Group Sequential Enrichment Design incorporating subgroup selection

Application to the optimization of the cutoff for a continuous predictive biomarker, with time-to-event endpoints

Émilie Gérard, Loïc Darchy

Group sequential enrichment design incorporating subgroup selection
- B.P. Magnusson and B.W. Turnbull, *Statistics in Medicine* (2013).

27th November, 2014

SANOFI

Contents

1. Setting the scene
2. Some results
3. Discussion

Contents

1 Setting the scene

2 Some results

3 Discussion

Clinical objective

to target the subpopulation who is more likely to respond to a given treatment

Why ?

- giving the right medicine to the right patient
- saving patients safety, costs and time

How ?

- Stage 1: removing the non responsive subsets at the interim analysis
- Stage 2: recruiting patients only in the remaining subsets

Publication's framework

➜ Partition of K disjoint pre-specified subsets ($K \geq 2$)

➜ Two distinct settings:

- no a priori ordering of the subsets,
- with a priori ordering of the subsets. Example with 4 subsets:

a priori ordered

Noting θ_i the treatment effect in subset i, $\theta_4 \geq \theta_3 \geq \theta_2 \geq \theta_1$.

➜ Assumption made by the authors:

- there is no opposite treatment effects across subsets
  - ➜ $\theta_4 \geq \theta_3 \geq \theta_2 \geq \theta_1 \geq 0$.

Our adaptation

➜ Two-stage study with time-to-event endpoints

➜ Subsets defined thanks to a continuous predictive biomarker

But no available cutoff value to distinguish responsive from unresponsive patients to a given treatment

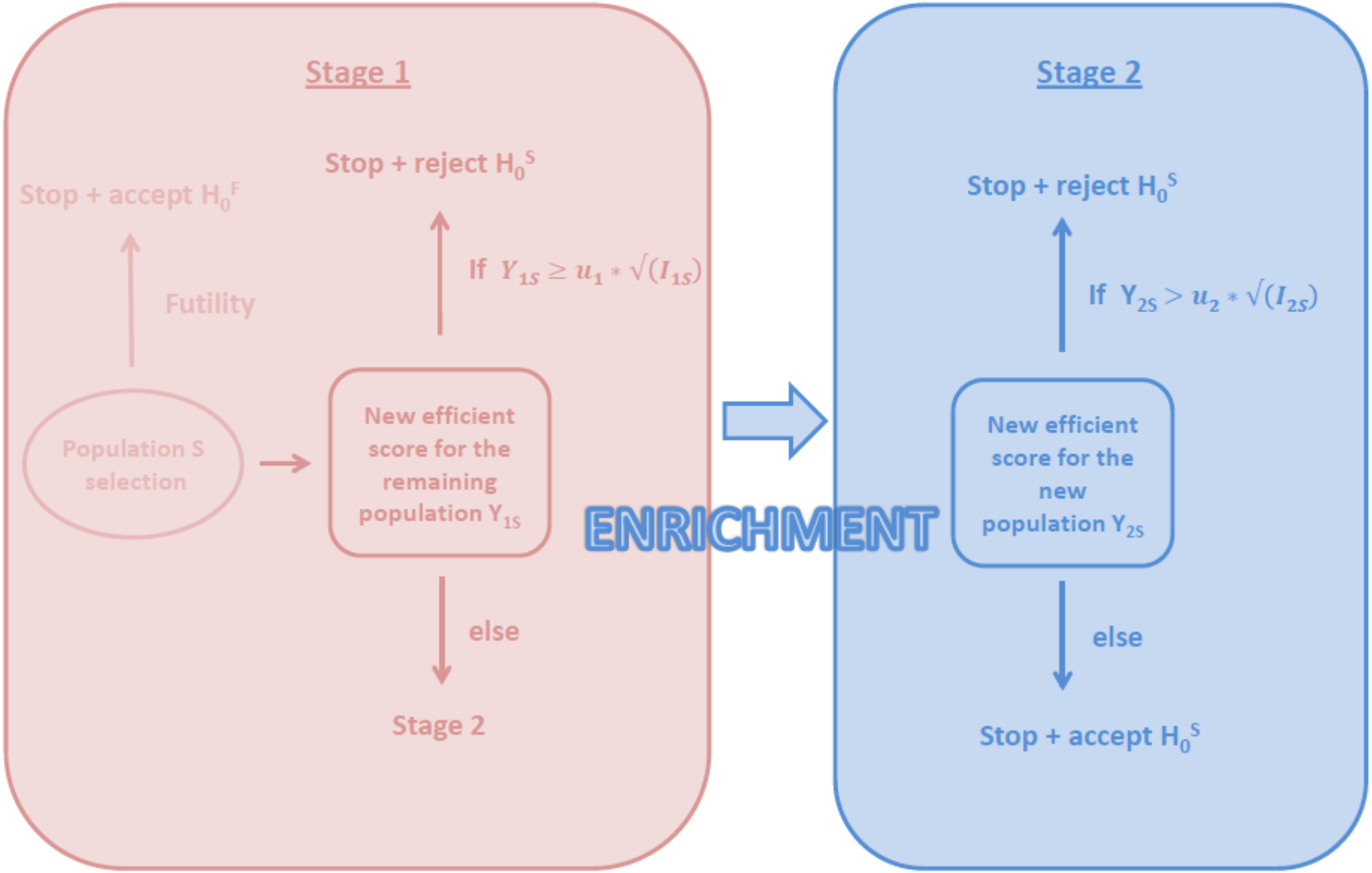
➜ 4 subsets of equal size based on quartiles ➜ **data driven**

➜ Quantitative treatment-by-biomarker interaction expected

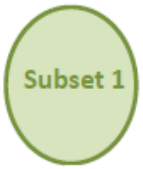
➜ Subsets a priori ordered by nature

H_0^F : there is no positive treatment effect in the full population
H_0^S : there is no positive treatment effect in the subpopulation S

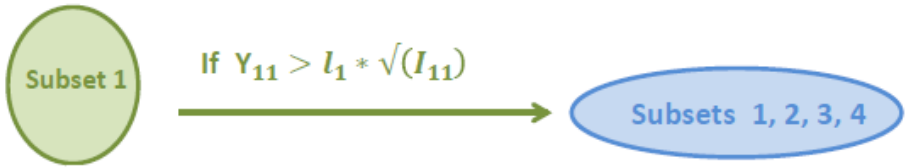
Reminder of GSED



Reminder of GSED

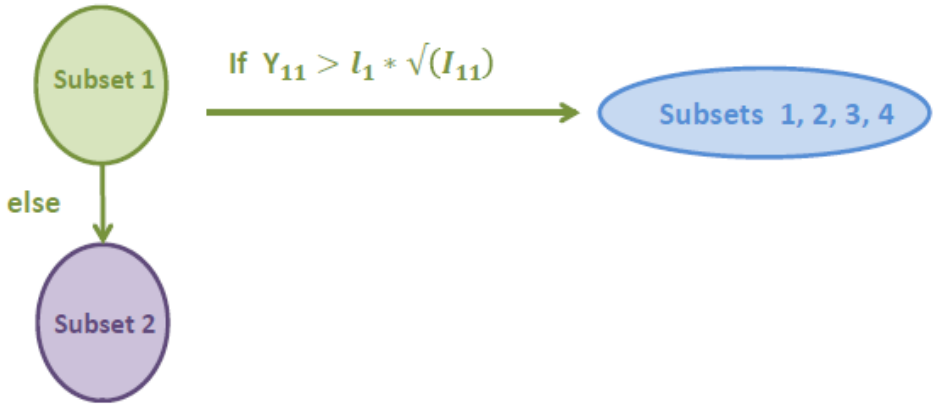


Reminder of GSED



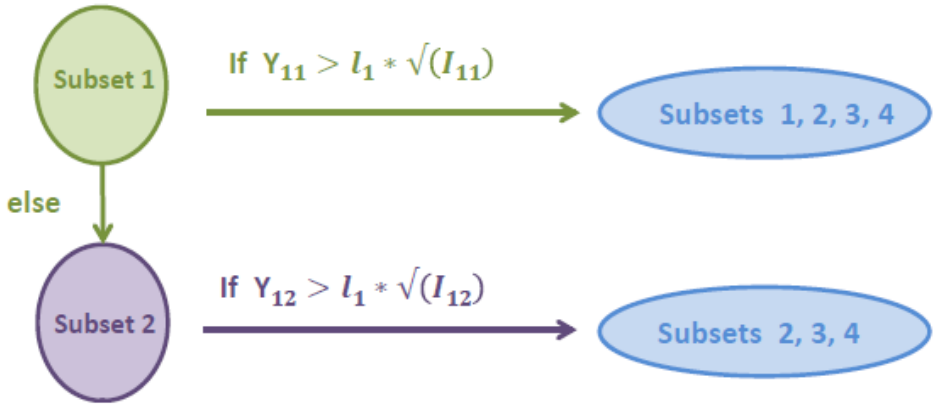
where Y_{1j} is the log-rank efficient score and I_{1j} is the Fisher's information of subset j at stage 1.

Reminder of GSED



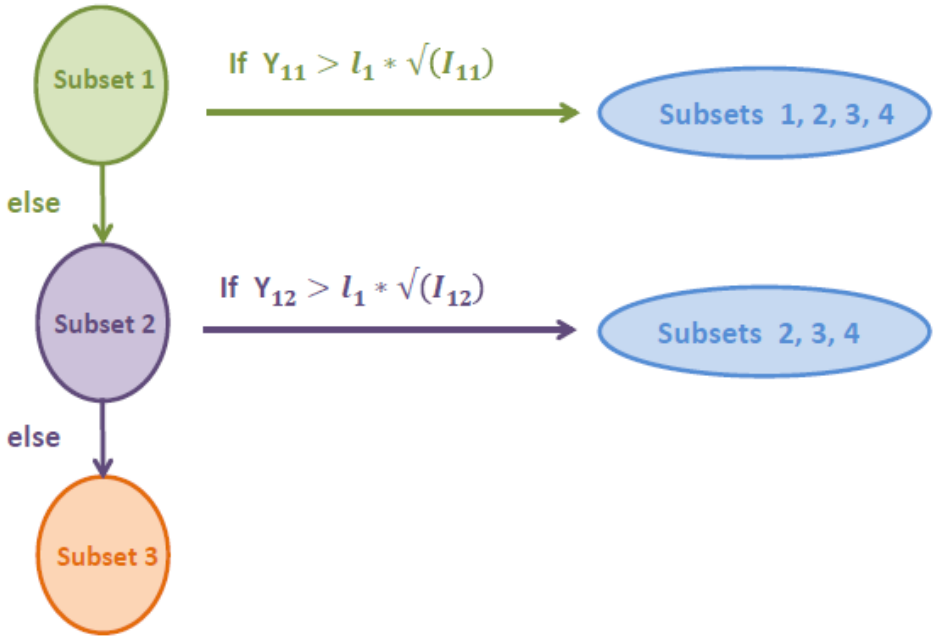
where Y_{1j} is the log-rank efficient score and I_{1j} is the Fisher's information of subset j at stage 1.

