A Sample-based Approach for Independent Review of PFS using Differential Discordance

Ohad Amit, PhD

Presented on behalf of the PFS Independent Review Working Group

GlaxoSmithKline (Ohad Amit, Frank Mannino)
Eli Lilly (Jonathan Denne)
AstraZeneca (Andrew Stone)
Merck (Cong Chen)
Hoffman-La Roche (Hans-Ulrich Burger)

ACKNOWLEDGEMENTS

William Bushnell, Paul Bycott, Steve Dahlberg, Andrew Strahs, Grant Williams, Vicki Goodman, Sam Givens, Nicole Blackman, Bill Capra, Bee Chen, Michelle Casey, Michael Durante, Dan Sargent, Karen Fitch, Lauren McCann, Jeff Helterbrand, Agnes Westelinck, Fabrice Marsicano, Soumi Lahiri, Liang Fang

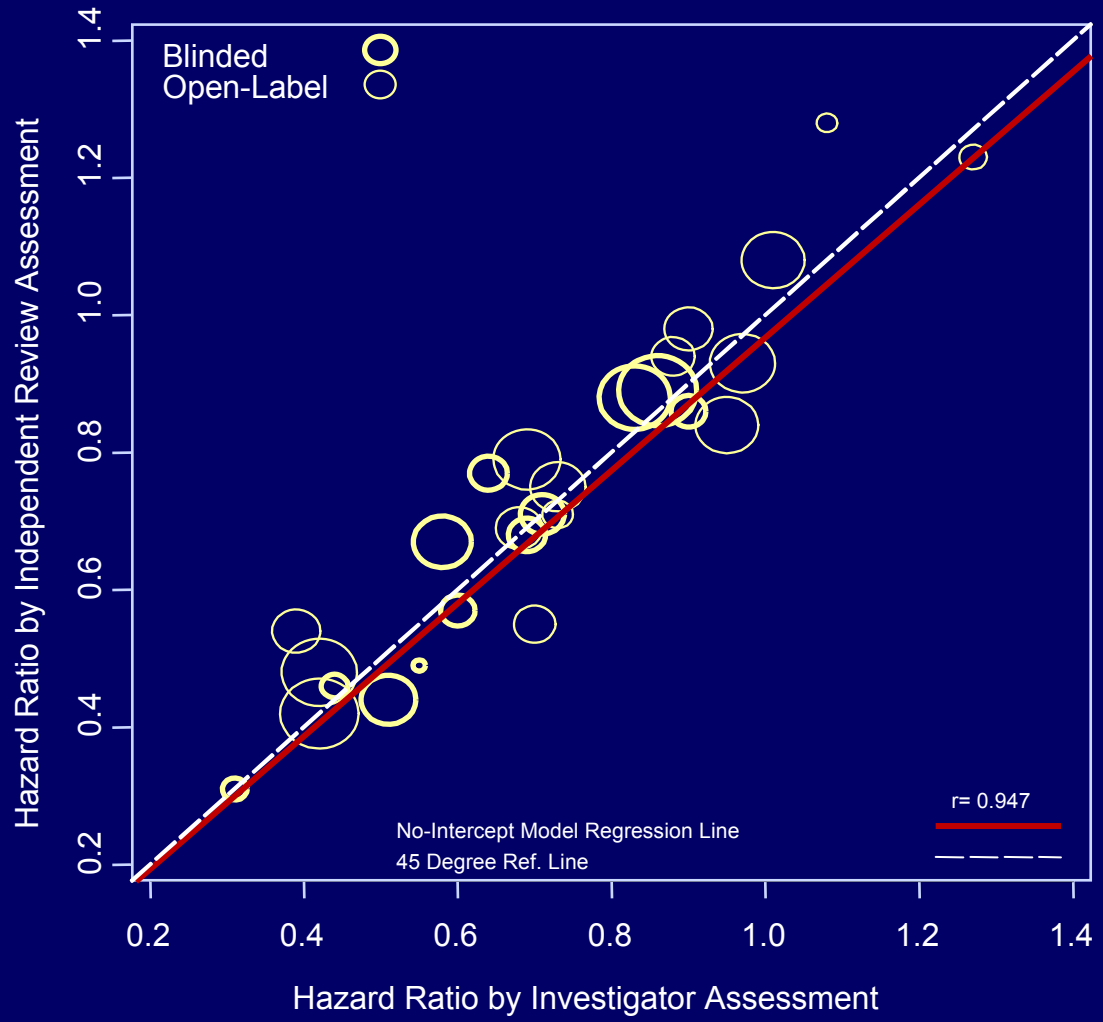
Key Points

- Local evaluation (LE) and blinded independent central review (BICR) provide comparable estimates of Tx effect in the great majority of clinical trials
- BICR is a mechanism for auditing the quality and reliability of the LE in the rare cases where evaluation bias may be present
- Desirable to lower the significant resource burden associated with a BICR by developing methods for detecting evaluation bias based on a sample of patients within the trial
- Differential discordance is an effective tool for assessing evaluation bias in a sample-based procedure

Introduction

- PFS working group formed in June 2008
- Meta-analysis undertaken to evaluate concordance in estimates of Tx effect between BICR and LE across multiple solid tumors
- Results from meta-analysis were used to motivate and develop methodology for a sample-based BICR
