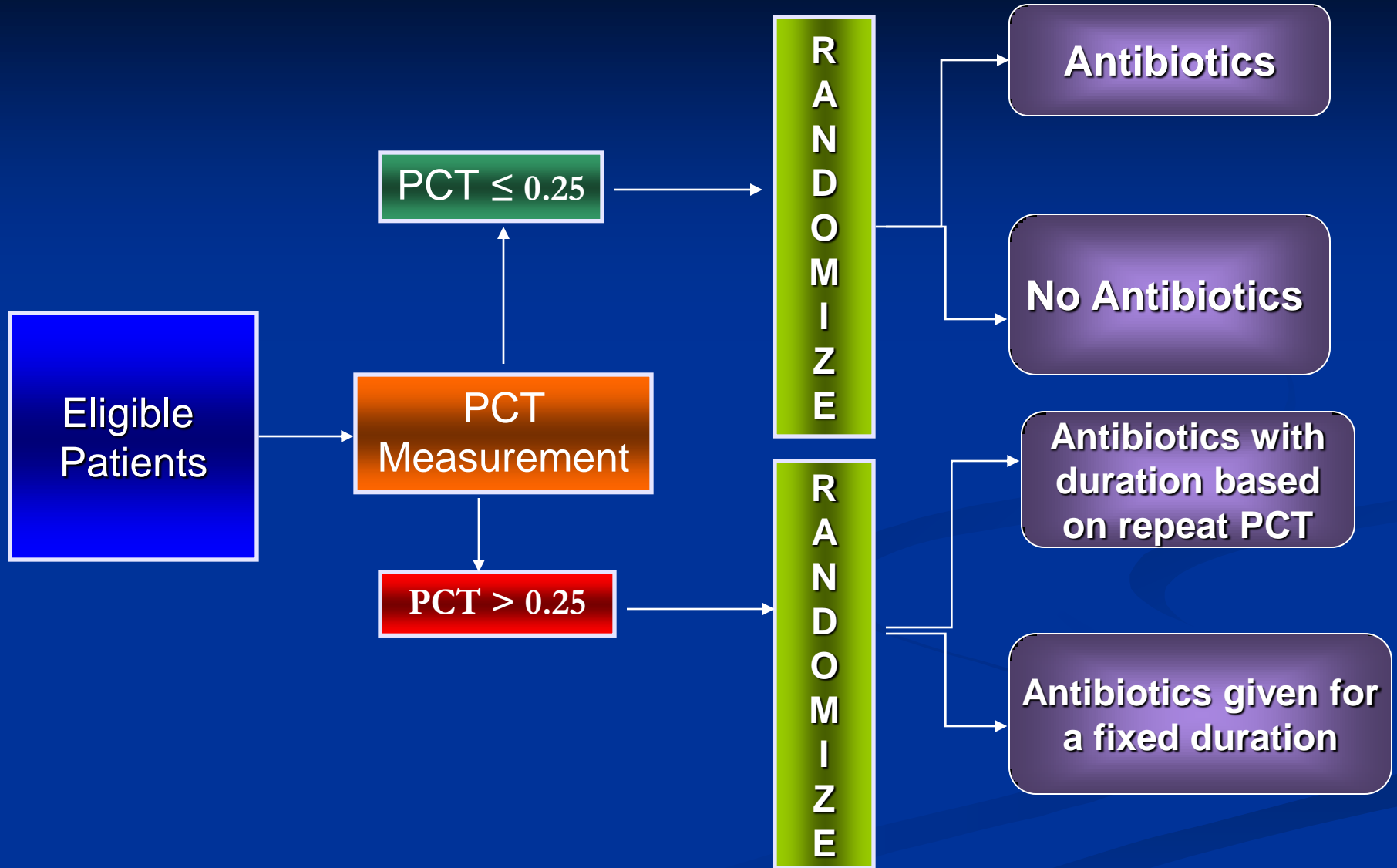
# Trial Design (JAMA article)



**Design features:** Non-inferiority design; composite adverse event outcome within 30 days

**Discussion Points:** 1) PCT not done on all pts, thus no further evaluation possible;  
2) Overlap of pts with same PCT values receiving similar treatment on both arms; thus not adequately powered for non-inferiority  
3) Variable duration of antibiotic therapy, impacting the primary outcome

