Power assessment in group sequential design with multiple biomarker subgroups for multiplicity problem

Lei Yang, Ph.D.
Statistical Scientist, Roche (China) Holding Ltd.
Aug 30th 2018, Shanghai Jiao Tong University

GROUP SEQUENTIAL DESIGN & EFFICACY SUPERIORITY INTERIM ANALYSIS

Group sequential design

- Potentially stop trial after interim analysis for futility or efficacy
- Look into the data multiple times while maintaining integrity

Early stopping for efficacy superiority at interim analysis

- All patients should have timely access to new effective treatment
- Launch new product early
- Early analysis may reveal problems (e.g. compliance, accrual rate)

HIERARCHICAL TESTING

Test primary population at full alpha

If significant, test next population with full alpha recycled and so forth

Fixed testing sequence ranking by clinical relevance or likelihood of success

Strong control of familywise error rate (FWER) due to (1) prospective specification of the testing sequence and (2) no further testing once the sequence breaks

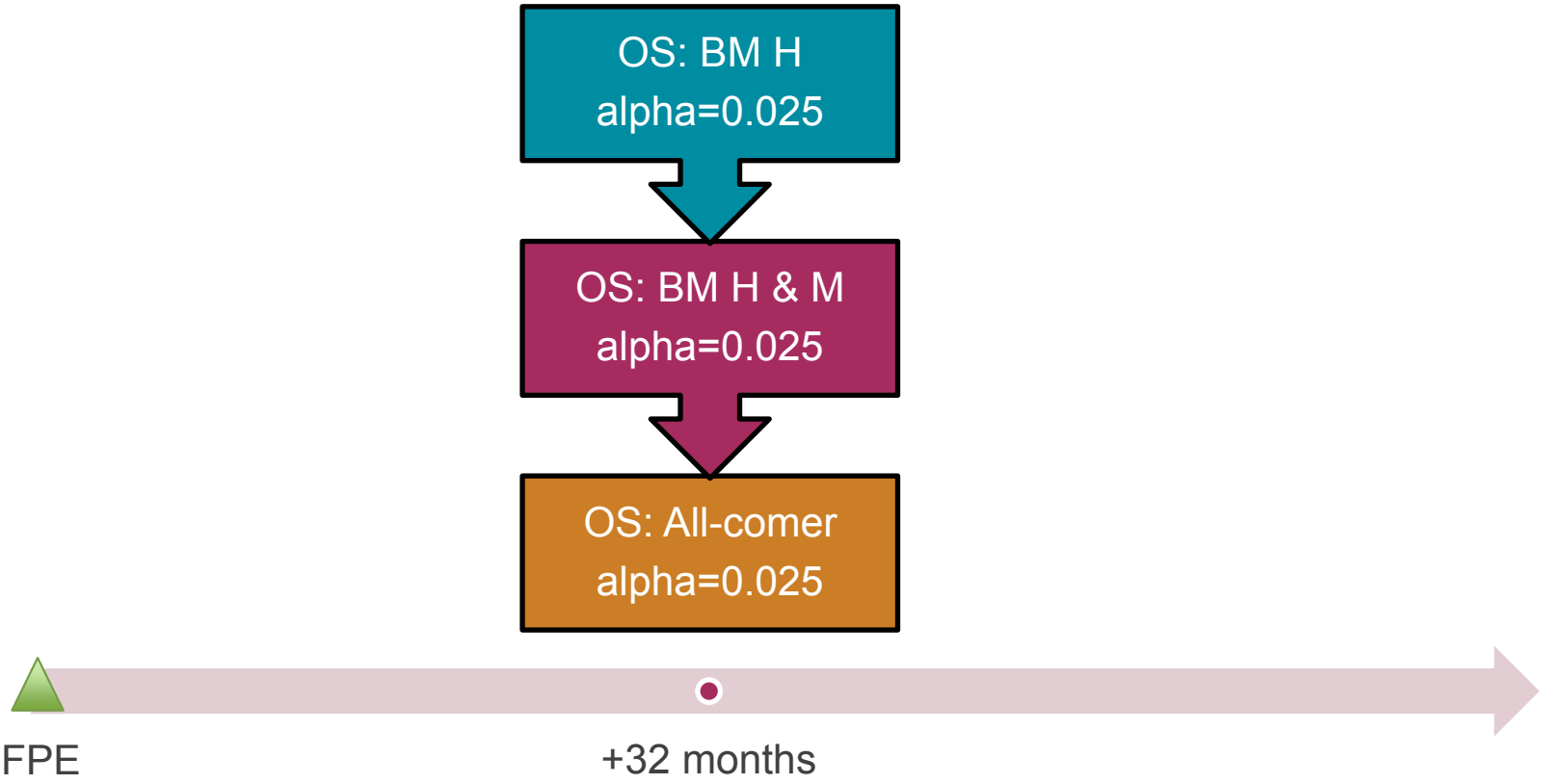
Carefully selecting the ordering of the tests is essential

FDA Guidance for Industry (2017); Hung et al (2007); Glimm et al (2010); Tamhane et al (2010)

