Informational Designs of Phase III Trials for Expedited Development of Immuno-oncology Therapies with a Putative Predictive Biomarker

Cong Chen, PhD
Director, BARDS, Merck & Co., Inc.

BASS XXII, 2015

MERCK

Outline

- Introduction to informational design
- Adaptive alpha-allocation
  - Without use of unblinded trial data
  - With use of unblinded trial data
- Adaptive biomarker population selection
  - Same endpoint for selection and final analysis
  - Different endpoints for selection and final analysis
- Discussion

Media buzz about PD-1/PD-L1 After ASCO

"Investigational immune-boosting therapies from Merck, BMS and Roche 'are producing such promising early results that doctors ASCO are openly speculating that some patients with the deadliest form of skin cancer may be cured.'"
– Robert Langreth

Bloomberg

"[Merck] is squarely in the race to bring to market one of what many experts view as the most promising class of drugs in years." – Andrew Pollack

The New York Times

The data presented Sunday, on drugs being developed by Merck & Co. and Bristol-Myers Squibb Inc., were the talk of a huge five-day conference of cancer specialists in Chicago, hosted by the American Society of Clinical Oncology.

AP Associated Press

"Results seem to indicate [lambro] is potent, maybe more so than its competitors."
– Matthew Herper

Forbes.com

"One member of the audience commented (half-jokingly) that the lambrolizumab data were so convincing, the drug should just be approved now without any further clinical trials necessary." – Adam Feuerstein

The Street

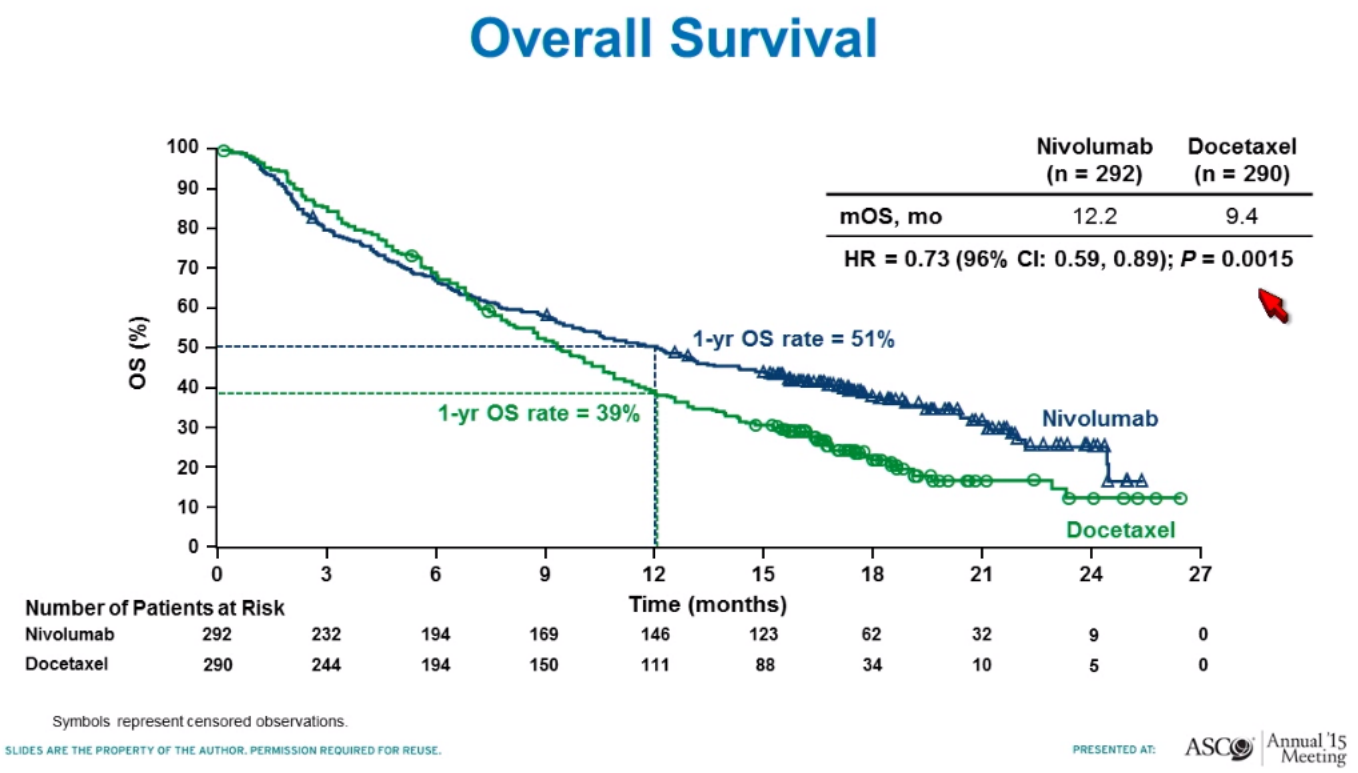
"I think all of you recognize this is a very special moment in oncology," Dr. Roger M. Perlmutter, head of research and development at Merck, told analysts Sunday at a standing-room-only meeting.

The New York Times

Expedited development amid uncertainties

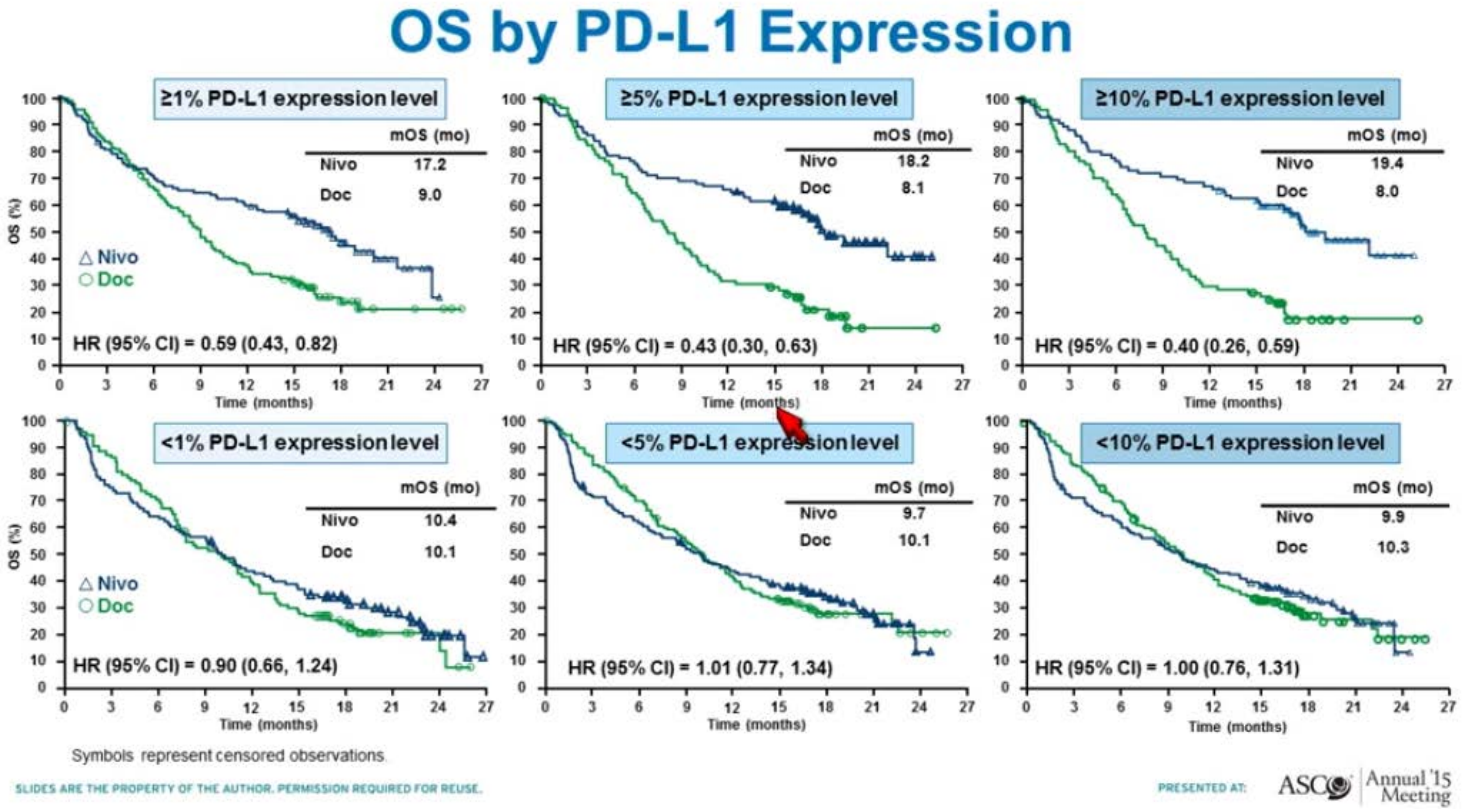
- Under fierce competition, Phase III confirmatory trials are often initiated at risk after preliminary anti-tumor activities are observed in small Phase I/II single arm studies.
  - Adjuvant or neo-adjuvant studies are often initiated w/o any data in same setting
- The preliminary data can hardly provide the much-needed information for selecting a biomarker subpopulation or prioritizing a biomarker hypothesis for Phase III testing
- The preliminary data seldom provides any insight on how the treatment effect evolves over time – a big headache!

Nivolumab in non-squamous lung



Phase III, randomized trial (CheckMate 057) of nivolumab (NIVO) versus docetaxel (DOC) in advanced non-squamous cell (non-SQ) non-small cell lung cancer (NSCLC).

Subgroup analysis of OS by PD-L1 expression



- How does hazard ratio evolve over time in PD-L1 high patients?
- What is the appropriate cutpoint for PD-L1 expression?
- Why KM curves overlap in patients with low PD-L1 expression?

