Panel 2 – Blinded Independent Central Review of PFS Endpoints

Dan Sargent, Ph.D
Mayo Clinic
September 14, 2009

Phase III Clinical Trial time to event endpoints

- Overall Survival – time from study entry to death
  - *Unambiguous*
  - *Easily assessed*
  - *Incontrovertibly important, a clear clinical benefit endpoint*
- Progression Free Survival (PFS) – time from study entry to first of disease prog or death

Merits of PFS as an endpoint

- Un-encumbered by cross-over
- Available more quickly than OS
- Variable demonstration of surrogacy for OS
  - *Colon – Yes – Buyse JCO 2007*
  - *Breast – No – Burzykowski, JCO 2008*
  - *Lung – Unclear – Buyse ASCO 2008*
- Limitations
  - *Clinical relevance?*
  - *Subjective measurement*
- Today: Not a debate about when PFS is appropriate, rather a discussion of how to ensure 'robustness' of findings when it is the endpoint

Recent FDA approvals based on PFS (partial list)

- sorafenib - renal cell
- gemcitabine - ovarian cancer
- ixabepilone - breast cancer
- bevacizumab - breast cancer
- rituximab - non-Hodgkin's lymphoma

When should a PFS result be considered ‘credible’?

- Large improvement (months, not weeks)
- Optimal: Placebo control
- Clear definitions for non-radiologic progressions
- Rational explanation for lack of survival impact
- Reliable, unbiased PFS assessments
  - *Often verified through the use of a blinded independent central review (BICR)*
  - *BICR adds substantial cost & complexity to clinical trials*

Today's goal

- When do we need blinded independent central review (BICR) to assess PFS endpoints?
  - *What is the concordance between local and BICR assessment of PFS?*
- In what cases could an 'audit' (a review of < 100% of cases) be acceptable, and what is an 'optimal' audit?
- Are there cases where major differences have been observed between PFS results based on local vs BIRC, and what was the

Panel 2: Agenda

- Will Bushnell, GlaxoSmithKline **(15 min)**
- Lori Dodd, National Cancer Institute **(10 min)**
- Ohad Amit, GlaxoSmithKline (10 min)
- Nancy Roach, C3: Colorectal Cancer Coalition **(5 min)**
- Richard Pazdur, Food and Drug Administration **(15 min)**
- Audience Questions and Comments **(30 min)**

ENGELBERG CENTER for
Health Care Reform
at BROOKINGS

FRIENDS of CANCER RESEARCH

Conference on Clinical Cancer Research

Supported by:

AACR CENTENNIAL 1907 – 2007 American Association for Cancer Research

ASCO American Society of Clinical Oncology

LIVESTRONG LANCE ARMSTRONG FOUNDATION

SUSAN G. KOMEN FOR THE CURE

September 14, 2009 • Washington, DC

An Overview of Independent Review of PFS and Proposal for an Audit Methodology

Will Bushnell

Presented on behalf of the PHRMA PFS Independent Review Working Group

Ohad Amit, Frank Mannino, Jon Denne, Steve Dahlberg, Andrew Strahs, Andy Stone, Paul Bycott, Will Bushnell, Bob Ford, Dmitri Pavlov, Sandra Chica, Rajeshwari Sridhara, Mark Rothmann, Aloka Chakravarty, Grant Williams, Hans Ulrich Burger, Bill Capra, Bee Chen, Nicole Blackman, Vicki Goodman

Outline

- Background
- Measurement Error and Bias
- Defining and Evaluating Discordance
- A “Meta-analysis” of 23 clinical trials

Background

- PHRMA PFS Independent Review Workstream formed in early 2008
- Team identified several objectives
  - Quantify value of independent review in the context for the PFS endpoint
  - Provide clarity on interpretation of discordance
  - Quantify impact of discordance on Tx effect estimates
  - Develop audit methodology

Key focus - differentiating measurement error from bias

- Questions
  - How does measurement error impact Tx effect?
  - How does bias impact Tx effect?
  - Should we be concerned if local and central evaluations provide same estimate of Tx effect in presence of significant discordance?
- Answers
  - Achieved through development and simulation from an error model
  - Simulation results confirmed through analysis of summary data from 23 trials

Overall Discordance is Indicative of Measurement Error

- Discordance is primarily a consequence of measurement error
  - Establishing progression is a highly complex process
  - High Discordance rates between 2 blinded independent reviewers has been observed in several retrospective analyses
  - Discordance between 2 blinded reviewers cannot be attributed to bias
  - Highly concordant estimates of Tx effect have been observed in the presence of high patient level discordance in the assessment of PD