Roche SBF/SJTU

STRAIGHTFORWARD TESTING WITH ONE PRIMARY ANALYSIS ONLY

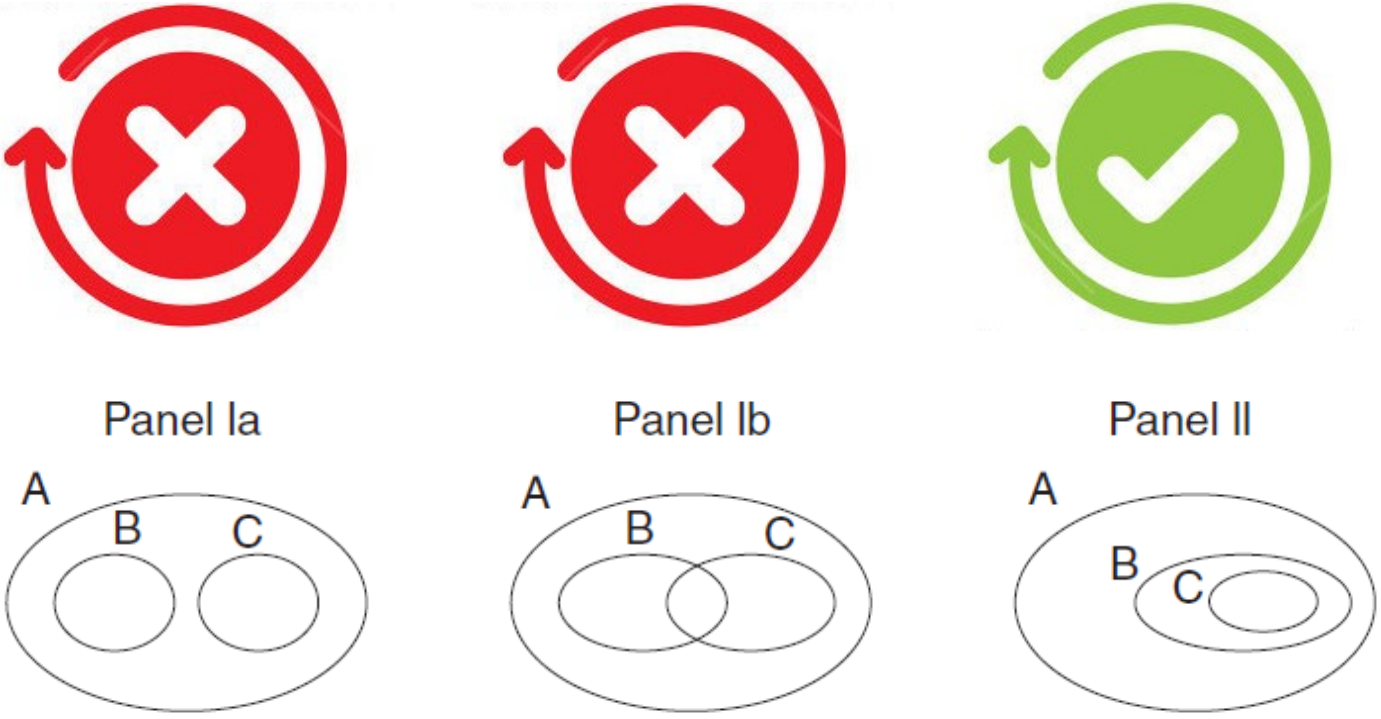
Primary Analysis



FPE: First patient enrolled

Roche SBF/SJTU

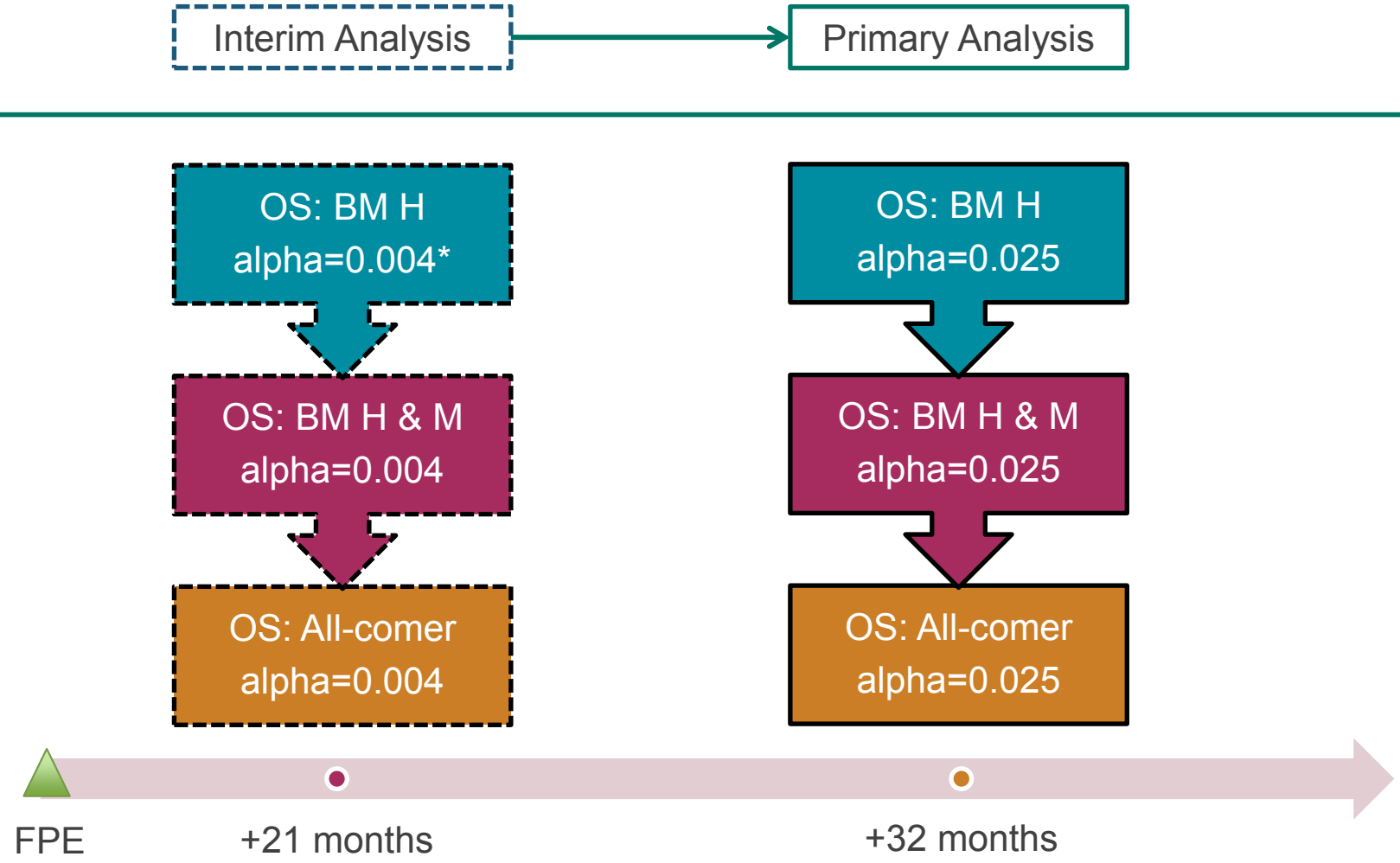
NESTED SUBGROUP ANALYSES WITHOUT IA



Possible scenarios with 2 BM subgroups: Panel I: non-nested subgroup analyses where subgroups can be disjoint (Panel Ia) or can intersect (Panel Ib); Panel II: nested subgroup analyses

Spiessens and Debois (2010)