Conventional designs

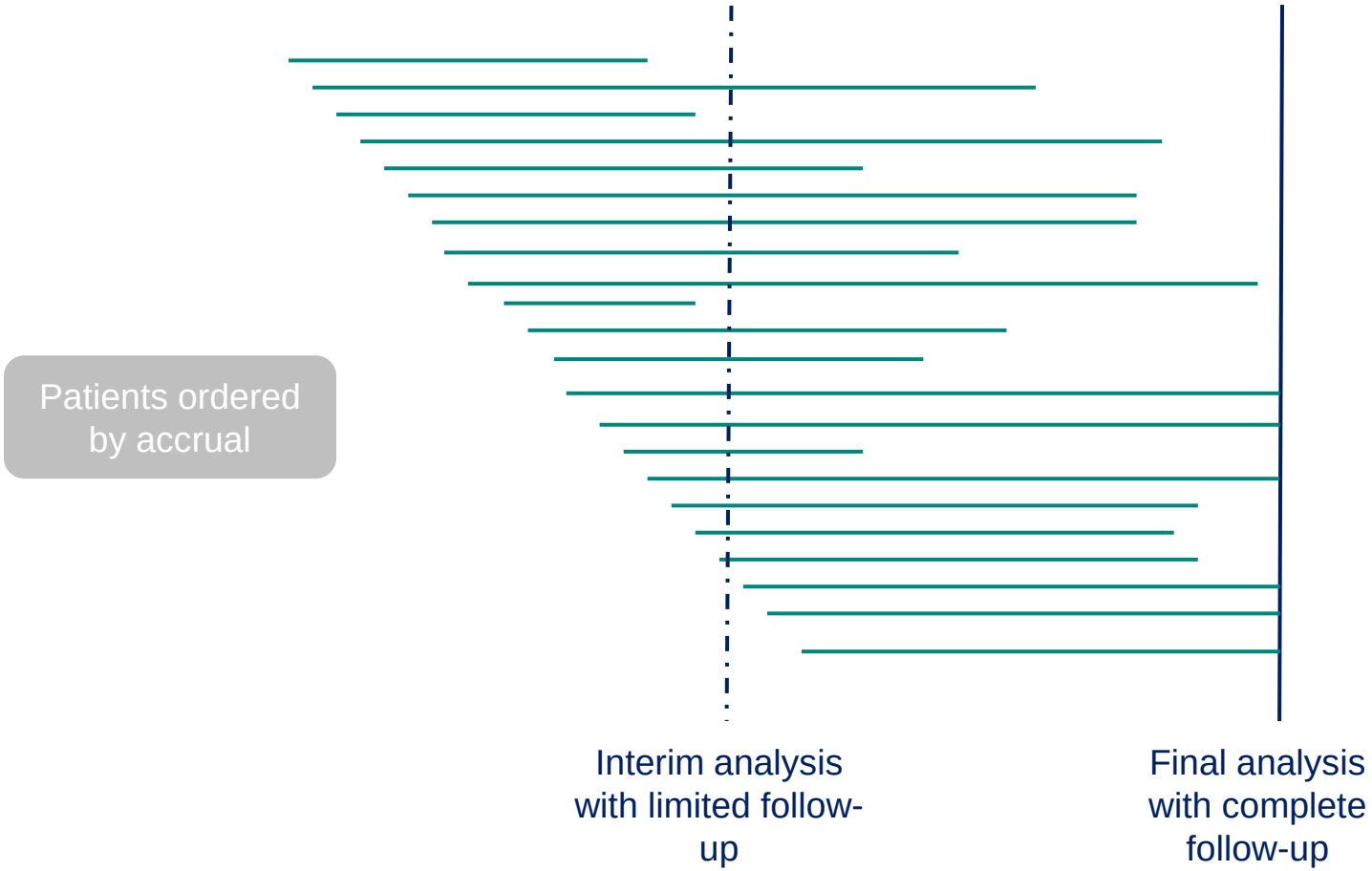
- Sequential Phase II followed by Phase III
  - Slow and susceptible to shift of treatment paradigm
- Seamless/adaptive Phase II/III
  - Treatment effect observed at an interim analysis may not be the same as in the final analysis due to mechanism of action, cross-over, or change in patient demographics
  - Use of an intermediate endpoint for decision may be unreliable because the predictive value of an intermediate endpoint is often unknown for drugs with a new mechanism of action, or in settings or populations with little experience

Informational Design

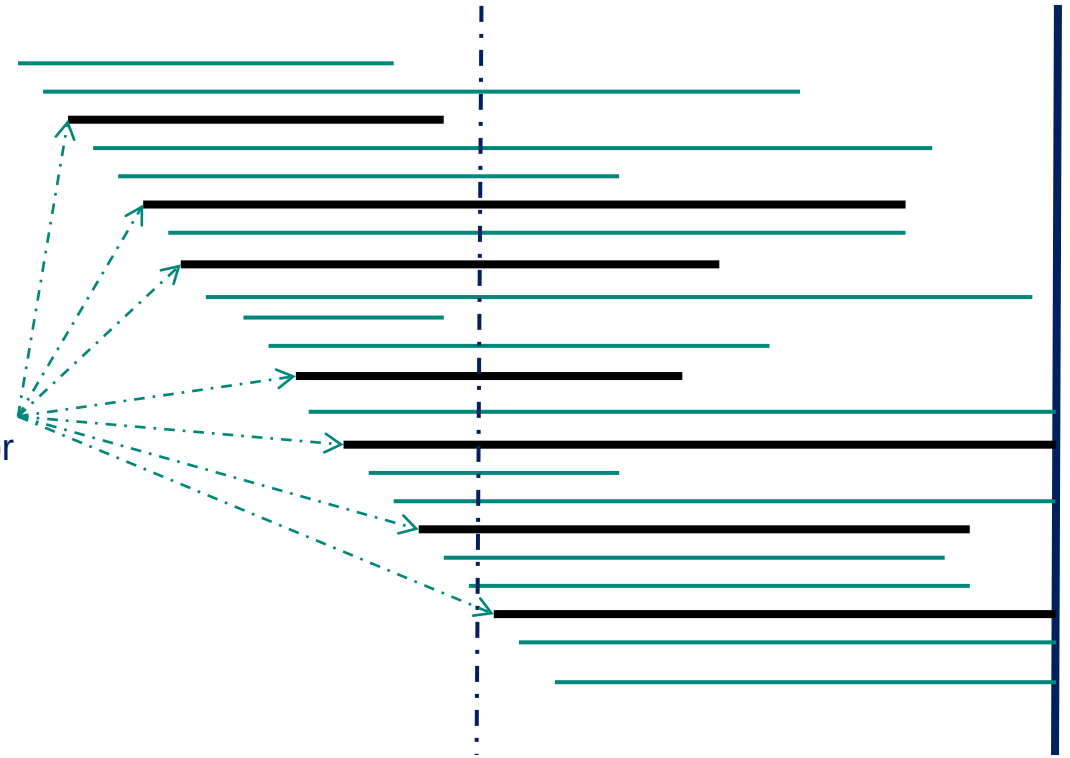
Informational analysis

- Add an analysis at end of the Phase III trial in a representative subset of patients (*sub-study*) for subpopulation selection and adaptive hypothesis adjustment
  - Two of every 10 patients are randomly selected if 20% of the trial information will be used in the analysis (e.g.)
  - The subgroup analysis is equivalent to a Phase II trial conducted under same clinical design at same time in same population at same sites as the Phase III trial
  - The informational analysis can be conduced earlier when an intermediate endpoint such as RR/PFS (vs OS as primary endpoint) is used for adaptive decision
- The patients in sub-study are included in final analysis

Conventional interim analysis



Informational analysis vs interim analysis



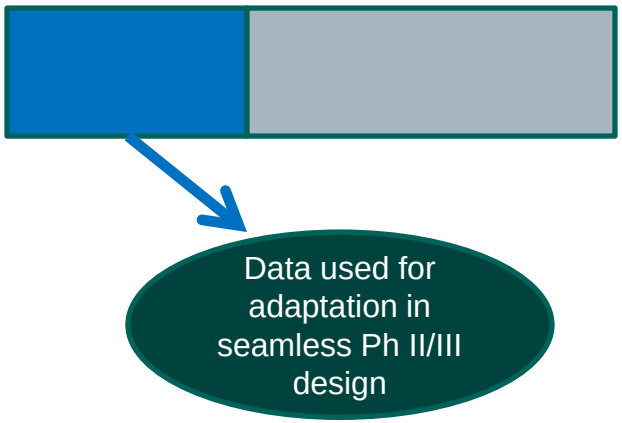
Patients selected for informational analysis

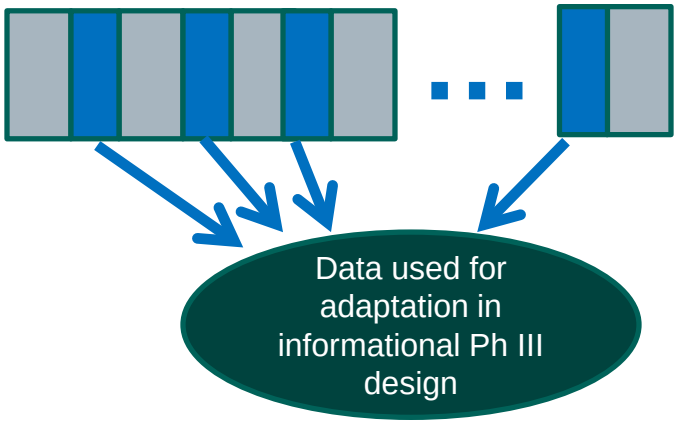
Interim analysis is conducted mid-trial in all enrolled patients

Informational analysis is conducted at end of the trial in a subgroup of patients

Informational design vs adaptive design

- Achilles's heel of a conventional adaptive design
  - Change of patients' characteristics after adaptation
- Information design is a type of ideal adaptive design
  - Some of the methods developed for adaptive design can be readily applied to informational design





