Decision Making in Early Stage Clinical Trials with Multiple Endpoints

Master's thesis in collaboration with AstraZeneca

Supervisors:
Marcus Millegård
Karin Nelander
Examiner: José Sánchez
Opponent: William Nilsén

Víctor López Juan

UNIVERSITY OF GOTHENBURG

Wednesday 27th April, 2022

AstraZeneca

Phases of clinical trials

	Phase I	Phase II		Phase III	
		IIa	IIb	III	NDA
Main focus	Safety	Dose finding, effectiveness		Approval	
Participants	Healthy volunteers	Afflicted patients			
	10s	10s-100s		100s-1000s	
Duration	Months	Months – 2 years		1–4 years	1 year
Cost (MUSD)	1–6	7–20		12–50	2
Success rate	64%	32%		60%	85%
					10%

Sources: FDA.gov (2018), Chow et al. (2003), Berlink et al. (2004), Hay et al. (2014)

- Randomized controlled trials: treatment and control arms
- Phase III: Large, costly, leads to approval.
- Phase II: Preliminary, smaller.

When do Phase II results justify a Phase III trial?

Surrogate endpoints and disease markers

- **Phase III endpoint**: Indication-specific, required for approval.

Example (Phase III endpoint for a heart drug)

For a heart drug, MACE(HR), i.e. Major Adverse Cardiovascular Events (Hazard Ratio)

$$\text{MACE(HR)} \approx \frac{\#\text{MACE in treatment arm}}{\#\text{MACE in control arm}}$$

- Ph3 endpoint requires large study $\Rightarrow$ Use surrogate endpoint for Ph2
- Not one clear surrogate $\Rightarrow$ Combine several disease markers across multiple domains:

Example (Phase II disease markers for heart disease by domain)

- Blood biomarkers domain: NT-proBNP
- Exercise capacity domain: VO2max, 6MWD
- Well-being domain: KCCQ-TSS
- Heart imaging domain: LVMI, LAVI, GLS, LVEF

How to combine multiple disease markers into a single decision? (Go/No-go)

A decision framework proposed by Lalonde (2007) for one endpoint

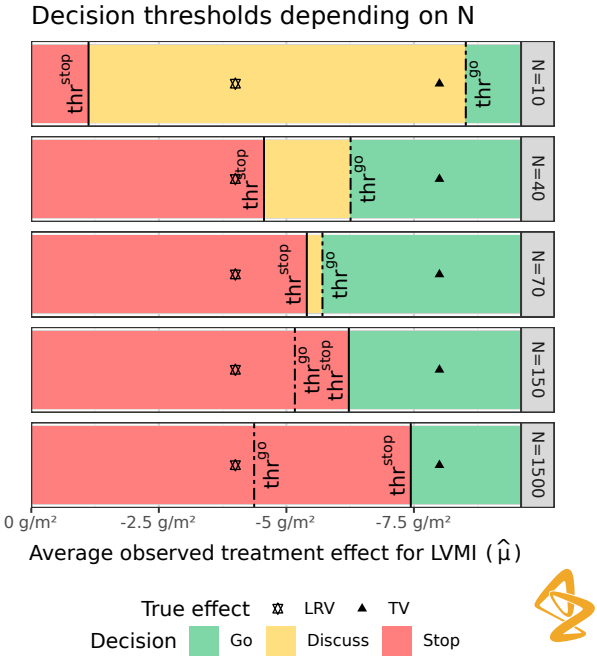
Example (Disease marker)

Left Ventricular Mass Index (LVMI)

- $\mu \in \mathbb{R}$: True effect
- TV: Target Value (desired)
- LRV: Lower Reference Value (clinically relevant)
- $0 < \text{LRV} < \text{TV}$ (negate otherwise)
- N: Patients per arm
- σ: Standard deviation per patient
- $\hat{\mu} \sim \mathcal{N}\left(\mu, \frac{2\sigma^2}{N}\right)$: Observed effect (wrt. placebo)
- thr^{go}, thr^{stop}: Thresholds

(i) If $\mu \leq \text{LRV}$, $\mathbb{P}(\hat{\mu} \geq \text{thr}^{\text{go}}) \leq 20\%$