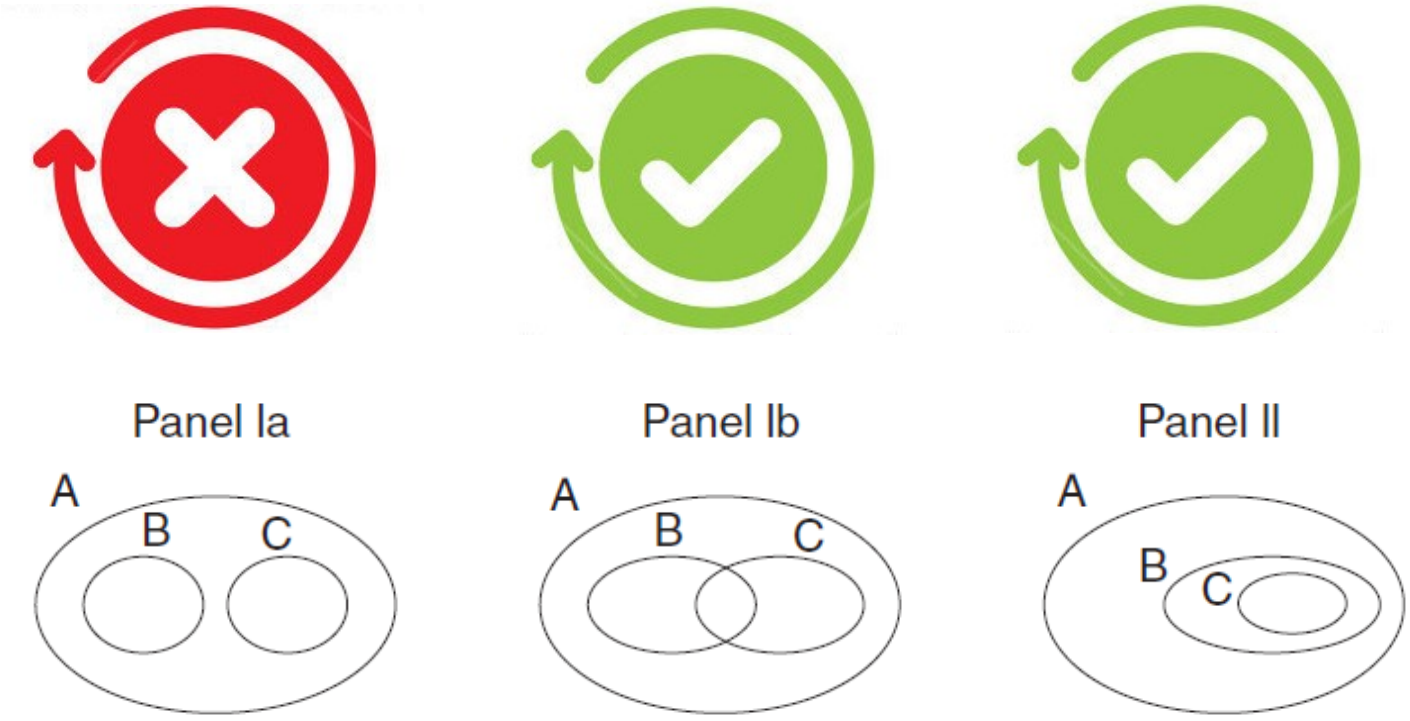
INTERIM ANALYSIS INTRODUCES MORE COMPLEXITY



**O'Brien-Fleming boundary used at 60% information fraction*
FPE: First patient enrolled

Roche SBF/SJTU

NESTED AND INTERSECTANT SUBGROUP ANALYSES WITH IA

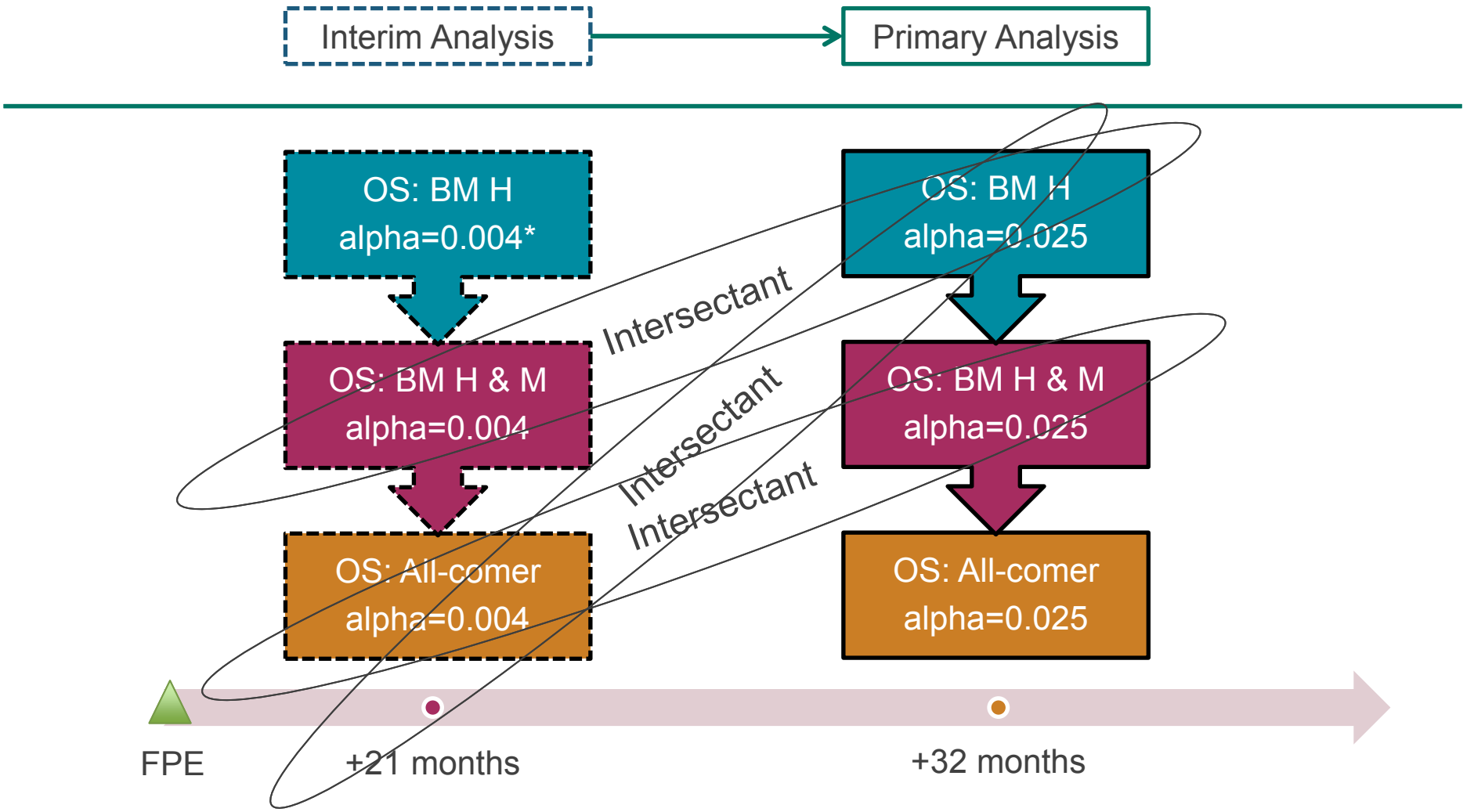


Possible scenarios with 2 BM subgroups: Panel I: non-nested subgroup analyses where subgroups can be disjoint (Panel Ia) or can intersect (Panel Ib); Panel II: nested subgroup analyses

Spiessens and Debois (2010)

Roche SBF/SJTU

A MIXTURE OF INTERSECTANT AND NESTED SUBGROUPS



**O'Brien-Fleming boundary used at 60% information fraction*
FPE: First patient enrolled

AN ARTIFICIAL PHASE III TIME-TO-EVENT CASE STUDY

- Drug A vs SOC by randomization ratio 1:1
- Population: All-comer & two subgroups defined by predefined cut-offs of CDx assays
  - All-comer (Group A) ⊃ BM high and medium (Group B) ⊃ BM high (Group C)
  - Can be extended to more biomarker subgroups
  - Prevalence rates are 75% for Group B and 50% for Group C
- Primary efficacy endpoint: OS
- HR assumptions (median OS of drug A vs SOC [months])