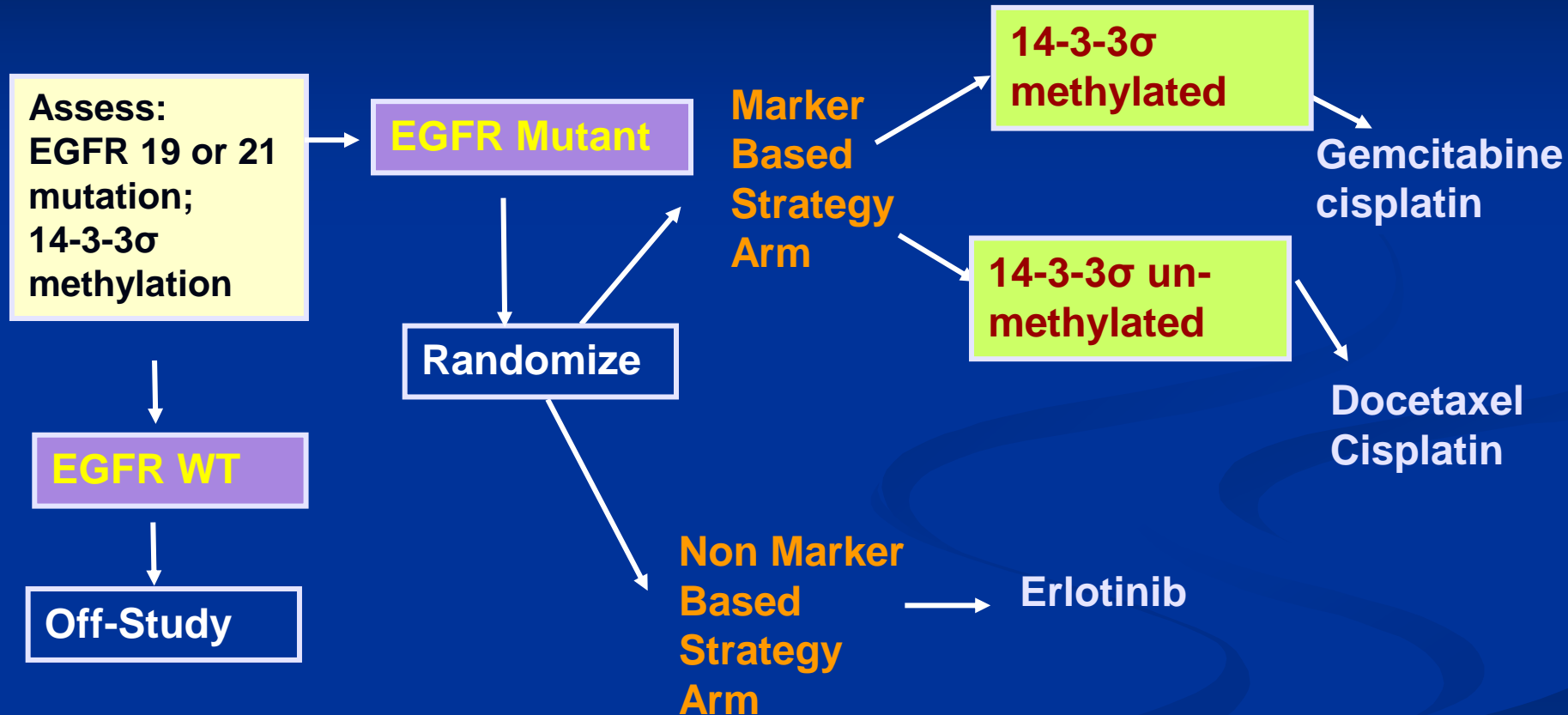
# Proposal for Trial Design



- Test effectiveness of antibiotic use in pts that have low PCT values
- Test effectiveness of the duration of antibiotic use in pts with high initial PCT values

