Similarly as in the power (1), the power (2) can be calculated by using multivariate normal integrals. For the details, please refer to Appendix A.1.

For simplicity, consider a two-stage group-sequential design with one interim and one final analysis. The probability of rejecting the null hypothesis at the interim analysis is the same for DF-1 and DF-2. The difference in power between DF-1 and DF-2 is due to whether or not the null hypothesis is rejected at the final analysis. The difference in decision making for DF-1 and DF-2 comes from the following two situations where the interim analysis result is inconsistent with the final analysis result even the alternative hypothesis is true, that is, (i) endpoint 1 is statistically significant at the interim, but not at the final analysis and similarly, and (ii) endpoint 2 is statistically significant at the interim, but not at the final analysis. Thus, DF-1 fails to reject the null hypothesis in both situations even if the alternative hypothesis is true, but DF-2 is able to reject the null hypothesis in both situations. However, the likelihood of this scenario occurring is quite low. Thus, there is little practical difference in the power and sample size determinations for DF-1 and DF-2. However, DF-2 offers the option of stopping measurement of an endpoint for which superiority has been demonstrated. Stopping measurement may be desirable if the endpoint is very invasive or expensive although stopping measurement may also introduce an operational difficulty into the trial. This will be illustrated in Section 3.

2.3. Maximum sample size and average sample number

We discuss two sample size concepts, that is, the maximum sample size (MSS) and the average sample number (ASN) based on DF-1 and DF-2, and the corresponding powers (1) and (2) discussed in the previous section.

The MSS is the sample size required for the final analysis to achieve the desired power $1-\beta$. The MSS is given by the smallest integer not less than n_L satisfying the power (1) or (2) for a group-sequential design at the prespecified δ_1, δ_2, ρ_{T} and ρ_{C}, with Fisher's information time for the interim analyses, n_l/n_L, $l = 1, \ldots, L$. To find a value of n_L, an iterative procedure is required to numerically solve for the power (1) or (2). This can be accomplished by using a grid search to gradually increase n_L until the power under n_L exceeds the desired power, although this often requires considerable computing resources. To reduce the computational resources, the Newton–Raphson algorithm in [14] or the basic linear interpolation algorithm in [15] may be utilized.

The ASN is the expected sample size under a specific hypothetical reference. Given these prespecifications, the ASN per intervention group for DF-1 is given by

$$\mathrm{ASN} = n_L\left(1 + \sum\nolimits_{l=1}^{L-1} \Pr\left[\left\{\bar{A}_{11} \cup \bar{A}_{21}\right\} \cap \cdots \cap \left\{\bar{A}_{1l} \cup \bar{A}_{2l}\right\}\right]\right) \Big/ L, \tag{3}$$

and for DF-2,

$$\mathrm{ASN} = n_L\left(1 + \sum\nolimits_{l=1}^{L-1} \Pr\left[\left\{\bar{A}_{11} \cap \cdots \cap \bar{A}_{1l}\right\} \cup \left\{\bar{A}_{21} \cap \cdots \cap \bar{A}_{2l}\right\}\right]\right) \Big/ L, \tag{4}$$

where $r_l = 1$ and $n_l = l n_1$, $l = 1, \ldots, L$. The representations for calculating ASN (3) and (4) are described in Appendix A.2.

The powers, MSS, and ASN will depend on the design parameters including differences between means, the correlation structure between the endpoints, the testing procedure (e.g., O'Brien–Fleming (OF) boundary [18], Pocock (PC) boundary [19]), the number of analyses, and the information time.

3. Evaluation of the sample size

3.1. Behavior of the sample size

In this section, we evaluate the behavior of the power, MSS, and ASN as the design parameters vary. Here, without loss of generality, $\sigma_1^2 = \sigma_2^2 = 1^2$ is chosen for simplicity, so that δ_1 and δ_2 are interpreted as (standardized) effect sizes.

