Evaluation of possible designs for a 3-arm clinical trial: Comparing a closed-testing design to alternatives

E. Asikanius[1], **K. Rufibach**[1], *J. Bahlo*[2], *G. Bieska*[1], *H.U. Burger*[1]

[1] *F. Hoffmann-La Roche Basel* [2] *German CLL Study Group, Cologne*

Joint Meeting of the IBS, Austro-Swiss and Italian Regions

Milan, June 16, 2015

Generated 2015-06-15 at 20:48:51.

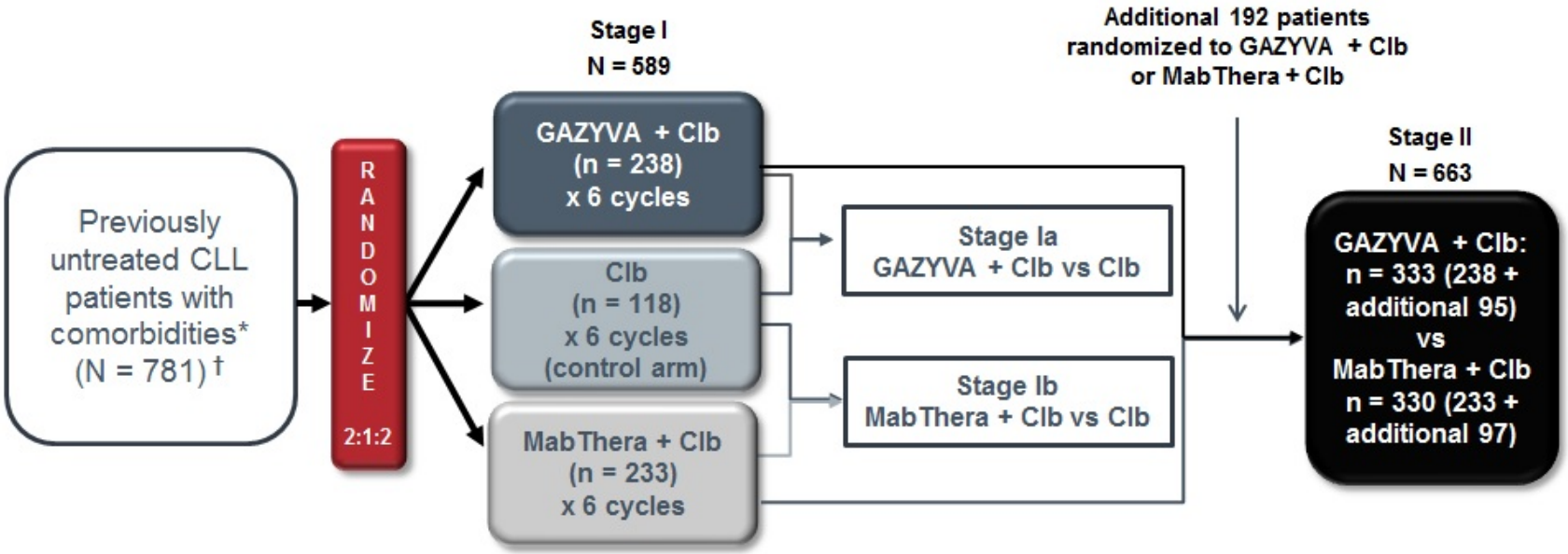
CLL11

Original publication: Goede et al. (2014) in **NEJM**.

Approval of **GAZYVA(RO)** in **chronic lymphocytic leukemia** (CLL).

1st Breakthrough Therapy-designated drug to receive FDA approval, Lee et al. (2014).

Primary endpoint: progression-free survival (PFS) $\Rightarrow$ time-to-event.



* Total CIRS score > 6 and/or CrCl < 70 mL/min; patients with CrCl < 30 mL/min or inadequate liver function excluded; age ≥18 years.
† Plus six additional GAZYVA + Clb patients in safety run-in[2]

Null hypotheses in CLL11

A: Chlorambucil (Cl): Chemotherapy, approved standard.

B: Cl + Rituximab: 1st generation anti-CD20 antibody. Off-label use, never approved in indication.

C: Cl + GAZYVA(RO): 2nd generation anti-CD20 antibody. New drug.