# **COMBINATION DESIGNS: ENRICHMENT FOLLOWED BY STRATEGY**

# Spanish Lung Cancer group (0601)



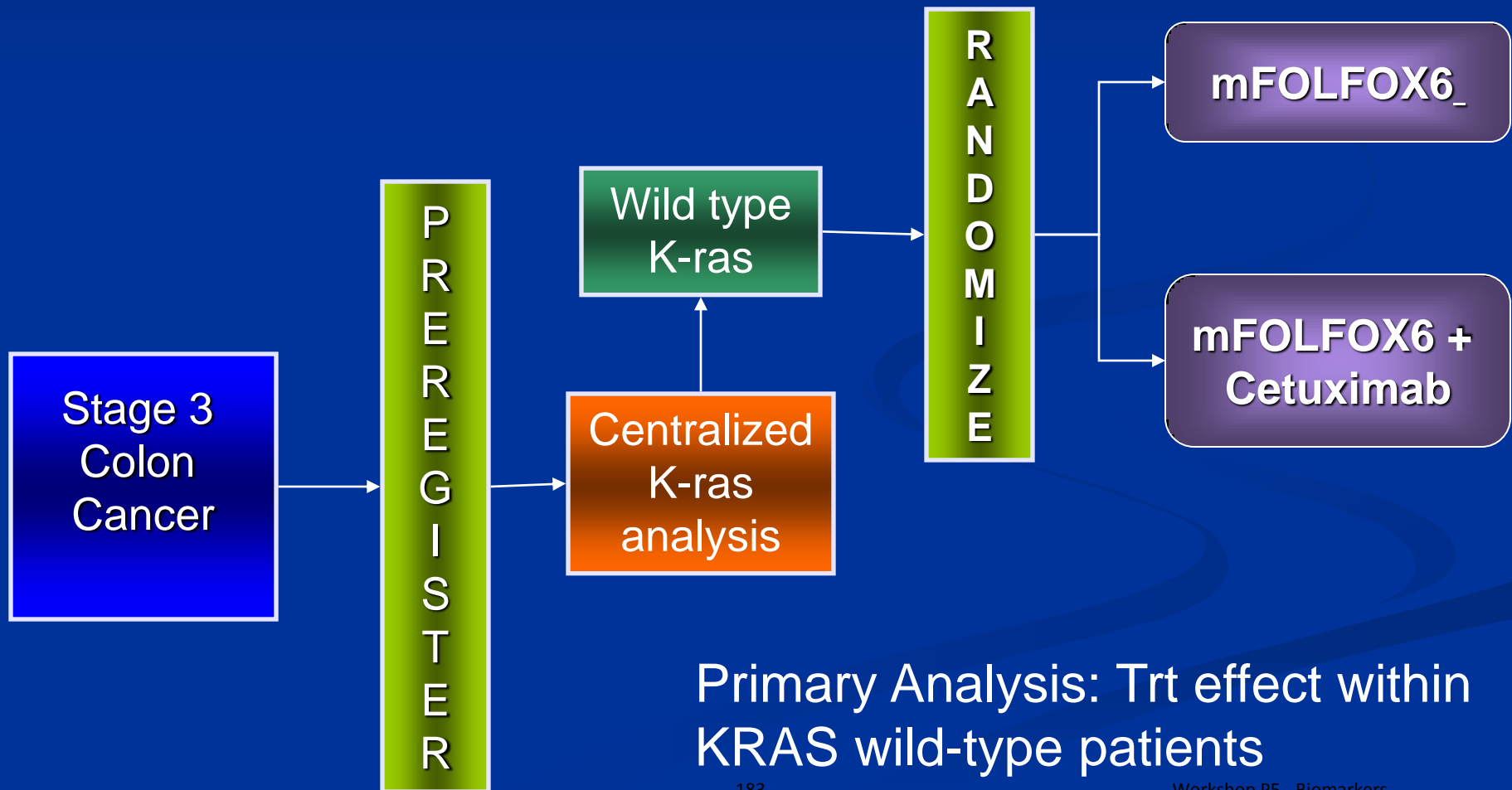
Compare outcomes between the marker based arm and the non-marker based arm

# SEQUENTIAL TESTING STRATEGY DESIGNS

# Closed Testing Procedure Example

## NCCTG Trial N0147

N0147 initially enrolled KRAS WT and mutant patients; modified en-route to randomize only WT patients.