- Sample-based BICR developed and validated through statistical simulation
- Key findings published in European Journal of Cancer (Amit, 2011)

Results of Meta-analysis of 27 Solid-tumor trials



Ratio of BICR HR to LE HR (95% CI): 1.02 (0.96,1.07)

Defining Discordance

- **Discordance**
  - **Disagreement at the patient level between LE and BICR regarding either the occurrence or timing of progression**
- **Discordance rate**
  - **The rate at which disagreements occur on occurrence or timing of progression within a treatment arm**
- **Differential Discordance (DD)**
  - **The difference between Tx arms in discordance rates**
- **Measures of discordance can be defined in multiple ways**
  - **Some more useful than others**

BICR and Discordance

- The goal of any independent review is to confirm the Tx effect:
  - Highly concordant estimates of Tx effect are observed in the presence of significant discordance at the patient level

- Discordance is a consequence primarily of measurement error but can also be induced by evaluation bias

Discordance and evaluation bias

- **Separating evaluation bias from measurement error is critical**

- **Mechanism for evaluation bias**
  - **Investigator systematically calls progression earlier (or later) on one arm relative to other**
  - **Leads to different discordance patterns/rates in the experimental and control arms: Differential Discordance**

Defining and Evaluating Discordance

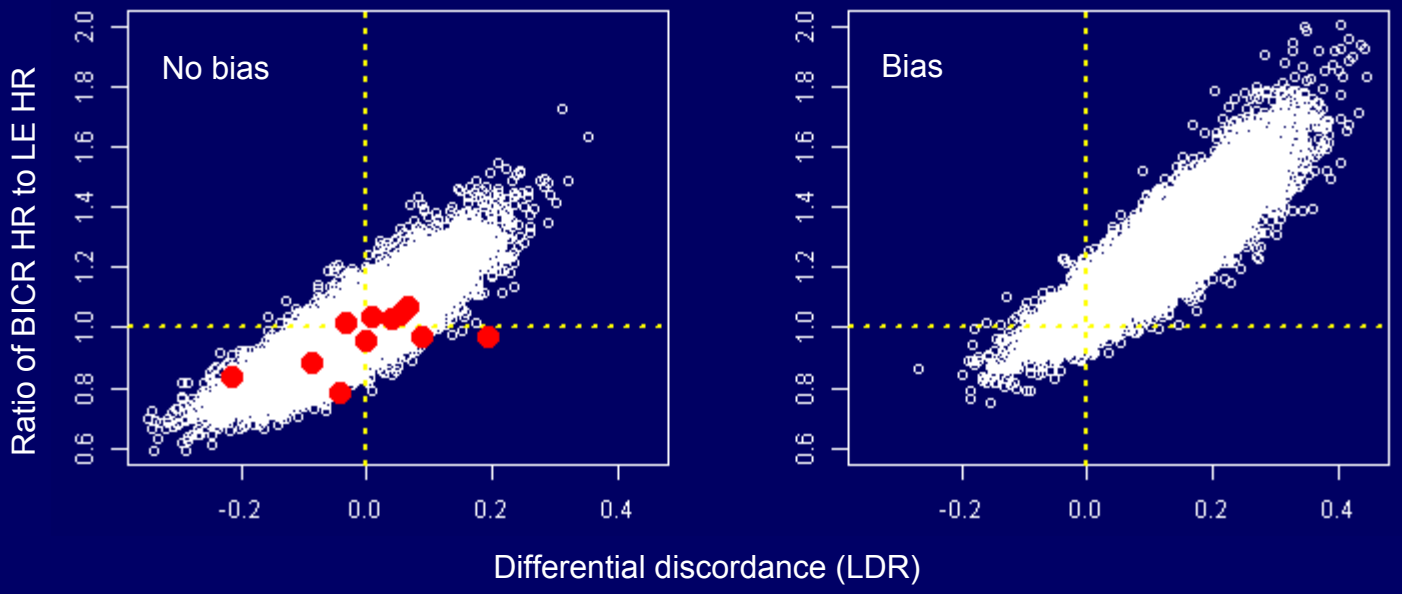
		BICR	
		PD	No PD
LE	PD	a=a1+a2+a3	b
	No PD	c	d