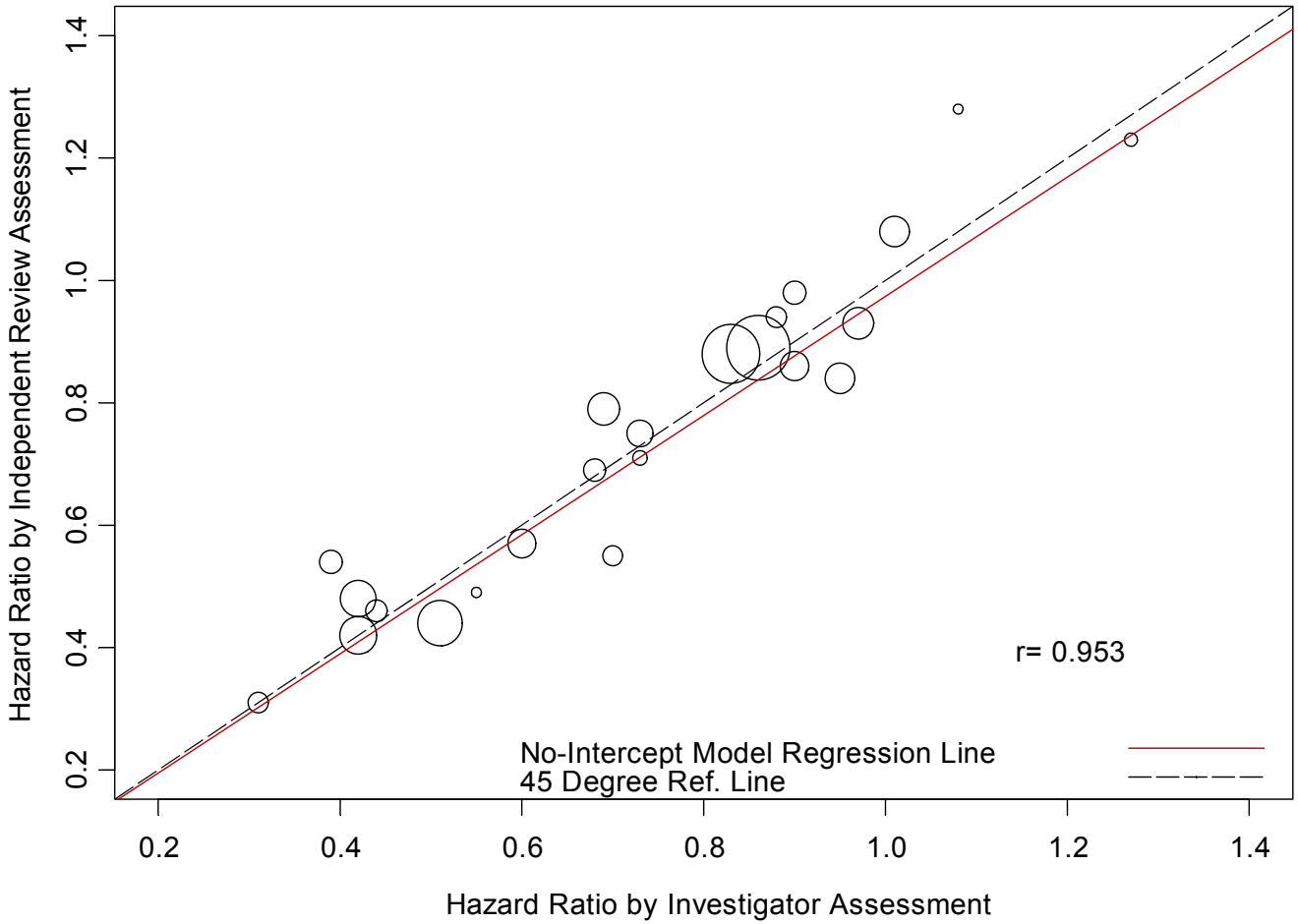
Meta-analysis of 23 Trials

- Data available on 23 trials where HR's were reported for IRC and investigator
  - Obtained through formal data collection exercise with sponsors and literature review
  - All trials in advanced metastatic disease
  - 7 Breast, 6 CRC, 6 RCC, 1 Mesothelioma, 1 GBM, 1 melanoma, 1 Ovarian
  - 8 of the 23 trials were blinded
  - Sample size ranged from 200 to ~1300 subjects
  - All but 5 trials had sample size >350 subjects