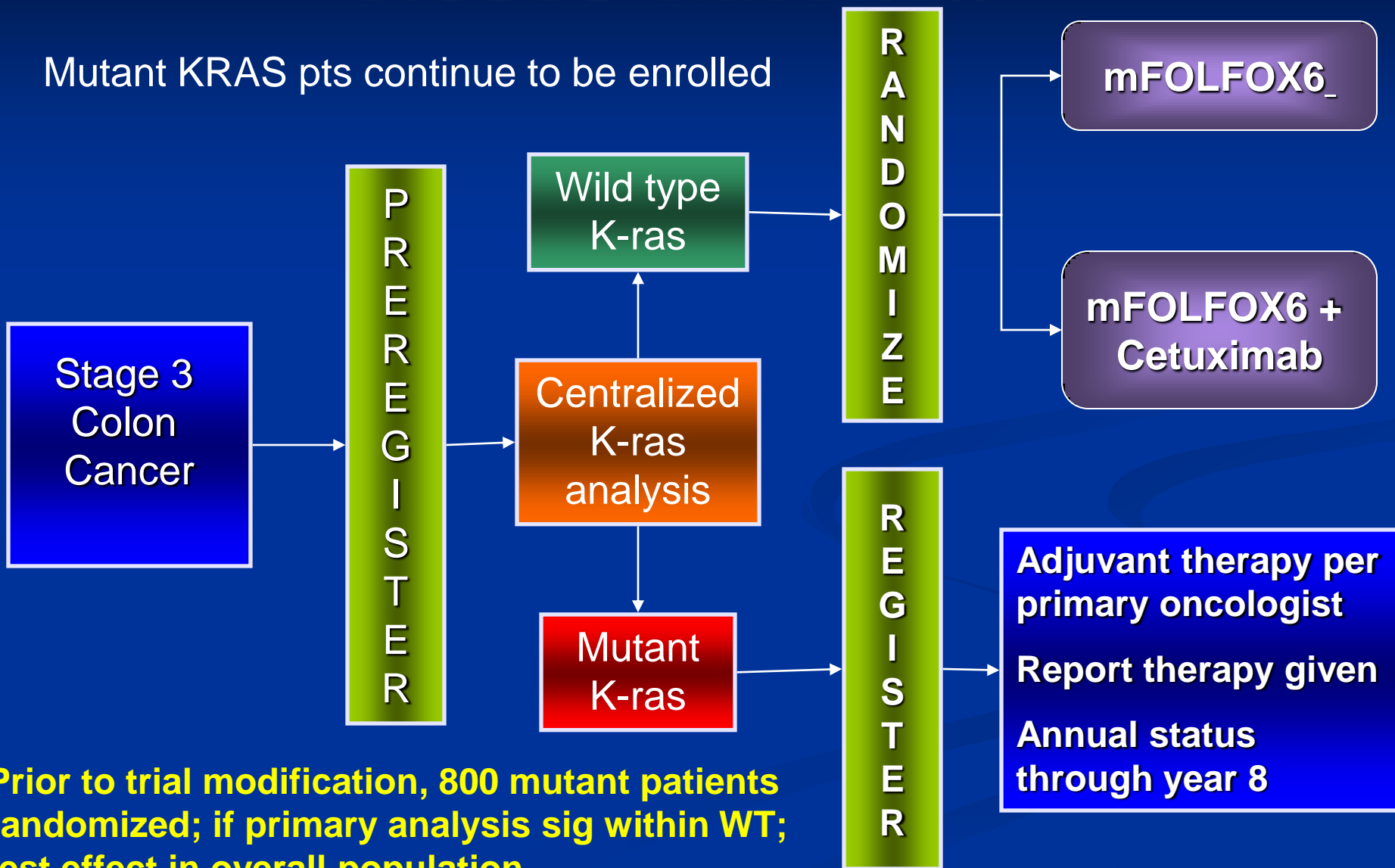




# Closed Testing Procedure Example

## NCCTG Trial N0147

Mutant KRAS pts continue to be enrolled



**Prior to trial modification, 800 mutant patients randomized; if primary analysis sig within WT; test effect in overall population**

# Hybrid Design: SWOG Lung Trial S0819

## S0819:

A Randomized Ph III Study Comparing  
Chemotherapy (Carboplatin/Paclitaxel/(Bevacizumab))  
+/- *Cetuximab*  
in Patients with Advanced  
Non-Small Cell Lung Cancer (NSCLC)

