Adaptive Biomarker Population Selection in Confirmatory Phase III Oncology Trials

Cong Chen and Nicole Li

FDA/Industry Statistics Workshop 2015

Outline

- Introduction
- Application of biomarker population selection to a hypothetical survival trial
  - Type I error control
  - Power
- Discussion

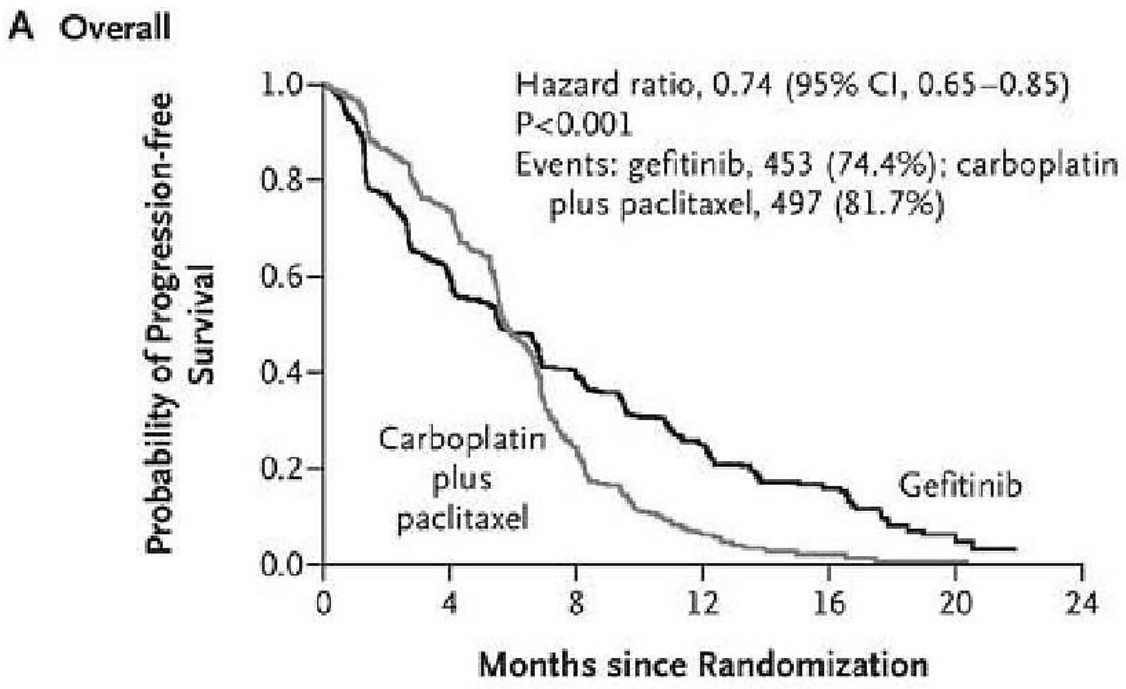
Predictive biomarkers in oncology

- Sometimes hit a jackpot
  - Vemurafenib in BRAF V600E mutant melanoma
  - Crizotinib in ALK-positive non-small cell lung cancer
- Sometimes get surprised
  - Gefitinib was approved for 2nd line lung cancer, but was found to only work in EGFR-mutated patients later
  - EGFR expression doesn’t predict response to Cetuximab as hypothesized, but K-ras wt genotype likely does

