

A background image of a globe showing the Americas, with a blue horizontal band across the middle.

A Staged Phase 2b/2c Trial Design to Accelerate pan-TB Drug Regimen Development

Robin Mogg (robin.mogg@takeda.com)
DIA Innovative Designs Scientific Working Group KOL lecture series
August 13, 2021

Acknowledgments



This is work from my time at the Bill & Melinda Gates Medical Research Institute and includes multiple contributors:

Gates MRI: Charles Wells, Nicole Frahm, Todd Bowser, Khisi Mdluli

BMGF: Dave Hermann, Debra Hanna

Otsuka: Jeff Hafkin

GSK: Gavin Koh, Alex Carlton

- TB is caused by bacteria (*Mycobacterium tuberculosis*) and most often affects the lungs; it is spread through the air similar to other respiratory diseases
- ~10 million people develop TB disease and ~1.5 million people die from TB **every year**; the top infectious killer every year until COVID-19
- About one-quarter of the world's population is estimated to be infected by TB. Only 5-15% of those infected will progress to develop TB disease, while the others remain only latently infected, not ill, and unable to transmit the disease
- Risk factors for progression to TB disease include: (1) recent infection; and (2) medical conditions that weaken the immune system (e.g., HIV)
- TB is curable using antibiotics, SOC treatment is a 6-month course of antibiotics (HRZE = isoniazid [H], rifampin [R], pyrazinamide [Z] and ethambutol [E]); adherence can be difficult
- Drug resistance is a global health and development threat; treatment is longer, more complex, and outcomes are less successful

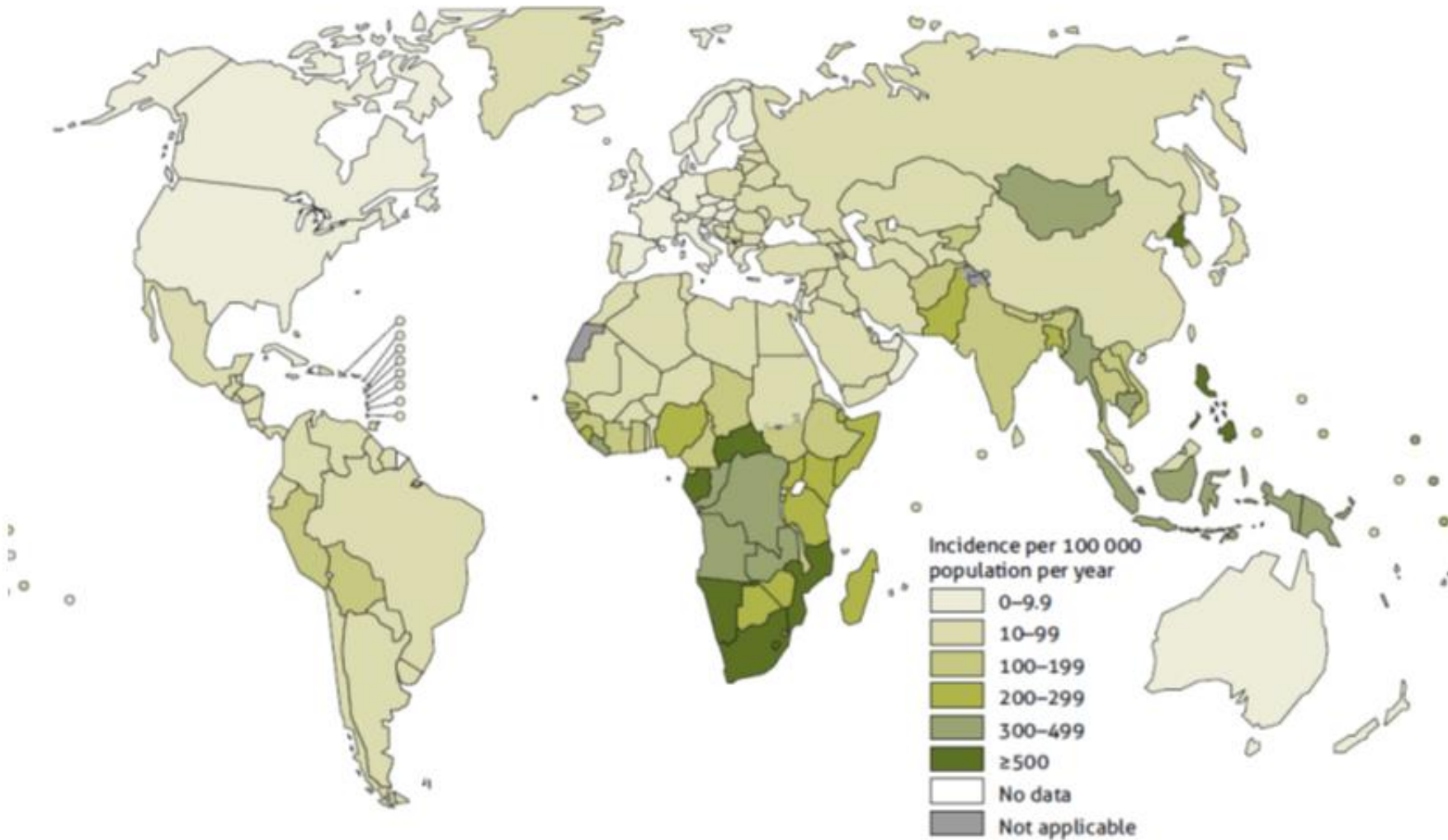
FDA NEWS RELEASE

FDA approves new drug for treatment-resistant forms of tuberculosis that affects the lungs

Approval marks the second drug approved under the Limited Population Pathway for Antibacterial and Antifungal Drugs and signals FDA's continued focus on facilitating development of new treatments to fight antimicrobial resistant infections

For Immediate Release: August 14, 2019

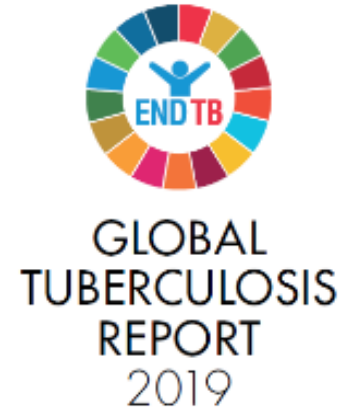
TB incidence rates in 2018



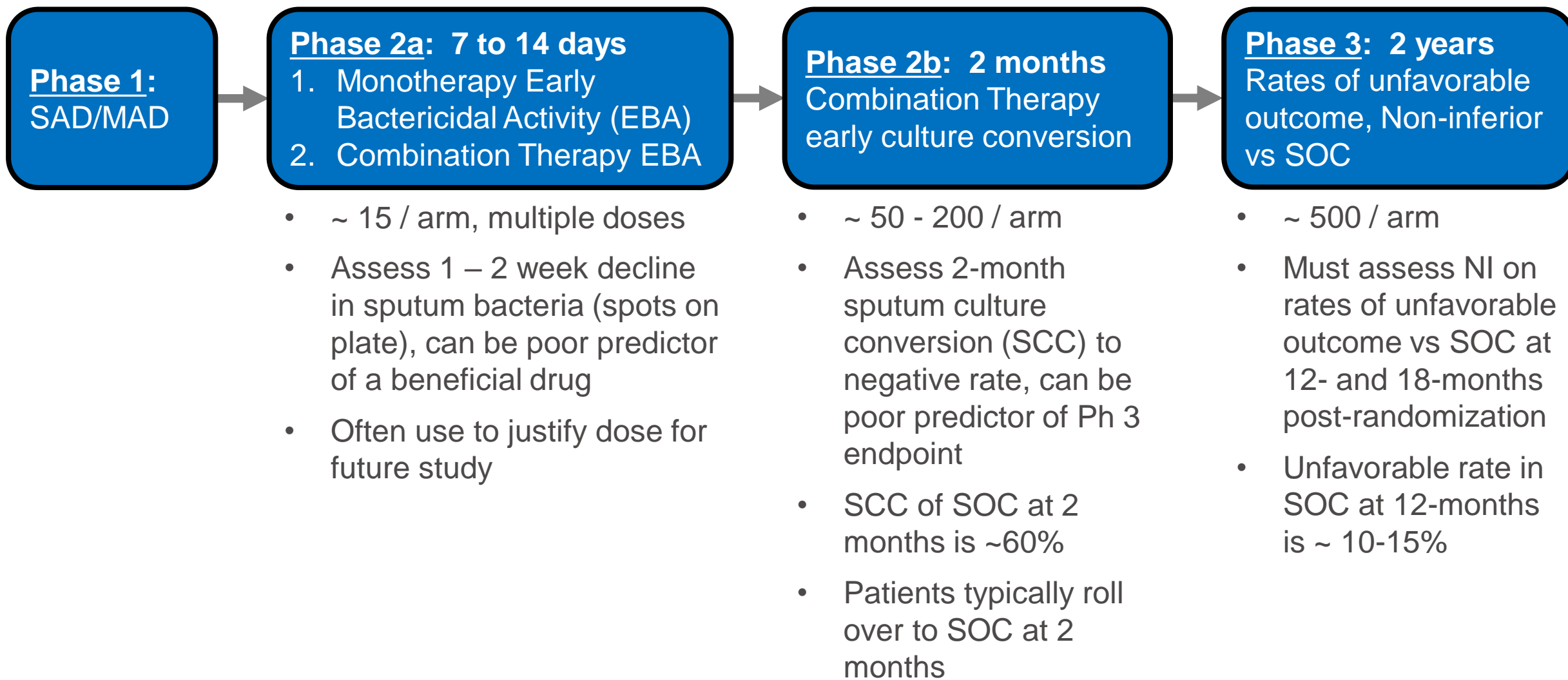
2018 Burden of TB Disease:

- 10 million people fell ill
 - 44% in SE Asia
 - 24% in Africa
 - 18% in WP
- 1.5 million people died
 - -27% vs 2000
 - 250,000 HIV+ (-60% vs 2000)

- The SDG and End TB Strategy targets set for 2030 cannot be met without intensified R&D
- Technological breakthroughs are needed by 2025, so that the annual decline in the global TB incidence rate can be accelerated to an average of 17% / year
- Priorities include:
 - A vaccine to lower the risk of infection
 - A vaccine or new drug treatment to cut the risk of TB disease in the 1.7 billion people already latently infected
 - Rapid diagnostics for use at the point of care
 - **Simpler, shorter drug regimens for treating TB disease**
- Recently, the M72/AS01E vaccine was found to be protective against TB disease in a Ph 2b trial among individuals with evidence of latent TB infection
 - If these findings are confirmed in a Ph 3 trial, this vaccine could transform global TB prevention efforts



Traditional TB Drug Clinical Development Pathway



Phillips et al. *BMC Medicine* (2016) 14:51
DOI 10.1186/s12916-016-0597-3



World TB Day

RESEARCH ARTICLE

Open Access



A new trial design to accelerate tuberculosis drug development: the Phase IIC Selection Trial with Extended Post-treatment follow-up (STEP)



Patrick P. J. Phillips^{1*}, Kelly E. Dooley², Stephen H. Gillespie³, Norbert Heinrich^{4,5}, Jason E. Stout⁶, Payam Nahid⁷, Andreas H. Diacon^{8,9}, Rob E. Aarnoutse¹⁰, Gibson S. Kibiki¹¹, Martin J. Boeree¹² and Michael Hoelscher^{4,5}

Rethinking non-inferiority: a practical trial design for optimising treatment duration

Matteo Quartagno^{1,2}, A Sarah Walker¹, James R Carpenter^{1,2}, Patrick PJ Phillips¹ and Mahesh KB Parmar¹