  - Biomarker high: 0.70 (14.3 vs 10)
  - Biomarker high and medium: 0.75 (13.3 vs 10)
  - All-comer: 0.80 (12.5 vs 10)
- Constant recruitment rate (40 pts/month for all-comer) and no drop-out assumed
- One-sided test with alpha = 0.025

WHEN CORRELATIONS AMONG BIOMARKER SUBGROUPS NOT CONSIDERED

		HR (median OS [months])	Sample size	Marginal power (IA)	Events (IA)	Event patient ratio, EPR (IA)
Without IA	BM H (Group C)	0.70 (10 vs 14.3)	451	87.6%	305	68%
	BM H&M (Group B)	0.75 (10 vs 13.3)	676	87.3%	465	69%
	All-comer (Group A)	0.80 (10 vs 12.5)	902	80.0%	631	70%
With IA	BM H (Group C)	0.70 (10 vs 14.3)	451	87.6% (40.8%)	307 (184)	68% (41%)
	BM H&M (Group B)	0.75 (10 vs 13.3)	676	87.3% (40.4%)	469 (281)	69% (42%)
	All-comer (Group A)	0.80 (10 vs 12.5)	902	80.0% (31.8%)	636 (382)	71% (42%)

Key message

- Sample size determined to give 80% power for all-comer with EPR=70% without considering biomarker subgroups
- If so, both biomarker subgroups are over-powered when assessed respectively
- Target number of events slightly larger with IA

Question

- How about if correlations are considered too?

