STATISTICAL DESIGNS AND STRATEGIES FOR ONCOLOGY DRUG DEVELOPMENT

Cong Chen

Executive Director and Head of Early Oncology Statistics, BARDS
Merck & Co., Inc., Kenilworth, NJ, USA

2020 ASA Biopharmaceutical Section Regulatory-Industry Workshop

Outline

- Phase IB Efficacy Screening
- 2-in-1 Adaptive Design and Extensions
- Phase 3 Programs with Biomarker Considerations
- Prediction of Treatment effect of Combination Therapies

PHASE IB EFFICACY SCREENING

Efficacy Screening Post Dose-finding

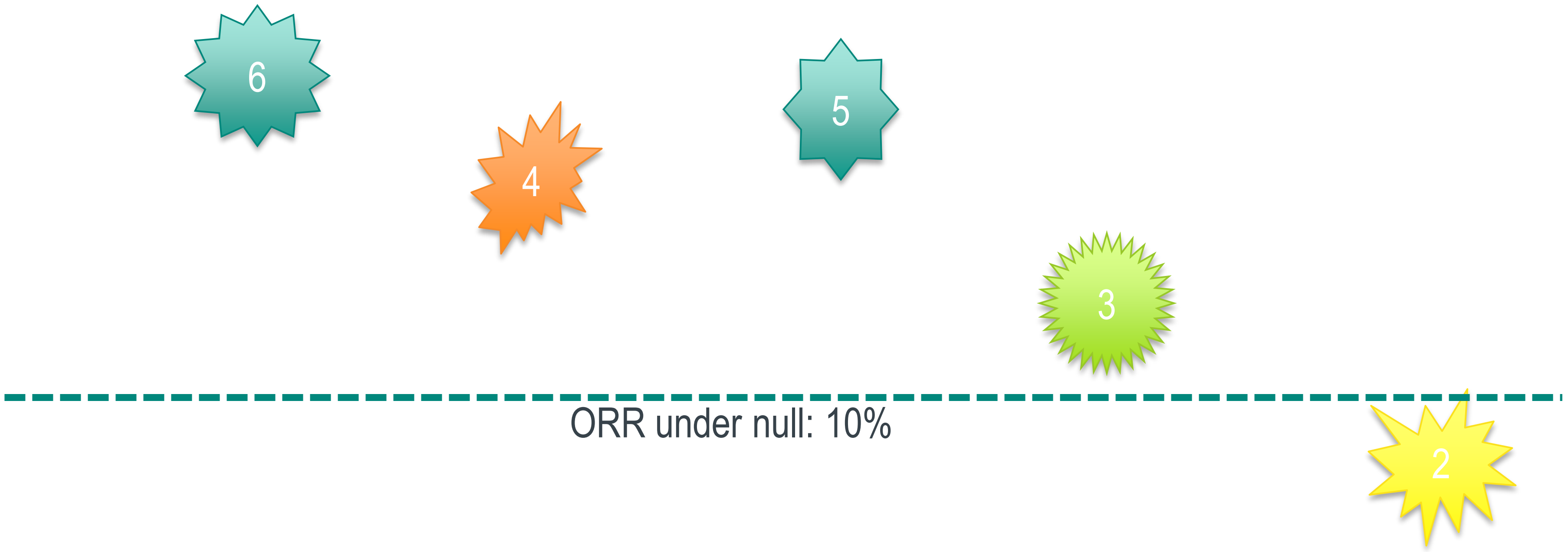
- How to test whether a new drug is active and worth further investigation most efficiently?
- A set of tumor types are often investigated simultaneously in a basket trial to account for *Type III error* of missed opportunities
  - **FDA definition**: patients defined by disease stage, histology, number of prior therapies, genetic or other biomarkers, or demographic characteristics

3-5 shots on goal

Chen C, Deng Q, He L, Mehrotra D, Rubin EH, Beckman RA. How many tumor indications should be initially studied in clinical development of next generation immunotherapies? *Contemporary Clinical Trials* 2017; 59:113-117.

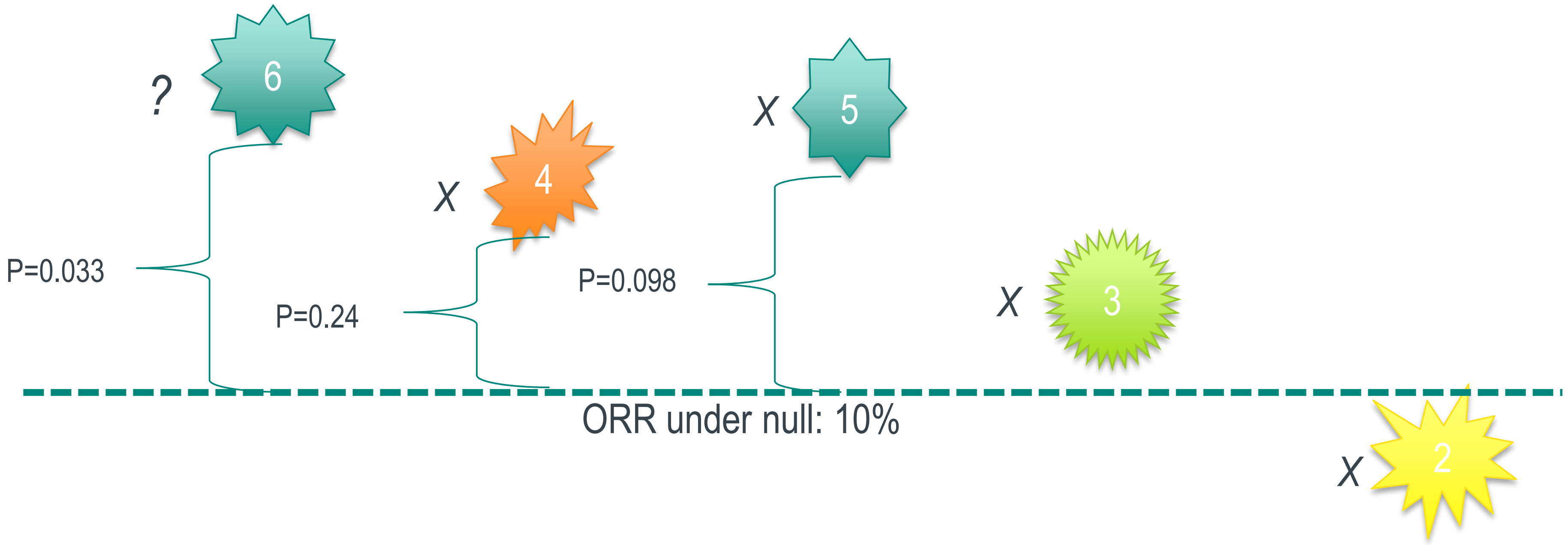
Hypothetical Outcome of a Simple Basket Trial

- Five tumor cohorts (n=25 each) in patients refractory to PD-1 treatment (null ORR: 10%)
- Number of responses range from 2 (8%) to 6 (24%)



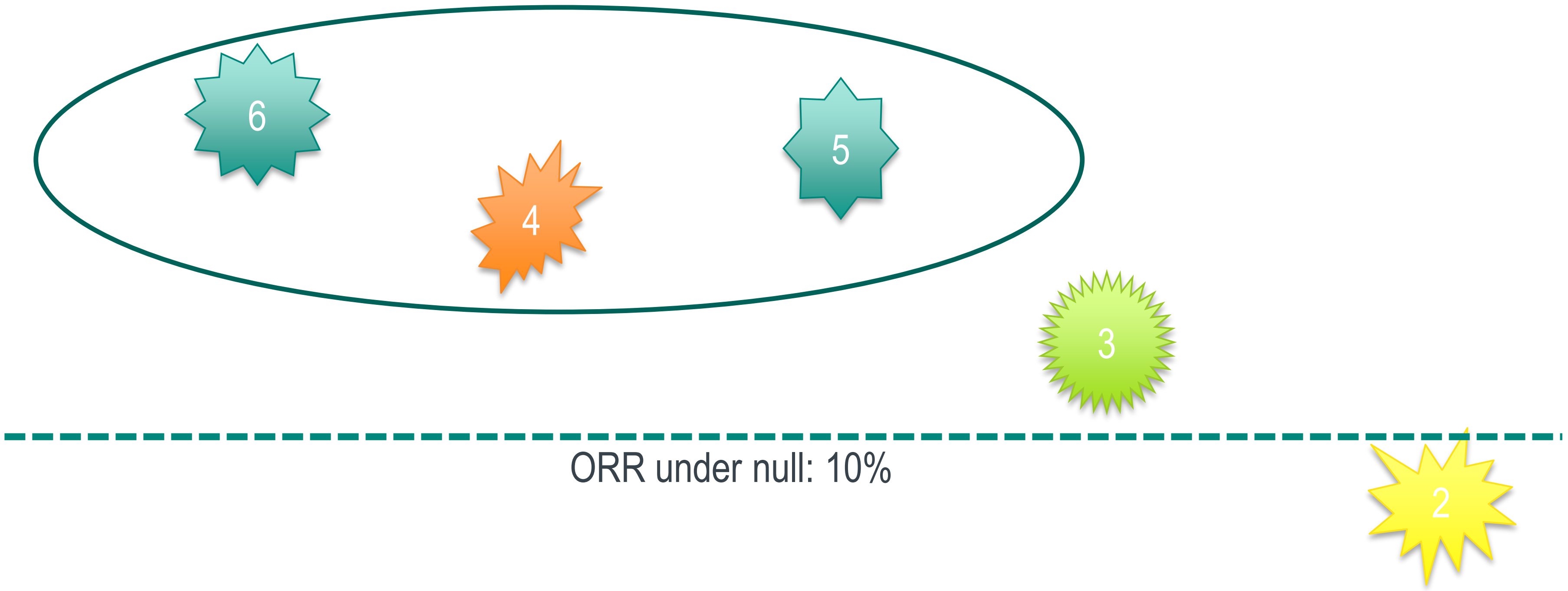
Independent Evaluation

- Each tumor cohort is evaluated separately, with or without multiplicity adjustment



Ad-hoc Assessment

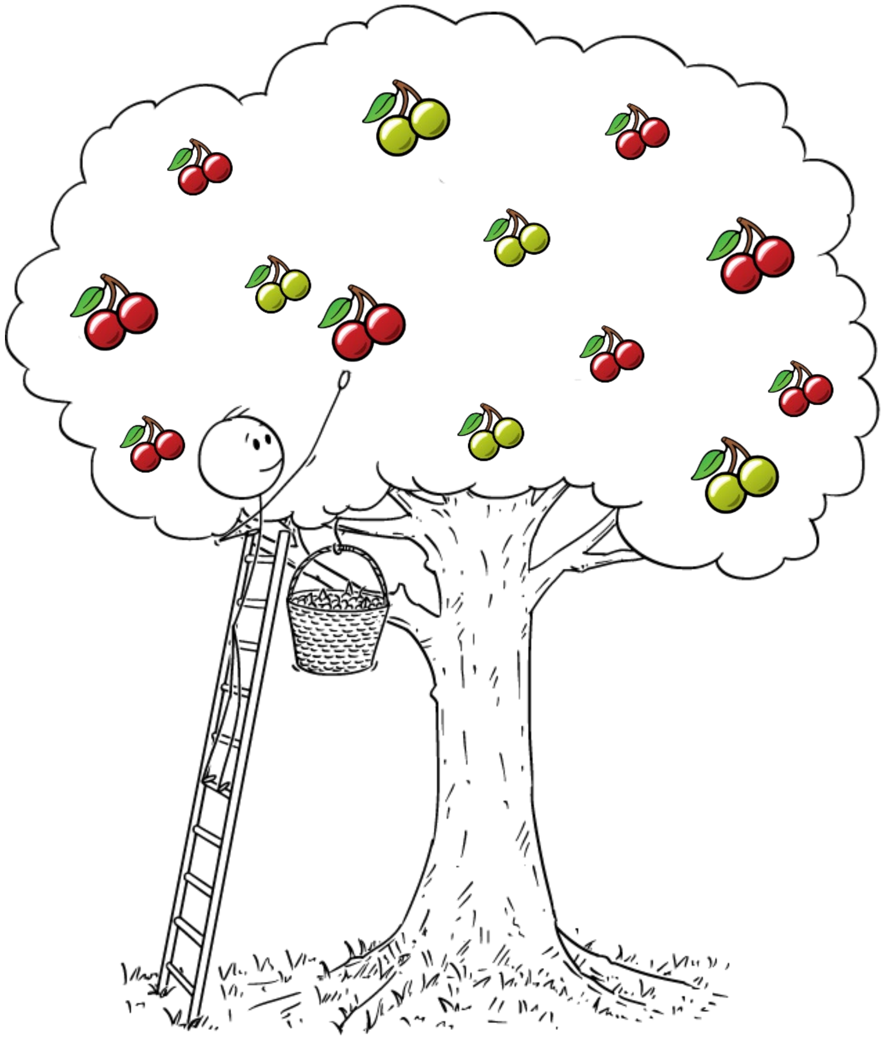
- *Clinical director 1:* Look at the 3 top ones! The drug is working!!
- *Clinical director 2:* This is cherry-picking.



Bayesian Information Borrowing

- Assumes some form of homogeneity on response rates across tumor cohorts
  - Thall et al. 2003, Berry et al. 2013, Simon et al., 2016, Cunanan et al., 2017
- *Clinical director 1*: I like Bayesian, but why does response to an active drug have to be homogeneous?
- *Clinical director 2*: It is too complicated for me. Can’t you just tell me how to cherry-pick properly?

Multiplicity Control for Cherry-picking



Chen C, Li N, Yuan S, Antonijevic Z, Kalamegham R, Beckman RA. Statistical design and considerations of a Phase 3 basket trial for simultaneous investigation of multiple tumor types in one study. *Statistics in Biopharmaceutical Research* 2016; 8 (3): 248-257.

Zhou H, Liu F, Wu C, Rubin EH, Giranda VL, Chen C. Optimal Two-stage Designs for Exploratory Basket Trials, Contemporary Clinical Trials 2019. DOI: 10.1016/j.cct.2019.06.021.

Wu C, Liu F, Zhou H, Rubin EH, Giranda VL, Chen C. Optimal Design and Analysis of Efficacy Expansion in Phase I Oncology Trials. Under review.

Chen C, Zhou H, Li W, Beckman RA. How Many Substudies Should be Included in a Master Protocol? Under review.

OPTIMAL BASKET TRIAL DESIGN WITH PRUNING AND POOLING

MERCK

Basket Designs with Cherry-picking

- Prune inactive ones and pool active ones in the pooled analysis (***pruning-and-pooling***)
  - Type I error is controlled at target level under global null
- Type II error is calculated under a non-informative prior for number of active tumors (i.e., uniform distribution)
  - Design parameters can be obtained similarly when an informative prior is available
- While sample size calculation is guided by the design parameters, interpretation of trial outcome may be based on totality of data to improve the quality of decision

Fit-for-purpose

One or two-stage?

Fix power or sample size?

Same or different hypotheses?

A One-stage Design Example with Same Hypotheses

- Design of a 5-tumor basket trial with minimal sample size targeting (α, β)=(0.05, 0.20)

P0	P1	r	α*	n
0.10	0.25	4	0.009	25

- Sample size in the hypothetical trial is optimal
  - The clinical intuition of pooling tumors with ≥4 responses would make sense
  - The pooled data should be tested at α*=0.009 to control α=0.05

- Scenarios of positive outcomes

Tumors	Sample size	Min #resp	Min ORR
1	25	8	32%
2	50	12	24%
3	75	15	20%
4	100	19	19%
5	125	22	18%

- With 4/5/6 responses in 3 tumor cohorts in the hypothetical trial, drug would be deemed active
  - May need more patients to further confirm the individual signals

A One-stage Design Example with Heterogenous Hypotheses

- Set-up for (H0, H1)
  - Mono in 3 tumor cohorts without SOC: (0.05, 0.2)
  - Combo with SOC in 2 tumor cohorts: (0.2, 0.35)
- Design features
  - Each has comparable probability to be pooled
  - Minimum overall sample size to achieve the desired Type I/II error rates
- Overall ORR for pooled tumor indications is compared to the weighted H0 by sample size

Design of the Hypothetical Trial

- Design parameters at (α, β)=(0.05, 0.20)
  - Total sample size=3*18+2*34=122

Cohorts	P0	P1	r	n	α*
3 (mono)	0.05	0.2	2	18	0.011
2 (combo)	0.2	0.35	9	34	

- Probability of pooling
  - (23%, 23%) under P0 for (mono, combo)
  - (90%, 89%) under P1 for (mono, combo)

- Examples of positive outcomes when one mono and one combo are left in pool (n=52=18+34)

Mono resp#	combo resp#	Overall	Wgted ORR (H0)	P-value
2 (11%)	13 (38%)	15(29%)	14.8%	0.0069 (<0.011)
4 (22%)	11 (32%)			
6 (33%)	9 (26%)			

Two-stage Optimal Basket Designs

- Design parameters of a two-stage 5-tumor basket trial with **minimal** sample size for same (P0, P1)=(0.1, 0.25) targeting (α, β)=(0.05, 0.20)
  - **A natural extension of Simon's designs for single arm trials to multi-arm basket trials**
  - N=43/40 when each applies a Simon's two-stage design independently at α=0.05, or much larger after multiplicity adjustment (α=0.01)

	r1	n1	α*	n
Optimal	2	9	0.019	33
Minimax	3	18	0.009	25
Tumor cohorts with ≥r1/n1 responders will be pooled for analysis at end of second stage and tested at α*				

Two-stage Design Under Fixed Sample Size

- Remaining sample size for terminated tumor cohorts is evenly distributed to continuing ones
- Design parameters of a two-stage 5-tumor basket trial with minimal sample size for same (P0, P1)=(0.1, 0.25) targeting (α, β)=(0.05, 0.20)
  - Planned sample size per arm (n=20) is smaller than under the optimal design (n=33)
  - A remaining arm may have more patients (e.g., n=35 if 3 arms are terminated earlier)

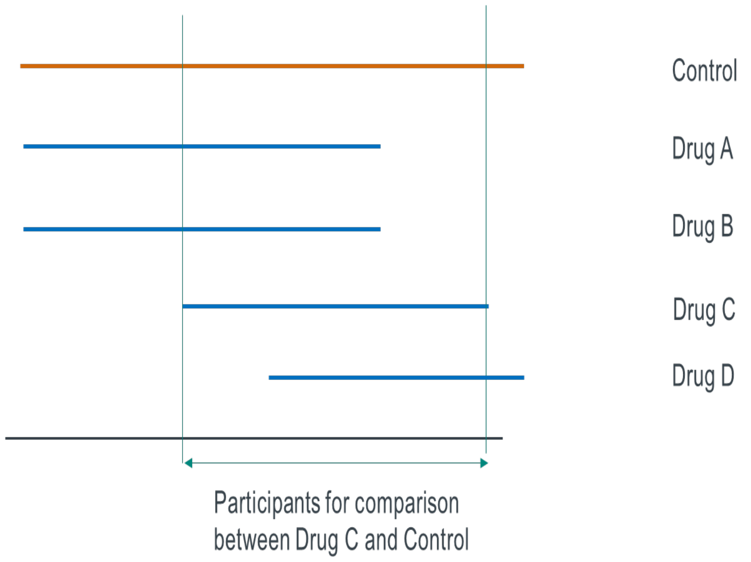
	r1	n1	α*	n
Minimal sample size	2	10	0.018	20
Tumor cohorts with ≥r1/n1 responders will be pooled for analysis at end of second stage and tested at α*				

VS INDEPENDENT EVALUATION (UMBRELLA DESIGN)