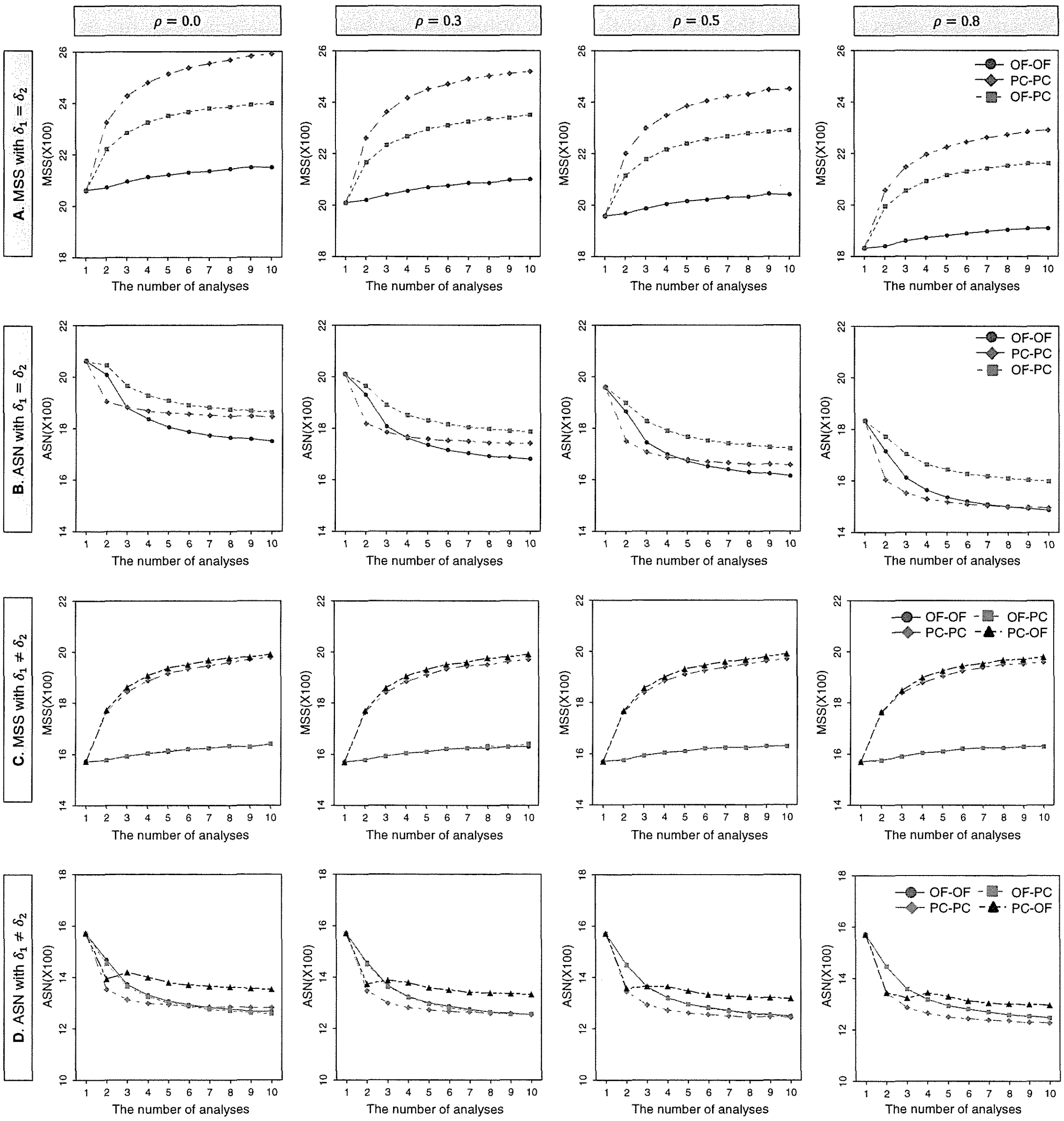

Figure 1. Behavior of MSS and ASN for DF-1 as the number of analyses and boundaries vary. The MSS and ASN per intervention group (equally-sized groups: $r_l = 1$) were calculated to detect the joint difference in the two endpoints with the overall power of 80% at the one-sided significance level of 2.5%, where $\delta_1 = \delta_2 = 0.1$ for A and B, and $\delta_1 = 0.1$ and $\delta_2 = 0.2$ for C and D; $\sigma_1^2 = \sigma_2^2 = 1^2$. When differences between means are equal, the critical values are determined by the three boundary combinations, that is, (i) the OF for both endpoints, (ii) the PC for both endpoints, and (iii) the OF for δ_1 and the PC for δ_2, with the LD alpha-spending method with equal information space. When differences between means are unequal, in addition to the three combinations, (iv) the PC for δ_1 and the OF for δ_2 is considered.

Figure 1 illustrates how the MSS and ASN per intervention group for DF-1 behave as a function of the number of analyses and the boundaries when effect sizes are equal and unequal, that is, $\delta_1 = \delta_2$ and $\delta_1 \neq \delta_2$ between the two endpoints. The MSS and ASN for DF-1 and DF-2 (equally-sized groups: $r_l = 1$) were calculated to detect the joint difference in the two endpoints with the overall power of 80% at the one-sided significance level of 2.5%, where $\delta_1 = \delta_2 = 0.1$ for equal effect sizes and $\delta_1 = 0.1$ and $\delta_2 = 0.2$ for unequal effect sizes; $\sigma_1^2 = \sigma_2^2 = 1^2$; and $\rho_T = \rho_C = \rho = 0.0, 0.3, 0.5$ and 0.8. The critical values are determined by the three boundary combinations, that is, (i) the OF for both endpoints (OF–OF), (ii) the PC for both endpoints (PC–PC), and (iii) the OF for δ_1 and the PC for δ_2 (OF–PC), with the LD alpha-spending method with equal information space.