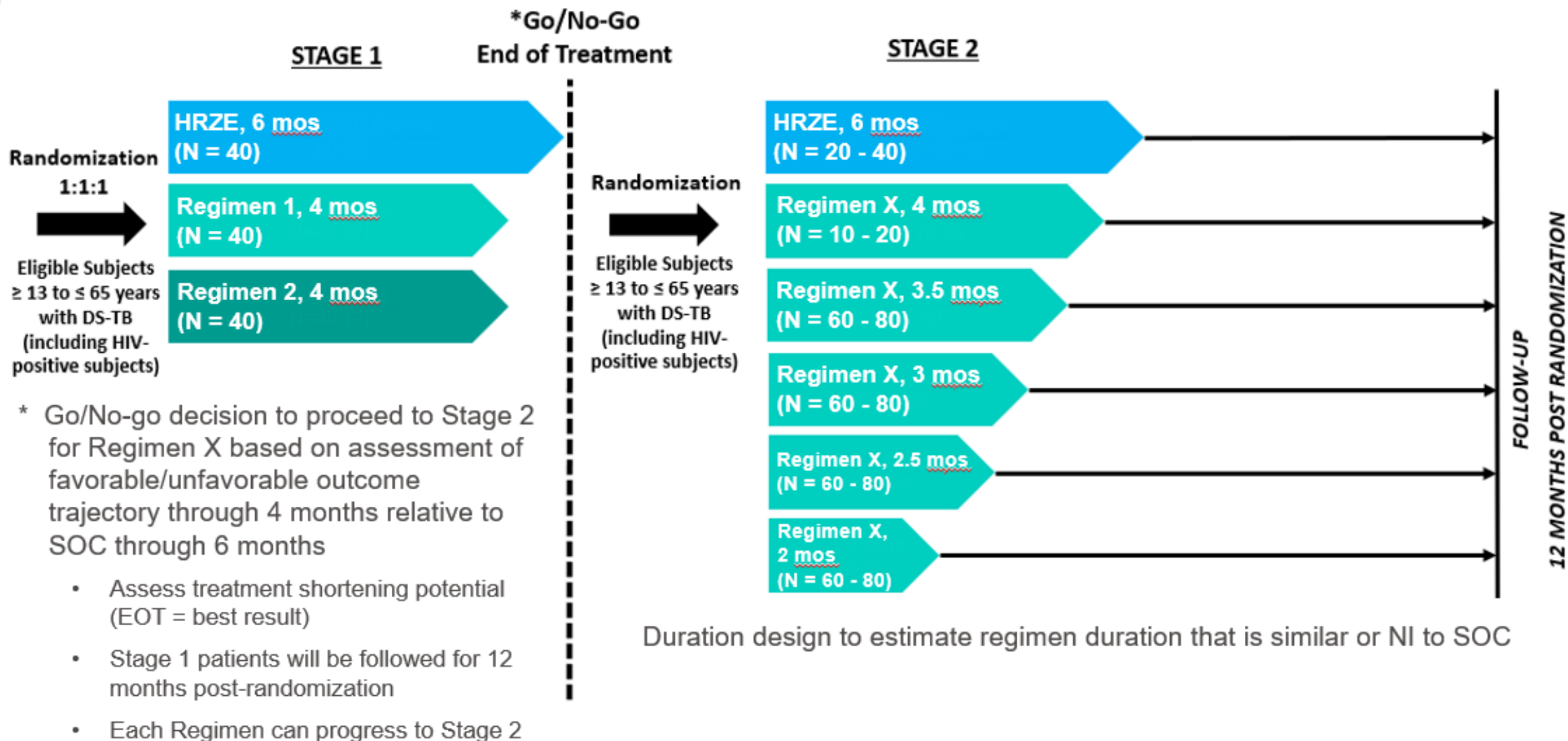
BMC Medicine

- Hybrid Phase 2/3 trial design where experimental regimen is given for the duration it will be studied in Phase 3 and patients are followed for clinical outcomes of treatment failure and TB relapse over 12 months post-randomization
- Collection of clinical outcome data in a relatively small number of participants provides valuable information about the likelihood of success in a future Phase 3 study

Clinical Trials
2018, Vol. 15(5) 477–488
© The Author(s) 2018

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1740774518778027
journals.sagepub.com/home/ctj


- “Durations design”: Recast the problem of finding an appropriate non-inferior treatment duration in terms of modelling the entire duration-response curve using a multi-arm randomized trial design, allocating patients to different treatment durations
- Fractional polynomials and spline-based methods
- Requires a total sample size of ~500 patients divided into a moderate number of durations (5-7)

Proposed Phase 2b/2c Study Design



- Stage 1: To assess treatment shortening potential by comparing the longitudinal favorable / unfavorable status rates **during treatment** with 4 months of a novel regimen relative to 6 months HRZE
- Stage 2: At 12 months post-randomization, compare the favorable / unfavorable outcome rates after treatment of varying durations (e.g., 2, 2.5, 3, 3.5, and 4 months) of a novel regimen relative to 6 months of HRZE
 - Primary Stage 2 Hypothesis: For at least one regimen duration, participants randomized to receive the novel regimen will have unfavorable outcome rates at 12 months post-randomization that are non-inferior to HRZE

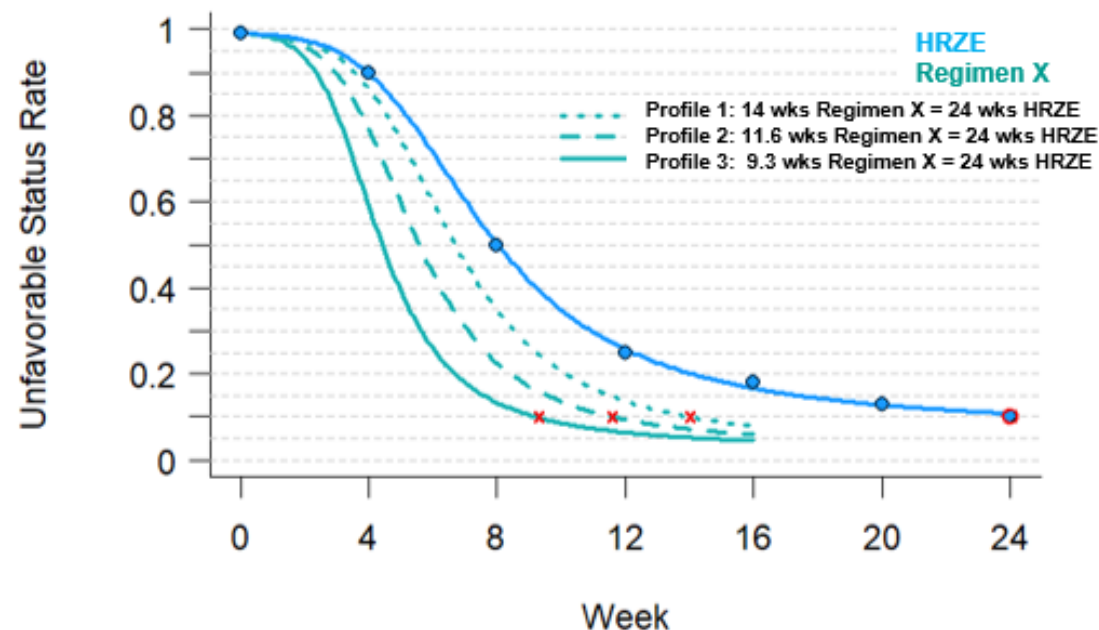
Stage 1 (Phase 2b)



Potential weekly (or subset of weekly) assessments



(Some) Potential Improved True Longitudinal Profiles



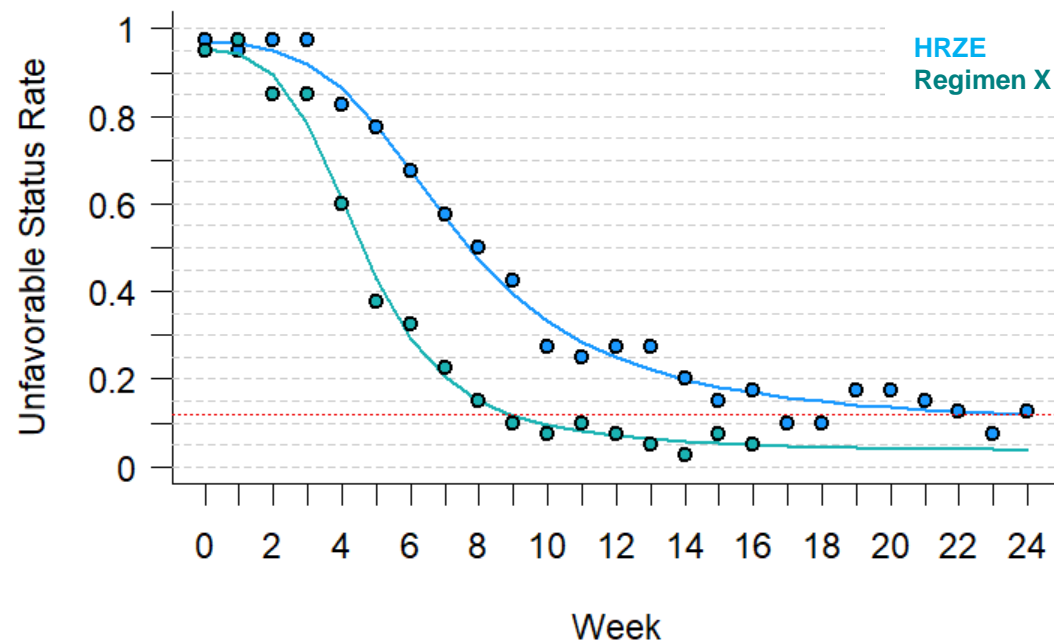
Analysis: Estimate and compare time-response profiles of unfavorable status rate between Regimen X and HRZE, accounting for within-subject correlation

Go/No-go: Evidence that Regimen X response profile is shifted favorably (e.g., to the left and/or lower) relative to the HRZE time-response profile suggests treatment shortening potential and “Go” to Stage 2

