MRC | Hubs for Trials Methodology Research

Statistical validation of surrogate outcome measures

Marc Buyse

IDDI, Louvain-la-Neuve, and

Universiteit Hasselt, Diepenbeek, Belgium

Clinical Trials Methodology Conference,

Bristol, UK

October 4-5, 2011

IDDI
Adding Value to Clinical Data

universiteit hasselt

Surrogate outcome measures?

Lagakos SW, Hoth DF. Surrogate markers in AIDS: Where are we? Where are we going? *Ann Intern Med* 1992; 116: 599.

Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? *Ann Intern Med* 1996; 125: 605.

DeGruttola V, Fleming TR, Lin DY, Coombs R. Validating surrogate markers – are we being naive? *J Infectious Dis* 1997; 175: 237.

Baker SJ. Surrogate endpoints: wishful thinking or reality? *J Natl Cancer Inst* 2006*; 98 : 502.*

Berger VW. Does the Prentice criterion validate surrogate endpoints? *Statist in Med* 2004; 23: 1571.

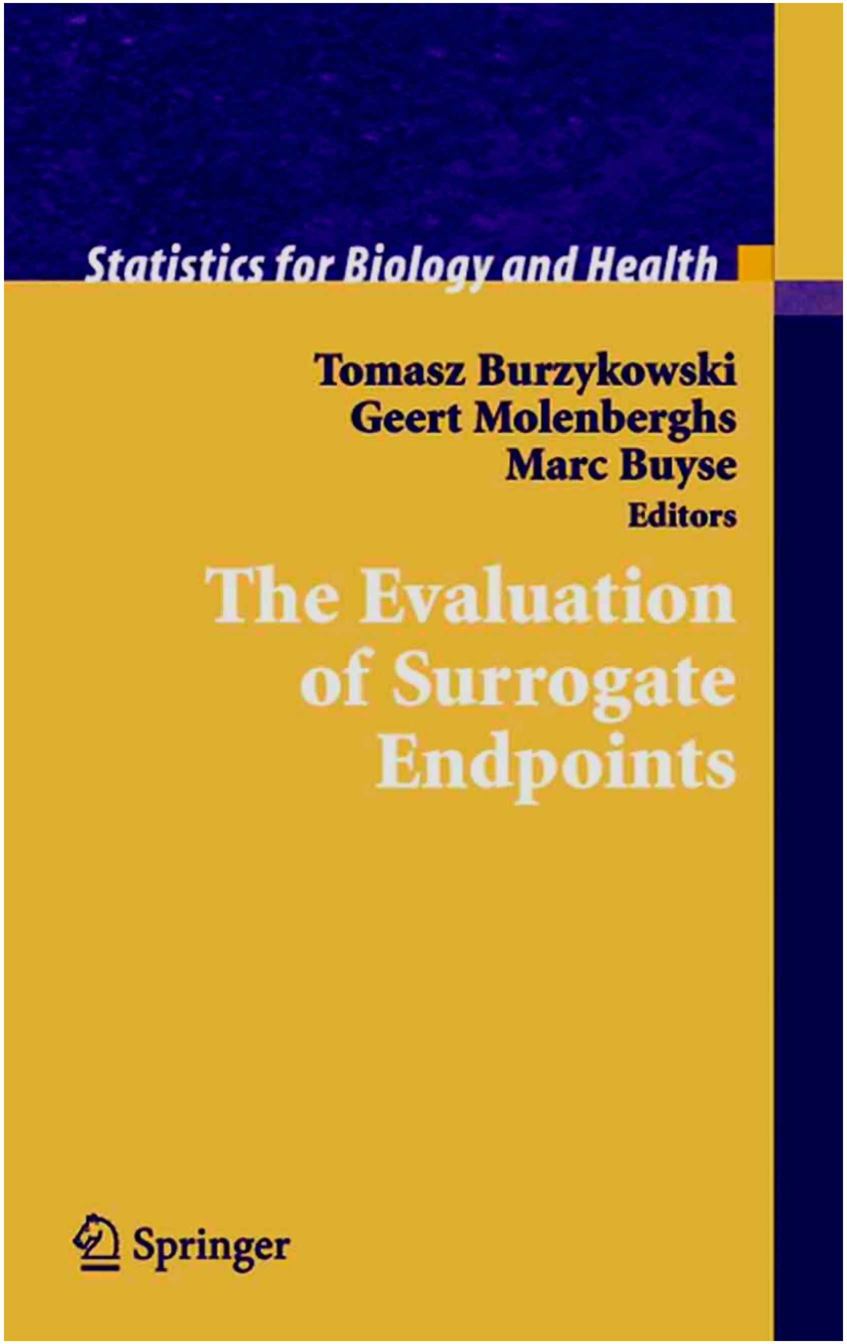
Burzykowski T. Surrogate endpoints: wishful thinking or reality? *Statist Meth Med Res* 2008; 917: 463.

Surrogate outcome measures?

The researches of many commentators have already thrown much darkness on this subject, and it is probable that if they continue, we shall soon know nothing at all about it.

Mark Twain

For the mathematically inclined…



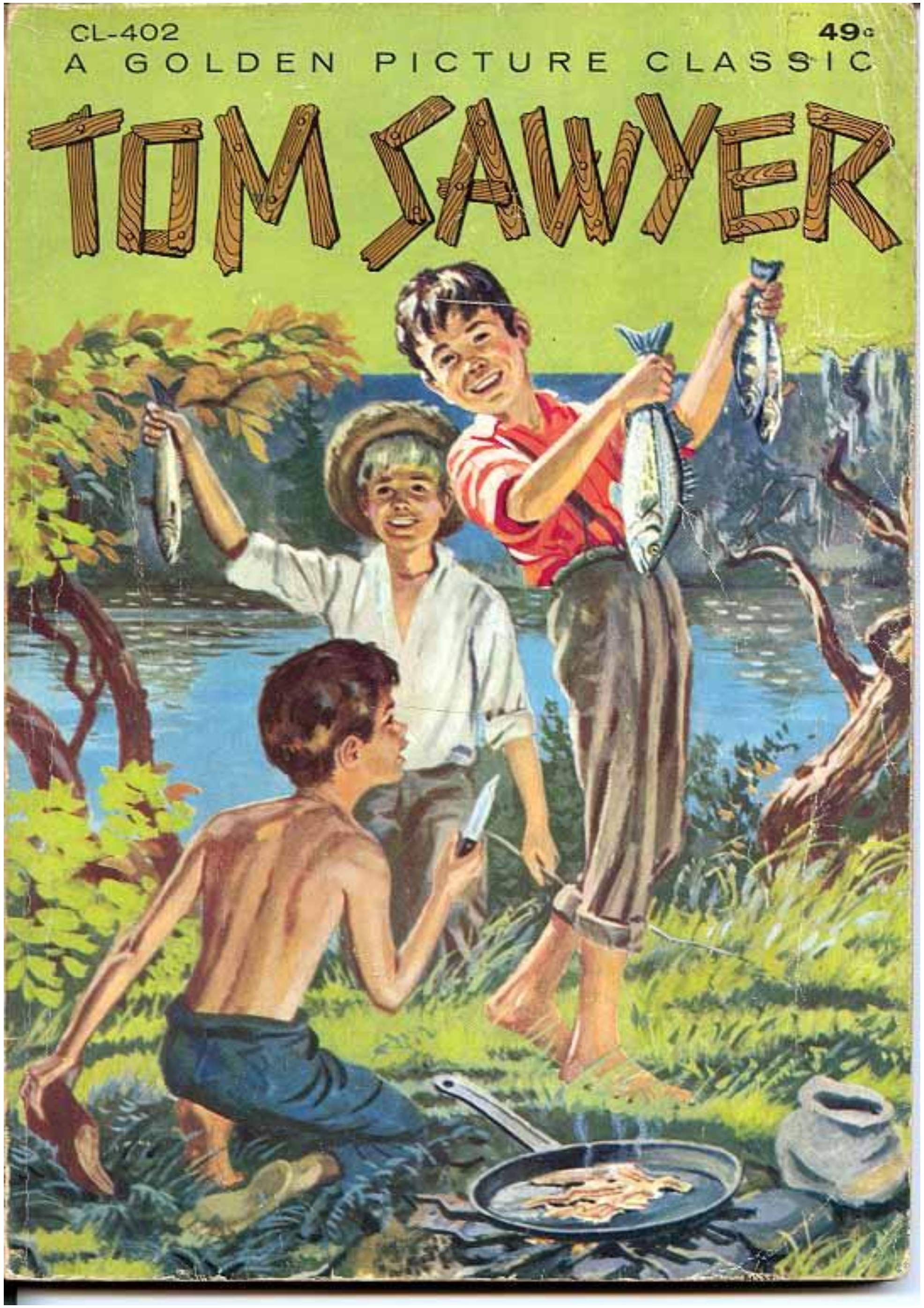
Statistics for Biology and Health

Tomasz Burzykowski
Geert Molenberghs
Marc Buyse
Editors

The Evaluation of Surrogate Endpoints

Springer

For the others…



Interest in surrogate endpoints / markers

- Feasibility / practicality of trials:
  - Shorter duration
  - Smaller sample size
  - Lower cost

- Availability of biomarkers that are potential surrogates:
  - Countless tissue, cellular, and hormonal factors
  - Advanced imaging techniques
  - Genomics, proteomics, metabolomics, other-ics

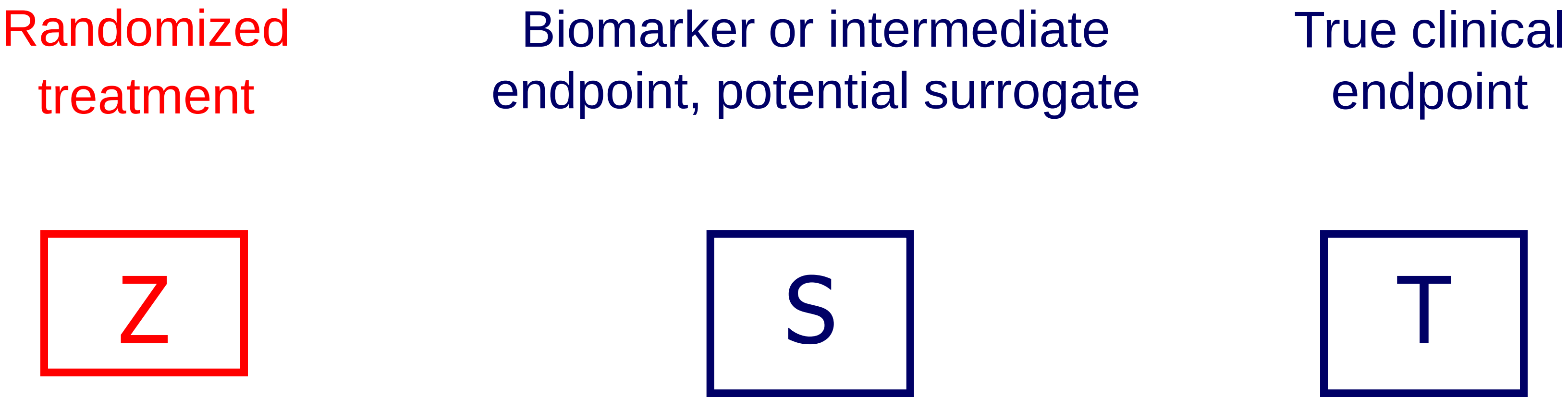
Ref: Schatzkin and Gail, Nature Reviews (Cancer) 2001, 3.