Reminder of GSED



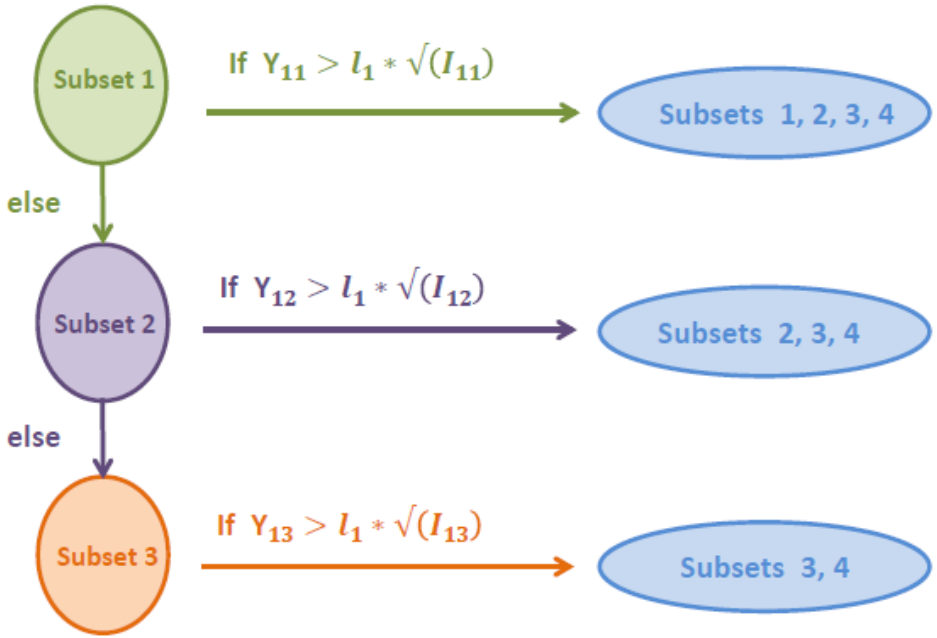
where Y_{1j} is the log-rank efficient score and I_{1j} is the Fisher's information of subset j at stage 1.

Reminder of GSED



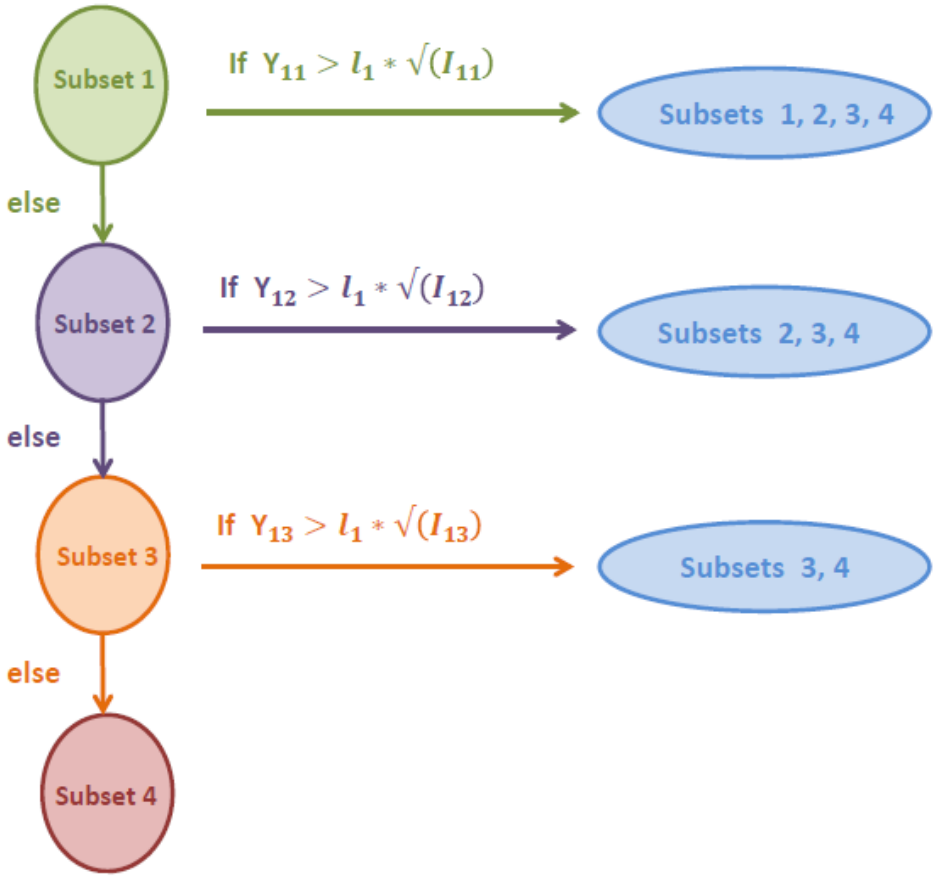
where Y_{1j} is the log-rank efficient score and I_{1j} is the Fisher's information of subset j at stage 1.

Reminder of GSED



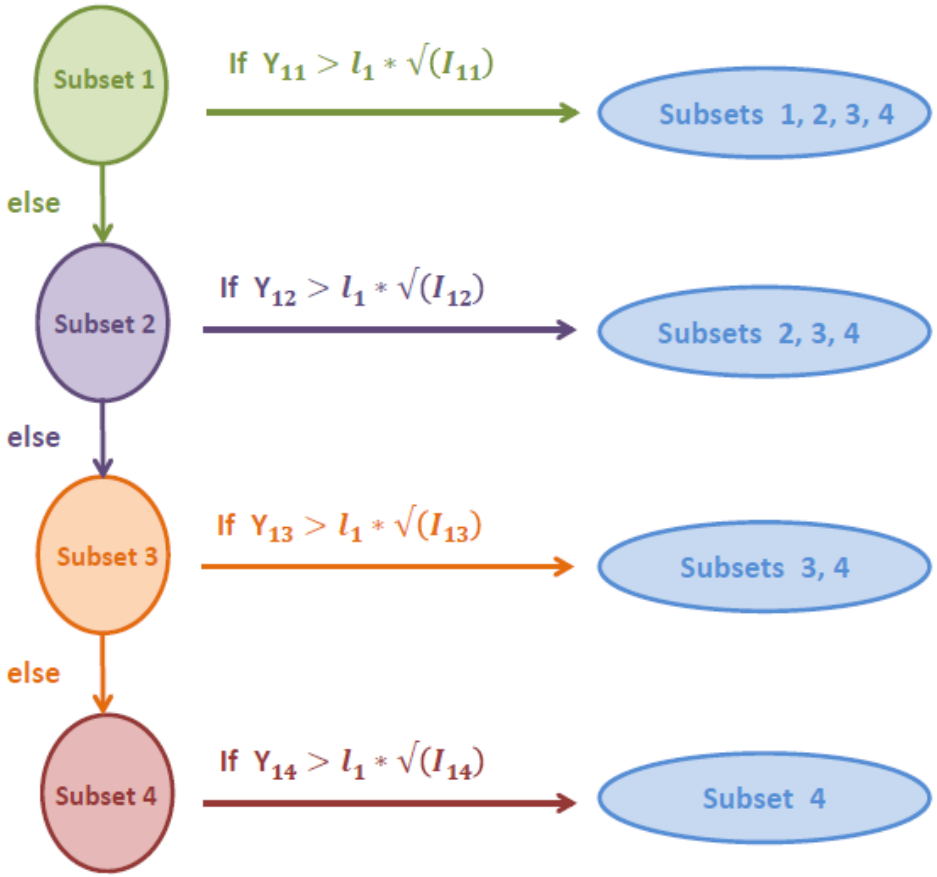
where Y_{1j} is the log-rank efficient score and I_{1j} is the Fisher's information of subset j at stage 1.

Reminder of GSED



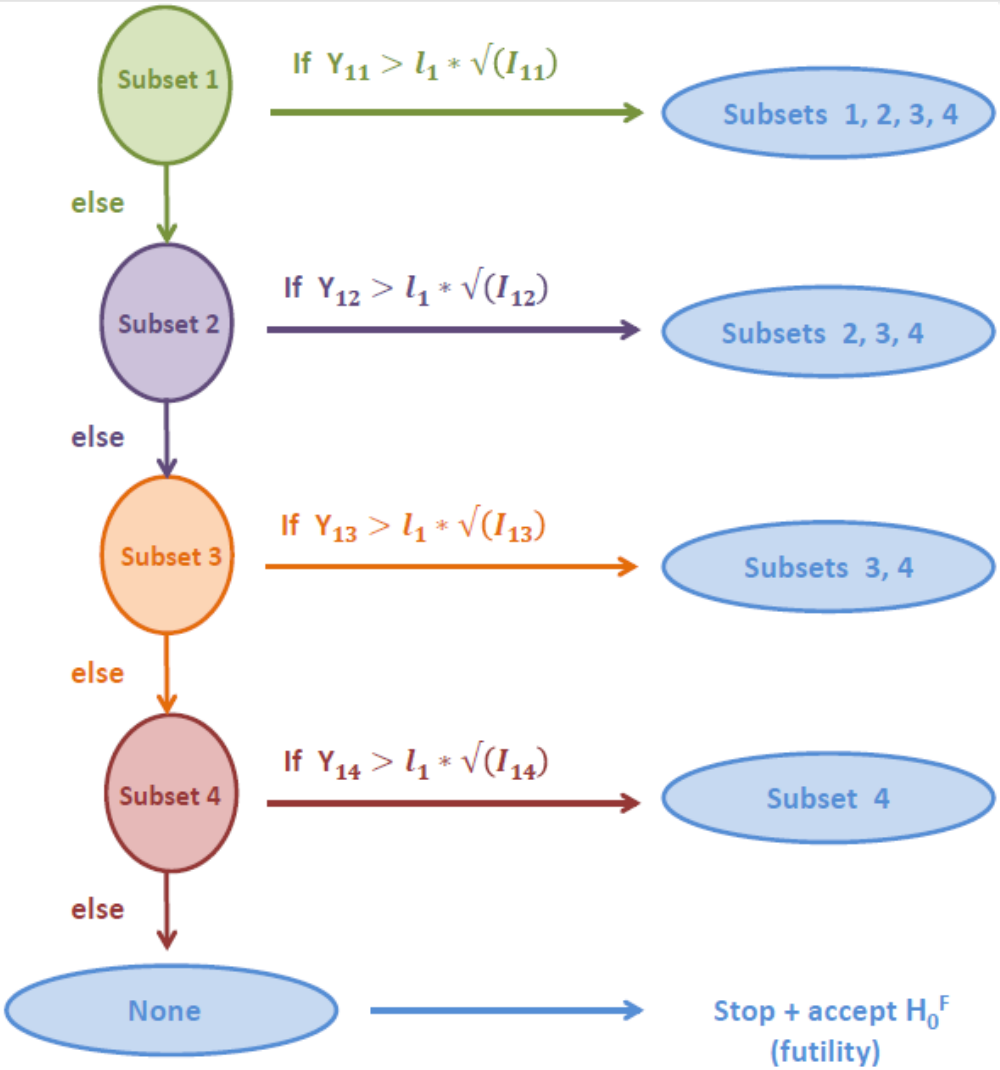
where Y_{1j} is the log-rank efficient score and I_{1j} is the Fisher's information of subset j at stage 1.