Illustration: Analysis of Example Data for Stage 1

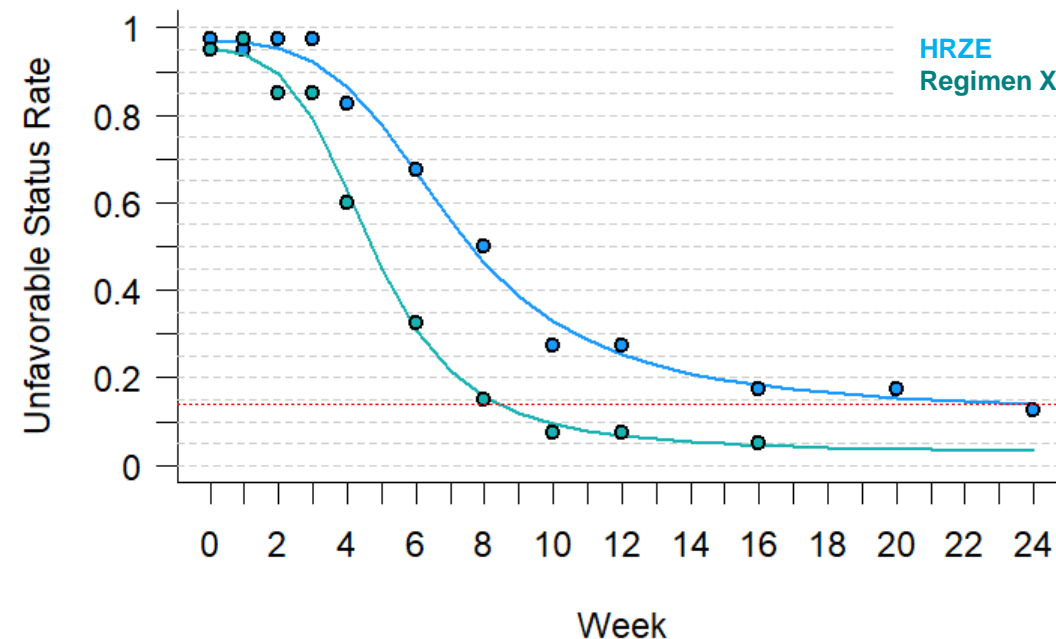


Weekly assessments



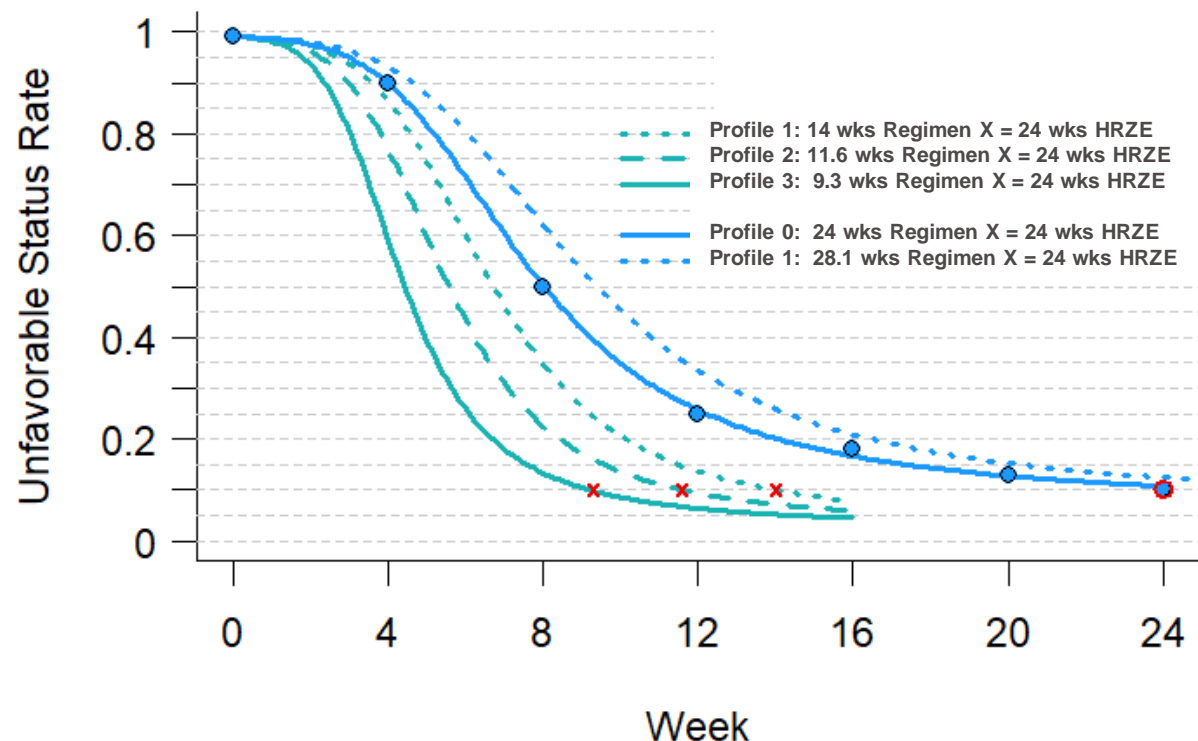
<u>HRZE 24 wks failure rate</u>	<u>Duration Reg X = HRZE 24 wks</u>	
11.8%	9.0 wks	
<u>HRZE 16 wks failure rate</u>	<u>Reg X 16 wks failure rate</u>	<u>P-value</u>
16.9%	5.0%	< 0.0001
<u>HRZE ED50</u>	<u>Reg X ED50</u>	<u>P-value</u>
6.4 wks	4.9 wks	0.079

Weekly through 4 wks, bi-weekly through 12 wks, monthly through 24 wks



<u>HRZE 24 wks failure rate</u>	<u>Duration Reg X = HRZE 24 wks</u>	
14.0%	8.4 wks	
<u>HRZE 16 wks failure rate</u>	<u>Reg X 16 wks failure rate</u>	<u>P-value</u>
18.3%	4.5%	0.003
<u>HRZE ED50</u>	<u>Reg X ED50</u>	<u>P-value</u>
6.0 wks	5.2 wks	0.294

True underlying response profiles



Simulation objectives: Evaluate operating characteristics to develop go/no-go criteria

- 5 underlying true Regimen X strategies
 - 3 sample sizes: N = 30, 40 or 50 / group
 - 3 sampling strategies: 12, 15 or 25 samples collected*
- * 12 samples = collection at 0, 1, 2, 3, 4, 6, 8, 10, 12, 16, 20, 24 wks
 15 samples = collection at 0, 1, 2, 3, 4, 5, 6, 7, 8, 10, 12, 14, 16, 20, 24 wks
 24 samples = weekly collection

Probability of advancing to Stage 2 should be **relatively low for Profiles -1 and 0**, and **relatively high for Profiles 2 and 3**

Potential questions to inform go/no-go:

- Is the estimated duration where Regimen X failure rate equals 24 wk HRZE failure rate < 16 wks?
 (Using data through wk 24 for HRZE and data through wk 16 for Regimen X)
- Is 16 wk Regimen X failure rate < 16 wk HRZE failure rate? Is Regimen X ED50 < HRZE ED50?
 (Using data through wk 24 OR using data through wk 16 for HRZE, and data through wk 16 for Regimen X)

Stage 1 Simulation Results [1]

Using data through wk 24 for HRZE and data through wk 16 for Regimen X



Prob(Est. duration*
< 16 wks) #samp #
 / subj subj

12.3	23.6	65.8	76.8	84	12	30
10.9	22.2	68.4	79.1	86.4	15	
9.5	18.6	67.8	80.2	87.9	25	
8.9	21.5	68.8	80.1	88	12	40
8	20.5	68.2	81.7	88.8	15	
6.8	17.3	73	83.9	90.9	25	
6.7	19.3	72.6	84	89.3	12	50
6	17.3	71.3	83.6	90.9	15	
5.9	15.2	70.8	83.7	91.4	25	

- Using a go/no-go criteria around estimated Regimen X duration that is comparable to 24 wks HRZE being < 16 wks has “type I error” around 20%, but power > 80%

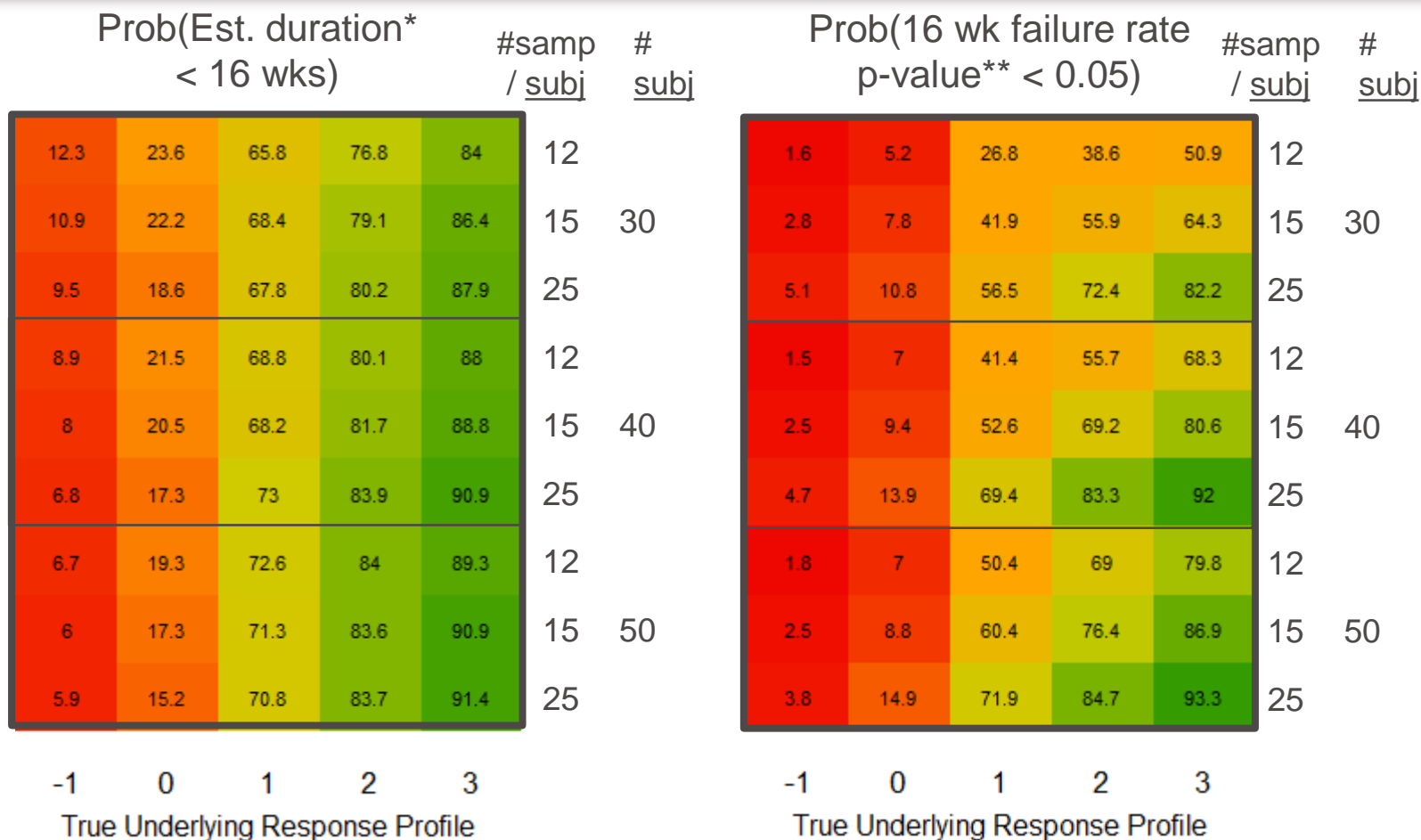
-1 0 1 2 3
True Underlying Response Profile

* where Regimen X failure rate equals 24 wk
HRZE failure rate

Profile -1: 28.1 wks Regimen X = 24 wks HRZE; Profile 0: 24 wks Regimen X = 24 wks HRZE;
Profile 1: 14.0 wks Regimen X = 24 wks HRZE; Profile 2: 11.6 wks Regimen X = 24 wks HRZE; Profile 3: 9.3 wks Regimen X = 24 wks HRZE

Stage 1 Simulation Results [2]

Using data through wk 24 for HRZE and data through wk 16 for Regimen X



- Using a go/no-go criteria around estimated Regimen X duration that is comparable to 24 wks HRZE being < 16 wks has “type I error” around 20%, but power > 80%
- Using a go/no-go criteria around a comparison of wk 16 failure rates has improved type I error but less than optimal power
- Increasing the number of assessments improves operating characteristics

* where Regimen X failure rate equals 24 wk HRZE failure rate

** Regimen X 16 wk failure rate significantly lower than HRZE 16 wk failure rate

Profile -1: 28.1 wks Regimen X = 24 wks HRZE; Profile 0: 24 wks Regimen X = 24 wks HRZE;
 Profile 1: 14.0 wks Regimen X = 24 wks HRZE; Profile 2: 11.6 wks Regimen X = 24 wks HRZE; Profile 3: 9.3 wks Regimen X = 24 wks HRZE