Challenges of Phase III oncology trials

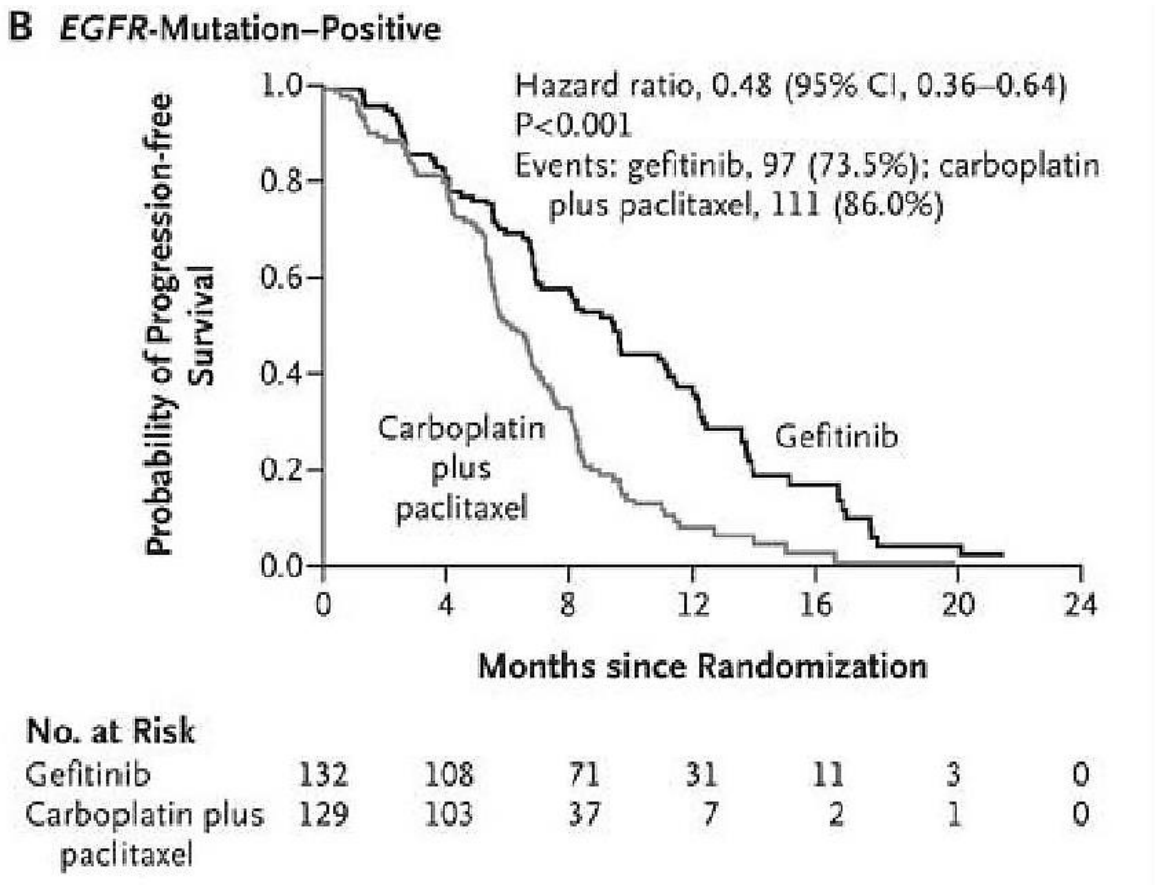
- Most Phase III confirmatory trials are initiated at risk after anti-cancer activities are seen in small phase I/II studies
  - Limited biomarker data, and biomarker effect is confounded with prognostic effect
- Inclusion of non-performing subpopulations makes study design less efficient/powerful
- A statistically significant outcome overall but clinically underwhelming outcome in biomarker subpopulation(s) present challenges to reimbursement and labeling