A similar concept

- Ideally a biomarker and cutpoint are available before Phase III to mitigate regulatory risk and avoid delay of approval, but it usually takes a long time to develop.
- Freidlin and Simon's adaptive signature design
  - Use a subset of patients in Phase III as training set to find a biomarker cutpoint
  - Split alpha between the biomarker positive population and all-comer population
  - Trial is positive if p-value <2% in all-comer population or <0.5% in biomarker positive population (excluding those in training set)
- We assume biomarker subpopulations are well-defined

Statistical issues of interest

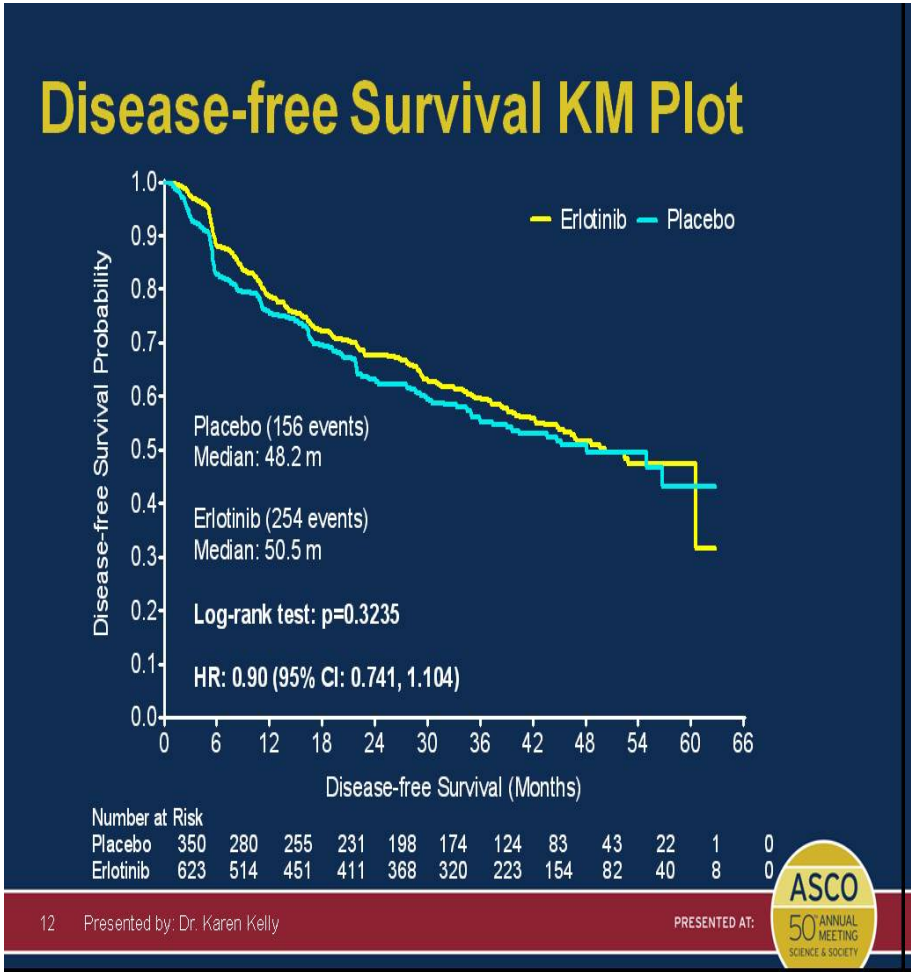
- How to test the co-primary hypotheses in overall population and a biomarker positive (BM+) subpopulation without any credible prior?
- Which biomarker subpop(s) to keep at final analysis?
  - Inclusion of non-performing subpopulations makes study design less efficient
  - A statistically significant outcome overall but clinically underwhelming outcome in biomarker subpopulation(s) present challenges to reimbursement

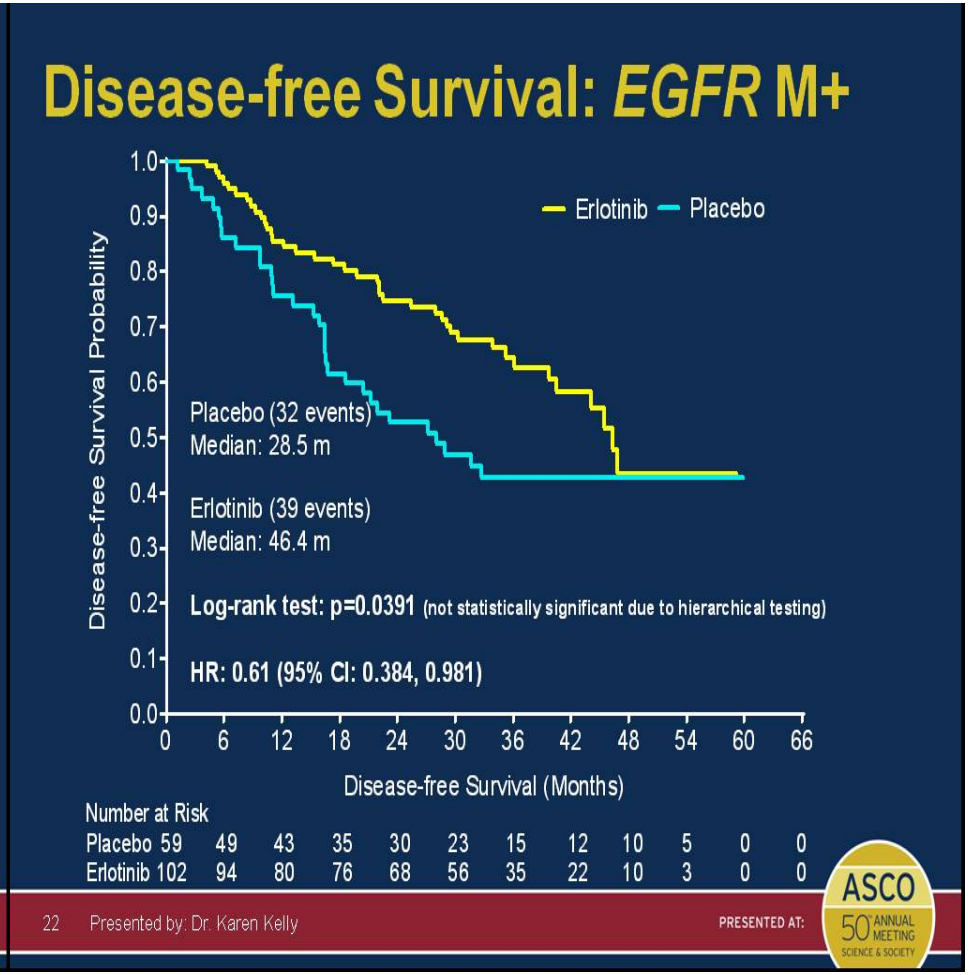
Adaptive alpha-allocation without use of unblinded trial-data

RADIANT – a motivating study

- Hypothesis testing
  - Erlotinib prolongs disease-free-survival (DFS) in completely resected patients with early stage (IB-IIIA) NSCLC whose tumor expressed EGFR by IHC or FISH
  - Step-down from all-randomized patients to a subpopulation with del19/L858R (EGFR M+)
- Sample size and timeline
  - 973 patients and 382 DFS events (~80% power for 0.75 hazard ratio)
  - Enrollment (9/2006 – 7/2010) and data cut-off (4/2013)

A missed opportunity





Adaptive alpha-allocation strategies

- Alpha-allocation as a function of blinded event ratio of a biomarker positive subpopulation in overall population
  - No penalty for multiplicity control

alpha-allocation function
=
alpha-spending function

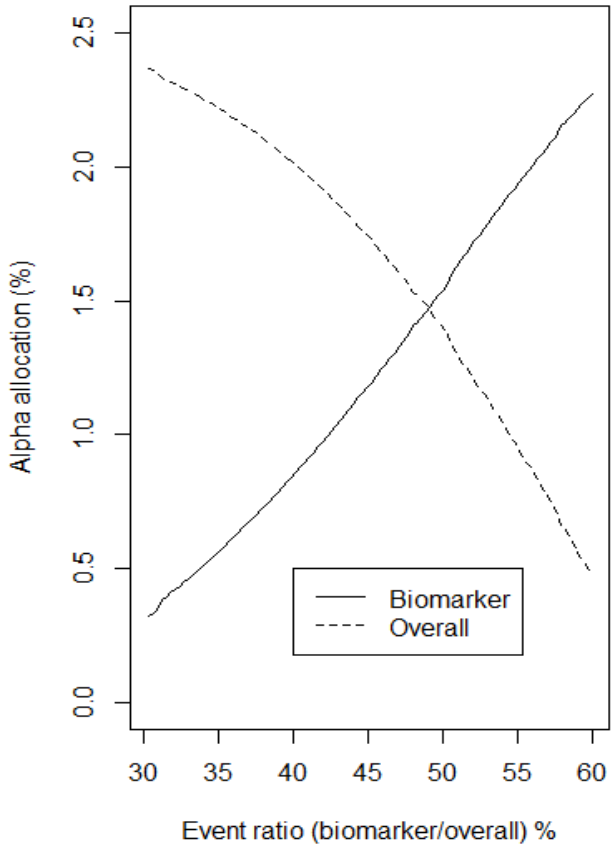
- Alpha-allocation not only as a function of blinded event ratio but also as a function of interim outcome
  - Pay penalty for multiplicity control