Outline

1. Capture of effect in a single randomized trial
2. Association measures in meta-analyses
3. Prediction
4. Causal inference
5. Conclusions

1. Capture of effect in a single randomized trial
2. Association measures in meta-analyses
3. Prediction
4. Causal inference

The single trial framework



Randomized treatment

Biomarker or intermediate endpoint, potential surrogate

True clinical endpoint

Z

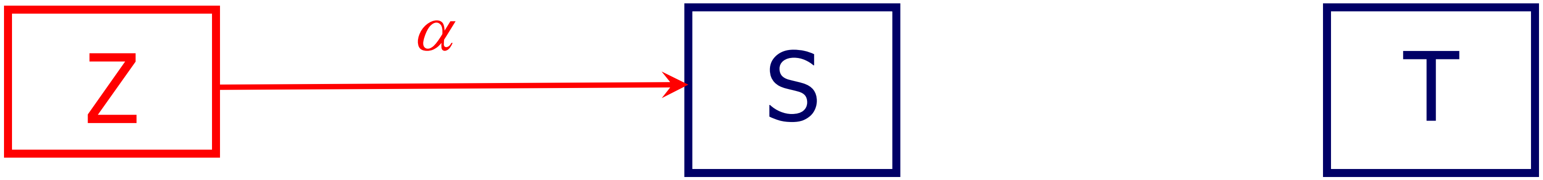
S

T

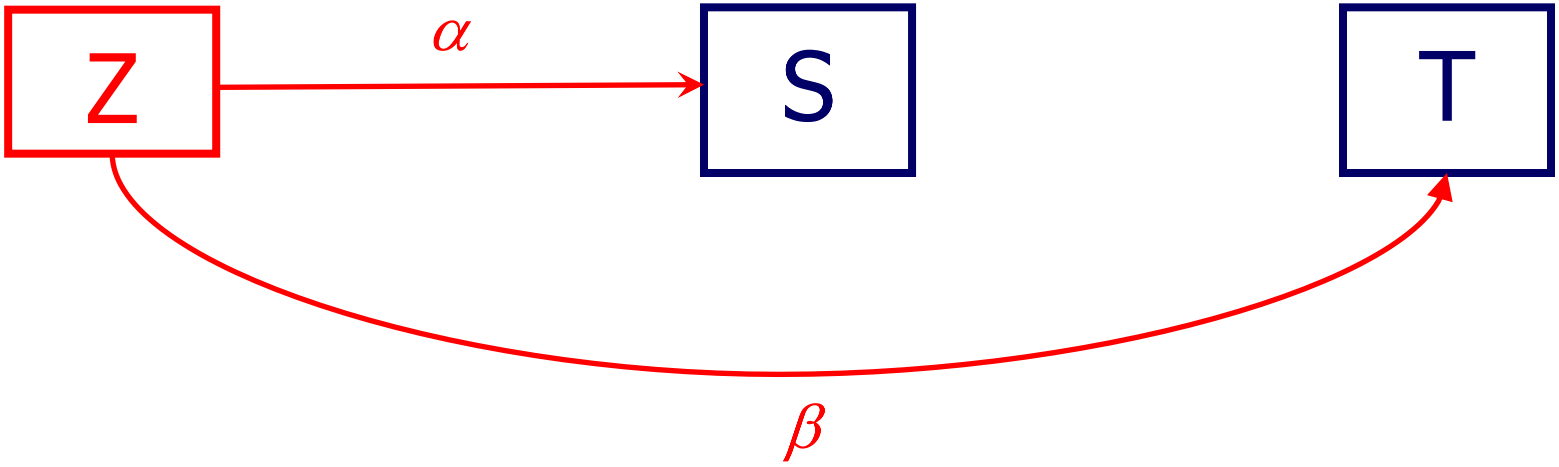
Ref: Buyse and Molenberghs, Biometrics 1998, 54: 1014

