Workshop on Health Care Studies, Institute for Mathematical Sciences.
July 3-7, 2017, Singapore

Multiplicity considerations in design, data monitoring and analysis of clinical trials with two semi-competing risks outcomes

Toshimitsu Hamasaki, PhD, Pstat®

National Cerebral and Cardiovascular Center, Suita, Osaka, Japan

National Cerebral and Cardiovascular Center

This research is partly supported by the Project Promoting Clinical Trials for Development of New Drugs (17lk0201061h0002/17lk0201061h0202) from the Japan Agency for Medical Research and Development (AMED)

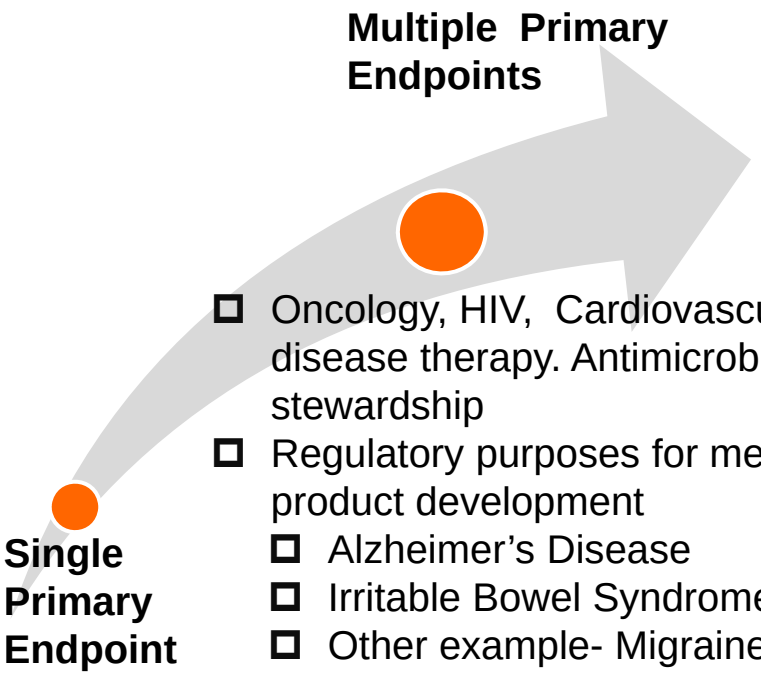
Acknowledgements

Tomoyuki Sugimoto, PhD
Kagoshima University, Kagoshima, Kagoshima, Japan

Scott R Evans, PhD
Harvard T.H. Chan School of Public Health, Boston, MA, USA

Koko Asakura, PhD
National Cerebral and Cardiovascular Center, Suita, Osaka, Japan

Multiple endpoints in clinical trials



- Oncology, HIV, Cardiovascular disease therapy. Antimicrobial stewardship
- Regulatory purposes for medical product development
  - Alzheimer's Disease
  - Irritable Bowel Syndrome
  - Other example- Migraine, Acute Pain, Fibromyalgia, Low Back Pain, Osteoarthritis etc

- Offer the opportunity of more completely characterizing intervention's multidimensional effects, especially in complex diseases
- Create challenges in design and analysis of clinical trials
- Extensive research and great methodological advance in this area over the last serval decade.
  - Many methods are available for continuous or binary
  - Methods for time-to-event outcomes are still limited although they are also common endpoints

CHMP Guideline on Medicinal Products for the Treatment Alzheimer's Disease and Other Dementias. EMEA: London, 2008; FDA . Daft Guidance for Industry. Irritable Bowel Syndrome: Clinical Evaluation of Products for Treatment. FDA: Rockville, MD, 2010; Offen et al, *Drug Inform J* 2007;**41**:31-46

Inferential goals for multiple endpoints

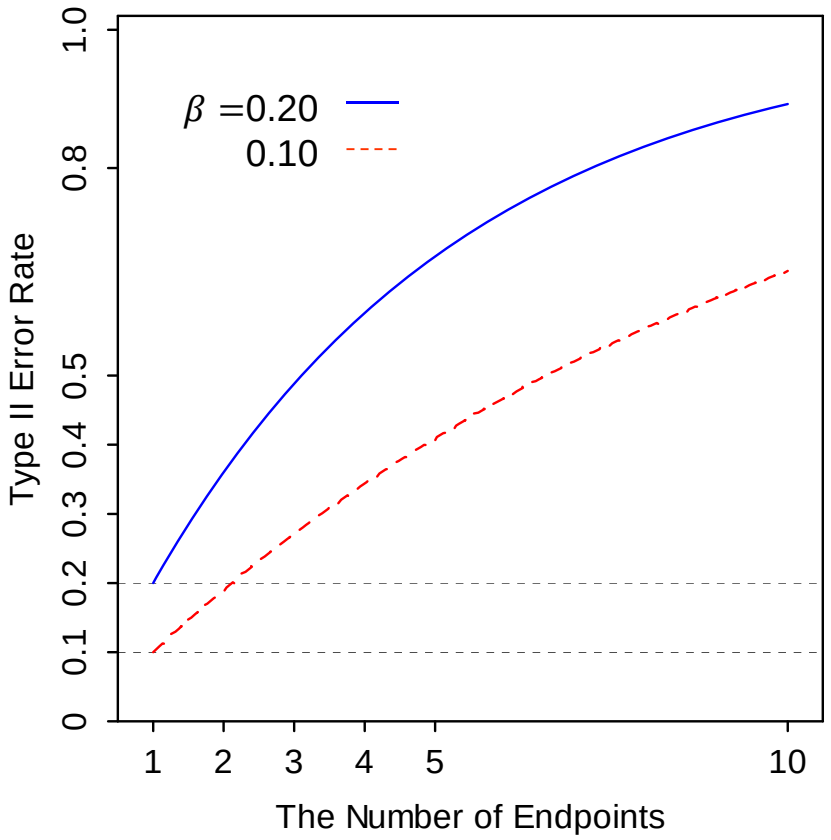
Inferential goals	Procedures
All-or-none: A trial is declared positive if statistical superiority effect is detected on ***ALL*** endpoints (**multiple co-primary endpoints: MCPE**)	Intersection union procedure
At-least one: A trial is declared positive if statistical superiority effect is detected on *AT-LEAST-ONE* endpoint (**multiple primary endpoints: MPE**)	Union-Intersection procedure Bonferroni and related procedure Fixed-sequence procedure Fallback procedure Adaptive alpha allocation …
Global: A trial is declared positive if statistical superiority effect is detected across the endpoints without necessarily a large significant effect any one endpoints	Normal theory model Likelihood ratio procedure
Superiority-noninferiority; A trial is declared positive if statistical superiority effect is detected on *AT-LEAST-ONE* endpoint, noninferior effect on all other endpoints	

Dmitrienko A et al (2010). *Multiple Testing Problems in Pharmaceutical Statistics*. CRC Press.

Co-primary endpoints and Type II error inflation

Multiple "Co-Primary" Endpoints

- No adjustment is needed to control Type I error rate as intersection-union test
- Type II error rate increases as the number of endpoints to be tested increases
- The marginal power must be increased for each endpoint to maintain the overall power at the design stage.
- But the sample size will result in too large sample size to conduct a clinical trial.

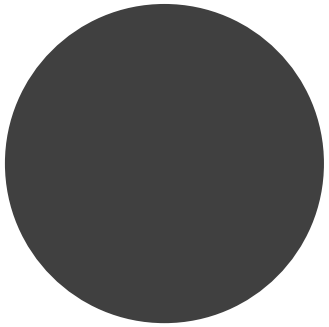


Zero correlations are assumed among the endpoints

Our research on co-primary endpoints

Outcome Scale	Fixed sample designs	Group-sequential designs
Continuous	Sozu et al (J Biopharm Stat 2011;21:650–668); Sugimoto et al (Pharma Stat 2012;11:118-128); Sozu et al (J Biopharm Stat 2016; 26, 631-643); Huang et al (PLoS ONE 2017 (in press))	Asakura et al (Stat Med 2014); Hamasaki et al (Stat Biopharm Res 2015; 7:36-54); Asakura et al (Biom J 2017 (in press))
Binary	Sozu et al (Stat Med 2010;29:2169–2179); Sozu et al (J Biopharm Stat 2011;21:650–668); Ando et al (Stat Biopharm Res 2015)	Asakura et al (Applied Statistics in Biomedicine and Clinical Trials Design, Chen Z et al (eds.), 235-262, Springer, 2015)
Time-to-event	Hamasaki et al (Pharm Statist 2013;12:28-34); Sugimoto et al (Biostat 2013;14:409-421); Sugimoto et al (Stat Med 2017;36: 1363-1382)	
Others	Sozu et al (Biomet J 2012; 54:716–729)	