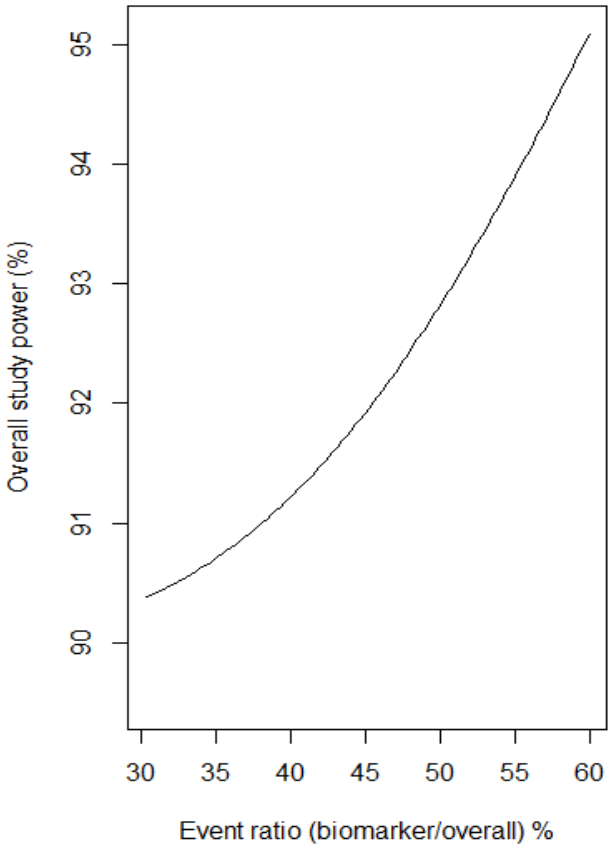
The penalty is also event ratio driven

Incorporate blinded event ratio only

- A trial is sized to have 90% power to detect a 0.7 hazard ratio at 2.5% (T: ~330 events)
- How to allocate alpha when target hazard ratio is 0.7 in overall population and 0.6 in BM+ population?
  - Overall alpha is controlled at 2.5% if alpha is controlled at 2.5% under each event ratio (notice $\int f(A|B)\partial B \leq \max\{f(A|B)\}$)
- A conservative but **optimized** Bonferroni approach
  - 0.5% to BM+, 2.0% to overall when event ratio (r)=40%
  - 1.4% to BM+, 1.1% to overall when event ratio (r)=50%
- Incorporate correlation into optimization
  - Customized alpha allocation function (Chen et al 2009)
  - Spiessens & Debois used an existing alpha-spending function as alpha-allocation function (2010)

Alpha-allocation function and power





What is the value of a 1% power increase?

- $1,000,000 if the drug has a net value of $1B in an indication over its life-time
  - A conservative assumption for a typical indication
- $1,000,000 savings in trial cost for a typical Phase III trial
  - Equivalent to the reduction of ~20 patients in a study with sample size of ~600
  - Average post per patient in an oncology trial is conservatively estimated to be $50,000
- An innovative statistical method is potentially worth millions of dollars to every study it is applied to!

Adaptive alpha-allocation with use of unblinded trial-data

Auto-adaptive alpha-allocation with trial data

- For each *t*, find the alpha-allocation that maximize the expected conditional power
  - Informational analysis provides an objective prior distribution of estimates for true treatment effects
  - Estimates of treatment effects based on external data can be further incorporated
- The adjusted alpha at *t*, $\alpha^*(t)$, is calculated to keep the actual Type I error controlled at α
  - The larger the *t* the smaller the $\alpha^*(t)$
- Is the α penalty worth it?
  - No if we have strong prior; Yes otherwise

Algorithm

- **Choose α_1 (overall study) and α_2 (subgroup) that maximize the expected conditional power**

- $$Q(\alpha_1, \alpha_2; \boldsymbol{x_{1,t}}, \boldsymbol{x_{2,t}}, \alpha_t) = \int \left\{ 1 - \Phi_{\sqrt{r}} \left(\frac{Z_{1-\alpha_1} - \sqrt{t}\boldsymbol{x_{1,t}}}{\sqrt{1-t}} - \sqrt{(1-t)I_3}\Delta_1, \right. \right.$$
$$\left. \left. \frac{Z_{1-\alpha_2} - \sqrt{t}\boldsymbol{x_{2,t}}}{\sqrt{1-t}} - \sqrt{(1-t)rI_3}\Delta_2 \right) \right\} g(\Delta_1, \Delta_2 | \boldsymbol{x_{1,t}}, \boldsymbol{x_{2,t}}) d\Delta_1 \Delta_2$$

  subject to the constraint by **nominal** type I error of :

$$1 - \Phi_{\sqrt{r}}(Z_{1-\alpha_1}, Z_{1-\alpha_2}) = \alpha_t, t \in [0,1]$$

- **Find α_t to keep overall alpha under control**
  - Denote $(\tilde{\alpha}_{1,t}, \tilde{\alpha}_{2,t}) = \arg\max Q(\alpha_1, \alpha_2 ; x_{1t}, x_{2t}, \alpha_t)$. The actual type I error under the global null hypothesis is:

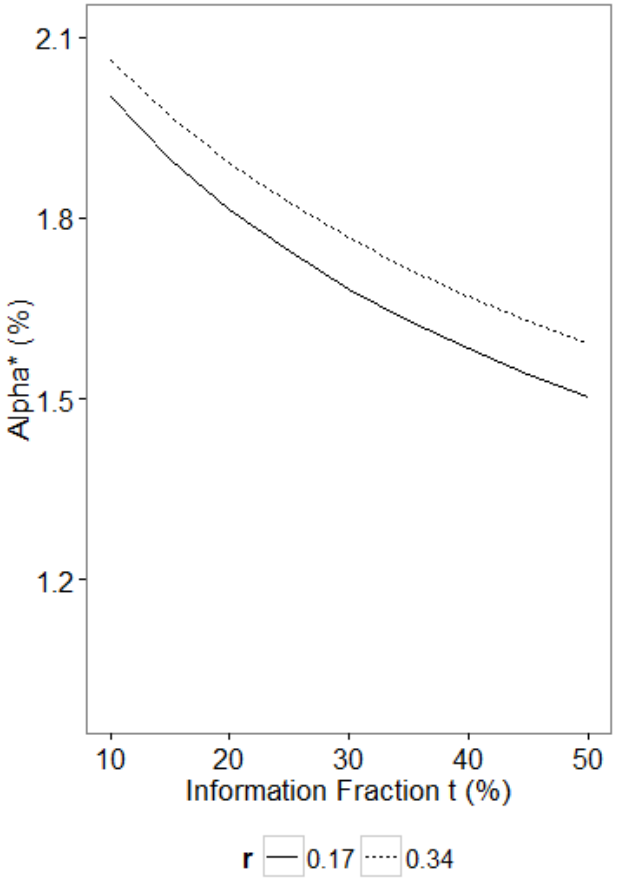
$$P(\alpha_t) = \int \left[1 - \Phi_{\sqrt{r}} \left(\frac{Z_{1-\tilde{\alpha}_{1,t}} - \sqrt{t}x_{1t}}{\sqrt{1-t}}, \frac{Z_{1-\tilde{\alpha}_{2,t}} - \sqrt{t}x_{2t}}{\sqrt{1-t}} \right) \right] \phi_{\sqrt{r}}(x_{1t}, x_{2t})\, dx_{1t} x_{2t}.$$

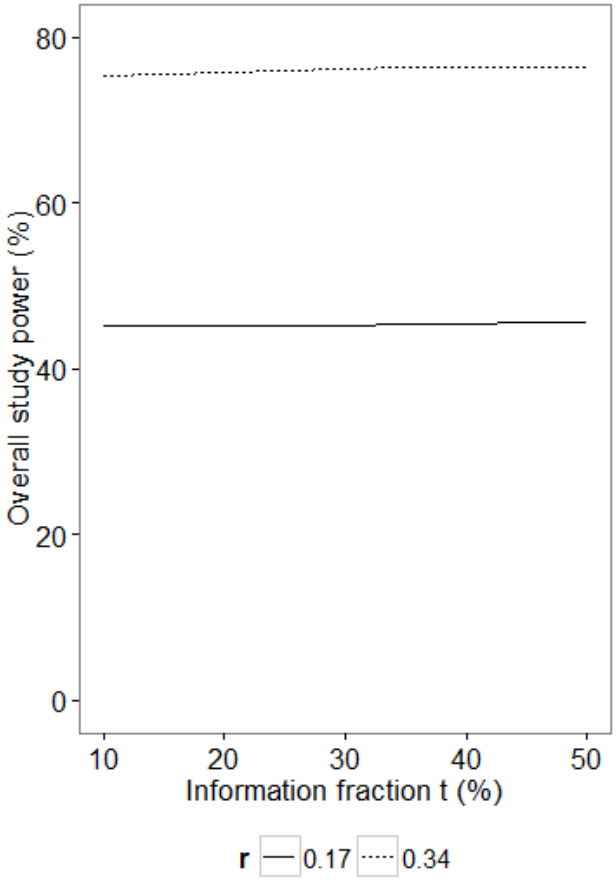
  - Iterative root finding for the equation $P(\alpha_t) = \alpha$.

Application to a RADIANT like study

- 1:1 randomization with a total 410 events
  - 83% power for detecting a 0.75 hazard ratio at 2.5% in overall population
  - The true (UNKNOWN) hazard ratio is 0.90 in overall population and 0.61 in the biomarker positive population
  - 17% or 34% of the events are assumed in the subpopulation
- Power comparison
  - The study has only 19% power if step-down from overall population (aka RADIANT approach)
  - Should the biomarker subpopulation be tested first, the study would have 54% power at r=17% and 83% power at r=34%
  - The informational design would have ~45% power at r=17% and ~75% power at r=34%
  - A little bit of information adds tremendous value. However, benefit of more information is offset by penalty on alpha.