Parameters of interest

Effect of treatment on surrogate

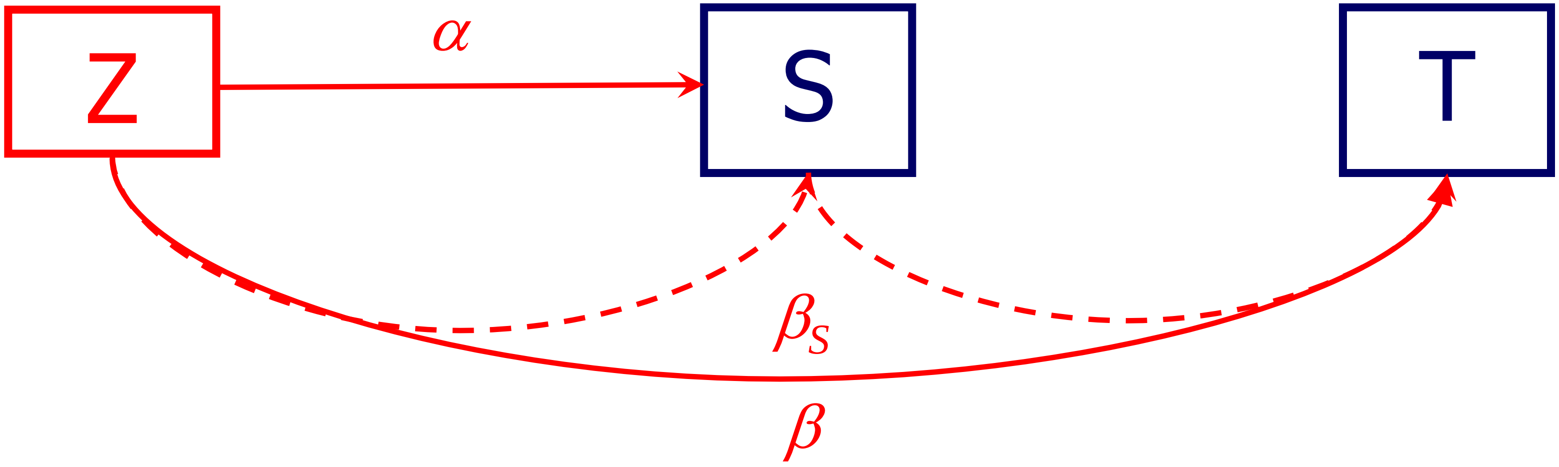


Parameters of interest



Effect of treatment on true endpoint

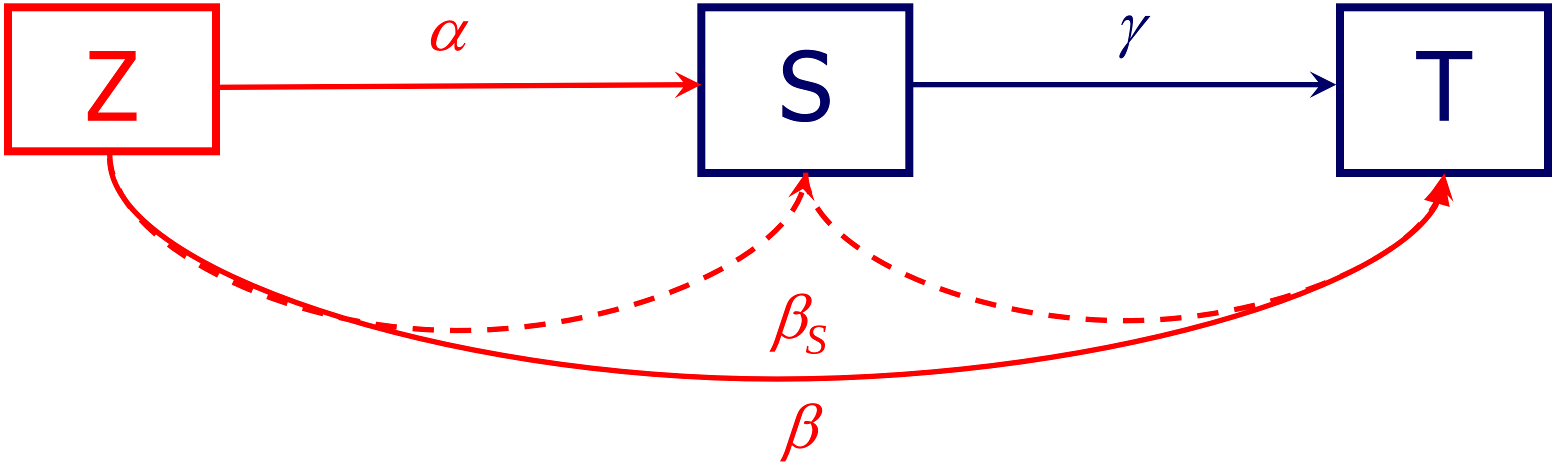
Parameters of interest



Effect of treatment on true endpoint,
adjusted for the surrogate

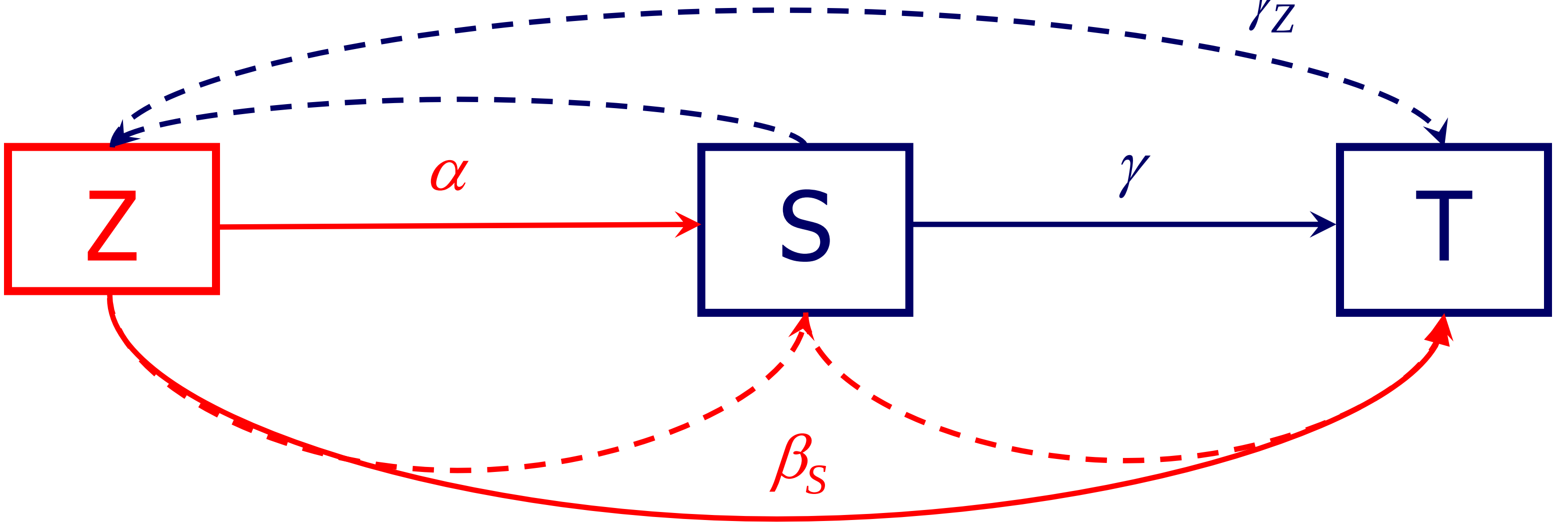
Parameters of interest

Effect of surrogate on true endpoint



Parameters of interest

Effect of surrogate on true endpoint,
adjusted for treatment



Correlation of endpoints is not enough

Key point: *“A correlate does not a surrogate make”*

⇒ A test of

H_0: $\gamma_Z = 0$

is not sufficient to establish validity

Refs: Fleming and DeMets, Ann Intern Med 1996, 125: 605

Biomarkers Definition Working Group, Clin Pharmacol Ther 2001, 69: 89.

A first set of criteria

A marker or endpoint can be used as a surrogate if

- it predicts the final endpoint:

$$H_0: \gamma_Z = 0$$

- it fully captures the effect of treatment upon the final endpoint:

$$H_0: \beta = 0 \ \underline{and} \ H_0: \beta_S \neq 0$$

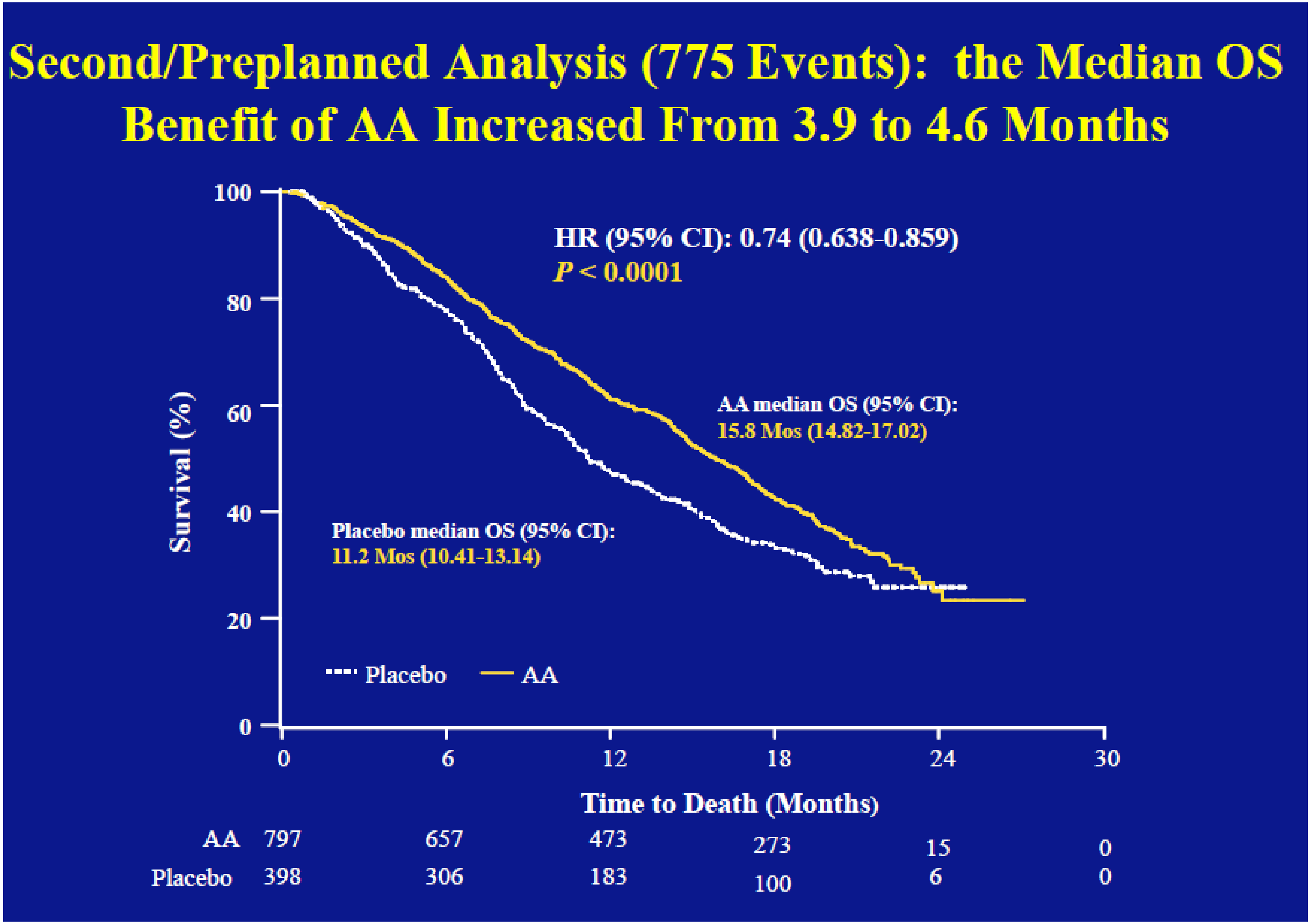
Ref: Prentice, Stat in Med 1989, 8: 431.

An example of Prentice’s approach

- Circulating tumor cells were measured among patients with metastatic castration-resistant prostate cancer.
- Patients randomized between abiraterone prednisone and placebo prednisone after failure of chemotherapy.
- CTC counts were taken at baseline, 4, 8 and 12 weeks after starting therapy.
- CTC conversion refers to a baseline count ≥ 5 cells and a count under treatment < 5 cells.