Umbrella Design for Exploratory Trials

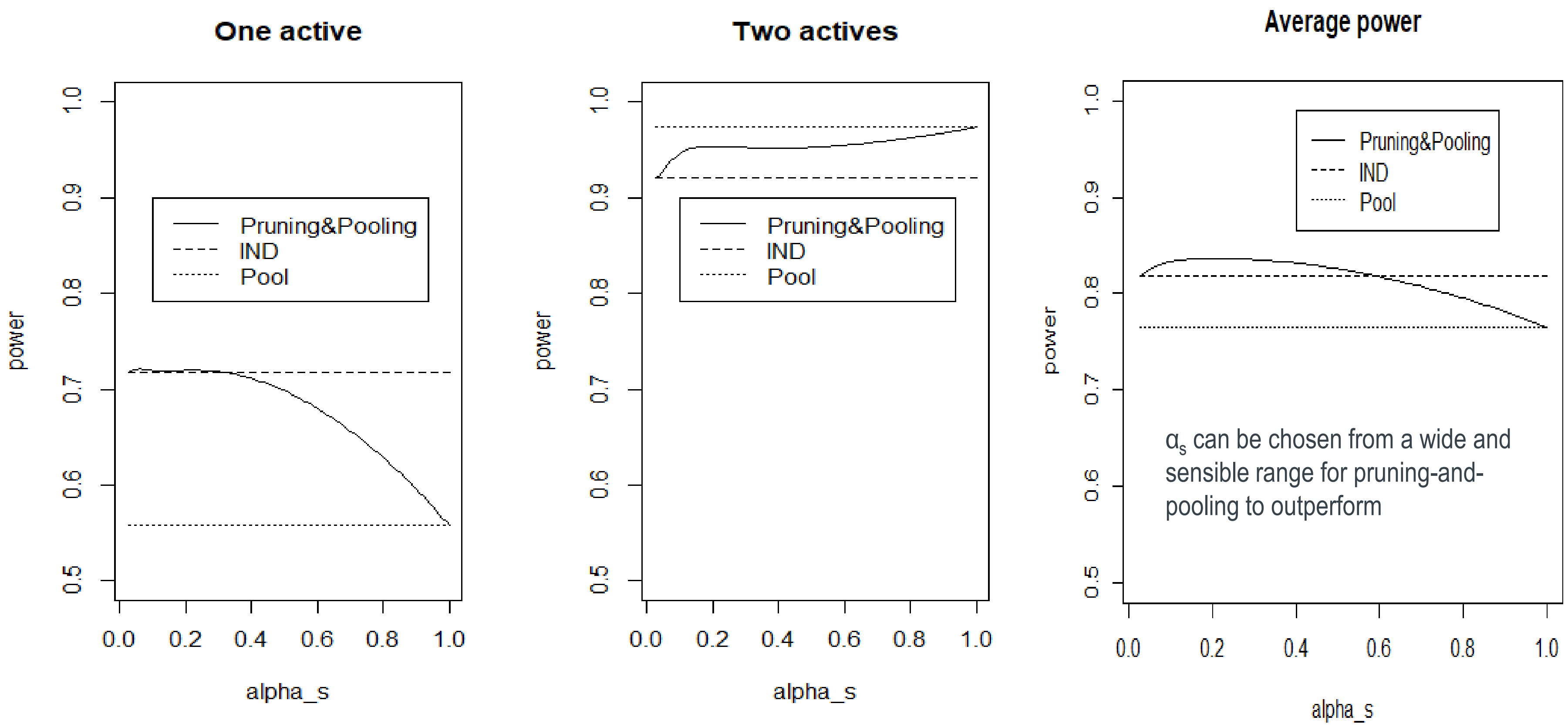
- Does any drug under the umbrella work?
  - Multiplicity is implicitly/explicitly adjusted to mitigate risk of subsequent investment
- When it is applied to a basket trial, it can be viewed as an extreme case of the basket design (i.e., high pruning bar and no pooling)
  - Another extreme type is to pool without pruning
  - An optimal basket design with maximum expected power under a non-informative prior has less extreme bars for pruning and pooling

However, each may be tested at full α in a confirmatory trial (Howard et al. SMMR 2018, Collignon et al. Clin Pharmacol Ther 2020)



Comparison in a Two-tumor Basket Trial

- Umbrella design (IND) works best when only one tumor cohort is active, pooling without pruning (Pool) works best when both are active, and pruning-and-pooling with less extreme bars works best when number of active tumors is uncertain



One active

Two actives

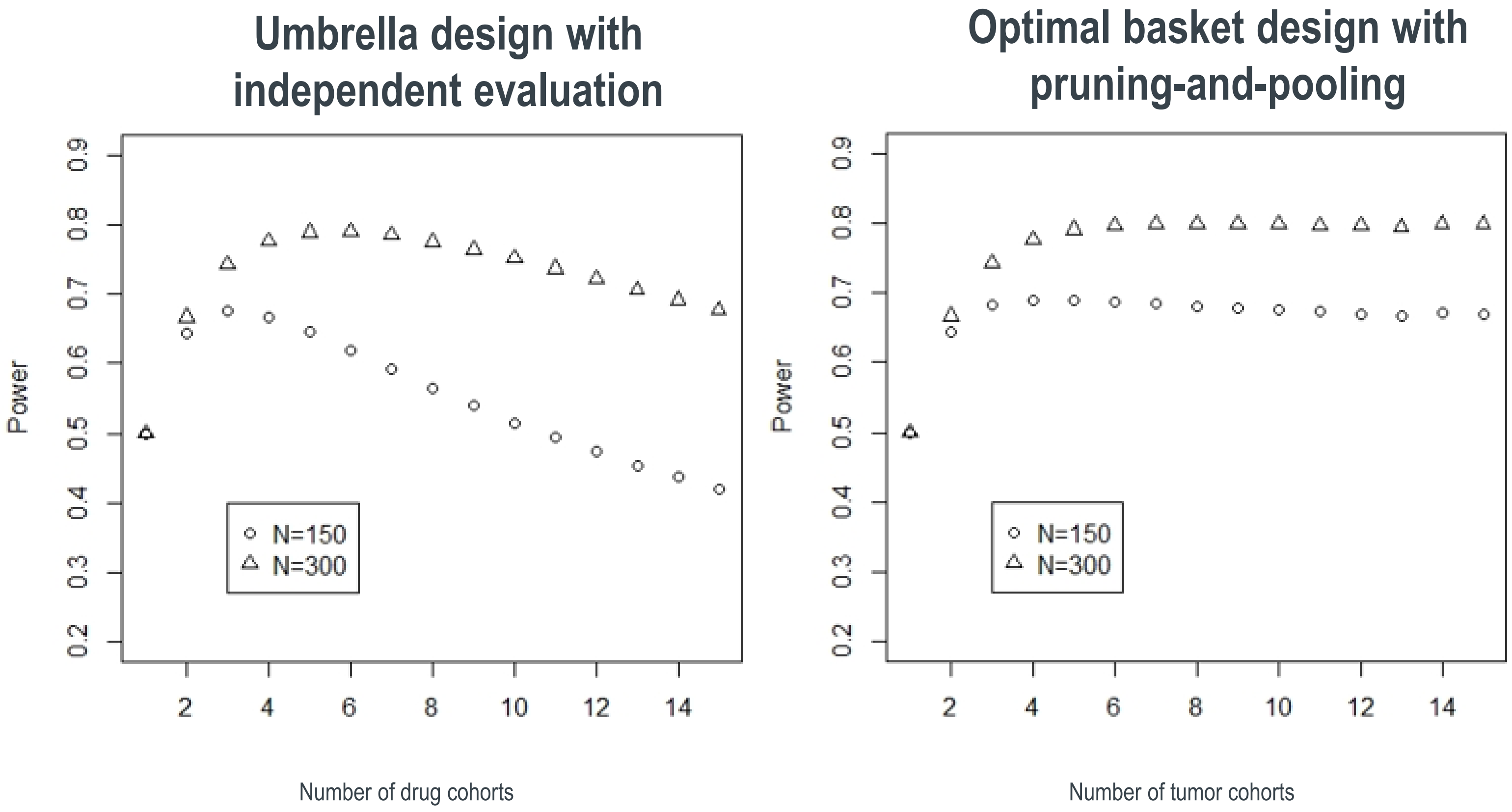
Average power

α_s can be chosen from a wide and sensible range for pruning-and-pooling to outperform

An Expected Power Analysis

- Total sample size N is equally distributed across the selected tumor cohorts
  - Type I error is controlled at 5% overall and one-stage designs considered for simplification
- Underlying probability of a cohort being active with target treatment effect *p*~Beta(a, b)
- How does the expected study power with respect to the prior distribution of *p* (UNKNOWN) change with number of study cohorts?

p~Beta(1, 1): a “non-informative” prior

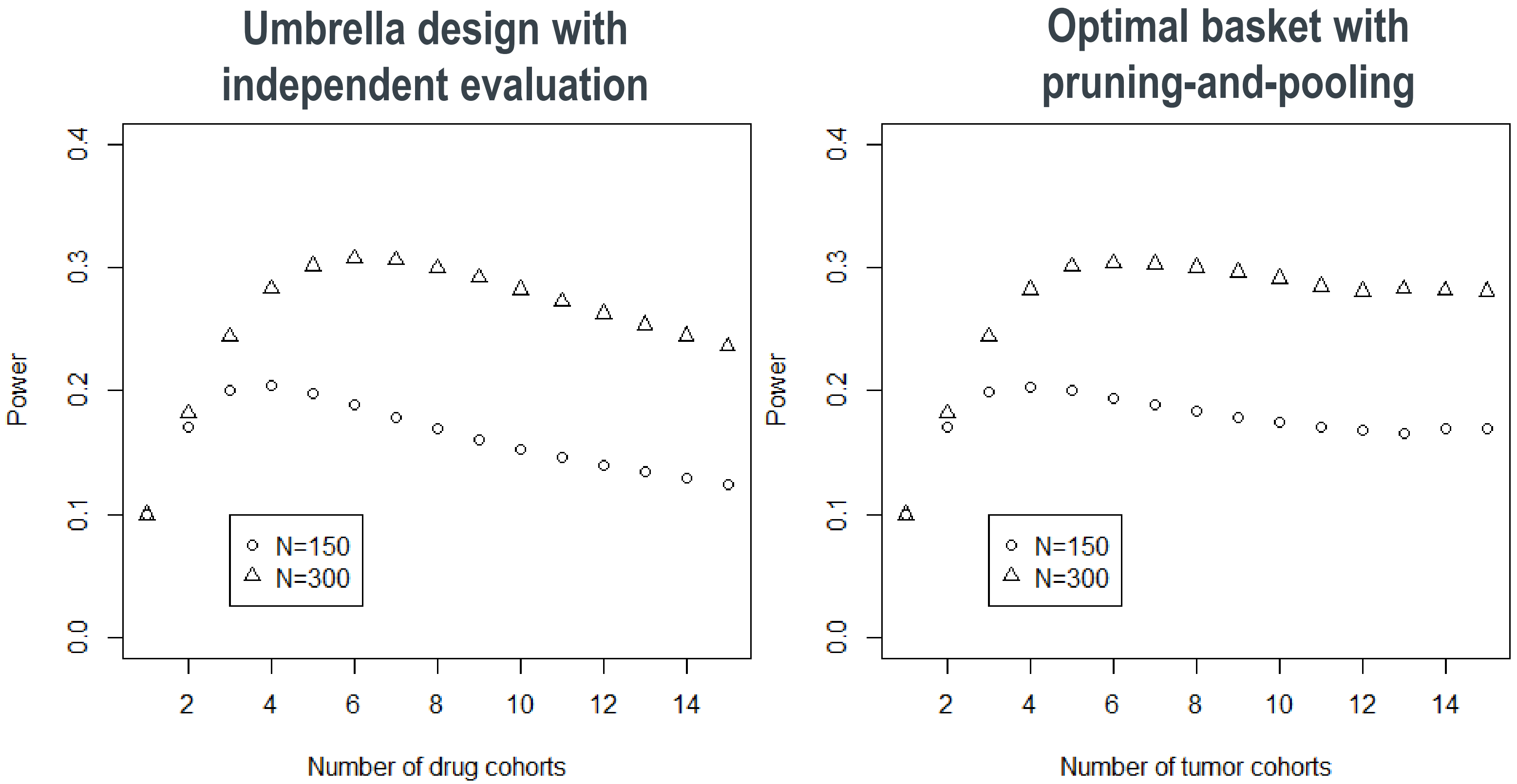


Umbrella design with independent evaluation

Optimal basket design with pruning-and-pooling

~3-4 cohorts under N=150 or ~5-7 cohorts under N=300
Pruning-and-pooling has more robust power curve

p~Beta(1, 9): a more realistic prior



~3-4 cohorts under N=150 or ~5-7 cohorts under N=300
Pruning-and-pooling has more robust power curve

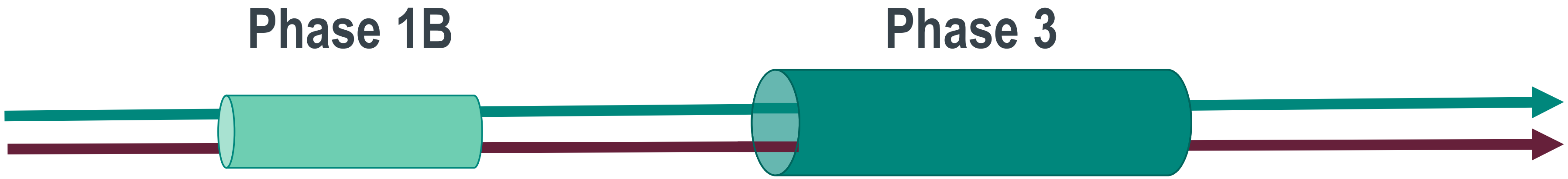
Discussions

- Optimal basket design with pruning-and-pooling is a more efficient method for testing global null hypothesis than independent evaluation (or pooled evaluation), given the unknowns
  - Although the assessment is based on a frequentist approach, general conclusion should apply to all established Bayesian approaches
  - Rejection of the global null ONLY means drug is active but paves the way for further investigation (e.g., add more patients to confirm)
  - Independent evaluation may be more appropriate when the primary objective is to nail down the active tumor indications
- A reasonably resourced exploratory master protocol (e.g., a umbrella trial for multiple drugs or a basket trial for multiple indications) may have ~30-50 patients per study cohort
  - Recommended number of cohorts is consistent with past work (Chen et al. 2017), despite of difference in utility function (predictive power vs benefit-cost ratio)

2-IN-1 ADAPTIVE DESIGN AND EXTENSIONS

Status Quo of Early-to-Late Transition in Oncology

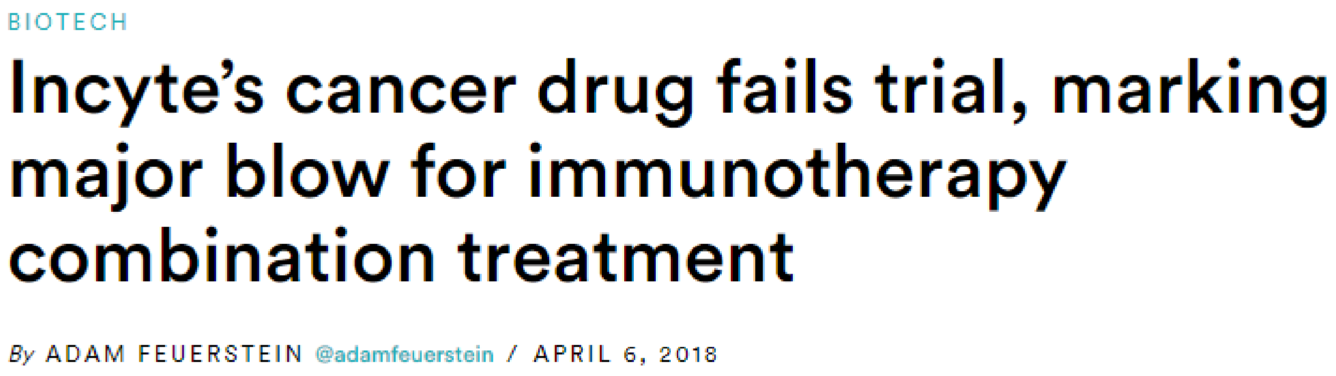
- A typical contemporary oncology program tests a new drug combination with an approved IO in Phase 1B and intends to go directly to Phase 3 once encouraging signal is observed



what are other words for unsustainable?
untenable, indefensible, unjustifiable, unviable, unsuitable, unendurable, unverifiable, weak, unfeasible
Thesaurus.plus

Keytruda+Epacadostat (IDO1) in Melanoma

- Considered the first major breakthrough post PD-1/PD-L1 but no monotherapy activity
- ECHO-202: Phase 1B in combo with Keytruda
  - ORR=**56%** vs ~**37%** for Keytruda alone based on historical data
- ECHO-301 (April 6, 2018)



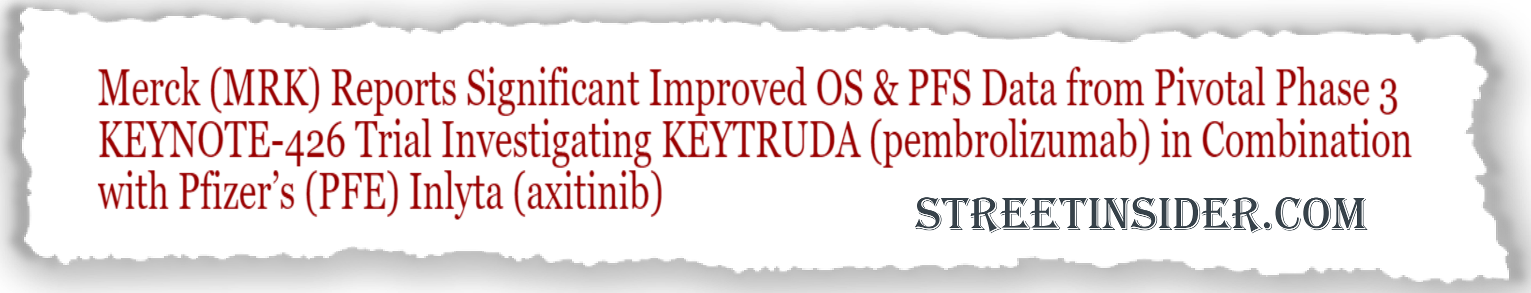
BIOTECH

Incyte's cancer drug fails trial, marking major blow for immunotherapy combination treatment