Stage 1 Simulation Results [3]

Using data through wk 24 for HRZE and data through wk 16 for Regimen X



Prob(Est. duration*
< 16 wks)

#samp
/ subj

subj

12.3	23.6	65.8	76.8	84	12	30
10.9	22.2	68.4	79.1	86.4	15	
9.5	18.6	67.8	80.2	87.9	25	
8.9	21.5	68.8	80.1	88	12	40
8	20.5	68.2	81.7	88.8	15	
6.8	17.3	73	83.9	90.9	25	
6.7	19.3	72.6	84	89.3	12	50
6	17.3	71.3	83.6	90.9	15	
5.9	15.2	70.8	83.7	91.4	25	

-1 0 1 2 3
True Underlying Response Profile

Prob(16 wk failure rate
p-value** < 0.05)

#samp
/ subj

subj

1.6	5.2	26.8	38.6	50.9	12	30
2.8	7.8	41.9	55.9	64.3	15	
5.1	10.8	56.5	72.4	82.2	25	
1.5	7	41.4	55.7	68.3	12	40
2.5	9.4	52.6	69.2	80.6	15	
4.7	13.9	69.4	83.3	92	25	
1.8	7	50.4	69	79.8	12	50
2.5	8.8	60.4	76.4	86.9	15	
3.8	14.9	71.9	84.7	93.3	25	

-1 0 1 2 3
True Underlying Response Profile

Prob(Est. duration* < 16
wks OR 16 wk failure
rate p-value** < 0.05)

#samp
/ subj

subj

12.3	23.6	65.8	77	84.3	12	30
10.9	22.4	68.8	79.3	86.9	15	
9.7	19.4	69.3	82.4	90	25	
8.9	21.6	69.1	80.4	88.4	12	40
8.1	21	69.2	83	89.9	15	
7.1	19	76.1	87.1	93.7	25	
6.7	19.4	73	84.7	89.8	12	50
6.1	17.5	72.8	85.2	92	15	
6.5	18.3	75.9	86.6	93.8	25	

-1 0 1 2 3
True Underlying Response Profile

* where Regimen X failure rate equals 24 wk
HRZE failure rate

** Regimen X 16 wk failure rate significantly
lower than HRZE 16 wk failure rate

Profile -1: 28.1 wks Regimen X = 24 wks HRZE; Profile 0: 24 wks Regimen X = 24 wks HRZE;
Profile 1: 14.0 wks Regimen X = 24 wks HRZE; Profile 2: 11.6 wks Regimen X = 24 wks HRZE; Profile 3: 9.3 wks Regimen X = 24 wks HRZE

Stage 1 Simulation Results [4]

Using data through wk 24 for HRZE and data through wk 16 for Regimen X



Prob(Est. duration*
< 16 wks)

#samp
/ subj

subj

12.3	23.6	65.8	76.8	84	12
10.9	22.2	68.4	79.1	86.4	15 30
9.5	18.6	67.8	80.2	87.9	25
8.9	21.5	68.8	80.1	88	12
8	20.5	68.2	81.7	88.8	15 40
6.8	17.3	73	83.9	90.9	25
6.7	19.3	72.6	84	89.3	12
6	17.3	71.3	83.6	90.9	15 50
5.9	15.2	70.8	83.7	91.4	25

-1 0 1 2 3
True Underlying Response Profile

Prob(16 wk failure rate
p-value** < 0.10)

#samp
/ subj

subj

4.4	10.9	43.2	57	67.8	12
5.4	13.7	56.2	69.7	78.4	15 30
8.6	17.8	68.1	81	88.6	25
4.2	12.9	56.4	70.4	80.9	12
5	15.6	64.6	80.2	88.6	15 40
7.1	21	77.5	88.8	94.2	25
3.6	13.5	65	80.7	88	12
4.8	14.8	71	84.4	91.8	15 50
6.7	20.3	79	88.7	95.5	25

-1 0 1 2 3
True Underlying Response Profile

** Regimen X 16 wk failure rate significantly
lower than HRZE 16 wk failure rate

Prob(Est. duration* ≤ 16
wks OR 16 wk failure
rate p-value** < 0.10)

#samp
/ subj

subj

12.5	24	66.3	77.3	85.2	12
11.1	22.7	70.3	80.9	88	15 30
11.1	21.6	72.7	85	91.5	25 30
9.2	21.9	70.4	81.5	89.3	12
8.5	22.1	71.6	85.2	91.5	15 40
8.3	22.6	79.7	89.8	95	25
7	20.5	74.6	86.4	91.3	12 50
6.7	19.1	75.9	87.3	93.5	15 50
8.2	22	80.4	89.3	95.6	25

-1 0 1 2 3
True Underlying Response Profile

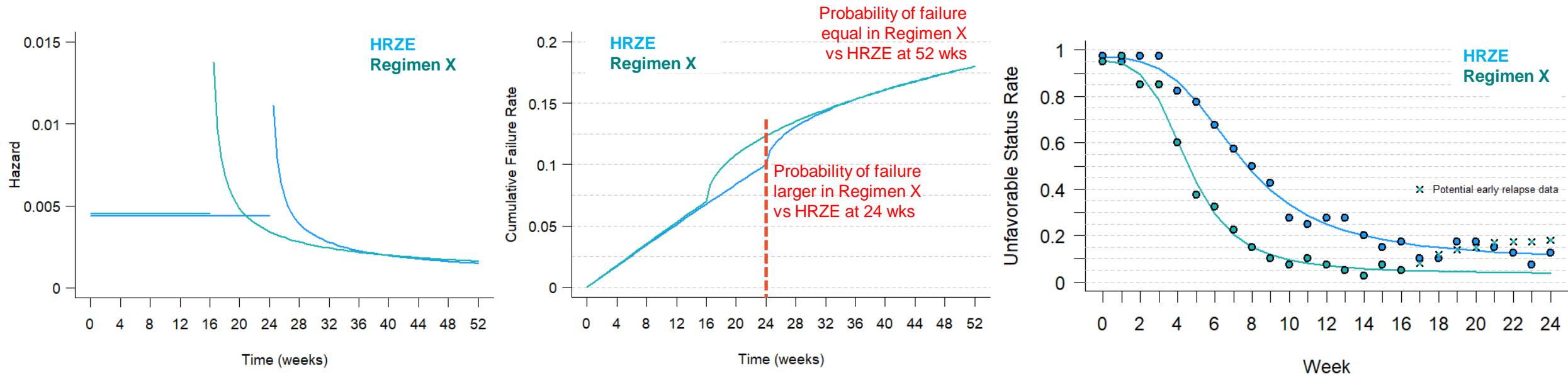
* where Regimen X failure rate equals 24 wk
HRZE failure rate

Profile -1: 28.1 wks Regimen X = 24 wks HRZE; Profile 0: 24 wks Regimen X = 24 wks HRZE;
Profile 1: 14.0 wks Regimen X = 24 wks HRZE; Profile 2: 11.6 wks Regimen X = 24 wks HRZE; Profile 3: 9.3 wks Regimen X = 24 wks HRZE

Cumulative probability of failure at 24 wks may be higher in Regimen X vs HRZE



- Example profile of true hazard of failure over time (left), cumulative true probability of failure (center), and potential impact on observed data (right)
 - Reasonable to assume that probability of failure is relatively constant during treatment, but after treatment the probability of relapse is highest immediately after treatment and then decreases over time

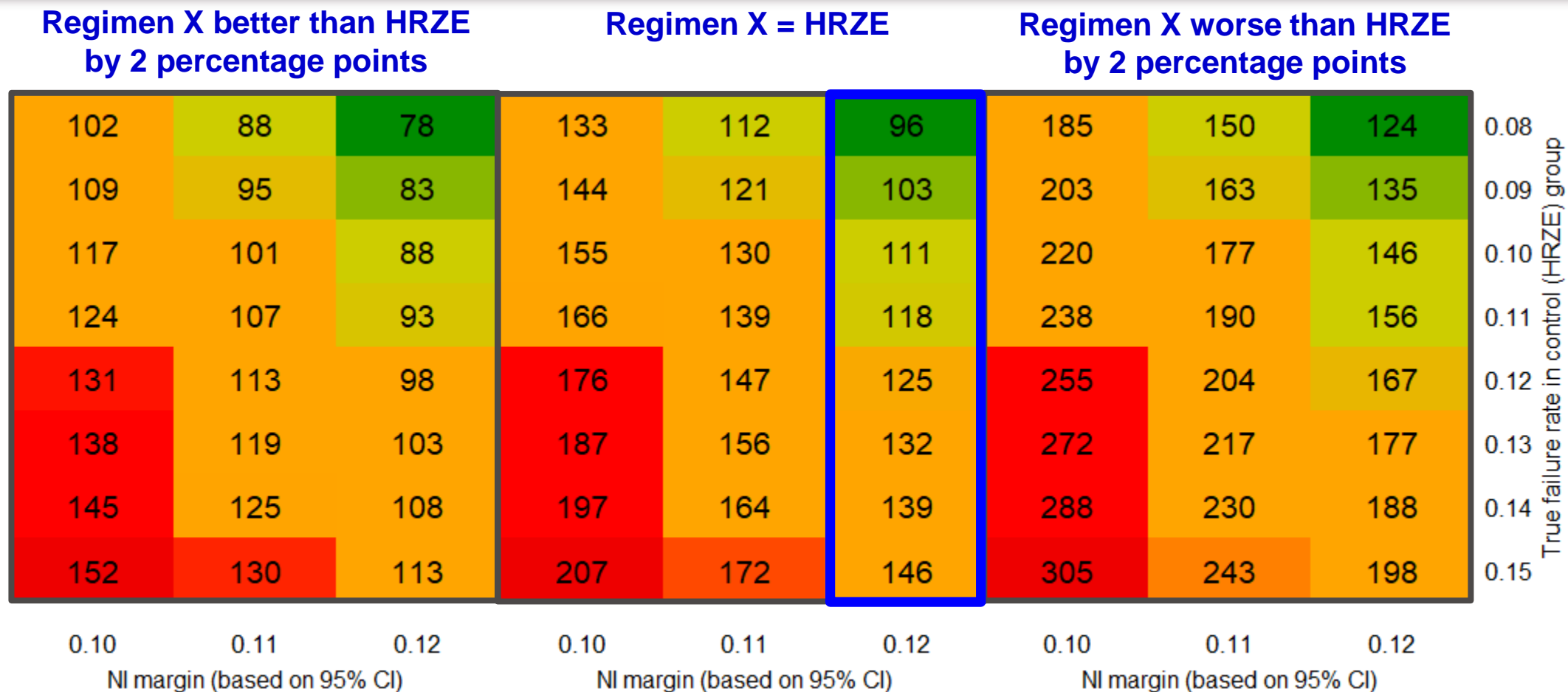


- Early relapse data should be considered cautiously
 - We will revisit this issue when monitoring futility in Stage 2

- Stage 1 sample sizes of 40 subjects / group with 15 samples / subject or 50 subjects / group with 12 samples / subject
- Recommend two-fold go/no-go criteria using HRZE data through 24 wks and Regimen X data through 16 wks:
 1. Based on statistical modeling, estimated Regimen X duration that has failure rate comparable to 24 wks HRZE is < 16 wks; OR
 2. P-value assessing if wk 16 Regimen X failure rate < wk 16 HRZE failure rate < 0.10
- Clinical trial simulations show following operating characteristics of two-fold go/no-go criteria:
 - < 10% probability of progressing to Stage 2 when Regimen X is worse than HRZE
 - ~ 20% probability of progressing to Stage 2 when Regimen X = HRZE
 - ~ 85% probability of progressing to Stage 2 when 11.6 wks Regimen X = 24 wks HRZE
 - ~ 90% probability of progressing to Stage 2 when 9.3 wks Regimen X = 24 wks HRZE
- Preliminary relapse data in Regimen X arm can also contribute to totality of the evidence for moving into Stage 2, but early relapse data should be considered cautiously

- Stage 2: At 12 months post-randomization, compare the favorable / unfavorable outcome rates after treatment of varying durations (e.g., 2, 2.5, 3, 3.5, and 4 months) of a novel regimen relative to 6 months of HRZE
 - Primary Stage 2 Hypothesis: For at least one regimen duration, participants randomized to receive the novel regimen will have unfavorable outcome rates at 12 months post-randomization that are non-inferior to HRZE
 - Possible strategies to assess NI:
 - Hierarchical pairwise comparisons: Start with longest duration (e.g., 4 months) of Regimen X and compare to HRZE, if successful “step-down” and compare next longest duration (3.5 months) to HRZE. Continue to step-down until a duration is unsuccessful (Type I error is spent) or until all durations are tested and successful.
 - MCP-Mod approach: Pre-specify possible duration response curves. Test for a significant duration response for one or more pre-specified response curves. If one or more response curves are significant, select “best” model (or perform model averaging of significant models) to determine the minimal duration that achieves NI. Requires at least 4 durations to be tested, 5 durations is even better (per-group sample sizes do not have to be equal).