Details of Formal Data Collection Exercise

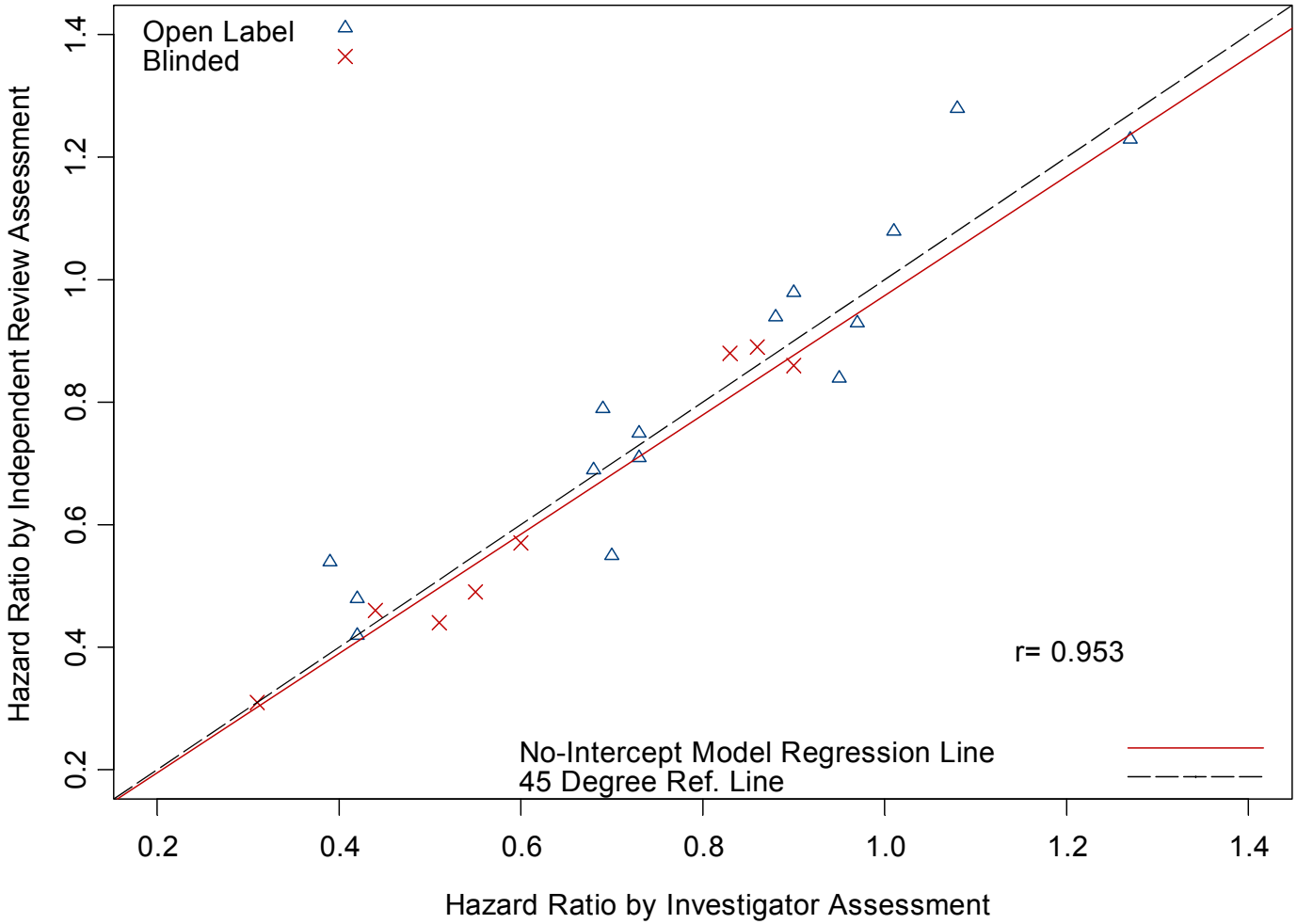
- Additional detailed data collected from 12 of the 23 studies
  - Contributed from 4 PHRMA member companies
  - 5 Breast, 3 CRC, 1 RCC, 1 Mesothelioma, 1 melanoma, 1 GBM
  - RECIST criteria used in 7 of 12 trials
  - Additional details collected on discordance and response criteria
  - Goal to confirm simulation results for identifying a robust discordance measure