Presentation outline

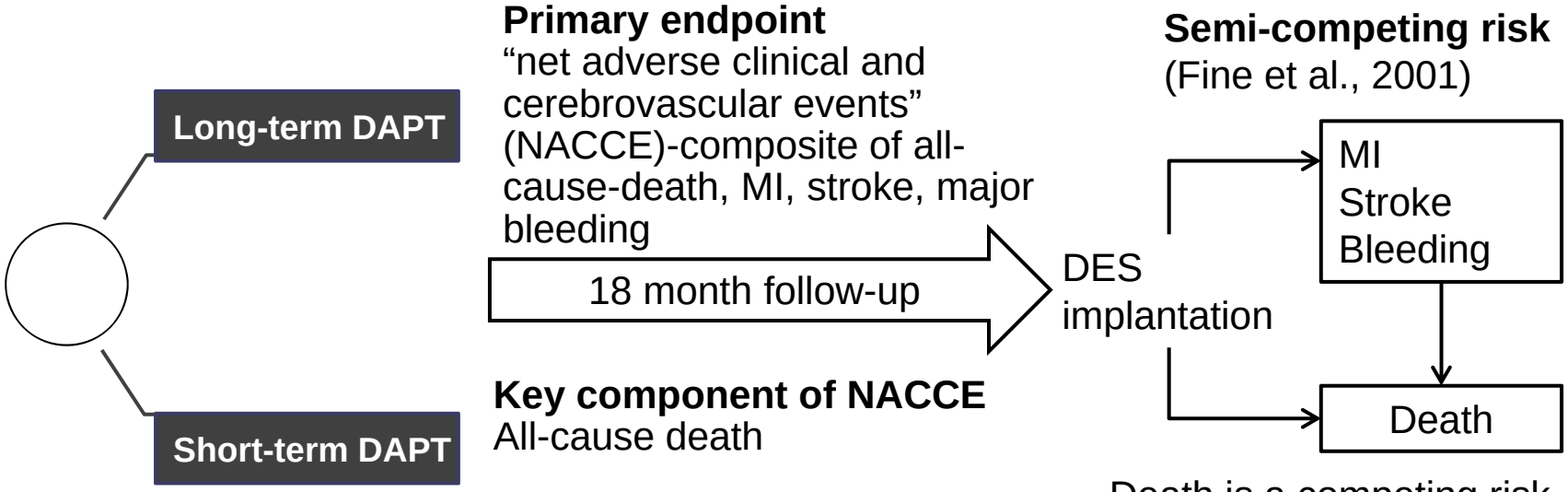


1. **Introduction**
2. **Technical development**
   Trial design, censoring scheme, and association measure
   logrank test statistics and its joint distribution
3. **co-primary endpoints**
   Strategies to rejecting null hypothesis, sample size, event numbers, empirical powers and Type I error rate
4. **multiple primary endpoints**
   Strategies to rejecting null hypothesis, and empirical powers
5. **Summary and further development**

1. Introduction

NIPPON study for a better DAPT duration

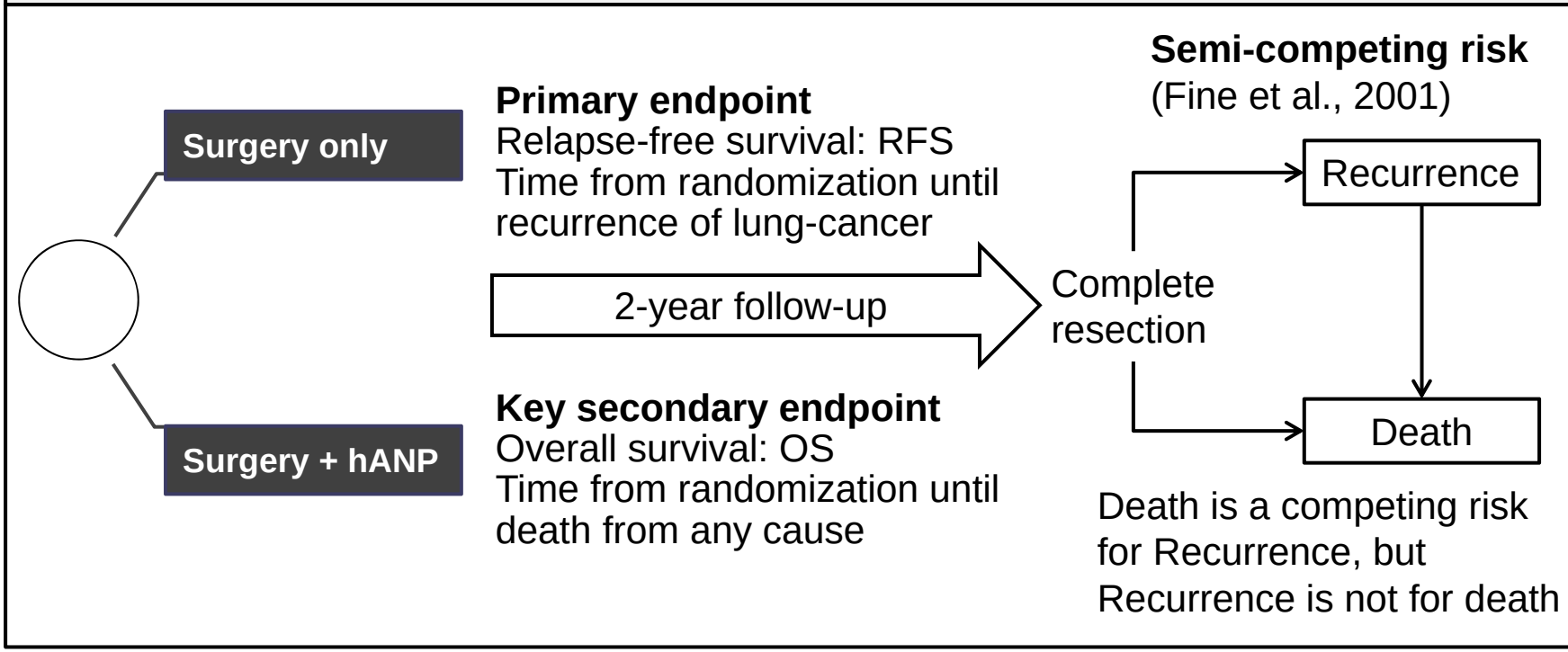
A randomized, 2-arm (1:1 ratio), multi-center, open-label, randomized noninferiority trial designed to evaluate the safety and efficacy of short-term (6 months) dual antiplatelet therapy (DAPT) versus long-term (18 months) DAPT after drug-eluting stent (DES) implantation in patients with coronary artery disease (Nakamura et al., 2017)



Nakamura et al. JACC Cardiovascular intervention 2017;10,1189-1198.

JANP Trial for preventing postoperative lung cancer recurrence

A randomized, 2-arm (1:1 ratio), multi-center, open-label, superiority, randomized phase II trial designed to evaluate the safety and efficacy of adding human atrial natriuretic peptide (hANP) to surgery versus surgery only, in patients with lung cancer (Nojiri et al, 2017)



Nojiri T et al. Trials 2017; **18**:183
Fine JP et al. Biometrika 2001;**88**:907-919

Multiple endpoints in cancer clinical trials

Table 1. A Comparison of Important Cancer Approval Endpoints

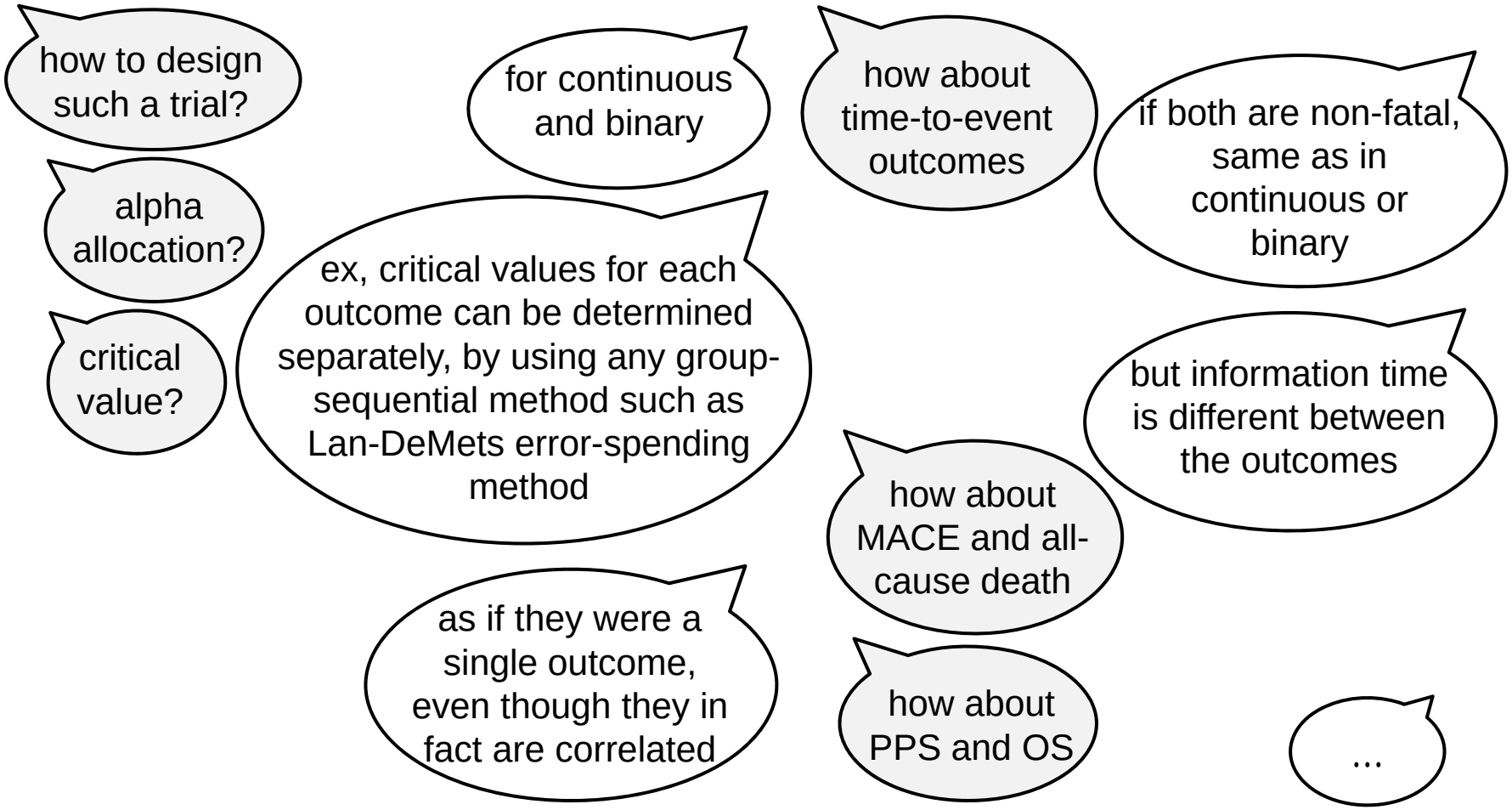
Endpoints	Regulatory Evidence	Study design	Advantages	Disadvantages
OS: Overall Survival	Clinical benefit for regulatory approval	Randomized clinical studies essential Blinding not essential	Universally accepted direct measure of benefit Easily measured Precisely measured	May involve larger studies May be affected by crossover therapy and sequential therapy Includes non-cancer deaths
TTP: Time to Progression or **PFS: Progression -Free Survival**	Surrogate for accelerated approval or regular approval	Randomized clinical studies essential Blinding preferred Blinded review recommend ed	Smaller sample size and shorter following-up necessary compared with survival studies Measurement of stable disease included Not affected by crossover or subsequent therapies Generally based on objective and quantitative assessment	Not statistically validated as surrogate for survival in all settings Not precisely measured; subject to assessment bias particularly in open-label studies Definitions vary among studies Frequent radiological or other assessment Involves balanced timing of assessments among treatment arms