a1= # of times agreed on timing and occurrence of progression
a2= # of times agreed on PD event but BICR has earlier time
a3= # of times agreed on PD event but LE has earlier time

Early Discrepancy Rate: EDR (**b/(a+b)**)
Rate that LE assessed PD is called earlier than BICR PD

Late Discrepancy Rate: LDR (**(a2+c)/(a2+b+c)**)
Proportion of disagreements where LE PD's occur later than BICR PD.

Correlation between Differential Discordance and Differences in HR between BICR and LE

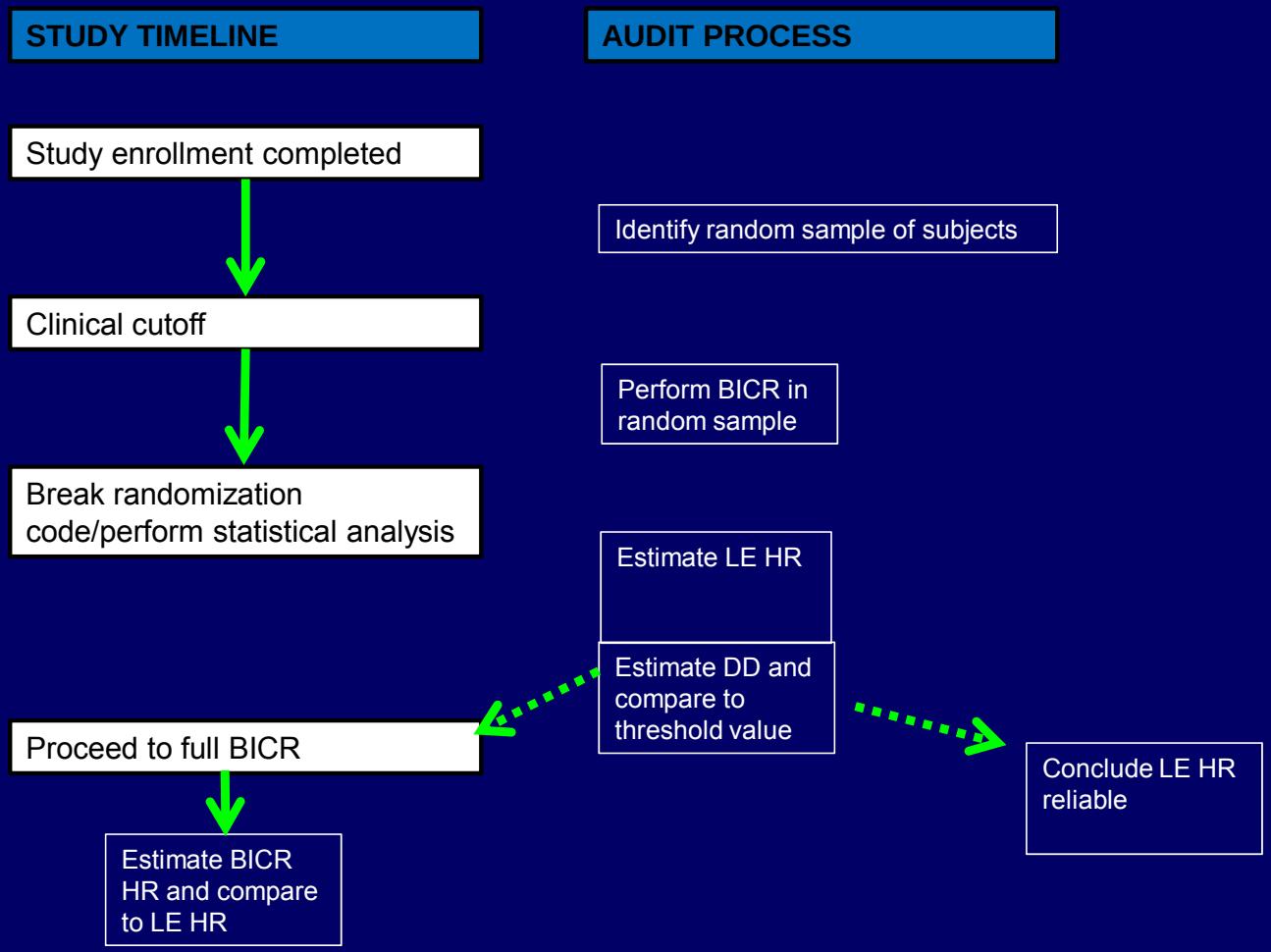


BICR as an Audit of LE using a Sample-based Procedure

- Goal of Sample-based Procedure
  - Increase confidence in integrity of trial and trial endpoints
  - Not intended to re-estimate treatment effect

- Key concepts supporting audit methodology:
  - Historically local evaluation provides reliable estimates of treatment effect
  - Want to reduce the burden of BICR while retaining a mechanism for detecting meaningful bias in the estimates of Tx
  - Meaningful bias in the estimates of treatment effect manifests as differences between treatment arms in discordance rates as measured by EDR/LDR

Audit Methodology: Sample-based Procedure



Operating Characteristics

- **Amit (2011)**
  - **Sensitivity: Probability of detecting bias in sample when it is truly present irrespective of outcome in full case review**
  - **Specificity: Probability of declaring LE reliable based on sample when no bias is present irrespective of outcome in full case review**
  - **Fixed threshold value regardless of sample size**

- **Per feedback from EMA define operating characteristics relative to the current practice for detecting evaluation bias**
  - **Comparison of hazard ratios from full case review**
  - **Relative difference of ~25% between LE and BICR HR's**
  - **Allow threshold value to reflect varying levels of uncertainty**