Estimates of Tx effect are Strongly Correlated



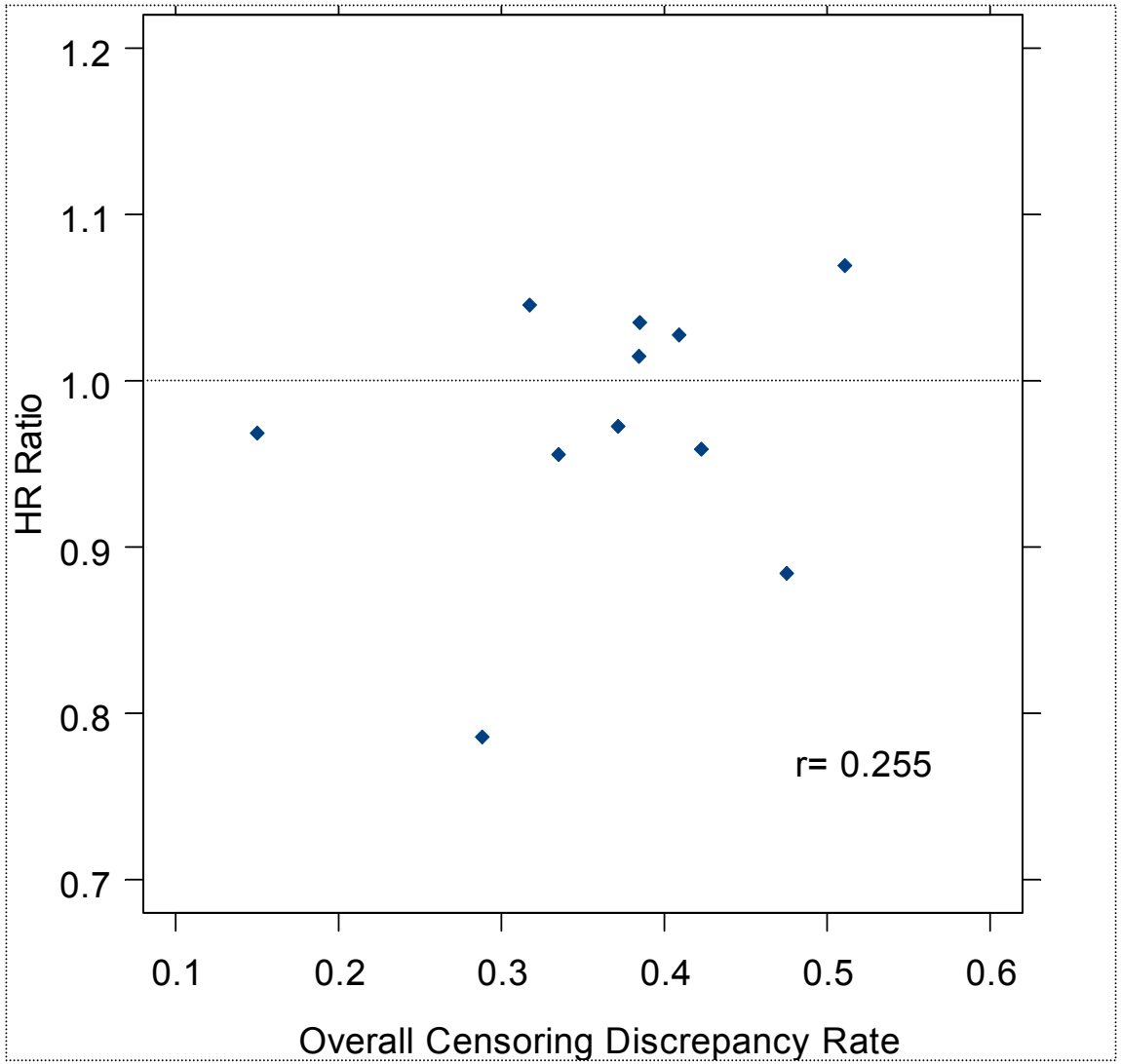
HR ratio (95% CI): 1.02 (0.96,1.07)

Effect of Blinding



P-value for effect of blinding: 0.193

The Level of Discordance Does not Impact Reliability of Tx Effect Estimates (N=11 trials)



Conclusions from "Meta-analysis" and Simulations

- Discordance – natural consequence of measuring complex system
- Observed strong agreement (local vs central) est. of Tx effect (HR) in meta-analysis
- Study blinding doesn't effect HR agreement
- HR agreement unaffected by discordance rates
- Differential discordance – is related to disagreement of HR and potentially indicative of bias

The utility of these findings

- Serves as a background for the development and conduct of BICR audits
- Informs the interpretation of results from trials using PFS and independent review