CANONICAL JOINT DISTRIBUTION WITHOUT IA

- Distribution of OS effect estimates

$$\hat{\theta}_A = \log\left(\widehat{HR}_A\right) \sim N\left(\theta_A, var\left(\hat{\theta}_A\right)\right), \mathcal{I}_A = \left(var\left(\hat{\theta}_A\right)\right)^{-1} = N_A/4$$
$$\hat{\theta}_B = \log\left(\widehat{HR}_B\right) \sim N\left(\theta_B, var\left(\hat{\theta}_B\right)\right), \mathcal{I}_B = \left(var\left(\hat{\theta}_B\right)\right)^{-1} = N_B/4$$
$$\hat{\theta}_C = \log\left(\widehat{HR}_C\right) \sim N\left(\theta_C, var\left(\hat{\theta}_C\right)\right), \mathcal{I}_C = \left(var\left(\hat{\theta}_C\right)\right)^{-1} = N_C/4$$

  with $\mathcal{I}_A, \mathcal{I}_B, \mathcal{I}_C$ are information levels in Group A, B and C and N_A, N_B, N_C are numbers of events respectively

- Under null hypotheses $H_{0,C}: \theta_C = 0; H_{0,B}: \theta_B = 0;\ H_{0,A}: \theta_A = 0$, logrank test statistics $(Z_C = \hat{\theta}_C\sqrt{\mathcal{I}_C}, Z_B = \hat{\theta}_B\sqrt{\mathcal{I}_B}, Z_A = \hat{\theta}_A\sqrt{\mathcal{I}_A})$ has approximately canonical joint distribution

$$\begin{pmatrix} Z_C \\ Z_B \\ Z_A \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_C\sqrt{\mathcal{I}_C} \\ \theta_B\sqrt{\mathcal{I}_B} \\ \theta_A\sqrt{\mathcal{I}_A} \end{pmatrix}, \begin{pmatrix} 1 & \sqrt{\tau_{CB}} & \sqrt{\tau_{CA}} \\ \sqrt{\tau_{CB}} & 1 & \sqrt{\tau_{BA}} \\ \sqrt{\tau_{CA}} & \sqrt{\tau_{BA}} & 1 \end{pmatrix}\right)$$

  where $\tau_{CB} = \mathcal{I}_C/\mathcal{I}_B, \tau_{CA} = \mathcal{I}_C/\mathcal{I}_A,\ \tau_{BA} = \mathcal{I}_B/\mathcal{I}_A$ are information fractions

- Define score statistics $(S_C, S_B, S_A) = (Z_C\sqrt{\mathcal{I}_{1C}}, Z_B\sqrt{\mathcal{I}_{1B}}, Z_A\sqrt{\mathcal{I}_{1A}})$, then $(S_C, S_B - S_C, S_A - S_B)$ are independent
- For the case study without IA, $\tau_{CB} = 0.64, \tau_{CA} = 0.46,\ \tau_{BA} = 0.72$

Spiessens and Debois (2010); Jennison and Turnbull (1997, 2000)

OUTCOMES WHEN WITHOUT IA

	PA			Outcome of label claiming	Probability of outcome	Accumulated probability	Hypothesis testing
	BM H	BM H & M	All-comer				
1	+	+	+	All-comer at PA	74.46%	74.5% (Marginal power for All-comer: **80.0%**)	Reject $H_{0,C}, H_{0,B}, H_{0,A}$ under alternative hypotheses at PA
2	+	+	-	BM H&M at PA	7.95%	82.4% (Marginal power for BM H&M: **87.3%**)	Reject $H_{0,C}, H_{0,B}$ under alternative hypotheses at PA
3	+	-		BM H at PA	5.18%	87.6% (Marginal power for BM H: **87.6%**)	Reject $H_{0,C}$ under alternative hypotheses at PA
4	-			No	12.41%	100.0%	-

Key message

- Powers are lower when taking correlations into consideration (-5.5% for all-comer and -4.9% for BM H&M), when there is no IA
- Difference might be larger if assumptions changed (HR, prevalence rates, etc)

CANONICAL JOINT DISTRIBUTION WITH IA

- Distribution of OS effect estimates

$$\hat{\theta}_{kA} = \log\left(\widehat{HR}_{kA}\right) \sim N\left(\theta_A, var(\hat{\theta}_{kA})\right), \mathcal{I}_{kA} = \left(var(\hat{\theta}_{kA})\right)^{-1} = N_{kA}/4$$
$$\hat{\theta}_{kB} = \log\left(\widehat{HR}_{kB}\right) \sim N\left(\theta_B, var(\hat{\theta}_{kB})\right), \mathcal{I}_{kB} = \left(var(\hat{\theta}_{kB})\right)^{-1} = N_{kB}/4$$
$$\hat{\theta}_{kC} = \log\left(\widehat{HR}_{kC}\right) \sim N\left(\theta_C, var(\hat{\theta}_{kC})\right), \mathcal{I}_{kC} = \left(var(\hat{\theta}_{kC})\right)^{-1} = N_{kC}/4$$

  with $\mathcal{I}_{kA}, \mathcal{I}_{kB}, \mathcal{I}_{kC}$ are information levels in Group A, B and C and N_{kA}, N_{kB}, N_{kC} are numbers of events respectively at kth stage, where $k \in \{1, 2\}$ with 1 for IA, 2 for PA