(ii) If $\mu \geq \text{TV}$, $\mathbb{P}(\hat{\mu} \leq \text{thr}^{\text{stop}}) \leq 10\%$



Combining multiple endpoints from the same domain

Example (from the "Imaging" domain)

LVMI and GLS (Global Longitudinal Strain)

- $\boldsymbol{\mu} \in \mathbb{R}^V$: True effect
- $\hat{\boldsymbol{\mu}} \sim \mathcal{N}\left(\boldsymbol{\mu}, \frac{2}{N}\Sigma\right)$: Observed effect

$$G(\hat{\boldsymbol{\mu}}) = \begin{cases} \text{stop} & \text{all } \hat{\mu}_i \leq \text{thr}^{\text{stop}}_{10\%,i} \\ \text{go} & \ldots \end{cases}$$

(i) $\mathbb{P}(\text{stop} \mid \text{some } \mu_i \geq \text{TV}_i) \leq 10\%$

(ii) $\mathbb{P}(\text{go} \mid \text{all } \mu_i \leq \text{LRV}_i) \leq 20\%$

How to choose "..." so that (i) and (ii) hold?

Two-variable case

$$\begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \frac{2}{N}\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right) \quad \begin{matrix} (\rho = 0.4, \\ N = 17) \end{matrix}$$

One variable case

Observed effect: $\hat{\mu}_1 \sim \mathcal{N}\left(\mu_1, \frac{2}{N}\sigma^2\right)$

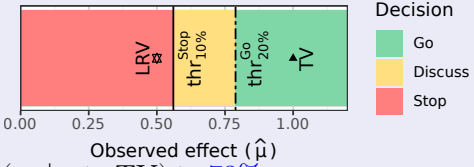
$$G(\hat{\boldsymbol{\mu}}) = \begin{cases} \text{stop} & \hat{\mu}_1 \leq \text{thr}^{\text{stop}}_{10\%} \\ \text{go} & \hat{\mu}_1 \geq \text{thr}^{\text{go}}_{20\%} \end{cases}$$

$\mathbb{P}(\text{stop} \mid \mu_1 \geq \text{TV}) \leq 10\%$

$\mathbb{P}(\text{go} \mid \mu_1 \leq \text{LRV}) \leq 20\%$

Synthetic endpoints

$\text{TV} = 1$, $\text{LRV} = 0.5$, $\sigma = 1$, $N = 17$

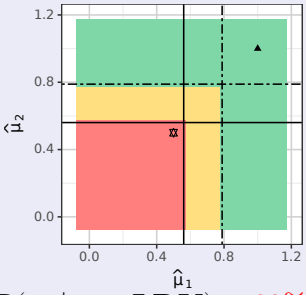


$\mathbb{P}(\text{go} \mid \mu \geq \text{TV}) \geq 73\%$

$G(\hat{\mu}) = \text{go}$ conditions for the two variable case

Disjunction

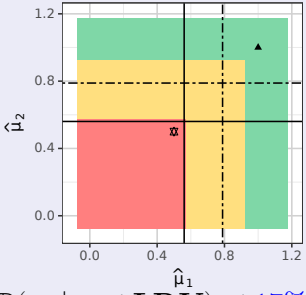
Either $\hat{\mu}_1$ or $\hat{\mu}_2 \geq \text{thr}^{\text{go}}_{20\%}$



$\mathbb{P}(\text{go} \,|\, \boldsymbol{\mu} \leq \mathbf{LRV}) \leq 32\%$
$\mathbb{P}(\text{go} \,|\, \boldsymbol{\mu} \geq \mathbf{TV}) \geq 88\%$
$\mathbb{P}(\text{go} | \mu_1 \text{ or } \mu_2 \geq \text{TV}) \geq 73\%$

Bonferroni correction

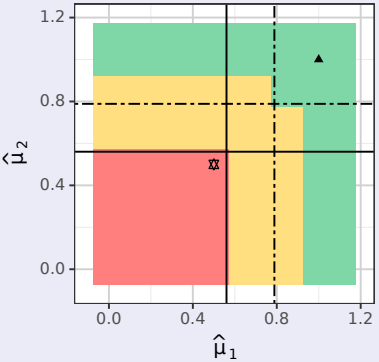
Either $\hat{\mu}_1$ or $\hat{\mu}_2 \geq \text{thr}^{\text{go}}_{10\%}$