ENGELBERG CENTER for
Health Care Reform
at BROOKINGS

FRIENDS of CANCER RESEARCH

Conference on Clinical Cancer Research

Supported by:

AACR CENTENNIAL 1907 – 2007 American Association for Cancer Research

ASCO American Society of Clinical Oncology

LIVESTRONG LANCE ARMSTRONG FOUNDATION

SUSAN G. KOMEN FOR THE CURE

September 14, 2009 • Washington, DC

An Audit Methodology for Central Review of PFS

Ohad Amit, PhD

Presented on behalf of the PHRMA PFS Independent Review Working Group

Ohad Amit, Frank Mannino, Jon Denne, Steve Dahlberg, Andrew Strahs, Andy Stone, Paul Bycott, Will Bushnell, Bob Ford, Dmitri Pavlov, Sandra Chica, Rajeshwari Sridhara, Mark Rothmann, Aloka Chakravarty, Grant Williams, Hans Ulrich Burger, Bill Capra, Bee Chen, Nicole Blackman, Vicki Goodman

Outline

- Bias and Differential Discordance
- Some Clinical Trial Results
- Audit Methodology
- Summary and Conclusions

Bias in the Evaluation of PFS

- Bias can manifest in several ways in the assessment of PD
- Directional Evaluation Bias is of primary concern in local evaluation
  - Investigator calls it early for the control arm
  - Investigator calls it late for the experimental arm
  - Either case can lead to optimistic estimate of Tx effect in local evaluation
  - Sponsor also concerned with directional evaluation bias in other direction – underestimation of magnitude of benefit
- Audit methodology developed to detect directional evaluation bias through a sensitive and specific measure

Differential Discordance as a Measure of Bias

- Differential discordance: Defined as the difference in discordance rate between treatment arms

- Large differences between Tx arms in discordance can raise suspicion of systematic directional evaluation bias

Differential Discordance as a Measure of Bias

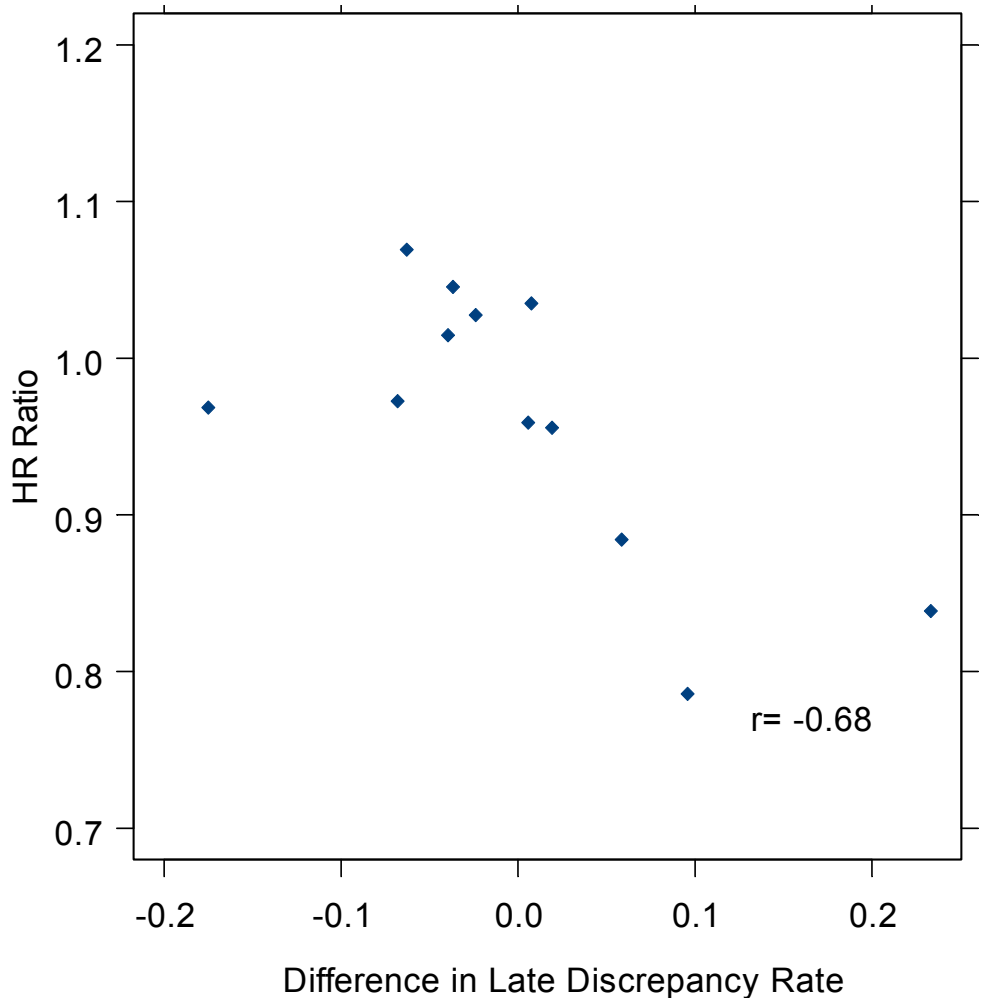
- No differential discordance → Absence of bias → Local evaluation provides a reliable estimate of treatment effect