FDA. Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics. 2007

Group-sequential designs for two event-time clinical trials

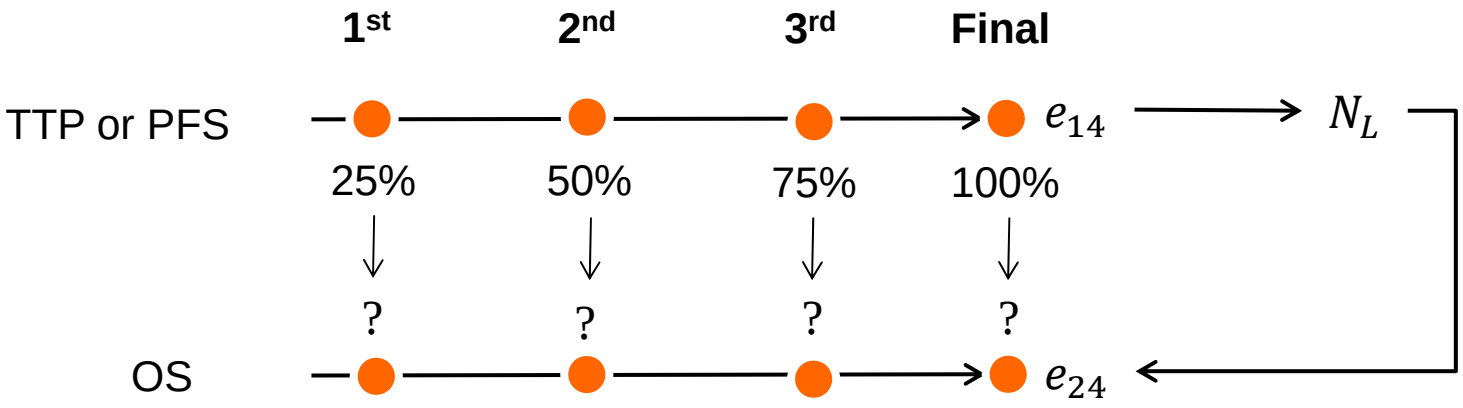
- Clinical trials with multiple event-time outcomes can be **expensive** and **resource intensive** as they often require
  - enrollment of large numbers of participants:
  - collection of massive amounts of data
  - long-term follow-up:
- **Group-sequential designs** can streamline clinical trials making them **more efficient**
  - offering potentially fewer required trial participants,
  - shortening the duration of clinical trials,
  - reducing costs
- Designing event-time trials is **more complex** **than continuous or binary outcome trials**, and considerable cares are needed especially in a group-sequential setting:

Questioning by myself



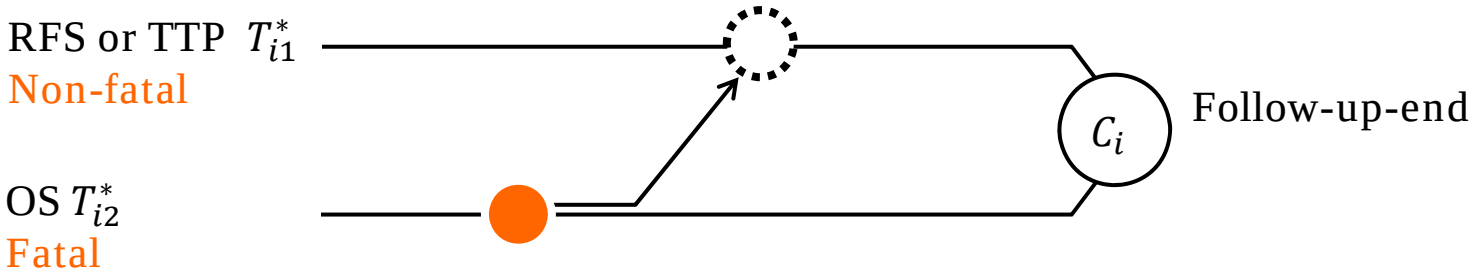
How to manage Type I error in two event-time clinical trials

- The information fraction (standardized internal time) for the two outcomes at a certain point in time will almost never be the same
  - How should α be allocated to each interim analysis for two endpoints?
  - What is a better strategy for early efficacy stopping in terms of efficiency (power, sample sizes, and event numbers)? How should events be monitored? Both or either of events?

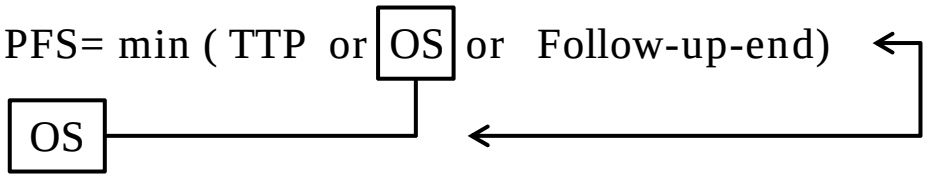


Censoring scheme and composite endpoint

- **Censoring schemes: Dependent censoring (Semi-competing risks)**



- **Censoring schemes: Dependent censoring (Semi-competing risks)**



should be correlated, depending on how much OS events are observed

- critical values for each outcome can be determined separately, by using any group-sequential method such as Lan-DeMets error-spending method

Research objectives

1	**To discuss group-sequential methods for clinical trials with semi-competing risks outcomes, as an extension of our previous works in fixed-sample designs (Hamasaki et al., 2013; Sugimoto et al., 2013, 2017)** ☐ Two intervention comparison ☐ Two situations: (1) non-fatal, non-composite outcome and fatal outcome, and (2) composite outcome including non-fatal and fatal outcomes and fatal outcome ☐ Normal approximation methods
2	**To discuss three strategies to monitor the two event-time outcomes in group-sequential clinical trials, with an illustration** ☐ Early stopping for efficacy ☐ **Multiple primary endpoints** and **co-primary** endpoints ☐ Critical boundary determination using alpha-spending ☐ A joint distribution defined by Clayton copula ☐ Maximum sample size, maximum events, and average events ☐ Evaluation by Simulation

Hamasaki T et al. Pharm Statist 2013;**12**:28-34. Sugimoto T et al. Biostatistics 2013;**14**:409-421. Sugimoto T et al. Statist Med 2017;**36**:1363-1382.

2. Technical development

A trial design and statistical settings

τ: Maximum follow-up time
$\tau = \tau_a + \tau_f$

Total sample size $N = n^{(1)} + n^{(2)}$

$n^{(1)} = rN$

1:Test

2: Control

$n^{(2)} = (1-r)N$

Allocation ratio $r\ (0 < r < 1)$

Non-Fatal (T_{i1}^*, C_{i1})
Fatal (T_{i2}^*, C_{i2})

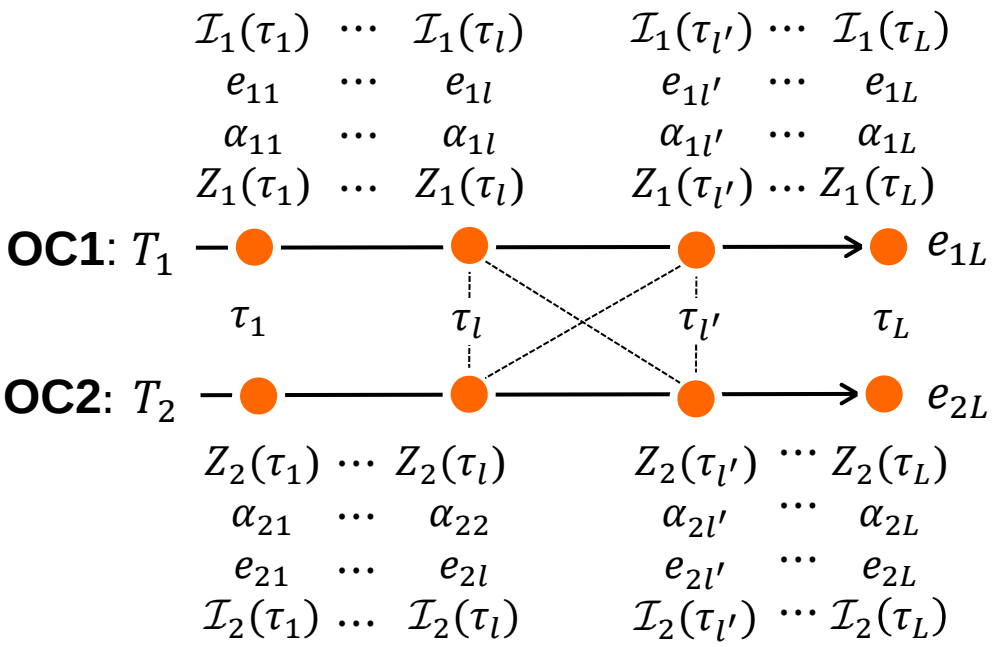
$T_{i1}^{NC} = \min(T_{i1}^*, C_i)$

$T_{i1}^{C} = \min(T_{i1}^*, T_{i2}^*, C_i)$

$T_{i2} = \min(T_{i2}^*, C_i)$

- T_{ik}^*: underlying continuous survival time
- C_{ik}: potential censoring time $(C_{i1} = C_{i2} = C_i)$
  - outcome k $(k = 1,2)$
  - subject i $(i = 1, \ldots, N)$
  - intervention group j $(j = 1,2)$
  - $j = 1 \rightarrow$ T; $j = 2 \rightarrow$ C
- Observed bivariate survival data
  $\{(T_{i1}, T_{i2}, \Delta_{i1}, \Delta_{i2}, g_i)\}_{i=1}^{N}$
  $T_{ik} = \min(T_{ik}^*, C_{ik})$
  $\Delta_{ik} = \mathbb{I}(T_{ik}^*, C_{ik})$: right-censoring indicator
  $\mathbb{I}(\cdot)$: the index function
  g_i: group index $g_i = 1 \rightarrow$ T; $g_i = 2 \rightarrow$ C