By ADAM FEUERSTEIN @adamfeuerstein / APRIL 6, 2018

Keytruda+Axitinib in 1L RCC

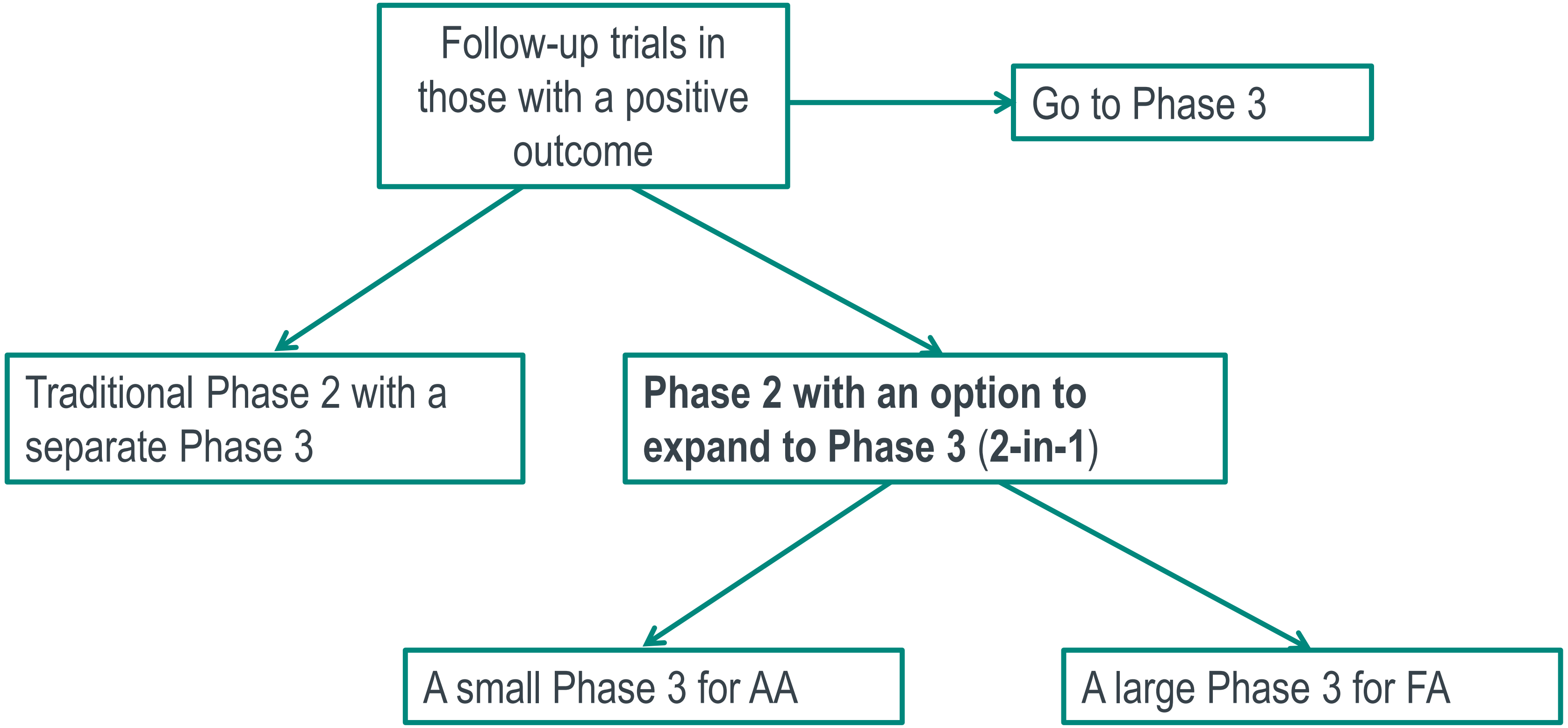
- Both Keytruda and axitinib were known to have monotherapy activity in RCC
- Phase 1B ORR for combo was 38/52 (**73%**; 95% CI 59·0-84·4) (vs **31%** for sunitinib)
  - The median PFS for combo was estimated as 21 months (vs 11 months for sunitinib)
- KN-426 (Oct 18, 2018)



Merck (MRK) Reports Significant Improved OS & PFS Data from Pivotal Phase 3 KEYNOTE-426 Trial Investigating KEYTRUDA (pembrolizumab) in Combination with Pfizer's (PFE) Inlyta (axitinib)

STREETINSIDER.COM

Options Post Phase 1B Efficacy Screening



Follow-up trials in those with a positive outcome

Go to Phase 3

Traditional Phase 2 with a separate Phase 3

Phase 2 with an option to expand to Phase 3 (2-in-1)

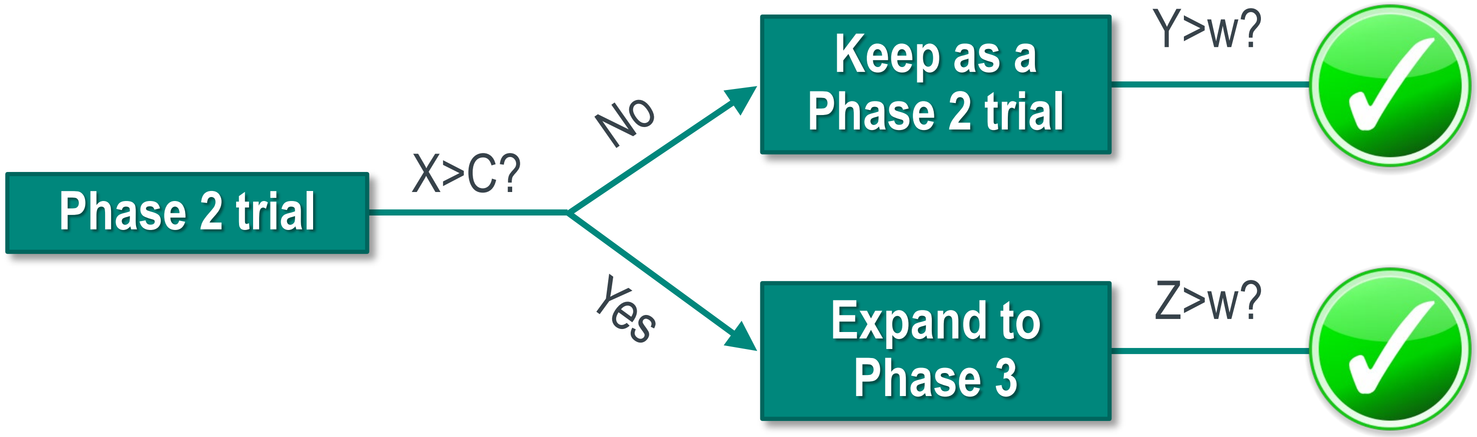
A small Phase 3 for AA

A large Phase 3 for FA

All data used in Phase 3 analysis

2-IN-1 DESIGN

A Generic Statistically Seamless 2-in-1 Design

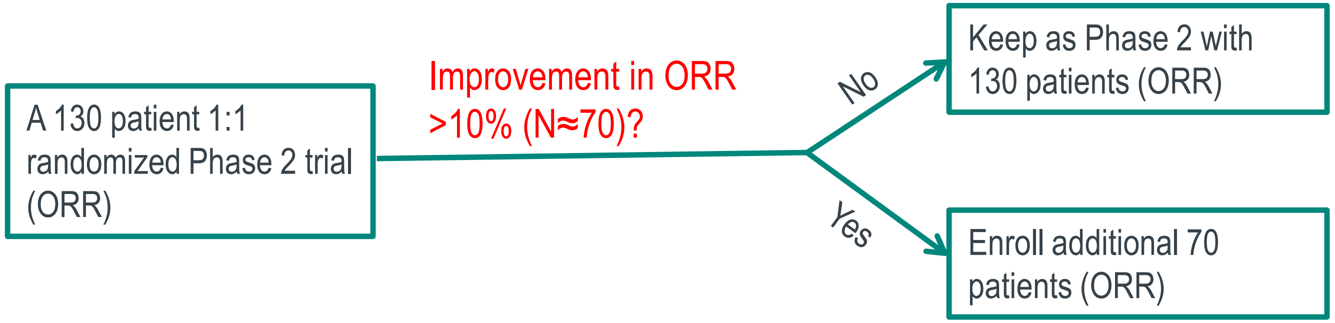


- The 3 endpoints that the standardized test statistics are based upon can be different
  - The expansion bar C is prespecified and binding
- No penalty for multiplicity control as long as $\rho_{XY} \geq \rho_{XZ}$ (automatically holds when Phase 2 endpoint is also used for expansion decision-making)
  - w=1.96 to keep alpha controlled at 2.5% (test Phase 2 at higher level if not for registration)

Chen C, et al. Adaptive Phase 2/3 Design for Expedited Oncology Drug Development. *Contemp Clin Trials.* 2018;64:238-242.

A Small Phase 3 Example

- A small Phase 1B trial of a combination therapy with SOC has demonstrated exciting ORR in 1st line H&N cancer
- A randomized Phase 2 trial based on 2-in-1 design is planned to confirm the signal with the upside for AA



A 130 patient 1:1 randomized Phase 2 trial (ORR)

Improvement in ORR >10% (N≈70)?

No

Keep as Phase 2 with 130 patients (ORR)

Yes

Enroll additional 70 patients (ORR)

- Probability of expansion is 82%, and 15% when true ORR improvement is 21%, and 0%, respectively

A Large Phase 3 Example

- A small Phase 1B trial of a combination therapy with SOC has demonstrated exciting ORR in 1st line gastric cancer
  - More patients are being added to confirm the signal
- A Phase 2/3 trial based on 2-in-1 design is planned at risk to trigger after confirmation
  - Phase 2 is oversized for AA
  - Faster development and fewer patients compared to separate Phase 2 and Phase 3
  - Less risky than straight Phase 3 by skipping Phase 2

Design Details

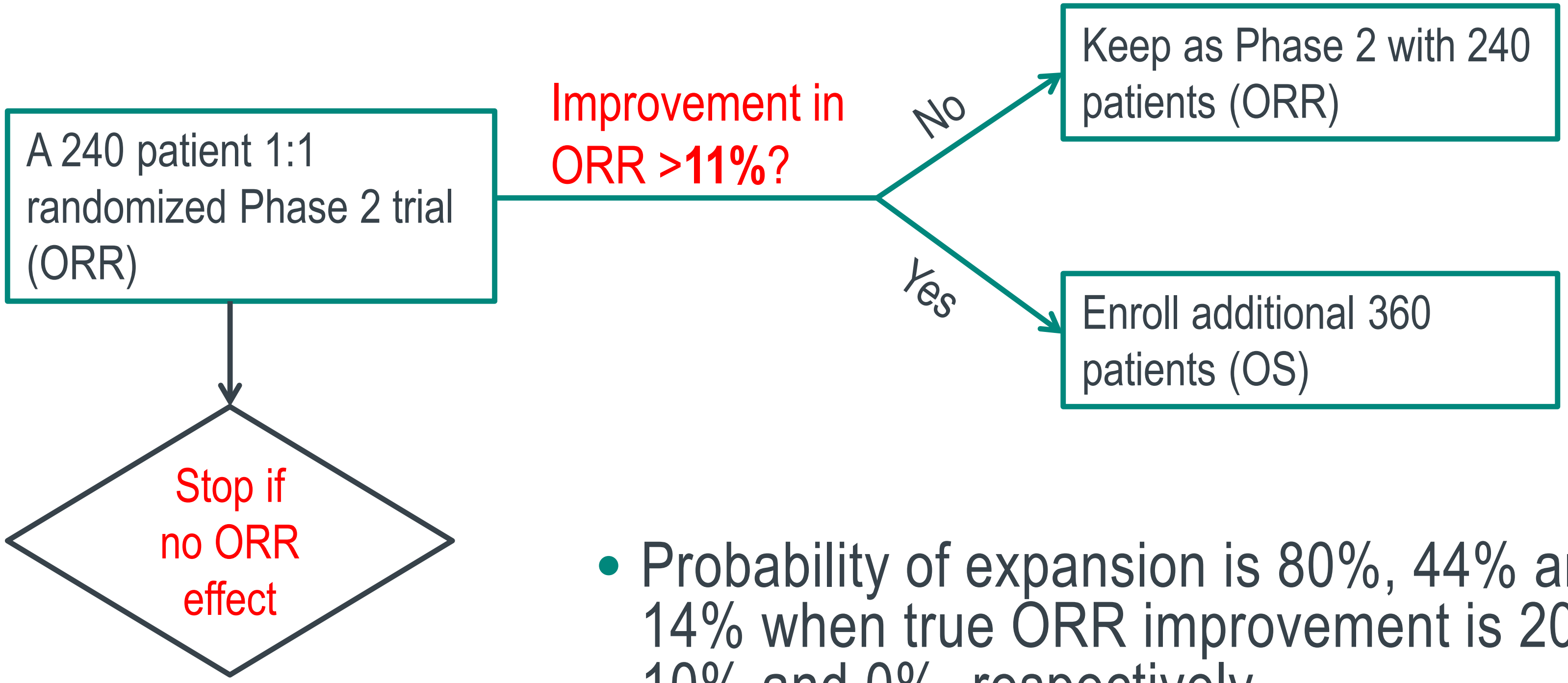
- Phase 2 (in case of no expansion)
  - With 240 patients, it has 88% power for detecting an ORR increase of **20%** at 2.5% (one-sided) alpha level
  - A futility analysis will be conducted to stop the trial early in case of no ORR improvement
  - P-value<0.025 for ORR leads to potential filing for AA
- Phase 3 (in case of expansion)
  - With 460 OS events (600 patients in total), it has 90% power for detecting a hazard ratio (HR) of **0.74** at 2.5% (one-sided) alpha level
  - P-value<0.025 for OS leads to potential filing for FA
- Expansion decision targets one month ahead of Phase 2 accrual completion to ensure seamless expansion

Expansion Bar Based on Benefit-Cost Ratio (BCR) Analysis

- Benefit: value adjusted probability of a positive trial
  - 1/4*prob(positive Phase 2)+3/4*prob(positive Phase 3)
- Cost: expected overall sample size for the study
  - 240+prob(expansion under null or alternative)*360
- Hypotheses with equal probability
  - **Null**: ORR difference=0, HR(OS)=1
  - **Alternative**: ORR difference=20%, HR(OS)=0.74

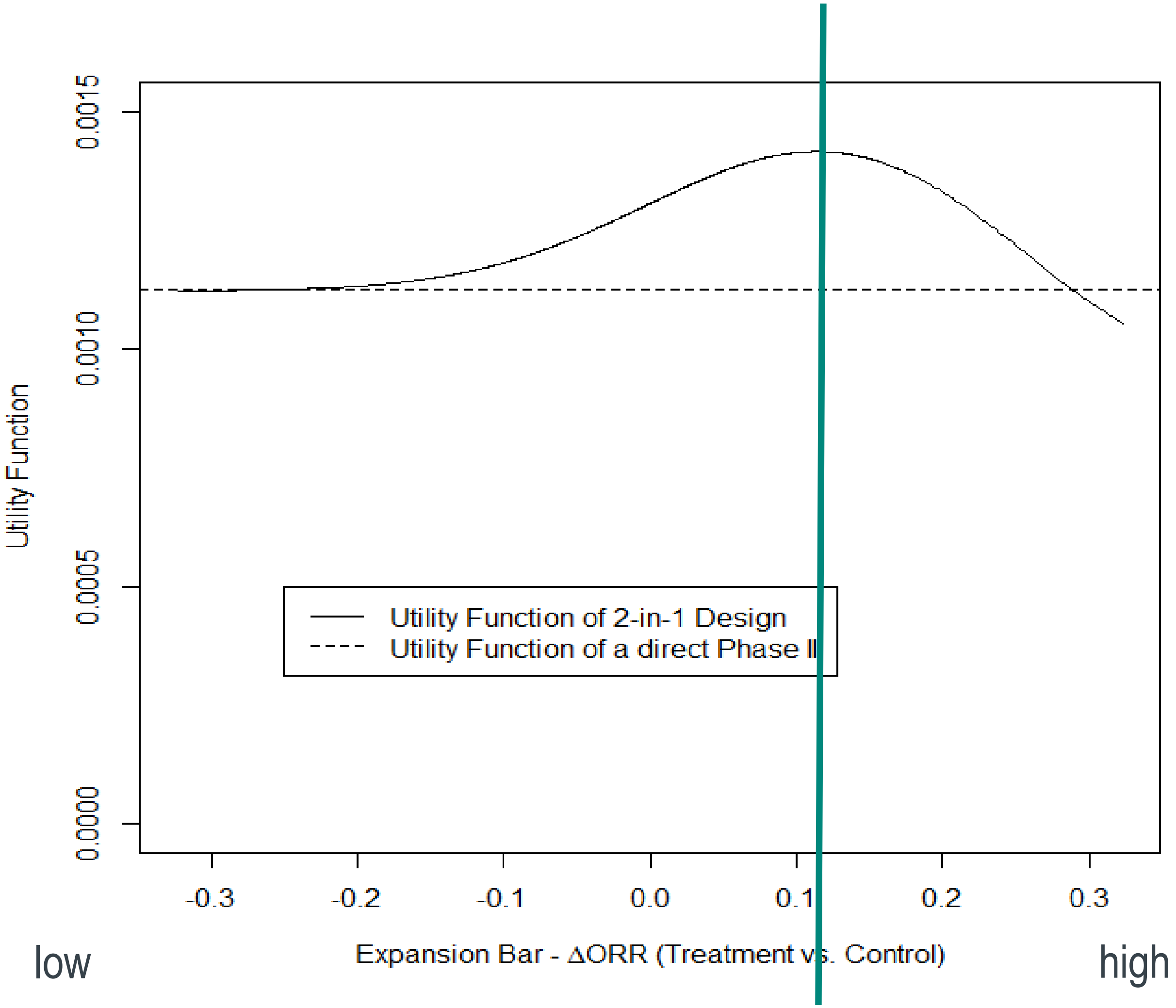
Type I error is controlled for any pre-specified bar

Resulting Design by Maximizing BCR


A 240 patient 1:1 randomized Phase 2 trial (ORR)
Improvement in ORR >11%?
No
Keep as Phase 2 with 240 patients (ORR)
Yes
Enroll additional 360 patients (OS)
Stop if no ORR effect

- Probability of expansion is 80%, 44% and 14% when true ORR improvement is 20%, 10% and 0%, respectively
- Probability of a positive Phase 2 is ~50% if true ORR is 11% but is potentially higher due to longer follow-up

BCR vs Expansion Bar



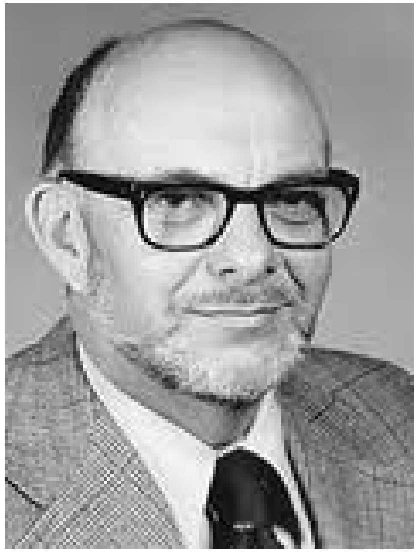
Robustness to Input Variables

Prior distribution of treatment effect for OS		Relative value of a positive Phase 2 vs. a positive Phase 3	Approximate optimal expansion bar in ΔORR
P(HR = 0.74)	P(HR = 1)		
1/3	2/3	1:3	12%
		1:5	10%
1/2	**1/2**	**1:3**	**11%**
		1:5	9%
2/3	1/3	1:3	10%
		1:5	8%

EXTENSIONS

Extensions

- Multiple Adaptive Decisions Overtime
- Multiple Cutpoints at Same Time
- Application of Group Sequential Method
- Multiple Intermediate Endpoints for Expansion Decision
- Multiple Clinical Endpoints



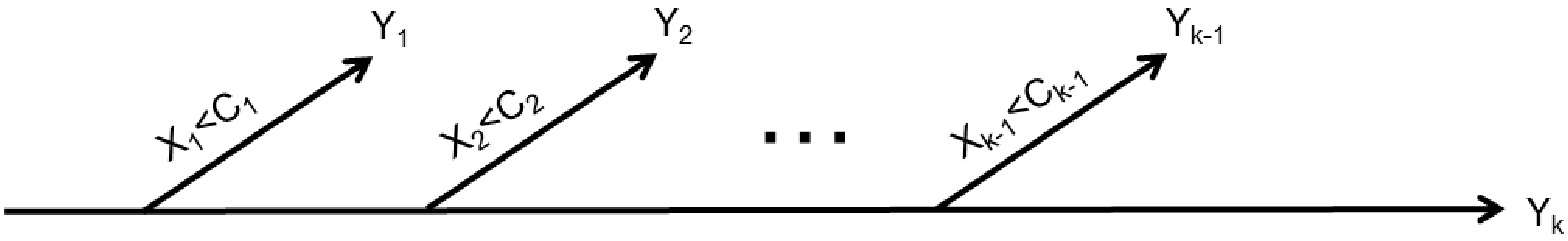
David Slepian 1923-2007

Keep overall Type I error under control under least assumptions

Chen C, Li W, and Deng Q. Extensions of the 2-in-1 design. *Contemp Clin Trials* 2020. *DOI: 10.1016/j.cct.2020.106053.*