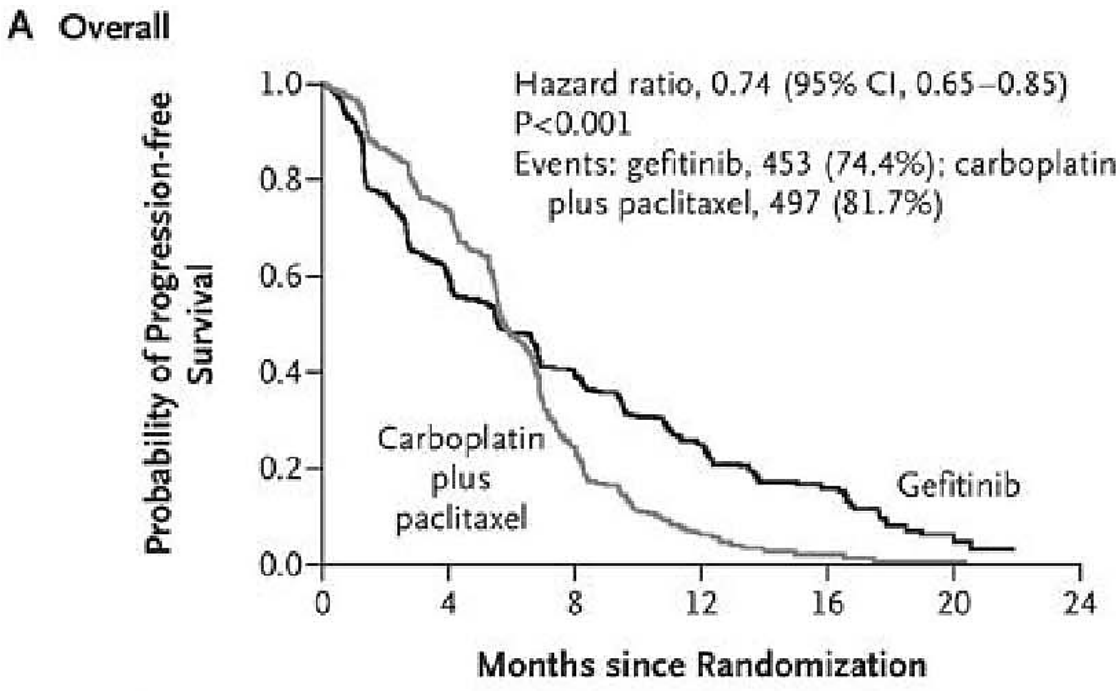
α* and power in a RADIANT like study





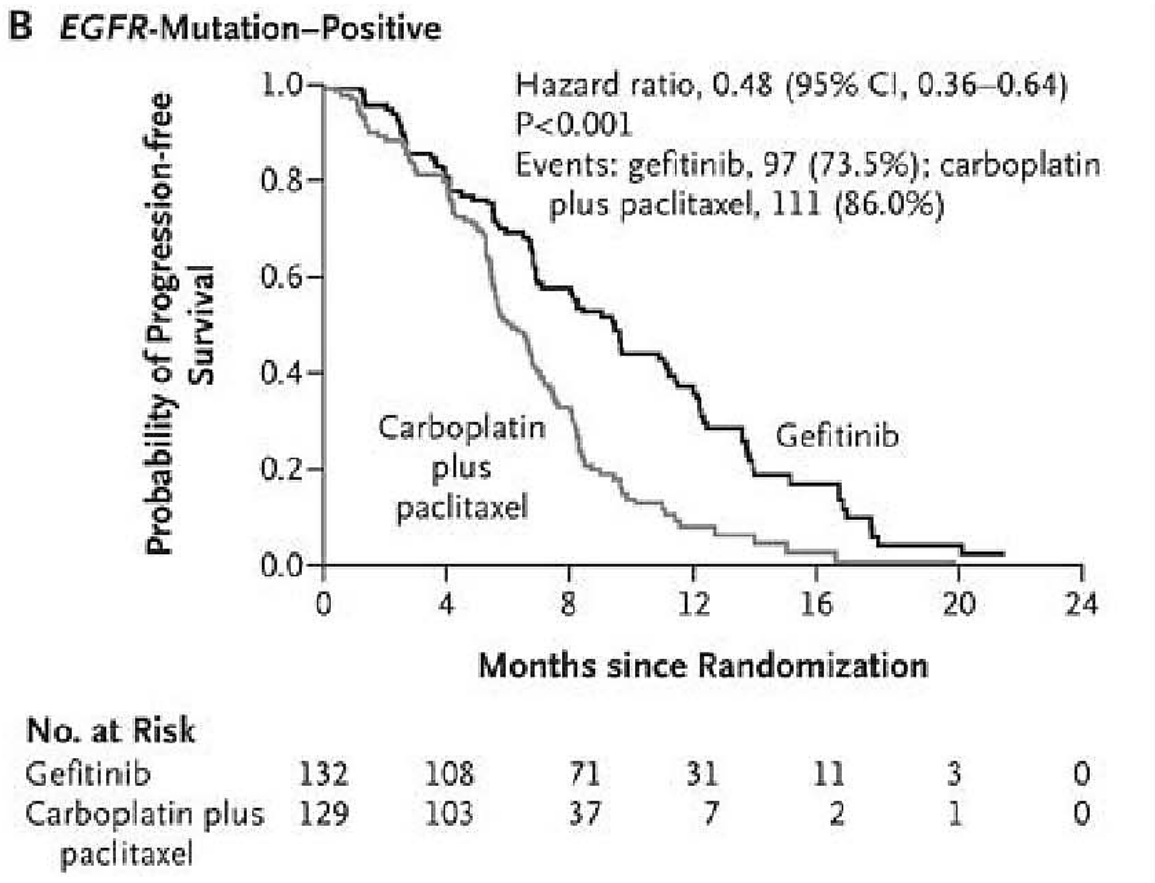
Adaptive population selection under same endpoint

IPASS – overall population

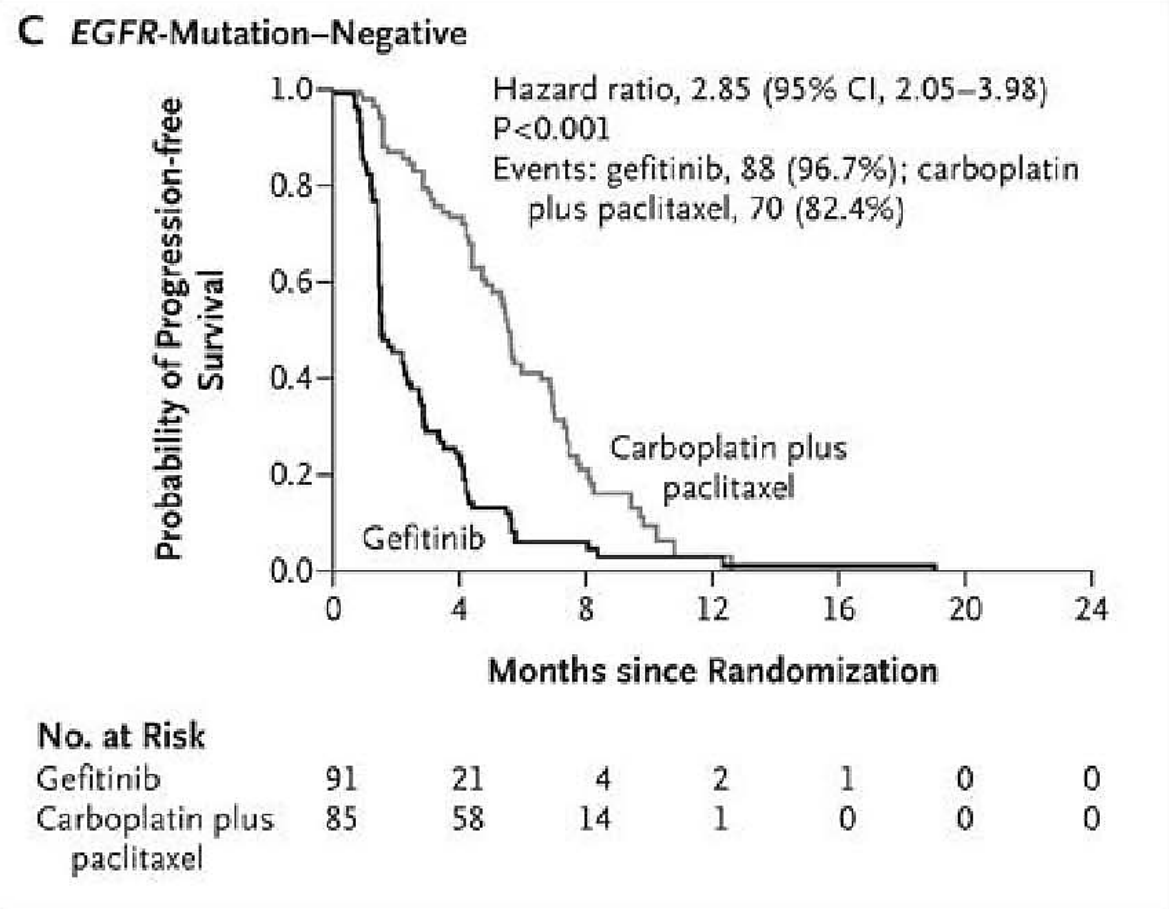


No. at Risk							
Gefitinib	609	363	212	76	24	5	0
Carboplatin plus paclitaxel	608	412	118	22	3	1	0

EGFR – mutation positive



EGFR – mutation negative



Set-up

- Suppose the overall population consists of k disjoint biomarker subpopulations and treatment effect increases with biomarker level
- A decision is made based on information fraction *t* to exclude subpopulations without a numerically positive treatment effect in a step-up process that starts from lowest biomarker level (least efficacious)
- Which Type I error rate (alpha*) should the hypothesis be tested in remaining patients?