Sensitivity and Specificity of Procedure

		Full Case Review	
		Concordant HR	Discordant HR
Audit	DD> threshold	w	x
	DD < threshold	y	z

"Sensitivity" (positive agreement) (x/x+z): *What proportion of the time do we detect evaluation bias in the audit given you would have detected it in the full case review?*

"Specificity" (negative agreement) (y/w+y): *What proportion of the time do we conclude LE is reliable based on DD in the audit given a similar conclusion based on HR's from the full case review*

Sample-size and Threshold Values for Audit

- Threshold value based on fixing sensitivity for detecting bias at >90% for a given sample size
- Threshold value is Sample-size dependent
- Depending on desired specificity, sample sizes of 100-200 subjects are needed

Operating Characteristics: Based on 10,000 Simulations

Metric	Sample-size	Specificity	Sensitivity	Threshold value
LDR	100	61.0%	89.7%	0.0143
	150	74.9%	89.7%	0.0557
	200	86.0%	89.7%	0.0866
EDR	100	45.6%	89.7%	0.0453
	150	58.7%	89.7%	0.0153
	200	68.9%	89.7%	-0.0024

When Should a Sample-based BICR be Done

- No BICR
  - Truly blinded trials
- Sample- based BICR (Audit)
  - Open-label trials
  - Trials where complete blinding is not possible
- 100% BICR
  - Trials with smaller sample size where audit is not feasible
  - Trials where there is a strong need to increase confidence in the LE of PFS (e.g. in tumors where RECIST criteria may be more difficult to apply)

Summary and Next Steps

- **Conclusions**
  - Local evaluation consistently provides reliable estimate of Tx effect
  - Differential Discordance is a useful tool for detecting evaluation bias
  - Differential Discordance can be used to design audits of a manageable size with good operating characteristics
- **Next Steps**
  - Retrospective application of audit procedure in completed clinical trials
  - Regulatory acceptance