Sample sizes for pairwise comparisons (80% power)



- For pairwise comparisons, a sample size of 100 – 150 per group is needed when using a NI margin of 12 percentage points (Regimen X = HRZE, 1-sided $\alpha = 2.5\%$, power = 80%)
 - With 3 durations vs control (i.e., Minimum Total N = 400), power to succeed on 1, 2 or 3 regimens = 80%, 64%, and 51%, respectively, using hierarchical method

What is MCP-Mod?



- Bretz et al, 2005 (Biometrics): A structured approach to model-based design and analysis of dose-finding studies under model uncertainty; positive opinions for Phase 2 use by EMA and FDA

EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINE

23 January 2014
EMA/CHMP/SAWP/757052/2013
Committee for Medicinal Products for Human Use (CHMP)

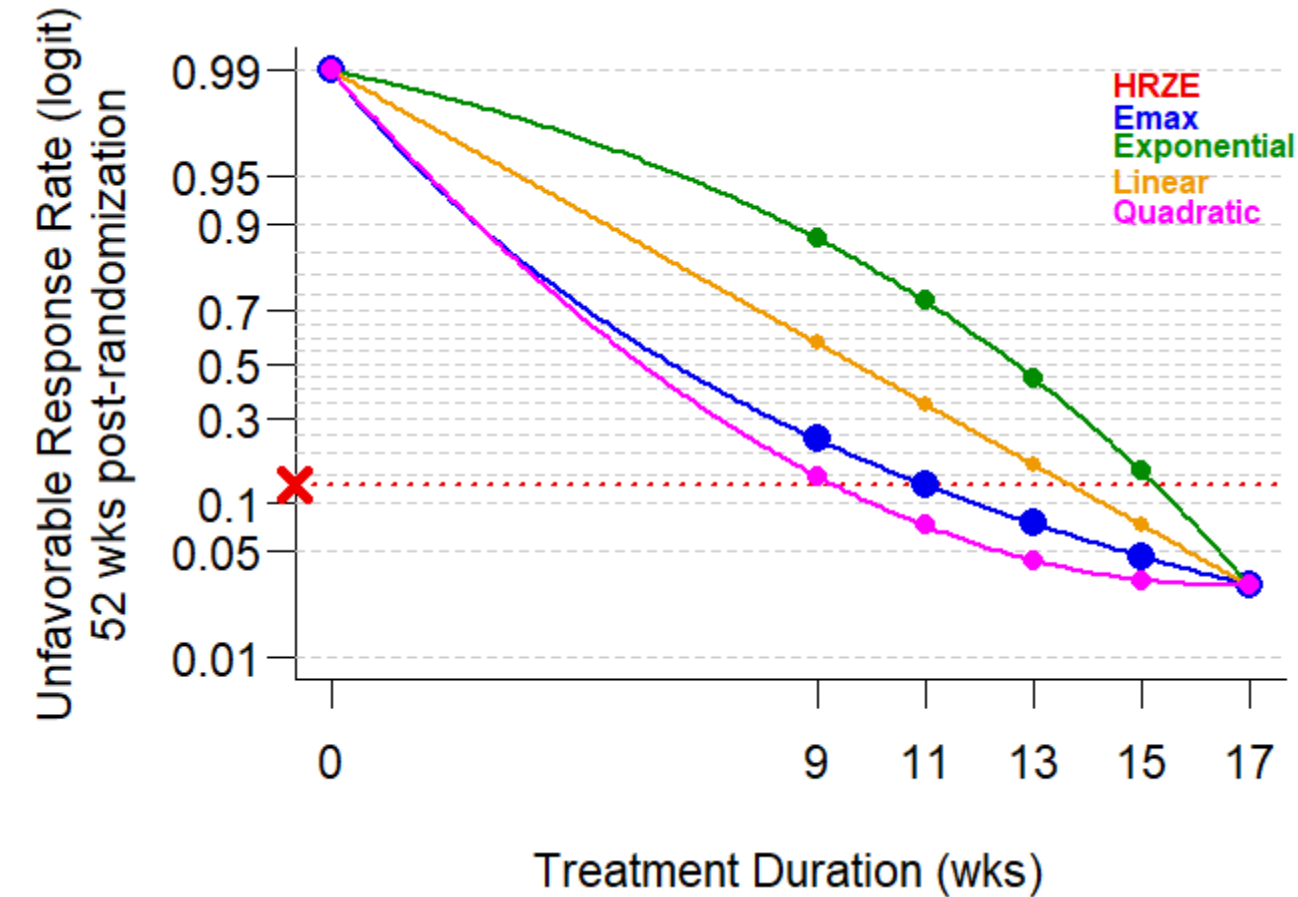
Qualification Opinion of MCP-Mod as an efficient statistical methodology for model-based design and analysis of Phase II dose finding studies under model uncertainty

FDA Determines *Multiple Comparison Procedure-Modeling (MCP-Mod)* Statistical Approach is Fit-for-Purpose

On May 26, 2016, FDA determined that the Multiple Comparison Procedure – Modeling (MCP-Mod) statistical approach is fit-for-purpose (FFP). FFP designation provides a pathway for FDA to assess and comment on the utility of drug development tools. Such tools are made [publicly available](#) in an effort to facilitate appropriate utilization in drug development.

- Can be used for duration-finding instead of dose-finding
- **MCP (= Multiple Comparisons Procedure)**: we are faced with 2 specific sources of multiplicity
 - Testing for a significant duration-response signal under model uncertainty (multiple plausible underlying dose-response profiles) can be done using “multiple contrasts test”
 - Confirmatory pairwise comparisons of individual durations against HRZE while borrowing strength from estimated duration response relationships (for duration selection)
- **Mod (= Modeling)**: assumes durations (d) and response (y) are related through a functional relationship (f) with model parameters (θ), i.e., $y_i = f(d_i, \theta) + \varepsilon_i$

Example candidate duration response profiles

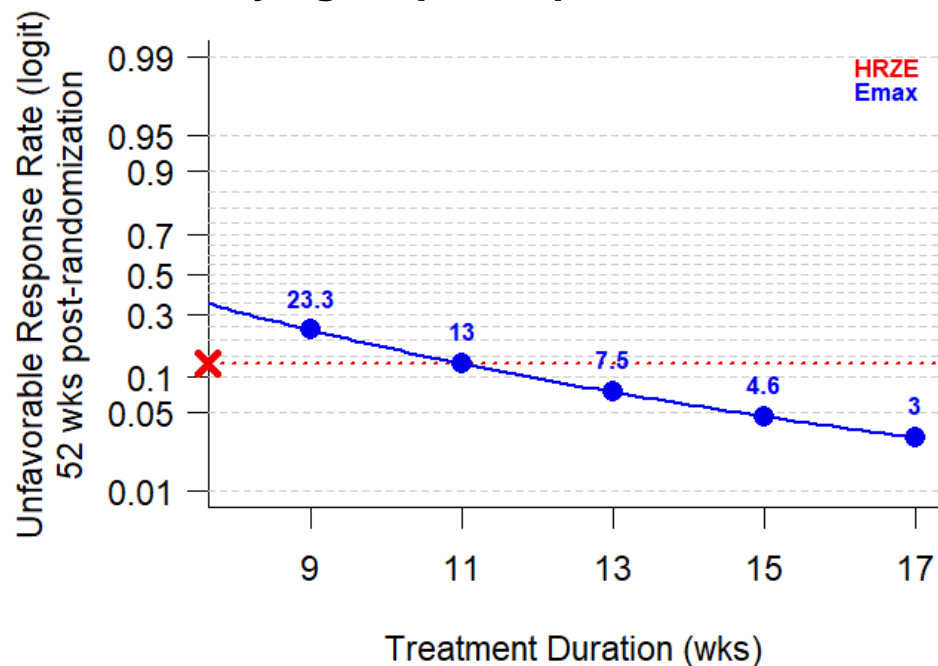


- In typical dose response studies, a dose of 0 (placebo) can be used in the modeling; here we do not have a duration = 0
 - With an active control, only the Regimen X groups can be used in the modeling
 - Comparisons to active control can be done based on the model-predicted unfavorable response rates from significant duration-response models

Illustration: Example duration response data



True underlying response profile



Hypothetical observed data from a trial with N = 70 / group (Total N = 420)

# failure / N (% failure)					
HRZE	9 wks Reg X	11 wks Reg X	13 wks Reg X	15 wks Reg X	17 wks Reg X
9 / 70 (12.9%)	18 / 70 (25.7%)	10 / 70 (14.3%)	3 / 70 (4.3%)	5 / 70 (7.1%)	1 / 70 (1.4%)

- **Question 1:** Is there a significant duration response?
 - Does not take into account the failure rate of HRZE
- **Question 2:** If there is a significant duration response, what is the minimum estimated duration that is NI to HRZE?
 - The minimum estimated duration depends on how NI to HRZE is defined
- **Let's analyze our observed data to answer the 2 questions**

Illustration Question 1: Is there a significant duration response?



