Some Considerations for Statistical Design, Analysis and Interpretation for Biomarker Classifier Based Clinical Trials in Establishing Efficacy in Support of Regulatory Marketing and Promotional Claims

Robert T. O'Neill Ph.D.

Director, Office of Biostatistics
Office of Translational Sciences, CDER, FDA

ODAC, Dec. 16, 2008

Outline of issues (1)

- Definitions of predictive / prognostic markers
- General principles for design and analysis of a clinical trial for subgroup differences
  - Control of the false positive/negative conclusions
  - Subgroup defined by pre-treatment baseline factor ascertained on all subjects (issue of randomization)
- Issues for a prospective/retrospective genomic clinical trial

Outline (2)

- The principle of replication to confirm hypotheses and lack thereof
- Examination of the KRAS studies -consistency, limitations and uncertainties that exist , P-mab vs C-mab
- Prospective study designs to evaluate a biomarker classified group do exist
- Consideration for levels of evidence needed for a biomarker subgroup

FDA comments to sponsors
Two options

- Optimal Approach: Conduct an adequate and well-controlled trial, prospectively designed to assess efficacy in subgroups based on KRAS testing by a validated assay.

- Pragmatic Approach: A retrospective analysis could be considered under the following conditions:

1. Adequate, well-conducted and well-controlled trial
2. Large sample size (approximate random allocation of factors not used as stratification variables for randomization, i.e., KRAS status)
3. KRAS biomarker status ascertained in a large portion of randomized subjects (> 90%)
4. Assay - acceptable analytical performance
5. Acceptable analysis plan

Definitions : prognostic of outcome / predictive of treatment effect

- Prognostic (is independent of treatment status)
  - A marker (or classifier) for which the magnitude of the event or outcome rate is related (eg. higher or lower)
- Predictive (depends upon treatment status and is relative to the control group)
  - A marker (or classifier) for which the magnitude of the *treatment effect* is related (eg. higher or lower)

Definition of Treatment Effect

Genomic Status*					Scenario C	
					Control	Drug C
g–					39%	49%
g+					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, O'Neill, Hung (2007, PS)

Performance of assays for marker classification

- What are the minimum performance characteristics (e.g., sensitivity, specificity, reproducibility) of the assay used to classify patient subgroups and what are the consequences of that performance for making correct inferences from the study
  - KRAS vs EGFR vs breast cancer assay
- In general, 'classifer' performance and marker prevalence (mix) may explain study to study heterogeneity and differences in results

Components of a good analysis plan for a retrospective evaluation

- Role of randomization to assure unbiased and fair comparisons
- Role of marker status classification - impact of convenience samples on biased estimates
- Marker classification performance
- Statistical control of false positive conclusions - how many hypotheses, which were primary, which failed
  - Accounting for multiplicity - how many outcomes, OS,PFS,RR
- Data to generate the hypothesis vs. data to confirm the hypothesis

General Strategy in Clinical Trials
Look for an overall treatment effect on the primary outcome in the intent to treat population (*everyone* randomized)

- If a statistically persuasive result, usually p value < 0.025 one-sided
- THEN, examine subgroups -options
  - Look at marker negative group
  - Look at marker positive group
  - Examine evidence for equal treatment effect in each groups
  - Examine evidence for differential treatment effect in each group - the 'interaction' test
- If no statistical significance on primary endpoint (hypothesis) - everything further is exploratory

If one changes the number, set or sequence of hypotheses of interest after start of the study

- Not an acceptable practice after observing the data
- Adaptive designs try to pre-specify this sequence

Prospective - Retrospective Study
What is it and
What are the concerns

- The classification factor is not known at the time of study initiation, and the study is, at first, not analyzed with that factor as part of the hypothesis (retrospective aspect)
- The initial hypothesis and endpoints for the study are not changed, except if pre-specified as part of a planned adaptive study design
- The control of the false positive conclusion from the study are appropriately dealt with
- The randomization is not stratified on a factor that itself is of interest as one of the hypotheses to be tested
- The factor of interest is ascertained at baseline on all subjects randomized to treatment groups - what if not ?

Prospective/Retrospective

Working Definition – In completed or post-interim-analysis trial where genomic samples were collected prior to treatment initiation, whether or not full ascertainment, the genomic hypothesis is '**prospectively specified**' prior to diagnostic assay testing. However, the clinical outcome data **without genomic information** have already been (partially) collected, unblinded, and analyzed. The genomic data analysis might be arguably 'prospectively' performed, which is a **retrospective analysis**.

Convenience Genomic Sample – The genomic samples may not be obtained on all randomized subjects if some refuse (often optional consent in current practice).