IPASS – overall population

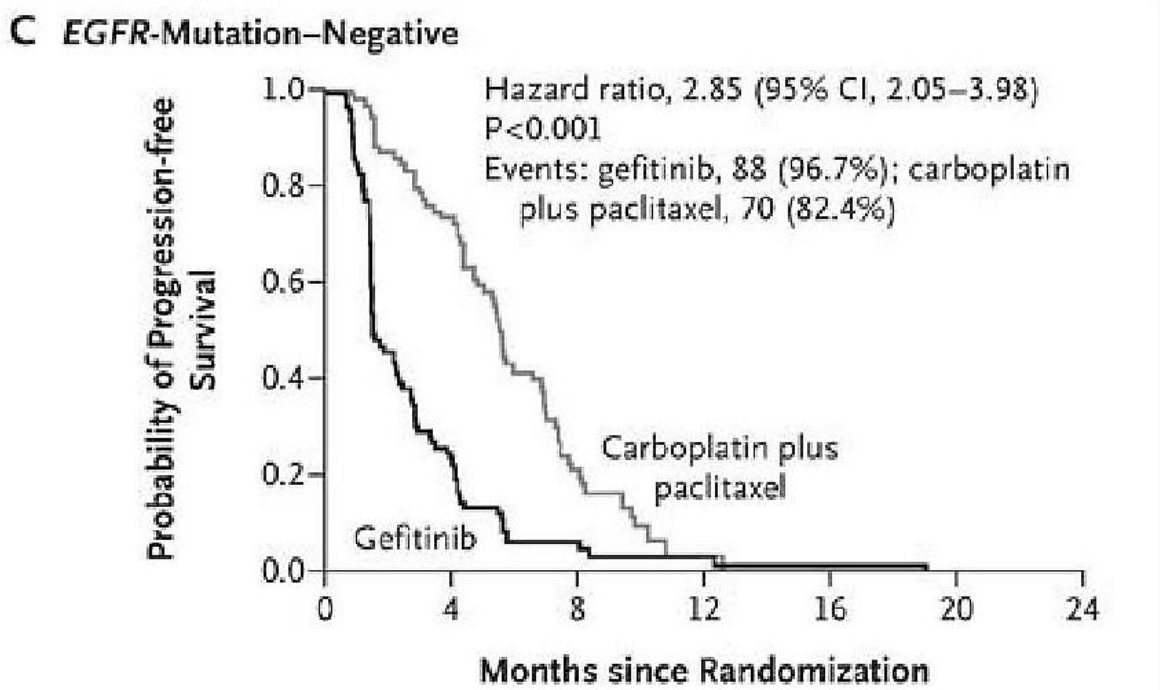


No. at Risk							
Gefitinib	609	363	212	76	24	5	0
Carboplatin plus paclitaxel	608	412	118	22	3	1	0

EGFR – mutation positive



EGFR – mutation negative



Biomarker population selection

- Conduct an interim analysis to de-select non-performing biomarker subpopulations
- Final analysis is conducted in remaining subpopulation(s) with option to step-down to overall
  - Data on overall population is hard to interpret if accrual for de-selected subpopulations discontinues, and/or sample size for remaining subpopulation(s) increases after the interim analysis (*enrichment*)
- Magnusson and Turnbull (2013) used same endpoint for interim analysis and final analysis, what if different endpoints are used?

A hypothetical example

- Treatment effect is assumed to be ordered by BM level from low to high with equal prevalence
- A randomized controlled trial targets 330 OS events overall (~600 patients) so that the study has 90% power to detect a 0.70 HR in OS at 2.5% alpha level
- An interim analysis of PFS is conducted when 165 deaths (t=50%) and 250 PFS events are observed
  - Exclude BM low in final analysis if p-value > α_t
  - Stop study if p-value for BM high is further > α_t
- What is the nominal alpha (α^*) at final analysis of OS to maintain overall type I error rate at 2.5%?

Nominal type I error at final analysis (α^*)

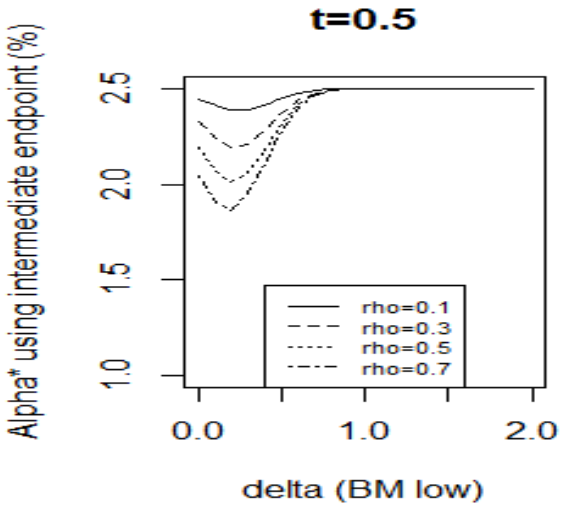
- Overall type I error under the null hypothesis of no OS effect ($\Delta_1 = \Delta_2 = 0$) without any constraint on PFS effects (δ_1, δ_2)
  - Jenkins, Stone and Jennison (2010) assumed no PFS effect but this assumption is too strong in practice
  - (d_1, d_2) are information units in (BM low, BM high) at interim