When effect sizes are equal, the MSS for the three boundary combinations increases as the number of analyses increases and the correlation is smaller. In all of $\rho = 0, 0.3, 0.5$ and 0.8, the largest MSS is given by PC–PC and the smallest MSS by OF–OF. On the other hand, the ASN for the three boundary combinations decreases as the number of analyses increases and the correlation is larger. In all of $\rho = 0.0, 0.3, 0.5$ and 0.8, the largest ASN is given by OF–PC.

When effect sizes are unequal $\delta_1 < \delta_2$, in addition to the three boundary combinations, one more combination of (iv) the PC for δ_1 and the OF for δ_2 (PC–OF) is considered, $\delta_1 = 0.1$ and $\delta_2 = 0.2$. Similarly as seen with equal effect sizes, the MSS for the four boundary combinations increases as the number of analyses increases, but it does not change as with the correlation varies. The largest MSS is given by PC–PC and PC–OF and the smallest MSS by OF–OF and OF–PC. On the other hand, the ASN for the four boundary combinations decreases as the number of analyses increases independently of the correlation. The largest ASN is given by OF–OF and OF–PC and the smallest ASN by PC–PC and PC–OF. When one effect size is smaller (or larger) than the other, the MSS and ASN will be driven by the smaller effect size. In this illustration, as the OF is selected for the smaller effect size and the PC for the larger, the MSS and ASN by OF–PC are approximately equal to those by OF–OF.

Figure 2 illustrates how the MSS and ASN per intervention group for DF-2 behave as a function of the number of analyses and the boundaries when effect sizes are equal $\delta_1 = \delta_2$ and unequal $\delta_1 \neq \delta_2$ between the two endpoints with the same parameter settings as in Figure 1. The MSS and ASN behaviors are similar to those observed for DF-1. The major difference between DF-1 and DF-2 is that the MSS and ASN for DF-2 are smaller than those for DF-1. They are notably smaller as the number of analyses increases, especially when the correlation is low.

If the trial was designed to detect effects on *at least one* endpoint with a prespecified ordering of endpoints, a choice of different boundaries for each endpoint (i.e., the OF for the primary endpoint and the PC for the secondary endpoint) can provide a higher power than using the same boundary for both endpoints [20, 21]. However, as shown in Figures 1 and 2, the selection of a different boundary has a minimal effect on the power.

3.2. Example

We provide an example to illustrate the sample size methods discussed in the previous sections. Consider the clinical trial, 'Effect of Tarenflurbil on Cognitive Decline and Activities of Daily Living in Patients With Mild Alzheimer Disease', a multicenter, randomized, double-blind, placebo-controlled trial in patients with mild Alzheimer disease (AD) [22]. Co-primary endpoints were cognition as assessed by the Alzheimer Disease Assessment Scale Cognitive Subscale (ADAS-Cog; 80-point scale) and functional ability as assessed by the Alzheimer Disease Cooperative Study activities of daily living (ADCS-ADL; 78-point scale). A negative change score from baseline on the ADAS-Cog indicates improvement while a positive change score on the ADCS-ADL indicates improvement. The original sample size per intervention group of 800 patients provided an overall power of 96% to detect the joint difference in the two primary endpoints between the tarenflurbil and placebo groups, by using a one-sided test at 2.5% significance level, with the standardized effect size of 0.2 for both endpoints. In addition, the correlation between the two endpoints was assumed to be zero in the calculation of the sample size although the two endpoints were expected to be correlated (for example, see Doraiswamy *et al.* [23]).

Table I displays the MSS and ASN per intervention group (equally-sized groups: $r_l = 1$) for the DF-1 and DF-2. The sample size was with an alternative hypothesis of a difference for both ADAS-Cog ($\delta_1 = 0.2$) and ADCS-ADL ($\delta_2 = 0.2$), with the overall power of 96% at the one-sided significance level of 2.5%, where $\rho = \rho_T = \rho_C = 0.0, 0.3, 0.5$, and 0.8 and $L = 1, 2, 3, 5, 8$, and 10. The critical values are determined by the three boundary combinations, that is, the OF for both endpoints (OF–OF), the PC for both endpoints (PC–PC), and OF for ADAS-Cog and the PC for ADCS-ADL (OF–PC).