Multiple Adaptive Decisions Overtime

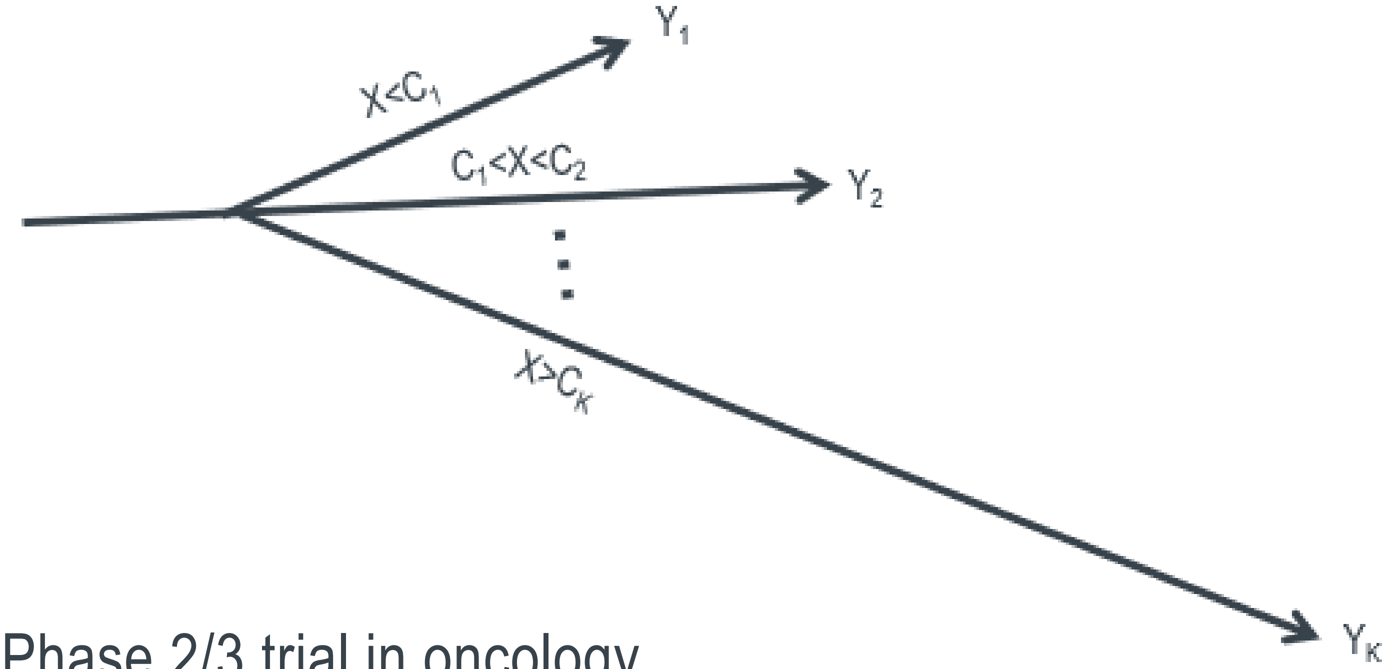
- Sample size increases each time an expansion bar is crossed
  - Y_j's can all be tested at the *α* level if corr(X_j, Y_l) is non-increasing in l $(j \leq l \leq K)$, which is generally expected to hold due to the nested structure of the study populations
  - Overall Type I error tends to decrease with *K*



- A hypothetical Phase 2/3 trial in oncology
  - Both X_1 and Y_1 may be based on objective response rate (ORR) while X_2 and Y_2 are based on progression-free-survival (PFS) and Y_3 is based on the overall survival (OS)

Multiple Cutpoints at Same Time

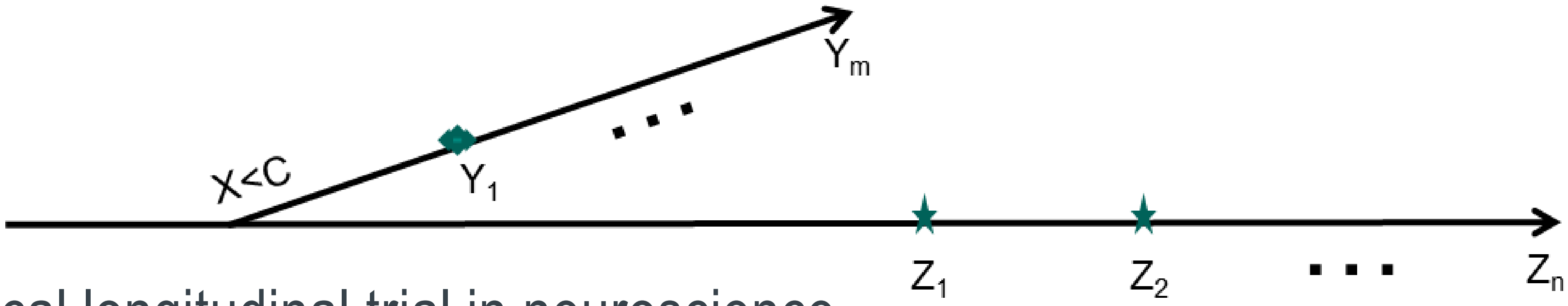
- Sample size increases with expansion bar
  - Y_j's can all be tested at the *α* level if corr(Y_j, X) is non-increasing in *j* $(1 \leq j \leq K)$, which is generally expected to hold, and overall Type I error tends to decrease with *K*



- A hypothetical Phase 2/3 trial in oncology
  - Both *X* and Y_1 may be based on ORR while Y_2 is based on PFS and Y_3 is based on OS

Group Sequential Design

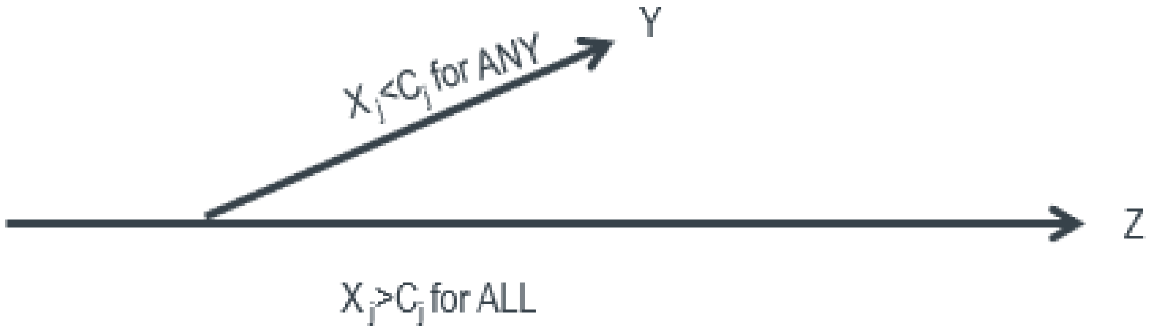
- An alpha-spending function is pre-specified for each scenario (expansion or not) that controls the Type I error under each at the α level.
  - The overall Type I error is controlled at the *α* level if $\rho_{XY_m} \geq \rho_{XZ_1}$, or roughly speaking the first interim analysis in case of expansion should be no sooner (or based on more information) than the final analysis in case of no expansion
  - A rigorous proof may require an extension of Slepian' Lemma, and is an open question



- A hypothetical longitudinal trial in neuroscience
  - Primary endpoint is continuous whereas measurement at an early time point (*X*) is used for adaptive decision and at later timepoints (*Y* and *Z*) are for hypothesis testing

Multiple Intermediate Endpoints for Expansion Decision

- To ensure robust control of Type I error, expand only when ALL expansion bars are crossed
  - Overall Type I error is controlled at the α level if corr(X_j, Y)≥corr(X_j, Z) for all j



- Hypothetical examples
  - In early stage Alzheimer disease, an improvement not only on the primary endpoint but also on other related cognitive scores and daily activities may be required to move forward
  - A new drug may need to be better than SOC in both safety and efficacy to be viable

Multiple Clinical Endpoints

- There are many ways to allocate alpha. The simplest is to apply a conservative Bonferroni approach to both scenarios but caution must be exerted.
- E.g., when corr(X, Y_j)≥ corr(X, Z_j) for 1≤j≤max{M, N}, same α_j can be used for Y_j and Z_j
  - In the special case of M=1, it can be tested at any α_j level if corr(X, Y_1)≥ corr(X, Z_j). But in order to enjoy full α, a nested correlation structure for corr(X, Z_j) may be needed.

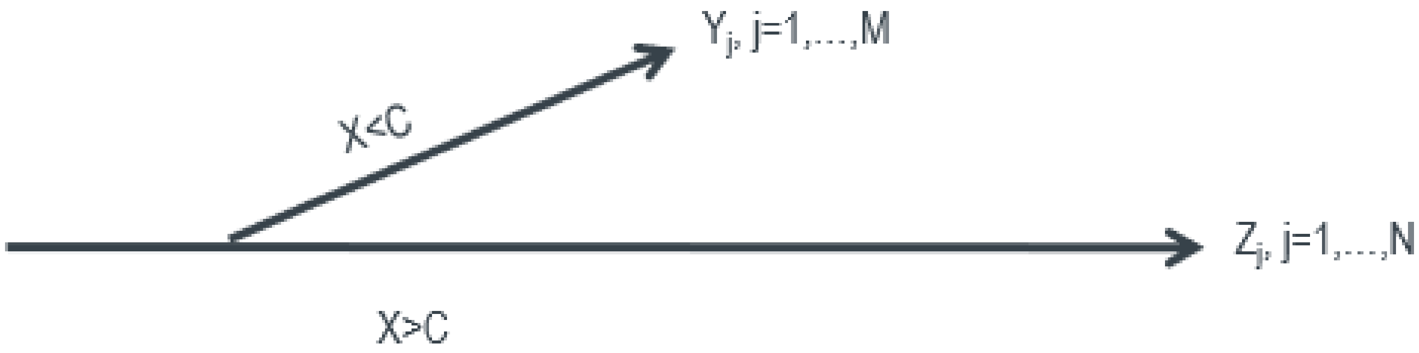
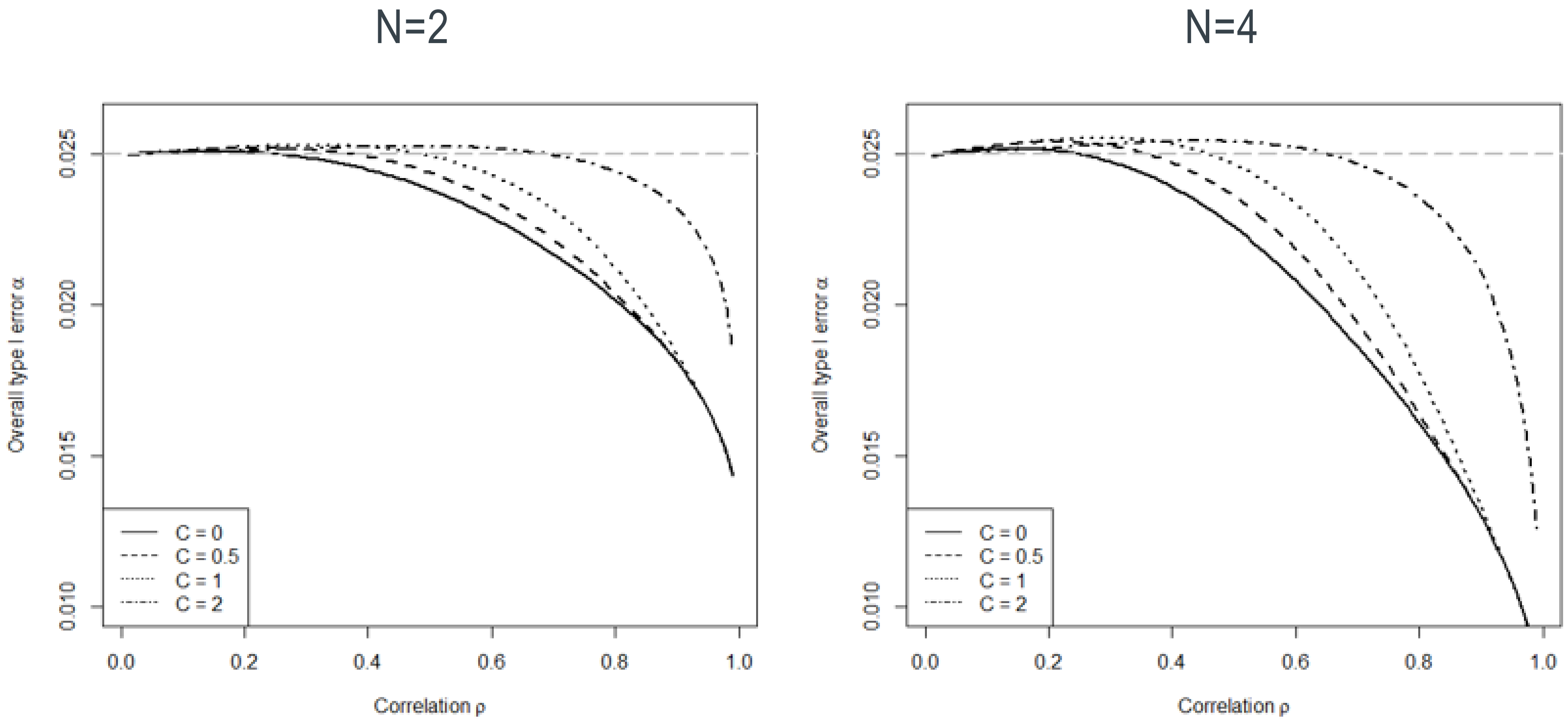


Illustration of a Counter Example Against Full α under M=1

- Despite of Bonferroni correction for the *N* primary endpoints, Y_1 cann’t tested at α when all involved test statistics have a common correlation *ρ* (a violation of the nested structure)



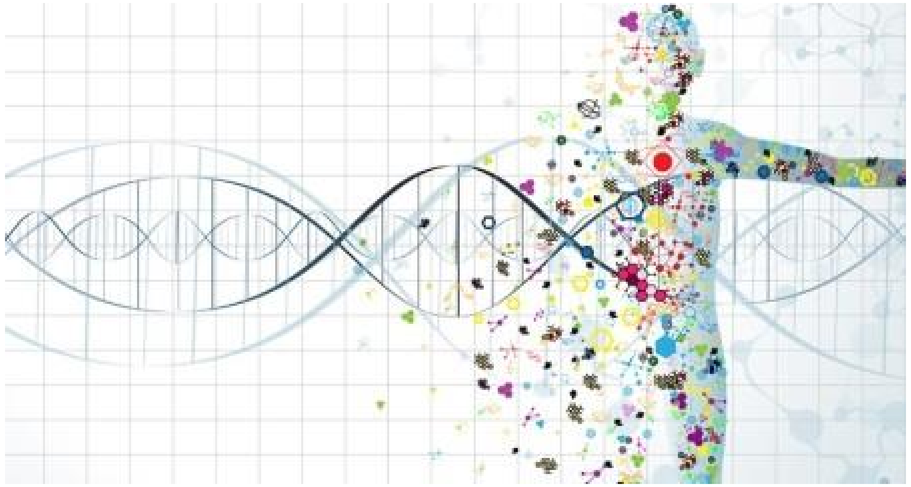
Discussions

- The designs are devised under minimum assumptions to remain robust and conservative
  - In practice, some of the conditions may be relaxed and more alpha may be recouped
  - Graphical approach may be incorporated to further improve the efficiency
  - Validity of the designs hinges upon relationship of correlations among test statistics, which is expected to hold in general but may require examination otherwise
- In practice, a clinical trial may contain a mixture of the extended features and other features
  - **Careful investigation may be needed to ensure Type I error control**
- A decision of no expansion is not the same as termination of futility
  - When it comes to deciding the expansion bars, both statistical operating characteristics and risk-adjusted cost-effectiveness should be considered.

PHASE 3 PROGRAMS WITH BIOMARKER CONSIDERATIONS

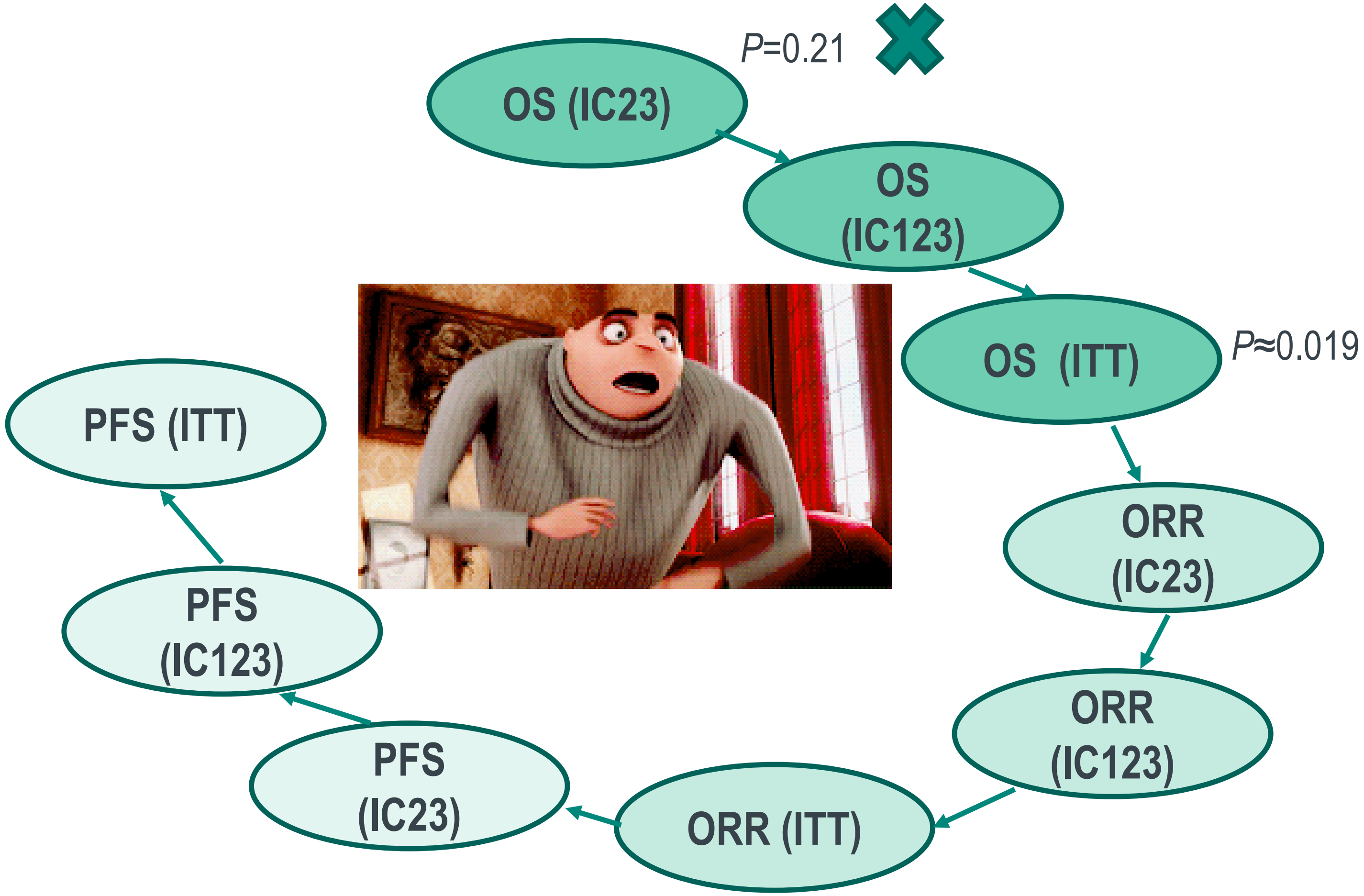
Status Quo

- A biomarker hypothesis is often built into a Phase 2/3 program after data from a Phase 1B single arm trial has shown stronger anti-tumor activity for an experimental drug in a biomarker+ population than in the biomarker- population
  - The uncertainty on the predictive biomarker is less characterized before entering into Phase 3 testing, and the risk is not well mitigated or sometimes totally ignored
  - Best opportunity for adaptive designs but rarely taken advantage of in practice