Ref: Danila et al, Clin Cancer Res 2011;17:3903

Phase III trial in metastatic prostate cancer

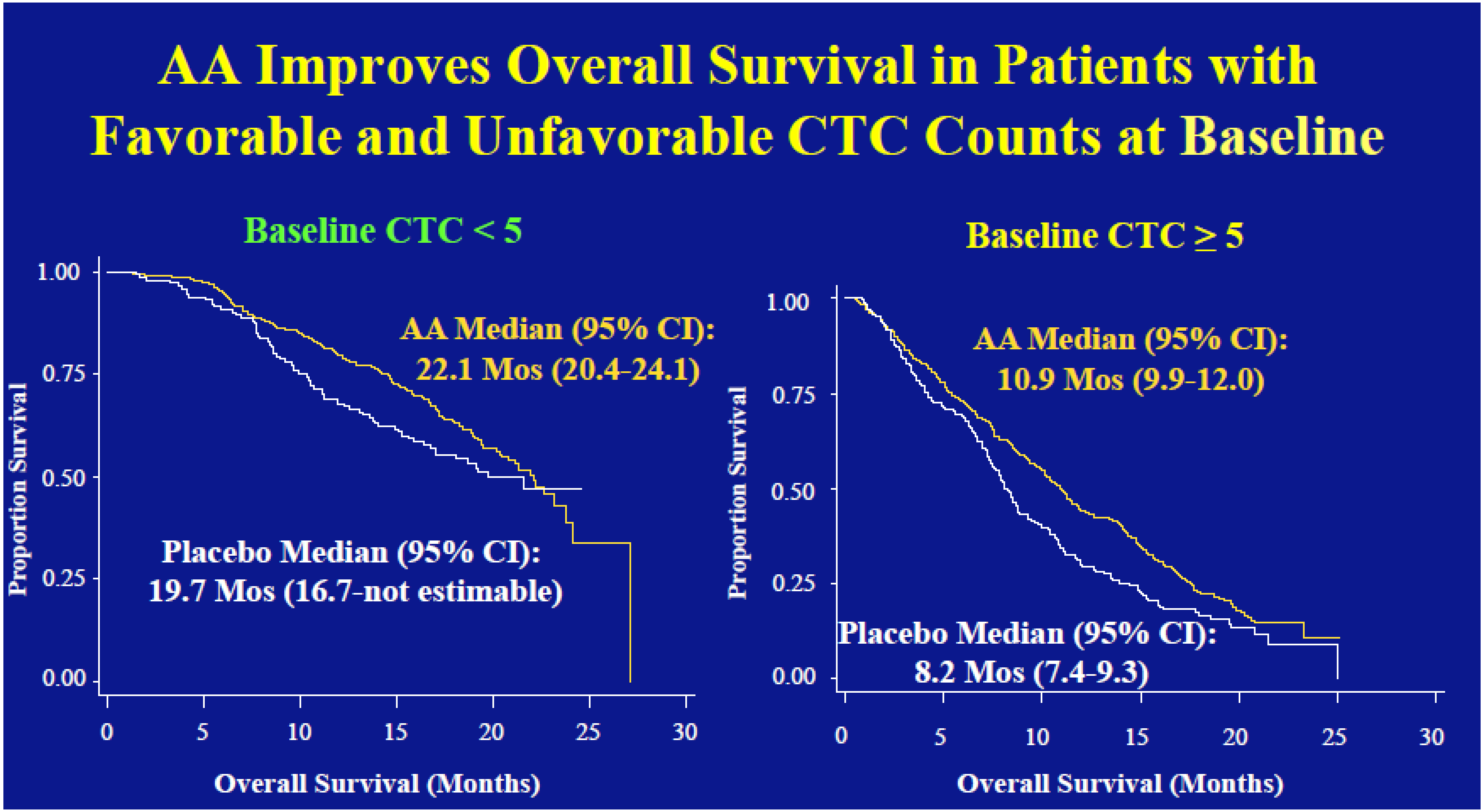


Second/Preplanned Analysis (775 Events): the Median OS Benefit of AA Increased From 3.9 to 4.6 Months

AA	797	657	473	273	15	0
Placebo	398	306	183	100	6	0

Ref: de Bono et al, NEJM 2011;364:1995; Scher et al, ASCO 2011

CTC counts are prognostic but not predictive



AA Improves Overall Survival in Patients with Favorable and Unfavorable CTC Counts at Baseline

Baseline CTC < 5

AA Median (95% CI): 22.1 Mos (20.4-24.1)

Placebo Median (95% CI): 19.7 Mos (16.7-not estimable)

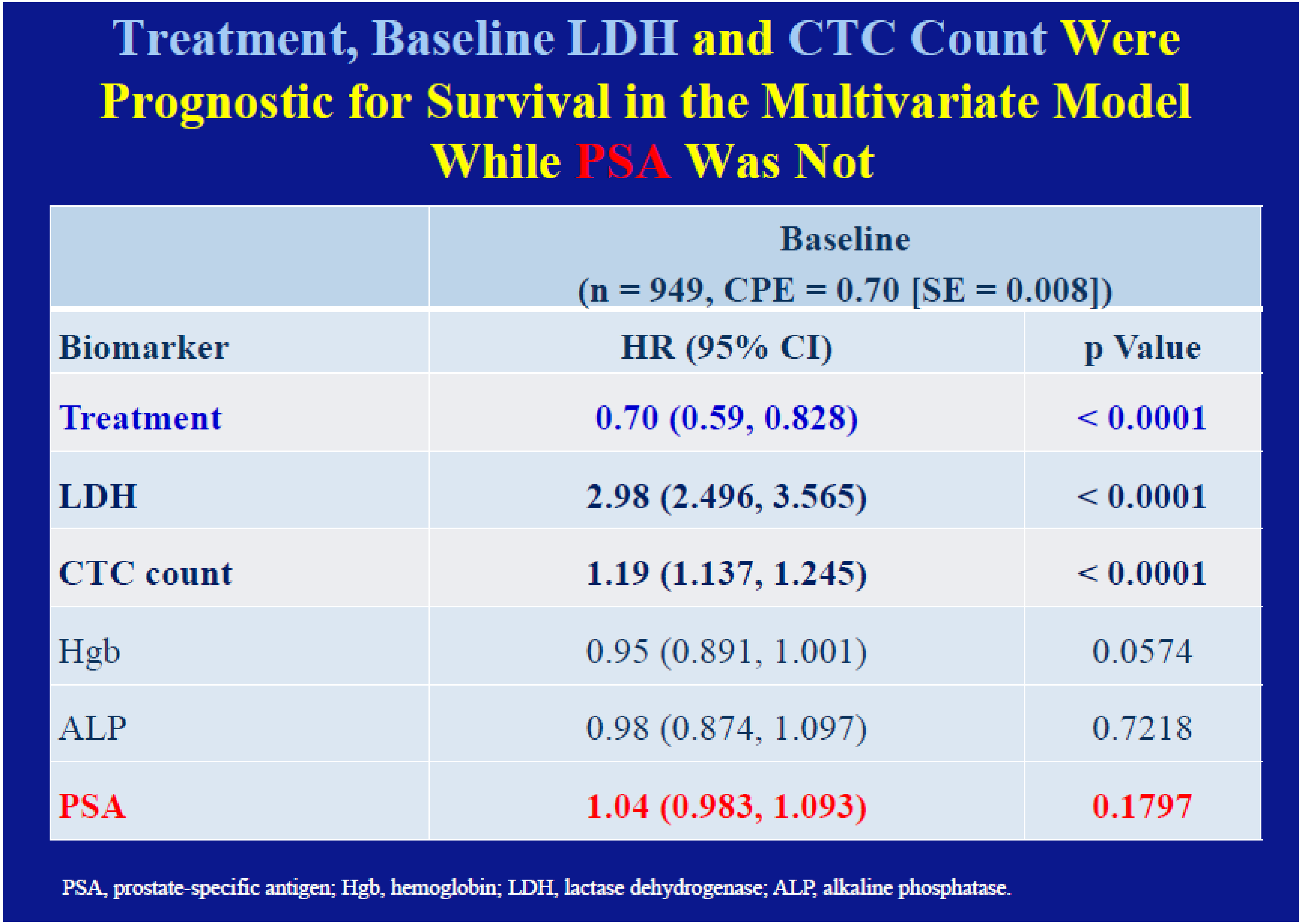
Baseline CTC ≥ 5

AA Median (95% CI): 10.9 Mos (9.9-12.0)

Placebo Median (95% CI): 8.2 Mos (7.4-9.3)

Ref: Scher et al, ASCO 2011

Model for treatment effect on overall survival



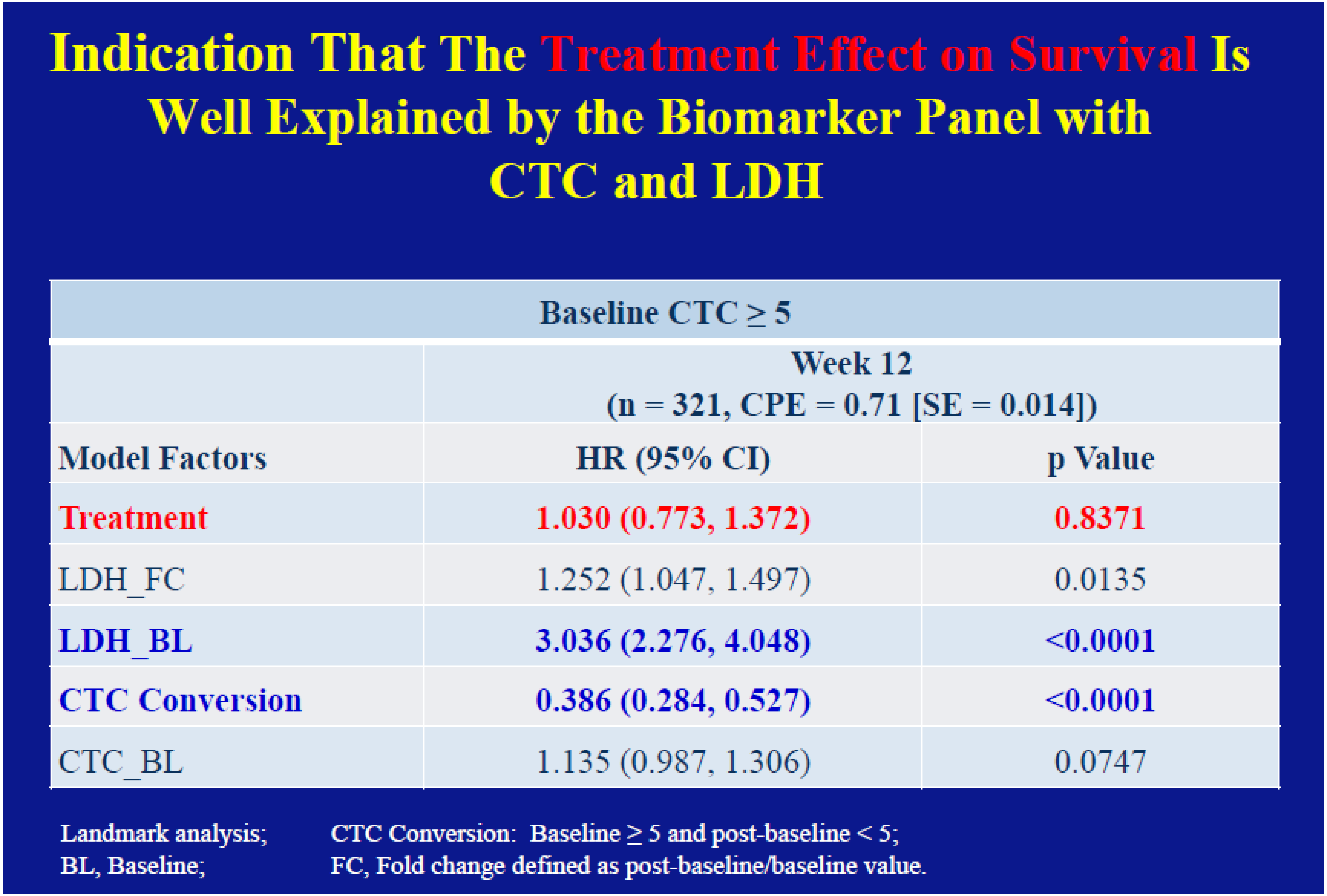
Treatment, Baseline LDH and CTC Count Were Prognostic for Survival in the Multivariate Model While PSA Was Not

	Baseline (n = 949, CPE = 0.70 [SE = 0.008])	
Biomarker	**HR (95% CI)**	**p Value**
Treatment	**0.70 (0.59, 0.828)**	**< 0.0001**
LDH	**2.98 (2.496, 3.565)**	**< 0.0001**
CTC count	**1.19 (1.137, 1.245)**	**< 0.0001**
Hgb	0.95 (0.891, 1.001)	0.0574
ALP	0.98 (0.874, 1.097)	0.7218
PSA	**1.04 (0.983, 1.093)**	**0.1797**

PSA, prostate-specific antigen; Hgb, hemoglobin; LDH, lactase dehydrogenase; ALP, alkaline phosphatase.

