



Marker evaluation tool for de-risking clinical trials

Motivation, examples, and live-demonstration

Stefan Thoma

Team: Claude Berge, Anton Belousov, Daria Rukina, Gian-Andrea Thanei

Who am I

- Background in Statistics and Psychology
- Working at Roche for 3 years
 - Blended role between statistical programming and statistics
 - Mostly Neuroscience
- First time at the Phuse Conference :)



Background: Multiple Sclerosis (MS) and how it is measured

- MS is a chronic autoimmune disease that affects the brain and spinal cord, damaging the protective myelin sheath around nerve fibers.
- Drugs aim to slow the progression of MS.
 - Specifically, they aim to increase the time to some concept of confirmed disease progression (CDP, generally based on EDSS).
- **Disease modifying treatments (DMT's) are available now, slowing the progression.**
- **Great for patients -> Time to CDP takes a lot longer.**
- **Not so great to develop new drugs -> Larger samples or larger follow up required for sufficient study power.**

Composite CDP (cCDP)

What?

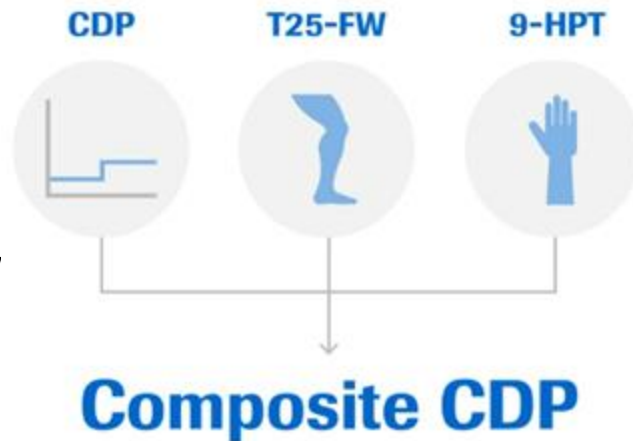
- Composite of three different measures
 - CDP is based on EDSS
 - Timed 25 Foot Walk
 - 9-Hole Peg Test
- Sustained worsening on any of the three is considered an event

Why?

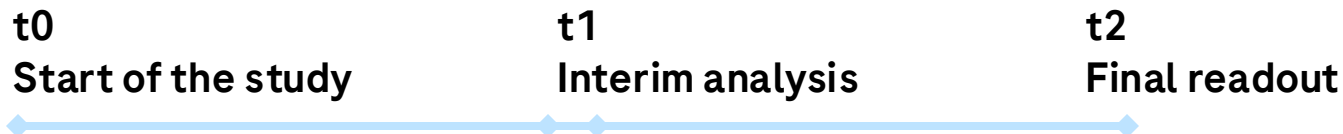
- Captures MS more broadly
- More sensitive – captures more events, more information

But!