Reminder of GSED



where Y_{1j} is the log-rank efficient score and I_{1j} is the Fisher's information of subset j at stage 1.

Reminder of GSED



where Y_{1j} is the log-rank efficient score and I_{1j} is the Fisher's information of subset j at stage 1.

➜ $S \in \{ \{1,2,3,4\}, \{2,3,4\}, \{3,4\}, \{4\}, \emptyset \}$

Type I error rate control

$H_0^S : \theta_j = 0, \forall$ subset j $\in S$, tested for all selectable populations
S $\in$ { {1,2,3,4}, {2,3,4}, {3,4}, {4}, $\emptyset$ }

$$\rightarrow \quad \sum_{S \subseteq \mathcal{P}} P_{\theta=0}[\text{Select S}] P_{\theta=0}[\text{Reject } H_0^S | P^* = S] = \alpha_{global} \qquad (1)$$

➜ *Weak control of type I error rate*

$$\mathcal{P} = \{1, 2, 3, 4\}$$

Weak control of type I error rate + assumption of no opposite treatment effects across subsets

➜ ***Strong control of type I error rate***

Accomplished work

Our purpose

to set up optimal recommendations, according to the scenario underneath which we are, about using GSED

This requires

1. Comparing the pros and cons of
   - GSED vs fixed design
   - GSED vs Group Sequential Design
   - GSED vs Combination Tests
2. Quantifying the impact of
   - population selection procedure
   - full vs partial enrichment

   on the GSED decision-making

Contents

1 Setting the scene

2 Some results

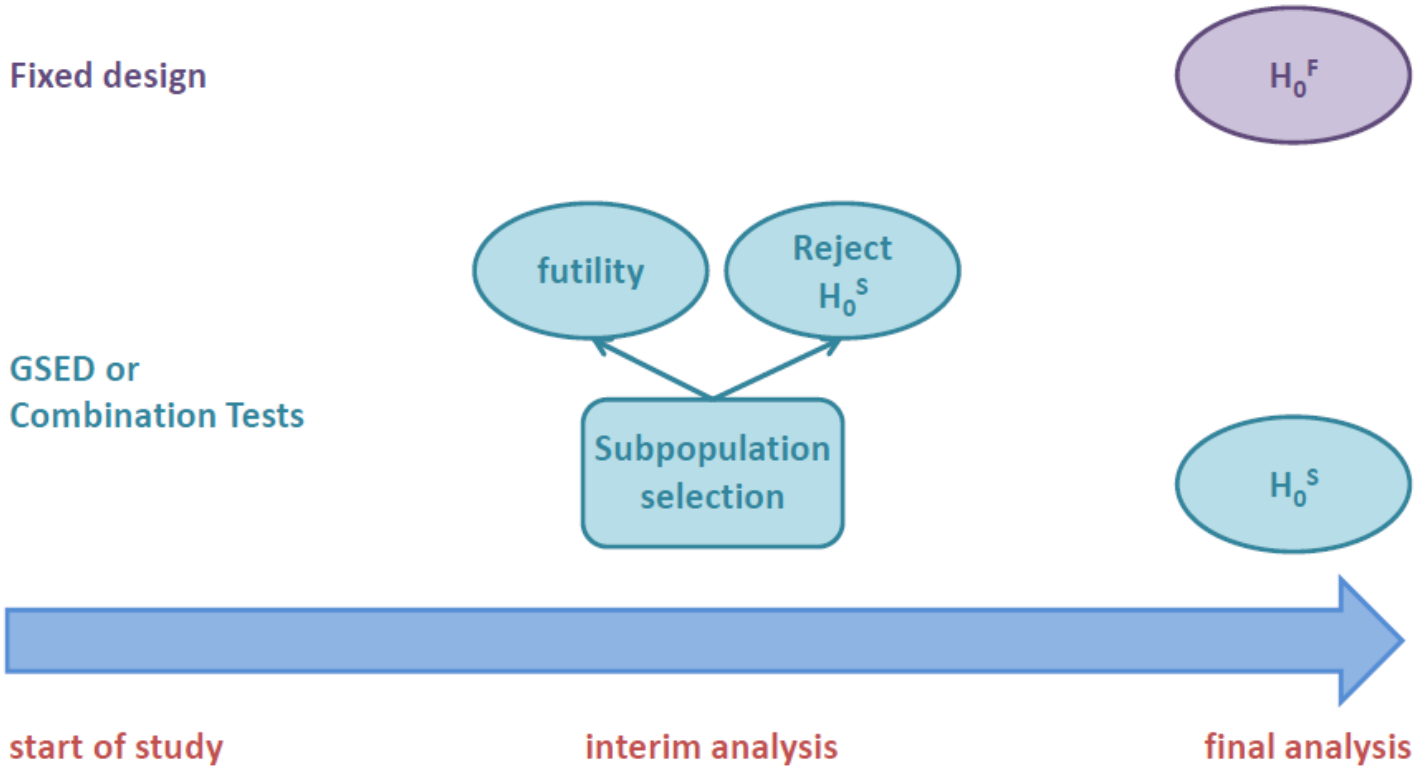
3 Discussion

Simulations framework

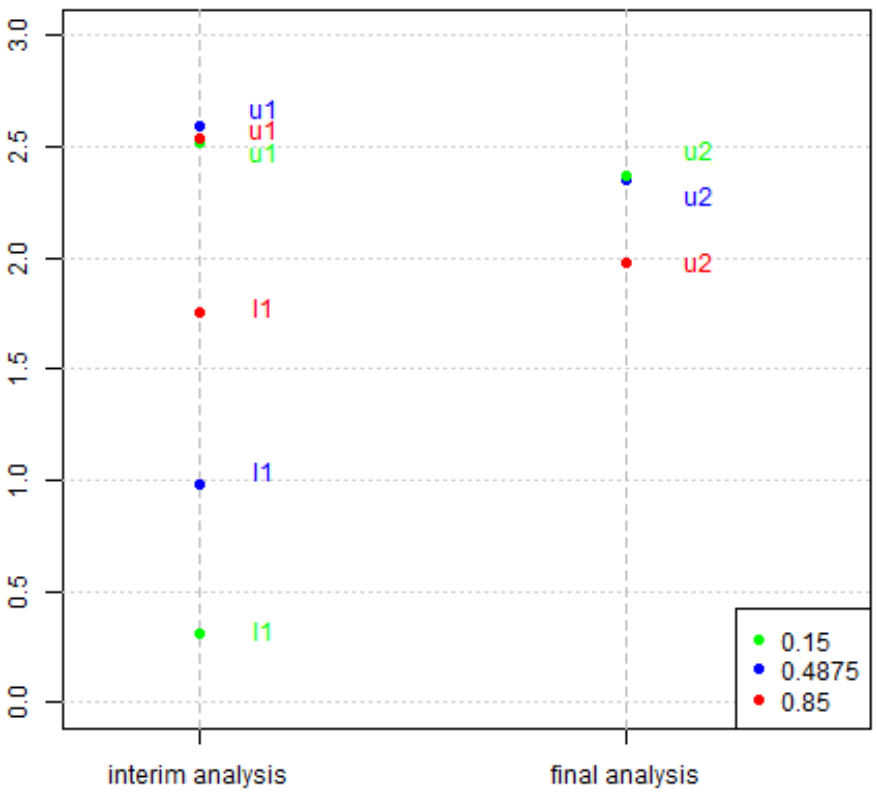
- H_0 : HR=1 vs HR=0.8 *5,000 trial simulations*
- $\alpha_{global} = 0.025$
- $\beta = 0.1$
- one-sided
- 2 arms : placebo vs treatment
- censoring rate = 0.1
- median time-to-event = 1 year in placebo group
- number of events expected = 847
- interim analysis at 424 events reached
- inclusion rate[1] $\approx$ 50 patients per month *(25 each arm)*
- duration of inclusion = until the Number of Subjects Required *(NSR = 1830)* is reached
- expected duration of a fixed design *(under HR=0.8)* $\approx$ 34 months

[1] If S={2,3,4} 85%, if S={3,4} 75%, if S={4} 65%

Considered designs



Stopping boundaries



Upper boundaries defined using an α_{global} spending function:

$$\begin{cases} \alpha^*_U(0) = 0 \\ \alpha^*_U(0.5) = 0.0125 \\ \alpha^*_U(1) = 0.025 \end{cases}$$

Lower boundaries defined using an $1 - \alpha_{global}$ spending function:

$$\begin{cases} \alpha^*_L(0) = 0 \\ \alpha^*_L(0.5) = 0.15, 0.4875 \text{ or } 0.85 \\ \alpha^*_L(1) = 0.975 \end{cases}$$

FWER control and power performance

Null hypothesis

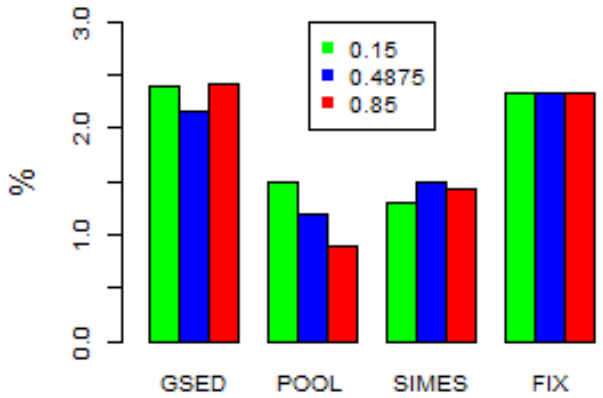


Figure: Probability to reject any H_0^S under the null hypothesis.