Based on the selected parameters described in [22], that is, $L = 1$ and $\rho = 0.0$, the sample size per intervention group is calculated as 804. If four interims and one final analysis are planned (i.e., $L = 5$) with DF-1, and conservatively assuming a zero correlation between the endpoints, then the MSS is 825 for OF–OF, 945 for PC–PC and 895 for OF–PC, and the ASN is 604 for OF–OF, 548 for PC–PC, and 608 for OF–PC. If the correlation is incorporated into the calculation when $\rho = 0.3, 0.5$, and 0.8, then the MSS are 820, 810, and 785 for OF–OF; 940, 930, and 900 for PC–PC; and 890, 885, and 860 for OF–PC. The ASN are 589, 574, and 543 for OF–OF; 525, 506, and 469 for PC–PC; and 593, 582, and 556 for OF–PC. When comparing DF-2 to DF-1, there are no major differences in MSS and ASN for all of the boundary combinations, although DF-2 provides a slightly smaller MSS and ASN than DF-1, for

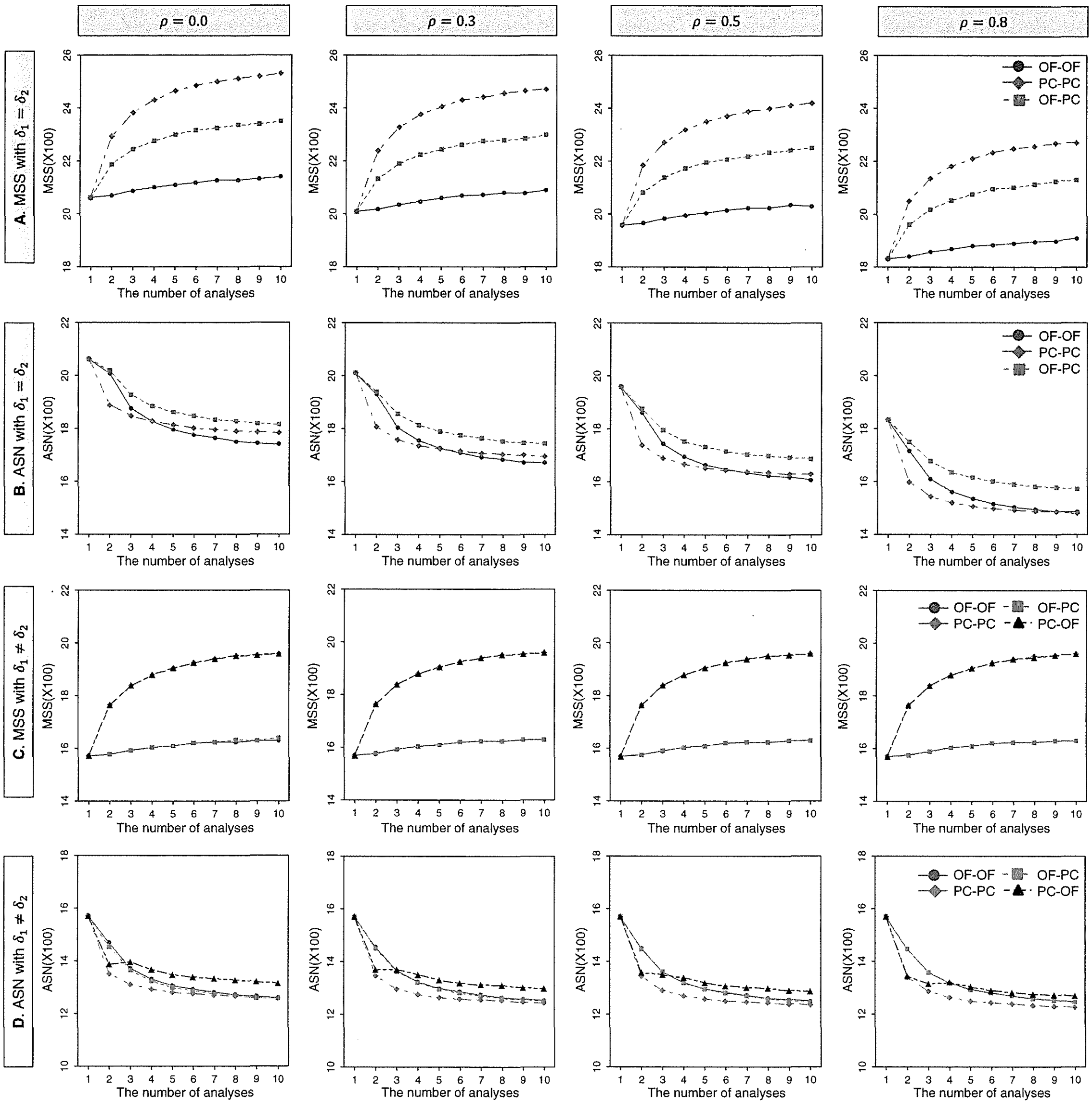

Figure 2. Behavior of MSS and ASN for DF-2 as the number of analyses and boundaries vary. The MSS and ASN per intervention group (equally-sized groups: $r_l = 1$) were calculated to detect the joint difference in the two endpoints with the overall power of 80% at the one-sided significance level of 2.5%, where $\delta_1 = \delta_2 = 0.1$ for A and B and $\delta_1 = 0.1$ and $\delta_2 = 0.2$ for C and D; $\sigma_1^2 = \sigma_2^2 = 1^2$. When differences between means are equal, the critical values are determined by the three boundary combinations, that is, (i) the OF for both endpoints, (ii) the PC for both endpoints, and (iii) the OF for δ_1 and the PC for δ_2, with the LD alpha-spending method with equal information space. When differences between means are unequal, in addition to the three combinations, (iv) the PC for δ_1 and the OF for δ_2 is considered.

PC–PC and OF–PC. However, if the endpoint is very invasive and thus stopping measurement may be ethically desirable, there is a benefit of using DF-2 as DF-2 offers the option of stopping measurement of an endpoint for which superiority has been demonstrated. For example, when four interims and one final analysis with DF-2 are planned (i.e., $L = 5$), the