- Differential Discordance → Potential for evaluation bias → Compare Tx effects by local and central evaluation

Details of Formal Data Collection Exercise

- Detailed data collected from 12 trials where Blinded Independent Central Review (BICR) and Local Evaluation (LE) data were available
  - Contributed from 4 PHRMA member companies
  - 5 Breast, 3 CRC, 1 RCC, 1 Mesothelioma, 1 melanoma, 1 GBM
  - RECIST criteria used in 7 of 12 trials
  - Details collected on differential discordance
  - Goal to confirm simulation results for identifying a robust discordance measure

Differential Discordance Impacts Reliability of Tx Effect Estimates (N=12 trials)



Central Review as an Audit of Local Evaluation

- Goal of Audit
  - Analogous to a “site audit” with goal of increasing confidence in integrity of trial and trial endpoints
  - Audit is not intended to re-estimate treatment effect
- Key concept underpinning audit methodology:
  - Local evaluation on the whole provides reliable estimates of treatment effect
  - Interested in detecting big anomalies
  - Big anomalies correlate with meaningful bias in the estimates of treatment effect
- Audit methodology developed using statistical simulation and results from meta-analysis

When Should a BICR Audit be Done

- No BICR
  - Blinded trials
  - Trials with a large treatment effect
- Partial BICR (Audit)
  - Open-label trials
  - When expected effect on PFS is not as large
- 100% BICR
  - Trials with smaller sample size or BICR not feasible
  - Trials where there is a strong need to increase confidence in the LE PFS

Audit Methodology

- Methodology accomplishes the following:
  - Detection of systematic, directional bias favoring the experimental treatment
  - Facilitate conclusions about reliability of Tx effect as estimated by local evaluation of PFS
- Some other key considerations
  - How do the identified measures of bias behave in a random sample of the full trial population
  - How large should the sample be
  - What is the threshold value at which we are concerned about the potential for bias
  - What level of uncertainty can we live with

Audit Operational Aspects

- Timing of audit – three options
  - Real time during study
  - At time of clinical cutoff but prior to breaking the randomization code
  - After Evaluation of LE Tx effect – no need for BICR with large Tx effects
- Central management of scans is desirable
- Two sampling schemes
  - Sample subset of LE PD events – limits the metrics but requires less subjects
  - Sample subset of Patients – requires more patients but facilitates calculation of multiple metrics

Audit Methodology Decision Rules

- Binary decision process:
  - Proceed to 100% BICR or
  - Conclude that local evaluation is reliable

- Decision should be based on weight of evidence
  - If differential discordance exceeds 10% consider a 100% BICR
  - If differential discordance is near 0 and no evidence for statistical difference then conclude local evaluation is reliable

- Two candidate measures of discordance identified for use in an audit
  - Early Discrepancy Rate (EDR): Measures imbalance in calling early progression on the control arm
  - Late Discrepancy Rate (LDR): Measures imbalance in calling late progression on the experimental arm

Audit Operating Characteristics

- Sensitivity of Audit
  - What proportion of the time do we detect evaluation bias when it is there?
  - This represents regulatory risk
- Specificity of Audit
  - What proportion of the time do we conclude LE is reliable in the absence of evaluation bias
  - This represents the sponsor's risk
- Desirable to have an audit with high sensitivity and specificity

Audit Operating Characteristics – (N=80 "events")

Discordance Measure	Cutoff Criteria	% Discordance Control/Experimental	Sensitivity	Specificity
LDR*	$\Delta \geq 0.10$	20%/40%	81%	87%
		25%/45%	83%	85%
		30%/50%	83%	84%
EDR	$\Delta \leq -0.10$	60%/40%	82%	82%
		65%/45%	82%	82%
		70%/50%	82%	81%