Ref: Scher et al, ASCO 2011

Is CTC conversion a surrogate endpoint?



Indication That The Treatment Effect on Survival Is Well Explained by the Biomarker Panel with CTC and LDH

	Baseline CTC ≥ 5	
	Week 12 (n = 321, CPE = 0.71 [SE = 0.014])	
Model Factors	**HR (95% CI)**	**p Value**
Treatment	**1.030 (0.773, 1.372)**	**0.8371**
LDH_FC	1.252 (1.047, 1.497)	0.0135
LDH_BL	**3.036 (2.276, 4.048)**	**<0.0001**
CTC Conversion	**0.386 (0.284, 0.527)**	**<0.0001**
CTC_BL	1.135 (0.987, 1.306)	0.0747

Landmark analysis; CTC Conversion: Baseline ≥ 5 and post-baseline < 5;
BL, Baseline; FC, Fold change defined as post-baseline/baseline value.

Ref: Scher et al, ASCO 2011

Is CTC conversion a surrogate endpoint?

The Prentice criteria are fulfilled by CTC conversion:

- Highly significant prognostic impact of CTC conversion on OS:

  $\gamma_Z = log\ (0.386) = -0.95\ (P < 0.0001)$

- Highly significant treatment effect on OS:

  $\beta\ = log\ (0.7) = -0.36\ (P < 0.0001)$

- No treatment effect on OS after adjustment for CTC conversion:

  $\beta_S = log\ (1.03) = 0.03\ (P = 0.83)$

Yet, do these results provide convincing evidence that CTC conversion is a valid surrogate for OS?

Problems with Prentice's approach

- Requires significant treatment effects on surrogate and true endpoints
- Rooted in hypothesis testing (impossible to prove the null H_0: $\beta_S \neq 0$ in finite samples)
- Does not quantify the predictive ability of a surrogate

Ref: Buyse and Molenberghs, Biometrics 1998, 54: 1014.

The proportion explained

The proportion explained is defined as

$$PE = 1 - \frac{\beta_S}{\beta}$$

For a good surrogate, $PE \cong 1$

Ref: Freedman et al, Stat in Med 1989, 8: 431.

Problems with the proportion explained

- Confidence limits for PE are wide
- PE is not a proportion
- PE can lie anywhere on the real line !

Refs: Lin et al, Stat in Med 1997, 16: 1515;
Buyse and Molenberghs, Biometrics 1998, 54: 1014.

Beyond the proportion explained

The proportion explained can be re-expressed as

$$PE = \lambda \cdot \frac{\gamma_Z}{RE}$$

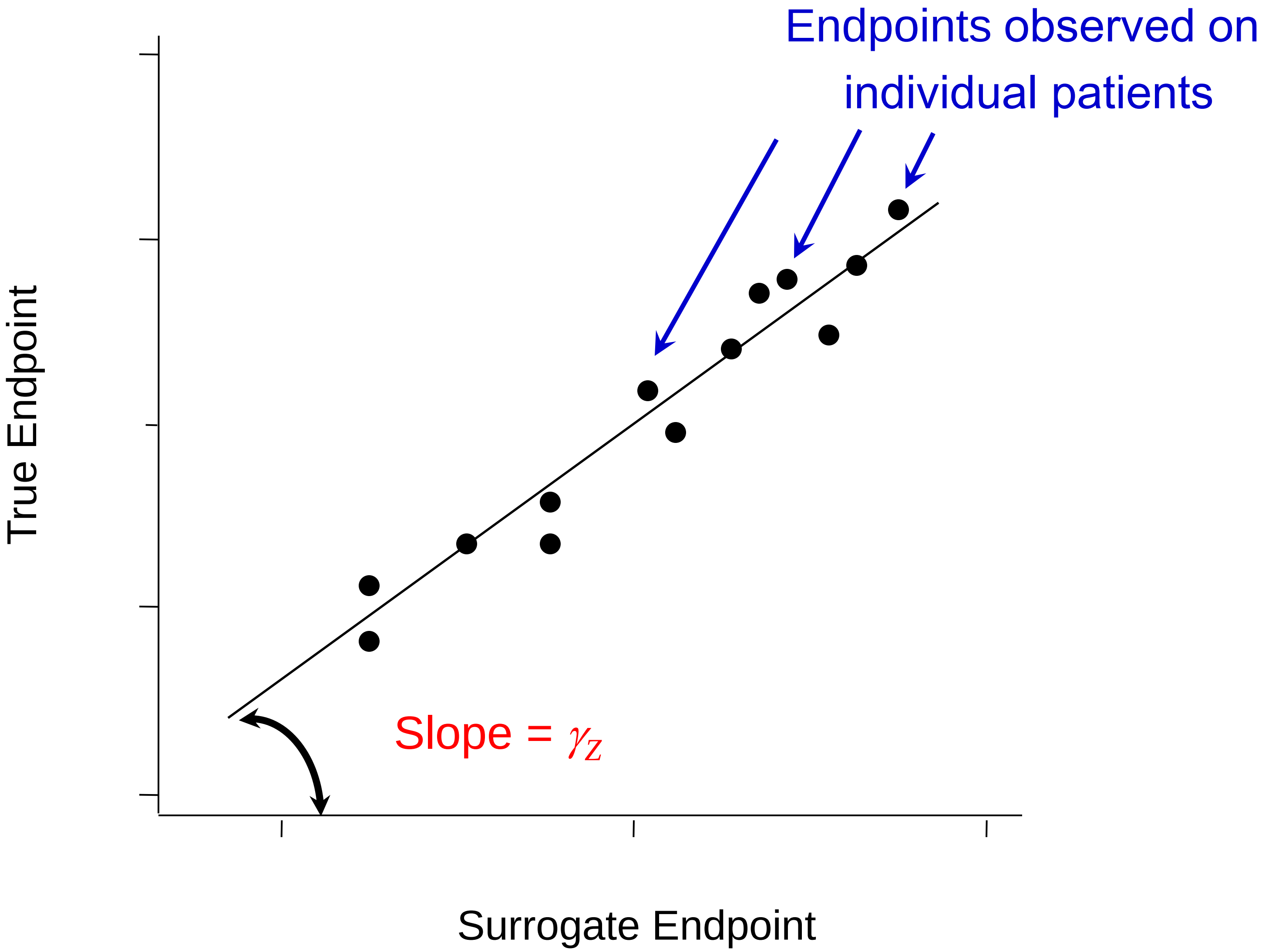
where RE is the relative effect and λ^2 a ratio of variances

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

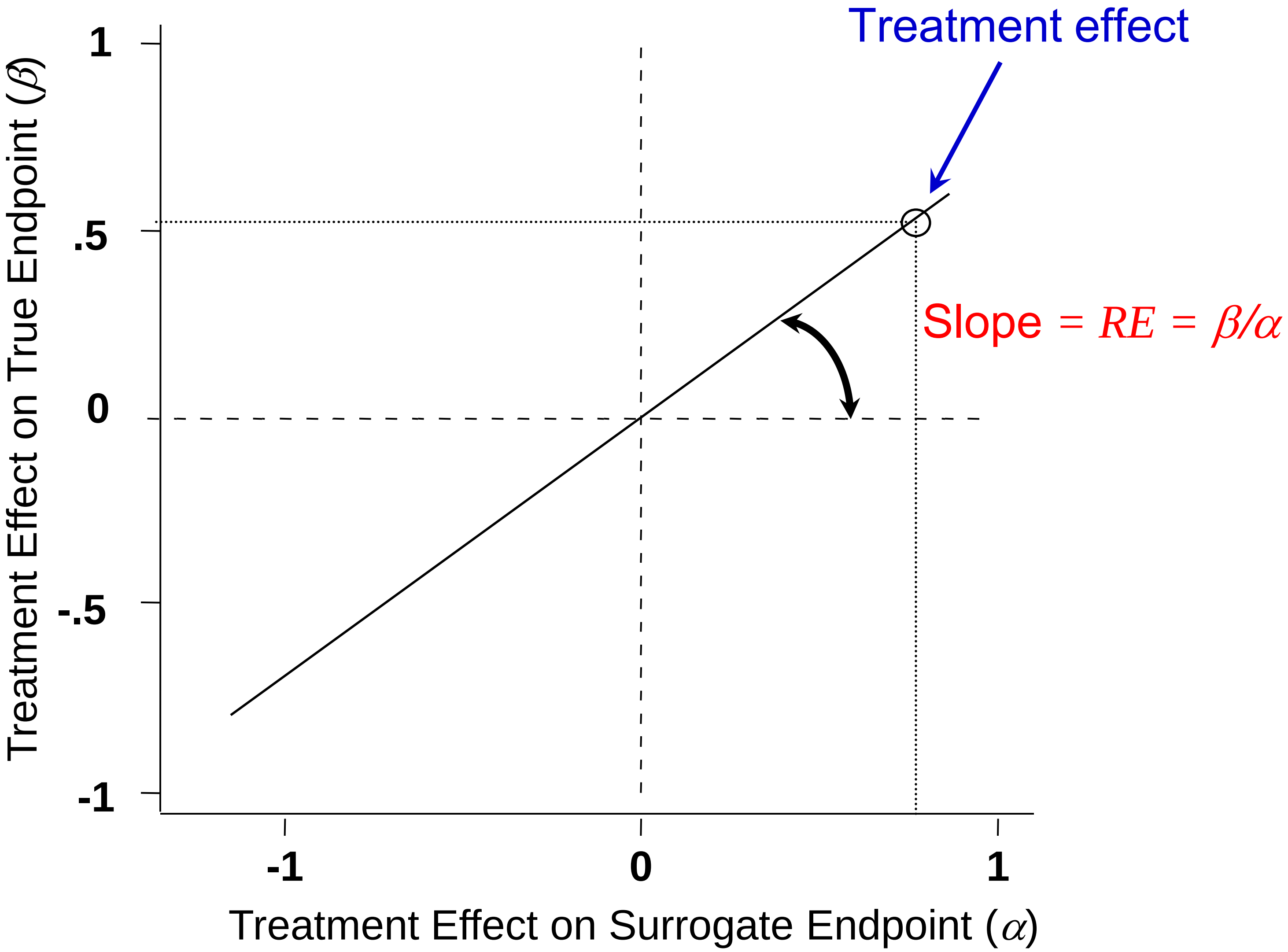
$$\lambda^2 = \frac{\sigma_{TT}}{\sigma_{SS}}$$

Ref: Molenberghs et al, Controlled Clin Trials 2002, 23: 607.