average total number of measurements for each intervention group are 1052, 1045, 1041, and 1021 for OF–OF; 846, 845, 841, and 831 for PC–PC; and 966, 961, 958, and 944 for OF–PC, corresponding to $\rho = 0.0, 0.3, 0.5$, and 0.8. They are smaller than those for DF-1 as the average total number of measurements for DF-1 are 1208, 1178, 1148, and 1086 for OF–OF; 1096, 1050, 1012, and 938 for PC–PC; and 1216, 1186, 1164, and 1112 for OF–PC.

Table I. MSS and ASN per intervention group (equally-sized groups) for detecting the joint difference for ADAS-Cog (0.2) and ADCS-ADL (0.2), with DF-1 and DF-2 and the overall power of 96% at the one-sided significance level of 2.5%.

			(i) OF–OF		(ii) PC–PC		(iii) OF–PC	
Decision-making framework	Correlation	Number of analyses	MSS	ASN	MSS	ASN	MSS	ASN
DF-1	0.0	1	804	804	804	804	804	804
		2	808	725	886	607	854	693
		3	816	647	918	572	876	652
		5	825	604	945	548	895	608
		8	832	579	968	535	912	587
		10	840	573	970	530	920	581
	0.3	1	799	799	799	799	799	799
		2	802	702	880	593	850	676
		3	810	633	912	552	870	638
		5	820	589	940	525	890	593
		8	824	563	960	511	904	571
		10	830	556	970	507	910	564
	0.5	1	791	791	791	791	791	791
		2	794	684	872	580	842	662
		3	801	620	903	536	864	627
		5	810	574	930	506	885	582
		8	816	549	952	492	896	558
		10	820	542	960	488	900	551
	0.8	1	764	764	764	764	764	764
		2	768	644	842	549	818	635
		3	774	588	873	501	840	603
		5	785	543	900	469	860	556
		8	792	520	920	453	872	533
		10	800	514	920	447	880	527
DF-2	0.0	1	804	804	804	804	804	804
		2	808	725	882	605	848	690
		3	813	645	912	569	867	646
		5	825	603	940	540	890	602
		8	832	578	960	524	904	579
		10	830	568	960	518	910	572
	0.3	1	799	799	799	799	799	799
		2	802	702	876	591	842	672
		3	807	632	906	549	861	632
		5	815	586	935	520	880	586
		8	824	562	952	503	896	564
		10	830	555	960	498	900	556
	0.5	1	791	791	791	791	791	791
		2	794	684	868	579	834	658
		3	801	620	897	533	855	621
		5	810	574	925	502	875	575
		8	816	549	944	486	888	552
		10	820	541	950	481	890	544
	0.8	1	764	764	764	764	764	764
		2	768	644	840	549	810	631
		3	774	588	870	499	831	597
		5	785	543	895	467	850	550
		8	792	520	912	450	864	528
		10	790	510	920	445	870	521