Technical development outline



$Z_l(\tau_l)$: logrank test statistics for for kth endpoint at lth analysis at calendar time τ_l

α_{kl}: allocated significance level allocated to lth analysis for kth endpoint

$\mathcal{I}_k(\tau_l)$: information for kth endpoint at lth analysis (time τ_l)

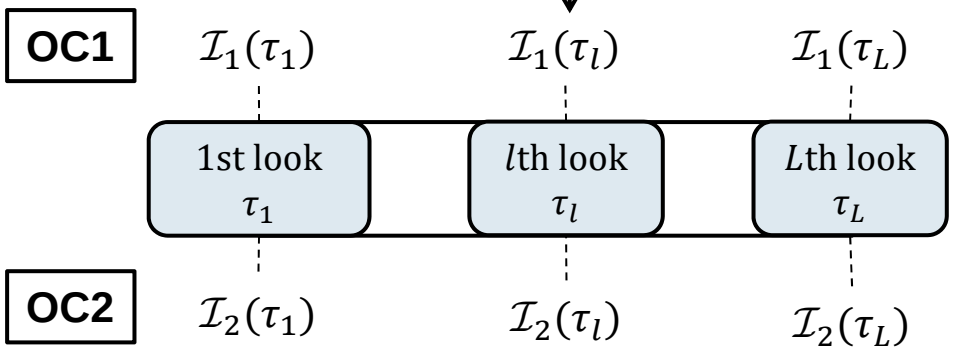
e_{kl}: cumlative number of events at lth analysis

- Assume that each marginal for outcomes is an exponential distribution
- Consider a sequence of two logrank statistics $\{Z_1(\tau_1), \ldots, Z_1(\tau_L), Z_2(\tau_1), \ldots, Z_2(\tau_L)\}$
- $\{Z_k(\tau_l)\}(l = 1, \ldots L, k = 1 \ldots 2)$ are **approximately multivariate-normally distributed** for large sample,
  - evaluating semi-competing risks and composited form relationships
  - Derive asymptotic variance and variance-covariance functions for two sequential logrank statistics

Information and standardized internal time

Asymptotic form of the Fisher's information- asymptotic variance

$$\{\mathcal{I}_{1l}(\tau_l)\}^{-1} = V_{11}^{(0)}(\tau_l) = \int_0^{\tau_l} \{H_1(t|\tau_l)\}^2 \left\{ \frac{\mathrm{d}\Lambda_1^{(1)}(t)}{a^{(2)}\mathcal{Y}_1^{(2)}(t|\tau_l)} + \frac{\mathrm{d}\Lambda_1^{(2)}(t)}{a^{(1)}\mathcal{Y}_1^{(1)}(t|\tau_l)} \right\}$$



The standardized internal time

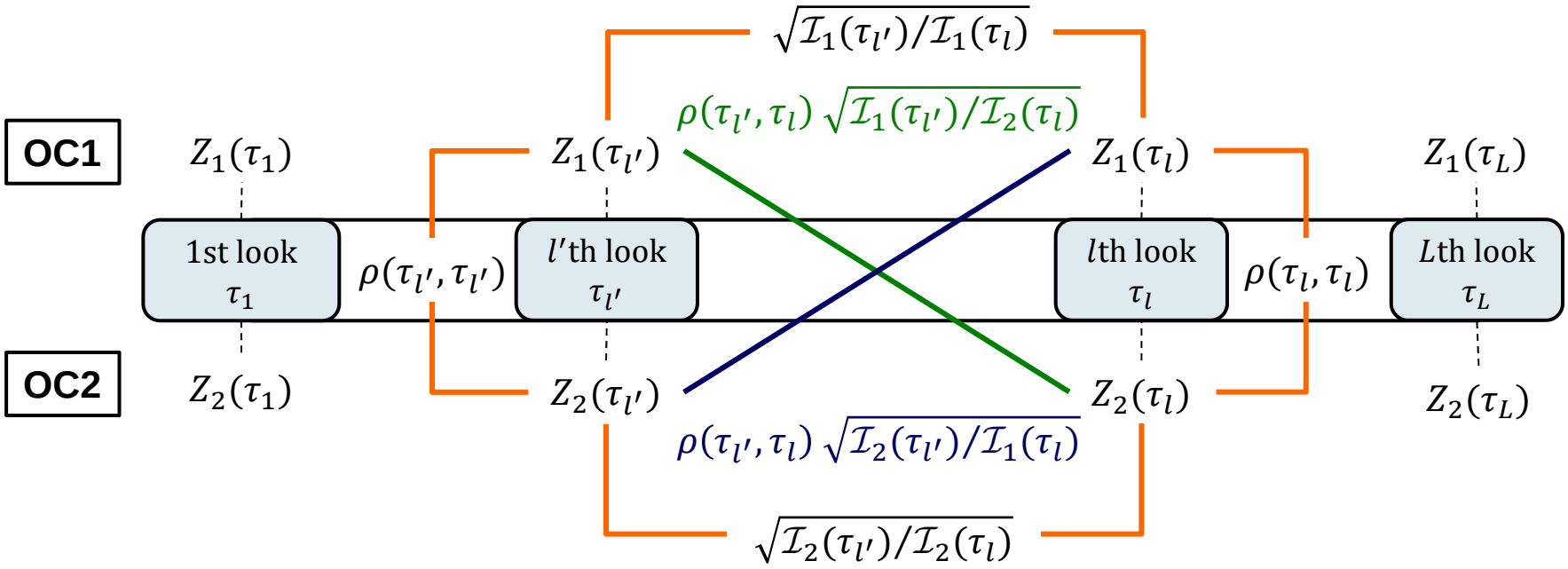
$$t_{kl} = \frac{\mathcal{I}_{kl}}{\mathcal{I}_{kL}} = \frac{V_{kk}^{(0)}(\tau_L)}{V_{kk}^{(0)}(\tau_l)}$$

$$\{\mathcal{I}_{2l}(\tau_l)\}^{-1} = V_{22}^{(0)}(\tau_l) = \int_0^{\tau_l} \{H_2(t|\tau_l)\}^2 \left\{ \frac{\mathrm{d}\Lambda_2^{(1)}(t)}{a^{(2)}\mathcal{Y}_2^{(2)}(t|\tau_l)} + \frac{\mathrm{d}\Lambda_2^{(2)}(t)}{a^{(1)}\mathcal{Y}_2^{(1)}(t|\tau_l)} \right\}$$

$\mathcal{I}_k(\tau_l)$: the information at τ_l for kth outcome

Calendar and information time

Correlation among the two sequential logrank test statistics



$$\rho(\tau_{l'},\tau_l)=\mathrm{corr}[Z_k(\tau_{l'}),Z_k(\tau_l)]\approx V_{kk}(\tau_{l'},\tau_l)/\sqrt{V_{kk}(\tau_{l'},\tau_{l'})V_{kk}(\tau_l,\tau_l)}$$

$$V_{kk}(\tau_l,\tau_{l'})=\sqrt{\frac{r_{l\wedge l'}}{r_{l\vee l'}}}\int_0^{\tau_l\wedge\tau_{l'}}H_k(t|\tau_l)H_k(t|\tau_{l'})\left\{\frac{\mathrm{d}\Lambda_k^{(1)}(t)}{a^{(1)}\mathcal{Y}_k^{(1)}(t|\tau_l\vee\tau_{l'})}+\frac{\mathrm{d}\Lambda_k^{(2)}(t)}{a^{(2)}\mathcal{Y}_k^{(2)}(t|\tau_l\vee\tau_{l'})}\right\}$$

An Illustration: calculated standardized internal time/OC1

ψ_1^{-1}	ψ_2^{-1}	$S_1^{(2)}(60)$	$S_2^{(2)}(60)$	ρ	Anal. #	Both non-fatal			One fatal			One fatal		
						C. time	OC1	OC2	C. time	OC1: NC	OC2: F	C. time	OC1: C	OC2: F
1.5	1.5	0.6	0.5	0.0	1	34.1	**0.5**	0.518	31.7	**0.5**	0.466	31.7	**0.5**	0.466
					2	60.0	**1.0**	1.0	60.0	**1.0**	1.0	60.0	**1.0**	1.0
				0.8	1	34.1	**0.5**	0.518	29.7	**0.5**	0.423	30.4	**0.5**	0.437
					2	60.0	**1.0**	1.0	60.0	**1.0**	1.0	60.0	**1.0**	1.0
		0.7	0.5	0.0	1	34.7	**0.5**	0.530	32.2	**0.5**	0.478	32.2	**0.5**	0.478
					2	60.0	**1.0**	1.0	60.0	**1.0**	1.0	60.0	**1.0**	1.0
				0.8	1	34.7	**0.5**	0.530	29.7	**0.5**	0.422	31.1	**0.5**	0.452
					2	60.0	**1.0**	1.0	60.0	**1.0**	1.0	60.0	**1.0**	1.0
2.0	1.5	0.6	0.5	0.0	1	34.4	**0.5**	0.524	32.0	**0.5**	0.472	31.9	**0.5**	0.471
					2	60.0	**1.0**	1.0	60.0	**1.0**	1.0	60.0	**1.0**	1.0
				0.8	1	34.4	**0.5**	0.524	29.8	**0.5**	0.425	30.6	**0.5**	0.443
					2	60.0	**1.0**	1.0	60.0	**1.0**	1.0	60.0	**1.0**	1.0

τ_a =24, τ_f =36. Bivariate exponential distribution is defined by Clayton copula (Clayton DG. *Biometrika* 1978; 65:14-151).