* Wang et al (2006 TPJ)

Part of a good statistical analysis plan - Control of false conclusions and of the multiplicity of hypotheses that may be tested

- The chance of erroneously concluding that there is a real treatment effect when in fact it is not true, or the chance of concluding there is no treatment effect when in fact one actually exists, are two critical concerns for the design and interpretation of study results of any clinical trial.
- Eg. endpoints (PFS,OS,RR); subgroups

Part of a good analysis plan: The importance of randomization on the full study entry population for marker subpopulation valid comparisons

- The role of randomization to assure comparable comparisons between two or more treatment groups , especially in small sample size subpopulation identified after completion of a clinical study.

CA 225006 (EPIC) Second-line

- cetuximab + irinotecan vs. irinotecan
  - EGFR-Positive Colorectal Carcinoma
- Evaluable KRAS samples from 300 of 1298 patients (23%)
  - "It should be noted that these reflect the US study population only, as it was not possible to collect tumor samples from other investigational sites in this study."
- Primary Endpoint - overall survival (OS)

Dramatic Difference Between ITT and Convenience Sample

- Primary Endpoint - overall survival (OS)
  - Overall: HR=0.98 (95% CI 0.85, 1.11) (1298 patients)
  - Evaluable patients: HR = 1.25 (95% CI 0.95, 1.66) (300 patients)

Inconsistency between endpoints in the convenience sample

- OS
  - WT KRAS HR = 1.29 (95% CI 0.89, 1.85) (192 patients)
  - Mutant KRAS HR = 1.28 (95% CI 0.81, 2.01) (108 patients)
- PFS
  - WT KRAS HR = 0.77 (95% CI 0.57, 1.04) (192 patients)
  - Mutant KRAS HR = 1.00 (95% CI 0.67, 1.49) (108 patients)
- Best Response
  - WT KRAS 10% vs. 7%, p = 0.61 (192 patients)
  - Mutant KRAS 12% vs. 5%, p = 0.29 (108 patients)

What evidence is there in the completed studies (C-mab, P-mab) based on the study reports

- Six randomized studies for which marker status is available only on a selected subset of the randomized study population, ranging from 23% to 92%
  - No documented evidence that the treatment groups with ascertained marker status are comparable for baseline variables
  - The P-mab study has the largest proportion of marker ascertainment

What do we need to know for a marker to be predictive of treatment effect (relative change in response)

- An unbiased comparison between the test treatment and control in each of the marker subgroups
  - Unbiased generally requires a randomized subset of subjects in each of the marker categories, not a convenience sample of subjects with marker status available

Cetuximab Trials

Clinical Trial	Line	Add'al Therapy	1° Endpt	Met 1° Endpoint	ITT Patients Tested for KRAS			Assay
					n	ITT	% of ITT	
CRYSTAL EMR 62202-013	1st	FOLFIRI	PFS	YES p = 0.048	540	1198	45	PCR based
NCIC-017 CA225025	3rd	BSC	OS	YES p = 0.005	394	572	69	sequencing
EPIC CA225006	2nd	irinotecan	OS	NO p = 0.71	300	1298	23	sequencing
OPUS EMR 62 202-047	1st	FOLFOX	RR	NO p = 0.06	233	337	69	PCR based

Panitumumab Trials

Clinical Trial	Line	Add'nl Therapy	1° Endpt	Met 1° Endpoint	ITT Patients Tested for KRAS			Assay
					n	ITT	% of ITT	
20020408	3rd	BSC	PFS	YES p < 0.0001	427	463	92	PCR based
PACCE 20040249	1st	chemo/ bev	PFS	NO Inferior P = 0.002	863	1053	82	PCR based