Solving for adjusted alpha (α^*)

- Let Y_{i1} be the test statistics based on information fraction t
  - The m-th subpopulation will not be included in final analysis if p-value based on Y_{i1} is > α_t for all $i \leq m$
- Suppose that m cohorts are excluded in the final analysis ($k > m \geq 0$), and let Z_{-m} be the corresponding test statistics. The probability of a positive outcome in pooled analysis is

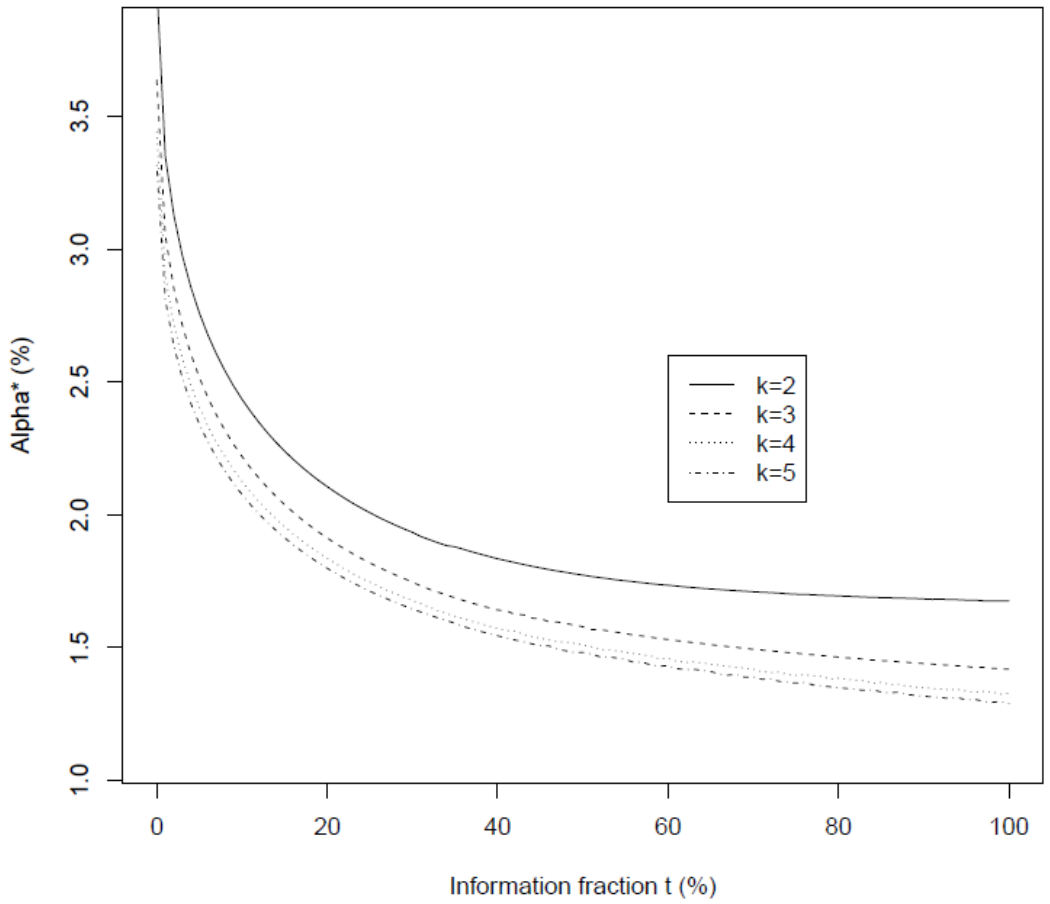
$$R(\alpha^* | \alpha_t, m) = \text{Prob}(Y_{i1} < Z_{1-\alpha_t} \text{ for } i=1,\ldots,m,\ Y_{m+1,1} > Z_{1-\alpha_t},\ Z_{-m} > Z_{1-\alpha^*})$$

- α^* is solved from below

$$\sum_{m=0}^{k-1} R(\alpha^* | \alpha_t, m) = \alpha$$

α* under different k

- Equal prevalence of events by biomarker level
- α_t=0.5 (binding)



A hypothetical example

- Consider a hypothetical study with 3 ordered biomarker subpopulations (i.e., low, intermediate, high)
- The study targets 410 events so that the study has 83% power for detecting a 0.75 hazard ratio at 2.5% (one-sided) in the overall population
- The study may drop low, low + intermediate, OR drop all ("early" termination) if empirical effect is negative
- Log-hazard ratios are $\log(0.75)+\delta$, $\log(0.75)$, $\log(0.75)-\delta$
  - When δ ranges from 0.2 to 0.4, hazard ratio ranges from 0.92 to 1.12 for the "low" group and from 0.50 to 0.61 for the "high" group

Operational characteristics

δ	t	α*	Prob (keep all)	Prob (drop low)	Prob (drop low/ intermediate)	Prob (drop all)	Overall study power
0.2	40%	0.0164	0.63	0.24	0.11	0.02	0.75
0.2	60%	0.0153	0.65	0.25	0.09	0.01	**0.84**
0.3	40%	0.0164	0.48	0.31	0.18	0.03	0.76
0.3	60%	0.0153	0.48	0.35	0.16	0.01	**0.87**
0.4	40%	0.0164	0.34	0.36	0.27	0.03	0.79
0.4	60%	0.0153	0.31	0.42	0.26	0.01	**0.91**

The overall study has 83% power w/o population de-selection. De-selection criterion or timing is not optimized.

Adaptive population selection under different endpoints

A hypothetical trial

- 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
  - Treatment effect is assumed to be ordered by BM level from low to high with equal prevalence
- 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 and sample size for BM high will increase as needed
- 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^*)

- Interim analysis is done when half of the survival information is available (t=50%) and accrual is about to complete
  - 250 PFS events at interim (vs 165 OS events)
  - BM low and BM high have same number of events
- Overall type I error under the null hypothesis of no treatment effect on OS without any constraint on PFS effect (δ_1, δ_2)

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

- Minimal α^* of entire (δ_1, δ_2) space that keeps above overall type I error at 0.025 is the nominal alpha for final analysis
  - Not needed when OS is used for biomarker selection

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

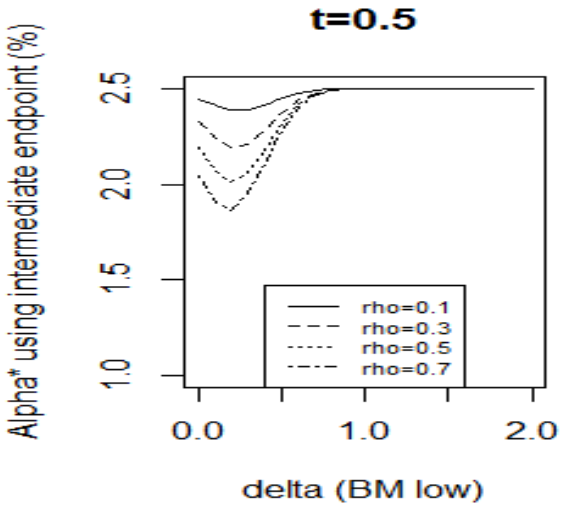
- Type I error under the null hypothesis of no OS effect ($\Delta_1 = \Delta_2 = 0$) **without** constraint on PFS effects (δ_1, δ_2)

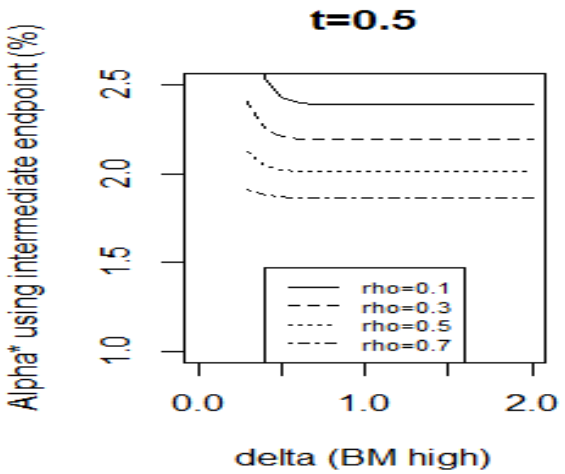
$$P\left(X_{l1} > Z_{1-\alpha_t}, V_{all} > Z_{1-\alpha^*} | \delta_1, \Delta_1 = \Delta_2 = 0\right) +$$
$$P(X_{l1} < Z_{1-\alpha_t}, X_{h1} > Z_{1-\alpha_t}, V_{h2} > Z_{1-\alpha^*} | \delta_1, \delta_2, \Delta_2 = 0)$$

- Minimal α^* of entire (δ_1, δ_2) space that keeps above overall type I error at 0.025 is the nominal alpha
  - $\delta_1 = \delta_2 = 0$ when OS is used for both analyses and in this case α^* can be greater than 2.5% due to binding futility stopping

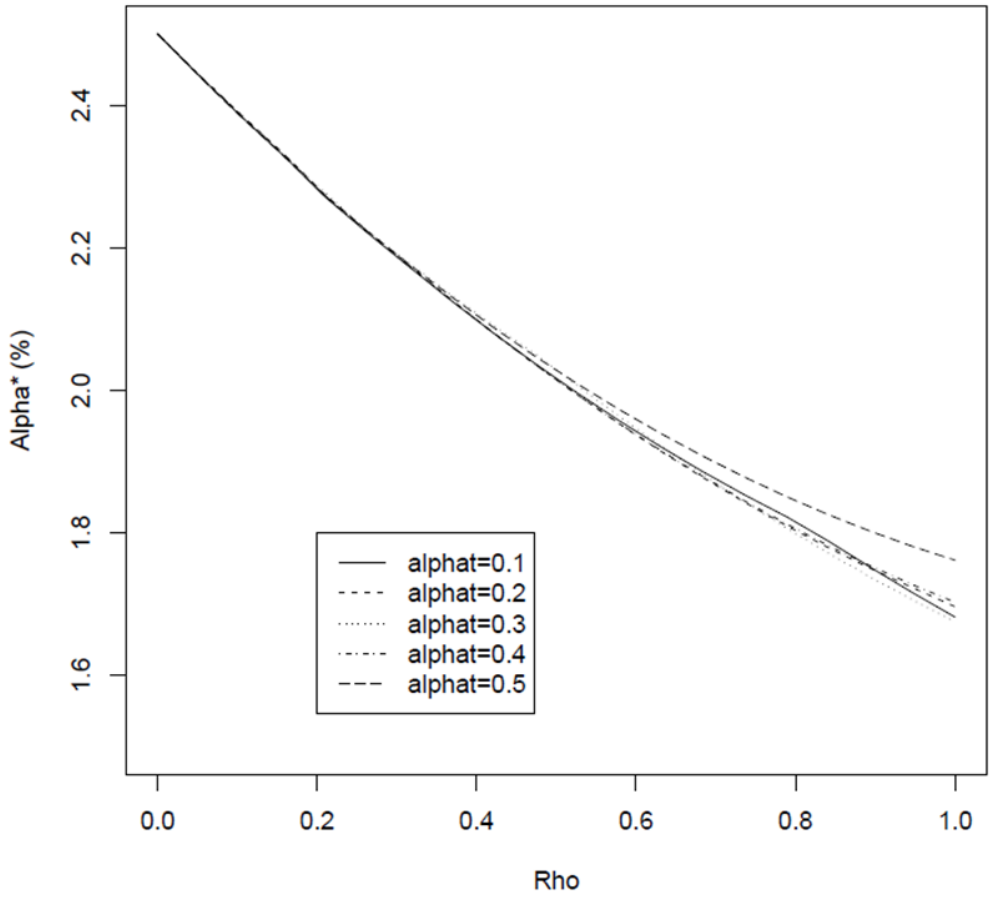
α^* by correlation between PFS and OS

- Each α^* is determined by correlation between PFS and OS which can be estimated from the trial data once study is over, and estimate of α^* is consistent as long as the correlation estimate is consistent
- Minimum α^* is reached at non-degenerate/non-trivial (δ_1, δ_2) due to the complicated interplay between cherry picking and futility stopping