- Still time-to-event endpoint
- Still very few events, due to efficacy of DMT's

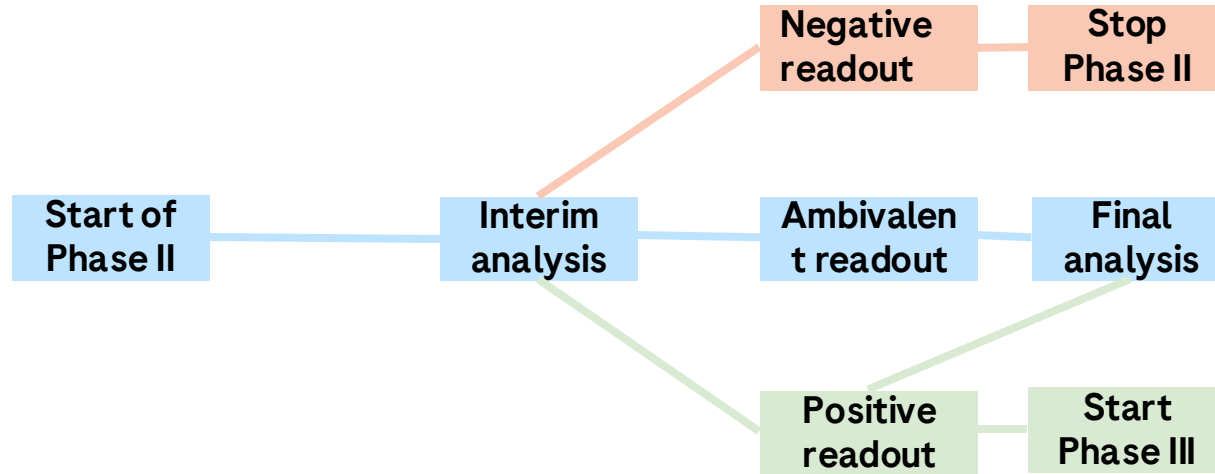


Problem Statement



- cCDP at t2
 - Current clinical endpoint of interest
 - Extracting events from these variables disregards a lot of data
 - Consequently, more data / time is required to get adequate information from endpoint
- We want to find some other criteria to make good decisions after one year (t1)
- The type of decisions, e.g.
 - Stop a Phase II trial early for futility (at t1)
 - Start a Phase III trial after early readout of Phase II trial (at t1)

Early Decisions



Proposal – Intermediate Endpoint (IE)

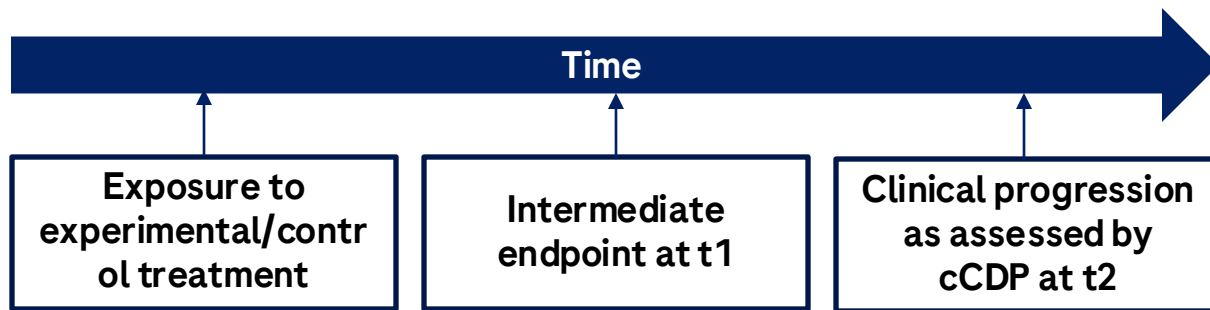
- **Problem:** Clinical time-to-event endpoint with small event rates require large trials.
- **Solution:** Intermediate Endpoint (IE) to de-risk the trial long before the final readout.

- **However:** Intermediate Endpoint must be suitable for de-risking clinical trials.
 - Teams evaluate IE's individually.
 - What makes an IE effective?

- **We propose:** A general framework to assess the suitability of an intermediate endpoint for your trial.
 - Easy to use but flexible shiny app
 - TA independent
 - Unified criteria across the board

What makes an effective intermediate endpoint?

Earlier effect of the drug & correlation with final clinical endpoint



- Early effect of the drug on the intermediate endpoint.
- Detecting an effect on the intermediate endpoint must give us information about the effect of the drug on the clinical endpoint at t2.
 - The intermediate endpoint at t1 must be correlated with the clinical endpoint at t2.