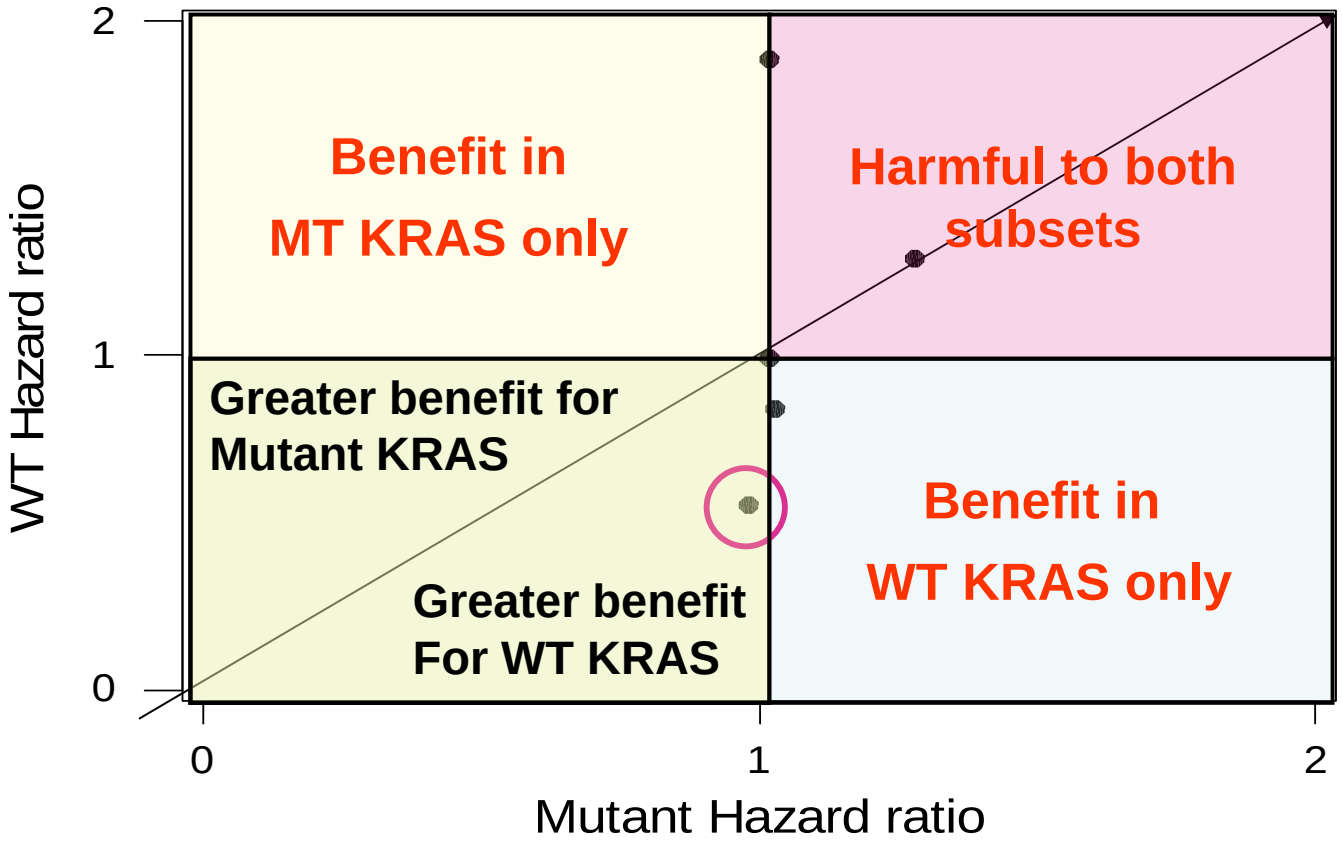
The evidence from the six studies is not as consistent as you might think

- Consider consistency - what might that mean
  - Line of therapy
  - Control group
  - Endpoints used as primary: PFS,OS, RR
  - Convenience samples
  - EPIC - US results may be very different