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



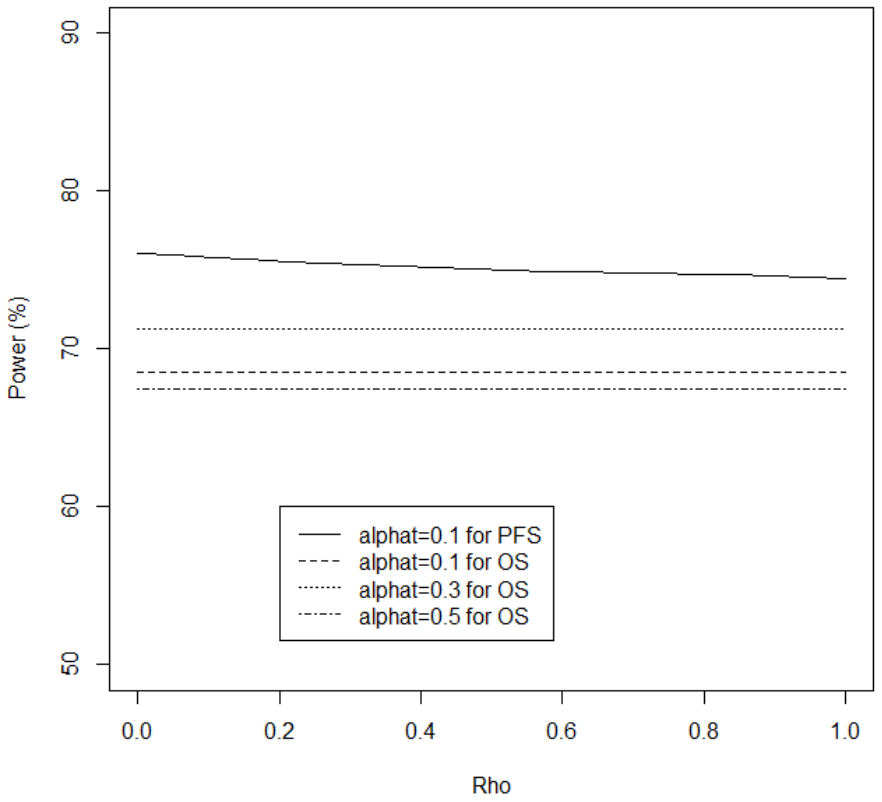
Minimal α* is robust to α_t in this hypothetical example

Set-up for power comparison

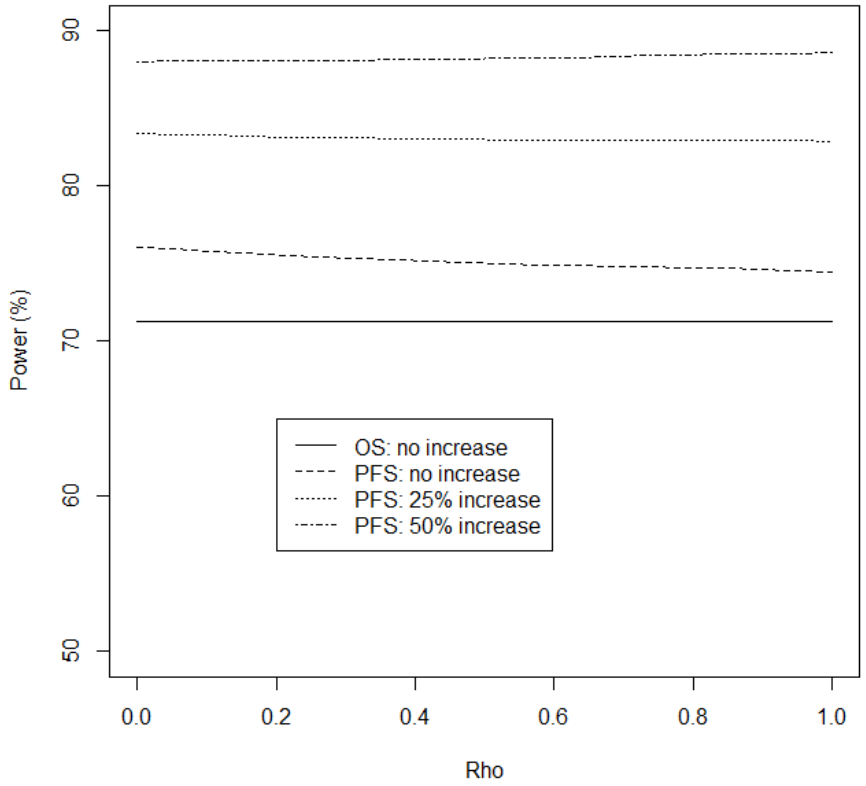
- True HR for OS is 0.6 in BM high and is 1 in BM low
  - The actual power without biomarker selection is 64%
- True HR for PFS is 0.45 in BM high and is 1 in BM low
  - Sensitivity of PFS for immunotherapies depends on tumor type and line of therapy, an may differ by biomarker level
  - It is unclear whether RR or PFS is a more sensitive intermediate endpoint, and how (not whether) RECIST should be modified to better predict clinical benefit

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 despite greater α*
- Power is robust to rho (more useful info ⇔ higher rho ⇔ higher penalty)



Impact of sample size increase on study power



Increase of sample size reduces the correction between PFS at interim and OS at final, and hence penalty

Discussion

- Uncertainty about biomarker effect and prevalence calls for data-driven and objective designs
- Uncertainty about treatment effect over time provides challenges to conventional adaptive designs
- Informational design provides a salvage plan, and is not meant to replace but to supplement conventional designs
  - However, it is the only option if the data on biomarkers are not available until the end of study

Key references

- Chen C, Li N, Shentu Y, Pang L, Beckman RA. Informational Design of confirmatory Phase III Trials for Expedited Development of Personalized Medicines. 2015, unpublished manuscript
- Chen C, Beckman RA. Hypothesis Testing in a Confirmatory Phase III Trial With a Possible Subset Effect. Statistics in Biopharmaceutical Research 2009; 1(4): 431–440.
- Spiessens B, Debois M. Adjusted significance levels for subgroup analysis in clinical trials. Contemporary Clinical Trials 2010; 31:647–656.
- Shentu Y, Chen C, Pang L, Beckman RA. Auto-adaptive Alpha Allocation: a strategy to mitigate risk on study assumptions. 2015, unpublished manuscript
- Freidlin B, Simon R. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. Clinical Cancer Research 2005; 11 (21): 7872-7878.

ORR of pembrolizumab in melanoma trials

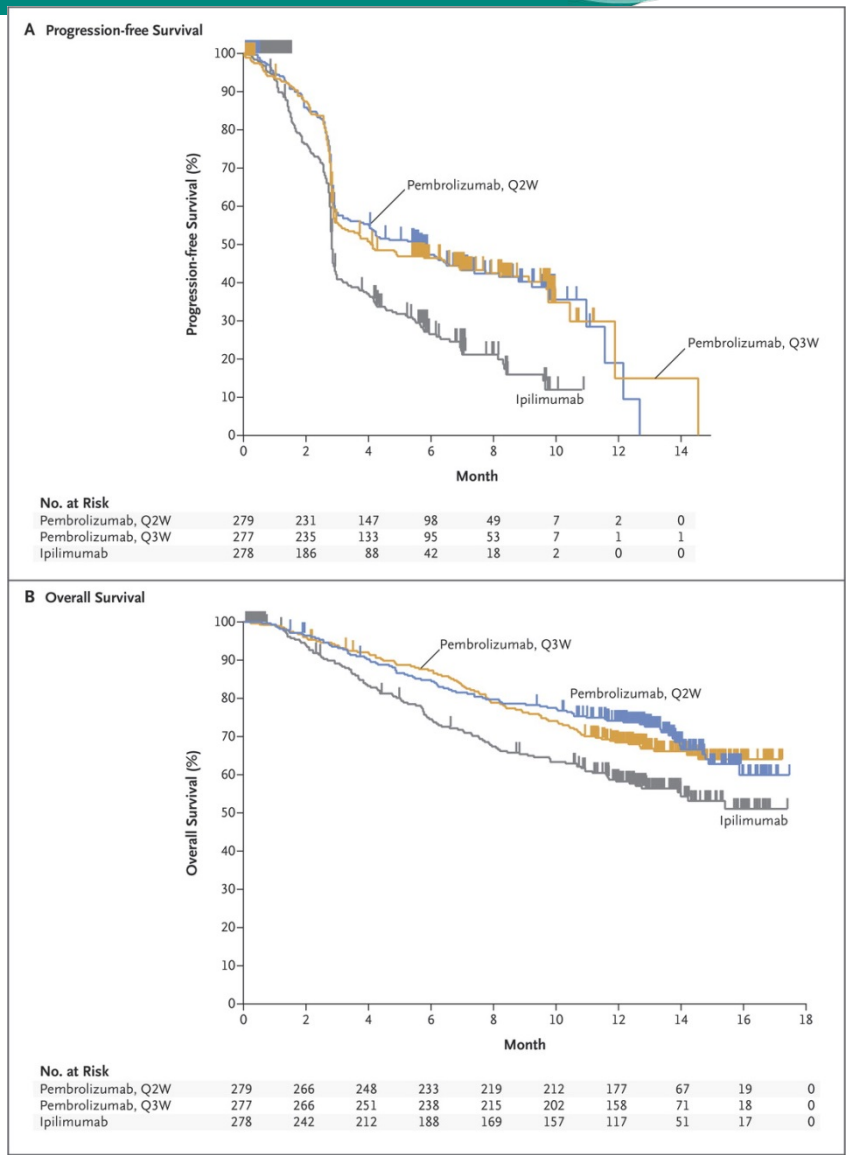
Dose	Prior IPI	Phase IB Exploratory		Phase III Confirmatory	
		N	ORR, % (95% CI)	N	ORR, % (95% CI)
10 mg/kg Q2W	Naive	39	49 (32–65)	279	34 (38–40)
	Treated	13	62 (32–86)		
10 mg/kg Q3W	Naive	19	26 (9–51)	277	33 (27–39)
	Treated	26	27 (12–48)		

