Adaptive design and SAP for interim analysis.

1 Trial design

1.1 General considerations

1.1.1 General design

ALEGORI is a two-stage adaptive randomized trial with dose selection at the interim analysis [1,2]. The first stage consists in a four parallel-arm randomized trial with a 1:1:1:1 randomization ratio, with a placebo and three doses of bevacizumab (50 mg, 75 mg, 100 mg). At the interim analysis, one dose is selected, and the trial proceeds to a second stage, where patients are randomized between placebo and bevacizumab at the selected dose with a 1:2 allocation ratio. It is therefore decided to include n_1 patients per arm at the first stage, and then n_2 patients in the placebo arm and $2 \times n_2$ in the treated arm at the second stage. In addition, it is planned that the number of patients n_2 at the second stage would be reassessed according to the results of the interim analysis (see 1.4.2).

At the interim analysis, no formal stopping rule for efficacy or futility is planned (see 1.4.4).

1.1.2 Primary outcome and test statistics

The primary outcome is the mean monthly epistaxis duration on the 3 months after the end of the treatment (thereafter denoted mean epistaxis duration). According to previous studies, it is assumed that this outcome has approximately a lognormal distribution. The logarithms of the mean epistaxis duration will therefore be analyzed by a Student test with Welch correction of degrees of freedom to allow for different variances between groups.

1.2 Testing strategy and control of the type I error rate

1.2.1 Null hypothesis and type I error rate

Several elementary null hypotheses can be defined, one by tested dose. If we note mi the mean of the logarithm of the mean epistaxis duration in arm *i* (*i*=0 for the placebo arm, and *i*=1, 2, or 3 for the different doses of bevacizumab), and $\theta_i = \mu_i - \mu_0$, i=1, 2, 3, then these null hypotheses are H_i: $\theta_i = 0$. We can additionally define three two-by-two intersection null hypotheses and the global null hypothesis H_{123}: $\theta_1 = \theta_2 = \theta_3 = 0$.

The global type I error rate of the trial is fixed at $\alpha = 0.025$ (with one-sided tests).

1.2.2 Combination test

The basic principle of adaptive designs relies on the combination at the end of the trial of outcomes observed at the different stages, in order to conclude on the possible rejection of the null hypothesis of no efficacy. The final test relies on a so-called combination test, that allows to compute a global test statistic using outcomes of patients included in the different study stages. Results of each stage can be equivalently summarized as test statistics or p-values. We will adopt the latter paradigm. Let us denote p_1 and p_2 the p-values obtained at the first and second stage, respectively. The combination function will be noted $C(p_1, p_2)$. To allow controlling the type I error rate α, one has to define a threshold c_α so that the null hypothesis H_0 shall be rejected at the final analysis if $C(p_1, p_2) \leq c_\alpha$ so that:

$$\int_0^1 \int_0^1 I[C(x,y) \leq c_\alpha] dy dx = \alpha$$

Where I(A) equals 1 if A is true and 0 otherwise.

Several choices of combinations have been published. It is decided to use Fisher's product combination test because of its simplicity, in particular with unbalanced second stages sample sizes [1]:

$$C(p_1, p_2) = p_1 \times p_2$$

The threshold c_α is then defined as $c_\alpha = \exp\{-0.5 \times q_{4,1-\alpha}\}$ where $q_{\nu,1-\alpha}$ is the$(1-\alpha)$-quantile of the chi-square distribution with ν degrees of freedom. Since no early stopping for efficacy is planned, no spending of the type I error rate occurs at the interim analysis, and there is no need to adjust the type I error rate α.

1.2.3 Handling of multiplicity

In this design, several null hypotheses can be defined (see 1.2.1). Since the trial is considered as a confirmatory trial, a strong control of the familywise error rate is deemed necessary [2]. This error corresponds to the maximal probability of rejecting at least one of the elementary null hypotheses that would be true [4]. To control this error, we will apply the closure principle, which allows to reject a null hypothesis H_i at the α level only if all intersection hypotheses that contain H_i are also rejected at the level α [5].

As a protection against multiplicity, the intersection null hypotheses, i.e. hypotheses such as H12 = H1 H2, should be tested using a correction for multiple testing. As part of the study planning, we have conducted a numerical study comparing several methods for correction for multiple testing, namely Bonferroni, Sidak, Simes and Dunnett [see 2 for a description of methods]. The method of Dunnett showed a very small increase of the type I error rate (2.53% instead of 2.5% on 150 000 simulations). This increase is not statistically significant but was found on three consecutive runs of 50 000 simulations. The method was thus not further considered. The method giving the best power while controlling α was Simes's [6] (figure 1), and was therefore retained.