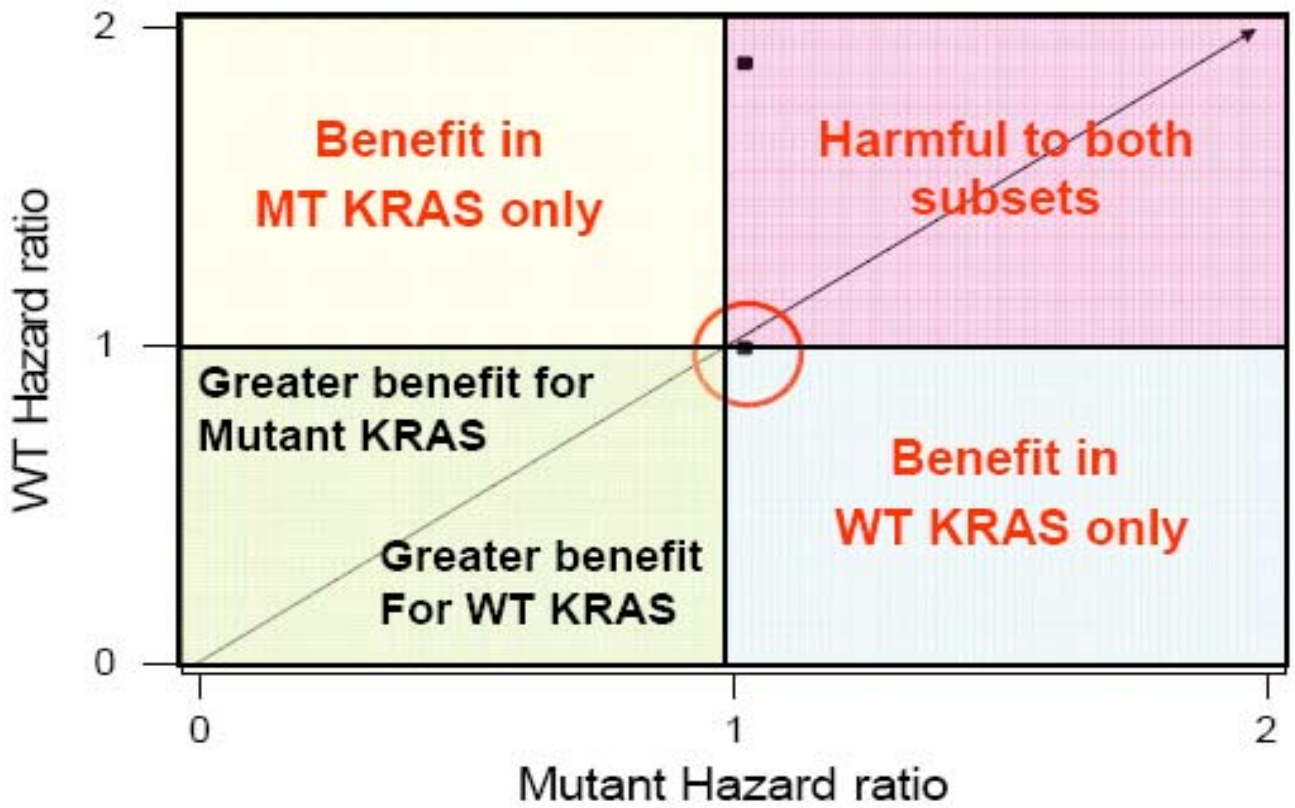
Overall survival

- The following graph provides a summary of overall survival for the five studies having overall survival comparisons for the WT KRAS and mutant KRAS subgroups
- Hazard ratios are used for overall survival.
- Points above the line correspond to larger effects for Cetuximab or Panitumumab for the mutant KRAS "subgroup" than for the wild-type "subgroup"