- Under null hypotheses $H_{0,kC}: \theta_C = 0; H_{0,kB}: \theta_B = 0; H_{0,kA}: \theta_A = 0$, logrank test statistics ($Z_C = \hat{\theta}_{1C}\sqrt{\mathcal{I}_{1C}}, Z_B = \hat{\theta}_{1B}\sqrt{\mathcal{I}_{1B}}, Z_A = \hat{\theta}_{1A}\sqrt{\mathcal{I}_{1A}}, Z_{C'} = \hat{\theta}_{2C}\sqrt{\mathcal{I}_{2C}}, Z_{B'} = \hat{\theta}_{2B}\sqrt{\mathcal{I}_{2B}}, Z_{A'} = \hat{\theta}_{2A}\sqrt{\mathcal{I}_{2A}}$) has approximately canonical joint distribution

$$\begin{pmatrix} Z_C \\ Z_B \\ Z_A \\ Z_{C'} \\ Z_{B'} \\ Z_{A'} \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_C\sqrt{\mathcal{I}_{1C}} \\ \theta_B\sqrt{\mathcal{I}_{1B}} \\ \theta_A\sqrt{\mathcal{I}_{1A}} \\ \theta_C\sqrt{\mathcal{I}_{2C}} \\ \theta_B\sqrt{\mathcal{I}_{2B}} \\ \theta_A\sqrt{\mathcal{I}_{2A}} \end{pmatrix}, \begin{pmatrix} 1 & \sqrt{\tau_{CB}} & \sqrt{\tau_{CA}} & \sqrt{\tau_{CC'}} & \sqrt{\tau_{CB'}} & \sqrt{\tau_{CA'}} \\ \sqrt{\tau_{CB}} & 1 & \sqrt{\tau_{BA}} & \sqrt{\tau_{BC'}} & \sqrt{\tau_{BB'}} & \sqrt{\tau_{BA'}} \\ \sqrt{\tau_{CA}} & \sqrt{\tau_{BA}} & 1 & \sqrt{\tau_{AC'}} & \sqrt{\tau_{AB'}} & \sqrt{\tau_{AA'}} \\ \sqrt{\tau_{CC'}} & \sqrt{\tau_{BC'}} & \sqrt{\tau_{AC'}} & 1 & \sqrt{\tau_{C'B'}} & \sqrt{\tau_{C'A'}} \\ \sqrt{\tau_{CB'}} & \sqrt{\tau_{BB'}} & \sqrt{\tau_{AB'}} & \sqrt{\tau_{C'B'}} & 1 & \sqrt{\tau_{B'A'}} \\ \sqrt{\tau_{CA'}} & \sqrt{\tau_{BA'}} & \sqrt{\tau_{AA'}} & \sqrt{\tau_{C'A'}} & \sqrt{\tau_{B'A'}} & 1 \end{pmatrix} \right)$$

  where

$$\tau_{CB} = \mathcal{I}_{1C}/\mathcal{I}_{1B}, \tau_{CA} = \mathcal{I}_{1C}/\mathcal{I}_{1A}, \tau_{CC'} = \mathcal{I}_{1C}/\mathcal{I}_{2C}, \tau_{CB'} = \mathcal{I}_{1C}/\mathcal{I}_{2B}, \tau_{CA'} = \mathcal{I}_{1C}/\mathcal{I}_{2A},$$
$$\tau_{BA} = \mathcal{I}_{1B}/\mathcal{I}_{1A}, \tau_{BC'} = (\mathcal{I}_{1C}/\mathcal{I}_{1B}) * (\mathcal{I}_{1C}/\mathcal{I}_{2C}), \tau_{BB'} = \mathcal{I}_{1B}/\mathcal{I}_{2B}, \tau_{BA'} = \mathcal{I}_{1B}/\mathcal{I}_{2A},$$
$$\tau_{AC'} = (\mathcal{I}_{1C}/\mathcal{I}_{1A}) * (\mathcal{I}_{1C}/\mathcal{I}_{2C}), \tau_{AB'} = (\mathcal{I}_{1B}/\mathcal{I}_{1A}) * (\mathcal{I}_{1B}/\mathcal{I}_{2B}), \tau_{AA'} = \mathcal{I}_{1A}/\mathcal{I}_{2A},$$
$$\tau_{C'B'} = \mathcal{I}_{2C}/\mathcal{I}_{2B}, \tau_{C'A'} = \mathcal{I}_{2C}/\mathcal{I}_{2A},$$
$$\tau_{B'A'} = \mathcal{I}_{2B}/\mathcal{I}_{2A},$$

  are information fractions

- Define score statistics $(S_C, S_B, S_A, S_{C'}, S_{B'}, S_{A'}) = (Z_C\sqrt{\mathcal{I}_{1C}}, Z_B\sqrt{\mathcal{I}_{1B}}, Z_A\sqrt{\mathcal{I}_{1A}}, Z_{C'}\sqrt{\mathcal{I}_{2C}}, Z_{B'}\sqrt{\mathcal{I}_{2B}}, Z_{A'}\sqrt{\mathcal{I}_{2A}})$, then $(S_C, S_B - S_C, S_A - S_B, S_{C'} - S_A, S_{B'} - S_{C'}, S_{A'} - S_{B'})$ are not independent anymore
- For the case study with IA, $\tau_{CB} = 0.64, \tau_{CA} = 0.46, \tau_{CC'} = 0.60, \tau_{CB'} = 0.38, \tau_{CA'} = 0.28,$
  $\tau_{BA} = 0.72, \tau_{BC'} = 0.38, \tau_{BB'} = 0.60, \tau_{BA'} = 0.43, \tau_{AC'} = 0.28, \tau_{AB'} = 0.43, \tau_{AA'} = 0.60, \tau_{C'B'} = 0.64, \tau_{C'A'} = 0.46, \tau_{B'A'} = 0.72$

Spiessens and Debois (2010); Jennison and Turnbull (1997, 2000)

A PRESPECIFIED TESTING RULE WITH IA

Overall hierarchical rule*

Start from testing BM H in IA

If significant, go to next subgroup in IA, and stop when All-comer in IA is significant