Statistical model

Multivariate normal distribution of endpoints

$$\begin{pmatrix} \log(H\hat{R}_{t2}) \\ \log(H\hat{R}_{t1}) \\ \Delta\hat{I}E_{t1} \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \log(HR) \\ \log(HR) \\ \Delta IE_{t1} \end{pmatrix}, \begin{pmatrix} \sigma^2(\log(H\hat{R}_{t2})) & \dots & \dots \\ \text{cov}(\log(H\hat{R}_{t2}), \log(H\hat{R}_{t1})) & \sigma^2(\log(H\hat{R}_{t1})) & \dots \\ \rho_{12} \cdot \sigma(\log(H\hat{R}_{t2})) \cdot \sigma(\Delta\hat{I}E_{t1}) & \rho_{11} \cdot \sigma(\log(H\hat{R}_{t1})) \cdot \sigma(\Delta\hat{I}E_{t1}) & \sigma^2(\Delta\hat{I}E) \end{pmatrix} \right)$$

- Delta IE must be observable early on
- Correlation between IE and HR is highly relevant, but hard to estimate

What correlation?

Three ways to estimate

Correlation is hard to estimate on a patient level, as time to cCDP will include censored observations.

No Data:

- Join forces with clinical sciences and come up with a reasonable range of plausible correlations.
- Ensure the lower end of the correlation is covered.

One RCT:

- Bootstrap. Draw from RCT and compute IE and HR.
- Repeat many times.
- Compute correlation coefficient on the bootstrapped estimates.
- Include CI for range of values.

Multiple RCTs:

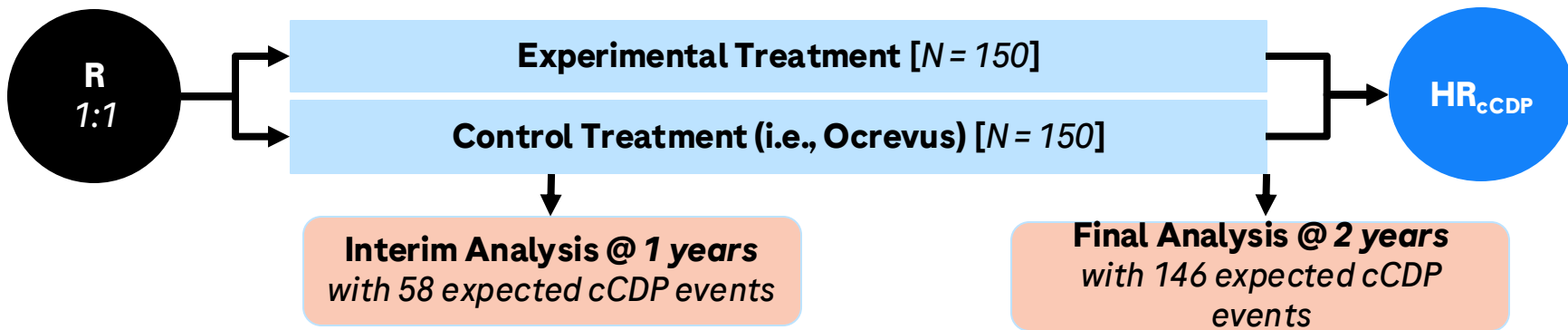
- Compute the IE and the HR on all trials separately.
- Compute correlation coefficient on the trials.
- Include CI for range of values

Potential Interim Endpoints

Variable name	Description
EDSS (clinical)	Expanded Disability Status Scale
wk25ft (clinical)	Timed 25 Foot Walk Test
pg9H (clinical)	Nine-Hole Peg Test
IgNfL (fluid biomarker)	Serum neurofilament light chain protein
CGMA (MRI)	Cortical grey matter atrophy
PBVC (MRI)	Percent brain volume change
T2LESV (MRI)	T2 lesion volume
THALVOLC (MRI)	Thalamic volume change
WMA (MRI)	White matter atrophy
CRBA (MRI)	Cerebellar atrophy
IgT1LESV (MRI)	Non-enhancing T1 lesion volume

Example in Multiple Sclerosis

Neurofilament light chain (NfL) as biomarker



Decision Criteria

1. Having a significant HR result with 2-sided alpha = 0.10 at 4 years without interim (usual study power)
2. Having both interim HR at 1 year and at 2 years significant
(with alpha = 0.20 each, two-sided test for the final analysis and 1-sided for the interim comparing to a threshold of 0.9)
1. Having both change in biomarker at 1 year and HR at 2 years significant
(with alpha = 0.20 each, two -sided for the final analyses, 1 sided for the interim comparing to a threshold of 0.1)

How Useful is Your Biomarker?