- Points below the line correspond to larger effects for Cetuximab or Panitumumab for the wild-type "subgroup" than for the mutant KRAS "subgroup"

Overall Survival: three trials showed no benefit or harmful effect to both subsets; only the circled trial shows clear benefit in WT KRAS patients only



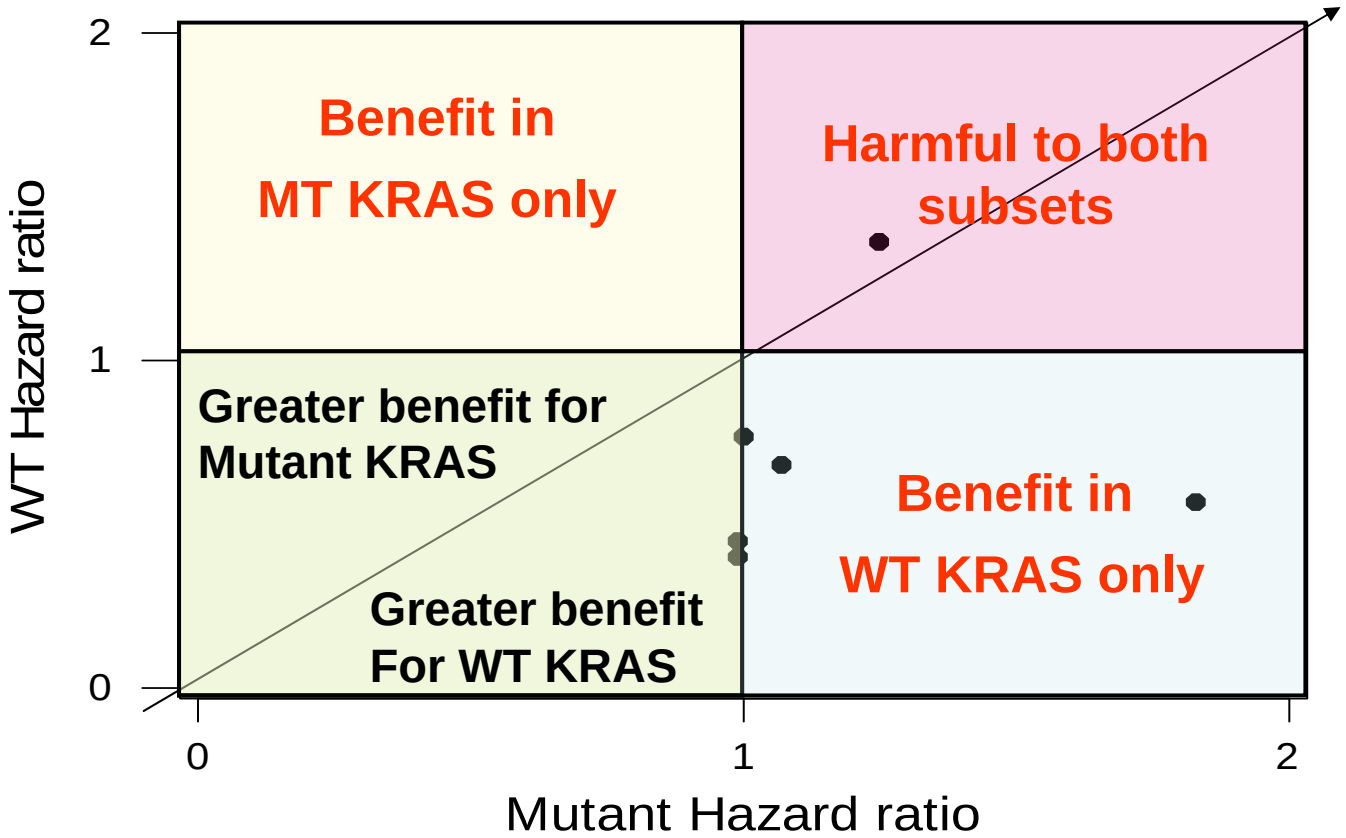
OS – P-mab studies: No benefit in MT KRAS patients; No or Harmful effect in WT KRAS patients