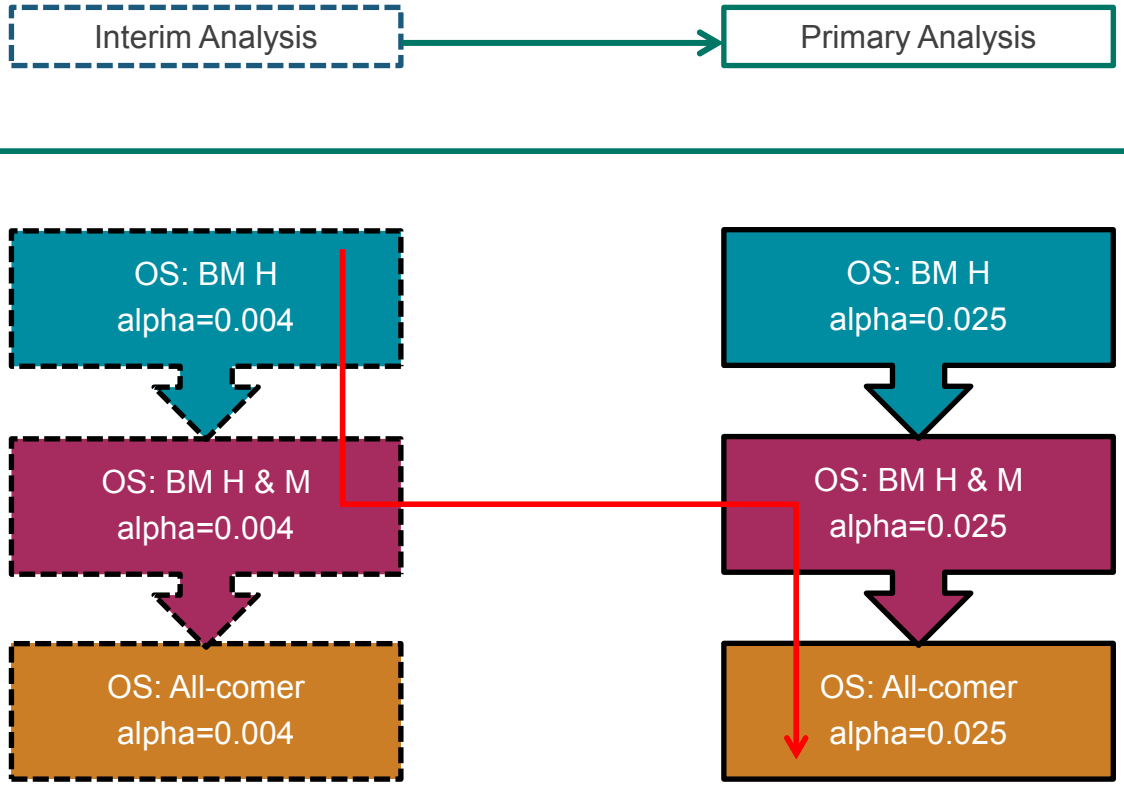
If not significant for any test in IA, go to the same subgroup in PA; then stop if not significant; otherwise move forward until All-comer in PA is significant

Key message

- The so-called 'overall hierarchical rule' achieves strong control of FWER in the case of with an IA

**Glimm et al (2010)*

SIMPLIFIED TESTING STRATEGY



Key message

- Dimensions reduced (n+1<2n*); nested subgroups only
- Score statistics with independent increment structure

**n is the number of tests in either IA or PA with always n>=2. In the artificial case study, n=3.*

OUTCOMES WHEN WITH IA

	IA			PA			Outcome of label claiming	Probability of outcome	Accumulated probability
	BM H	BM H & M	All-comer	BM H	BM H & M	All-comer			Power of IA
1	+	+	+				All-comer at IA	22.39%	22.4% (Marginal power for All-comer: **31.8%**)
2	+	+	-			+	BM H&M at IA, All-comer at PA	8.52%	30.9%
3	+	+	-			-	BM H&M at IA	0.02%	30.9% (Marginal power for BM H&M: **40.4%**)
4	+	-			+	+	BM H at IA, All-comer at PA	9.75%	40.7%
5	+	-			+	-	BM H at IA, BM H&M at PA	0.10%	40.8%
6	+	-			-		BM H at IA	0.02%	40.8% (Marginal power for BM H: **40.8%**)

Key message

- Powers are lower when taking correlations into consideration with IA (-9.4% for all-comer and -9.5% for BM H&H respectively at IA)
- Difference might be larger if assumptions changed

NON-PROPORTIONAL HAZARD RATIO – MARGINAL POWER

		Sample size	Event patient ratio, EPR (IA)	Events (IA)	PH: Marginal power (IA)	NPH(assume 3M efficacy effect delay): Marginal power (IA)
Without IA	BM H (Group C)	451	68%	305	87.6%	62.7%
	BM H&M (Group B)	676	69%	465	87.3%	62.2%
	All-comer (Group A)	902	70%	631	80.0%	56.8%
With IA	BM H (Group C)	451	68% (41%)	307 (184)	87.6% (40.8%)	62.7% (13.9%)
	BM H&M (Group B)	676	69% (42%)	469 (281)	87.3% (40.4%)	62.2% (12.1%)
	All-comer (Group A)	902	71% (42%)	636 (382)	80.0% (31.8%)	56.8% (8.1%)

Key message

- Power decrease are 23%~25% for PA when assuming 3 months efficacy effect delay while larger for IA (27%~49%)
- Decrease depends on specific assumption of treatment effect delay

NON-PROPORTIONAL HAZARD RATIO – A TENTATIVE STRATEGY

- How about if correlations are considered too?
- Unfortunately NO theory established yet when assuming non-proportional HR!