IMvigor211 in 2L UC

- Previous data from a single arm study seems to support a step-down approach



OS (IC23)

P=0.21 ✖

OS (IC123)

OS (ITT)

P≈0.019

ORR (IC23)

ORR (IC123)

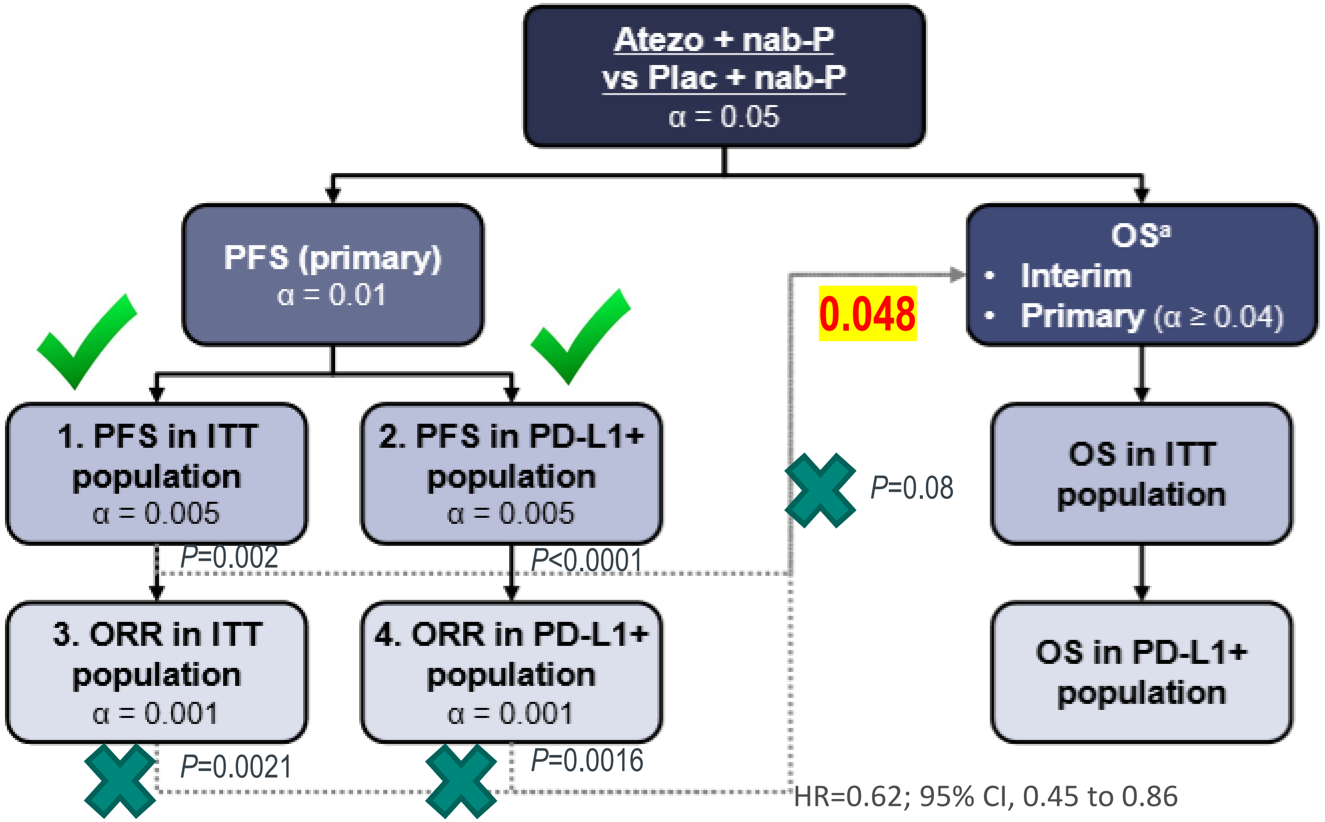
ORR (ITT)

PFS (IC23)

PFS (IC123)

PFS (ITT)

IMpassion130 in 1L TNBC



Atezo + nab-P
vs Plac + nab-P
α = 0.05

PFS (primary)
α = 0.01

OS[a]
- Interim
- Primary (α ≥ 0.04)

0.048

1. PFS in ITT population
α = 0.005

2. PFS in PD-L1+ population
α = 0.005

OS in ITT population

OS in PD-L1+ population

3. ORR in ITT population
α = 0.001

4. ORR in PD-L1+ population
α = 0.001

P=0.002

P<0.0001

P=0.08

P=0.0021

P=0.0016

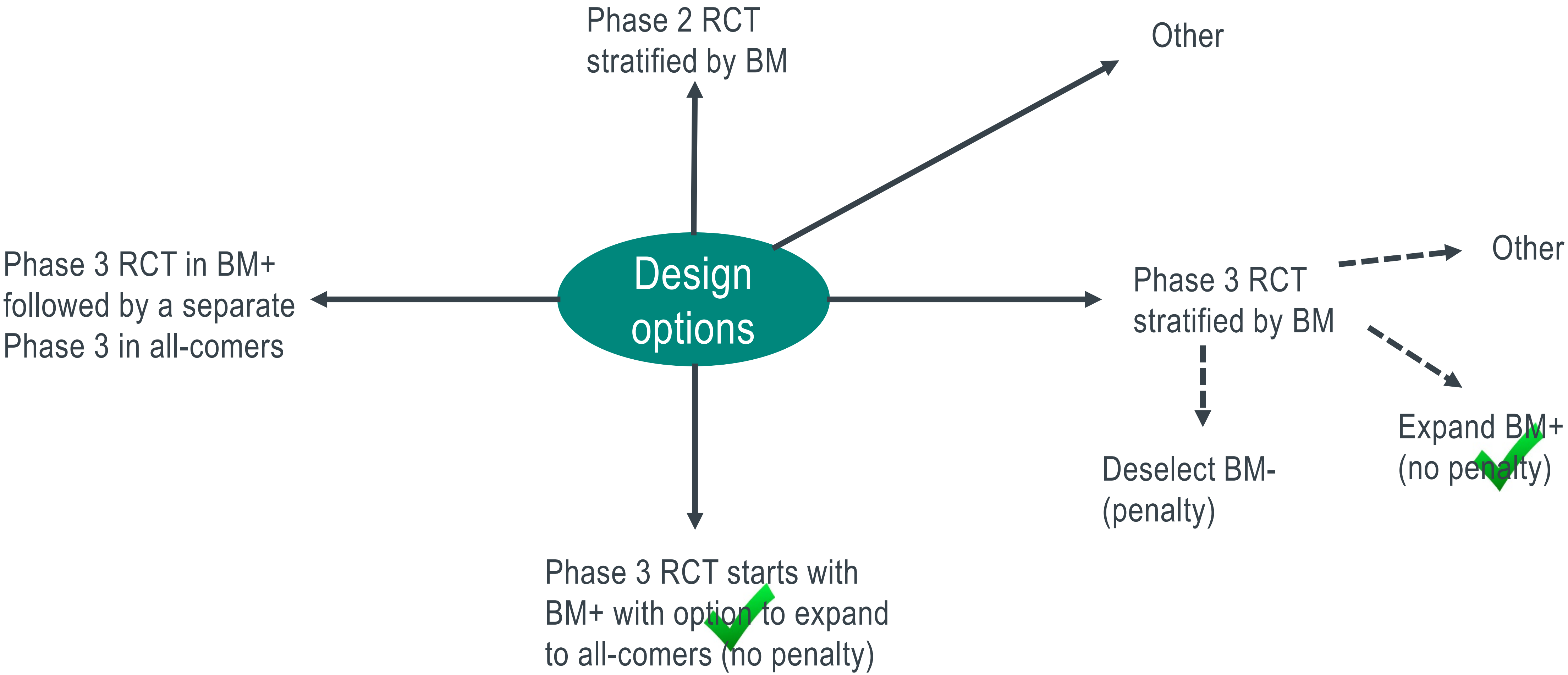
HR=0.62; 95% CI, 0.45 to 0.86

- Primary PFS analysis (PFS tested in ITT and PD-L1+ populations)
- First interim OS analysis (OS tested in ITT population, then, if significant, in PD-L1+ population)

P≈0.0034 and OS would be positive in PD-L1+ should some alpha be allocated

ADAPTIVE POPULATION EXPANSION

What to Do with An Early Biomarker Signal?



Design options

Phase 2 RCT stratified by BM

Other

Phase 3 RCT in BM+ followed by a separate Phase 3 in all-comers

Phase 3 RCT stratified by BM

Other

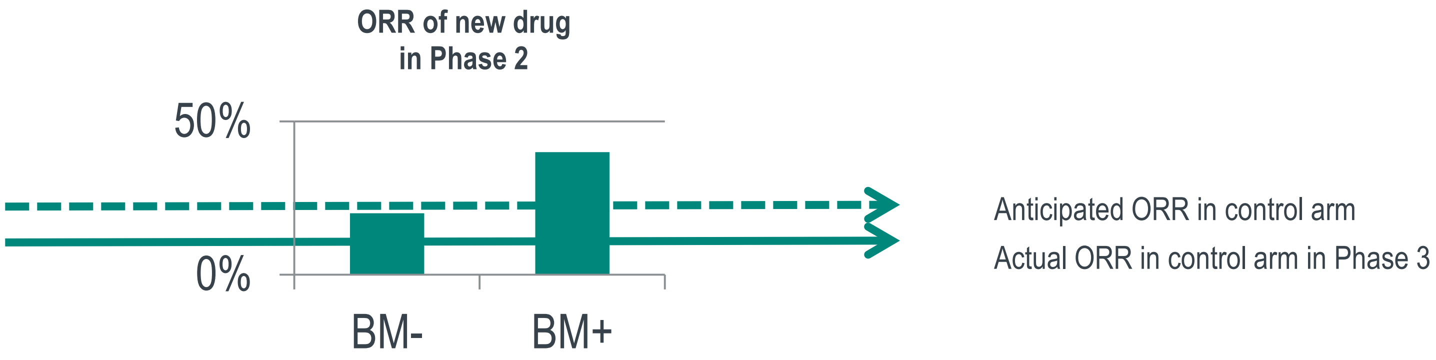
Deselect BM- (penalty)

Expand BM+ (no penalty) ✓

Phase 3 RCT starts with BM+ with option to expand to all-comers (no penalty) ✓

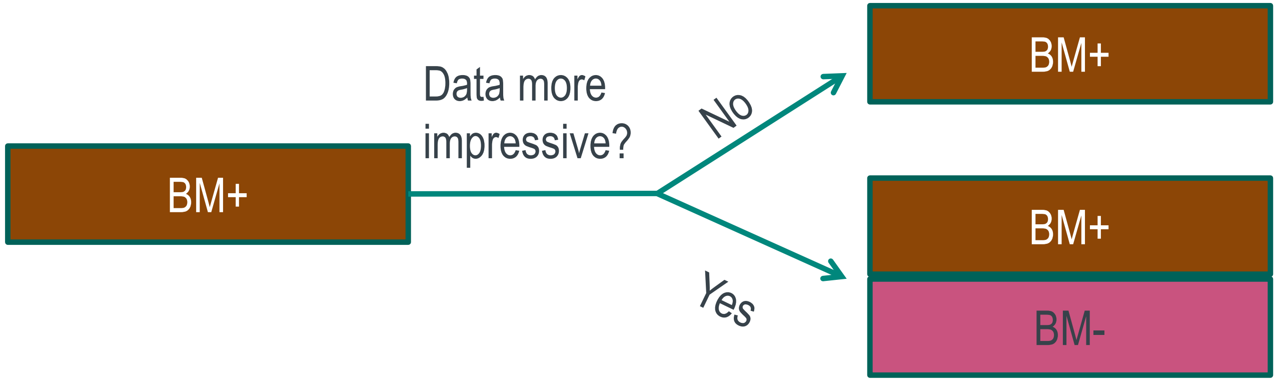
Expansion of BM+ Patients to All-comers

- In a single arm Phase IB study, an investigational new drug showed **similar** ORR overall to SOC based on historical data but **higher** ORR in a BM+ population
- A biomarker enrichment study is justifiable, but upside for a broader label can't be totally ruled out given the preliminary data



An Adaptive Approach With One Study

- Enroll BM+ patients first, expand to all-comers if interim data more promising than expected, suggesting likely broader activity
  - Patients used for expansion decision are included in primary analysis of BM+ population but not in the all-comer population

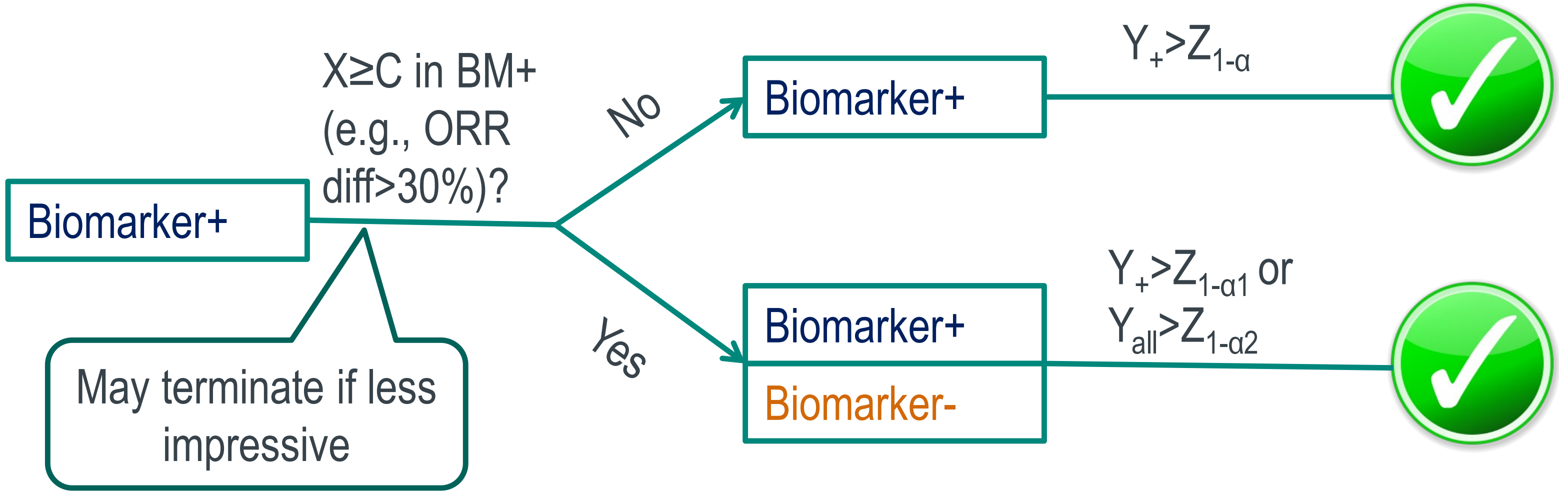


- Any penalty for multiplicity control? No but…

Chen C, Li X, Li W, Beckman RA. Adaptive Expansion of Biomarker Populations in Phase 3 Clinical Trials. *Contemporary Clinical Trials* 2018;71:181-185.

A General Design

- X: test statistics based on the endpoint for adaptive decision
- Test statistics based on the primary endpoint
  - Y_{all}: based on the all-comers enrolled POST-adaptation
  - Y_{+}: based on the BM+ population as planned



$\alpha_1+\alpha_2=\alpha$

Multiplicity Control

- Overall Type I error is controlled at α as long as $\alpha_1+\alpha_2 \leq \alpha$ w/o any constraint on E{X} and C when $\mathrm{Corr}(X, Y_+) \geq 0$
- Correlation assumption automatically holds when decision is based on the primary endpoint, and also automatically holds when the two endpoints have a positive correlation
  - For IOs, responders clearly tend to live longer (i.e., evidence of positive correlation, which can be validated with trial data as needed)

Application to A Hypothetical Trial

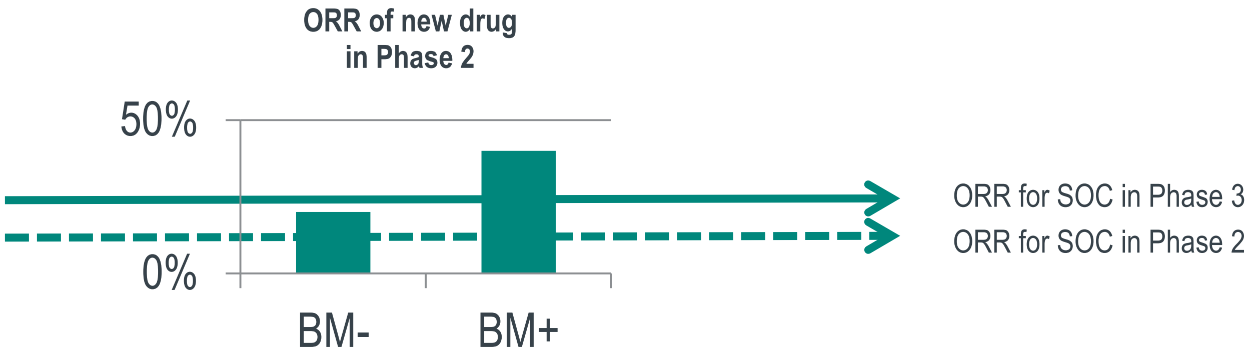
- The study targets to enroll 350 BM+ patients in 15 months and completes after 230 death events are observed
- An interim analysis is conducted after 150 patients are enrolled, ~400 all-comers will be enrolled if treatment effect is greater than expected
  - Half are expected to be BM+

Approximate sample size for overall program

Approaches	No-Go to all-comers based on BM+ data	Go to all-comers based on BM+ data
Sequential	350	350+400
Staggered/Parallel	350+400	350+400
Adaptive	350	150+400

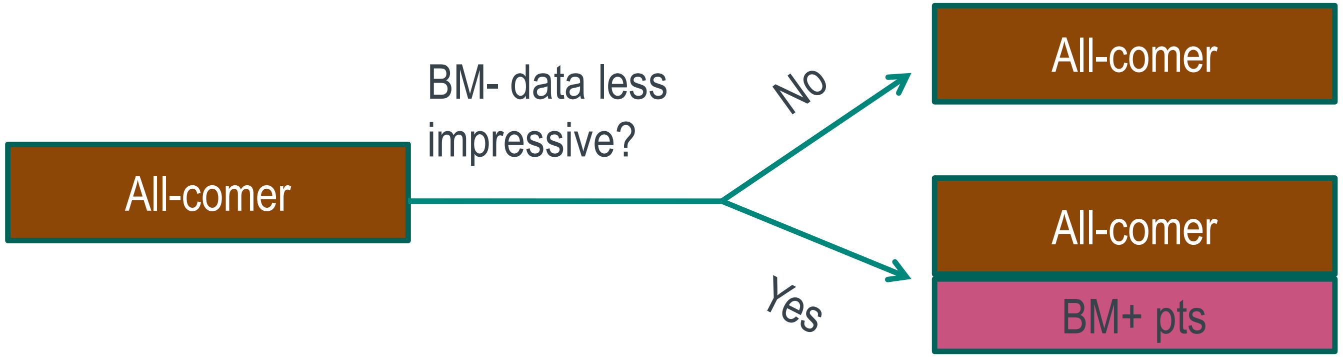
All-comer Study With A Biomarker Hypothesis

- In a small Phase 2 randomized study, an investigational new drug **improved** ORR over SOC in **both** biomarker subpopulations but more so in BM+ population
- An all-comer study is justifiable, can we add more BM+ patients in case data is less promising in BM- population?