## Hypotheses:

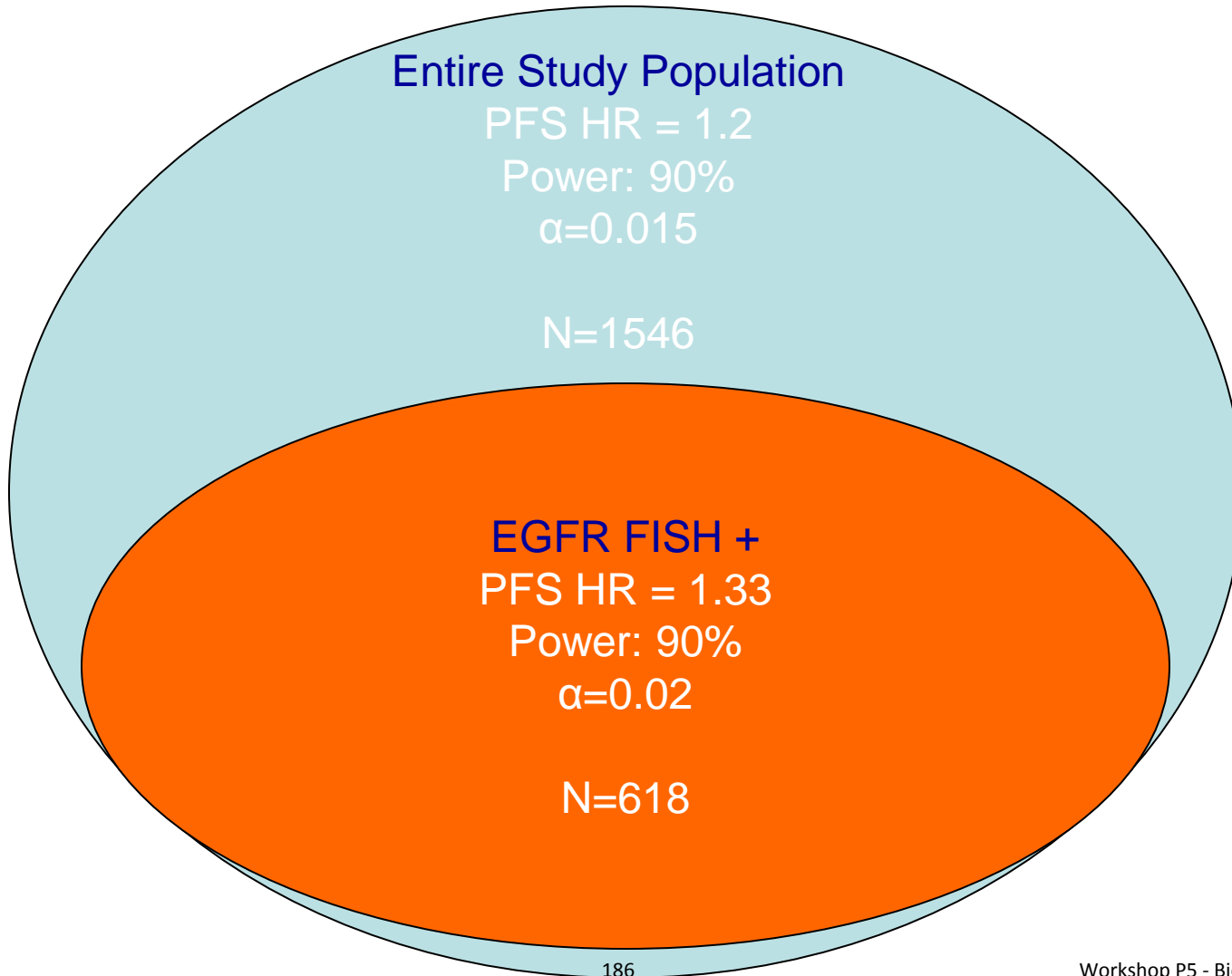
- ❖ *Cetuximab* will increase the efficacy of concurrent chemotherapy in patients with advanced NSCLC.
- ❖ EGFR FISH is a better predictor of benefit than EGFR IHC.

## Questions

➔ Should all NSCLC patients be treated with a targeted agent or should only EGFR FISH positive patients be so treated?

➔ What is the most appropriate trial design to validate the new tumor markers and to determine subgroups of patients most likely to benefit from a new therapy?

# SWOG S0819



# SWOG S0819

Prevalence of FISH+ ~ 50%, power = 92%, overall alpha=.025 (1-sided)

## Hypotheses to be tested:

**H1: Entire cohort:** Addition of Cetuximab increases median PFS by 20%.

**H2: FISH+ cohort:** Addition of Cetuximab increases median PFS by 33%.

**H-strategy:** Strategy of (Chemo+Cetuximab for FISH+ cohort) versus Chemo only for everyone superior: Increase of median PFS in strategy arm by 15%.

Design	N
All Comers Design with split alpha (H1 and H2)	618/1546
All Comers Design (H1 only)	1418
Marker Positive Design (H2 only)	584
Marker Strategy Design	2406