Prediction of true endpoint from surrogate endpoint



Prediction of treatment effect (regression through the origin !)



Need for multiple trials

For a marker to be used as a surrogate, we need *"repeated demonstrations of a strong correlation between the marker and the clinical outcome".*

And,

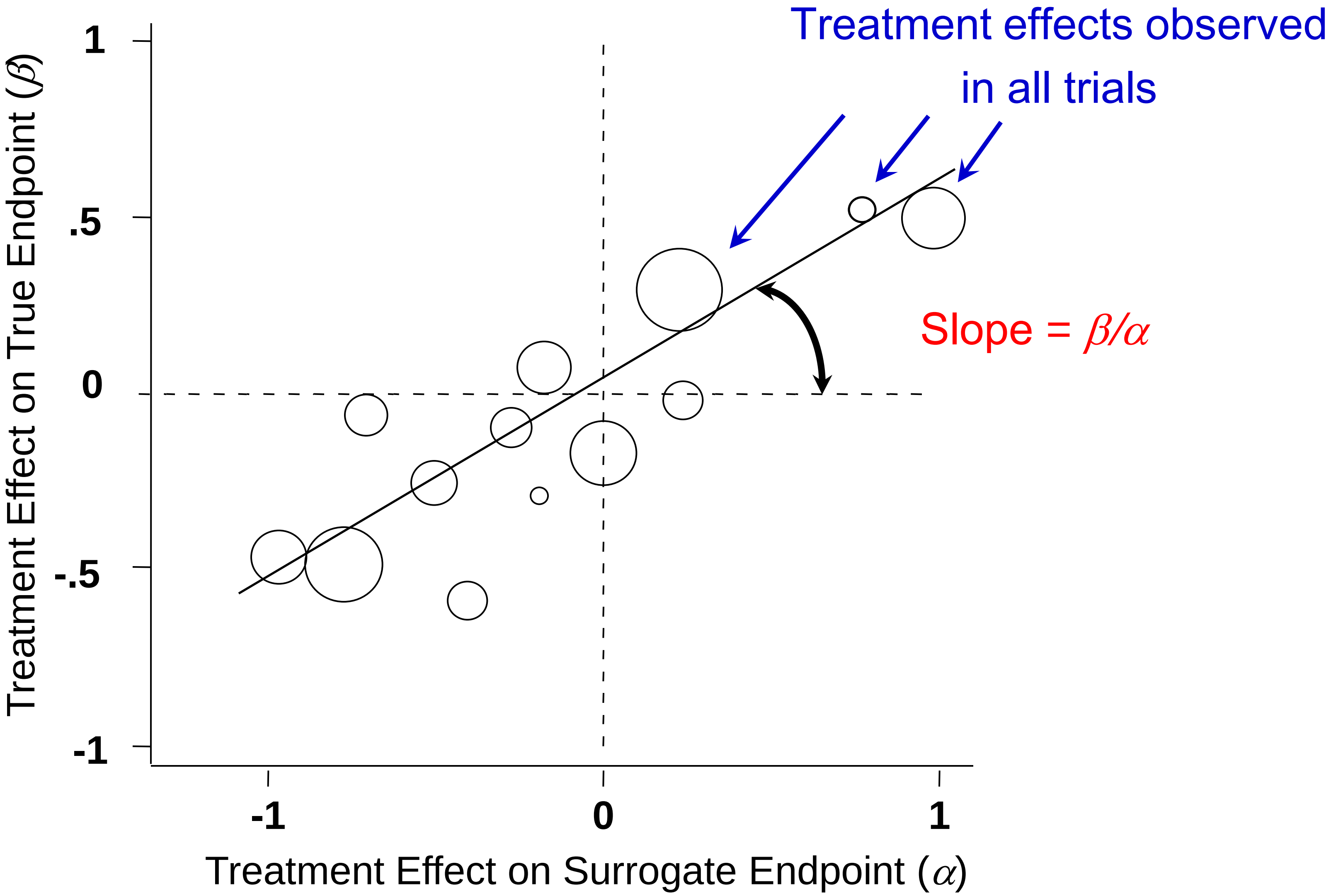
"there has been little work on alternative statistical approaches. A meta-analysis approach seems desirable to reduce variability".

Refs: Holland, 9th EUFEPS Conference on "Optimising Drug Development: Use of Biomarkers", Basel, 2001;

Albert et al, Stat in Med 1998,17: 2435.

1. Capture of effect in a single randomized trial
2. Association measures in meta-analyses
3. Prediction
4. Causal inference

Prediction of treatment effect: multiple trials



Measures of association at two levels

At the individual level, correlation between the endpoints (generalizes γ_Z)