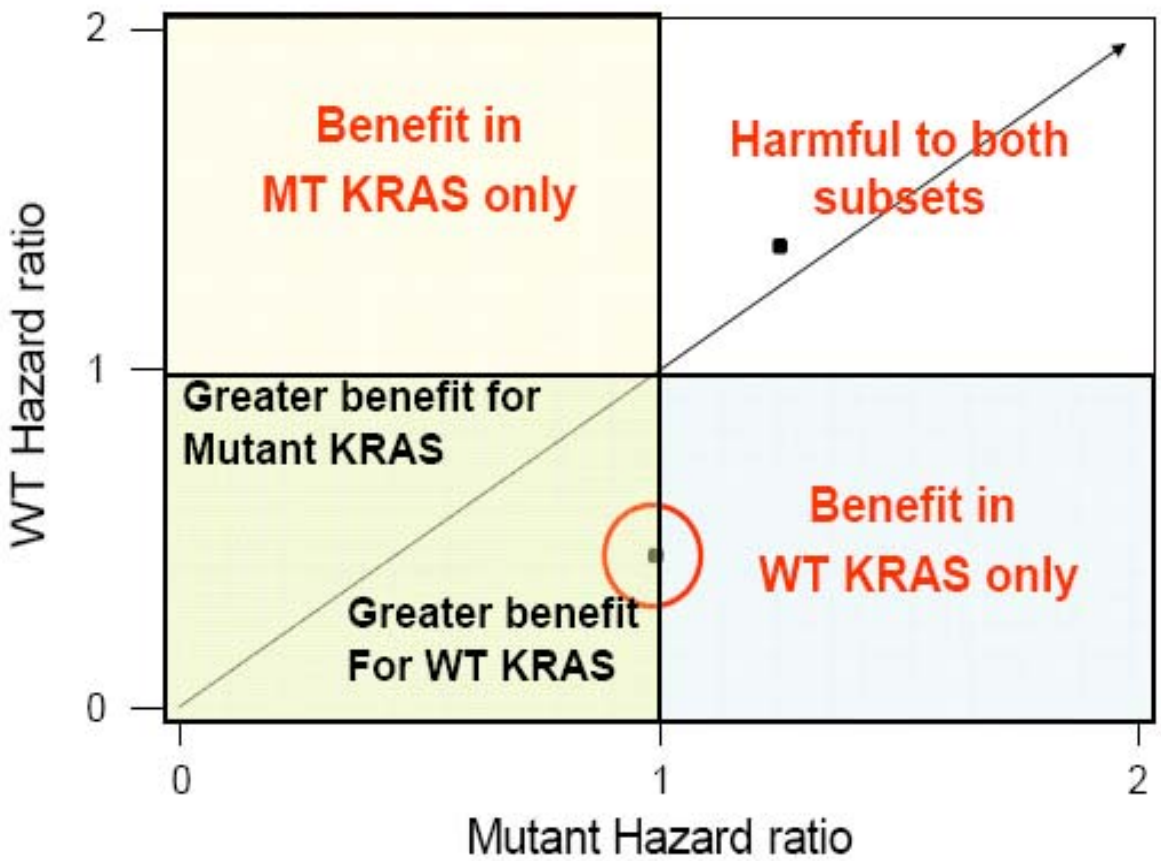
Progression-free survival

- The following graph provides a summary for the six studies of the progression-free survival comparisons for the WT KRAS and mutant KRAS subgroups
- Hazard ratios are used for progression-free survival.
- Points above the line correspond to larger effects for Cetuximab or Panitumumab for the mutant KRAS "subgroup" than for the wild-type "subgroup"
- Points below the line correspond to larger effects for Cetuximab or Panitumumab for the wild-type "subgroup" than for the mutant KRAS "subgroup"

PFS: five trials show benefit in WT KRAS only, one trial shows harmful effect to both subsets



PFS – P-mab studies: harmful effect to both subsets in one trial; benefit in WT KRAS only in one trial



Clinical study designs do exist that can prospectively address the marker predictor hypothesis

Prospective study design options

- A two stage design that reserves some type 1 error for testing a subgroup yet to be specified - (biological plausibility)
  - Fixed study design with no adaptation to increase samples size overall or in subgroups
- An adaptive study design that can increase sample size and pre-specifies the 'win criteria' or study 'success' criteria
- Test the efficacy of a strategy that screens for the classifier

Trial to demonstrate minimizing risk

The Abacavir 'PREDICT -1' trial

N Engl J Med 2008;358:568-79.

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

HLA-B*5701 Screening for Hypersensitivity to Abacavir

Simon Mallal, M.B., B.S., Elizabeth Phillips, M.D., Giampiero Carosi, M.D., Jean-Michel Molina, M.D., Cassy Workman, M.B., B.S., Janez Tomažič, M.D., Eva Jägel-Guedes, M.D., Sorin Rugina, M.D., Oleg Kozyrev, M.D., Juan Flores Cid, M.D., Phillip Hay, M.B., B.S., David Nolan, M.B., B.S., Sara Hughes, M.Sc., Arlene Hughes, Ph.D., Susanna Ryan, Ph.D., Nicholas Fitch, Ph.D., Daren Thorborn, Ph.D., and Alastair Benbow, M.B., B.S., for the PREDICT-1 Study Team*

- Same treatment in both randomized groups
- Treatment groups differ by screen strategy and entrance criteria into the trial
  - HLA-B*5701 screening - exclude positive subjects in one of the randomized arms
  - Goal: demonstrate screening reduces incidence of serious adverse event
  - Provides estimates of SE and SP for the classifier

The NEW ENGLAND JOURNAL of MEDICINE

N Engl J Med 2008;358:568-79.

ORIGINAL ARTICLE

HLA-B*5701 Screening for Hypersensitivity to Abacavir

Simon Mallal, M.B., B.S., Elizabeth Phillips, M.D., Giampiero Carosi, M.D., Jean-Michel Molina, M.D., Cassy Workman, M.B., B.S., Janez Tomažič, M.D., Eva Jägel-Guedes, M.D., Sorin Rugina, M.D., Oleg Kozyrev, M.D., Juan Flores Cid, M.D., Phillip Hay, M.B., B.S., David Nolan, M.B., B.S., Sara Hughes, M.Sc., Arlene Hughes, Ph.D., Susanna Ryan, Ph.D., Nicholas Fitch, Ph.D., Daren Thorborn, Ph.D., and Alastair Benbow, M.B., B.S., for the PREDICT-1 Study Team*

METHODS

This double-blind, prospective, randomized study involved 1956 patients from 19 countries, who were infected with human immunodeficiency virus type 1 and who had not previously received abacavir. We randomly assigned patients to undergo prospective HLA-B*5701 screening, with exclusion of HLA-B*5701–positive patients from abacavir treatment (prospective-screening group), or to undergo a standard-of-care approach of abacavir use without prospective HLA-B*5701 screening (control group). All patients who started abacavir were observed for 6 weeks. To immunologically confirm, and enhance the specificity of, the clinical diagnosis of hypersensitivity reaction to abacavir, we performed epicutaneous patch testing with the use of abacavir.