1.2.4 Multiplicity and combination test

The mutlipel testing procedure is only formally used at the first stage, when three elementary null hypotheses are tested (each dose of bevacizumab versus placebo). We then obtain three p-values denoted $p_{1,1}$, $p_{1,2}$ and $p_{1,3}$. The p-values for the intersection null hypotheses are obtained as follows. For $p_{1,ij}$, we first order $p_{1,i}$ and $p_{1,j}$, and we denote their rank by r. The value for $p_{1,ij}$ is then $2 \times \min(p_{1,r}/r)$. Then $p_{1,1}$, $p_{1,2}$ and $p_{1,3}$.are ordered and, still noting r their order, $p_{1,123} = 3 \min (p_{1,r}/r)$.

Only one dose is retained for the second stage, that we denote h. A single p-value can thus be computed, $p_{2,h}$, that is used to test H_h and all intersection hypotheses involving H_h. By denoting k and l the two non-selected doses, four p-values are obtained by combination:

- $C(p_{1,h}, p_{2,h})$
- $C(p_{1,hk}, p_{2,h})$
- $C(p_{1,hl}, p_{2,h})$
- $C(p_{1,123}, p_{2,h})$

We conclude at the efficacy versus placebo of bevacizumab at the dose h is all four combinations p-values are lower than c_α.

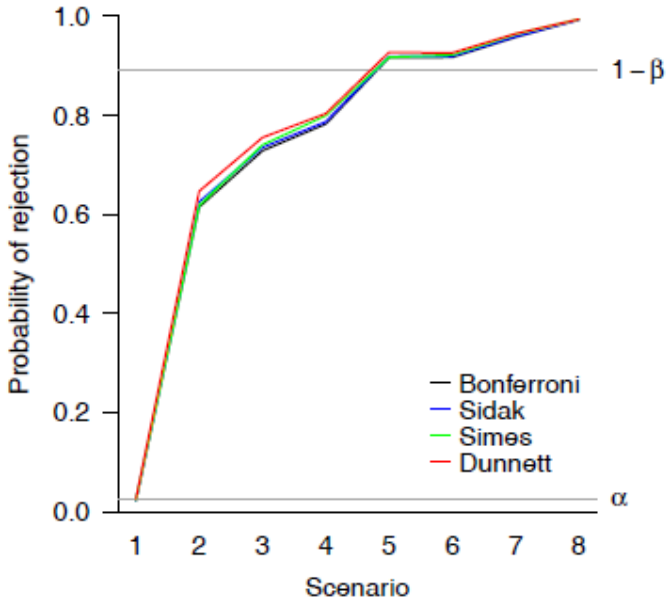

Figure 1. Comparaison of the performances of different methods correcting for multipel testing uder several simulation scenarios.

1.3 Strategy for dose selection

The dose that shows the highest standardized difference against placebo at the interim analysis will be selected for the second stage. This dose is thus the one for which the p-value p_1 is the smallest among $p_{1,1}$, $p_{1,2}$ and $p_{1,3}$.

1.4 Power and sample size

1.4.1 Initial sample size

The sample size was first computed so that each test of the elementary null hypothesis Hi would have a power close to $1 - \beta = 0.9$ to show a relative decrease of 40% in the mean of the mean epistaxis duration as compared to placebo. Given the assumed lognormal distribution of the outcome and data in the literature, this would correspond to showing an absolute difference in the means of the logarithms of the mean epistaxis duration of $\delta = -0.52$ as compared to placebo. We assume that the standard deviation of the logarithm of the mean epistaxis duration is $\sigma = 0.63$ in each group. Assuming a Bonferroni correction (more stringent than Simes correction) for multiple testing, and without accounting for the unbalanced allocation ratio 1:2 at the second stage, a global number of patients of 40 per arm insures a power of 89%. With a 1:2 allocation ratio at the second stage, selecting $n_1 = n_2 = 20$ without interim analysis gives an overall power of 94.4%, thus more than 90%. It is however decided to keep this number of patients to account for the power loss induced by the use of Fisher's combination test, and to maintain a better power in the case bevacizuamb was slightly less effective than expected or if the standard deviation of the outcome was slightly higher than planned.

It is decided to perform the interim analysis when 20 patients per arm will have been randomized.

1.4.2 Sample size reassessment

At the interim analysis, one dose will be selected for the second stage of the trial. It can also be decided to reassess the sample size for the second stage.

This reassessment will be based on conditional power, i.e. the probability to conclude at the efficacy of bevacizumab at the end of the trial given results observed at the interim analysis, if the efficacy is equal to that planned (decrease difference $\delta = -0.52$).