Two issues in the method

Normal approximation-based method

- How much does the method work?: Evaluate the practical utility of the normal approximation method via Monte-Carlo simulation in terms of power and Type I error

Standardized internal time for non-fatal or composite outcome

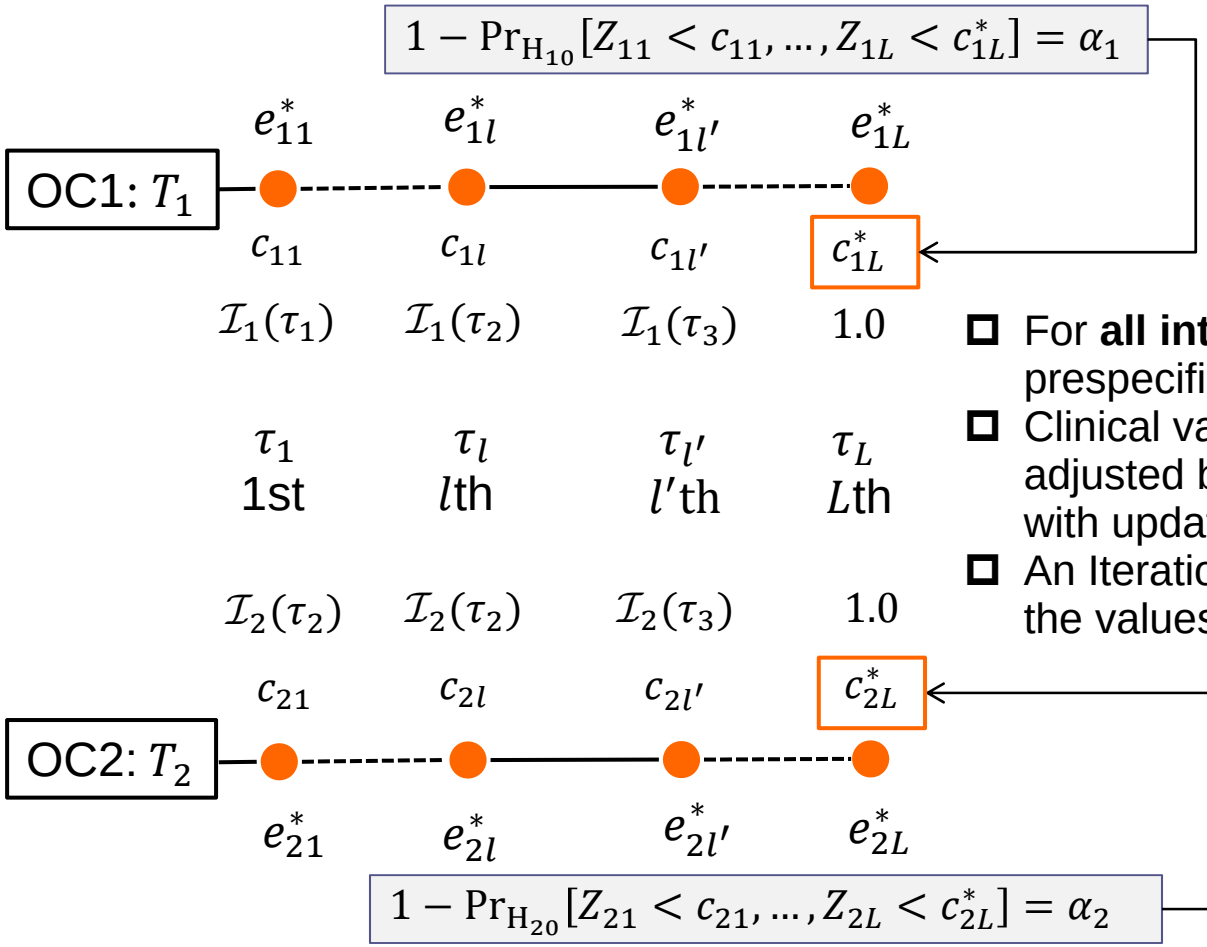
- Standardized internal time for non-fatal outcome (TTP) or composite outcome (MACE, PFS) is effected by censoring scheme and composite form with the parameters (e.g., cumulative survival, hazard ratio) of fatal outcome, but standardized internal time for fatal outcome is not.
- **At the planning stage of a trial**, by using the method with two outcomes association structure, critical boundary can be prespecified, and the power, sample size, maximum events and average events can be evaluated
- **During the trial**, how can the method be implemented?
  - Miss-specification of two outcomes association structure may be a issue in controlling the Type I error
  - Need to update the critical value based on the observed events, but how?

An Illustration: calculated standardized internal time/calendar time

ψ_1^{-1}	ψ_2^{-1}	$S_1^{(2)}(60)$	$S_2^{(2)}(60)$	ρ	Anal. #	Calen. time	Both nonfatal		One fatal		One fatal	
							OC1	OC2	OC1	OC2 Fatal	OC1 Comp.	OC2 Fatal
1.5	1.5	0.6	0.5	0.0	1	36	0.543	0.558	0.599	0.558	0.599	0.558
					2	60	1.0	1.0	1.0	1.0	1.0	1.0
				0.8	1	36	0.543	0.558	0.636	0.558	0.621	0.558
					2	60	1.0	1.0	1.0	1.0	1.0	1.0
		0.7	0.5	0.0	1	36	0.530	0.558	0.587	0.558	0.587	0.558
					2	60	1.0	1.0	1.0	1.0	1.0	1.0
				0.8	1	36	0.530	0.558	0.638	0.558	0.607	0.558
					2	60	1.0	1.0	1.0	1.0	1.0	1.0
2.0	1.5	0.6	0.5	0.0	1	36	0.542	0.558	0.600	0.558	0.599	0.558
					2	60	1.0	1.0	1.0	1.0	1.0	1.0
				0.8	1	36	0.542	0.558	0.639	0.558	0.619	0.558
					2	60	1.0	1.0	1.0	1.0	1.0	1.0

τ_a =24, τ_f =36. Bivariate exponential distribution is defined by Clayton copula (Clayton DG. *Biometrika* 1978; 65:14-151).

critical value adjustment based on observed events: our approach



$$\text{corr}[Z_{1k}, Z_{1k'}] = \sqrt{\frac{e_{1k}^*}{e_{1k'}^*}}$$

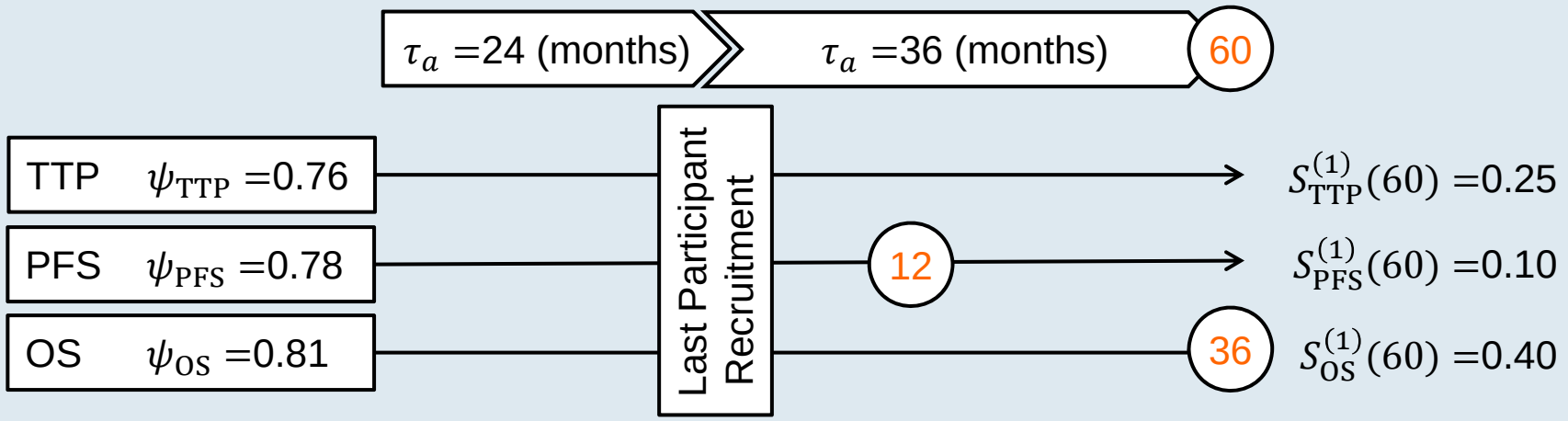
$$\text{corr}[Z_{2k}, Z_{2k'}] = \sqrt{\frac{e_{2k}^*}{e_{2k'}^*}}$$

- For **all interim analyses**, prespecified critical values are used
- Clinical value **at the final analysis** is adjusted based on observed events, with updated correlation matrix.
- An Iteration procedure is need to find the values c_{1L}^* and c_{2L}^*

3. Co-primary endpoints

An illustration: ICON7 trial

A randomized (1:1 ratio), 2-arm, multi-center, open-label phase III trial designed to evaluate the safety and efficacy of adding bevacizumab, a humanized monoclonal antibody against Vascular Endothelial Growth Factor (VEGF), to standard chemotherapy with carboplatin and paclitaxel, in patients with ovarian cancer (Perren et al, 2011)



- At a 5% significance level of two-sided test, 90% power PFS (674 events) and 80% power for OS (715 events) (**1520 participants recruited**)
- Implicitly assumed PPS and OS are independent ---90% × 80% → 72% power

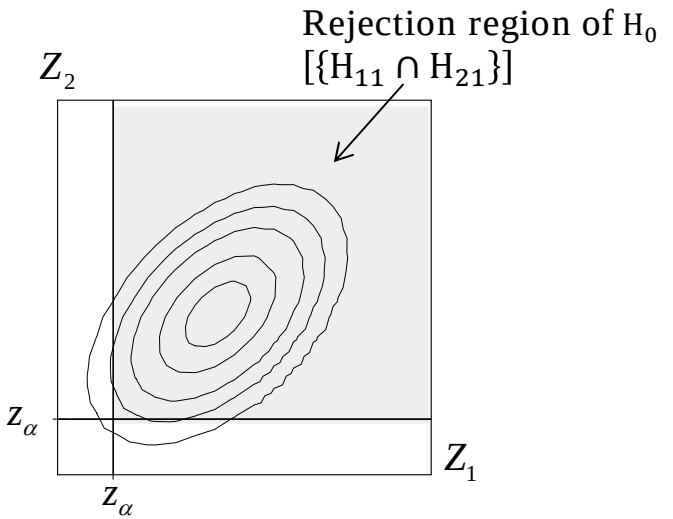
Perren TJ et al. N Engl J Med 2011; **365**:2484-2496.