$\mathbb{P}(\text{go} \,|\, \boldsymbol{\mu} \leq \mathbf{LRV}) \leq 17\%$
$\mathbb{P}(\text{go} \,|\, \boldsymbol{\mu} \geq \mathbf{TV}) \geq 74\%$
$\mathbb{P}(\text{go} | \mu_1 \text{ or } \mu_2 \geq \text{TV}) \geq 57\%$

Simes/Benjamini-Hochberg

Either $\hat{\mu}_1$ or $\hat{\mu}_2 \geq \text{thr}^{\text{go}}_{10\%}$;
or both $\hat{\mu}_1$ and $\hat{\mu}_2 \geq \text{thr}^{\text{go}}_{20\%}$



$\mathbb{P}(\text{go} \,|\, \boldsymbol{\mu} \leq \mathbf{LRV}) \leq 19\%$
$\mathbb{P}(\text{go} \,|\, \boldsymbol{\mu} \geq \mathbf{TV}) \geq 77\%$
$\mathbb{P}(\text{go} \,|\, \mu_1 \text{ or } \mu_2 \geq \text{TV}) \geq 57\%$

Conclusion

- Multiple comparisons should be accounted for.
- Improves over one endpoint ($\mathbb{P}(\text{go} | \mu_1 \geq \text{TV}) = 73\%$).

When is a disease marker informative?

- Clinically (assumed): Effective (resp. ineffective) in Ph3 $\leftrightsquigarrow$ Marker $\mu = \mathrm{TV}$ (resp. $\mu = 0$)
- Statistically informative: Low variance/High statistical power

Ph3 endpoint itself is clinically informative (MACE), but not necessarily statistically.

Definition (Significant at $\alpha = 0.05$)

$\underline{\hat{\mu} \geq \mathrm{thr}^{\mathrm{sig}}_{2.5\%}}$ or $\underline{\hat{\mu} \leq -\mathrm{thr}^{\mathrm{sig}}_{2.5\%}}$, with
$\mathbb{P}(\hat{\mu} \geq \mathrm{thr}^{\mathrm{sig}}_{2.5\%}$ or $\hat{\mu} \leq -\mathrm{thr}^{\mathrm{sig}}_{2.5\%} \mid \mu = 0) = 5\%.$

Definition (Power)

$\mathbb{P}(\hat{\mu} \geq \mathrm{thr}^{\mathrm{sig}}_{2.5\%} \mid \mu = \mathrm{TV}) \in [0, 1]$
Reminder: $\hat{\mu} \sim \mathcal{N}(\mu, \frac{2\sigma^2}{N})$

- Power increases with N, decreases with increasing σ.

	all cases $\mu = 0$	Power $= 0.8$ $\mu = \mathrm{TV}$	Power $= 0.2$ $\mu = \mathrm{TV}$	Power $= 0.06$ $\mu = \mathrm{TV}$
$\mathbb{P}(\hat{\mu} \geq \mathrm{thr}^{\mathrm{sig}}_{2.5\%})$	2.5%	80%	20%	6%
$\mathbb{P}(\hat{\mu} < 0)$	50%	0.2%	11%	31%

- Low power $\Rightarrow$ Endpoint becomes noise.

Considering an additional endpoint in the same domain

Example

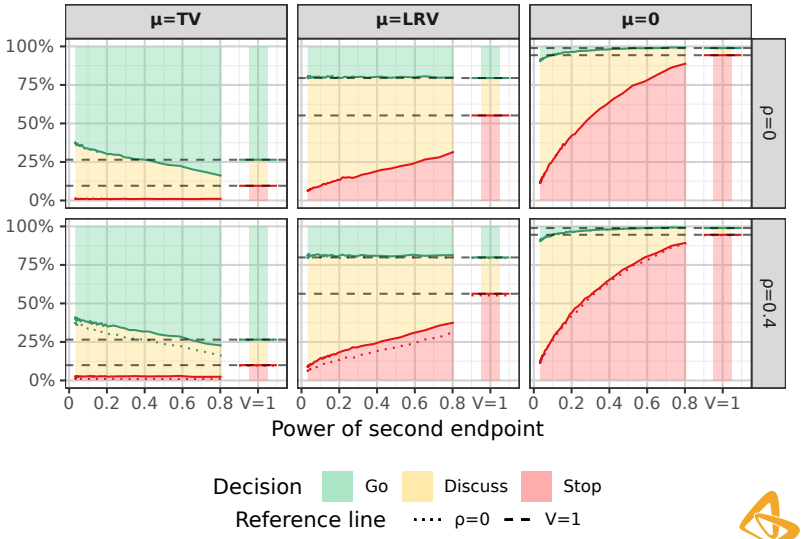
Include one or two exercise endpoints? (e.g. 6MWD, VO2max)