$$R^2_{indiv}$$

At the trial level, correlation between the treatment effects on the endpoints

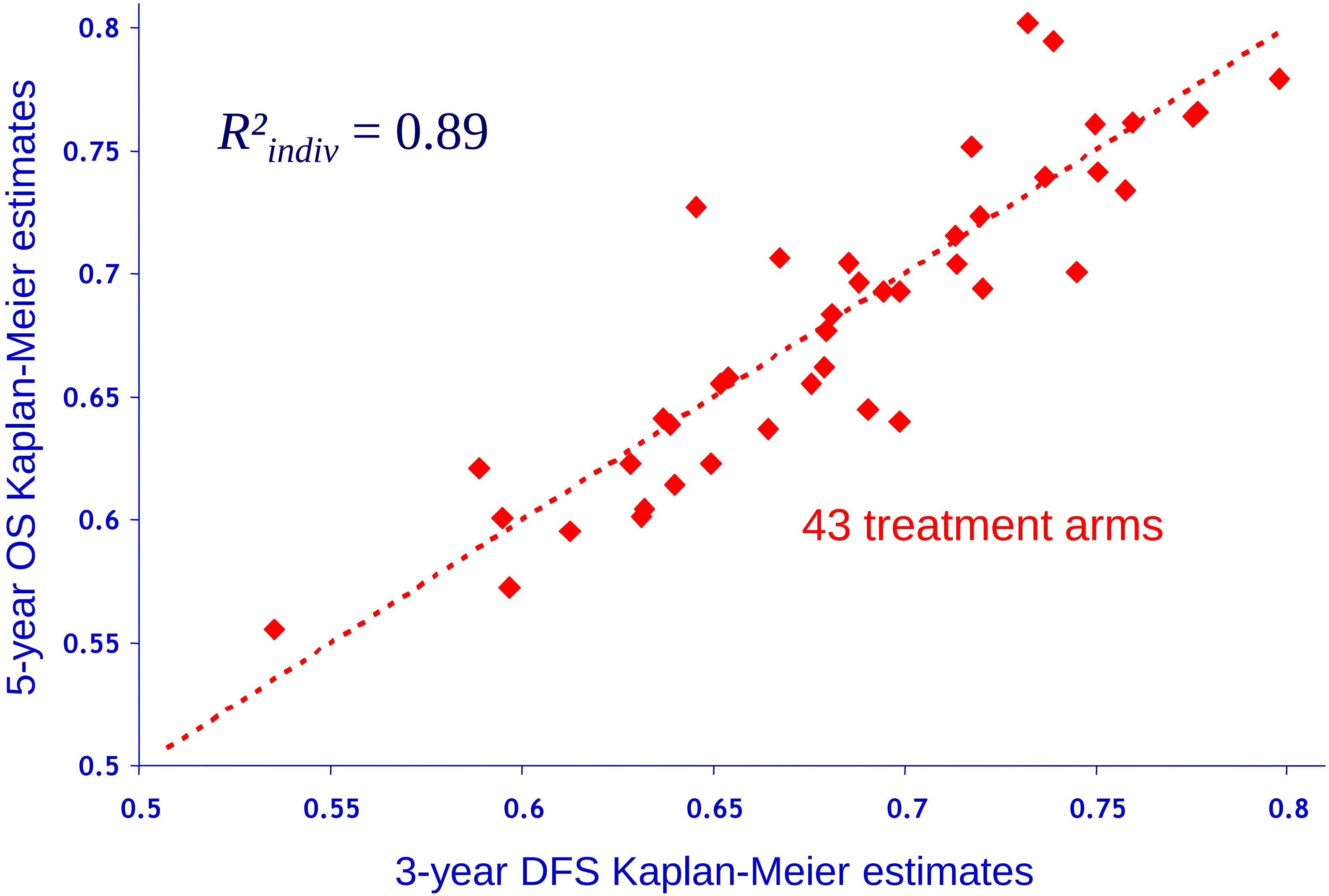
$$R^2_{trial}$$

Early colorectal cancer :
DFS as a surrogate for OS

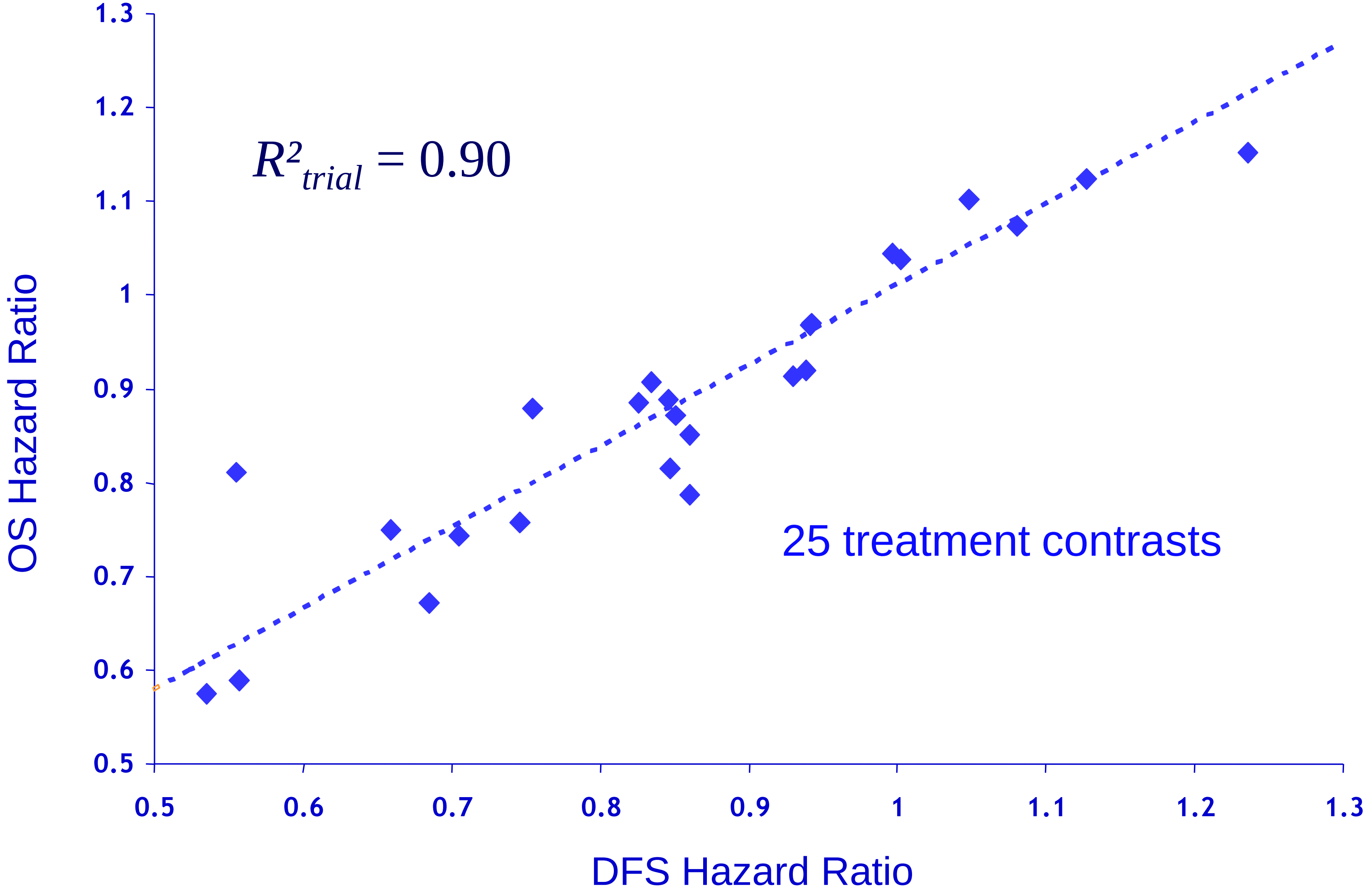
- Patients with early colorectal cancer, after resection of primary tumor
- Units of analysis : 20,898 patients in 18 randomized trials (25 treatment contrasts)
- Treatments: 5FU-based therapy *vs.* control or another 5FU-based therapy (43 treatment arms)
- Surrogate endpoint: disease-free survival (DFS)
- True endpoint: survival (OS)

Ref: Sargent et al, JCO 2005, 23: 8664.

Correlation between endpoints



Correlation between treatment effects



1. Capture of effect in a single randomized trial
2. Association measures in meta-analyses
3. Prediction
4. Causal inference

The “Surrogate Threshold Effect” (STE)

The “Surrogate Threshold Effect” is the treatment effect on the surrogate that would predict a statistically significant treatment effect on the true endpoint.

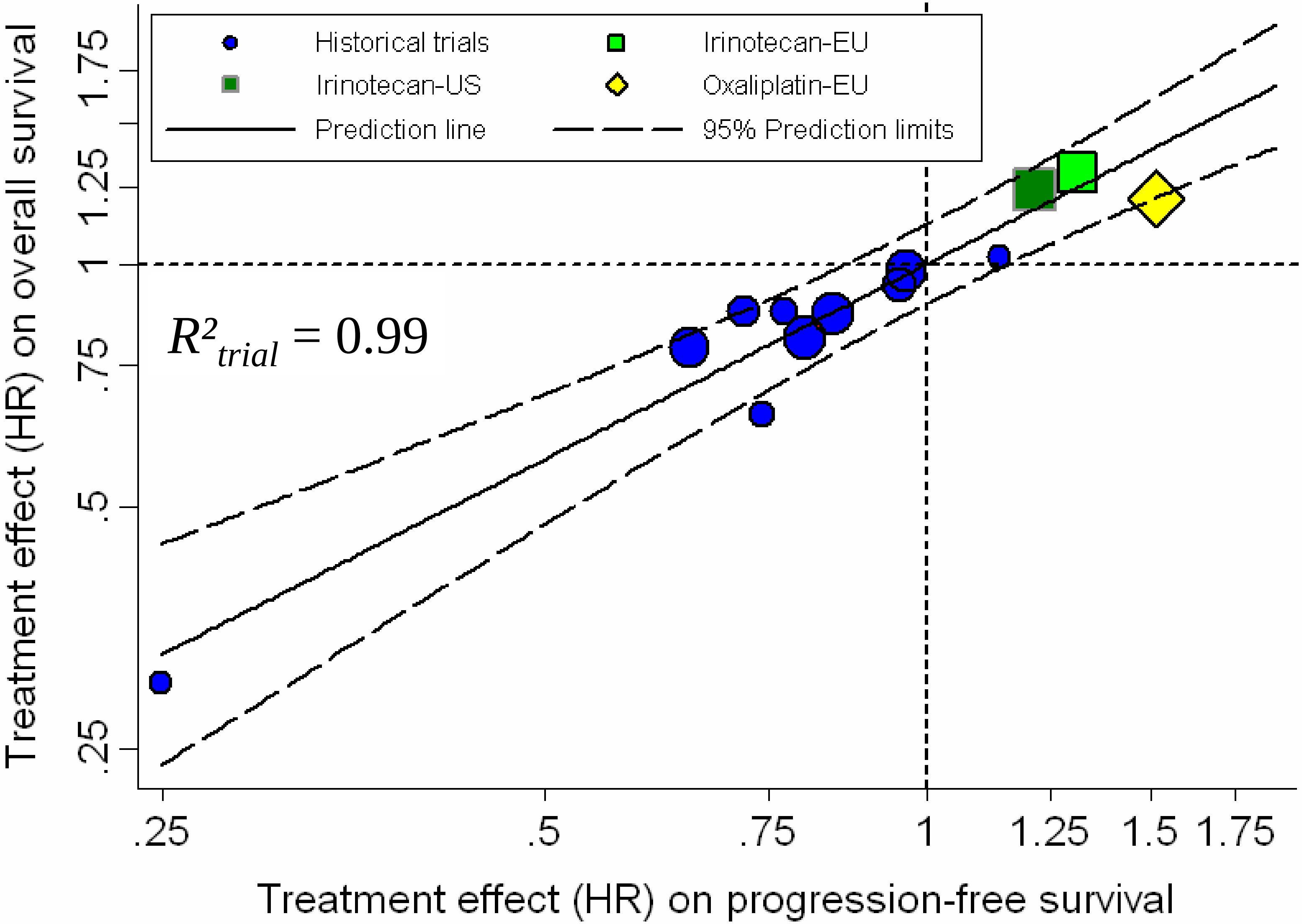
Ref: Burzykowski and Buyse, Pharmaceutical Stat 2006, 5: 173.