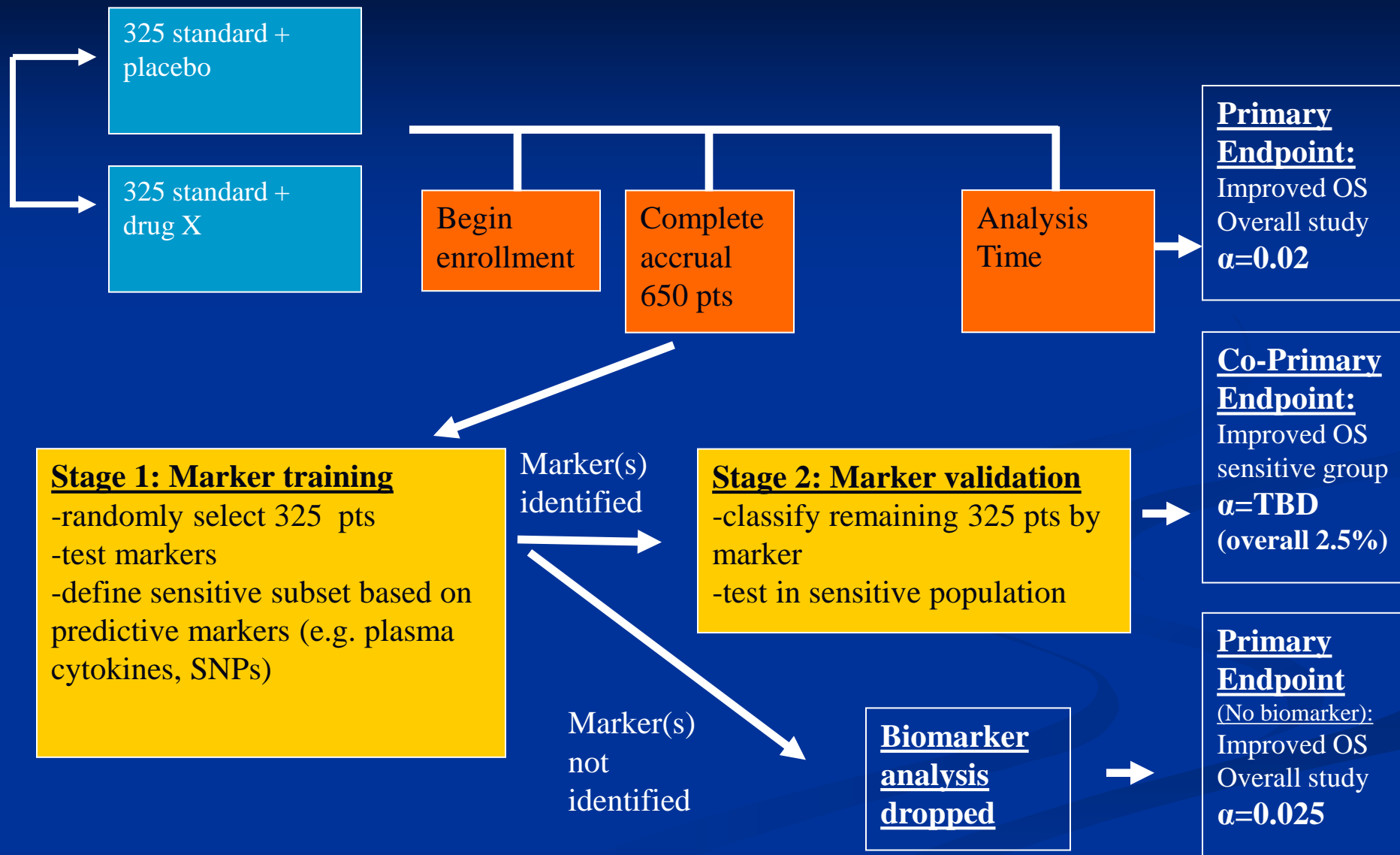
# Adaptive Signature Designs (CVASD and ASD)

## Features:

- Candidate predictive biomarkers known
- Both overall signature and threshold for determining positive/negative unknown
- Eligibility not restricted based on marker status
- ASD/CVASD test for an overall effect as well as prospective development of a signature to identify subsets that benefit most from treatment

Freidlin, Jiang, Simon CCR 2010

# S1114: ASD design example



# Scenario 1: T benefits a small subset, M+

Prevalence of M+: 10%;

Response rates: 25% in control arm; 25% to T in M-

Test	% of times the test is significant	
	M+ Response rate to T: 90%	M+ Response rate to T: 60%
Overall testing: 0.04	26%	11%
Subset testing: 0.01	88%	14%
Overall Power for CVASD	91%	23%
Traditional Design 0.05 level	30%	12%

For smaller treatment effects within M+, CVASD better but not optimal

## Scenario 2: T benefits M+ (40%)

Prevalence of M+: 40%;