Assume **proportional hazards**. Pairwise null hypotheses:

$$
\begin{aligned}
H_{0,\text{A vs. C}} &: \mathrm{HR}_{\text{A/C}} = 1, \\
H_{0,\text{A vs. B}} &: \mathrm{HR}_{\text{A/B}} = 1, \\
H_{0,\text{B vs. C}} &: \mathrm{HR}_{\text{B/C}} = 1.
\end{aligned}
$$

Null hypotheses in CLL11

All hypotheses of interest $\Rightarrow$ design must protect **familywise error rate** (FWER) **strongly**, i.e.

$$P(\text{reject at least one true null hypothesis}) \quad \leq \quad \alpha$$

irrespective of which null hypothesis are true.

CLL11 used **closed test**: reject pairwise null hypothesis at $\alpha = 0.05$ only if **global** null hypothesis

$$H_{0,\text{global}} \; : \; \text{HR}_{\text{A/C}} \; = \; \text{HR}_{\text{A/B}} \; = \; \text{HR}_{\text{B/C}} \; = \; 1.$$

(that implies all pairwise nulls) is rejected at $\alpha = 0.05$.

Time-to-event endpoint $\Rightarrow$ when to perform global test and pairwise tests?

Goal of this talk

Assume generic scenario reminiscent of CLL11.

Propose different inference strategies.

Questions:

1. **Time to market** is determined by first cutoff A vs. C. Quantify differences?
2. Closed test in Strategies 3 and 4 induces a **power loss** for each pairwise comparison. Quantify power loss, mainly for B vs. C?

Recruitment assumptions:

- $n = 640$ patients in each strategy.
- Randomize 1:2:2.
- 20pts/m for 2m, 40pts/m for 15m.

Assumptions for sample size planning

Global significance level: $\alpha = 0.05$.

Median PFS and corresponding hazard ratios used as alternatives in power computation (assuming **Exponentiality**):

- $\mathrm{HR}_{\mathrm{A/C}} = 12/27 = \mathbf{0.444}$,
- $\mathrm{HR}_{\mathrm{A/B}} = 12/20 = \mathbf{0.600}$,
- $\mathrm{HR}_{\mathrm{B/C}} = 20/27 = \mathbf{0.741}$.

Power:

- A vs. C: $\sim$**98%**. Why overpower? In CLL11 reasons were:
  - Futility and efficacy interim for B vs. C at final analysis of A vs. C $\Rightarrow$ 30% adequate information fraction to perform interim at,
  - length of safety follow up for C had to be large enough to be able to assess benefit-risk,
  - randomization to arm A expected to have terminated at A vs. C analysis cutoff.
- A vs. B: 80%.
- B vs. C: 80%.

Strategies to answer scientific questions

1. **Three separate trials**:
   - Each at $\alpha = 0.05$.
   - Distribute patients on three trials $\Rightarrow$ use each patient for **one** comparison only.
2. **One 3-arm trial with Bonferroni correction**:
   - Each comparison at $\alpha = 0.0167$.
   - All patients in same trial $\Rightarrow$ use each patient for **two** comparisons.
3. **One 3-arm trial with closed testing, wait until last comparison mature**:
   - Wait until targeted number of events for **latest** comparison is reached.
   - Test $H_{0,\text{global}}$.
   - If $H_{0,\text{global}}$ is rejected $\Rightarrow$ perform all pairwise comparisons.
4. **One 3-arm trial with closed testing, each comparions analyzed once mature**:
   - Wait until targeted number of events for **first** comparison is reached.
   - Test $H_{0,\text{global}}$.
   - If $H_{0,\text{global}}$ is rejected $\Rightarrow$ perform first pairwise comparison.
   - Perform other pairwise comparisons once targeted number of events reached.
   - Strategy used in CLL11.

Choice of strategies and power loss

Choice of strategies for comparison:

- All protect FWER.
- Generic approaches to answer scientific questions. Can be fine-tuned in a given application.
- Alternative strategies presumably have operational characteristics somewhere between chosen strategies.

Power loss of pairwise tests in Strategy 4: Test $H_{0,\text{global}}$ when first cutoff (A vs. C) is reached $\Rightarrow$ how much power do we lose for B vs. C?

Methods

Strategies 1, 2:

- **Compute** number of necessary events.
- **Compute** cutoffs for analyses based on that.

Strategies 3, 4:

- **Unadjusted** analysis: **Compute** number of necessary events and cutoff.
- **Adjusted** analysis: Global test gates pairwise tests. Increase number of necessary events from unadjusted analysis until **simulations** (10^6 runs) yield targeted power.