4. Sample size recalculation

Clinical trials are designed based on assumptions often constructed based on prior data. However, prior data may be limited or an inaccurate indication of future data, resulting in trials that are over/underpowered. Interim analyses provide an opportunity to evaluate the accuracy of the design assumptions and potentially make design adjustments (i.e., to the sample size) if the assumptions were markedly inaccurate. The tarenflurbil trial mentioned in the previous section, failed to demonstrate a beneficial effect of tarenflurbil on both ADAS-Cog and ADCS-ADL. The observed treatment effects were smaller than the assumed effects. Group-sequential designs allow for early stopping when there is sufficient statistical evidence that the two treatments are different. However, more modern adaptive designs may also allow for increases in the sample size if effects are smaller than assumed. Such adjustments must be conducted carefully for several reasons. Challenges include the following: (i) maintaining control of statistical error rates, (ii) developing a plan to make sure that treatment effects cannot be inferred via back-calculation of a resulting change in the sample size, (iii) consideration of the clinical relevance of the treatment effects, and (iv) practical concerns such as an increase in cost and the challenge of accruing more trial participants. In this section, we discuss sample size recalculation based on the observed intervention's effects at an interim analysis with a focus on control of statistical error rates.

4.1. Test statistics and conditional power

Consider that the maximum sample size is recalculated to n'_L based on the interim data at the Rth analysis. Suppose that n'_L is subject to $n_R < n'_L \leqslant \lambda n_L$, where λ is a prespecified constant for the maximum allowable sample size. For simplicity, assume a common correlation between the treatment groups, that is, $\rho_\mathrm{T} = \rho_\mathrm{C} = \rho$. Let $(\tilde{\delta}_1, \tilde{\delta}_2)$ and let (δ_1^*, δ_2^*) be the mean differences used for planned sample size and for recalculated sample size, respectively.

Here, we consider the Cui–Hung–Wang (CHW) statistics [24] for sample size recalculation in group-sequential designs with two co-primary endpoints to preserve the overall Type I error rate at a prespecified alpha level even when the sample size is increased and conventional test statistics are used. The CHW statistics are

$$Z'_{km} = \sqrt{\frac{n_R}{n_m}} Z_{kR} + \sqrt{\frac{n_m - n_R}{n_m}} \frac{\sum_{i=n_R+1}^{n'_m} Y_{\mathrm{T}ki} - \sum_{j=n_R+1}^{n'_m} Y_{\mathrm{C}kj}}{\sqrt{2\left(n'_m - n_R\right)}},$$

where $n'_m = (n_m - n_R)(n'_L - n_R) \big/ (n_L - n_R) + n_R$ and $r_R = r_m = 1 (k = 1, 2; \quad R = 1, \ldots, L-1; m = R+1, \ldots, L)$. The same critical values utilized for the case without sample size recalculation are used.

The sample size is increased or decreased when the conditional power evaluated at the Rth analysis is lower or higher than the desired power $1 - \beta$. Under the planned maximum sample size and a given observed value of (Z_{1R}, Z_{2R}), for DF-1, the conditional power is defined by

$$CP = \Pr\left[\bigcup_{m=R+1}^{L} \{A_{1m} \cap A_{2m}\} \,|a_{1R}, a_{2R}\right] \tag{5}$$

if $Z_{1l} \leqslant c_{1l}$ or $Z_{2l} \leqslant c_{2l}$ for all $l = 1, \ldots, R$, where (a_{1R}, a_{2R}) is a given observed value of (Z_{1R}, Z_{2R}). On the other hand, the conditional power for DF-2 is given by

$$CP = \begin{cases} \Pr\left[\bigcup_{m=R+1}^{L} A_{1m} \,|a_{1R}, a_{2l'}\right] \\ \quad \text{if } Z_{1l} \leqslant c_{1l} \text{ for all } l = 1, \ldots, R \text{ and } Z_{2l'} > c_{2l'} \text{ for some } l' = 1, \ldots, R, \\ \Pr\left[\bigcup_{m=R+1}^{L} A_{2m} \,|a_{2R}, a_{1l'}\right] \\ \quad \text{if } Z_{2l} \leqslant c_{2l} \text{ for all } l = 1, \ldots, R \text{ and } Z_{1l'} > c_{1l'} \text{ for some } l' = 1, \ldots, R, \\ \Pr\left[\left\{\bigcup_{m=R+1}^{L} A_{1m}\right\} \cap \left\{\bigcup_{m=R+1}^{L} A_{2m}\right\} \,|a_{1R}, a_{2R}\right] \\ \quad \text{if } Z_{1l} \leqslant c_{1l} \text{ and } Z_{2l} \leqslant c_{2l} \text{ for all } l = 1, \ldots, R. \end{cases} \tag{6}$$

The detailed calculation of the conditional powers for DF-1 and DF-2 are provided in Appendix A.3. Because (δ_1, δ_2) is unknown, it is customary to substitute $(\hat{\delta}_1^*, \hat{\delta}_2^*)$, the estimated mean differences at the Rth analysis $(\hat{\delta}_{1R}, \hat{\delta}_{2R})$ or the assumed mean differences during trial planning $(\tilde{\delta}_1, \tilde{\delta}_2)$. We consider the conditional power based on $(\delta_1^*, \delta_2^*) = (\hat{\delta}_{1R}, \hat{\delta}_{2R})$, which allows evaluation of behavior of power independent of $(\tilde{\delta}_1, \tilde{\delta}_2)$.