Upload parameters from previous analysis

Browse...

No file selected

Trial design

Participants per arm, N



Type of clinical readout

Time to event

Recruitment period in months, t_R



Minimal exposure before marker assessment in months



Study month of interim readout, t_1

12

Study month of final readout, t_2

24

Introduction

Definitions

Inputs

Power $\sim \Delta BM(t_1)$

Power $\sim \rho$

Conditional power

Futility Summary

Authors:

Stefan Thoma (App responsible), Anton Belousov, Claude Berge, Gian-Andrea Thanei, Daria Rukina

Introduction

When it takes a long time to observe the primary endpoint, e.g., a time-to-event (TTE) endpoint with long follow-up, utilizing information on an early intermediate endpoint or a surrogate readout at interim analysis (IA) can be more efficient for drug development. For instance, observing confirmed/sustained disability progression in subjects with multiple sclerosis based on a composite primary endpoint may take several years. In such a situation it is appealing to use outcomes with relatively short follow-up that are closely related to the primary endpoint for early decision-making, e.g., moving a drug development program into phase 3 or stopping for futility.

Within a clinical trial setting, a surrogate endpoint is a marker - a laboratory measurement, radiographic image, physical sign or other measure that can be used instead of the main clinical outcome of interest in the evaluation of an experimental treatment. While the treatment effect associated with a surrogate endpoint might not be of any direct value to a patient, it can be used to predict the corresponding clinical benefit that would have been addressed by the key clinical outcome.

In situations where the evidence linking the surrogate to the main clinical outcome is not deemed sufficient for regulatory acceptance, it could still be very valuable in non-confirmatory phases of drug development. This is particularly the case, when the primary clinical endpoint is costly to obtain or requires a substantial amount of time to observe. However, due to the increased uncertainty associated with the use of a surrogate, additional risks are inherent in development decision-making.

In drug development, each decision on whether to continue with the development program for a given drug is associated with a risk of false positive or false negative choice. Every decision, correct or otherwise, has an associated cost and possibly a net reward. For any stage in the process, use of a surrogate endpoint rather than the clinical outcome may reduce costs but increase the chance of an incorrect decision to stop or continue developing the compound. The risks of using surrogates include an increased risk of making an incorrect decision regarding stopping or continuing development of a particular compound. Use of a valid surrogate may yield benefits in terms of reduced costs and duration of individual segments of the development process. Therefore, it is essential to gain a quantitative understanding of the evidence linking the surrogate to the main outcome of interest that can then be utilized in the design of clinical studies and the formulation of development decision criteria.

The proposed tool should assist in considering and weighting these risks and benefits in a formal decision analysis framework to provide informed guidance of drug development. In this framework a theoretical model is utilized to estimate and compare the conditional (at interim) and the final power (or the probability of success, PoS) of a PhII / proof of concept or PhIII / confirmatory study. The following decision rules are considered:

- Standard assessment at the end of the study using clinical readout

Trial design

Participants per arm



Type of clinical readout

Time to event

Recruitment period in months



Minimal exposure before months



Study month of interim readout, t_1

12

desc

Total N subjects with t_1

N events at t_1 (H_0)	14.00
N events at t_2 (H_0)	54.00
N events at t_1 (H_1)	12.00
N events at t_2 (H_1)	47.00

Hypotheses

Null hypothesis (H_0)

Fraction of events

0.5

Alternative hypothesis (H_1)

Hazard ratio (HR, treatment vs control)



Expected differential of the biomarker at t_1



Expected range of correlation of biomarker at t_1 and t_2

$$\rho = \text{Corr}(\Delta BM(t_1), HR(t_2))$$

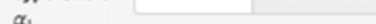


Decision rules

Type one error at the final analysis, α_2



Type one error at interim, α_1



Hazard ratio at interim



Biomarker threshold at interim



Upload parameters from previous analysis

Browse...