"Co-primary" endpoints

Hypothesis for co-primary

$$\begin{cases} H_0: H_{10} \cup H_{20} \\ H_1: H_{11} \cap H_{21} \end{cases}$$

$$\begin{cases} H_{k0}: \psi_k(t) \geq 1, & \text{for all } t \\ H_{k1}: \psi_k(t) < 1, & \text{at some } t \end{cases}$$



$\psi_k \cdots$ hazard ratio for Endpoint $k (k = 1,2)$

$Z_k \cdots$ logrank test statistics for Endpoint k

$\alpha \cdots$ significant level for hypothesis testing

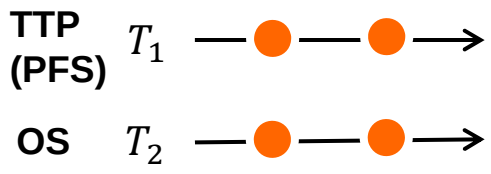
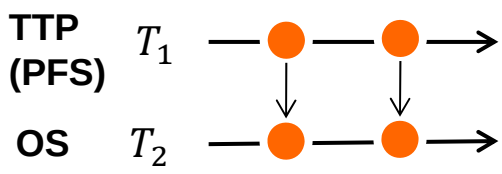
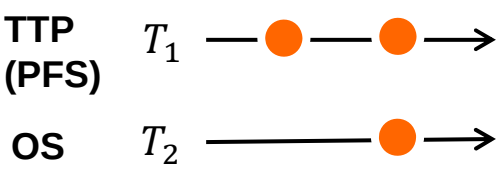
$z_\alpha \cdots$ the upper α-th percent point of Z_k

- ✓ Significance on **BOTH** endpoints being sufficient for proof of effect
- ✓ **No adjustment** for control of the Type I error rate between the endpoints, but **need for adjustment** among the analyses.
- ✓ Type II error rate increases as the number of endpoints to be tested increases- need sample size adjustment to maintain the overall power

FDA. Multiple Endpoints in Clinical Trials 2017. CPMP. Multiplicity Issue in Clinical Trials. 2017

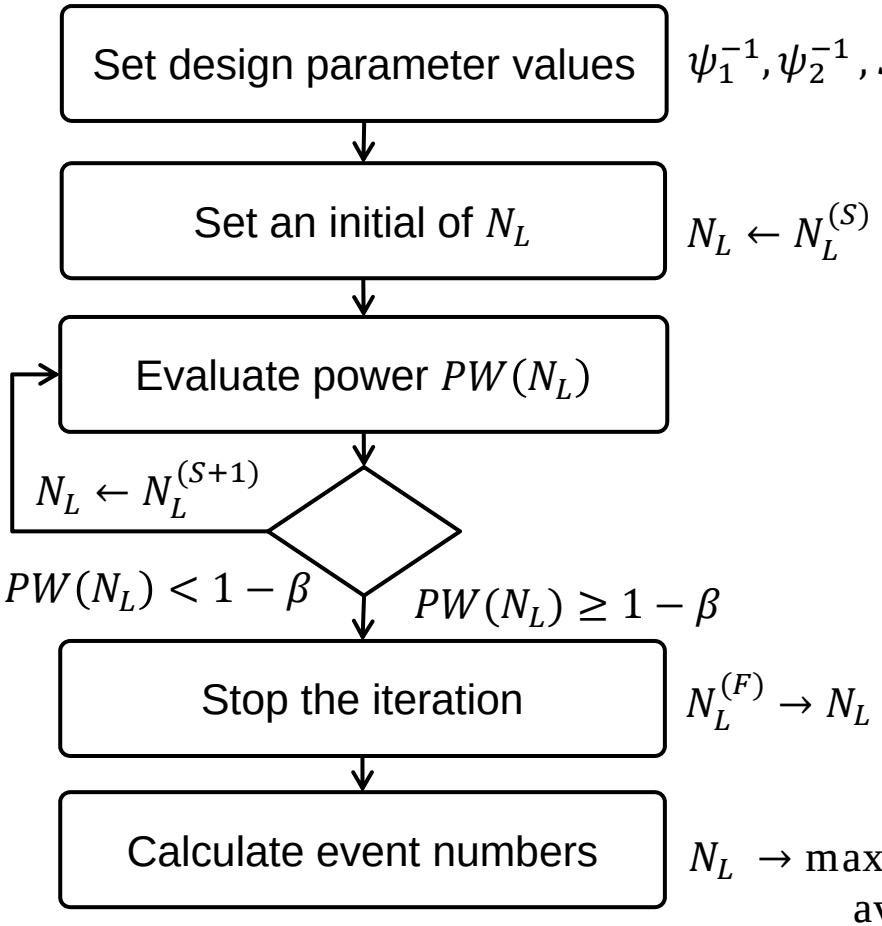
Strategies for rejecting null hypothesis: co-primary endpoints

- Co-primary hypothesis: $H_0: H_{10} \cup H_{20}$ versus $H_1: H_{11} \cap H_{21}$
- Two analyses: first (36M) & final (60M)/Fixed calendar time
- One sided test at α =2.5%; Power $1-\beta$ of 80%
- Lan-DeMets error-spending method (Lan and DeMets, 1983), using O'Brian-Fleming (OF)-type function for both endpoints
- **Evaluate sample size and event numbers**

ST1	ST2	ST3
		
☐ Monitor Both: TTP (PFS) and OS ☐ TTP and OS are rejected at any interim, not necessarily simultaneously at the same analysis	☐ Monitor TTP (PFS) first ☐ Test OS if TTP (PFS) has been rejected (TTP (or PFS) is not tested again)	☐ Monitor TTP (PFS) only ☐ Test OS only at the final ☐ TTP (PFS) is not tested again at the final if TTP (PFS) has been rejected at the interim

Lan KKG, DeMets DL. Biometrika 1983; **70**:659–663

Calculation for sample size and event numbers



Ex. Power for ST1

$$PW(N_L) = \Pr_{\boldsymbol{\psi}=\boldsymbol{\psi}^*}\left[\left\{\bigcup_{l=1}^{L} A_{1l}\right\} \cap \left\{\bigcup_{l=1}^{L} A_{2l}\right\}\right]$$

$$A_{kl} = \{Z_{kl}(\tau_l) > c_{kl}(\alpha)\}$$

$$\bar{A}_{kl} = \{Z_{kl}(\tau_l) \leq c_{kl}(\alpha)\}$$

ICON7: Calculated internal time and corresponding critical boundary

ρ	Calen. time	ST 1 and 2				ST 3			
		Information time		OF-type bound.		Information time		OF-type bound.	
		TTP	**OS**	**TTP**	**OS**	**TTP**	**OS**	**TTP**	**OS**
0.0	36 60	0.6886 1.0	0.5799 1.0	2.4619 1.9974	2.4619 1.9974	0.6886 1.0	0.5799 1.0	2.4619 1.9974	– 1.9600
0.5	36 60	0.6850 1.0	0.5799 1.0	2.4695 1.9966	2.4619 1.9974	0.6850 1.0	0.5799 1.0	2.4695 1.9966	– 1.9600
0.8	36 60	0.5758 1.0	0.5799 1.0	2.7330 1.9773	2.4619 1.9974	0.5758 1.0000	0.5799 1.0	2.7330 1.9773	– 1.9600
ρ	Calen. time	**PPS**	**OS**	**PPS**	**OS**	**PPS**	**OS**	**PPS**	**OS**
0.0	36 60	0.6883 1.0	0.5799 1.0	2.4624 1.9973	2.4619 1.9974	0.6883 1.0	– 1.0	2.4624 1.9973	– 1.9600
0.5	36 60	0.6925 1.0	0.5799 1.0	2.4537 1.9982	2.4619 1.9974	0.6925 1.0	– 1.0	2.4537 1.9982	– 1.9600
0.8	36 60	0.6221 1.0000	0.5799 1.0	2.6130 1.9845	2.4619 1.9974	0.6221 1.0000	– 1.0	2.6130 1.9845	– 1.9600

Bivariate distribution is given by Clayton copula and correlation between cumulative hazards is defined by Pearson-type correlation (Hsu L, Prentice RL. Biometrika 1996; **83**:491–506)

Calculated sample sizes and event numbers: TTP and OS

		Max. sample size	Max. events		Ave. events		Empirical power (%)		
			TTP	OS	TTP	OS	Joint	TTP	OS
$\rho = 0.0$	Fixed sample design	1628	1005	784	1005	784	80.1	96.4	83.1
	ST1	1638	1011	789	803	689	80.0	96.1	83.2
	ST2: TTP→OS	1639	1012	790	803	707	80.2	96.2	80.2
	ST3: TTP→OS	1630	1006	785	799	785	80.0	96.1	80.0
$\rho = 0.5$	Fixed sample design	1693	1045	816	1045	816	80.2	94.0	84.7
	ST1	1703	1051	821	840	712	80.1	93.9	84.6
	ST2: TTP→OS	1704	1052	821	841	733	80.2	94.0	80.2
	ST3: TTP→OS	1695	1046	817	837	817	80.0	93.9	80.0
$\rho = 0.8$	Fixed sample design	1658	1023	1023	799	799	80.0	93.9	83.9
	ST1	1671	1031	832	699	699	80.1	94.0	84.0
	ST2: TTP→OS	1674	1033	833	725	725	80.0	94.0	80.0
	ST3: TTP→OS	1663	1026	828	801	801	80.1	93.9	80.1