Response rates: 25% in control arm; 25% to T in M-

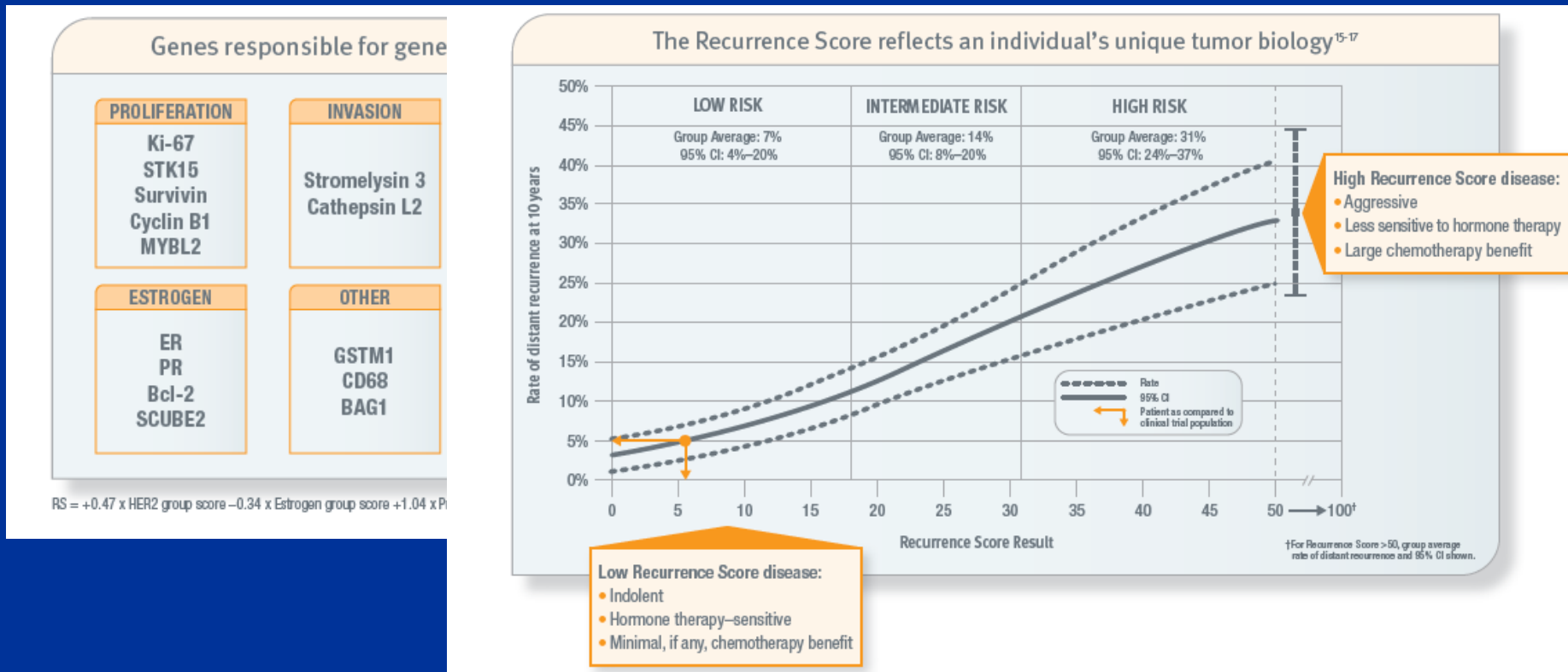
Test	% of times the test is significant	
	M+ Response rate to T: 70%	M+ Response rate to T: 60%
Overall testing: 0.04	96%	83%
Subset testing: 0.01	81%	27%
Overall Power for CVASD	97%	84%
Traditional Design 0.05 level	96%	86%

As the fraction of M+ increases, i.e., treatment is broadly effective - less difference between CVASD and traditional design



# The Oncotype DX Breast Cancer Assay

- A 21 gene expression that provides a Recurrence Score unique to each patient -
  - Predicts chemotherapy benefit, and the 10 year risk of distant recurrence.



## Individualized Recurrence Score® (RS) results assess the potential benefit of chemotherapy and the likelihood of distant breast cancer recurrence

Treatment decisions were changed even when definitive treatment decisions had already been made for these patients\*††

- 33% of the overall population switched from CT + HT to HT alone based on a low Recurrence Score result<sup>‡</sup>
- 4% of the overall population switched from HT only to CT + HT based on a high Recurrence Score result<sup>‡</sup>

Studies have shown that Recurrence Score results reduce chemotherapy use, spare patients the negative health and quality of life impact of unnecessary chemotherapy, and reduce the costs to society and the healthcare system<sup>1,6,9,10</sup>