No file selected



Download Parameters

Run Analysis

Trial design

Participants per arm, N 

Type of clinical readout

Time to event

Recruitment period in months, t_R 

Minimal exposure before marker assessment in months

Study month of interim readout, t_1

12

Study month of the final readout, t_2

24

desc

value

Total N subjects with biomarker data at t1

50.00

N events at t1 (H0)

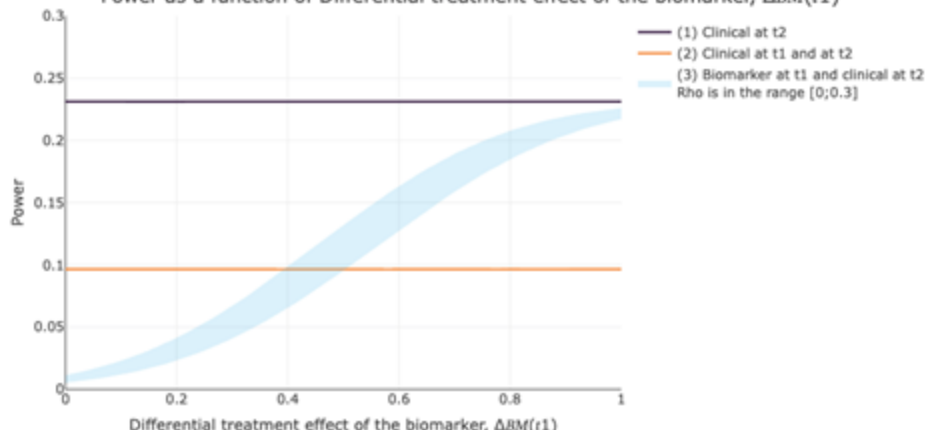
14.00

N events at t2 (H0)

54.00

N events at t2 (H1)

17.00

Power as a function of Differential treatment effect of the biomarker, $\Delta BM(t_1)$ Differential treatment effect of the biomarker, $\Delta BM(t_1)$

The graph reports the resulting power (Y axis) of several complete decision strategies (with and without including interim assessment; with and without evidence from the surrogate at interim time point) as function of signal to noise ratio (standardized differential treatment effect, SNR) $\Delta BM(t_1)$ of a biomarker (X axis):

The following descriptions illustrate the key features of the graph:

- Black line corresponds to the power of the study that utilizes only the final clinical assessment at t_2 . This is the best possible power that can be achieved and it does not depend on the properties of biomarker readout as it does not use the latter.
- Orange line corresponds to the power of the study that utilizes both the interim (at t_1) and the final clinical assessments (at t_2).
- Shaded blue region represents the decision rule that combines biomarker readout at t_1 and the final clinical assessment at t_2 . The width of this region reflects the range of correlations between biomarker and clinical readouts $\rho = \text{Corr}(\Delta BM(t_1), HR(t_2))$.

The opportunity window for utilizing a BM readout at t_2 as a surrogate of the final efficacy measure at t_2 appears at biomarker signals (SNR) where the power of the decision [BM at t_1 & Clinical at t_2] = the blue curves, exceeds the power of the design plan [Clinical at t_1 & Clinical at t_2], see the orange line.

Note that as we change the value of differential treatment effect at interim (OX axis) the value of the final response (HR) readout stays the same and corresponds to the pre-specified value under the alternative hypothesis. Supposed to be constant is also the strength of correlation between the interim biomarker and final response readouts. This scenario does not reflect the (expected) situation when the interim biomarker and the final clinical readout are interdependent via a particular functional form. So, this plot should be understood as follows: Every point at OX axis re-defines the H1 hypothesis by (re-)setting the expected value of the differential BM readout at t_1 at the otherwise fixed values of $HR(t_2)$ and ρ .