- A dose response seen in Phase 1B disappears in Phase III
- Patients are never i.i.d in oncology trials, especially in the field of highly competitive immunotherapies

N Engl J Med 2013, 169: 134-144; N Engl J Med 2015; 372:2521-2532

Immuno-oncology therapies (pembrolizumab vs ipilimumab) in advanced melanoma

- Flatter tails than normally seen
- Much longer survival than before when the median was 8-10 months
- Many patients may live for >5 years (“cured”)

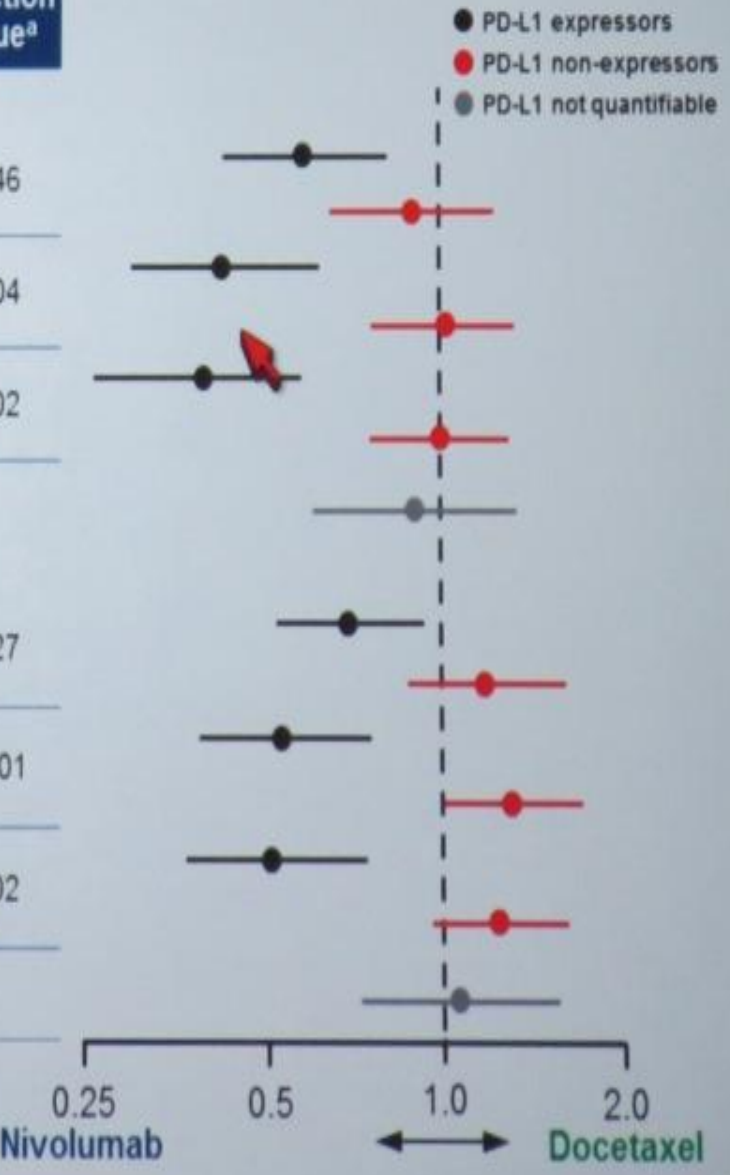



CANCER
IT'S PERSONAL
THE RIGHT PATIENT. THE RIGHT TREATMENT.

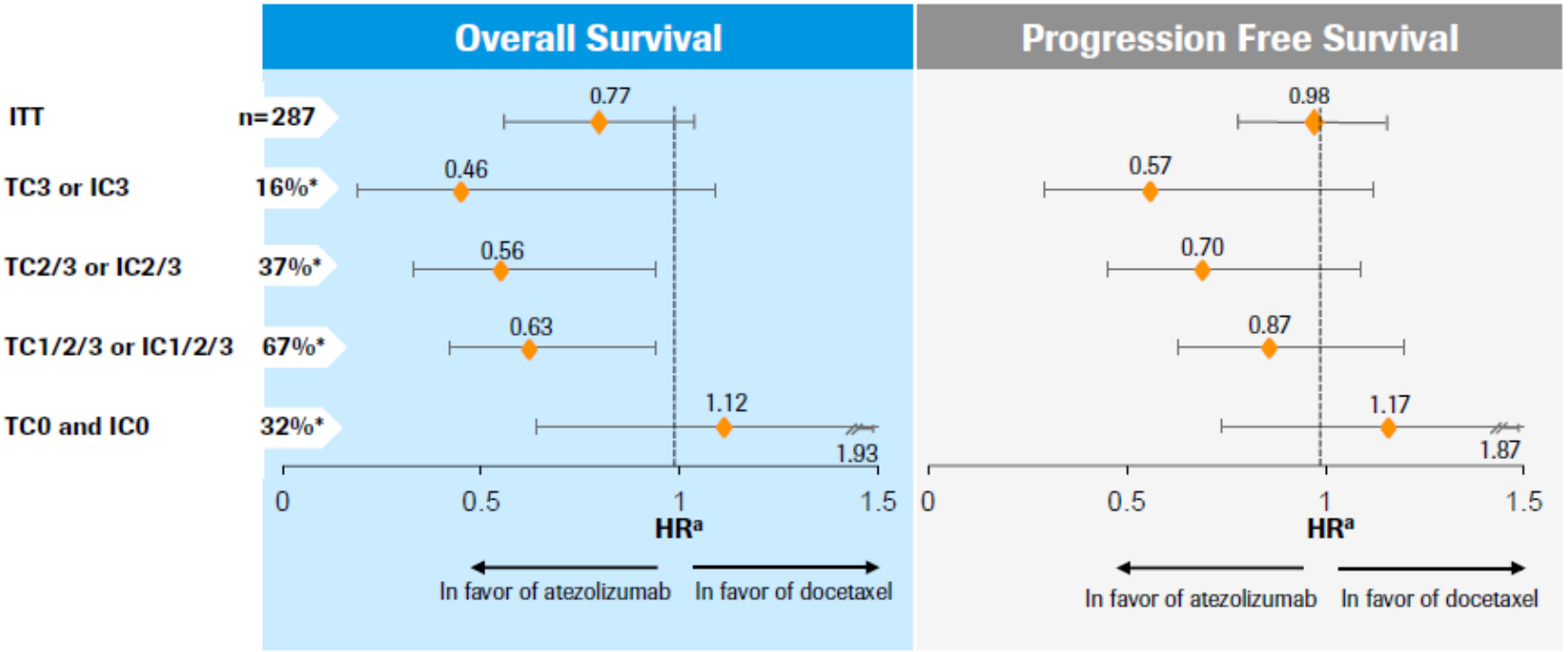
OS and PFS Hazard Ratios by Baseline PD-L1 Expression

PD-L1expression level	Nivolumab n	Docetaxel n	Unstratified HR (95% CI)	Interaction *P*-value[a]
OS				
≥1%	123	123	0.59 (0.43, 0.82)	0.0646
<1%	108	101	0.90 (0.66, 1.24)	
≥5%	95	86	0.43 (0.30, 0.63)	0.0004
<5%	136	138	1.01 (0.77, 1.34)	
≥10%	86	79	0.40 (0.26, 0.59)	0.0002
<10%	145	145	1.00 (0.76, 1.31)	
Not quantifiable at baseline	61	66	0.91 (0.61, 1.35)	
PFS				
≥1%	123	123	0.70 (0.53, 0.94)	0.0227
<1%	108	101	1.19 (0.88, 1.61)	
≥5%	95	86	0.54 (0.39, 0.76)	<0.0001
<5%	136	138	1.31 (1.01, 1.71)	
≥10%	86	79	0.52 (0.37, 0.75)	0.0002
<10%	145	145	1.24 (0.96, 1.61)	
Not quantifiable at baseline	61	66	1.06 (0.73, 1.56)	

[a] Interaction p-value from Cox proportional hazard model with treatment, PD-L1 expression and treatment by PD-L1 expression interaction



POPLAR


Overall Survival
Progression Free Survival
ITT
n=287
TC3 or IC3
16%*
TC2/3 or IC2/3
37%*
TC1/2/3 or IC1/2/3
67%*
TC0 and IC0
32%*
0.77
0.46
0.56
0.63
1.12
1.93
0.98
0.57
0.70
0.87
1.17
1.87
0
0.5
1
1.5
HR[a]
In favor of atezolizumab
In favor of docetaxel