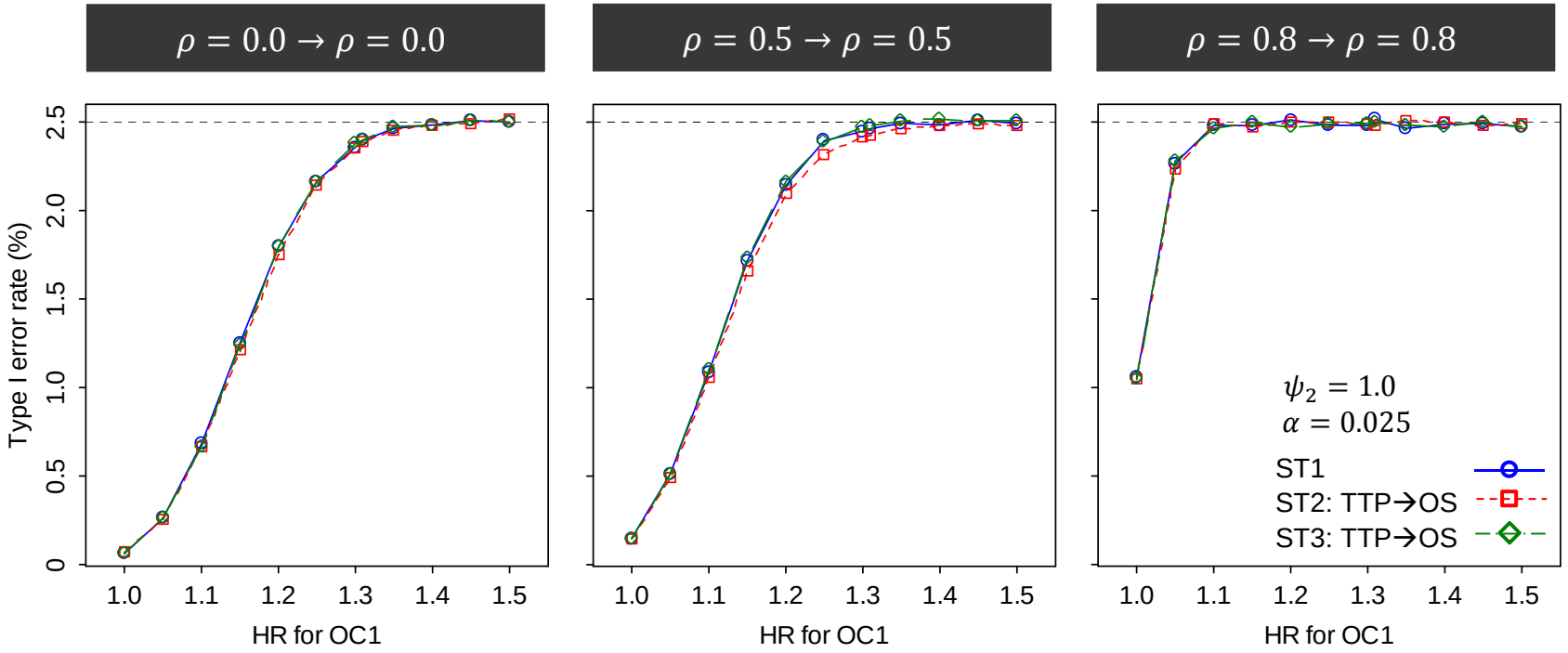
Empirical power is evaluated with 100,000 runs. Bivariate distribution is given by Clayton copula (Clayton, 1976). Correlation between cumulative hazards is defined by Pearson-type correlation (Hsu and Prentice 1996)

Calculated sample sizes and event numbers: PPS and OS

		Max. sample size	Max. events		Ave. events		Empirical power (%)		
			TTP	OS	TTP	OS	Joint	TTP	OS
$\rho = 0.0$	Fixed sample design	1510	932	728	932	728	79.9	98.9	80.2
	ST1	1520	938	732	717	645	80.1	98.9	80.4
	ST2: TTP→OS	1521	939	733	717	646	80.0	98.9	80.0
	ST3: TTP→OS	1510	932	728	712	728	80.2	98.8	80.2
$\rho = 0.5$	Fixed sample design	1543	952	744	952	744	80.1	96.0	81.3
	ST1	1550	957	747	744	656	79.9	95.9	81.0
	ST2: TTP→OS	1551	957	747	743	658	80.1	96.0	80.1
	ST3: TTP→OS	1541	951	743	739	743	79.9	95.8	79.9
$\rho = 0.8$	Fixed sample design	1563	965	753	965	753	80.1	93.4	81.7
	ST1	1570	969	757	765	663	80.0	93.3	81.5
	ST2: TTP→OS	1572	970	757	765	666	79.9	93.3	79.9
	ST3: TTP→OS	1562	964	753	761	753	79.9	93.2	79.9

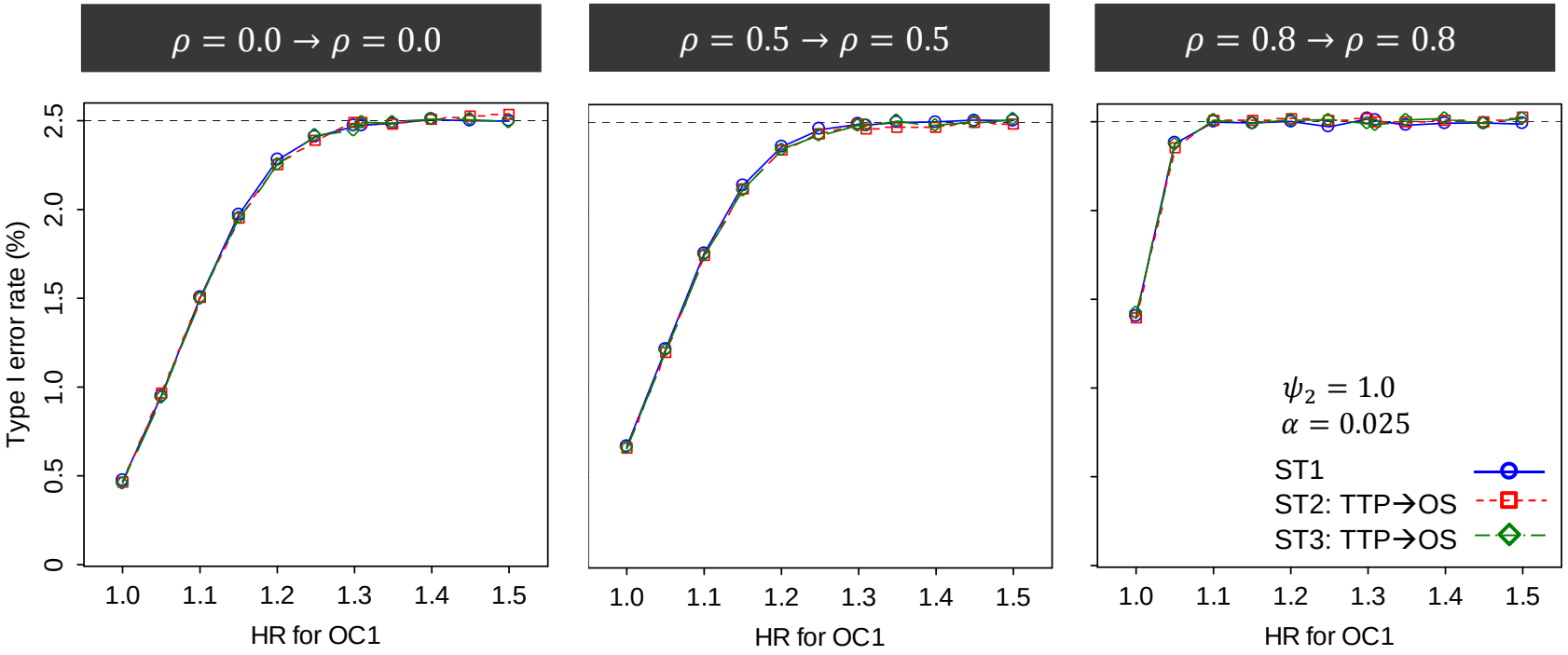
Empirical power is evaluated with 100,000 runs. Bivariate distribution is given by Clayton copula (Clayton, 1976). Correlation between cumulative hazards is defined by Pearson-type correlation (Hsu and Prentice 1996)

Type I error rate behaviors: TTP and OS



Type I error rate is evaluated with 1,000,000 runs. Bivariate data is generated by Clayton copula (Clayton, 1976). Correlation between cumulative hazards is defined by Pearson-type correlation (Hsu and Prentice, 1996)

Type I error rate behaviors: PFS and OS



Type I error rate is evaluated with 1,000,000 runs. Bivariate data is generated by Clayton copula (Clayton, 1976). Correlation between cumulative hazards is defined by Pearson-type correlation (Hsu and Prentice 1996)