Alternative hypothesis

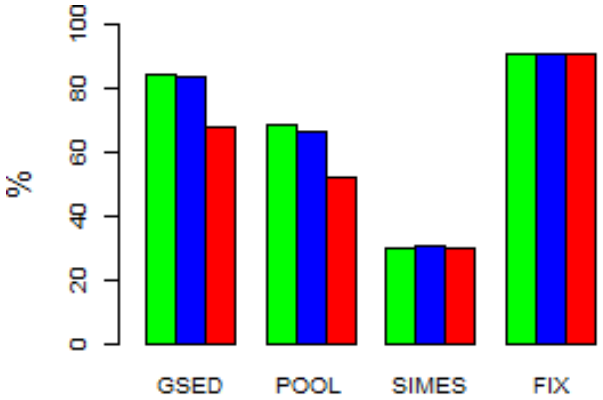


Figure: Probability to reject any H_0^S under the constant 0.8 scenario.

Assess the studied designs

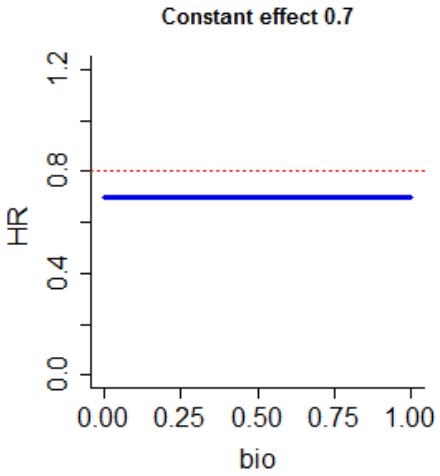
Primary criteria

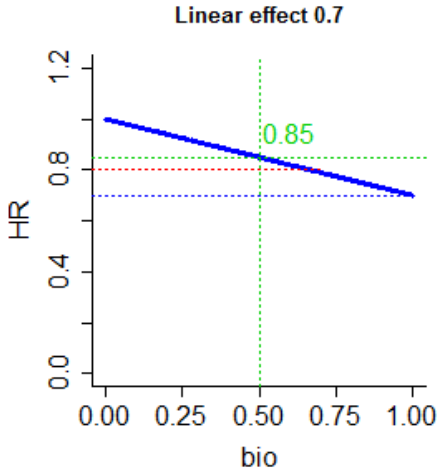
- control of FWER
- power under various scenarios
- averaged sample size
- averaged duration of study

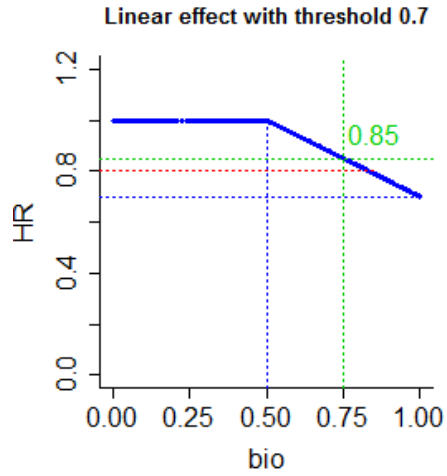
Secondary criteria

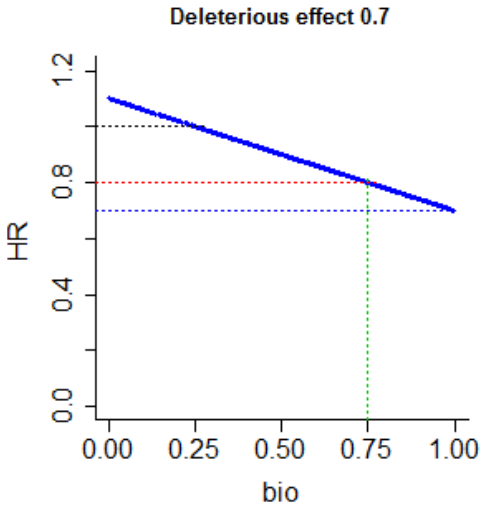
- probability of stopping at stage 1
- averaged final analysis date *(=averaged maximum study duration)*

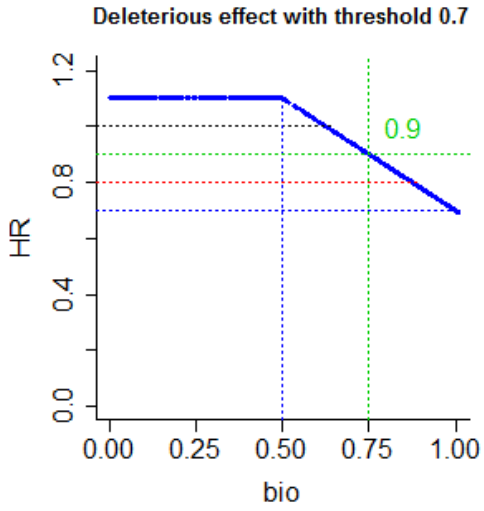
Various scenarios











Soundness of the population selection

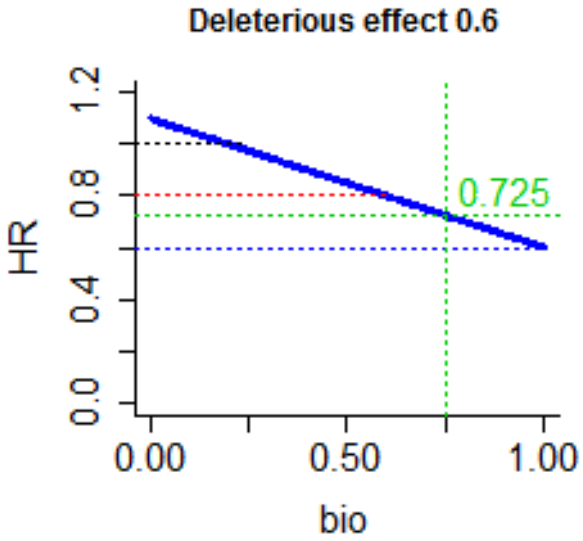
If HR $\leq$ 0.8, then there is a positive treatment effect.

Good and bad outcomes defined with Sanofi's panel experts:

- If the top 25% of patients has a Hazard Ratio (HR) very close to 0.8 or better, it is at least worth not stopping at the interim analysis.
- If a wider subpopulation also meets this standard, then it is good not to stop in that population.
- A desirable feature is to retain the larger subpopulation whose averaged HR is lower than 0.85.
- Conversely, an undesirable feature is to retain an individual subset having a deleterious treatment effect (i.e. HR $>$ 1).

Calibration of l_1 value

Move from one l_1 value to another one ?



GSED 0.4875 ➜ GSED 0.85

Probability to reject any H_0^S	-15%	✗
Good outcomes rate	+10%	✓
Bad outcomes rate	$\approx$ 10% ➜ $<$ 5%	✓
Averaged duration	-5%	✓
Averaged sample size	-5%	✓

GSED vs Combination Tests

PROPERTIES	GSED	CT
strong FWER control	+/++	++
power *(reject any H_0^S)*	++	-
flexibility	-	+
small sample size	+	+
small duration of study	+	+
likelier to stop at stage 1	+	-
minimizes bad outcomes	+	+/++
maximizes good outcomes	+	-

If GSED bad outcomes rate $\geq$ GSED good outcomes rate, taking a CT design instead of GSED may be an alternative.

Nevertheless

CT is inadequate for scenarios requiring enrichment because of a substantial loss of power imputable to the closed testing principle.

Contents

1 Setting the scene

2 Some results

3 Discussion

In a nutshell

➜ No *gold standard configuration*. *goal and constraint-dependent*
But globally, GSED more successful than CT.

➜ Not surprisingly, GSED more suitable than fixed design under scenarios with threshold and/or deleterious effect.

Comparative appraisal very challenging

- many indicators which sometimes contradict one another
  e.g. power vs sample size and duration
- assessment of the soundness of population selection ✓
  but subjective and scenario-dependent
- utility score bringing together the various key indicators ✗

Generalization of our results is hazardous

- closely related to inclusion dynamics
- percentile transformation run on a continuous biomarker
  ➜ test-to-test variations in case of high variability in its value

GSED use in a pivotal study ?

Premature to express a firm opinion

- strong control of type I error rate in a general setting is a conjecture
- difficulty in correctly addressing the bias estimation issues
- implementation of a dedicated software package to derive the GSED stopping boundaries is desirable
- lack of flexibility in changing population selection rules
  stopping boundaries rapidly intractable
- full respect of the futility rule is mandatory
- well-known relationship between biomarker and treatment effect *known thresholds or representative population to pre-specify subsets*
- need/pertinence for selecting a subpopulation
- no trouble in recruiting specific patients *inclusion dynamics*

Thanks for listening.