- Pre-specified candidate models for duration-response shape of Regimen X: linear, quadratic, exponential, emax
- Use “generalized” MCP-Mod procedure: (1) use logistic regression model fit with duration as a “factor” (no assumption of duration-response shape) to get estimates and variances at each duration; (2) use estimates and variances to test significance of varying duration-response curve

Model	t-Stat	Adj p-value
Emax	4.255	< 0.001
Quadratic	4.249	< 0.001
Linear	4.234	< 0.001
Exponential	4.177	< 0.001
Critical t-Stat value = 2.042		

Several strategies can be taken for model selection (or averaging across statistically significant candidate models), for simplicity we will select model based on the maximum t-Statistic

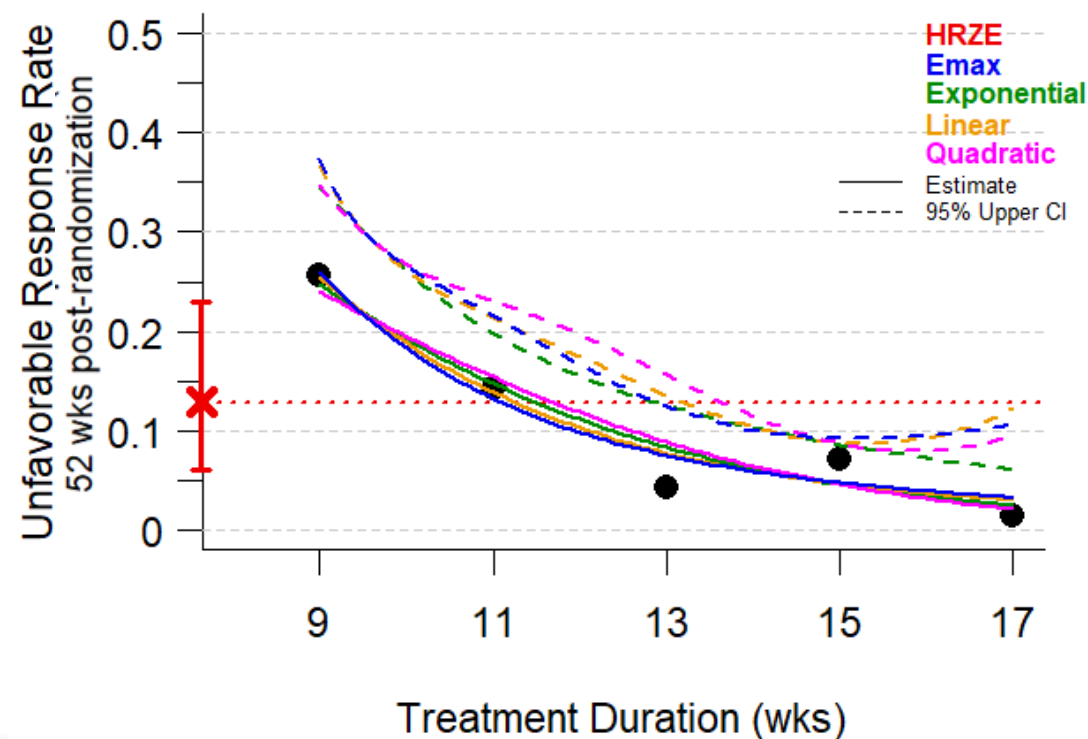


Illustration Question 2: What is the minimum estimated duration that is NI to HRZE?



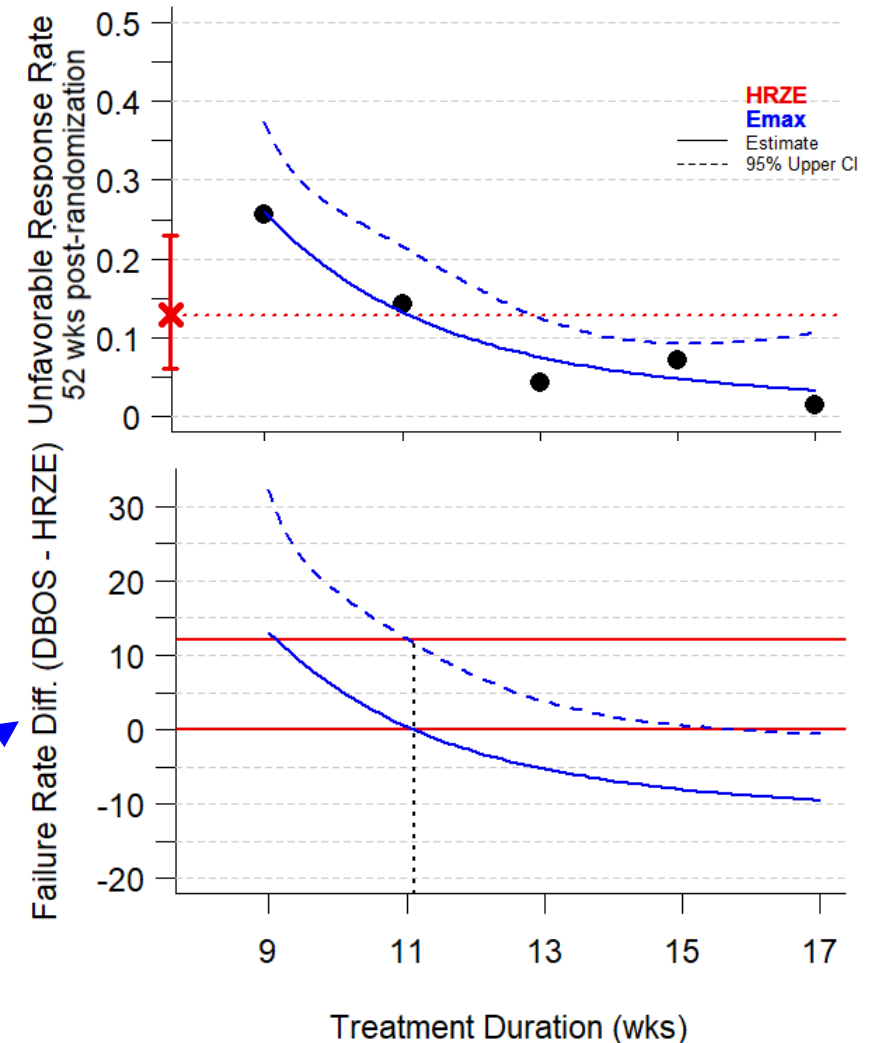
Duration (wks)	Est. Failure Rate	95% Wald CI*	Est. Diff from HRZE	95% Wald CI**
17	3.3	(1.0, 10.6)	-9.6	(-18.5, -0.6)
15	4.8	(2.4, 9.2)	-8.1	(-16.7, 0.5)
13	7.5	(4.4, 12.5)	-5.3	(-14.4, 3.8)
11	13.2	(7.8, 21.5)	0.4	(-11.4, 12.2)
9	26.0	(17.1, 37.4)	13.1	(-6.0, 32.2)

HRZE estimated failure rate = 12.9, 95% exact CI = (6.1, 23.0)

* Based on model fit (back-transformed from logit scale)

** Based on delta method to estimate failure rate variance from model fit on the logit scale

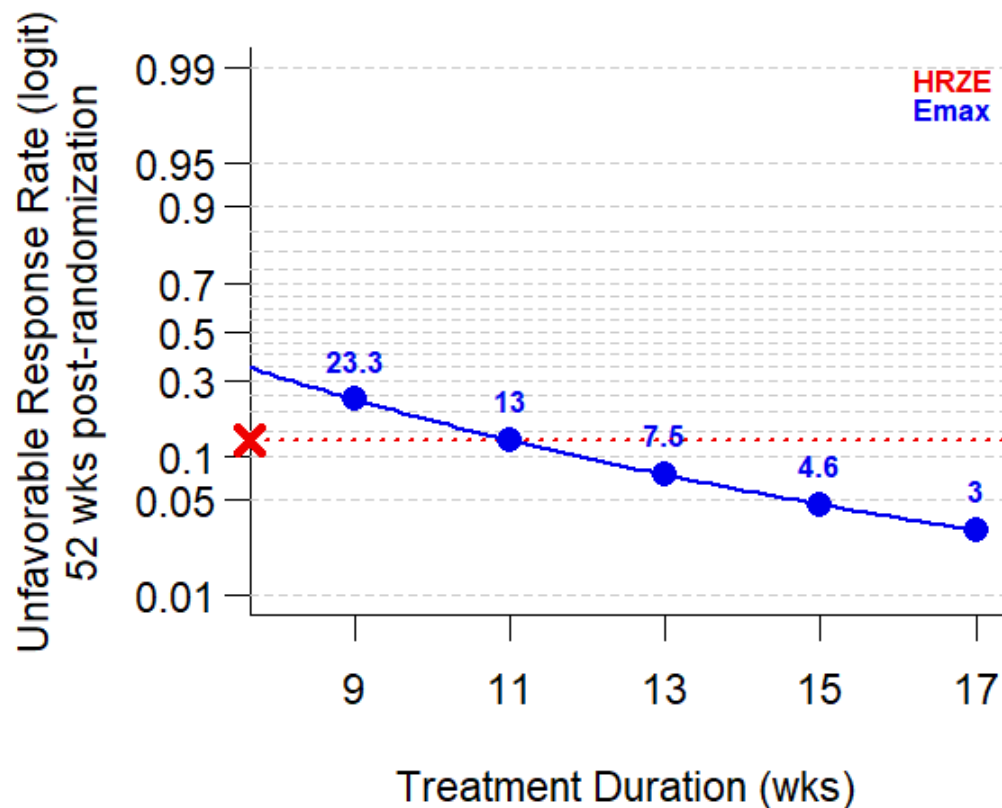
- Of the doses tested, durations ≥ 13 weeks achieve NI using a NI margin of 12 percentage points
- Based on the model, we estimate the minimum duration to achieve NI using a NI margin of 12 percentage points to be 11.1 weeks



Simulation Setup: True Response Profile = EMAX [1]



True underlying response profile



Scenario	Per group sample size						Total N
	HRZE	9 wks Reg X	11 wks Reg X	13 wks Reg X	15 wks Reg X	17 wks Reg X	
1	70	70	70	70	70	70	420
2	60	60	60	60	60	60	360
3	50	50	50	50	50	50	300
4	50	50	25	50	25	50	250
5	60	60	30	60	30	60	300
6	60	60	20	60	20	60	280
7	60	60	0	60	30	60	270
8	60	60	0	60	60	30	270

Questions:

- What is power to detect a duration response?
- What is estimated minimum duration to achieve NI based on 95% CI Upper Bound ≤ 12
- How does power and estimation compare to using hierarchical pairwise comparisons?

EMAX [1]: HRZE failure rate = 13%, 11 wks Reg X = HRZE

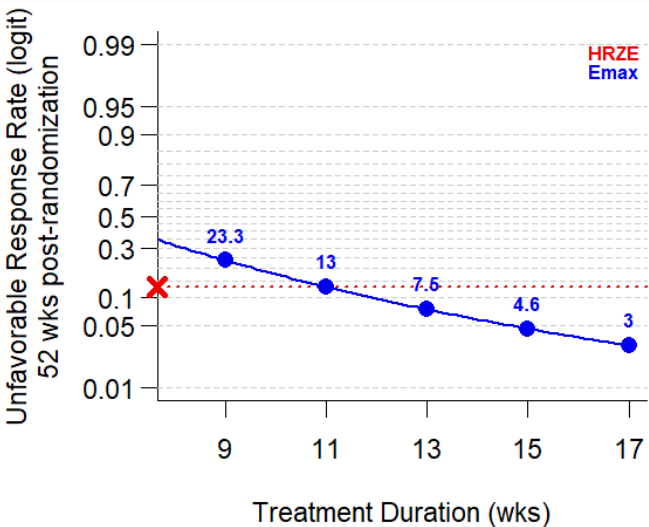
Evaluate later

EMAX [2]: HRZE failure rate = 13%, 15 wks Reg X = HRZE

EMAX [3]: HRZE failure rate = 18%, 13 wks Reg X = HRZE

EMAX [0]: HRZE failure rate = 13%, 17 wks Reg X = HRZE + 12 (Null)

Simulation Results: True Response Profile = EMAX [1]



Scenario	Per group sample size						Total N
	HRZE	9 wks Reg X	11 wks Reg X	13 wks Reg X	15 wks Reg X	17 wks Reg X	
1	70	70	70	70	70	70	420
2	60	60	60	60	60	60	360
3	50	50	50	50	50	50	300
4	50	50	25	50	25	50	250
5	60	60	30	60	30	60	300
6	60	60	20	60	20	60	280
7	60	60	0	60	30	60	270
8	60	60	0	60	60	30	270