When recalculating the sample size, three options are possible: (i) only allowing an increase in the sample size, (ii) only allowing a decrease in the sample size, and (iii) allowing an increase or decrease in sample size. For all the cases, we assign Z'_{km} and n'_m instead of Z_{km} and n_m in the conditional powers (5) and (6) for the conditional power with sample size recalculation. Consider the rule for determining the recalculated sample size n'_L, when the sample size may be increased only, which is

$$n'_L = \begin{cases} n_L, & \text{if } CP \geqslant 1-\beta \text{ or } \min(\hat{\delta}_{1R}, \hat{\delta}_{2R}) \leqslant 0, \\ \min\left(n''_L, \lambda n_L\right), & \text{otherwise,} \end{cases}$$

where n''_L is the smallest integer $n'_L (> n_R)$, where the conditional power achieves the desired power $1-\beta$. When the sample size may be decreased only, the recalculated sample size n'_L is

$$n'_L = \begin{cases} n''_L, & \text{if } CP > 1-\beta, \\ n_L, & \text{otherwise.} \end{cases}$$

When the sample size may be increased or decreased, the recalculated sample size n'_L is

$$n'_L = \begin{cases} n''_L, & \text{if } CP > 1-\beta, \\ n_L, & \text{if } CP = 1-\beta \text{ or } \min\left(\hat{\delta}_{1R}, \hat{\delta}_{2R}\right) \leqslant 0, \\ \min\left(n''_L, \lambda n_L\right), & \text{otherwise.} \end{cases}$$

4.2. Simulation study

A simulation study was performed to evaluate the impact of sample size recalculation based on DF-1 and DF-2 on the power and Type I error rate. We consider group-sequential designs with a single interim, that is, one interim and one final analyses, and with multiple interims, that is, three interims and one final analyses. In addition, we discuss the three options of (i) only decreasing the sample size, (ii) only increasing the sample size, and (iii) increasing or decreasing the sample size, based upon the observed intervention's effect. The planned MSS per intervention group is calculated to detect the joint difference for two endpoints with the overall power of 80% at the one-sided significance level of 2.5%, where $(\tilde{\delta}_1, \tilde{\delta}_2) = (0.2, 0.2)$, $\sigma_1^2 = \sigma_2^2 = 1^2$ and the correlation is assumed to be known correlation at the design stage, that is, $\rho = 0.0, 0.3, 0.5$, and 0.8. For the evaluation of the Type I error rate, the two pairs of the mean differences $(\delta_1, \delta_2) = (0.0, 0.0)$ and (0.0, 0.2) are considered under H_0. For the designs with a single interim, the timing of the interim analysis for sample size recalculation is evaluated at 0.25, 0.50, and 0.75 of information time. For designs with multiple interims, one sample size recalculation is considered, and the timing is evaluated at the first, second, and third of interim analysis. The critical values are determined by the OF boundary for both endpoints with the LD alpha-spending method, with equal information space. The upper limit of the recalculated sample size is set to $n'_2 = \lambda n_2$ with $\lambda = 1.5$. The number of replications for the simulation is set to 1,000,000 for the evaluation of the Type I error rate and 100,000 replications for the power. These number of replications for the simulation was determined based on the precision, where a sample size of 1,000,000 provides a two-sided 95% confidence interval with a width equal to 0.001 when the proportion is 0.025, and a total number of replications of 100,000 provides a two-sided 95% confidence interval with a width equal to 0.005 when the proportion is 0.80.

Suppose that the sample size recalculation is based on the interim estimates of (δ_1, δ_2). Note that the value of correlation assumed at the design stage is retained for the sample size recalculation, that is, without updating based on observed correlation at the interim as the correlation is a nuisance parameter in hypothesis testing. All results are summarized in Tables S1–S4 in the Supporting information. As there are no significant differences between DF-1 and DF-2 with respect to the Type I error rates and empirical powers, we limit the discussion to the behavior of the Type I error rates and power for DF-1.