The study confirmed the hypothesis that screening will reduce severe adverse reaction

Table 2. Incidence of Hypersensitivity Reaction to Abacavir.*

Hypersensitivity Reaction	Prospective Screening	Control	Odds Ratio (95% CI)*	P Value
	no. of patients/total no. (%)			
Clinically diagnosed				
Total population that could be evaluated	27/803 (3.4)	66/847 (7.8)	0.40 (0.25–0.62)	P<0.001
White subgroup	24/679 (3.5)	61/718 (8.5)	0.38 (0.23–0.62)	P<0.001
Immunologically confirmed				
Total population that could be evaluated	0/802	23/842 (2.7)	0.03 (0.00–0.18)	P<0.001
White subgroup	0/679	22/713 (3.1)	0.03 (0.00–0.19)	P<0.001

* P values, odds ratios, and 95% confidence intervals (CIs) were calculated by means of logistic-regression analysis and adjusted for self-reported race (white vs. nonwhite), history of receipt of antiretroviral therapy (none vs. any), introduction of a new nonnucleoside reverse-transcriptase inhibitor (yes or no), and concurrent use or nonuse of a protease inhibitor. The model-based incidences of clinically diagnosed hypersensitivity reaction to abacavir in the total population that could be evaluated and in the white subgroup were 3.3% and 3.5%, respectively, for the prospective-screening group and 7.9% and 8.6%, respectively, for the control group. The white subgroup included the two and three patients reporting both categories of white ancestry in the prospective-screening group and the control group, respectively. The odds ratios for immunologically confirmed hypersensitivity reaction to abacavir were obtained by means of exact methods, owing to the absence of immunologically confirmed hypersensitivity reaction in the prospective-screening group. The model (involving a median, unbiased estimate of the odds ratio) estimated the odds of hypersensitivity reaction in the prospective-screening group versus the control group to be 1:33 ($1 \div 0.03 = 33$). (Although a simple point estimate of the odds ratio from the raw data yields a more intuitive value of 0, it also implies an infinite reduction in the odds, which is problematic for linear regression modeling in that it introduces error from division by 0.)

The study design also provided estimates of SE and SP

Table 4. Performance Characteristics of HLA-B*5701 Screening for Hypersensitivity Reaction to Abacavir in the Control Group.*

Subgroup	Positive for HLA-B*5701	Negative for HLA-B*5701	Total	Performance Characteristic for Hypersensitivity Reaction
		number of patients		*percent (95% CI)*
Clinically diagnosed hypersensitivity reaction				
Total population that could be evaluated				
Hypersensitivity reaction	30	36	66	Sensitivity: 45.5 (33.1–58.2)
No hypersensitivity reaction	19	762	781	Specificity: 97.6 (96.2–98.5) PPV: 61.2 (46.2–74.8) NPV: 95.5 (93.8–96.8)
White subgroup				
Hypersensitivity reaction	29	32	61	Sensitivity: 47.5 (34.6–60.7)
No hypersensitivity reaction	19	638	657	Specificity: 97.1 (95.5–98.3) PPV: 60.4 (45.3–74.2) NPV: 95.2 (93.3–96.7)
Immunologically confirmed hypersensitivity reaction				
Total population that could be evaluated				
Hypersensitivity reaction	23	0	23	Sensitivity: 100 (85.2–100)
No hypersensitivity reaction	25	794	819	Specificity: 96.9 (95.5–98.0) PPV: 47.9 (33.3–62.8) NPV: 100 (99.5–100)
White subgroup				
Hypersensitivity reaction	22	0	22	Sensitivity: 100 (84.6–100)
No hypersensitivity reaction	25	666	691	Specificity: 96.4 (94.7–97.6) PPV: 46.8 (32.1–61.9) NPV: 100 (99.4–100)

* The white subgroup included the two and three patients reporting both categories of white ancestry in the prospective-screening group and the control group, respectively. NPV denotes negative predictive value, and PPV positive predictive value.

Other designs might be more efficient (less subjects)

PHARMACEUTICAL STATISTICS
Pharmaceut. Statist. (2007)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/pst.300

WILEY InterScience®
DISCOVER SOMETHING GREAT

Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset[‡,§]

MAIN PAPER

Sue-Jane Wang[1,*,†], Robert T. O'Neill[1] and H. M. James Hung[2]

[1] *Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD, USA*

[2] *Division of Biometrics I/OB, Office of Translational Sciences, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD, USA*

STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:329–339
Published online 18 November 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.1975

On the efficiency of targeted clinical trials

A. Maitournam and R. Simon*,†

Perspective

Clinical Cancer Research *Vol. 10, 6759–6763, October 15, 2004*

Evaluating the Efficiency of Targeted Designs for Randomized Clinical Trials

Richard Simon and Aboubakar Maitournam
Biometric Research Branch, National Cancer Institute, Bethesda, Maryland

these targeted designs. As discussed in this article, w
the efficiency of targeted designs in comparison with
randomized designs with broader eligibility criteria.
ated efficiency in the context of a binary outcome