## Treatment decisions\*†

**37%**  
changed

### The Oncotype DX assay consistently changed treatment decisions across 7 independent studies<sup>§2-8</sup>

\*This analysis included ER-positive

†Patients from this analysis who could not be classified as CT or HT

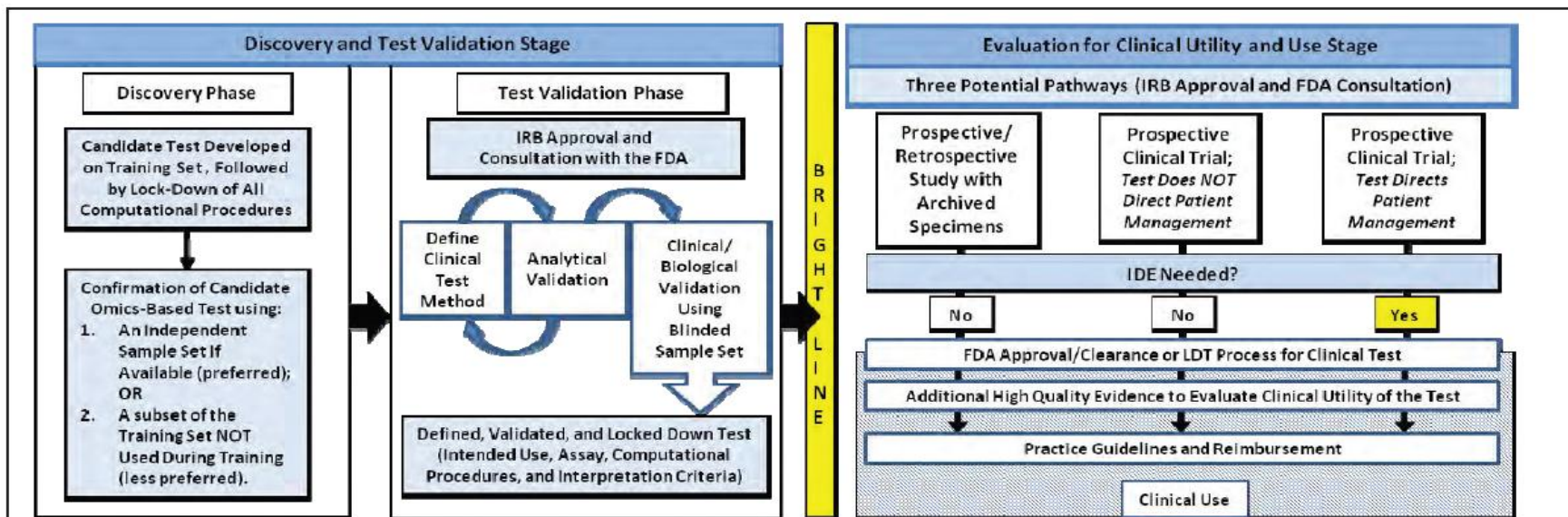
‡CT=chemotherapy

Before RS → After RS	CT+ HT → HT (n=297)	HT → HT (n=303)	CT+ HT → CT+ HT (n=271)	HT → CT+ HT (n=41)
(TOTAL=912)	(n=600)		(n=312)	
Asad et al <sup>2</sup> (n=81)	36	13	24	8
Henry et al <sup>3</sup> (n=29)	7	14	6	2
Kiang et al <sup>4</sup> (n=313)	105	119	69	20
Liang et al <sup>5</sup> (n=260)	85	47	125	3
Lo et al <sup>6</sup> (n=83)	20	40	20	3
Oratz et al <sup>7</sup> (n=68)	14	32	19	3
Thanasoulis et al <sup>8</sup> (n=78)	30	38	8	2
<b>% with changed treatment decisions</b>	<b>33%</b>			<b>4%</b>

§All studies except Henry et al had statistically significant differences in CT recommendation before and after Recurrence Score testing.

# Closing Comments

- Choice of trial design depends on
  - Biological rationale
  - Marker prevalence
  - Assay performance
  - Strength of preliminary evidence
  - Incremental benefit of marker-based selection
- An optimal design can help to predict which patient is likely to benefit from a treatment and/or requires intensive treatment. This helps to:
  - Improve the success rate of clinical drug development
  - Bring down trial costs in terms of patients and resources
  - Prevents patients from being exposed to toxic treatments that may not benefit them.



**FIGURE S-1 Omics-based test development process.** In the first stage of omics-based test development, there are two phases: discovery and test validation. In the discovery phase, a candidate test is developed and confirmed. The fully specified computational procedures are locked down in the discovery phase and should remain unchanged in all subsequent development steps. Ideally, confirmation should take place on an independent sample set. Under exceptional circumstances it may be necessary to move into the test validation phase without first confirming the candidate test on an independent sample set if using an independent test set in the discovery phase is not possible, but this increases the risk of test failure in the validation phase. In the test validation phase, the omics-based test undergoes analytical and clinical/biological validation. The bright line signifies the point in test development where a fully defined, validated, and locked down clinical test (analytical and clinical/biological validation) is necessary. Changes to the test after the bright line is crossed require a return to the test validation phase, approval by the Institutional Review Board, and possibly consultation with the Food and Drug Administration. In the second stage of test development, the fully defined, validated, and locked down omics-based test undergoes evaluation for its intended clinical use. Evaluation of clinical utility and use is a process that often continues after initial adoption into clinical use. Statistics and bioinformatics validation occurs throughout the discovery and test validation stage as well as the stage of evaluation for clinical utility and use.

NOTE: FDA = Food and Drug Administration, IDE = investigational device exemption, IRB = Institutional Review Board, LDT = laboratory-developed test.

A diagram from the report's executive summary, detailing the two-stage omics test validation process.

**Although molecular profiling is expensive, not doing so is ultimately far more expensive and gives the wrong answer**

**(Stewart et al., JCO 2010)**

**Treating “unselected” populations with regimens that benefit only a subset of patients is less economically sustainable with expensive molecularly targeted therapeutics**

# Acknowledgements

- Dan Sargent, Mayo Clinic
- John Crowley, CRAB

# Thank You!