Formulas in backup.

Results - analysis cutoffs

Detailed results in backups.

			A vs. C	A vs. B	B vs. C
Hazard ratio			0.444	0.600	0.741
Strategy 1:		computed #required events	111	136	349
Three separate trials		computed cutoff (months)	**34.4**	39.2	–
Strategy 2:		computed #required events	136	181	465
3-arm with Bonferroni		computed cutoff (months)	**21.4**	24.3	90.1
Strategy 3:	unadj.	computed #required events	275	303	349
3-arm with		computed cutoff (months)	47.2	47.2	47.2
closed testing	adj.	ass. (B vs. C)/resulting (A vs. C/B) #events	276	303	350
		cutoff (months) corresponding to #events	**47.4**	47.4	47.4
Strategy 4:	unadj.	computed #required events	111	136	349
3-arm with		computed cutoff (months)	18.6	19.4	47.2
closed testing	adj.	assumed #required events	111	136	366
		cutoff (months) corresponding to #events	**18.6**	19.4	51.0
	power	simulated power corresponding to #events	0.974	0.807	0.800
		simulated unadj. power corresp. to #events	0.988	0.809	0.817

Patients for each comparison:

- Strategy 1: 64/128; 64/128; 128/128.
- Strategies 2-4: 128/256; 128/256; 256/256.

Results - power loss

Detailed results in backups.

			A vs. C	A vs. B	B vs. C
Hazard ratio			0.444	0.600	0.741
Strategy 1: **Three separate trials**		computed #required events	111	136	349
		computed cutoff (months)	34.4	39.2	–
Strategy 2: **3-arm with Bonferroni**		computed #required events	136	181	465
		computed cutoff (months)	21.4	24.3	90.1
Strategy 3: **3-arm with closed testing**	unadj.	computed #required events	275	303	349
		computed cutoff (months)	47.2	47.2	47.2
	adj.	ass. (B vs. C)/resulting (A vs. C/B) #events	276	303	350
		cutoff (months) corresponding to #events	47.4	47.4	47.4
Strategy 4: **3-arm with closed testing**	unadj.	computed #required events	111	136	349
		computed cutoff (months)	18.6	19.4	47.2
	adj.	assumed #required events	111	136	366
		cutoff (months) corresponding to #events	18.6	19.4	51.0
	power	simulated power corresponding to #events	0.974	0.807	**0.800**
		simulated unadj. power corresp. to #events	0.988	0.809	**0.817**

Patients for each comparison:

- Strategy 1: 64/128; 64/128; 128/128.
- Strategies 2-4: 128/256; 128/256; 256/256.

Results

Results: with CLL11 strategy,

- save between $\sim$ 3**m** and $\sim$ 29**m** to first cutoff,
- $\sim$ 2% **power loss** for B vs. C, corresponding to **17 events** or $\sim$ 4**m**.

Explore strategy based on closed testing in multi-arm trials.

Paper compares strategies with respect to

- operational complexity,
- operational bias,
- difficulty of inference in pairwise comparisons,
- type I error protection for secondary endpoints.
- Sensitivity analysis: CLL11 assumed quite large effect sizes. Strategy also feasible for smaller effect sizes?

Operational aspects in CLL11

Operational bias: Information from ongoing CT causes changes to participant pool, investigator or patient behavior, or other clinical aspects that affect conduct such that conclusions about efficacy or safety are impacted by differences in data collected post public availability of interim results.