*For the LDR approximately 160 patients will need to be sampled to get 80 "events"

Underlying rates based are based on clinical trial data and simulation results

Summary, Conclusions and Next Steps

- Several operational aspects to consider in an audit including timing, scan management and sampling scheme
- Differential Discordance is a useful tool for detecting presence of evaluation bias
- Differential Discordance can be used to design audits of a manageable size with good operating characteristics
- There is need to pilot an audit in a current trial

ENGELBERG CENTER for
Health Care Reform
at BROOKINGS

FRIENDS of CANCER RESEARCH

Conference on Clinical Cancer Research

Supported by:

AACR CENTENNIAL 1907 – 2007 American Association for Cancer Research

ASCO American Society of Clinical Oncology

LIVESTRONG LANCE ARMSTRONG FOUNDATION

SUSAN G. KOMEN FOR THE CURE

September 14, 2009 • Washington, DC

An audit of locally-determined progression-free survival using blinded independent central review

Lori E. Dodd (NIAID)
Edward L. Korn (NCI)
Boris Freidlin (NCI)
Robert Gray (Harvard)
Suman Bhattacharya (Genentech)
Rajeshwari Sridhara (FDA)

Sept 14, 2009

Background on BICR

- Motivation: Reader variability.

Background on BICR

- Motivation: Reader variability.
  - In a firstline therapy trial of metastatic breast cancer (E2100):

Background on BICR

- Motivation: Reader variability.
  - In a firstline therapy trial of metastatic breast cancer (E2100):
  - Discrepancy rate between Blinded Independent Central Review (BICR) and Local Evaluations (LE) discrepancy rate: 36 %.

Background on BICR

- Motivation: Reader variability.
  - In a firstline therapy trial of metastatic breast cancer (E2100):
  - Discrepancy rate between Blinded Independent Central Review (BICR) and Local Evaluations (LE) discrepancy rate: 36 %.
  - Two BICR reviewers discrepancy rate: 34 %.†

† Discrepancy rates were computed on 649 subjects for whom an image was available for BICR and included a +/- 6 week window.

Background on BICR

- Motivation: Reader variability.
  - In a firstline therapy trial of metastatic breast cancer (E2100):
  - Discrepancy rate between Blinded Independent Central Review (BICR) and Local Evaluations (LE) discrepancy rate: 36 %.
  - Two BICR reviewers discrepancy rate: 34 %.†
- Issue: Potential presence of informative censoring.

† Discrepancy rates were computed on 649 subjects for whom an image was available for BICR and included a +/- 6 week window.

Background on BICR

- Motivation: Reader variability.
  - In a firstline therapy trial of metastatic breast cancer (E2100):
  - Discrepancy rate between Blinded Independent Central Review (BICR) and Local Evaluations (LE) discrepancy rate: 36 %.
  - Two BICR reviewers discrepancy rate: 34 %.†
- Issue: Potential presence of informative censoring.
  - Patients are taken off protocol at point of local progression and no other scans are taken.

† Discrepancy rates were computed on 649 subjects for whom an image was available for BICR and included a +/- 6 week window.

Background on BICR

- Motivation: Reader variability.
  - In a firstline therapy trial of metastatic breast cancer (E2100):
  - Discrepancy rate between Blinded Independent Central Review (BICR) and Local Evaluations (LE) discrepancy rate: 36 %.
  - Two BICR reviewers discrepancy rate: 34 %.†
- Issue: Potential presence of informative censoring.
  - Patients are taken off protocol at point of local progression and no other scans are taken.
  - If BICR progression not determined by this time point, patient is censored.

† Discrepancy rates were computed on 649 subjects for whom an image was available for BICR and included a +/- 6 week window.

Background on BICR

- Motivation: Reader variability.
  - In a firstline therapy trial of metastatic breast cancer (E2100):
  - Discrepancy rate between Blinded Independent Central Review (BICR) and Local Evaluations (LE) discrepancy rate: 36 %.
  - Two BICR reviewers discrepancy rate: 34 %.†
- Issue: Potential presence of informative censoring.
  - Patients are taken off protocol at point of local progression and no other scans are taken.
  - If BICR progression not determined by this time point, patient is censored.
  - Further, loss of events reduces power.

† Discrepancy rates were computed on 649 subjects for whom an image was available for BICR and included a +/- 6 week window.

Results from E2100

- In spite of these concerns, hazard ratios from BICR and LE were similar.