Advanced colorectal cancer:
PFS as a surrogate for OS

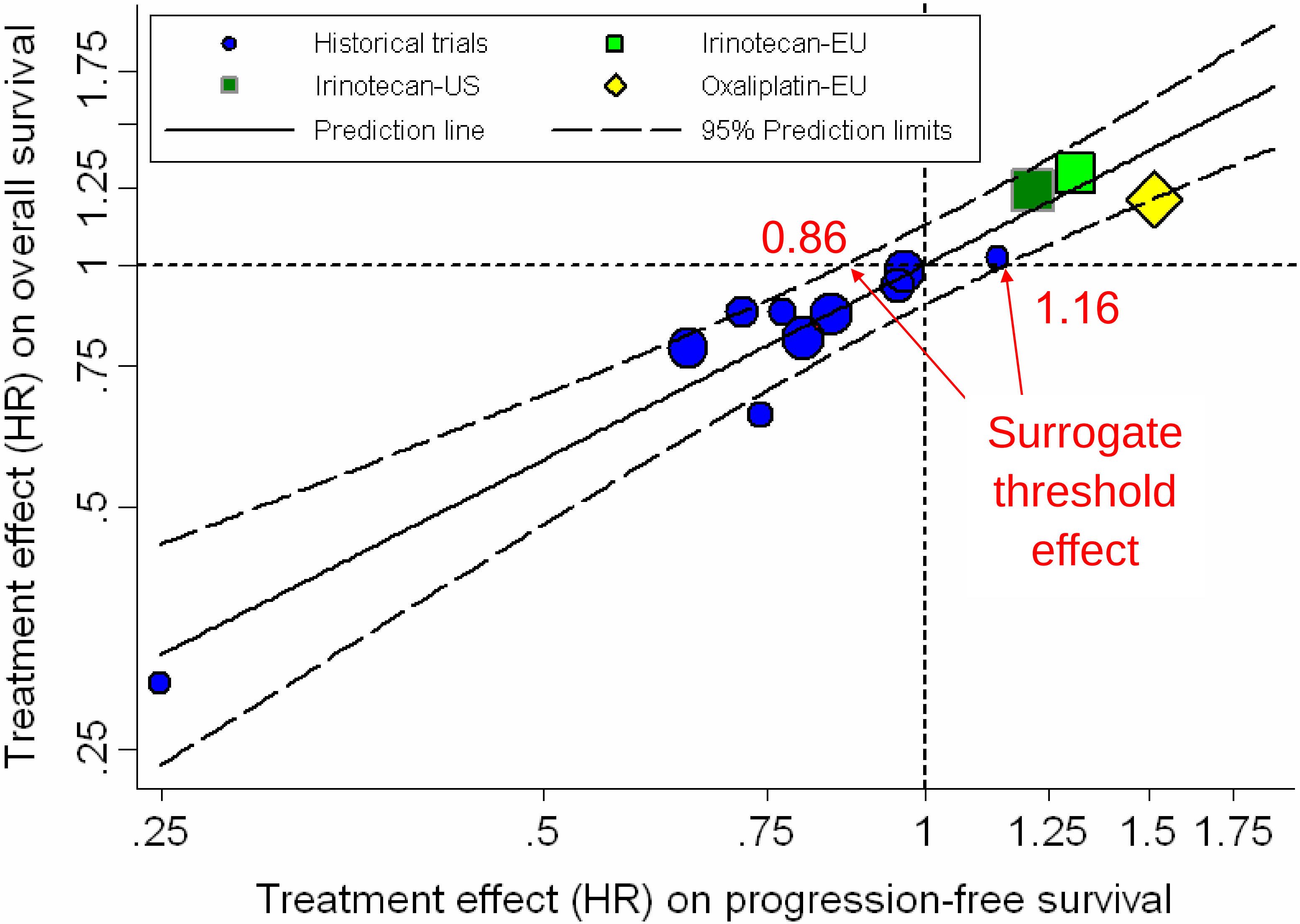
- Patients with advanced (metastatic) colorectal cancer
- Units of analysis: 4,352 patients in 13 trials
- Treatments (5FU/LV common arm):
  - 10 historical trials
    5FU *vs.* 5FU/L
  - 3 validation trials
    oxaliplatin or irinotecan + 5FU/LV *vs.* 5FU/LV
- Surrogate endpoint: PFS
- True endpoint: OS

Ref: Buyse et al, J Clin Oncol 2007, 25: 5218.

Correlation between treatment effects



The "Surrogate Threshold Effect" (STE)



0.86

1.16

Surrogate threshold effect

1. Capture of effect in a single randomized trial
2. Association measures in meta-analyses
3. Prediction
4. Causal inference

Return to early colorectal cancer

Randomized treatment	Z	0	Control (no treatment or standard chemotherapy)
		1	Experimental chemotherapy
Surrogate endpoint: DFS at 3 years	S	0	Recurrent disease or death within 3 years
		1	Alive without recurrence at 3 years
True endpoint: OS at 5 years	T	0	Dead within 5 years
		1	Alive at 5 years

Ref: Li et al, Biometrics 2010;66:523.

Counterfactual outcomes

Each subject has four *potential outcomes*, denoted $T_i(0)$ and $T_i(1)$ for the true endpoint, and $S_i(0)$ and $S_i(1)$ for the surrogate endpoint.

Subject	Z	$T(0)$	$T(1)$	$S(0)$	$S(1)$
1	0	1	?	1	?
2	1	?	1	?	1
3	1	?	0	?	0
4	0	0	?	1	?
5	1	?	1	?	1
…					

Counterfactual probabilities

Denote p_{11} to p_{44} the counterfactual probabilities :

	$(T(0), T(1))$			
$(S(0), S(1))$	$(0,0)$	$(0,1)$	$(1,1)$	$(1,0)$
$(0,0)$	p_{11}	p_{12}	p_{13}	p_{14}
$(0,1)$	p_{21}	p_{22}	p_{23}	p_{24}
$(1,1)$	p_{31}	p_{32}	p_{33}	p_{34}
$(1,0)$	p_{41}	p_{42}	p_{43}	p_{44}

Principal stratification

Frangakis and Rubin (*Biometrics* 2002) define the *principal stratification* for the surrogate endpoint:

	$(T(0), T(1))$				
$(S(0), S(1))$	(0,0)	(0,1)	(1,1)	(1,0)	Principal stratification
(0,0)	p_{11}	p_{12}	p_{13}	p_{14}	Never responders
(0,1)	p_{21}	p_{22}	p_{23}	p_{24}	Improved
(1,1)	p_{31}	p_{32}	p_{33}	p_{34}	Always responders
(1,0)	p_{41}	p_{42}	p_{43}	p_{44}	Harmed

Ref: Frangakis and Rubin, Biometrics 2002;58:21.

Principal surrogate

For a good surrogate, subjects who are improved (or harmed) on the surrogate must also be improved (or harmed) on the true endpoint

		$(T(0), T(1))$			
$(S(0), S(1))$	(0,0)	(0,1)	(1,1)	(1,0)	Principal stratification
(0,0)	p_{11}	p_{12}	p_{13}	p_{14}	Never responders
(0,1)	p_{21}	p_{22}	p_{23}	p_{24}	Improved
(1,1)	p_{31}	p_{32}	p_{33}	p_{34}	Always responders
(1,0)	p_{41}	p_{42}	p_{43}	p_{44}	Harmed

For a « principal » surrogate, p_{22} / p_{2+} and p_{44} / p_{4+} must be close to 1 (p_{i+} denotes the number of subjects in principal stratum *i*).

Ref: Frangakis and Rubin, Biometrics 2002;58:21.

Associative proportion

Surrogate associative proportion: $(p_{22} + p_{42} - (p_{24} + p_{44}))/(p_{2+} - p_{4+})$

Associative proportion: $(p_{22} + p_{42} - (p_{24} + p_{44}))/(p_{+2} - p_{+4})$

	$(T(0), T(1))$				
$(S(0), S(1))$	(0,0)	(0,1)	(1,1)	(1,0)	Principal stratification
(0,0)	p_{11}	p_{12}	p_{13}	p_{14}	Never responders
(0,1)	p_{21}	p_{22}	p_{23}	p_{24}	Improved
(1,1)	p_{31}	p_{32}	p_{33}	p_{34}	Always responders
(1,0)	p_{41}	p_{42}	p_{43}	p_{44}	Harmed

Ref: Frangakis and Rubin, Biometrics 2002;58:21.

Associative proportion assuming monotonicity

Surrogate associative proportion: $p_{22} \,/\, p_{2+}$

Associative proportion: $p_{22} \,/\, p_{+2}$

	$(T(0), T(1))$				
$(S(0), S(1))$	(0,0)	(0,1)	(1,1)	(1,0)	Principal stratification
(0,0)	p_{11}	p_{12}	p_{13}	0	Never responders
(0,1)	p_{21}	p_{22}	p_{23}	0	Improved
(1,1)	p_{31}	p_{32}	p_{33}	0	Always responders
(1,0)	0	0	0	0	Harmed

Note: $p_{i4} = 0 \ \forall \ i$ and $p_{4j} = 0 \ \forall \ j$

Ref: Li et al, Biometrics 2010;66:523.

Early colorectal cancer

Surrogate associative proportion = .54 (-1.33; 2.19)

Associative proportion = .83 (-2.02; 3.19)

	$(T(0), T(1))$				
$(S(0), S(1))$	(0,0)	(0,1)	(1,1)	(1,0)	Principal stratification
(0,0)	.259	.001	.026	.001	Never responders
(0,1)	.001	.011	.021	.000	Improved
(1,1)	.021	.014	.619	.011	Always responders
(1,0)	.001	.000	.014	.001	Harmed

Ref: Li et al, Biostatistics 2011;12:478.

Early colorectal cancer with monotonicity

Surrogate associative proportion, $p_{22} \,/\, p_{2+} = .10\ (.00 - .50)$

Associative proportion, $p_{22} \,/\, p_{+2} = .12\ (.00 - .64)$

	$(T(0), T(1))$				
$(S(0), S(1))$	(0,0)	(0,1)	(1,1)	(1,0)	Principal stratification
(0,0)	.259	.003	.037	0	Never responders
(0,1)	.003	.002	.012	0	Improved
(1,1)	.029	.008	.647	0	Always responders
(1,0)	0	0	0	0	Harmed

Note: $p_{i4} = 0\ \forall\ i$ and $p_{4j} = 0\ \forall\ j$

Ref: Li et al, Biostatistics 2011;12:478.

Problems with causal inference

- Estimation of counterfactual probabilities through complex model (e.g. Bayesian)
- Results sensitive to restrictive assumptions (e.g. monotonicity)
- Poor estimates (large confidence intervals)
- No agreed upon measure of surrogacy
- Without monotonicity, the net associative proportion is a ratio of differences; its values span [–∞, +∞]

Ref: Li et al, Biostatistics 2011;12:478.

1. Capture of effect in a single randomized trial
2. Association measures in meta-analyses
3. Prediction
4. Causal inference
5. Conclusions

Conclusions

- There are no absolute standards for surrogacy
- Even so, some intermediate endpoints (DFS, PFS in colorectal cancer) or biomarkers (CTCs in prostate cancer) have undergone « validation »
- Principal surrogacy is more principled than statistical surrogacy, but causal inference is challenging
- Large sets of randomized data are required (typically, meta-analyses of RCTs)
- If a surrogate is shown valid under specific conditions (treatment / environment), is it still valid under different conditions (e.g. an experimental treatment)?
- Good (let alone perfect) surrogates are hard to find!

Surrogate outcome measures?

Get your facts first, and then you can distort them as much as you please.

Mark Twain