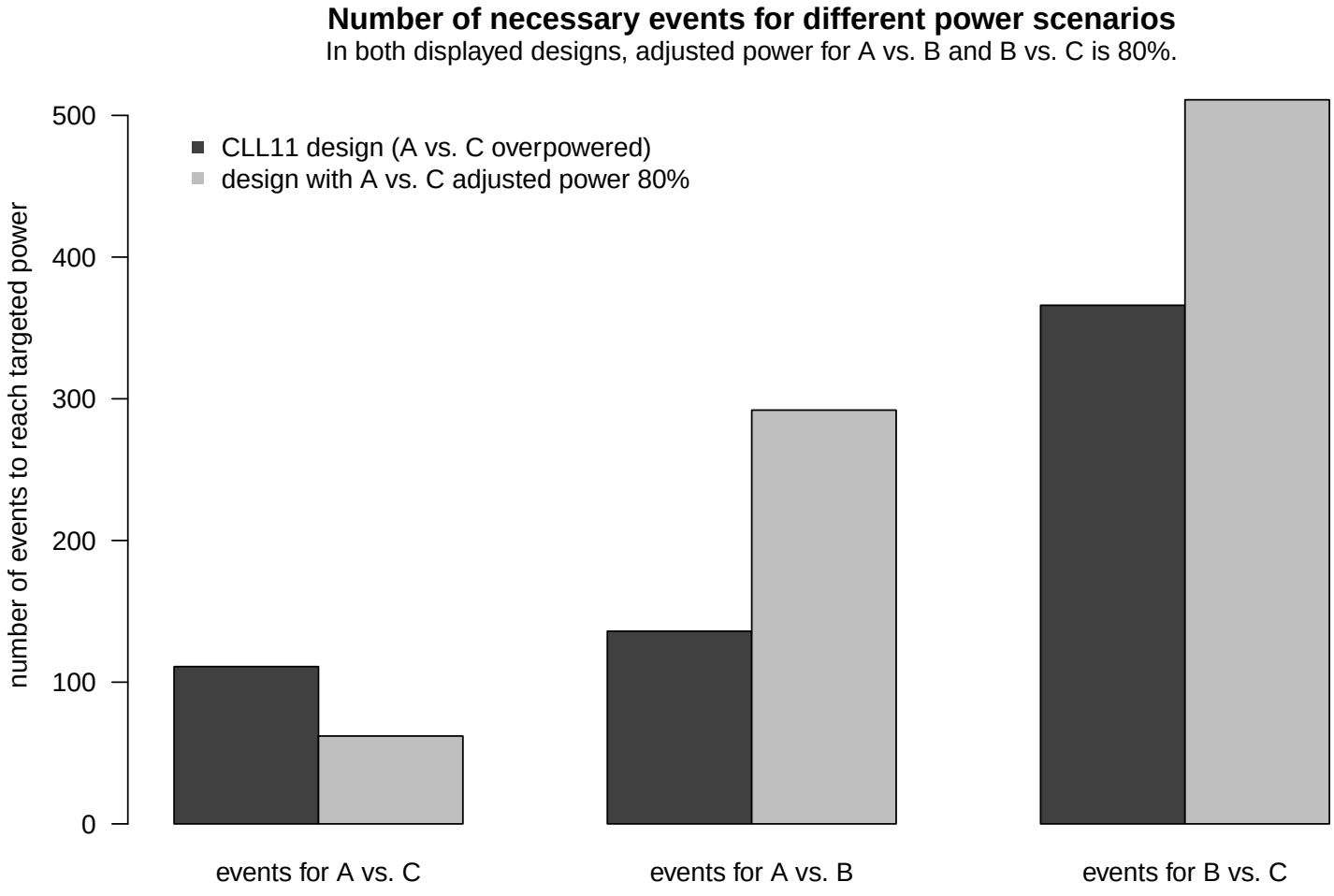
CLL11:

- A vs. C became available quickly.
- Treatment schedule in CLL11 rather fixed once started.
- Define analysis timepoints not only through PFS cutoffs: e.g. all patients needed to be randomized to A prior to cutoff for A vs. C.

Further operational aspects:

- Multiple final / interim analyses on different sets of patients.
- iDMC for interim analyses in B vs. C.
- Independent response review: even more important after A vs. C was unblinded.

Why not power A vs. C with 80%?



* Power for B vs. C for 80% adjusted power design is 76% only, since this is the power corresponding to the maximum number of possible events (= total number of patients) that can be reached with chosen recruitment.

Acknowledgments: We thank Jörg Maurer, Carol Ward, and Daniel Sabanés Bové for helpful discussions and proofreading of paper.

Thank you for your attention.