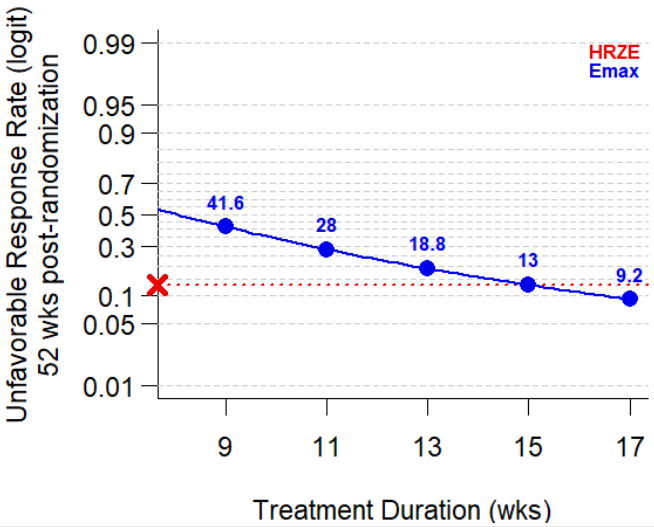
EMAX [1]: HRZE failure rate = 13%, 11 wks Reg X = HRZE

Scenario	Hierarchical Pairwise Comparisons (HPC)		MCP-Mod	
	Power	If Significant, Est. Min. Duration where 95% UB Reg X – HRZE ≤ 12*	Power	If Significant, Est. Min. Duration where 95% UB Reg X – HRZE ≤ 12*
1	99.8	13 (11, 15)	97.2	10.8 (9.4, 12.8)
2	99.4	13 (11, 15)	93.8	10.9 (9.6, 13.1)
3	97.8	13 (11, 17)	86.3	11.2 (9.9, 13.6)
4	97.4	13 (11, 17)	78.5	11.5 (10.0, 14.1)
5	99.2	13 (11, 17)	87.4	11.2 (9.8, 13.5)
6	99.0	13 (11, 17)	84.9	11.4 (9.9, 13.8)
7	99.2	13 (13, 17)	87.7	11.7 (10.3, 13.8)
8	86.0	13 (13, 15)	84.4	11.8 (10.6, 13.4)

* Median (90% lower, upper bound); Based on 2000 simulations per row
Power HPC = Probability 95% UB Reg X – HRZE ≤ 12 for 17 wks duration;
Power MCP-Mod = Probability p-value for dose response ≤ 0.025 and 95% UB Reg X – HRZE ≤ 12 for 17 wks duration

HRZE failure rate = 13%, 11 wk Reg X = HRZE: HPC has higher power than MCP-Mod; HPC overestimates duration

Simulation Results: True Response Profile = EMAX [2]



Scenario	Per group sample size						Total N
	HRZE	9 wks Reg X	11 wks Reg X	13 wks Reg X	15 wks Reg X	17 wks Reg X	
1	70	70	70	70	70	70	420
2	60	60	60	60	60	60	360
3	50	50	50	50	50	50	300
4	50	50	25	50	25	50	250
5	60	60	30	60	30	60	300
6	60	60	20	60	20	60	280
7	60	60	0	60	30	60	270
8	60	60	0	60	60	30	270

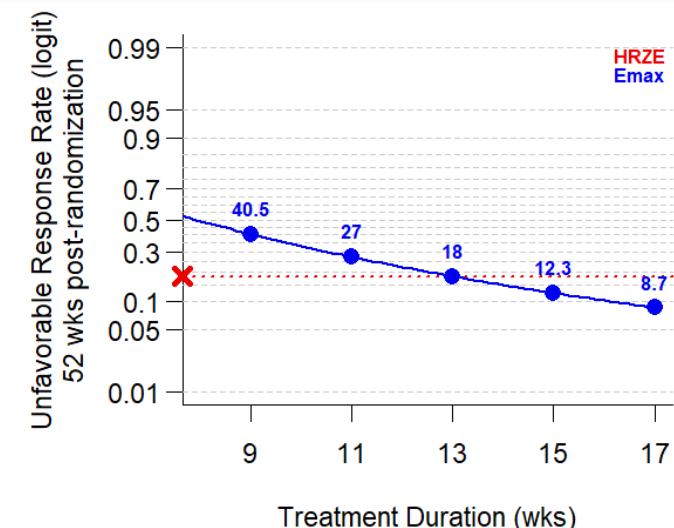
EMAX [2]: HRZE failure rate = 13%, 15 wks Reg X = HRZE

Scenario	Hierarchical Pairwise Comparisons (HPC)		MCP-Mod	
	Power	If Significant, Est. Min. Duration where 95% UB Reg X – HRZE ≤ 12*	Power	If Significant, Est. Min. Duration where 95% UB Reg X – HRZE ≤ 12*
1	82.2	15 (13, 17)	84.7	15.2 (12.1, 16.1)
2	74.9	15 (13, 17)	79.6	14.4 (12.2, 16.3)
3	67.3	17 (13, 17)	70.5	14.6 (12.2, 16.3)
4	65.0	17 (13, 17)	64.7	14.7 (12.3, 16.5)
5	74.4	17 (13, 17)	74.5	14.5 (12.3, 16.3)
6	75.8	17 (13, 17)	75.3	14.6 (12.3, 16.4)
7	74.0	17 (13, 17)	72.6	14.5 (12.5, 16.4)
8	44.5	15 (13, 17)	58.4	14.3 (12.5, 16.1)

* Median (90% lower, upper bound); Based on 2000 simulations per row
Power HPC = Probability 95% UB Reg X – HRZE ≤ 12 for 17 wks duration;
Power MCP-Mod = Probability p-value for dose response ≤ 0.025 and 95% UB Reg X – HRZE ≤ 12 for 17 wks duration

HRZE failure rate = 13%, 15 wk Reg X = HRZE: MCP-Mod has higher power than HPC; MCP-Mod slightly underestimates duration

Simulation Results: True Response Profile = EMAX [3]



Scenario	Per group sample size						Total N
	HRZE	9 wks Reg X	11 wks Reg X	13 wks Reg X	15 wks Reg X	17 wks Reg X	
1	70	70	70	70	70	70	420
2	60	60	60	60	60	60	360
3	50	50	50	50	50	50	300
4	50	50	25	50	25	50	250
5	60	60	30	60	30	60	300
6	60	60	20	60	20	60	280
7	60	60	0	60	30	60	270
8	60	60	0	60	60	30	270

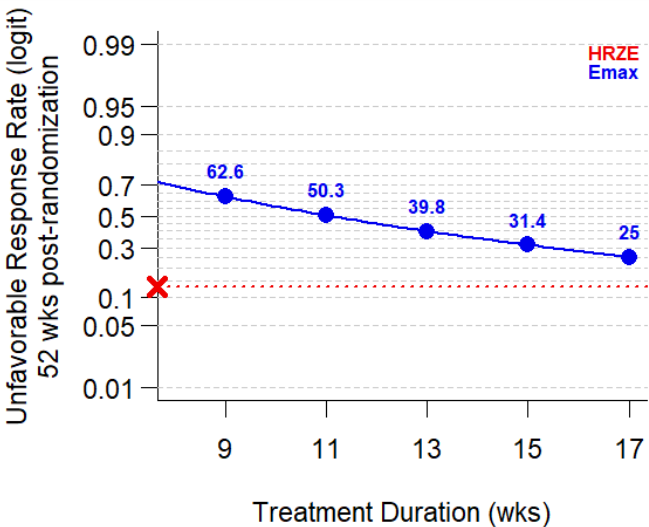
EMAX [3]: HRZE failure rate = 18%, 13 wks Reg X = HRZE

Scenario	Hierarchical Pairwise Comparisons (HPC)		MCP-Mod	
	Power	If Significant, Est. Min. Duration where 95% UB Reg X – HRZE ≤ 12*	Power	If Significant, Est. Min. Duration where 95% UB Reg X – HRZE ≤ 12*
1	95.2	15 (11, 17)	96.3	13.1 (11.2, 15.1)
2	92.2	15 (13, 17)	93.4	13.3 (11.3, 15.4)
3	86.8	15 (13, 17)	88.4	13.5 (11.3, 15.8)
4	85.9	17 (13, 17)	82.8	13.7 (11.4, 16.1)
5	91.4	15 (13, 17)	90.4	13.5 (11.4, 15.7)
6	91.6	17 (13, 17)	90.5	13.6 (11.5, 15.8)
7	90.7	15 (13, 17)	89.2	13.6 (11.9, 15.8)
8	70.3	15 (13, 17)	79.3	13.3 (11.9, 15.3)

* Median (90% lower, upper bound); Based on 2000 simulations per row
 Power HPC = Probability 95% UB Reg X – HRZE ≤ 12 for 17 wks duration;
 Power MCP-Mod = Probability p-value for dose response ≤ 0.025 and 95% UB Reg X – HRZE ≤ 12 for 17 wks duration

HRZE failure rate = 13%, 13 wk Reg X = HRZE: MCP-Mod has higher power than HPC; HPC overestimates duration

Simulation Results: True Response Profile = EMAX [0] (NULL)



Scenario	Per group sample size						Total N
	HRZE	9 wks Reg X	11 wks Reg X	13 wks Reg X	15 wks Reg X	17 wks Reg X	
1	70	70	70	70	70	70	420
2	60	60	60	60	60	60	360
3	50	50	50	50	50	50	300
4	50	50	25	50	25	50	250
5	60	60	30	60	30	60	300
6	60	60	20	60	20	60	280
7	60	60	0	60	30	60	270
8	60	60	0	60	60	30	270

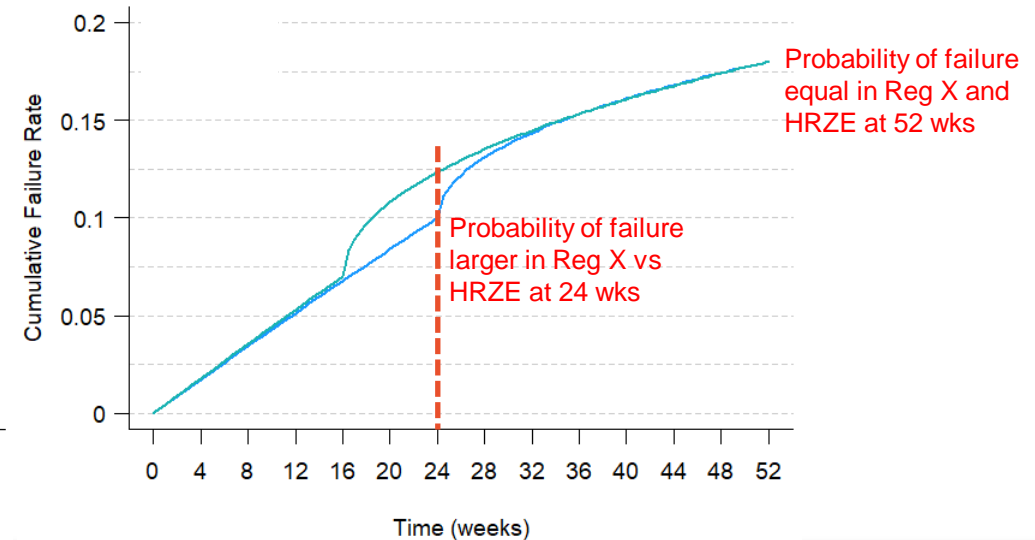
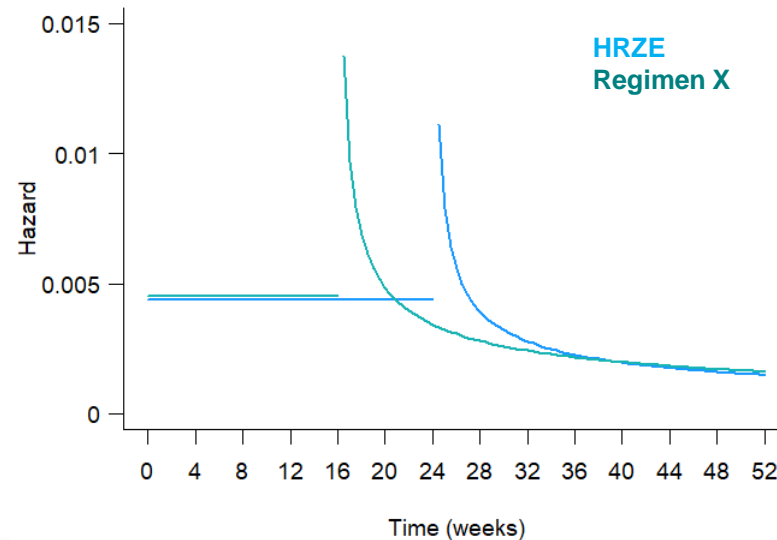
EMAX [0]: HRZE failure rate = 13%, 17 wks Reg X = HRZE + 12 (Null)