Contents

1. Setting the scene
2. Some results
3. Discussion

Efficient score

Log-rank statistics

$$S = \frac{\sum_{k=1}^{K}(D_{pbo,k} - E_{pbo,k})}{\sqrt{\sum_{k=1}^{K} V_k}} \sim \mathcal{N}(0,1) \tag{2}$$

where

- $D_{pbo,k}$ = number of deaths in the control arm at death time t_k
- $N_{pbo,k} - D_{pbo,k}$ = number of living patients in the control arm
- $E_{pbo,k} = \frac{N_{pbo,k} D_{pbo,k}}{N_k}$
- $V_k = V_{test,k} = V_{pbo,k} = \frac{N_{pbo,k} N_{test,k} D_k (N_k - D_k)}{N_k^2 (N_k - 1)}$

Efficient log-rank score

$$X = \sum_{k=1}^{K}(D_{pbo,k} - E_{pbo,k}) \tag{3}$$

Fisher's information

- d = number of expected events
- 4 subsets with equal proportions
- I_{max} = the maximal Fisher's information = $\frac{d}{4}$

At the interim analysis

$I_{1j} = \frac{1}{4} \times \frac{I_{max}}{2}$

$I_{1\{1,2,3,4\}} = I_{11} + I_{12} + I_{13} + I_{14}$

$I_{1\{2,3,4\}} = I_{12} + I_{13} + I_{14}$

$I_{1\{3,4\}} = I_{13} + I_{14}$

At the final analysis

$I_{2\{1,2,3,4\}} = \frac{I_{max}}{2} + I_{11} + I_{12} + I_{13} + I_{14}$

$I_{2\{2,3,4\}} = \frac{I_{max}}{2} + I_{12} + I_{13} + I_{14}$

$I_{2\{3,4\}} = \frac{I_{max}}{2} + I_{13} + I_{14}$

$I_{24} = \frac{I_{max}}{2} + I_{14}$

Null hypothesis discussion

How to deal with null hypotheses intersections ?

A weighting of the involved subsets

If S={3,4}, $H_0^{\{3,4\}} : w_3\theta_3 + w_4\theta_4 \leq 0$,
where w_3 and w_4 are the respective sizes of subsets 3 and 4.

But, how formalizing the intersection of $H_0^{\{3,4\}}$ and $H_0^{\{4\}}$?

$\{\theta_j\}$ cannot be negative for any j i.e. there cannot be opposite direction treatment effects *(cf B. P. Magnusson and B. W. Turnbull)*
➜ $H_0^{\{j\}} : \theta_j = 0$

Thus, if S={3,4}, $H_0^{\{3,4\}} : \theta_3 = \theta_4 = 0$
➜ intersections of null hypotheses are obvious

Stopping boundaries choice

Upper boundaries obtained using an α spending function:

$$\alpha_U^*(0)=0 < \alpha_U^*(0.5)=\alpha_1 < \alpha_U^*(1)=\alpha_{global}$$

$$\begin{cases} P_{\theta=0}[\text{Stop trial at stage 1 with rejection of some } H_0^S] = \alpha_U^*(0.5) - \alpha_U^*(0) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad = \alpha_1 \\ P_{\theta=0}[\text{Stop trial at stage 2 with rejection of some } H_0^S] = \alpha_U^*(1) - \alpha_U^*(0.5) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad = \alpha_{global} - \alpha_1 \end{cases}$$

⚠ $H_0^S : \theta_j = 0, \forall j \in S$, tested for all selectable subpopulations S

➜ $$\sum_{S \subseteq \mathcal{P}} P_{\theta=0}[\text{Select S}] P_{\theta=0}[\text{Reject } H_0^S | P^* = S] = \alpha_{global} \qquad (4)$$

= Weak control of type I error rate $\qquad \mathcal{P} = \{1, 2, 3, 4\}$

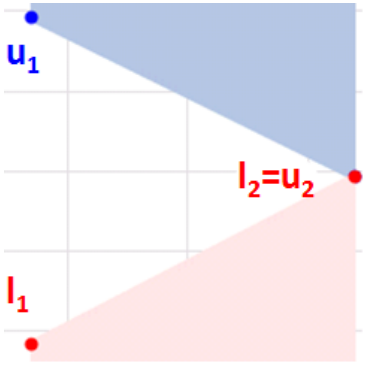
Stopping boundaries choice

Lower boundaries obtained using a $1-\alpha$ spending function:

$\alpha_L^*(0)=0 < \alpha_L^*(0.5)=\alpha_1' < \alpha_L^*(1)=1\text{-}\alpha_{global}$

$$\begin{cases} P_{\theta=0}[\text{Stop trial exactly at stage 1 with } \textbf{no} \text{ rejection of any } H_0^S] \\ \qquad = \alpha_L^*(0.5) - \alpha_L^*(0) = \alpha_1' \\ P_{\theta=0}[\text{Stop trial exactly at stage 2 with } \textbf{no} \text{ rejection of any } H_0^S] \\ \qquad = \alpha_L^*(1) - \alpha_L^*(0.5) = 1 - \alpha_{global} - \alpha_1' \end{cases}$$

- $P_{\theta=0}[\text{Select } \emptyset] P_{\theta=0}[\text{Accept } H_0^S \text{ at stage 1} | P^* = \emptyset] = P_{\theta=0}[P^* = \emptyset]$
$$= \alpha_L^*(0.5) = \alpha_1' \qquad (5a)$$
- $P_{\theta=0}[P^* = \emptyset] + \sum_{S \subseteq \mathcal{P}} P_{\theta=0}[\text{Select S}] P_{\theta=0}[\text{Accept } H_0^S \text{ at stage 2} | P^* = S]$
$$= \alpha_L^*(1) = 1 - \alpha_{global} \qquad (5b)$$



Condition 5b ➜ $l_2 = u_2$

Condition $u_2=l_2$

$$P_{\theta=0}[\text{Stop at stage 1 with rejection of some } H_0^S] = \alpha_1 \quad (6)$$

$$P_{\theta=0}[\text{Stop at stage 1 with no rejection of any } H_0^S] = \alpha_1' \quad (7)$$

$$(6) + (7) \rightarrow P_{\theta=0}[\text{Do not stop at stage 1}] = 1 - \alpha_1 - \alpha_1' \quad (8)$$

And

$$P_{\theta=0}[\text{Stop at stage 2 with rejection of some } H_0^S] = \alpha_{global} - \alpha_1 \quad (9)$$

$$P_{\theta=0}[\text{Stop at stage 2 with no rejection of any } H_0^S] = 1 - \alpha_{global} - \alpha_1' \quad (10)$$

$$(9) + (10) = (8) \rightarrow u_2=l_2$$

Familywise type I error rate control of GSED

The FWER control

$$\begin{aligned} \text{FWER} &= sup_\theta P_\theta[\text{Reject at least one } H_0^S, S \subseteq P^0(\theta)] \\ &= sup_\theta \sum_{S \subseteq P^0(\theta)} P_\theta[P^* = S \text{ and subsequently reject } H_0^S] \end{aligned}$$

where $P^0(\theta) = \{j \in \mathcal{P} : \theta_j = 0\}$, the index set of subsets for which there is no treatment effect.

So, for $\theta = 0$ i.e. when $\theta_j = 0, \forall j \in \mathcal{P}$, FWER $= \alpha_{global}$ by design.

strong control of FWER

if, for any arbitrary θ,

$$FWER = \sum_{S \subseteq P^0(\theta)} P_\theta[\text{Select S and reject } H_0^S] \leq \alpha_{global}$$

Familywise type I error rate control of GSED

We assume that $\{\theta_j\}$ cannot be negative for any j i.e. there cannot be opposite direction treatment effects ➜ $H_0^{\{3,4\}} : \theta_3 = \theta_4 = 0$

and θ^+ such that $P^0(\theta^+) \neq \emptyset$ and $P^0(\theta^+) \subset \mathcal{P}$ but $P^0(\theta^+) \neq \mathcal{P}$
➜ at least one subset with a positive treatment effect
+ one with a null effect

Hence,

$$FWER = \sum_{S \subseteq P^0(\theta^+)} P_{\theta^+}[\text{Select S}] P_{\theta^+}[\text{Reject } H_0^S | P^* = S]$$

$$< \sum_{S \subseteq P^0(\theta^+)} P_{\theta=0}[\text{Select S}] P_{\theta=0}[\text{Reject } H_0^S | P^* = S]$$

Familywise type I error rate control of GSED

We assume that $\{\theta_j\}$ cannot be negative for any j i.e. there cannot be opposite direction treatment effects ➜ $H_0^{\{3,4\}} : \theta_3 = \theta_4 = 0$ and θ^+ such that $P^0(\theta^+) \neq \emptyset$ and $P^0(\theta^+) \subset \mathcal{P}$ but $P^0(\theta^+) \neq \mathcal{P}$

➜ at least one subset with a positive treatment effect

\+ one with a null effect

Hence,

Under θ^+ configuration, at least one $\theta_j > 0 \notin S = \{r, ..., 4\}, j < r$

➜ less likely to select S *(cf population selection process)*

$$FWER = \sum_{S \subseteq P^0(\theta^+)} P_{\theta^+}[\text{Select S}] P_{\theta^+}[\text{Reject } H_0^S | P^* = S]$$

$$<$$

$$< \sum_{S \subseteq P^0(\theta^+)} P_{\theta=0}[\text{Select S}] P_{\theta=0}[\text{Reject } H_0^S | P^* = S]$$

Familywise type I error rate control of GSED

We assume that $\{\theta_j\}$ cannot be negative for any j i.e. there cannot be opposite direction treatment effects ➜ $H_0^{\{3,4\}} : \theta_3 = \theta_4 = 0$ and θ^+ such that $P^0(\theta^+) \neq \emptyset$ and $P^0(\theta^+) \subset \mathcal{P}$ but $P^0(\theta^+) \neq \mathcal{P}$

➜ at least one subset with a positive treatment effect

\+ one with a null effect

Hence,

$\theta_S = 0$ under both configurations

$$FWER = \sum_{S \subseteq P^0(\theta^+)} P_{\theta^+}[\text{Select S}] P_{\theta^+}[\text{Reject } H_0^S | P^* = S]$$

$$= $$

$$< \sum_{S \subseteq P^0(\theta^+)} P_{\theta=0}[\text{Select S}] P_{\theta=0}[\text{Reject } H_0^S | P^* = S]$$

Familywise type I error rate control of GSED

We assume that $\{\theta_j\}$ cannot be negative for any j i.e. there cannot be opposite direction treatment effects ➜ $H_0^{\{3,4\}} : \theta_3 = \theta_4 = 0$ and θ^+ such that $P^0(\theta^+) \neq \emptyset$ and $P^0(\theta^+) \subset \mathcal{P}$ but $P^0(\theta^+) \neq \mathcal{P}$