Adaptive Expansion of BM+ Population

- Enroll all-comer patients, add more BM+ patients if interim data in BM- patients is less promising which suggests lower POS in BM+ population
  - – Patients used for expansion decision are included in primary analysis of all analyses, but the additional BM+ patients are excluded in all-comer analysis

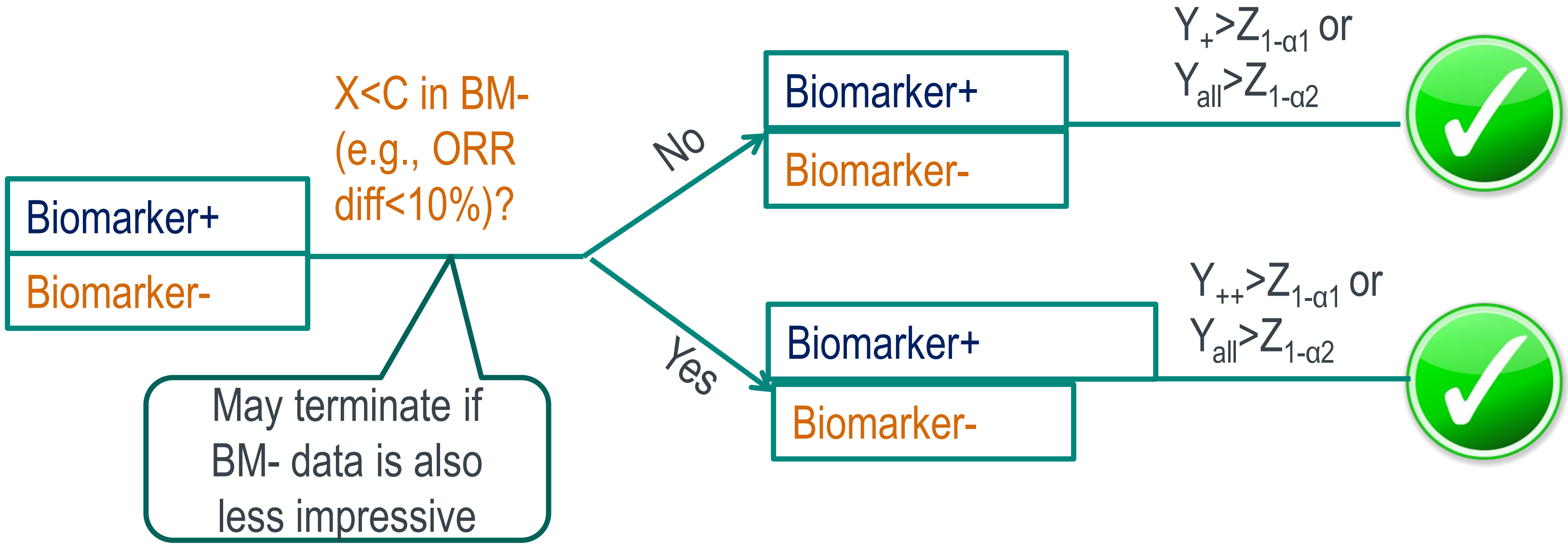


- Any penalty for multiplicity control? No!

Chen C, Li X, Li W, Beckman RA. Adaptive Expansion of Biomarker Populations in Phase 3 Clinical Trials. *Contemporary Clinical Trials* 2018;71:181-185.

A General Design

- X: test statistics based on the endpoint for adaptive decision
- Test statistics based on the primary endpoint
  - Y_{all}: based on all-comer population as planned
  - Y_{+}: based on BM+ patients in all-comers
  - Y_{++}: based on ALL BM+ patients



Biomarker+

Biomarker-

X<C in BM- (e.g., ORR diff<10%)?

May terminate if BM- data is also less impressive

No

Biomarker+

Biomarker-

$Y_{+}>Z_{1-\alpha 1}$ or $Y_{all}>Z_{1-\alpha 2}$

Yes

Biomarker+

Biomarker-

$Y_{++}>Z_{1-\alpha 1}$ or $Y_{all}>Z_{1-\alpha 2}$

$\alpha_1+\alpha_2=\alpha$

Multiplicity Control

- Overall Type I error is controlled at α **irrespective** of E{X}, C and the correlation structure among the test statistics
  - While the BM- population used for adaptation decision is also included in the analysis of the all-comers, there is no modification on sample size or hypothesis testing strategy for the all-comer population
  - The decision to increase sample size in BM+ population is driven by BM- patients, an independent data source

Application to A Hypothetical Trial

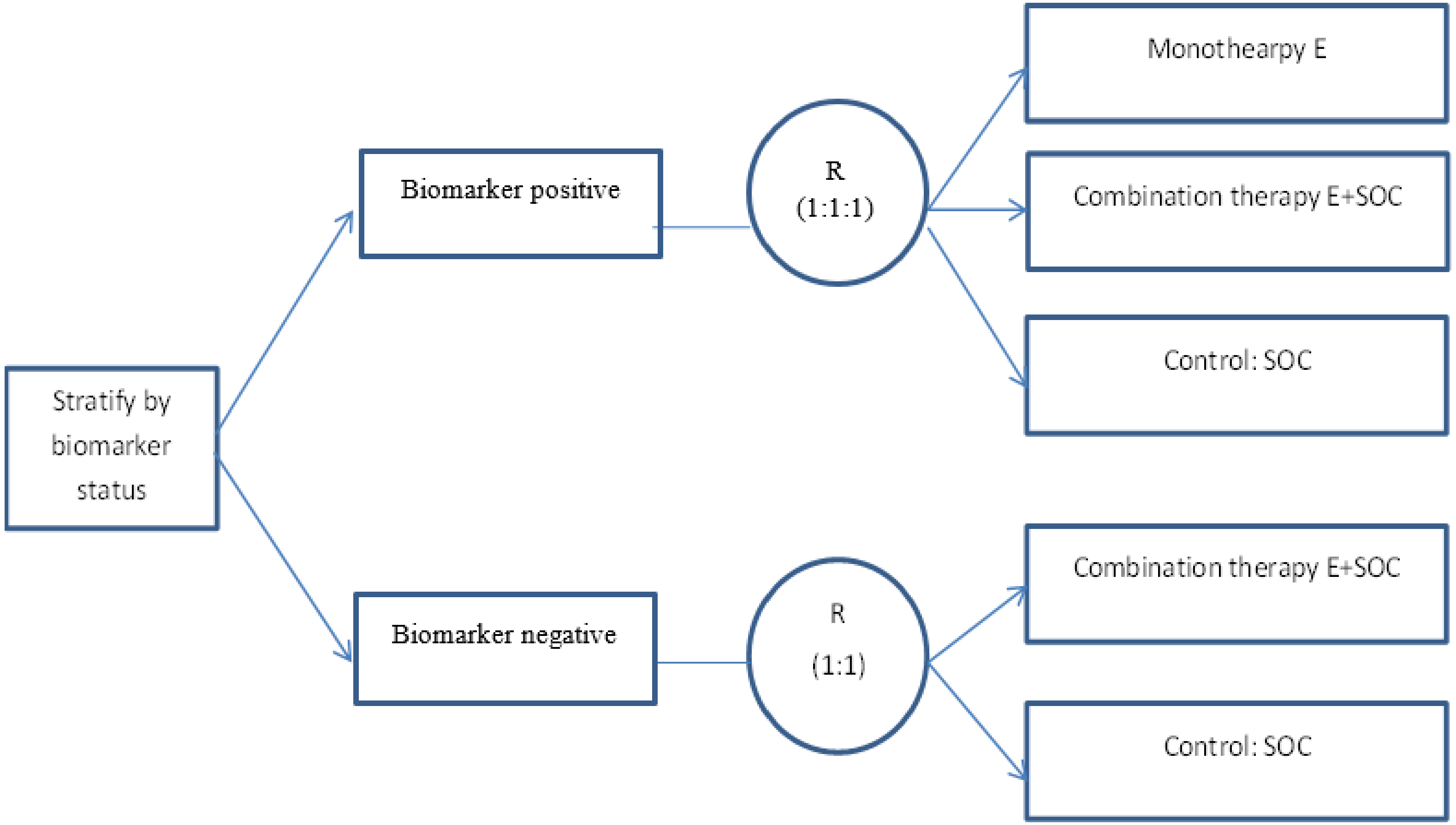
- The all-comer study targets to enroll 510 patients and completes after 300 death events are observed
  - With ~100 events in the BM+ population (1/3 overall), the study has 80% power to detect a hazard ratio of 0.50 at α_1=0.005
- An interim analysis is conducted after ~210 patients are enrolled, and ~100 BM+ patients will be added if data in BM- population is less promising
  - With events expected to increase from 100 to 150, it now has 85% power to detect a smaller treatment effect (hazard ratio of 0.55) in this population at same alpha

TEST MONO IN BM+ AND COMBO IN ALL-COMERS

Mono for BM+ and Combo for All-comers

- There is often an interest in testing the monotherapy of an investigational new drug vs SOC in a BM+ population, and combination with SOC vs SOC in all-comers
- The conventional approach conducts two separate trials
  - Less efficient as both trials enroll BM+ patients to the SOC arm but data are not shared
  - Unfair to BM- patients who failed to meet eligibility criterion for the BM+ study, and potential to skew biomarker prevalence in the all-comer study
  - If the two trials are conducted at the same time at same sites, which trial should a BM+ patient participate?

One Trial Design



Sun L, Kang SP, Chen C. Testing of Monotherapy and Combination Therapy in One Trial with Biomarker Consideration, *Contemporary Clinical Trials* 2019. DOI: 10.1080/19466315.2019.1665578.

Statistical Analyses

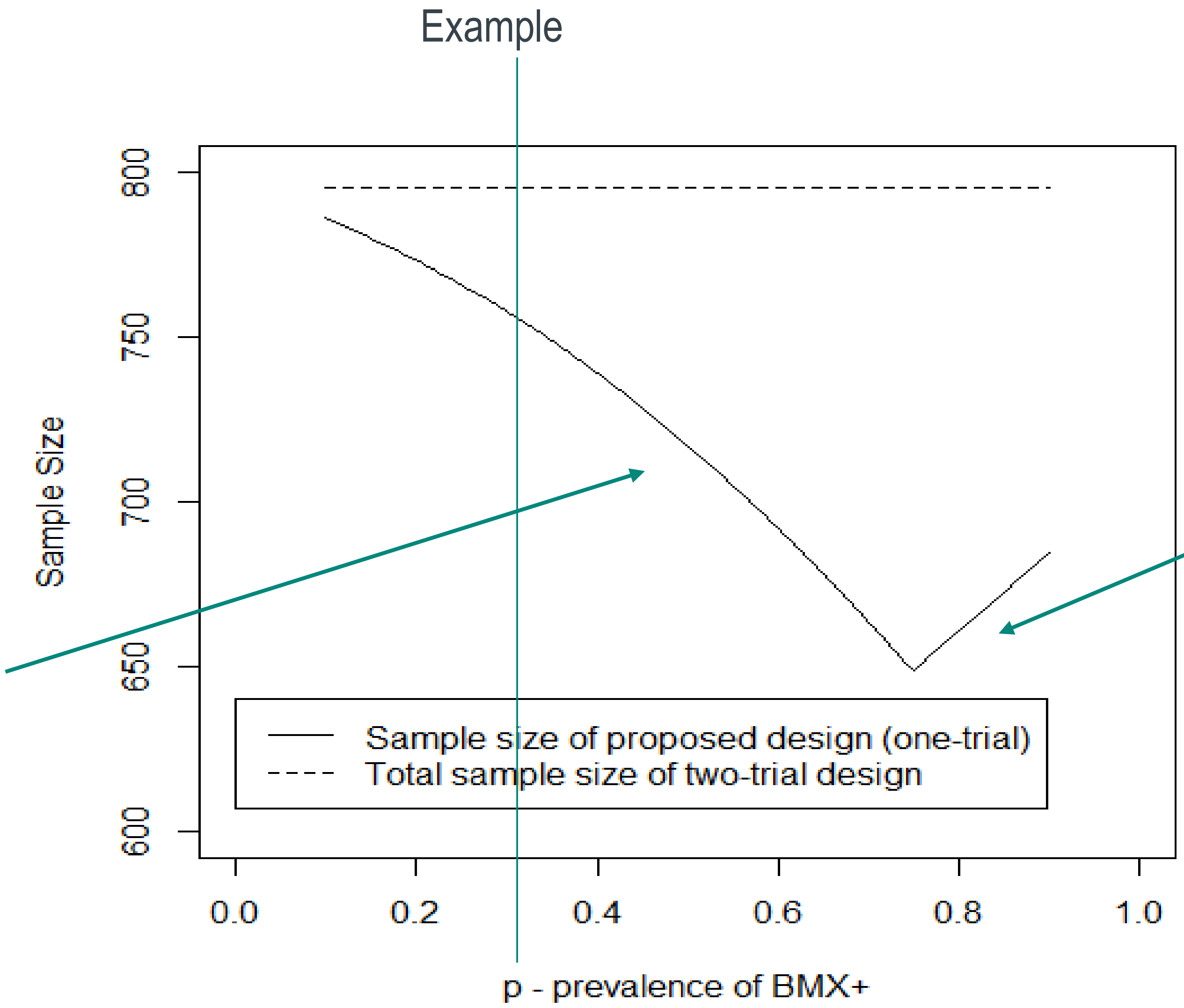
- **No multiplicity adjustment is needed as mono in BM+ population and combo in all-comers address two separate efficacy**
- Regular log-rank test and Cox-regression method applicable to mono vs. SOC in BM+
- Two-step log-rank test and Cox-regression method applicable to combo vs SOC in all comers
  - Analyze BM+ and BM- patients separately, and combine in a weighted sum
  - Weight of BM+ vs BM- strata = 3:2, pre-determined by randomization ratio (no estimation)
  - Minor loss of efficiency with weighted log-rank test is offset by gain of sharing SOC

Sample Size Comparison

	One-Trial Design	**Two-Trial Design**
Sample size for monotherapy vs. SOC	326	326
Sample size for combination vs. SOC	489	472
Shared sample size in control arm	61	0
Total sample size	**754**	**798**
Number of screened patients who cannot be enrolled solely because of being biomarker negative	408	652
Biomarker prevalence = 1/3, HR (OS) mono vs SOC = 0.65, HR (OS) combo vs SOC = 0.70, one-sided alpha = 0.025, power = 90%, 70% randomized patients have events by the time of final analysis.		

Trial cost per patient > $100K

Sample Size Under Different BM+ Prevalence



Additional BM+ patients needed

Sample size driven by all-comers and over-powered for BM+

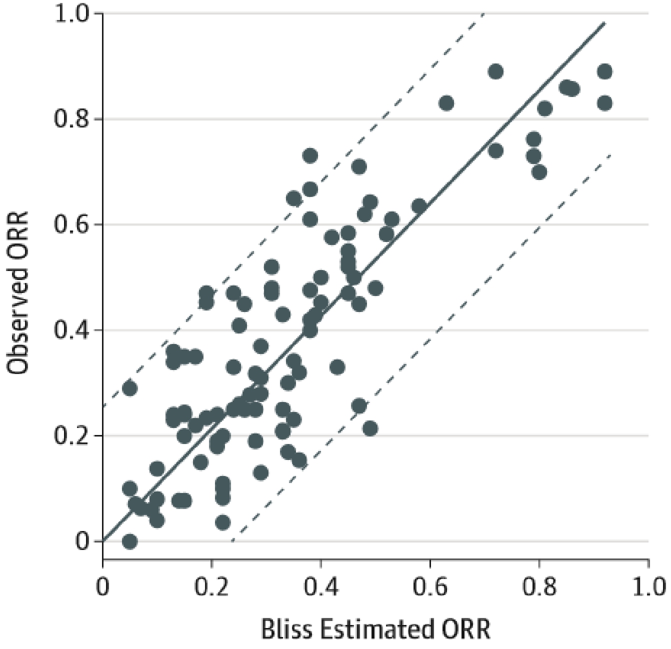
Discussions

- Biomarker hypotheses add a new dimension to the trial design and monitoring
  - The analysis time can be different between all-comers and BM+ patients
  - Prevalence of BM+ events may deviate from initial projection
- The three designs can be further enhanced and modified to meet the practical need
  - Various adaptive features can be added to the single trial design

PREDICTION OF TREATMENT EFFECT OF COMBINATION THERAPIES WITH INDEPENDENT DRUG ACTION MODEL

Synergistic Effect

- Many drugs are brought to clinical testing after a synergistic effect is observed in preclinical tumor models, i.e., combination therapy can kill tumor cells at a faster rate than projected by the additive effect of two constituents of the combination
- However, synergistic effect is rarely seen in clinical trials at population level, and even worse most investigational oncology drugs fail despite encouraging preclinical data
  – Schmidt et al. (2020) showed that ORRs of PD-1 checkpoint inhibitor combinations are consistent with Bliss independence model at population level (i.e., $R=R_1+R_2-R_1*R_2$)



Independent Drug Action at Individual Level

- **Definition**: a patient's response to a combination therapy of two constituents is the best of the two potential responses (i.e., best response = response to either one)
- The two responses may have a (small) positive correlation (*ρ*) due to cross-resistance of the constituents (Gao et al 2015; Palmer and Sorger 2017), which lead to (slightly) lower ORR by ρτ than the Bliss model prediction (i.e., antagonistic effect at population level)
  - When responses are *independent*, predicted ORR same as from Bliss independence model
  - A negative correlation at individual level implies a synergistic effect at population level