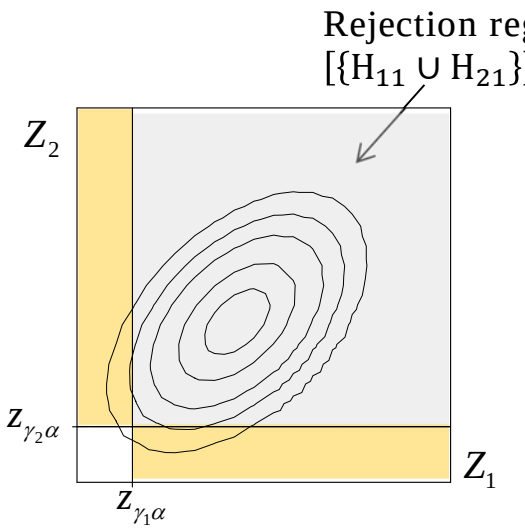
4. Multiple primary endpoints

Multiple primary endpoints

Hypothesis for at least one

$$\begin{cases} H_0: H_{10} \cap H_{20} \\ H_1: H_{11} \cup H_{21} \end{cases}$$

$$\begin{cases} H_{k0}: \psi_k(t) \geq 1, & \text{for all } t \\ H_{k1}: \psi_k(t) < 1, & \text{at some } t \end{cases}$$



$\psi_k \cdots$ hazard ratio for Endpoint $k(k = 1,2)$

$Z_k \cdots$ logrank test statistics for Endpoint k

$\alpha \cdots$ significant level for hypothesis testing

$z_{\gamma_k a} \cdots$ the upper $\gamma_k\alpha$-th percent point of Z_k

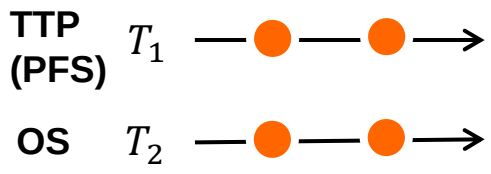
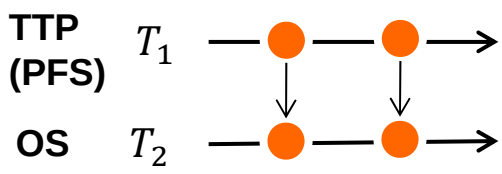
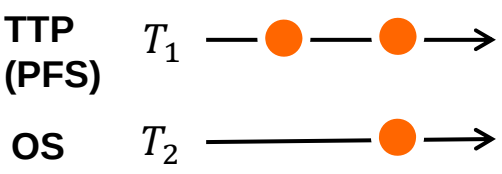
$\gamma_k \cdots$ weight $\gamma_1 + \gamma_2 = 1$

- ✓ Significance on **at least one** endpoint being sufficient for proof of effect
- ✓ **Need adjustment** for control of the Type I error rate between the endpoints, and **need for adjustment** among the analyses.

FDA. Multiple Endpoints in Clinical Trials 2017. CPMP. Multiplicity Issue in Clinical Trials. 2017

Strategies for rejecting null hypothesis: Multiple primary endpoints

- Multiple primary hypothesis: $H_0: H_{10} \cap H_{20}$ versus $H_1: H_{11} \cup H_{21}$
- Two analyses: first (36M) & final (60M)/Fixed calendar time
- One sided test at α =2.5%
- Lan-DeMets' Error-spending method (Lan and DeMets, 1983), using O'Brian-Fleming (OF)-type function for both endpoints
- **Evaluate empirical power under 1520 subjects**

ST1	ST2	ST3
		
☐ Weighted Bonferroni procedure ☐ Monitor both ☐ Two outcomes are rejected at any interim, ☐ Weight: w_1+w_2 = 1	☐ Fixed-sequence procedure ☐ Monitor TTP (PPS) first ☐ Test OS if TTP has been rejected ☐ Consider other order (OS→TTP/PFS)	☐ Monitor TTP (PPS) ☐ Test OS only at the final ☐ TTP (PPS) will be not tested again at the final if TTP has been rejected at the interim ☐ Consider other order (OS→TTP/PFS)

Lan KKG, DeMets DL. Biometrika 1983; **70**:659–663

Empirical power for the strategies: TTP and OS

ρ	Strategy	ALO	Both	TTP	OS
0.0	ST1: w_1=0.3 for TTP	97.1	66.9	88.2	75.9
	ST1: w_1=0.5 for TTP	97.7	65.7	91.5	71.9
	ST1: w_1=0.8 for TTP	97.6	56.3	93.9	60.0
	ST2: TTP → OS	95.0	76.3	95.0	76.3
	ST3: TTP → OS	95.0	76.6	95.0	76.6
	ST2: OS → TTP	63.1	62.1	62.1	63.1
	ST3: OS → TTP	63.2	62.3	62.3	63.2
0.5	ST1: w_1=0.3 for TTP	94.0	63.3	81.3	76.0
	ST1: w_1=0.5 for TTP	94.9	63.2	86.3	71.8
	ST1: w_1=0.8 for TTP	94.8	54.7	89.7	59.8
	ST2: TTP → OS	91.4	74.1	91.4	74.1
	ST3: TTP → OS	91.3	74.7	91.3	74.7
	ST2: OS → TTP	40.7	40.3	40.3	40.7
	ST3: OS → TTP	41.2	40.8	40.8	41.2
0.80	ST1: w_1=0.3 for TTP	92.8	65.4	82.1	76.1
	ST1: w_1=0.5 for TTP	93.6	65.1	86.7	72.1
	ST1: w_1=0.8 for TTP	93.6	56.2	90.2	59.7
	ST2: TTP → OS	91.6	75.5	91.6	75.5
	ST3: TTP → OS	91.6	75.6	91.6	75.6
	ST2: OS → TTP	23.7	23.6	23.6	23.7
	ST3: OS → TTP	23.6	23.5	23.5	23.6

Empirical power is evaluated with 100,000 runs. Bivariate distribution is given by Clayton copula (Clayton, 1976). Correlation between cumulative hazards is defined by Pearson-type correlation (Hsu and Prentice 1996)

Empirical power for the strategies: PPS and OS

ρ	Strategy	ALO	Both	PPS	OS
0.0	ST1: w_1=0.3 for TTP	97.7	75.0	96.6	76.1
	ST1: w_1=0.5 for TTP	98.3	71.4	97.8	71.8
	ST1: w_1=0.8 for TTP	98.7	59.8	98.6	59.9
	ST2: TTP → OS	98.9	79.8	98.9	79.8
	ST3: TTP → OS	98.9	80.4	98.9	80.4
	ST2: OS → TTP	98.9	97.7	97.7	98.9
	ST3: OS → TTP	98.9	97.7	97.7	98.9
0.5	ST1: w_1=0.3 for TTP	92.8	72.4	89.3	75.9
	ST1: w_1=0.5 for TTP	94.2	70.2	92.5	71.9
	ST1: w_1=0.8 for TTP	95.4	59.4	94.8	60.0
	ST2: TTP → OS	95.6	79.1	95.6	79.1
	ST3: TTP → OS	95.7	79.5	95.7	79.5
	ST2: OS → TTP	95.6	95.1	95.1	95.6
	ST3: OS → TTP	95.8	95.3	95.3	95.8
0.80	ST1: w_1=0.3 for TTP	88.7	71.0	83.6	76.1
	ST1: w_1=0.5 for TTP	90.4	69.5	88.0	72.0
	ST1: w_1=0.8 for TTP	91.8	58.9	91.1	59.5
	ST2: TTP → OS	92.7	78.6	92.7	78.6
	ST3: TTP → OS	92.5	78.9	92.5	78.9
	ST2: OS → TTP	92.6	92.5	92.5	92.6
	ST3: OS → TTP	92.6	92.5	92.5	92.6

Empirical power is evaluated with 100,000 runs. Bivariate distribution is given by Clayton copula (Clayton, 1976). Correlation between cumulative hazards is defined by Pearson-type correlation (Hsu and Prentice 1996)

5. Summary and further development

Summary

- Designing multiple event-time outcomes trials that include interim analyses may provide efficiencies by detecting trends prior to planned completion of the trial.
- In such trials, one challenge is how to monitor multiple event-time outcomes in a group-sequential setting as the information fraction for the outcomes may differ at any point in time.
  - discuss logrank test-based methods for monitoring two event-time outcomes in group-sequential trials that compare two interventions when testing if a test intervention is superior to a control intervention on: (i) all event-time outcomes (MCPE) or (ii) at least one of the event-time outcomes (MCP).
  - evaluate two semi-competing risk situations: (a) both events are non-composite but one event is fatal, and (b) one event is composite but the other is fatal and non-composite.
  - derive asymptotic form of variance-covariance function of two sequential logrank test statistics to determine standardized internal time and corresponding critical boundaries, and probability of rejecting the null hypotheses
  - evaluate several strategies for rejecting null hypothesis in early efficacy stopping in clinical trials with MCP and MCPE

Findings

- **The normal approximation-based methods are valid in most practical situation**
  - Based on the result from Monte-Carlo simulation, the methods are valid in most practical situation as long as the sample sizes are not extremely small or unbalanced between the group. All strategies can the control the Type I error and achieve the desired power adequately. In small-sized or unbalanced-sized trials, the exact methods may be considered.
- **Co-primary endpoints**
  - There is **no major difference in power, sample size and event numbers** among the three strategies: the strategy with either outcome being tested only at the final analysis slightly improve the power and decrease the maximum sample size and maximum event numbers, but provides smaller expected number for the outcome monitored during the a trial, while larger expected event numbers for the outcome tested at final, compared with other strategy
- **For multiple endpoints**
  - There is **some difference in disjunctive and conjunctive powers** among the three strategies: the weight to testing, or the order of testing is important to maximize disjunctive and conjunctive powers. Monitoring a log-term outcomes is good idea to maximize the success of a trial.

Summary: advantage of the methods

- The developed method is complicated, but…
  - Can provide the opportunity of evaluating how the relationship between two outcomes impacts on the decision-making for rejecting null hypothesis, in terms of Type I error, power, and sample size and event numbers.
  - Can provide some insight to choose a better strategy for monitoring two event-time outcomes
- An extension to futility assessment, sample size recalculation and conditional power assessment, sensitive subgroup identification, multi-arm trials….

Thank you for your kind attention

If you have any questions, please **e-mail to**

toshi.hamasaki@ncvc.go.jp