➜ at least one subset with a positive treatment effect

\+ one with a null effect

Hence,

If there are some subsets with treatment effect in S

➜ likelier to select S

$$FWER = \sum_{S \subseteq P^0(\theta^+)} P_{\theta^+}[\text{Select S}] P_{\theta^+}[\text{Reject } H_0^S | P^* = S]$$

$$< \sum_{S \subseteq P^0(\theta^+)} P_{\theta=0}[\text{Select S}] P_{\theta=0}[\text{Reject } H_0^S | P^* = S]$$

$$< \sum_{S \subseteq \mathcal{P}} P_{\theta=0}[\text{Select S}] P_{\theta=0}[\text{Reject } H_0^S | P^* = S]$$

Familywise type I error rate control of GSED

We assume that $\{\theta_j\}$ cannot be negative for any j i.e. there cannot be opposite direction treatment effects ➜ $H_0^{\{3,4\}} : \theta_3 = \theta_4 = 0$ and θ^+ such that $P^0(\theta^+) \neq \emptyset$ and $P^0(\theta^+) \subset \mathcal{P}$ but $P^0(\theta^+) \neq \mathcal{P}$

➜ at least one subset with a positive treatment effect

\+ one with a null effect

Hence,

$$FWER = \sum_{S \subseteq P^0(\theta^+)} P_{\theta^+}[\text{Select S}] P_{\theta^+}[\text{Reject } H_0^S | P^* = S]$$

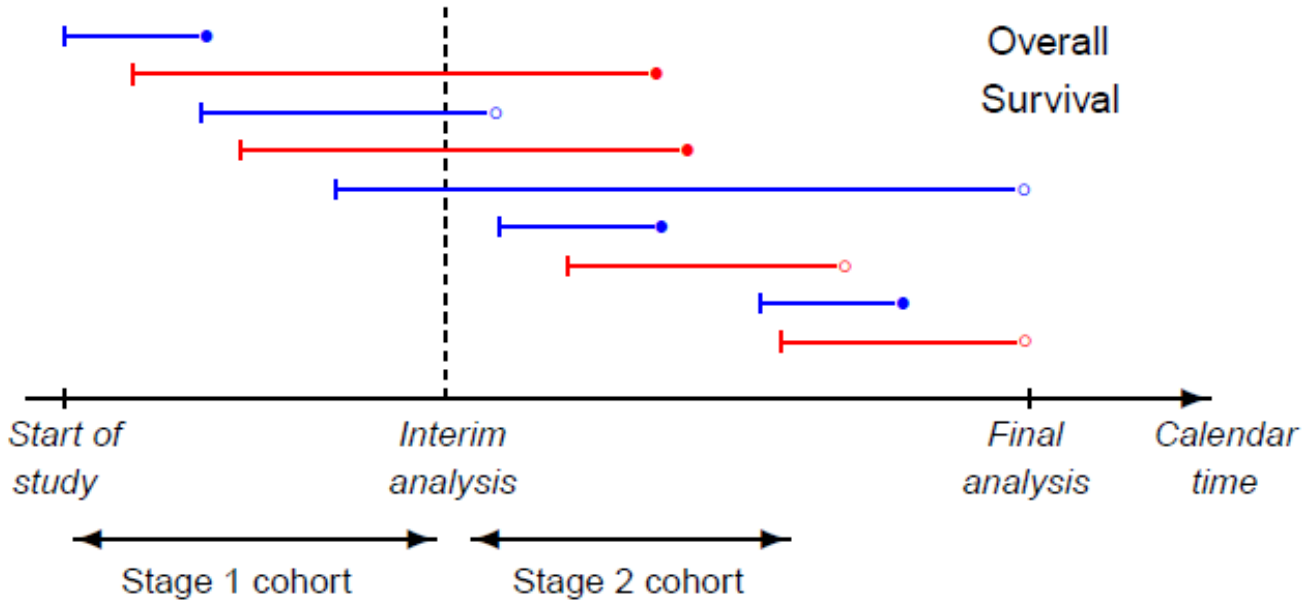
$$< \sum_{S \subseteq P^0(\theta^+)} P_{\theta=0}[\text{Select S}] P_{\theta=0}[\text{Reject } H_0^S | P^* = S]$$

$$< \sum_{S \subseteq \mathcal{P}} P_{\theta=0}[\text{Select S}] P_{\theta=0}[\text{Reject } H_0^S | P^* = S] = \alpha_{global}$$

by design

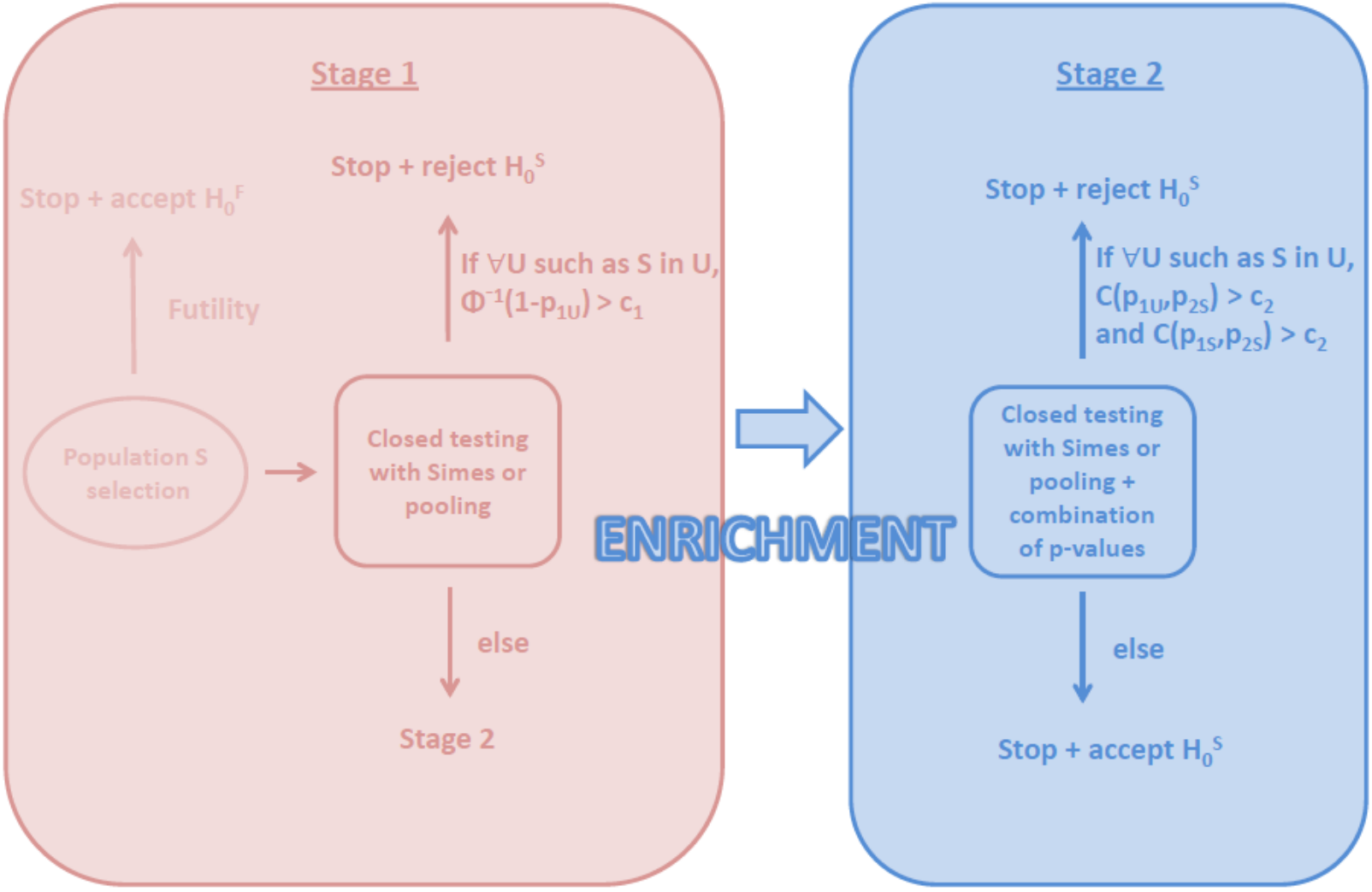
Reminder of Combination Tests

two independent cohorts



Key:
- Subjects randomised to Exp. Treatment
- Subjects randomised to Control
- ● Death observed
- ○ Censored observation.

Reminder of Combination Tests



Reminder of Combination Tests

Stage 1

If S={4} :
➜ if $\Phi^{-1}(1-p_{1,4}) > c_1$ & $\Phi^{-1}(1-p_{1,34}) > c_1$ & $\Phi^{-1}(1-p_{1,234}) > c_1$ & $\Phi^{-1}(1-p_{1,1234}) > c_1$, then stop the trial by rejecting H_0^S
➜ else proceed to stage 2

If S={3,4} :
➜ if $\Phi^{-1}(1-p_{1,34}) > c_1$ & $\Phi^{-1}(1-p_{1,234}) > c_1$ & $\Phi^{-1}(1-p_{1,1234}) > c_1$, then stop the trial by rejecting H_0^S
➜ else proceed to stage 2

If S={2,3,4} :
➜ if $\Phi^{-1}(1-p_{1,234}) > c_1$ & $\Phi^{-1}(1-p_{1,1234}) > c_1$, then stop the trial by rejecting H_0^S
➜ else proceed to stage 2

If S={1,2,3,4} :
➜ if $\Phi^{-1}(1-p_{1,1234}) > c_1$, then stop the trial by rejecting H_0^S
➜ else proceed to stage 2

Reminder of Combination Tests

Stage 2

If S={4} :

➜ if $C(p_{1,4}, p_{2,4}) > c_2$ & $C(p_{1,34}, p_{2,4}) > c_2$ & $C(p_{1,234}, p_{2,4}) > c_2$ & $C(p_{1,1234}, p_{2,4}) > c_2$ then stop the trial by rejecting H_0^S

➜ else stop the trial by accepting H_0^S

If S={3,4} :

➜ if $C(p_{1,34}, p_{2,34}) > c_2$ & $C(p_{1,234}, p_{2,34}) > c_2$ & $C(p_{1,1234}, p_{2,34}) > c_2$ then stop the trial by rejecting H_0^S

➜ else stop the trial by accepting H_0^S

If S={2,3,4} :

➜ if $C(p_{1,234}, p_{2,234}) > c_2$ & $C(p_{1,1234}, p_{2,234}) > c_2$ then stop the trial by rejecting H_0^S

➜ else stop the trial by accepting H_0^S

If S={1,2,3,4} :