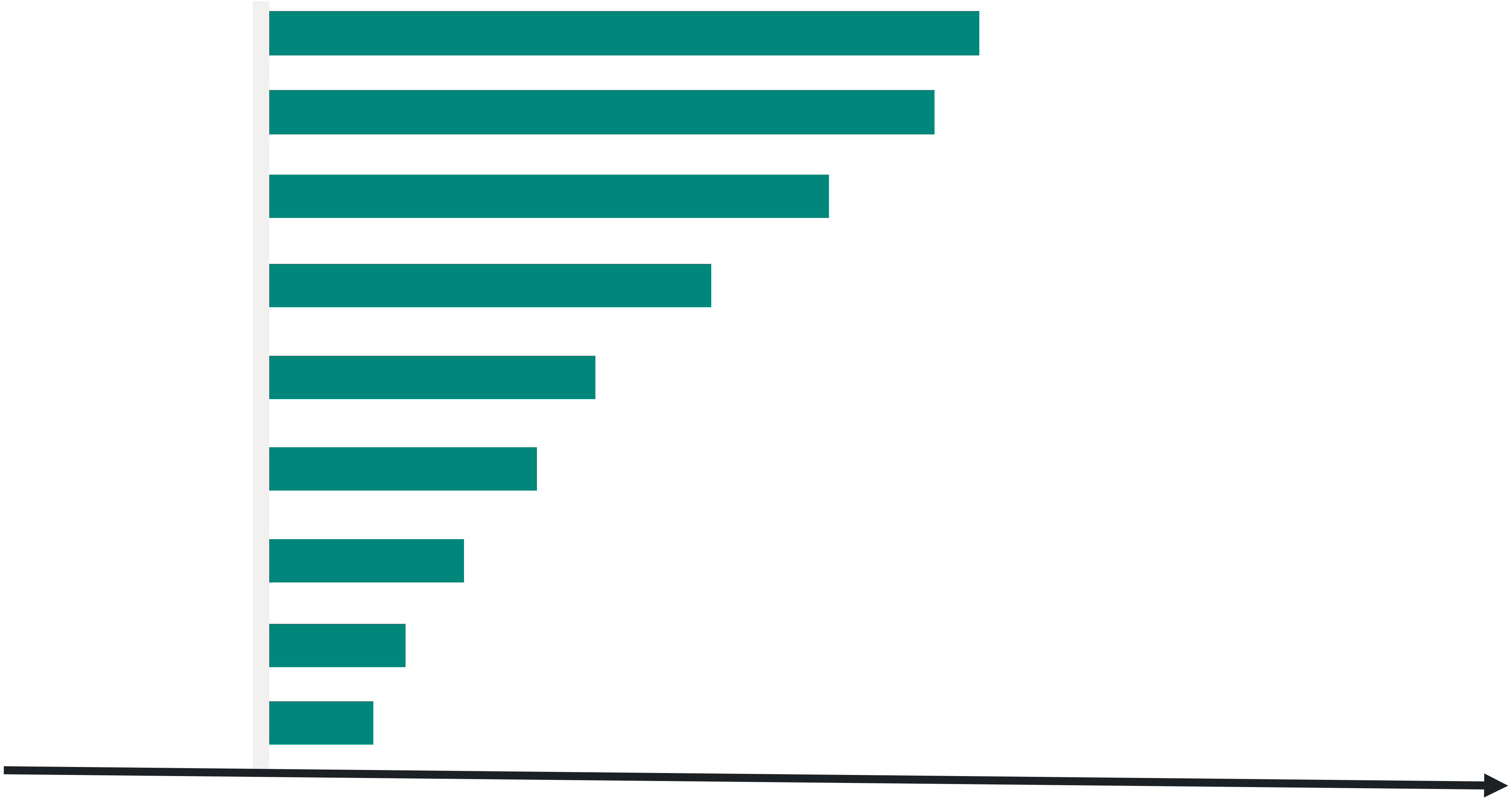
	Response	**No**
Response	$R_1R_2+\rho\tau$	$R_1(1-R_2)-\rho\tau$
No	$R_2(1-R_1)-\rho\tau$	$(1-R_1)(1-R_2)+\rho\tau$

$$\tau = \sqrt{R_1(1-R_1)R_2(1-R_2)}$$

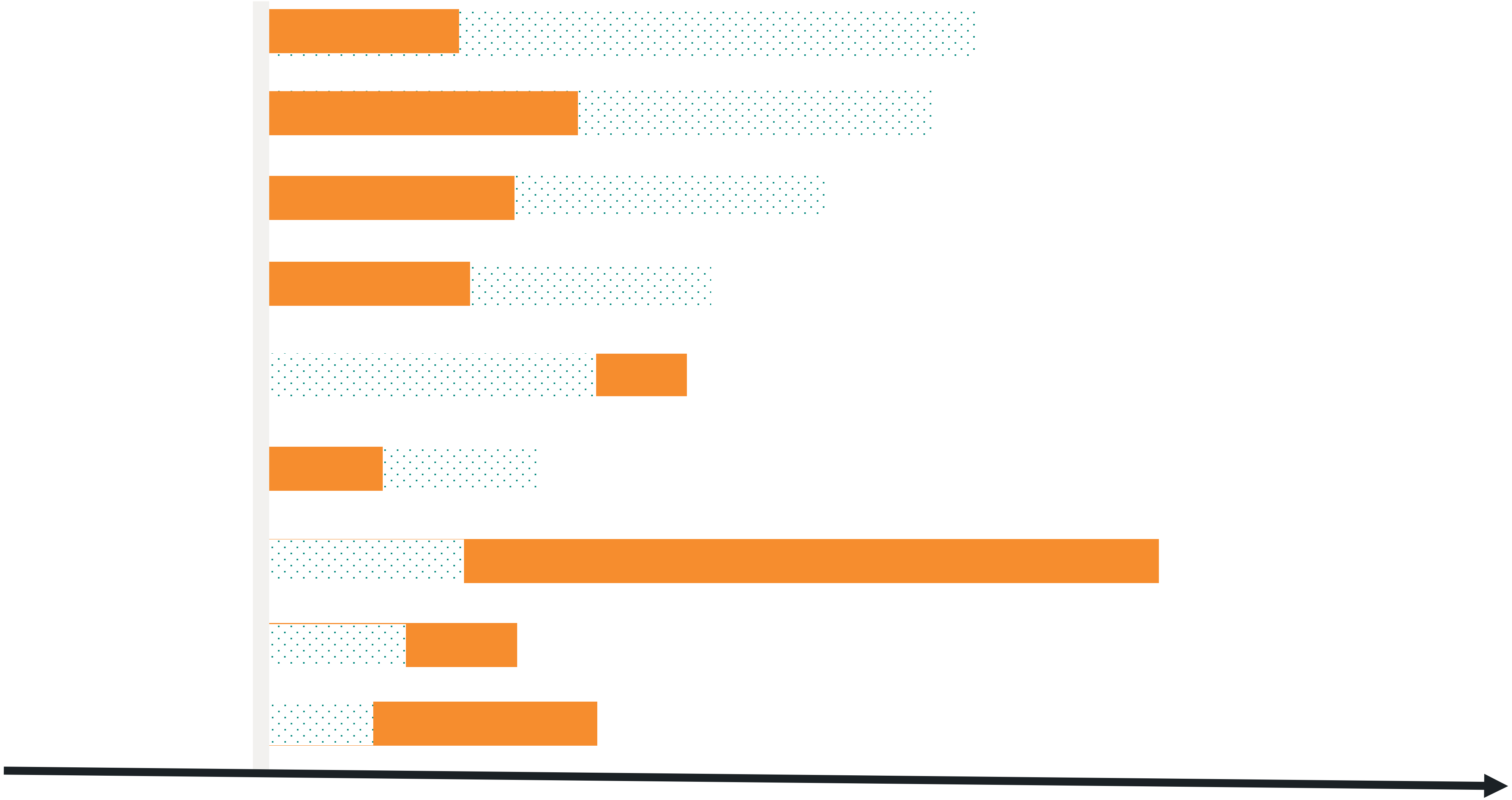
What About PFS?

- Palmer et al. [2020] showed that PFS outcomes for drug combinations with immune checkpoint inhibitors were largely predictable from the independent drug action model
  - Digitally construct survival functions for constituents from published KM curves
  - Draw samples of hypothetical PFS times from each survival function
  - Add noise to the rank-ordered PFS times to achieve intended Spearman's correlation
  - Form pairs of PFS times by the reshuffled rank-order ("responses" to constituents)
  - Find the maximum of each pair (predicted "response" to combination)
  - Generate survival function for the predicted PFS time from predicted "responses"
- Non-parametric and robust in nature, but difficult to derive statistical properties

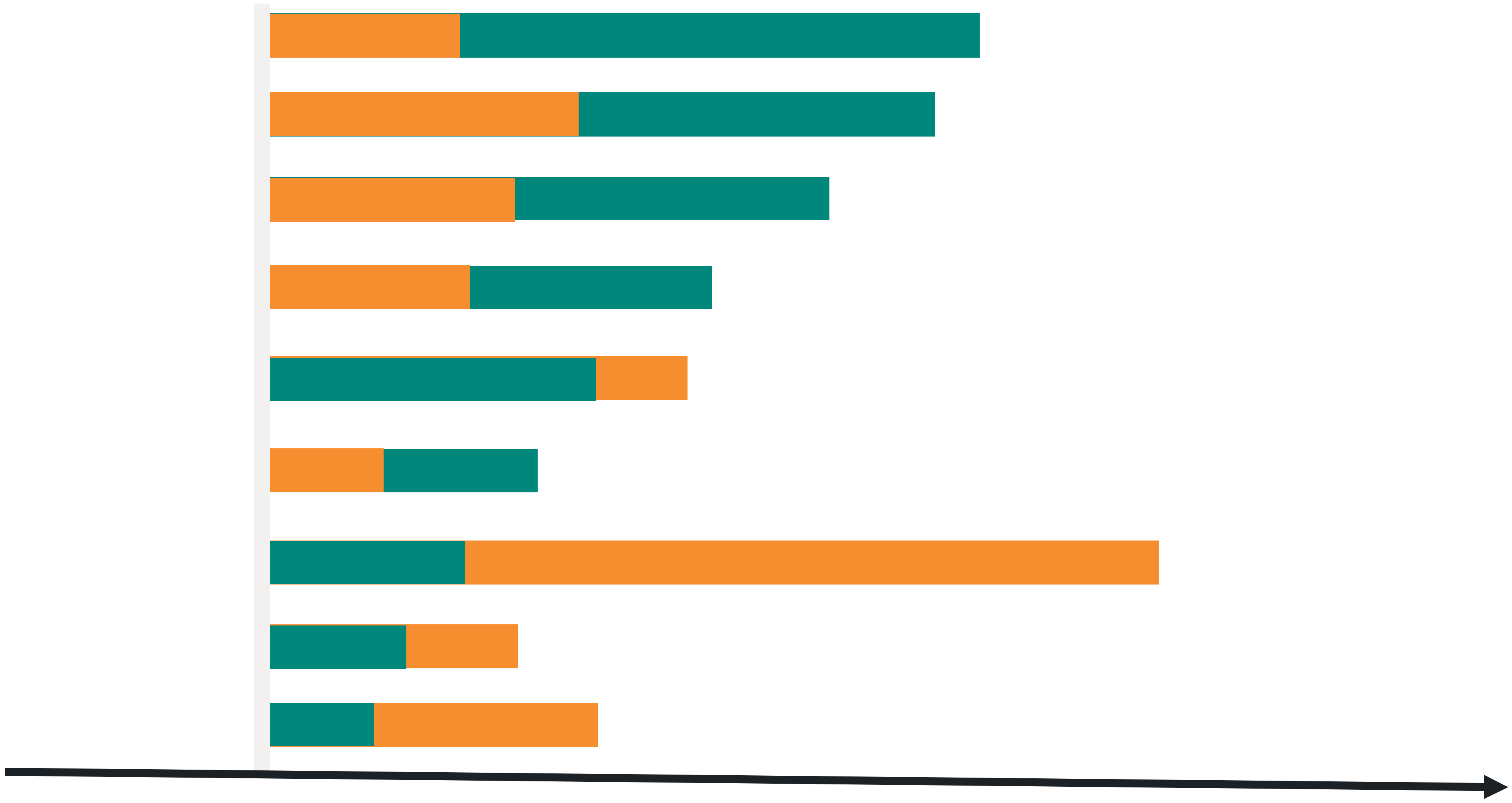
PFS: Drug 1



PFS: Drug 2 (vs Drug 1)



PFS: Drug 1 + Drug 2 = max (Drug 1, Drug 2)



PREDICTION OF PFS EFFECT

Our Proposed Approach

- Apply independent drug action model to the **bivariate indicator variable** $I_{\{T_i > t\}}$ which takes value 1 (“response”) when PFS time for *i*-th constituent $T_i > t$ or 0 otherwise $(i = 1,2)$
  - **Semi-parametric and robust in nature, and easier to derive statistical properties**
  - Same as Palmer’s approach when T_1 and T_2 are independent (i.e., $\varphi(t)$=0), and both degenerate to Bliss independence model with survival rate behaving like ORR

$$S(t) = Pr(max(T_1, T_2) > t) = Pr(I_{\{T_1 > t\}} = 1 \text{ or } I_{\{T_2 > t\}} = 1)$$

$$= 1 - Pr(I_{\{T_1 > t\}} = 0 \text{ and } I_{\{T_2 > t\}} = 0)$$

$$= S_1(t) + S_2(t) - S_1(t)S_2(t) - \varphi(t)\sqrt{S_1(t)(1 - S_1(t))S_2(t)(1 - S_2(t))}$$

Chen et al. Independent Drug Action and Its Statistical Implications for Development of Combination Therapies. *Contemporary Clinical Trials* 2020. https://doi.org/10.1016/j.cct.2020.106126.

On Correlation $\varphi(t)$

- With the inclusion of a time-varying correlation coefficient, our approach can account for any joint parametric distribution for a bivariate TTE variable
- The flexibility of having a time-varying correlation coefficient is especially desirable for predicting the treatment effect of combination immunotherapies as it may evolve over time
- When it is expected to be small, the impact of mis-specification is negligible
  - Accurate estimates of $\varphi(t)$ come from proper meta-analysis of relevant trials (e.g., trials for drugs with the same class of action in the same disease setting), OR deep understanding of MOAs of involved drugs (e.g., non-overlapping MOAs may imply small correlation)

Proposed Estimates of Predicted Survival Functions

- Predicted survival function for the combination therapy (*t* is suppressed)

$$\hat{S} = \hat{S}_1 + \hat{S}_2 - \hat{S}_1\hat{S}_2 - \varphi\sqrt{\hat{S}_1(1-\hat{S}_1)\hat{S}_2(1-\hat{S}_2)}$$

$$Var(\hat{S}) \approx \left[(1-S_2) - \frac{\sqrt{S_2(1-S_2)}}{2\sqrt{S_1(1-S_1)}}(1-2S_1)\varphi\right]^2 \sigma_{S_1}^2 + \left[(1-S_1) - \frac{\sqrt{S_1(1-S_1)}}{2\sqrt{S_2(1-S_2)}}(1-2S_2)\varphi\right]^2 \sigma_{S_2}^2$$

- Predicted survival function for a constituent drug (*t* is suppressed)

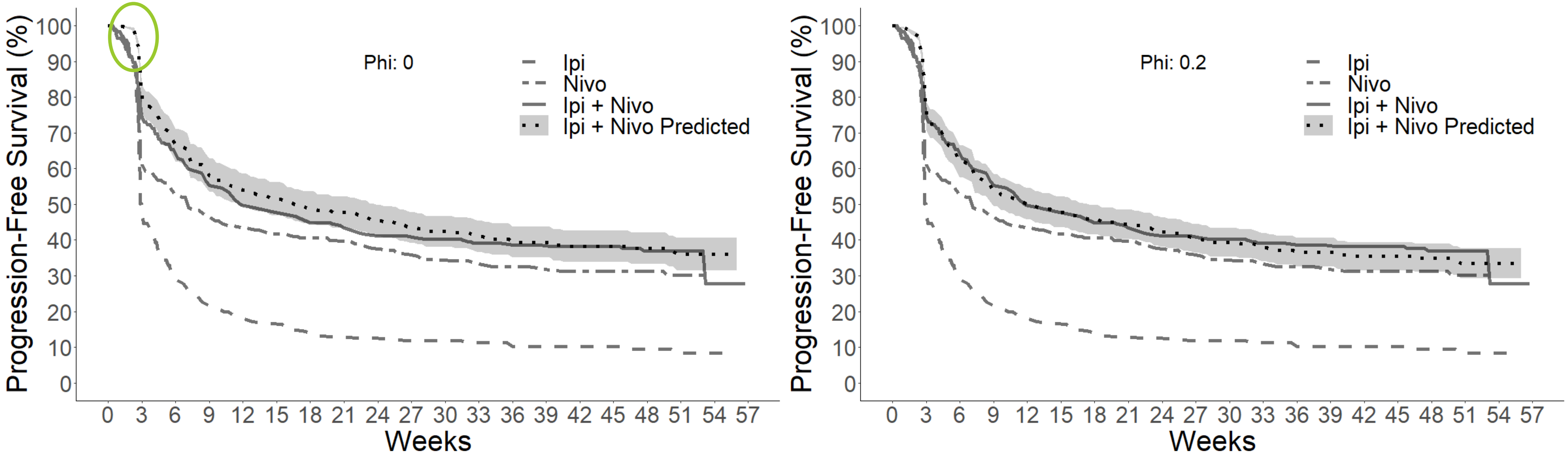
$$\hat{S}_2 = \frac{(\hat{S}-\hat{S}_1)}{(1-\hat{S}_1)} + \varphi\frac{\sqrt{\hat{S}_1(1-\hat{S}_1)(1-\hat{S})(\hat{S}-\hat{S}_1)}}{(1-\hat{S}_1)^2}$$

$$Var(\hat{S}_2) \approx \frac{1}{(1-S_1)^2}\left[1 + \frac{\varphi S_1(1-2S+S_1)}{2\sqrt{A}}\right]^2 \sigma_S^2 + \frac{(1-S)^2}{(1-S_1)^4}\left[1 + \frac{\varphi\left(S_1^2 + 2S_1 - 2SS_1 - S\right)}{2\sqrt{A}}\right]^2 \sigma_{S_1}^2$$

where $A = S_1(1-S_1)(1-S)(S-S_1)$

Prediction of PFS for IPI-Nivo Combo in 1L melanoma

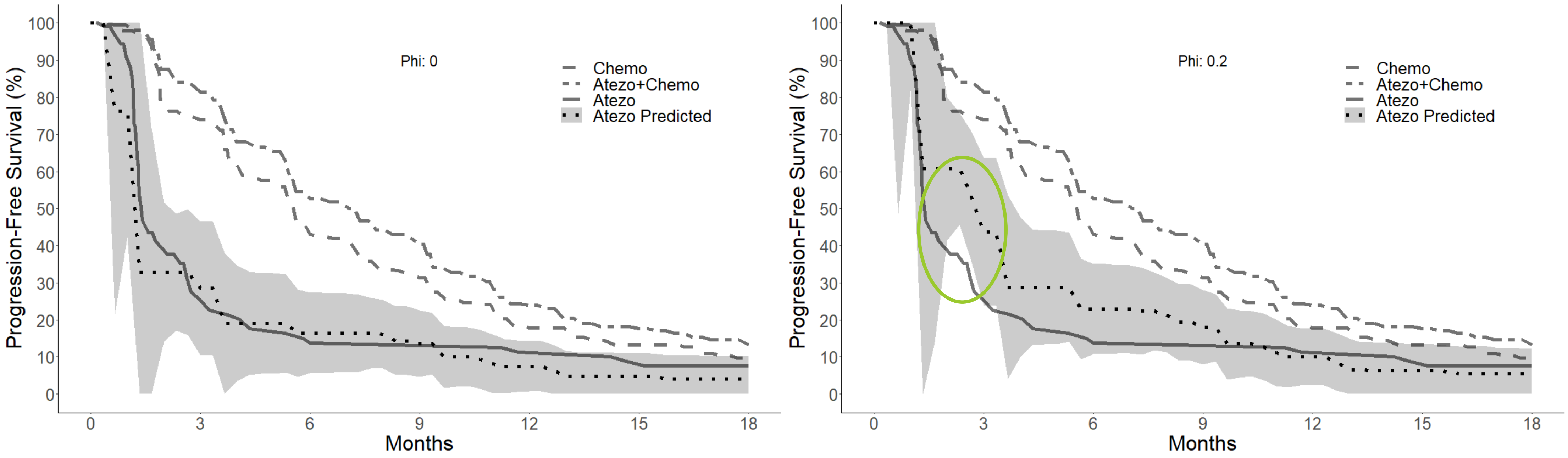
- Estimation of predicted PFS rate and 95%CI for the combo based on Kaplan Meier estimates for the individual drugs in untreated melanoma from CheckMate 067
  - Combo data from Checkmate 067 is used for comparison



left panel: $\varphi(t)$= 0; right panel: $\varphi(t)$= 0.2

Prediction of PFS for Atezolizumab in 1L TNBC

- Estimation of predicted PFS rate and 95% CI for Atezo based on Kaplan Meier estimates for chemo and combo in advanced triple-negative breast cancer from the Impassion 130 study
  - Single arm data from a Phase 1 study is used for comparison