Setup

- Simes method
- **Endpoints:**
  Synthetic
  1st: Power $= 0.8$
  2nd: Power ≤ 0.8

Conclusions

- Low power $\Rightarrow$ Lower $\mathbb{P}(\text{Go} \mid \boldsymbol{\mu} = \mathbf{TV})$
- Lower $\mathbb{P}(\text{Stop} \mid \boldsymbol{\mu} = \mathbf{0})$



Decision policies for endpoints in different domains

Example

3 domains	4 endpoints (1,1,2)
(a) Events	$(i = 1)$ MACE (Ph3)
(b) Biomarker	$(i = 2)$ NT-proBNP
(c) Exercise	$(i = 3)$ VO2Max, $(i = 4)$ 6MWD

Definition (Policy implem.)

$G : (\mathbb{R}^V, \leq^V) \to (\text{Decision}, \leq)$
$G(\hat{\boldsymbol{\mu}}) \mapsto (\text{Stop} \leq \text{Discuss} \leq \text{Go})$

1. Obtain one decision for each domain (as discussed)
2. Combine domain-level decisions into an overall decision (respecting definition).

Domain-level decisions: $G_a(\hat{\mu}_1), G_b(\hat{\mu}_2), G_c(\hat{\mu}_3, \hat{\mu}_4) \in \{\text{Go}, \text{Discuss}, \text{Stop}\}$

Overall decision: $G(\hat{\boldsymbol{\mu}}) = \begin{cases} \text{go} & 2 \text{ or more domains are Go} \\ & \quad \text{and no significant negative effects} \\ \text{stop} & 0 \text{ domains are Go} \end{cases}$

Significant negative effect: $\hat{\mu}_i \leq -\text{thr}^{\text{sig}}_{5\%,i}$

Including a new endpoint in a domain of its own

Example

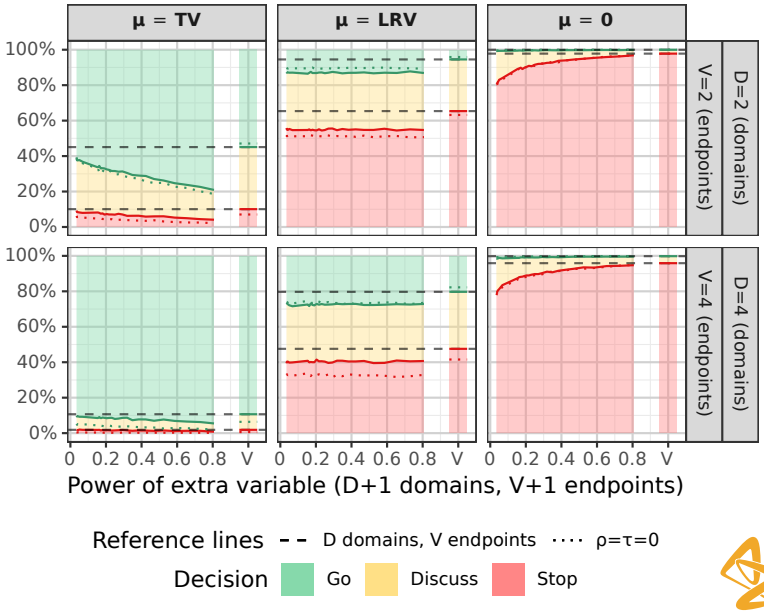
Include Ph3 endpoint?