➜ if $C(p_{1,1234}, p_{2,1234}) > c_2$ then stop the trial by rejecting H_0^S

➜ else stop the trial by accepting H_0^S

Reminder of Combination Tests

closed testing principle (Marcus et al., 1976)

$H_0^S : \theta_S \leq 0$ is rejected overall at level α_{global}
$\Leftrightarrow$ each H_0^U is rejected at level α_{global} for every subpopulation U containing S, based on stage 1 and stage 2 data

➜ test $H_0^S : \bigcap_{S \subseteq U} H_0^U$ at each stage

$H_0^{\{i\}} : \theta_i = 0$ ➜ $H_0^U : \theta_U = 0$ means $\forall$ subset $i \in U, \theta_i = 0$

Hence,

- if S={1,2,3,4}, $H_0^S : \theta_1 = \theta_2 = \theta_3 = \theta_4 = 0$,
- if S={2,3,4}, $H_0^S = H_0^{\{1,2,3,4\}} \bigcap H_0^{\{2,3,4\}} : \theta_2 = \theta_3 = \theta_4 = 0$,
- if S={3,4}, $H_0^S = H_0^{\{1,2,3,4\}} \bigcap H_0^{\{2,3,4\}} \bigcap H_0^{\{3,4\}} : \theta_3 = \theta_4 = 0$,
- if S={4}, $H_0^S = H_0^{\{1,2,3,4\}} \bigcap H_0^{\{2,3,4\}} \bigcap H_0^{\{3,4\}} \bigcap H_0^{\{4\}} : \theta_4 = 0$.

Reminder of Combination Tests

➜ test H_0^U : $\forall$ subset $i \in U, \theta_i = 0$ at each stage

Simes' method (Simes, 1986)

Noting m = number of subsets of U at stage 1, m' = number at stage 2 and $p_{k,U}$ = p-value of U at stage k with Φ the standard normal distribution,

Stage 1

- calculate $p_{1,U} = min_{j=1,\dots,m}(\frac{m * p_{1,(j)}}{j})$,
  where $p_{1,(j)}$ denotes the j^{th} p-value of cohort 1 in increasing order

Stage 2

- calculate $p_{1,U} = \min\limits_{j=1,\dots,m}(\frac{m * p_{1,(j)}}{j})$,
  where $p_{1,(j)}$ denotes the j^{th} p-value of cohort 1 in increasing order
- calculate $p_{2,U} = \min\limits_{j=1,\dots,m'}(\frac{m' * p_{2,(j)}}{j})$,
  where $p_{2,(j)}$ denotes the j^{th} p-value of cohort 2 in increasing order

⚠ *$p_{1,U}$ at stage 2 may differ from $p_{1,U}$ at stage 1*

$p_{2,U} = p_{2,S}$ because enrichment = cohort 2

Reminder of Combination Tests

➜ test H_0^U : $\forall$ subset $i \in U, \theta_i = 0$ at each stage

Once p-values are estimated by Simes' method or by a mere pooling of the considered subsets, if we proceeded to stage 2:

➜ combine the two stages p-values of each population U

Inverse Normal Combination Function (Lehmacher et Wassmer, 1999)

H_0^U is rejected if $C(p_{1,U}, p_{2,U}) = w_1 Z_{1,U} + w_2 Z_{2,U}$
$= w_1 \Phi^{-1}(1 - p_{1,U}) + w_2 \Phi^{-1}(1 - p_{2,U}) > c_2$,

where Φ is the standard normal distribution
and w_1 and w_2 have to be pre-specified such that $w_1^2 + w_2^2 = 1$

Usually, we set $w_1 = w_2 = 0.5$.
Note that $C(p_{1,U}, p_{2,U}) = w_1 Z_{1,U} + w_2 Z_{2,U} \sim \mathcal{N}(0,1)$ under H_0^U.

Proposals of GSED: 2 options

Once the population selection performed at stage 1, how to plan the final analysis ?

GSED A: *Partial enrichment*

All events stage 1 + events in S after IA = number of expected events

- Stage 1 *(Before IA)* : events of S **and events of S^c**
- Stage 2 *(After IA)* : events of S

(considered by B.P. Magnusson & B. W. Turnbull)

GSED B: *Full enrichment*

Only events in S stage 1 + 2 = number of expected events

- Stage 1 *(Before IA)* : events of S
- Stage 2 *(After IA)* : events of S

(better power in case of small S but more costly)

Proposals of GSED: 2 checks

How to improve our design ?

Check 1 : futility test for S

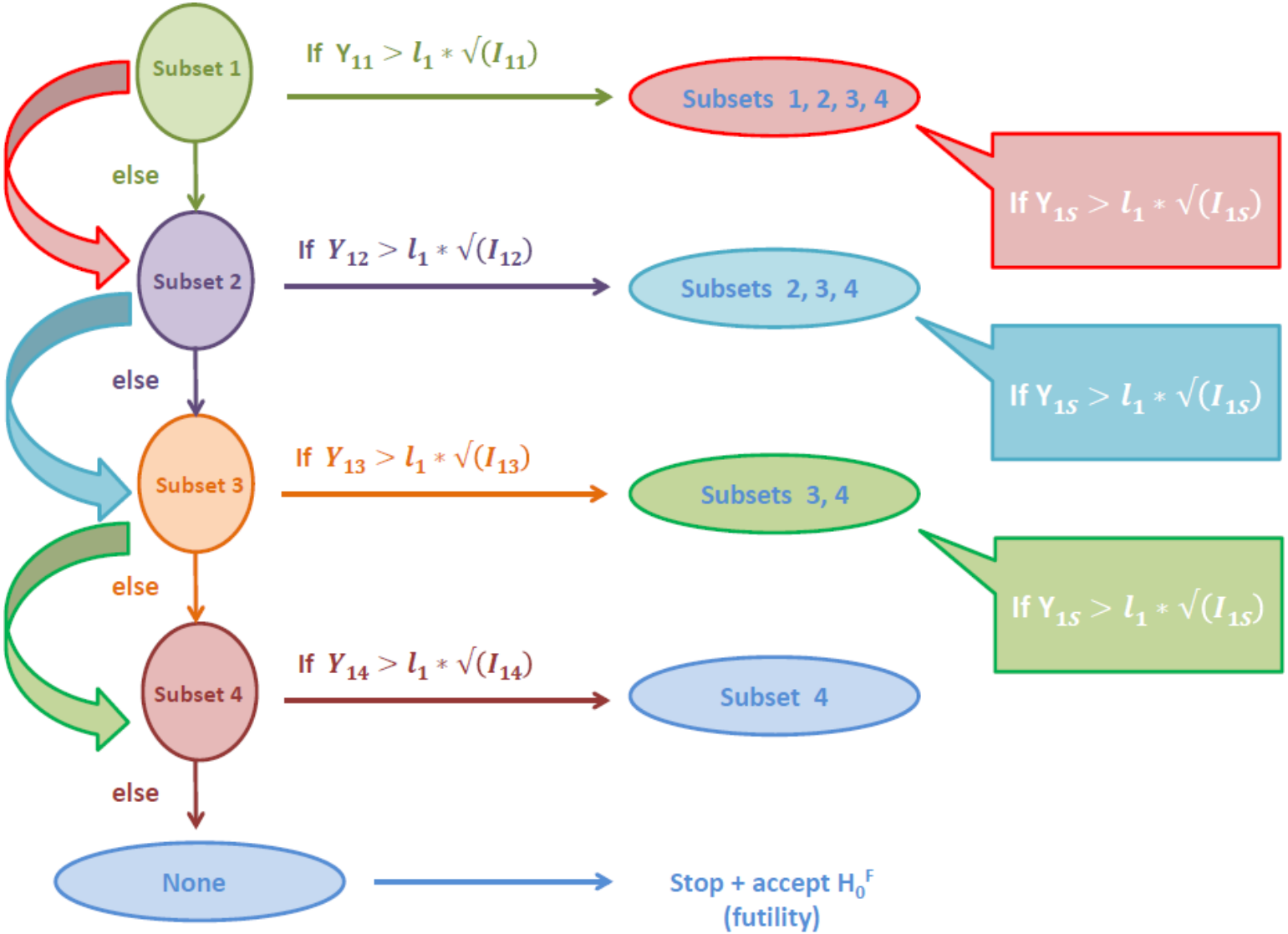
- add the condition : if $Y_{1S} \leq l_1 * \sqrt{l_{1S}}$, remove the smaller index subset and follow the S selection