References

- Cook, T. D. and DeMets, L., David (2008). *Introduction to Statistical Methods for Clinical Trials*. Chapman & Hall.
- Goede, V., Fischer, K., Busch, R., Engelke, A., Eichhorst, B., Wendtner, C. M., Chagorova, T., de la Serna, J., Dilhuydy, M. S., Illmer, T., Opat, S., Owen, C. J., Samoylova, O., Kreuzer, K. A., Stilgenbauer, S., Dohner, H., Langerak, A. W., Ritgen, M., Kneba, M., Asikanius, E., Humphrey, K., Wenger, M. and Hallek, M. (2014). Obinutuzumab plus chlorambucil in patients with CLL and coexisting conditions. *N. Engl. J. Med.* **370** 1101–1110.
- Lee, H. Z., Miller, B. W., Kwitkowski, V. E., Ricci, S., DelValle, P., Saber, H., Grillo, J., Bullock, J., Florian, J., Mehrotra, N., Ko, C. W., Nie, L., Shapiro, M., Tolnay, M., Kane, R. C., Kaminskas, E., Justice, R., Farrell, A. T. and Pazdur, R. (2014). U.s. Food and drug administration approval: obinutuzumab in combination with chlorambucil for the treatment of previously untreated chronic lymphocytic leukemia. *Clin. Cancer Res.* **20** 3902–3907.
- Schoenfeld, D. (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika* **68** 316–319.

Backup slides.

Number of required events

Number of required events d to detect assumed log hazard ratio θ using 2-sided significance level α and power $1 - \beta$. Use Schoenfeld's formula:

$$d \geq \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\kappa(1-\kappa)\theta^2},$$

where z_α is the α-quantile of a standard Normal distribution.

$\kappa \in (0, 1)$ is proportion of patients randomized to arm A.

Schoenfeld (1981) or Section 4.4 in Cook and DeMets (2008).

Computation of cutoff

For given $t_0 > 0$, compute number of events $m = m(t_0)$ expected in cohort recruited at times $a_1 \leq \ldots \leq a_n$ via

$$\begin{aligned} m \;=\; \mathbb{E}(\#\text{events}) \quad &= \quad \mathbb{E}\Big(\sum_{i=1}^{K(t_0)} 1\{\text{event in } (a_i, t_0]\} \Big) \\ &= \quad \sum_{i=1}^{K(t_0)} P\Big(\text{event in } (a_i, t_0]\Big) \\ &= \quad \sum_{i=1}^{K(t_0)} F_\lambda(t_0 - a_i). \end{aligned}$$

F_λ pre-specified CDF with parameter (vector) λ and $K(t_0) \;:=\; \#\{i : a_i \leq t_0\}$.

Throughout, we assume F_λ exponential with rate computed from assumed medians.

No drop-out assumed.

Detailed results

			A vs. C	A vs. B	B vs. C
Hazard ratio			*0.444*	*0.600*	*0.741*
Three separate trials (assuming same total #patients)		significance level	0.05	0.05	0.05
		#patients in each arm	64/128	64/128	128/128
		assumed power	0.980	0.800	0.800
		computed #required events	111	136	349
		computed cutoff (months)	34.4	39.2	–
3-arm with Bonferroni correction		significance level	0.017	0.017	0.017
		#patients in each arm	128/256	128/256	256/256
		assumed power	0.980	0.800	0.800
		computed # required events	136	181	465
		cutoff (months)	21.4	24.3	90.1
3-arm with closed testing (simultaneous analysis)		significance level	0.05	0.05	0.05
		#patients in each arm	128/256	128/256	256/256
	unadjusted analysis	computed # required events	275	303	349
		computed cutoff (months)	47.2	47.2	47.2
		computed power corresponding to #events	1.000	0.987	0.800
	adjusted analysis	#events global test		465	
		assumed (B vs. C) / resulting (A vs. C/B) #events	276	303	350
		cutoff (months) corresponding to #events	47.4	47.4	47.4
		simulated power corresponding to #events	1.000	0.980	0.801
3-arm with closed testing (staggered analysis, as in CLL11)		significance level	0.05	0.05	0.05
		#patients in each arm	128/256	128/256	256/256
	unadjusted analysis	assumed power	*0.980*	*0.800*	*0.800*
		computed # required events	111	136	349
		computed cutoff (months)	18.6	19.4	47.2
	adjusted analysis	#events global test		185	
		assumed # required events	111	136	366
		cutoff (months) corresponding to #events	*18.6*	*19.4*	*51.0*
		simulated power corresponding to #events	0.974	0.807	*0.800*
		simulated unadjusted power corresponding to #events	0.988	0.809	*0.817*

Doing now what patients need next

R version and packages used to generate these slides:

R version: R version 3.1.1 (2014-07-10)

Base packages: stats / graphics / grDevices / utils / datasets / methods / base

Other packages: reporttools / xtable

This document was generated on 2015-06-15 at 20:48:51.