Trial design

Participants per arm, N 

Type of clinical readout

Time to event

Recruitment period in months, t_R 

Minimal exposure before marker assessment in months

Study month of interim readout, t_1

12

Study month of the final readout, t_2

24

desc

value

Total N subjects with biomarker data at t_1

50.00

N events at t_1 (H_0)

14.00

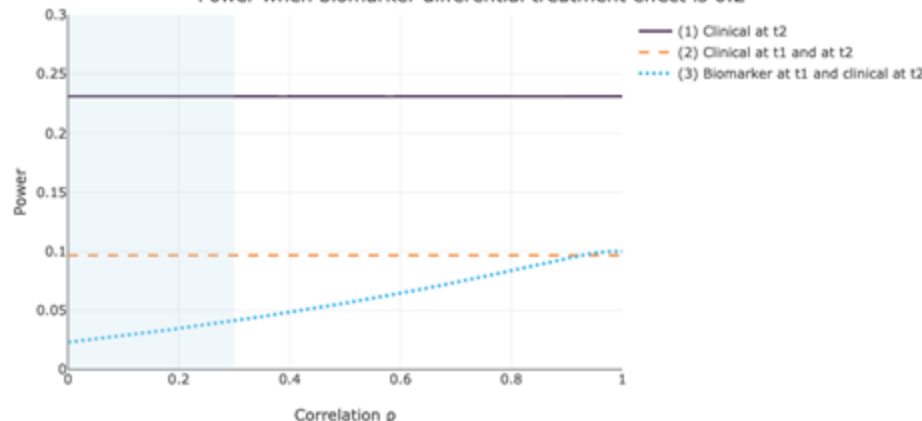
N events at t_2 (H_0)

54.00

N events at t_1 (H_1)

14.00

Power when biomarker differential treatment effect is 0.2



The graph reports the resulting power (Y axis) of several decision strategies (with and without including interim assessment; with and without evidence from the surrogate at interim time point) as a function of study-specific correlation between biomarker at interim and final clinical readouts $\rho = \text{Corr}(\Delta BM(t_1), HR(t_2))$ (X axis):

The following descriptions illustrate the key features of the graph:

- Black line corresponds to the power of the study that utilizes only the final clinical assessment at t_2 . This is the best possible power that can be achieved and it does not depend on the properties of biomarker readout as it does not use the latter.
- Orange line corresponds to the power of the study that utilizes both the interim (at t_1) and the final clinical assessments (at t_2).
- Blue line represents the decision rule that combines biomarker readout at t_1 and the final clinical assessment at t_2 . The curve corresponds to the particular, pre-specified (H_1 hypothesis) strength of differential treatment effect $\Delta BM(t_1)$ for biomarker readout at t_1 .
- The pre-specified range of correlation strength ρ is shaded.

Note that as we change the value of correlation (OX axis) the value of the differential biomarker effect at interim and the value of the final response (HR) readout stay the same and correspond to the pre-specified values under the alternative hypothesis. This scenario does not reflect the (expected) situation when the interim biomarker and the final clinical readout are interdependent via a particular functional form. So, this plot should be understood as follows: Every point at OX axis re-defines the H_1 hypothesis by (re-)setting the expected value of ρ at the otherwise fixed values of $HR(t_2)$ and differential BM readout at t_1 .