Check 2 : heterogeneity check for S

- make sure there is no opposite treatment effects between subsets

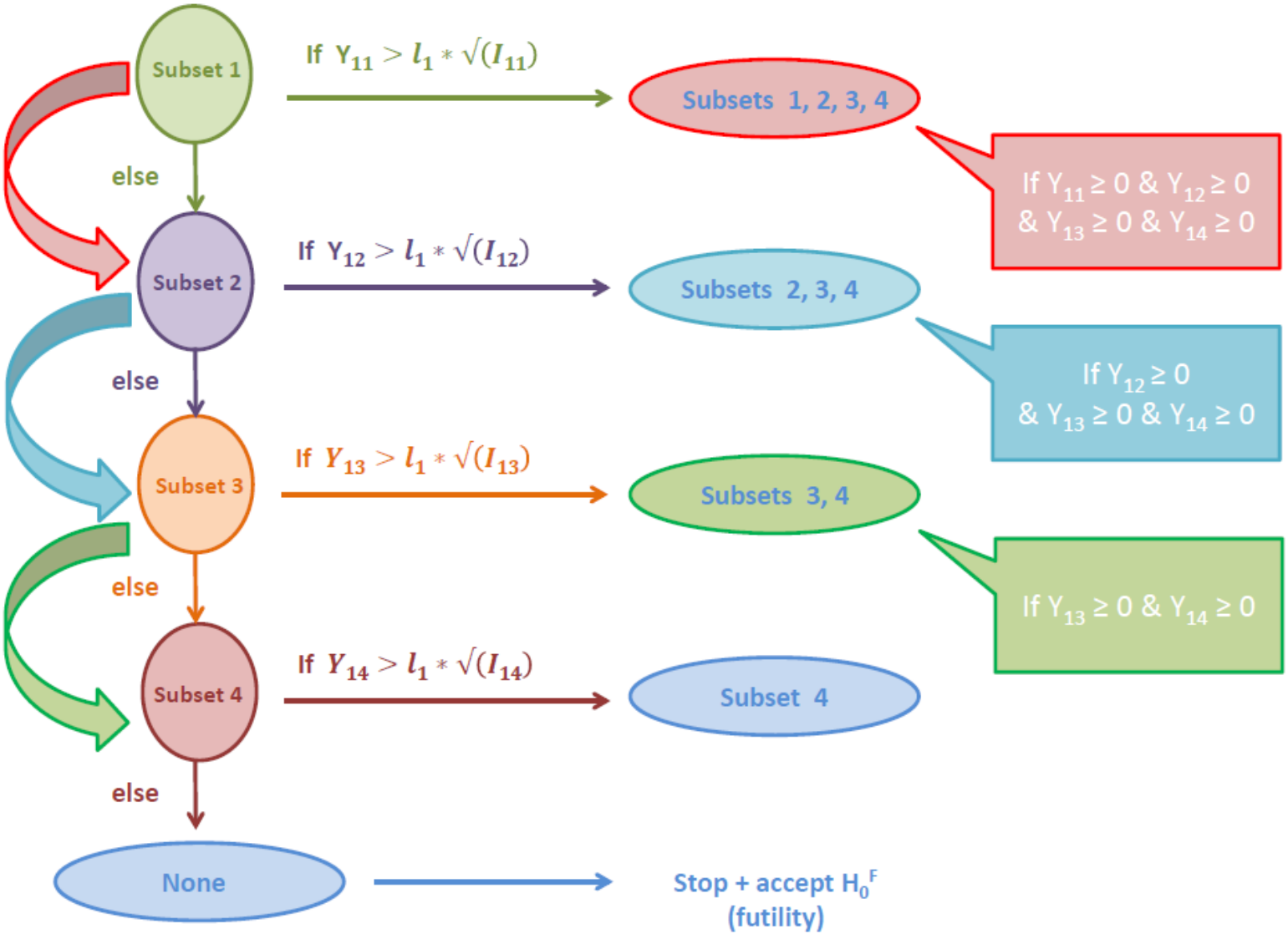
⚠ *Computation of stopping boundaries must be adjusted according to option A or B and the checks chosen !*

Check 1



where Y_{1j} is the log-rank efficient score and I_{1j} is the Fisher's information of subset j at stage 1.

Check 2



where Y_{1j} is the log-rank efficient score and I_{1j} is the Fisher's information of subset j at stage 1.

Checks utility

scenario	l_1 value	futility (%)
H_0	small	28.98
	medium	22.88
	high	7.48
Constant 0.8	high	7.86
Linear 0.8	small	7.16
	medium	10.00
	high	7.98
Linear+threshold 0.7	small	6.20
	medium	8.00
Linear+threshold 0.8	small	12.84
	medium	12.10
	high	5.78
deleterious 0.8	small	9.38
	medium	10.00
deleterious+threshold 0.6	small	6.30
	medium	5.02
deleterious+threshold 0.7	small	10.96
	medium	7.32
deleterious+threshold 0.8	small	16.42
	medium	10.14

futility is the amount of times the selected population S does not pass the futility test: $Y_{1S} > l_1 * \sqrt{l_{1S}}$.
(5,000 trial simulations)

A major drawback of GSED

$\star$ **not flexible:** stopping boundaries derivation extremely arduous

- $P_{\theta=0}[P^* = S]$ modifications
- l_1 modifications
- u_1 and u_2 modifications

➜ checks 1 & 2 withdrawn

GSED vs GSD

H_0^F : there is no positive treatment effect in the full population
H_0^S : there is no positive treatment effect in the subpopulation S

GSED

Stage 1

- select the subpopulation S, using futility boundary l_1
- reject H_0^S or proceed to stage 2, using efficacy boundary u_1

Enrichment

Stage 2

- conclude about H_0^S, using efficacy boundary u_2

GSD

Stage 1

- futility option added, using l_1 of GSED
- reject H_0^F or proceed to stage 2, using efficacy boundary c_1

Usual accrual

Stage 2

- conclude about H_0^F, using efficacy boundary c_2

General trend regarding designs

constant scenarios :

- **GSED A = GSD** 0.5 & 0.6
- **GSD > GSED A** 0.7 & 0.8
- **Simes > pooling** 0.5
- **Simes = pooling** 0.6
- **pooling > Simes** 0.7 & 0.8

linear scenarios :

- **GSD > GSED A** duration, sample size
- **GSED A > GSD** power, good, bad
- **pooling > Simes** power, good

linear+threshold scenarios :

- **GSD > GSED A** duration, sample size
- **GSED A > GSD** power, good, bad
- **Simes > pooling** all

deleterious scenarios :

- **GSD > GSED A** duration, sample size
- **GSED A > GSD** power, good, bad
- **Simes > pooling** duration, sample size, bad
- **pooling > Simes** power

deleterious+threshold scenarios :

- **GSD > GSED A** duration, sample size
- **GSED A > GSD** power, good, bad
- **Simes > pooling** all

GSED B < GSED A ➜ too small rises in power and good outcomes for huge averaged sample size and study duration

General trend regarding l_1 values

- **averaged duration of study** ↘ when l_1 value ↗
- **final analysis date** ↘ when l_1 value ↘
- **probability to stop at stage 1** ↗ when l_1 value ↗
- **averaged sample size** ↘ when l_1 value ↗
- **probability to reject any** H_0^S ↗ when l_1 value is medium

constant scenarios :

- **good outcomes** ↗ when l_1 values ↘

linear scenarios :

- **good outcomes** ↗ when l_1 value is medium or low (HR=0.5 & 0.6), medium or high (HR=0.7 & 0.8)
- **bad outcomes** ↘ when l_1 values ↗ (not defined for HR=0.5 & 0.6 & 0.7)

linear+threshold scenarios :

- **good outcomes** ↗ when l_1 value is medium or high
- **bad outcomes** ↘ when l_1 values ↗

deleterious scenarios :

- **good outcomes** ↗ when l_1 value is medium or high
- **bad outcomes** ↘ when l_1 values ↗

deleterious+threshold scenarios :

- **good outcomes** ↗ when l_1 value ↗
- **bad outcomes** ↘ when l_1 values ↗

Our recommendations

scenario	design	l_1 value	Comment
H_0			all designs control the FWER
constant 0.8	A/pool	low	↗ good rate,power/↘ duration,sample size
0.7	A/pool	low	↗ good rate,power/↘ duration,sample size
0.6	A/pool/Simes	low	medium l_1 values are also good
0.5	A	low / medium	Simes, high l_1 values are also good
linear 0.8	Simes	high	greatly ↘ bad rate *(GSED A greatly ↗ power)* disappointing performance on good/bad rates
0.7	A	medium / high	↗ good rate,power/↘ duration,sample
0.6	A	low / medium	↗ slightly good rate/↗ slightly power
0.5	A	low / medium	low l_1 value to minimize the maximum duration
linear+th 0.8	Simes	high	↘ bad rate *(GSED A ↗ power)* disappointing performance on good/bad rates
0.7	A	high	*(Simes ↘ bad rate,* *medium l_1 value ↗ power, good rate)*
0.6	A	medium / high	↗ power,good rate/↘ bad rate,duration,sample
0.5	A	medium / high	↗ power/ ↘ bad rate, duration, sample,↗ good rate

Our recommendations

scenario	design	l_1 value	Comment
deleterious 0.8	A	high	↗ power with respect to Simes disappointing performance on good/bad rates
0.7	A	medium / high	↗ good rate,power/ ↘ duration,sample,bad rate
0.6	A	medium / high	↗ good rate,power/ ↘ duration,sample,bad rate
0.5	A	medium / high	↗ good rate,power/ ↘ duration,sample,bad rate
deleterious+th 0.8	A/Simes/pool	high	many endings for futility
0.7	A	high	optimizing all criteria disappointing performance on good/bad rates *(medium l_1 value ↗ power)*
0.6	A	medium / high	↗ power/ ↘ bad rate,duration,sample,↗ good rate *(Simes unfortunately ↘ power)*
0.5	A	medium / high	↗ power/ ↘ bad rate,duration,sample,↗ good rate

Estimation bias

Bias source

population selection process at stage 1

Configuration

4 biomarker subsets, two-stage study

Accomplished work

- averaged unconditional bias per subset
- averaged bias per subset conditional on the selected population
- Mean Squared Error of unadjusted and adjusted estimators
- empirical coverage probabilities of 90% confidence intervals

Estimation bias methodology

Simulate B clinical trials {
  For every trial {
    $\forall j \in [\![1,4]\!]$, calculate Maximum Likelihood Estimator (MLE) $\hat{\theta}_j$
    Then simulate B samples, using the estimated $\hat{\theta}_j$ as θ_j {
      $\forall j \in [\![1,4]\!]$, calculate the MLE of the b^{th} trial $\hat{\theta}^*_{bj}$
    }

    Which leads to the mean MLE for θ_j : $\bar{\theta}^{*1}_{Bj} = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}^*_{bj}$

    Hence $\hat{\theta}^{*1}_j = \hat{\theta}_j - (\bar{\theta}^{*1}_{Bj} - \hat{\theta}_j)$, where $(\bar{\theta}^{*1}_{Bj} - \hat{\theta}_j)$ is the simulated bias estimate.
  }
  The green part can be repeated with $\hat{\theta}^{*1} = (\hat{\theta}^{*1}_1, ..., \hat{\theta}^{*1}_4)$ instead of $\hat{\theta}_j$ in the B samples simulation, which leads to $\hat{\theta}^{*2}_j = \hat{\theta}_j - (\bar{\theta}^{*2}_{Bj} - \hat{\theta}^{*1}_j)$.
}
Finally, compute the Mean-Squared Error (MSE) for the B trials.

Its 90% confidence intervals: using the 5% and 95% percentiles of the B samples simulations.

Probability whether the true treatment effect θ_j is within them: based on all the B trial simulations.

Estimation bias results

- our attempts have not materialized into convincing results
- according to the publication: *'Concerning estimation, topics for further research include development of improved bootstrap procedures for unbiased estimation and confidence intervals [...]'*
  ➜ the authors agree that estimation issues remain unanswered and very difficult to address
- bias and confidence intervals are known to be a very difficult component of adaptive designs
- not clear to us at which extent estimation issues should be explored and satisfactorily solved for endorsement by regulators in pivotal settings