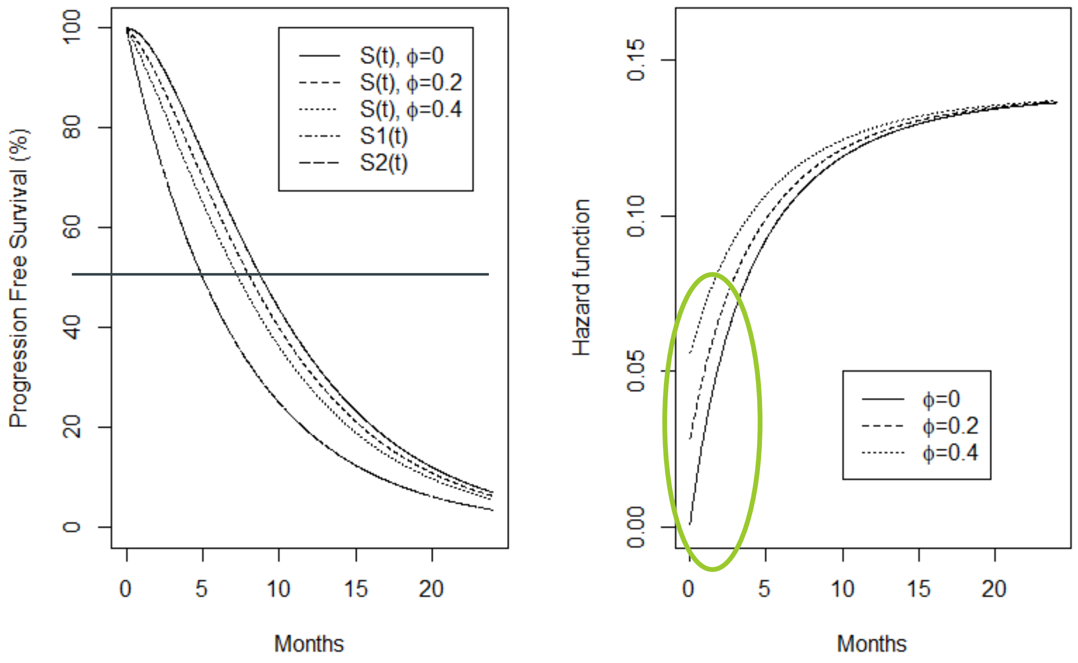
left panel: $\varphi(t) = 0$; right panel: $\varphi(t) = 0.2$

APPLICATION TO DESIGN AND MONITORING OF A HYPOTHETICAL TRIAL

Status Quo of Choosing Δ in Trial Design

- When medians for the two constituents are not available, target hazard ratio is often based on ***clinical interest*** with little statistical justification
- When available, median for combo is predicted to be sum of the two medians and the target hazard ratio is chosen under the exponential distribution assumption
  - However, the sum overestimates the true median and the true hazard ratio is not constant

Predicted survival function and hazard function for combo when PFS for a constituent follows an exponential distribution with median of 5 months



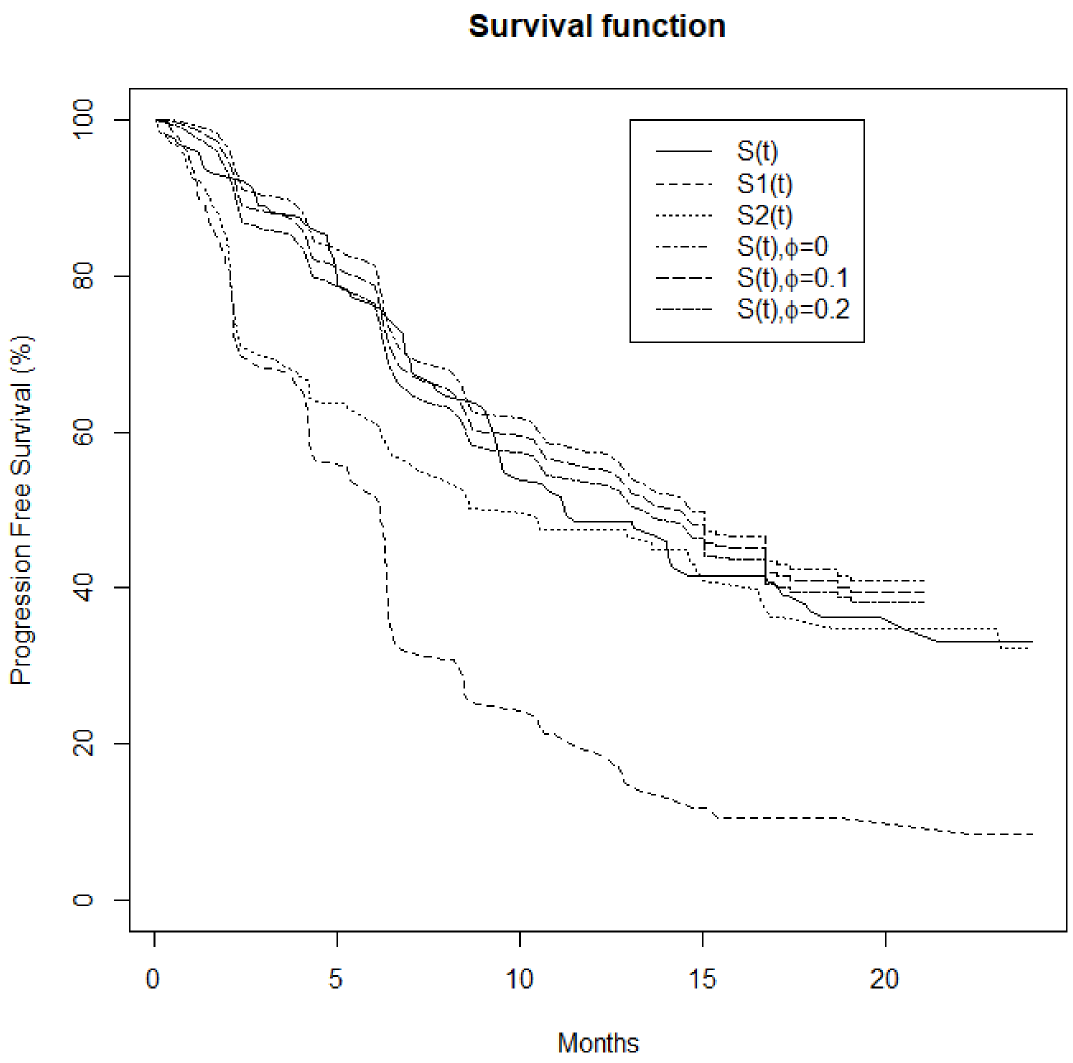
A Hypothetical Trial

- Pembrolizumab+chemo vs chemo (1:1) in 1L NSCLC patients with IHC PD-L1 TPS≥50%
  - **What would the hazard ratio and overall event rate be at an analysis time?**
- Data source for predicting PFS of pembrolizumab+chemo
  - Pembrolizumab and chemo: from KN-024 (pembrolizumab vs chemo in same population)
- Actual PFS data from KN-189 (pembrolizumab+chemo vs chemo in all-comers) in same subpopulation is used for comparison

Can we predict Δ???

Survival Functions for PFS

- Observed survival functions for chemo $S_1(t)$ and pembro $S_2(t)$ in KN-024 are digitally constructed from published trial data
- Predicted survival functions ($S(t)$, φ=0, 0.1 or 0.2) match well with observed survival function for pembro-chemo combination $S(t)$ in KN-189



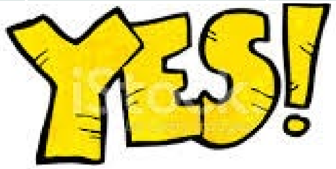
Prediction of Treatment Effect in the Hypothetical Trial

- Generate enrollment time for each patient
  - Patients are assumed to be enrolled in 12 months at constant rate
- Generate event time according to the survival function of each treatment arm and censor at analysis time if event occurs later
  - KM curve for chemo as observed in KN-024
  - Predicted survival function for combo based on the independent drug action model
- Fit generated data to Cox-regression model to calculate the hazard ratio

Performance of Predicted Event Rate and Hazard Ratio

- Observed KM curves in KN-189 are used to estimate the “true” event rate and hazard ratio
  - The actual accrual schedule in KN-189 was different, and partly because of that the estimated outcome may be slightly different from the actual outcome of the study

Analysis time (months)	Est. outcome from KN-189 data		Predicted outcome under independent drug action					
			$\varphi = 0$		φ=0.1		φ=0.2	
	Event rate (%)	Hazard ratio	Event rate (%)	Hazard ratio	Event rate (%)	Hazard ratio	Event rate (%)	Hazard ratio
12 (accrual)	38	0.33	37	0.30	38	0.33	39	0.36
18	64	0.36	61	0.32	62	0.34	63	0.37
24	75	0.37	74	0.35	74	0.37	75	0.39

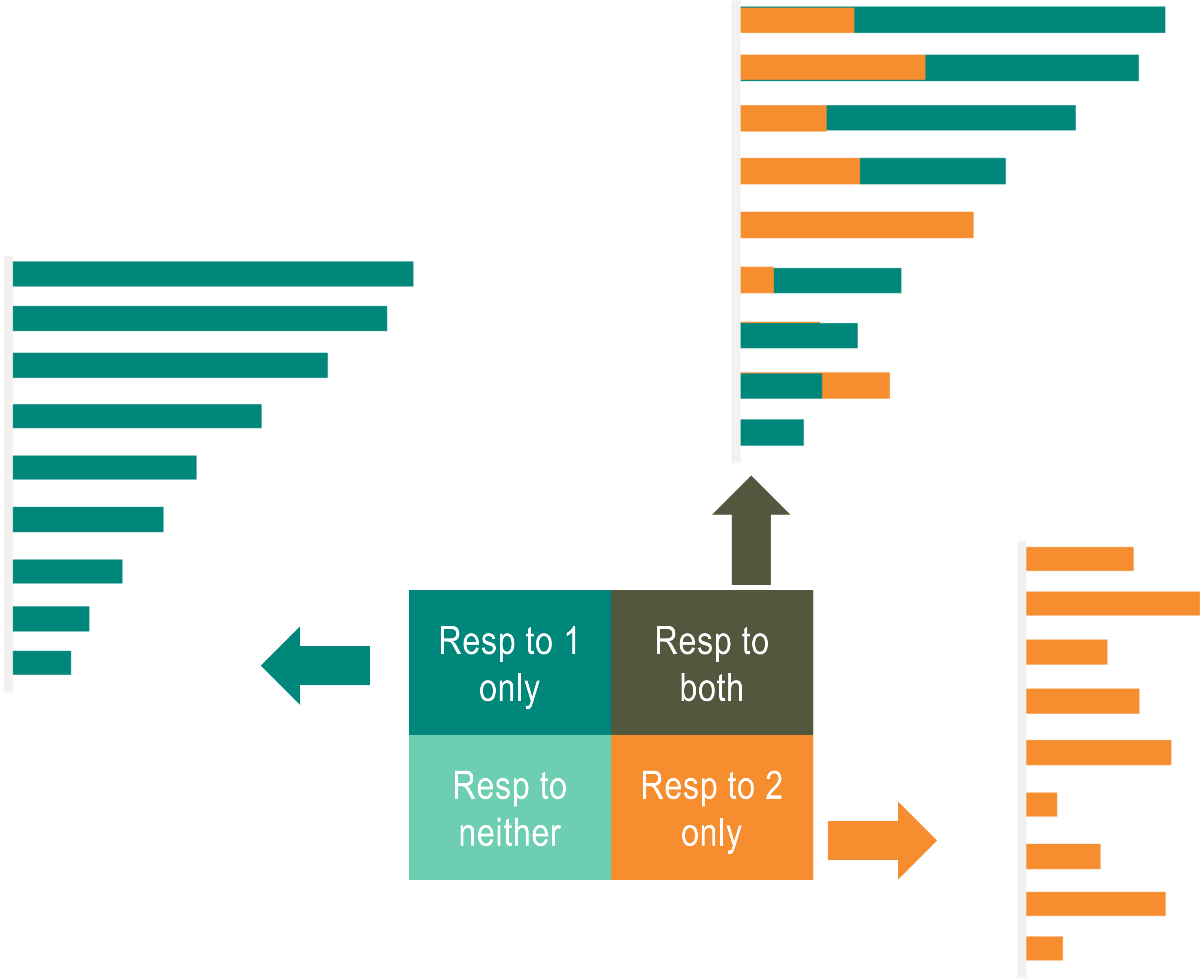


DISCUSSION

Discussion Points

- **All models are wrong but some are useful**
  - The independent drug action model is no exception, but it represents a reasonable working assumption based on empirical evidence in absence of a viable alternative
- Our proposed approach is potentially very helpful for trial design and monitoring
  - The uncertainty of predicted survival function may be incorporated into estimation of PoS (i.e., integrated power over the estimated distribution of hazard ratio)
  - Ongoing work on other endpoints and adjustment of baseline difference across trials
- The independent drug action model can help explain some of the perplexing questions
  - Why response duration for combination therapy may be shorter than for a constituent?
  - Which combination is better (strong+weak or moderate+moderate)?
  - Why is it so difficult to develop effective combination therapies?

Response Duration of Combo (Assuming Drug 1> Drug 2)



Resp to 1 only

Resp to both

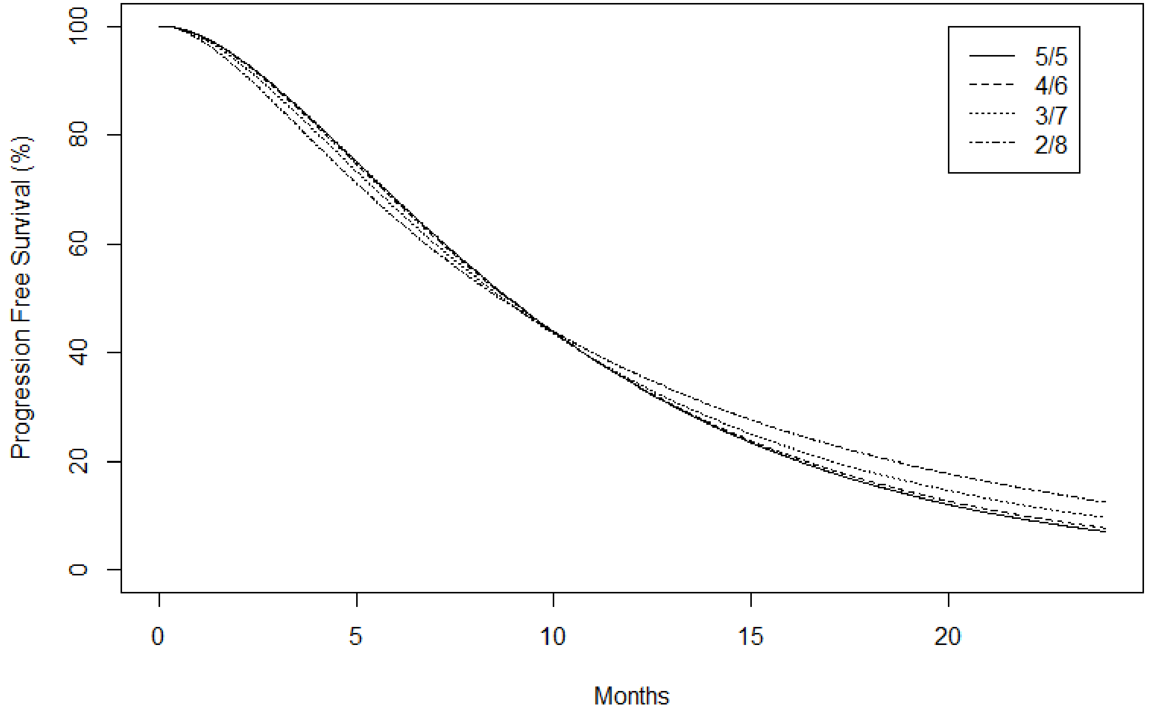
Resp to neither

Resp to 2 only

Depending on % of patients in each cell, taken together, duration for combo can be longer or shorter than for Drug 1, but will be longer than for Drug 2

Comparison among Different Combinations

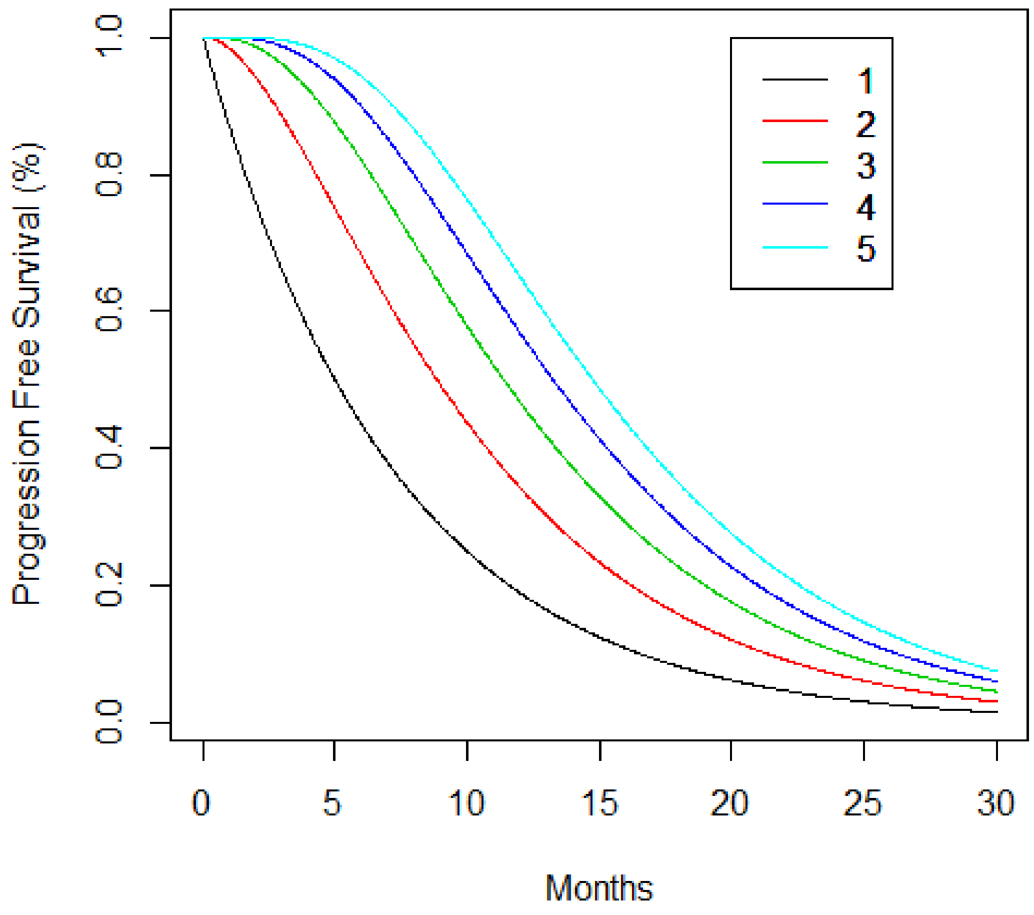
- Median PFS of constituents (months), all under exponential distribution assumption
  - 5+5
  - 4+6
  - 3+7
  - 2+8
- While comparable overall, 5+5 performs best early on and 2+8 performs best later on
  - Implies different optimal α-allocation strategy



$\varphi(t) = 0$

Diminishing Treatment Effect of Combination Therapies

- Median PFS increases with number of constituents but at a slower rate
  - For example, the medians are predicted to be 8.9, 11.4, 13.2, and 14.7 months when it increases from 2 to 5 assuming each has an independent exponential distribution with median of 5 months
- **Deep understanding of MOAs and biomarker guided drug development are key to future success (Palmer and Sorger 2017)**



$\varphi(t) = 0$