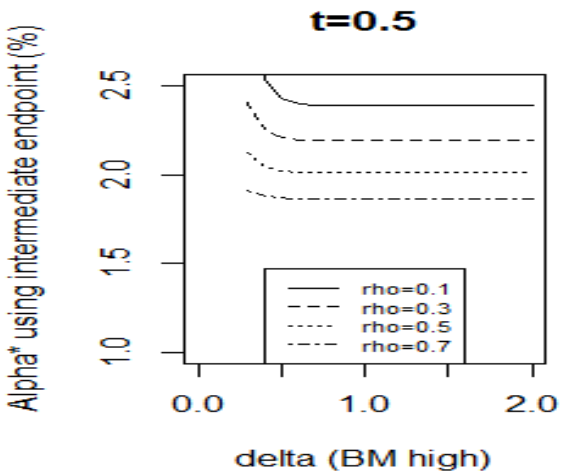
$$P\left(X_{l1} > Z_{1-\alpha_t}, V_{all} > Z_{1-\alpha^*} | (\delta_1\sqrt{d_1 t}, \Delta_1 = \Delta_2 = 0)\right) +$$
$$P(X_{l1} < Z_{1-\alpha_t}, X_{h1} > Z_{1-\alpha_t}, V_{h2} > Z_{1-\alpha^*} | (\delta_1\sqrt{d_1 t}, \delta_2\sqrt{d_2 t}, \Delta_2 = 0)$$

- Minimal α^* in the entire (δ_1, δ_2) space that keeps above overall type I error at 0.025 is the nominal alpha for final analysis
  - $\delta_1 = \delta_2 = 0$ when OS is used for both interim analysis (for biomarker population de-selection) and final analysis

α^* by correlation between *X* and *V* (ρ)

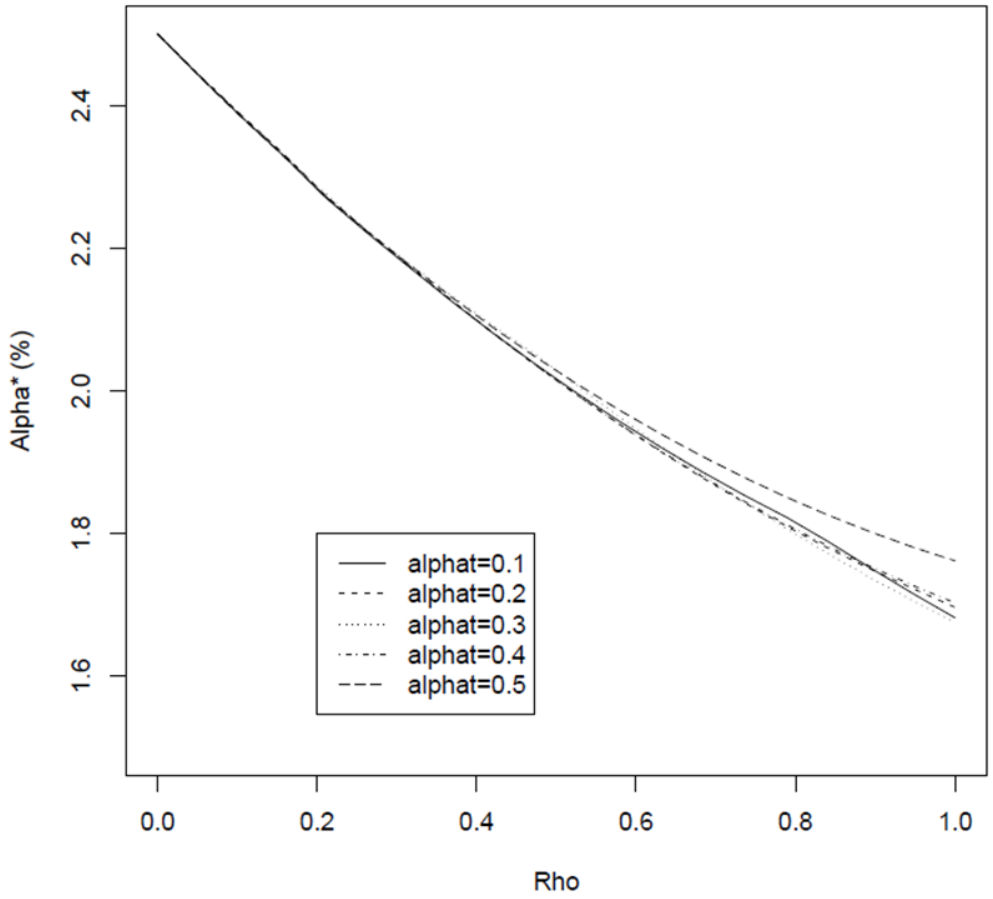
- Minimal α^* is reached at nontrivial $\delta_1 < \delta_2$
- Lower α^* is associated with higher ρ
- ρ can be estimated from trial data and alpha is controlled as long as a consistent estimator of ρ (e.g., WLW method) is used for calculation of α^*