Trial design

Participants per arm, N 

Type of clinical readout

Time to event

Recruitment period in months, t_R 

Minimal exposure before marker assessment in months

Study month of interim readout, t_1

12

Study month of final readout, t_2

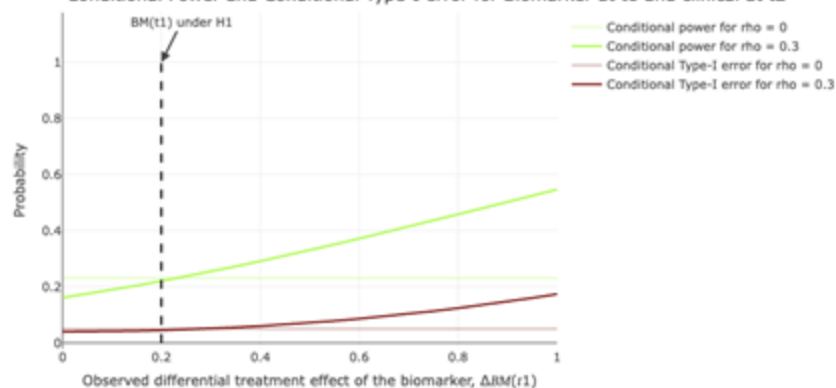
24

desc	value
Total N subjects with biomarker data at t1	50.00
N events at t1 (H0)	14.00
N events at t2 (H0)	54.00
N events at t1 (H1)	12.00
N events at t2 (H1)	47.00

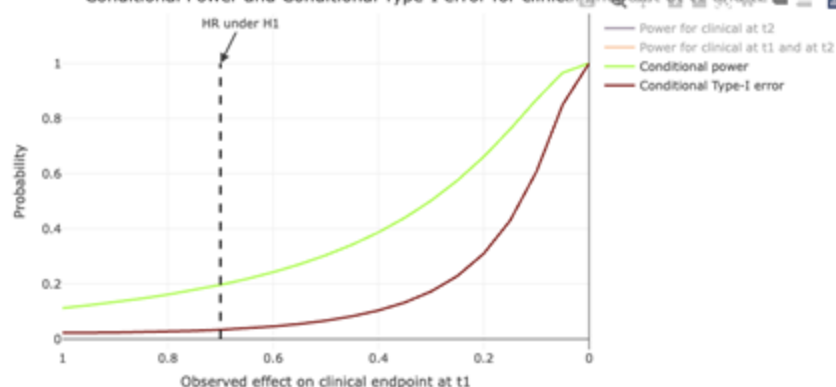
The graph reports on the Y axis the conditional power (CP), the probability to correctly declare trial success at t_2 given observed BM differential treatment effect at t_1 (X axis). Put another way: conditional on the interim readout power, CP, is the probability of the null hypothesis of no effect at final being rejected when the trial continues to its planned final readout and under the assumption of the unchanged treatment effect.

This conditional power corresponds to the original, unchanged formulation of the alternative hypothesis H1 (expected effect sizes at t_1 and t_2).

Conditional Power and Conditional Type-I error for Biomarker at t1 and clinical at t2



Conditional Power and Conditional Type-I error for clinical endpoint at t1 and t2



Upload parameters from previous analysis

Browse...

No file selected

Trial design

Participants per arm, N 

Type of clinical readout

Time to event

Recruitment period in months, t_R 

Minimal exposure before marker assessment in months

Study month of interim readout, t_1

12

Study month of the final readout, t_2

24

desc

value

Total N subjects with biomarker data at t1

50.00

N events at t1 (H0)

14.00

N events at t2 (H0)

54.00

N events at t1 (H1)

12.00

N events at t2 (H1)

47.00

Futility using (bio)marker threshold

(Bio)marker futility thresholds

0.5 0.2

Threshold	ρ	Power w/o Futility	CP	Power loss
0.20	0.00	0.23	0.23	0.21
0.20	0.30	0.23	0.22	0.19
0.50	0.00	0.23	0.23	0.23
0.50	0.30	0.23	0.33	0.23

Futility using clinical threshold

Clinical futility thresholds

0.5 0.2

Threshold	ρ	Power w/o Futility	CP	CP(d)	Power loss
0.20	0.05	0.23	0.66	0.47	0.23
0.50	0.05	0.23	0.30	0.27	0.21

Conclusion

General

- Drug development is getting harder.
- More sensitive endpoints are required
 - Ideally at final readout
 - At least for early (internal) decision making

Application

- Application makes different endpoints comparable.
- Framework covers most important operating characteristics.
- Framework is extensible.
- Small details make the application much more user friendly.