Scenario	Hierarchical Pairwise Comparisons (HPC)		MCP-Mod	
	Power	If Significant, Est. Min. Duration where 95% UB Reg X – HRZE ≤ 12*	Power	If Significant, Est. Min. Duration where 95% UB Reg X – HRZE ≤ 12*
1	2.5	17 (17, 17)	1.4	16.6 (15.7, 17.0)
2	2.7	17 (17, 17)	1.6	16.7 (15.9, 17.0)
3	2.5	17 (17, 17)	1.4	16.4 (15.4, 17.0)
4	3.1	17 (17, 17)	2.1	16.6 (16.0, 16.9)
5	3.3	17 (17, 17)	1.6	16.6 (15.4, 16.9)
6	3.0	17 (17, 17)	1.6	16.5 (15.8, 17.0)
7	3.5	17 (17, 17)	2.0	16.6 (15.9, 17.0)
8	1.9	17 (17, 17)	2.2	16.7 (15.8, 17.0)

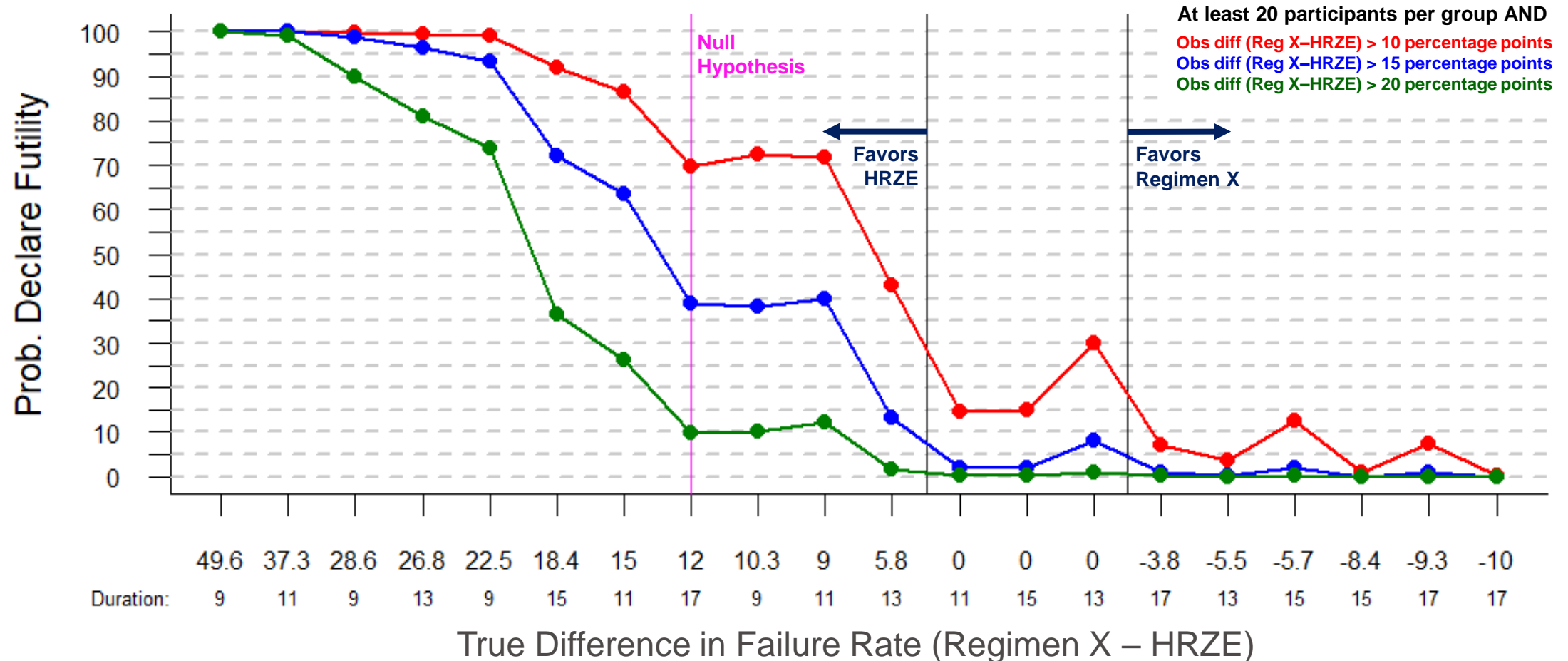
* Median (90% lower, upper bound); Based on 2000 simulations per row
Power HPC = Probability 95% UB Reg X – HRZE ≤ 12 for 17 wks duration;
Power MCP-Mod = Probability p-value for dose response ≤ 0.025 and 95% UB Reg X – HRZE ≤ 12 for 17 wks duration

- Continuous assessment of futility: Count and accrue observed failures in real time
 - Each time a failure occurs, compare probability of failure between DBOS at each duration and HRZE
 - Declare futility if difference in probability of failure (DBOS – HRZE) exceeds a pre-specified threshold (e.g.) 5, 10, 15, or 20 percentage points, when there are at least (e.g.) 20 participants in each group
- Assess operating characteristics of continuous assessment using clinical trial simulation
 - Assumptions: 66 months of accrual, constant hazard (i.e., probability) of failure over treatment period (lower for Reg X vs HRZE), and Weibull hazard (i.e., probability) of failure during follow-up high early and decreases over time

- Example:

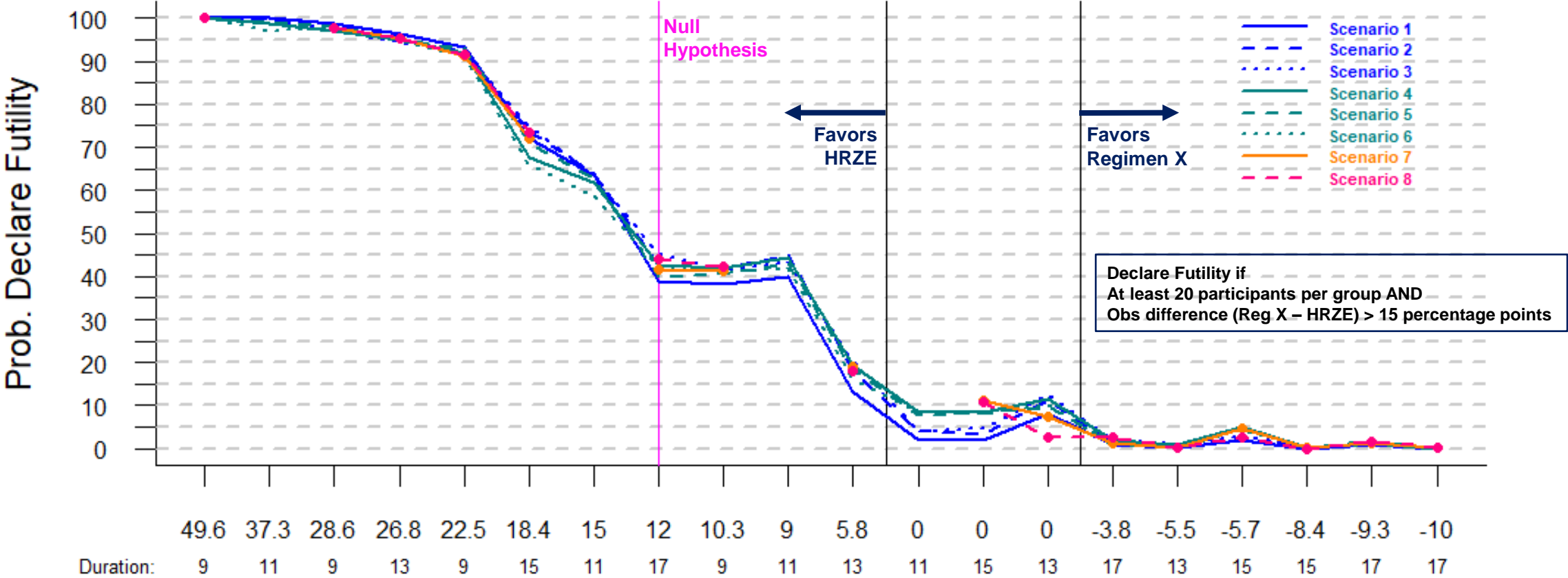


Operating characteristics of continuous monitoring for futility (Scenario 1)



- Using observed difference in failure rates > 15 percentage points as a rule, there is low probability to declare futility when failure rates are equal or favor DBOS, and reasonably high probability to declare futility under the null and when failure rates clearly favor HRZE

Operating characteristics of continuous monitoring for futility



Scenario	Per group sample size						Total N
	HRZE	9 wks Reg X	11 wks Reg X	13 wks Reg X	15 wks Reg X	17 wks Reg X	
1	70	70	70	70	70	70	420
2	60	60	60	60	60	60	360
3	50	50	50	50	50	50	300
4	50	50	25	50	25	50	250
5	60	60	30	60	30	60	300
6	60	60	20	60	20	60	280
7	60	60	0	60	30	60	270
8	60	60	0	60	60	30	270

True Difference in Failure Rate (Regimen X – HRZE)

- No meaningful impact of sample sizes examined on probability of declaring futility with continuous monitoring

- At least 400 – 600 participants required to properly power NI using pairwise comparisons; MCP-Mod allows for smaller overall sample sizes and unbiased estimates of durations to achieve NI
 - MCP-Mod provides framework to rigorously assess NI, enables estimating optimal duration to achieve NI in Phase 3
 - MCP-Mod requires ~300-360 participants; sample sizes are best anchored at 2, 3 and 4-month treatment durations while smaller numbers can be allocated to 2.5 and 3.5-month durations
- Continuous monitoring for futility can be done with low probability of stopping when regimens favor Regimen X and relatively high probability of stopping when regimens clearly favor HRZE
 - Can easily adapt “observed difference” continuous futility rule into one using Bayesian posterior probabilities (e.g., high posterior probability true difference > 10 percentage points)

- Two-stage designs require important statistical considerations if we want to fold Stage 1 data into overall estimation and modeling
 - Data from Stage 1 (4 mos treatment duration) may be biased (potentially overly good for experimental arm and/or overly poor for HRZE) relative to Stage 2 data, because observing Stage 2 data is conditional on Stage 1 looking promising with respect to the difference
 - Statistical methods exist, regulatory acceptance uncertain
 - Many elegant designs can be further examined and considered: evaluation via clinical trial simulation and early regulatory buy-in will be incredibly important

Thank you!