Changing a sample size according to the conditional power is feasible from a statistical point-of-view, without challenging the validity of the trial. Nevertheless, it is preferable that (1) the second stage sample size would not be decreased – in case such a property was deemed desirable, then it should be preferred to use formal stopping rules for efficacy, (2) the planned difference δ would be used rather than the observed difference at the interim analysis and (3) a maximum number of patients beyond which the sample size cannot be increased should be determined [7, 8].

It is planned to reassess the number of patients at the second stage as n'_2 instead of n_2 to maintain the conditional power at 90% at the second stage, but n'_2 should not be larger than $2 \times n_2$. The total number of patients included in the trial should thus not be larger than 200, and would be comprised between 140 (4×20 at the first stage, 20 and 40 at the second stage) and 200 (4×20 at the first stage, 40 and 80 at the second stage).

The conditional power is given by:

$$\varphi(p_1) = 1 - \Phi\left[\Phi^{-1}(1 - c_\alpha/p_1) - \frac{\delta}{\sigma}\sqrt{\frac{2n_2}{3}}\right]$$

Instead of reassessing the sample size in each case, the strategy of the promising zone design is adopted [8]. The sample size will be reassessed only if the conditional power $\varphi(p_1)$ is comprised between a lower bound φ_f and an upper bound φ_u. Several choices for φ_f and φ_u have been studied by numerical simulation, and compared to a strategy without sample size reassessment.

1.4.3 Power considerations

For this type of adaptive design, several definitions of the power can be considered. To compute the sample size, we have used the probability of rejection of elementary null hypotheses, that is to say on a trial that would conclude at the efficacy of any dose that would actually be better than the placebo. Other definitions could be to compute the probability to select the most effective dose ad to conclude at its efficacy, for instance.

Similarly, several types of alternative hypotheses to the global null hypothesis H_{123} can be considered. For instance all θ_i could be equal to $\delta < 0$, or $(\theta_1, \theta_2, \theta_3) = (\delta_1, \delta_2, \delta_3)$ where $\delta_i < 0$ are different from each other. It is also possible to consider cases where $(\theta_1, \theta_2, \theta_3) = (0, \delta_2 < 0, \delta_3 < 0)$ or $(\theta_1, \theta_2, \theta_3) = (0, 0, \delta_3 < 0)$. Different scenarios have thus been studied in a numerical simulation study, and the operating characteristics of the design assessed under these scenarios.

1.4.4 Non-binding futility stopping

No binding stopping rule for futility at the interim analysis is planned. This choice allows a greater flexibility at the second stage. Nonetheless, it will possible at the interim analysis to decide to stop the trial for futility, upon recommendation of the independent data monitoring committee. Conditional power considerations can be used to motivate this decision, in particular by contrasting the conditional power to the conditional error, obtained under H_0. Such a decision cannot lead to inflate the type I error rate, since no early stopping for efficacy (and thus with rejection of the null hypothesis) is allowed.

1.5 Conclusion

The methodology used for this trial allows a strong control of the familywise error rate in the trial. It is also compliant with regulatory recommendations [9, 10].

The slight loss in power associated to fisher's combination test is compensated by a larger planned sample size.

Operational characteristics of the trial have been investigated by numerical simulations.

2 Operational characteristics

The operational characteristics have been obtained by simulating 10 000 fictive trials for each scenario. The standard deviation of the outcome (on the logarithm scale) was fixed at 0.63, and the treatment effect for doses 1, 2 and 3, $\theta = (\theta_1, \theta_2, \theta_3)$ is given in the tables.

Table 1: Performance of the proposed design.

	Without reassessment of n_2	With reassessment of n_2	
$(\varphi_f;\varphi_u)$	(−;−)	(0.7;0.9)	(0.6;0.9)
Simulation Θ=(0,0,0)			
α	2.4%	2.4%	2.4%
Average N	140	150	170.3
Selected dose			
50 mg	33.4%	33.4%	33.4%
75 mg	33.7%	33.7%	33.7%
100 mg	32.9%	32.9%	32.9%
Dose considered as effective			
50 mg	0.8%	0.8%	0.8%
75 mg	0.8%	0.8%	0.8%
100 mg	0.8%	0.8%	0.8%
Simulation Θ=(−0.37,−0.52,−0.7)			
Power	99.4%	99.5%	99.5%
Average N	140	140.4	140.4
Selected dose			
50 mg	3.6%	3.6%	3.6%
75 mg	18.3%	18.3%	18.3%
100 mg	78.2%	78.2%	78.2%
Dose considered as effective			
50 mg	3.3%	3.3%	3.3%
75 mg	18.0%	18.1%	18.1%
100 mg	78.1%	78.1%	78.1%
Simulation Θ=(−0.37,−0.52,−0.52)			
Power	96.4%	97.2%	97.2%
Average N	140	141.2	141.2
Selected dose			
50 mg	11.3%	11.3%	11.3%
75 mg	44.9%	44.9%	44.9%
100 mg	43.8%	43.8%	43.8%
Dose considered as effective			
50 mg	10.1%	10.3%	10.3%
75 mg	43.7%	44.0%	44.0%
100 mg	42.6%	42.9%	42.9%
Simulation Θ=(−0.37,−0.37,−0.37)			
Power	80.0%	82.5%	82.8%
Average N	140	143.6	144.1