Results from E2100

- In spite of these concerns, hazard ratios from BICR and LE were similar.
  - LE hazard ratio is 0.42 (95% CI: 0.34 to 0.51).

Results from E2100

- In spite of these concerns, hazard ratios from BICR and LE were similar.
  - LE hazard ratio is 0.42 (95% CI: 0.34 to 0.51).
  - BICR hazard ratio is 0.48 (95% CI: 0.39 to 0.61).

Results from E2100

- In spite of these concerns, hazard ratios from BICR and LE were similar.
  - LE hazard ratio is 0.42 (95% CI: 0.34 to 0.51).
  - BICR hazard ratio is 0.48 (95% CI: 0.39 to 0.61).
- This provides reassurance that the LE result was not driven by reader-evaluation bias.

Results from E2100

- In spite of these concerns, hazard ratios from BICR and LE were similar.
  - LE hazard ratio is 0.42 (95% CI: 0.34 to 0.51).
  - BICR hazard ratio is 0.48 (95% CI: 0.39 to 0.61).
- This provides reassurance that the LE result was not driven by reader-evaluation bias.
- Could we have arrived at a similar conclusion if we had audited a subset with BICR?

A proposal for a BICR audit

- Goals of audit: provide assurance about lack of LE bias in the treatment effect.

A proposal for a BICR audit

- Goals of audit: provide assurance about lack of LE bias in the treatment effect.
- Our approach: estimate the BICR hazard ratio and determine statistical and clinical significance.

A proposal for a BICR audit

- Goals of audit: provide assurance about lack of LE bias in the treatment effect.
- Our approach: estimate the BICR hazard ratio and determine statistical and clinical significance.
- We developed a more efficient and unbiased estimator that utilizes the correlation between LE and BICR.

Application of audit approach to E2100

- We have a complete BICR in E2100.

Application of audit approach to E2100

- We have a complete BICR in E2100.
- What if we had performed an audit sampling 20% of subjects for BICR?

Application of audit approach to E2100

- We have a complete BICR in E2100.
- What if we had performed an audit sampling 20% of subjects for BICR?
- Resample "audits" of size 20% many times (1000).

Application of audit approach to E2100

- We have a complete BICR in E2100.
- What if we had performed an audit sampling 20% of subjects for BICR?
- Resample "audits" of size 20% many times (1000).
- Test whether BICR hazard ratio is:

Application of audit approach to E2100

- We have a complete BICR in E2100.
- What if we had performed an audit sampling 20% of subjects for BICR?
- Resample "audits" of size 20% many times (1000).
- Test whether BICR hazard ratio is:
  - different from 1, or

Application of audit approach to E2100

- We have a complete BICR in E2100.
- What if we had performed an audit sampling 20% of subjects for BICR?
- Resample "audits" of size 20% many times (1000).
- Test whether BICR hazard ratio is:
  - different from 1, or
  - not greater than some threshold (the "clinical significance factor").

Application to E2100, cont'd

Results of audit approach applied to E2100:

Audit size	20% audit	50% audit
Proportion of times the audit:		
Excludes the null hypothesis	0.894	1.0
Rules out an improvement less than 1 month	0.741	0.999
Rules out an improvement less than 2 months	0.502	0.905

Important feature of audit strategy:

E2100 was an example with a relatively large treatment effect. What about more moderate effect sizes?

LE Conclusion		
Small, statistically significant	Moderate, statistically significant	Large, statistically significant
Implications for BICR audit		
Not clinically meaningful & results likely not confirmed by full audit	Full audit may be necessary	Audit of subset may be sufficient

Conclusions

- Large treatment effects more clinically meaningful and robust to reader evaluation bias.

Conclusions

- Large treatment effects more clinically meaningful and robust to reader evaluation bias.
- Similar hazard ratios from BICR and LE offers assurance about lack of evaluation bias.

Conclusions

- Large treatment effects more clinically meaningful and robust to reader evaluation bias.
- Similar hazard ratios from BICR and LE offers assurance about lack of evaluation bias.
- However, if the hazard ratio from BICR is not significant may not mean there was no true effect.

Conclusions

- Large treatment effects more clinically meaningful and robust to reader evaluation bias.
- Similar hazard ratios from BICR and LE offers assurance about lack of evaluation bias.
- However, if the hazard ratio from BICR is not significant may not mean there was no true effect.
- Re-analysis of the E2100 trial indicated that 89% of time an audit of 20% would have been sufficient to demonstrate a significant treatment effect.

Acknowledgements

- ECOG
- Genentech