The data and the studies

- FDA does not have access to the KRAS level data
- We use these data for educational purpose and for illustration of issues that can occur
- Concordance or consistency of results across studies is confounded by multiple factors
  - 1st,2nd ,3rd line, endpoints, convenience sample

Scientific principle of Replication / Confirmation for evidence

- The likelihood that a subgroup effect is real
- The likelihood that the statistical evidence for reproducing a treatment effect identified in a subpopulation in a single clinical trial can be demonstrated in another independent study
- P-mab has two studies with potentially conflicting evidence - not initially designed with the hypothesis of interest
- C-mab has no studies prospectively planned with this hypothesis and all with convenience samples

Subgroup finding have always been looked at cautiously

The NEW ENGLAND JOURNAL of MEDICINE

SPECIAL REPORT

Statistics in Medicine — Reporting of Subgroup Analyses in Clinical Trials

Rui Wang, M.S., Stephen W. Lagakos, Ph.D., James H. Ware, Ph.D., David J. Hunter, M.B., B.S., and Jeffrey M. Drazen, M.D.

N ENGL J MED 357;21 WWW.NEJM.ORG NOVEMBER 22, 2007

- Concern has usually been that subgroup findings are exploratory at best or false positives unless further evidence available (strength of prior evidence)
- Strategy has been to:
  - adjust for multiple analyses, test for interactions, confirm with another study with a sufficient sample size since subgroups are usually small in size

Trials whose findings were reversed upon completion of a second study planned to specifically test the hypothesis generated in the first study

Perspectives

The Fragility of Cardiovascular Clinical Trial Results

LEMUEL A. MOYÉ, MD, PhD,* ANITA DESWAL, MD†

Journal of Cardiac Failure Vol. 8 No. 4 2002

ABSTRACT

Background: Clinical trials that have their prospective analysis plan altered are difficult to interpret.

Why ?

The analysis plan , in the initial study, changed after seeing the data: it placed new emphasis on a subgroup finding, or on a secondary endpoint raised in prominence leading to false discoveries that were not replicated

3 Examples: Vesnarinone, Amlodipine, Losartan

PRAISE 1

Advanced heart failure patients

PRAISE (Prospective Randomized Amlodipine Survival Evaluation): a randomized, placebo-controlled trial with more than 1100 cardiomyopathy patients, randomized to amlodipine (a calcium channel blocker), or placebo.

The primary endpoint of the study (all-cause mortality or cardiovascular morbidity) was not significantly different between the 2 groups.

Unexpected results

Amlodipine was beneficial for patients with non-ischemic cardiomyopathy — a 31% risk reduction (95% CI 2%–51%; p=0.04) in the primary endpoint.

All-cause mortality (a secondary endpoint) was reduced by 46% (95% CI 21%–63%; p<0.001), a benefit not seen in the ischemic heart disease patients receiving amlodipine.

PRAISE 2

Conclusions

The results of PRAISE 2 do not confirm the survival benefit of amlodipine seen in non-ischemic cardiomyopathy in the PRAISE 1 trial.

The favorable survival benefit of amlodipine seen in PRAISE 1 was likely due to chance, despite the fact that mortality is an unequivocal endpoint; the benefit was seen in a prespecified subgroup and the p value for the subgroup was very small.

The combined results of PRAISE 1 and PRAISE 2 indicate that long-term treatment with amlodipine is neither beneficial nor harmful in patients with severe chronic heart failure.

The results of PRAISE 2 emphasize the need for replication, even when the results define a mortality benefit and are associated with low p values.

PRAISE 1

Trial results

The study results were considered by many to be highly credible because the reduction in mortality was seen in a prospectively defined subgroup and the p value was very small (<0.001), in addition to the fact that mortality is an unbiased and incontrovertible endpoint.

Critics argued that although prospectively defined, the subgroup was small (only 119 patients) and all-cause mortality was a secondary endpoint.

Moreover, no plausible mechanisms could explain the risk reduction in patients with non-ischemic cardiomyopathy; risk reduction had been expected in the ischemic heart disease population.

Back to the questions

- The adequacy of analysis plans and analyses to data
- The available studies
- The limitations to conclude a marker predictive of treatment benefit

How much evidence is needed to establish or support a predictive marker claim ?

- A collection of observational associations external to any study that is used to interpret effects in a separate study that is not initially designed for that purpose - is it real and is it repeatable
- Two independent studies, both of which are prospectively designed to test the marker hypothesis
- The strength of the statistical evidence
  - Effect size, consistency across studies, sufficient sample size, randomization, control for false positive conclusions (multiplicity)