Setup

$$\begin{cases} \text{go} & \geq 2 \text{ domains Go} \\ & \text{and no sig. neg.} \\ \text{stop} & 0 \text{ domains Go} \end{cases}$$

Endpoints: Synthetic
Base: $V{=}D$, Power $= 0.8$
New $(+1)$: Power ≤ 0.8
Correlation: Same domain: $\rho = 0.4$; Other: $\tau = 0.2$

- High D, low power $\Rightarrow$ Diminishing benefit
- Low power $\Rightarrow$ Same drawbacks for all D



Impact of a domain for which the drug has no effect

Example

Drug shows no effect in exercise endpoints.

Setup

V endpoints: Power $= 0.8$

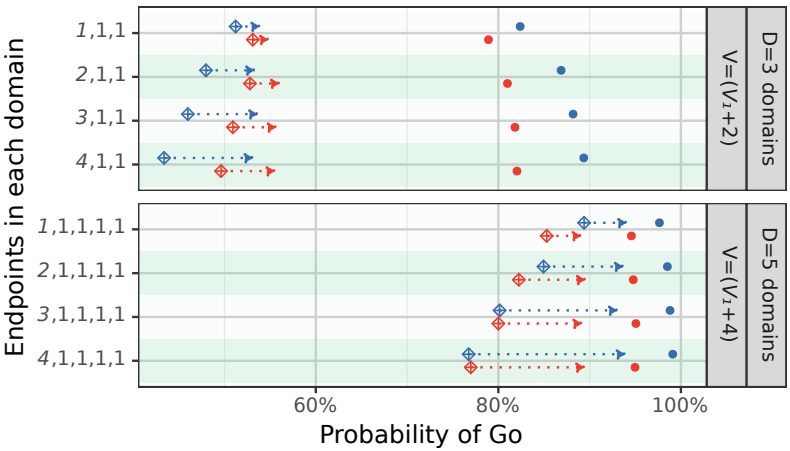
"with.cond": $G(\hat{\boldsymbol{\mu}}) =$

$$\begin{cases} \text{go} & \geq 2 \text{ domains Go} \\ & \text{and no sig. neg.} \\ \text{stop} & 0 \text{ domains Go} \end{cases}$$

"w/o.cond": $G(\hat{\boldsymbol{\mu}}) =$

$$\begin{cases} \text{go} & \geq 2 \text{ domains Go} \\ \text{stop} & 0 \text{ domains Go} \end{cases}$$

- Use fewer endpoints in domains with **0** effect.



Putting it together: a case study

5 domains	9 endpoints
Events	MACE (Ph3)
Biomarker	NT-proBNP
Exercise	VO2Max, 6MWD
Imaging	LAVI, LVEF, LVMI, GLS
Well-being	KCCQ-TSS

Policy

$$G(\hat{\boldsymbol{\mu}}) = \begin{cases} \text{Go} & \geq 2 \text{ domains are Go} \\ & \text{and no sig. neg.} \\ \text{Discuss} & 1 \text{ domain is Go} \\ \text{Stop} & 0 \text{ domains are Go} \end{cases}$$

Desired probabilities

True effect ($\boldsymbol{\mu}$)	Probability
All endpoints $\geq \mathbf{TV}$	$\mathbb{P}(\text{Go}) \geq 90\%$
All endpoints in 4 domains (inc. Ph3) $\geq \mathbf{TV}$	$\mathbb{P}(\text{Go}) \geq 80\%$
All endpoints $\leq \mathbf{0}$	$\mathbb{P}(\text{Stop}) \geq 80\%$
All endpoints $\leq \mathbf{0}$	$\mathbb{P}(\text{Go}) \leq 5\%$
...	...