Selected dose			
50 mg	33.4%	33.4%	33.4%
75 mg	33.2%	33.2%	33.2%
100 mg	33.4%	33.4%	33.4%
Dose considered as effective			
50 mg	26.7%	27.4%	27.5%
75 mg	26.7%	27.5%	27.7%
100 mg	26.6%	27.5%	27.6%

175

176

177 **Table 2: Performance of the proposed design (additional scenarios).**

	Without reassessment of n_2	With reassessment of n_2	
$(\varphi_f;\varphi_u)$	(−;−)	(0.7;0.9)	(0.6;0.9)
Simulation Θ=(0,−0.37,−0.52)			
Power	92.2%	93.7%	94.1%
Average N	140	142.5	143.1
Selected dose			
50 mg	0.2%	0.2%	0.2%
75 mg	23.3%	23.3%	23.3%
100 mg	76.5%	76.5%	76.5%
Dose considered as effective			
50 mg	0.1%	0.1%	0.1%
75 mg	19.2%	19.7%	19.7%
100 mg	73.0%	74.0%	74.4%
Simulation Θ=(0,−0.37,−0.37)			
Power	74.0%	77.4%	78.3%
Average N	140	144.9	146.5
Selected dose			
50 mg	0.8%	0.8%	0.8%
75 mg	49.2%	49.2%	49.2%
100 mg	49.9%	49.9%	49.9%
Dose considered as effective			
50 mg	0.1%	0.1%	0.1%
75 mg	37.0%	38.4%	38.8%
100 mg	37.1%	39.0%	39.5%
Simulation Θ=(0,0,−0.52)			
Power	91.7%	93.8%	94.7%
Average N	140	143.6	145
Selected dose			
50 mg	0.5%	0.5%	0.5%
75 mg	0.5%	0.5%	0.5%
100 mg	99.0%	99.0%	99.0%
Dose considered as effective			
50 mg	0.1%	0.1%	0.1%
75 mg	0.1%	0.1%	0.1%
100 mg	91.7%	93.8%	94.7%

Simulation Θ=(0,0,−0.37)			
Power	61.9%	66.5%	69.3%
Average N	140	147.1	152.3
Selected dose			
50 mg	2.7%	2.7%	2.7%
75 mg	2.9%	2.9%	2.9%
100 mg	94.4%	94.4%	94.4%
Dose considered as effective			
50 mg	0.2%	0.2%	0.2%
75 mg	0.2%	0.2%	0.2%
100 mg	61.9%	66.5%	69.3%

References

1. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. Biometrics 1994; 50:1029-1041.
2. Bretz F, Koenig F, Brannath W, Glimm E, Posch M. Adaptive designs for confirmatory clinical trials. Statistics in Medicine 2009; 28:1181-1217. URL:http://dx.doi.org/10.1002/sim.3538.
3. Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. Biometrics 1999;55:1286-1290.
4. Hochberg Y, Tamhane A. Multiple comparison procedures . Wiley: New York, 1987.
5. Marcus R, Eric P, Gabriel K. On closed testing procedures with special reference to ordered analysis of variance. Biometrika 1976; 63:655-660.
6. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. Biometrika 1986; 73:751-754.
7. Desseaux K, Porcher R. Flexible two-stage design with sample size reassessment for survival trials. Statistics in Medicine 2007; 26:5002-5013. URL:http://dx.doi.org/10.1002/sim.2966.
8. Mehta C, Pocock S. Adaptive increase in sample size when interim results are promising: A practical guide with examples. Statistics in Medicine 2010; 30:3267-3284. URL:http://dx.doi.org/10.1002/sim.4102.
9. FDA. Guidance for industry: Adaptive design clinical trials for drugs and biologics 2010. URL:http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm201790.pdf.
10. Porcher R, Lecocq B, Vray M. Les méthodes adaptatives: quand et comment les utiliser dans les essais cliniques? Thérapie 2011; 66:309-317.