A tentative strategy	• **To approximate the NPH model with a PH model by re-estimating a rough proportional HR* from NPH simulation**
Pros	• Very straightforward for audience to understand impacts of NPH • Make cross-table comparison possible
Cons	• Lack of solid theory support and good accuracy • Need further investigation on the HR re-estimation

** e.g., with a relatively stable proportional HR around the timing of PA and beyond*

NON-PROPORTIONAL HAZARD RATIO – POWER FINDINGS

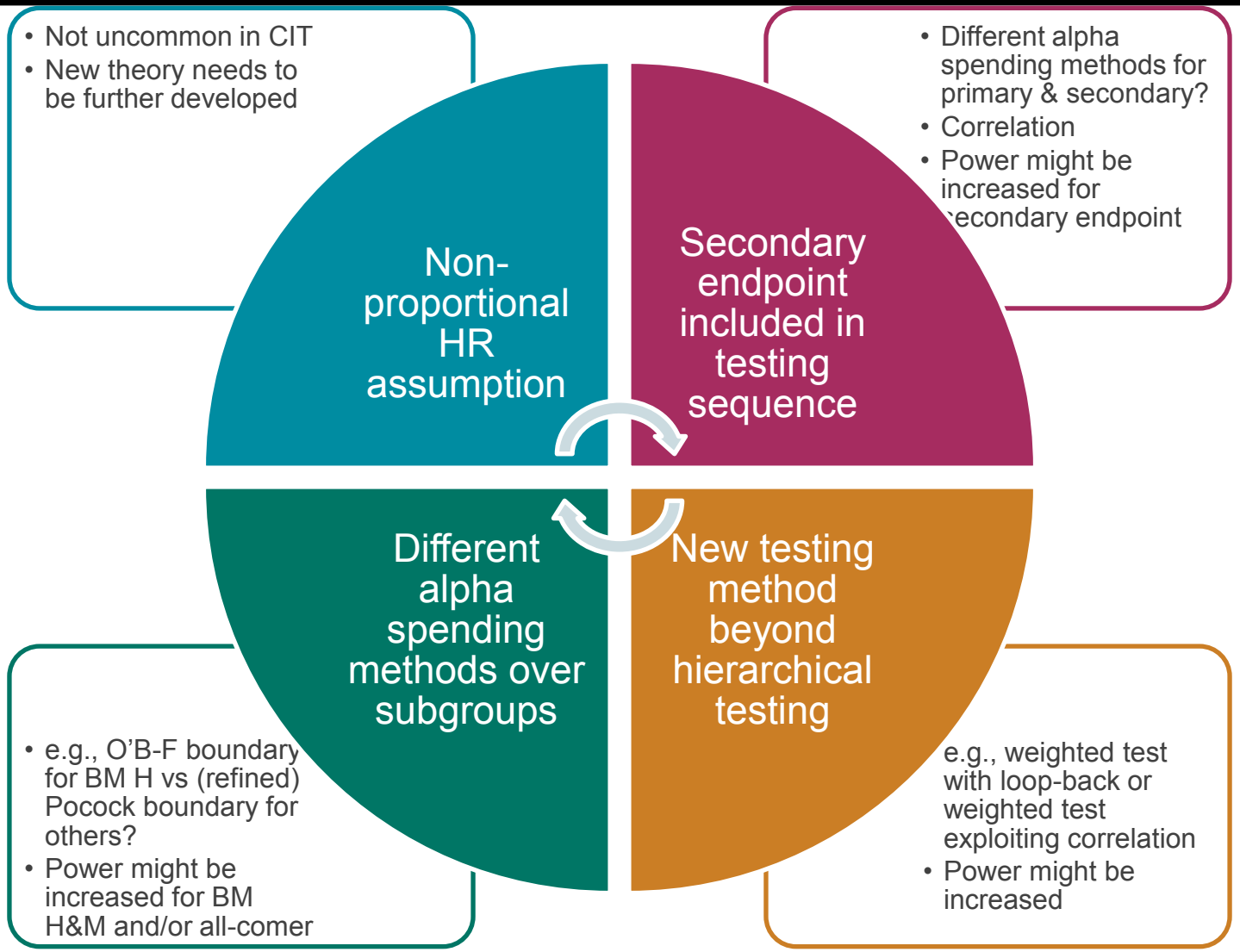
	PA/IA			Outcome of label claiming	PH	NPH*
	BM H	BM H & M	All-comer		Power of PA (IA)/Assumed HR	Power of PA (IA)/Re-estimated HR
1	+	+	+	All-comer at PA	74.5% (22.4%)/0.70	43.0% (7.9%)/0.77 ↓
2	+	+	-	BM H&M at PA	82.4% (30.9%)/0.75	53.0% (12.2%)/0.81 ↓
3	+	-		BM H at PA	87.6% (40.8%)/0.80	62.7% (19.0%)/0.85 ↓

Key message

- Either taking correlation into account or NPH will lower down powers for both IA and PA
  - When taking correlations into account, power decrease from PH to NPH is usually larger than the case of marginal power for PA (for IA, it is on the contrary)
  - When assuming the NPH model, power decrease from marginal power to correlated one is usually larger than the case of PH model for PA (for IA, it is on the contrary)
  - IA power for NPH considering correlation >= marginal IA power for NPH is most-likely due to poor approximation of the tentative strategy (intuitively it should be "<")
- More investigation needed to understand the 'why'

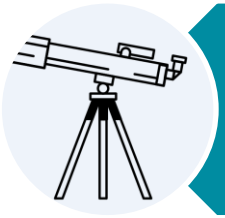
** Assuming there is a 3 months efficacy effect delay*

TOPICS FOR FURTHER EXTENSION

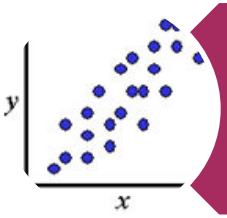


Maurer and Bretz (2013); Xi et al (2017); Tamhane et al (2018)

TAKE-AWAY MESSAGES



Interim analysis brings more complexity to group sequential design with multiple biomarker subgroups



Correlations among subgroups should be considered to calculate power

Scenario planning for claiming study success at IA/PA is be pre-discussed prior to study readouts

Group sequential testing method, non-proportional HR and alpha spending methods need to be fully discussed prior to study conduct

THANK YOU