Questions

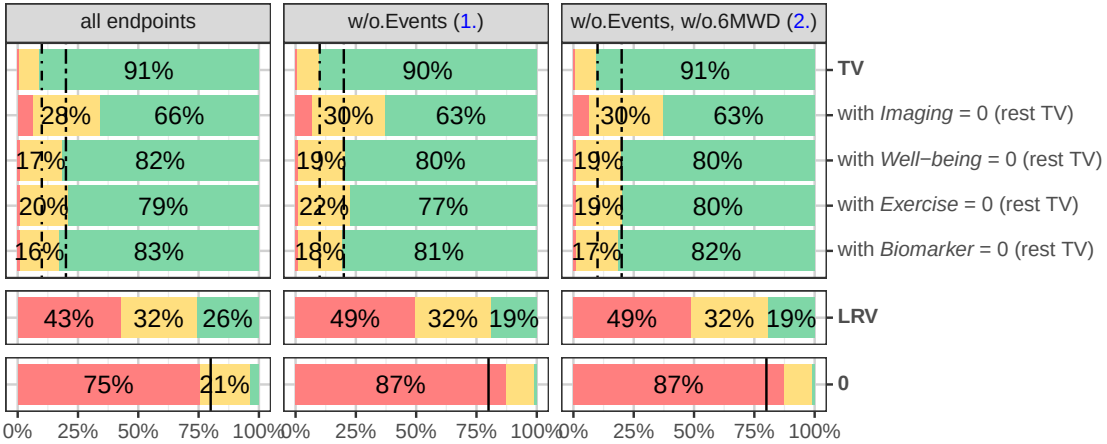
(i) How many patients per arm (N)?

(ii) Which endpoints to include?

Procedure

1. Initial number of patients ($N = 155$).
2. Remove endpoints (step-wise).

Endpoint selection by step-wise removal



Threshold

- 90% Go
- 80% Go
- 80% Stop

Decision

- Go
- Discuss
- Stop

1. Events: very low power ($\ll 0.1$) $\Rightarrow$ Remove
2. 6MWD: Adds complexity, lower power than VO2max $\Rightarrow$ Remove
3. Imaging (4 endpoints): Important; but $\mathbb{P}(\text{Go}|\text{Imaging is } 0)$ is low partly due to "no sig. neg." condition (see Chapter 4) $\Rightarrow$ Perhaps include fewer Imaging endpoints

Limitations/Future work

- Allowed conditions: e.g. go if $0 \leq \text{thr}^{\text{go}} \leq \hat{\mu}$;
  (or, with transformation, go if $\hat{\mu} \leq \text{thr}^{\text{go}} < 0$)
  Out of scope: Range conditions on observed effects (e.g. go if $\text{thr}_1^{\text{go}} \leq \hat{\mu} \leq \text{thr}_2^{\text{go}}$)
- Data assumed normal (log-normal using transformation, HR approximated)
  Out of scope: Non-normal data (e.g. time-to-event)
- No missing/incomplete data
- Restricted to Lalonde-style thresholds

Conclusions

- The decision framework by Lalonde can be extended to multiple endpoints (but with care!)
- Multiple comparisons need to be accounted for; Simes/Benjamini-Hochberg gains from additional endpoints in a domain.

Some endpoints may decrease probability of correct $\mathrm{Go}/\mathrm{Stop}$ decision:

- Low power endpoints $\Rightarrow$ Remove if Power $\ll 0.2$, or if costly and redundant.
- Endpoints potentially unaffected by an otherwise effective drug $\Rightarrow$ Minimize redundancies.

Also in the thesis

- Hierarchical policies.
- Covariance matrix based on data.
- Curvilinear thresholds (e.g. Hotelling's T^2, measures).
- More simulations.

See `https://lopezjuan.com/project/gonogo`

Acronyms

- MACE (Major Adverse Cardiovascular events): For instance, hospitalizations or death due to heart disease.
- LVMI (Left Ventricular Mass Index): Mass of left ventricle relative to patient's surface area.
- GLS (Global Longitudinal Strain): Change in length of heart muscle during systole.
- LAVI (Left-atrial Volume Index): Volume of the left atrium relative to the patient's surface area.
- LVEF (Left-ventricular Ejection Fraction): Percentage of blood volume collected in the left ventricle that is ejected during diastole.
- 6MWD (6-minute walking distance): Distance walked by the patient in 6 minutes in a controlled setting. VO2max (Volume of molecular oxygen, maximum): Peak oxygen consumption during exercise of increasing intensity.
- KCCQ-TSS (Kansas City Cardiomyopathy Questionnaire - Total Symptom Score): Self-reported well-being by the patient using a standardized questionnaire.
- NT-proBNP: A hormone precursor which is predictive of heart failure.