Minimal α^* by different de-selection rule (α_t)

- Interim analysis serves role of cherry-picking which inflates alpha but also shares similarity for futility stopping which deflates alpha
- Minimal α^* is robust to α_t in this hypothetical example

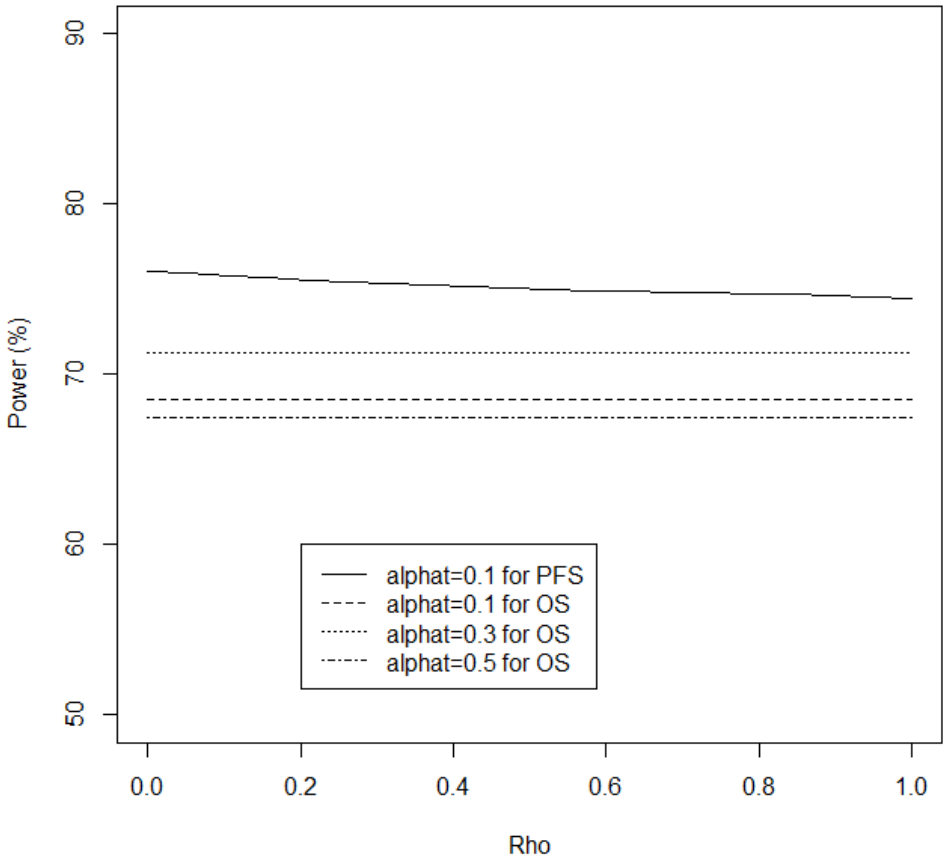


Set-up for power comparison

- True HR for OS is 0.6 in BM high and is 1 in BM low
  - The actual power without biomarker selection is 64%
- True HR for PFS is 0.45 in BM high and is 1 in BM low
- Why is it likely more powerful to use PFS for de-selection?
  - More sensitive than OS
  - More events

Power comparison

- Use of OS for de-selection
  - Highest power is achived at α_t=0.3 (~0.9 hazard ratio)
- Use of PFS for de-selection
  - α_t is conveniently chosen at 0.1 (~0.8 hazard ratio)
- All have higher power than no-selection (64%), and use of PFS has higher power than use of OS



Summary

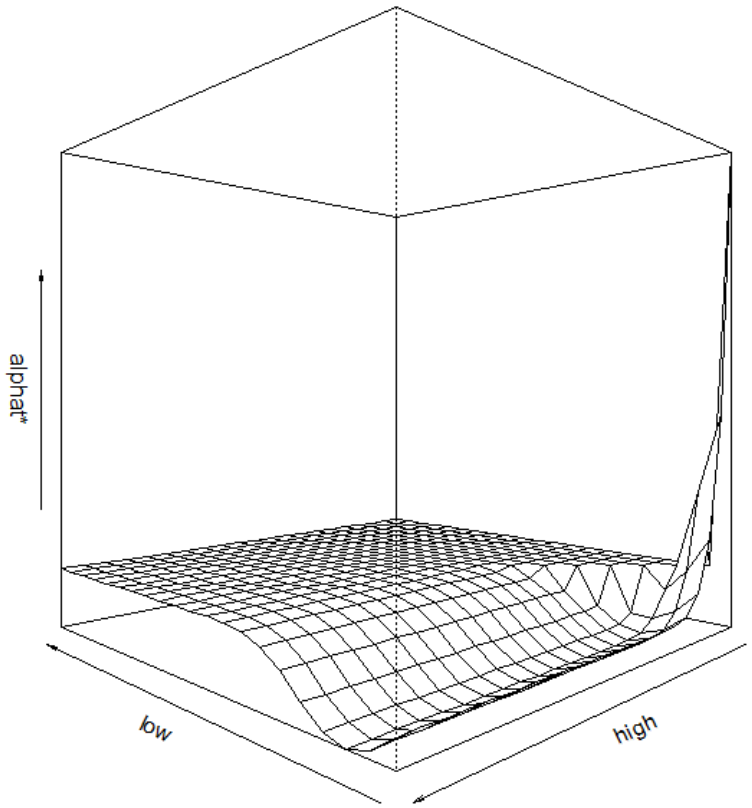
- De-selection of non-performing biomarker subpopulations helps increase the study power even after paying the penalty for multiplicity control
  - Type I error rate can be appropriately controlled whether same or different endpoints are used
- Use of a sensitive intermediate endpoint for selection helps further increase the study power in presence of non-performing subpopulation(s)
  - Treatment effects can be treated as nuisance parameters in type I error calculation
  - (It is trivial extension when sample size for BM high is increased after de-selection of BM low, or when there are more than 2 BM subpopulations)

Discussion

- A sensitive intermediate endpoint may not always exist
- Different approaches may be applied to biomarker population selection depending on prior information on biomarker effect
  - E.g., no ordering, no futility stopping
- When there is uncertainty about how treatment effect evolves over time, an *informational analysis* may be conducted in a subgroup of patients representing the overall population at end of the trial for biomarker population selection
  - Chen et al. *Informational Design of Confirmatory Phase III Trials for Expedited Development of Personalized Medicines,* unpublished manuscript

Back-up

Pattern of α^* under different (δ_1, δ_2)



Impact of sample size increase on study power