Figure 3 illustrates how the Type I error rates and powers behave as a function of the correlation, the timing of the interim analysis for sample size recalculation, and the sample size recalculation options for DF-1 in the single-interim case. In all three recalculation options, the Type I error rates increase as the correlation increases, but they do not exceed the targeted 2.5%. There is no practical difference in

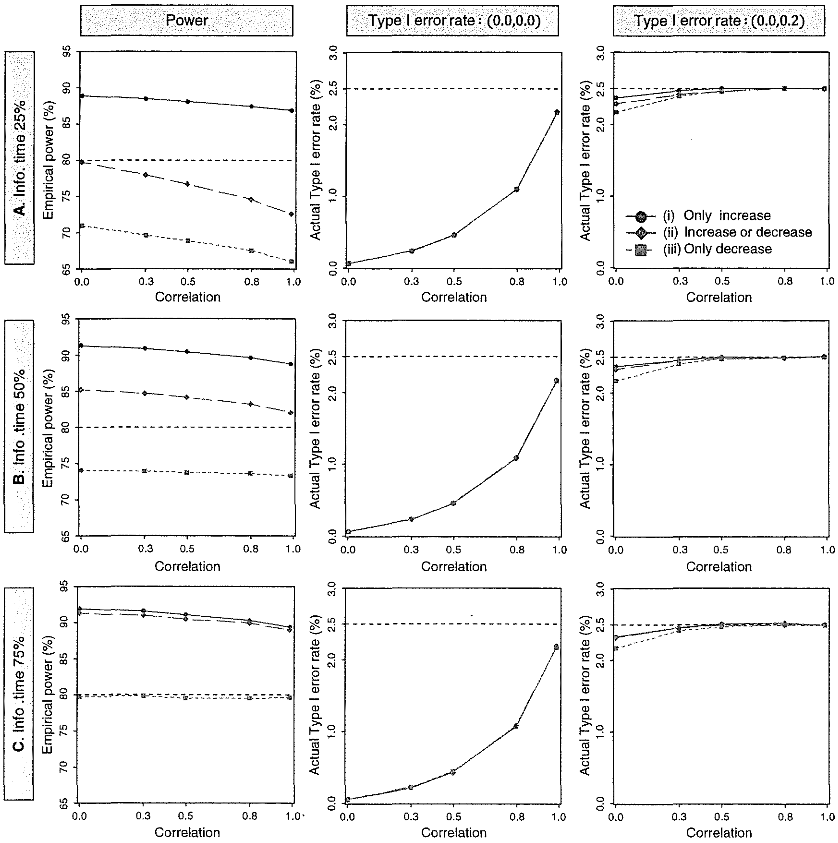

Figure 3. Behavior of the power and Type I error rate as a function of the correlation with sample size recalculation in two-stage group-sequential designs, where the information times of 0.25, 0.50, and 0.75 were selected as the timing of the sample size recalculation. The planned MSS per intervention group is calculated to detect the joint difference for two endpoints with the overall power of 80% at the one-sided significance level of 2.5%, where one interim and one final analysis are to be performed. The critical values are determined by the OF boundary for both endpoints, with the LD alpha-spending method. The upper limit of recalculation sample size is $n_2' = \lambda n_2$ with $\lambda = 1.5$. The number of replications for simulation is set to 1,000,000 for evaluation of the Type I error rate and 100,000 replications for the power (DF-1).

the behavior of the Type I error rates depending on the timing of the interim analysis for sample size recalculation. On the other hand, for the behavior of the power, when only allowing an increase in the sample size, the empirical powers are higher than the desired power of 80% in all of the three timings of sample size recalculation, although the power is slightly decreased with higher correlation. When allowing an increase or a decrease in the sample size, if the timing for sample size recalculation is at 25% information time, then the empirical power is lower than the desired power of 80%, especially with higher correlation. However, if the timing for sample size recalculation is 50% or 75%, then the empirical powers are higher than in all three timings of sample size recalculation. When only allowing a decrease in the sample size, if the timing for the sample size recalculation is at 25% or 50% information time, then the empirical powers are always lower than the desired power, especially with higher correlation. If the timing for sample size recalculation is at 75% information time, then the empirical power is almost achieved at the desired power of 80%.

Figure 4 illustrates how the Type I error rates and powers behave as a function of the correlation, the timing of the interim analysis for sample size recalculation, and the sample size recalculation options for DF-1, in the multiple-interim case. The results are similar to those in the single-interim case; when only allowing an increase in the sample size, compared with the desired power of 80%, the empirical powers are improved in all of the three timings for the sample size recalculation, but the empirical power is much lower than the desired power if the sample size recalculation is conducted early in the study, especially when allowing a decrease in the sample size.

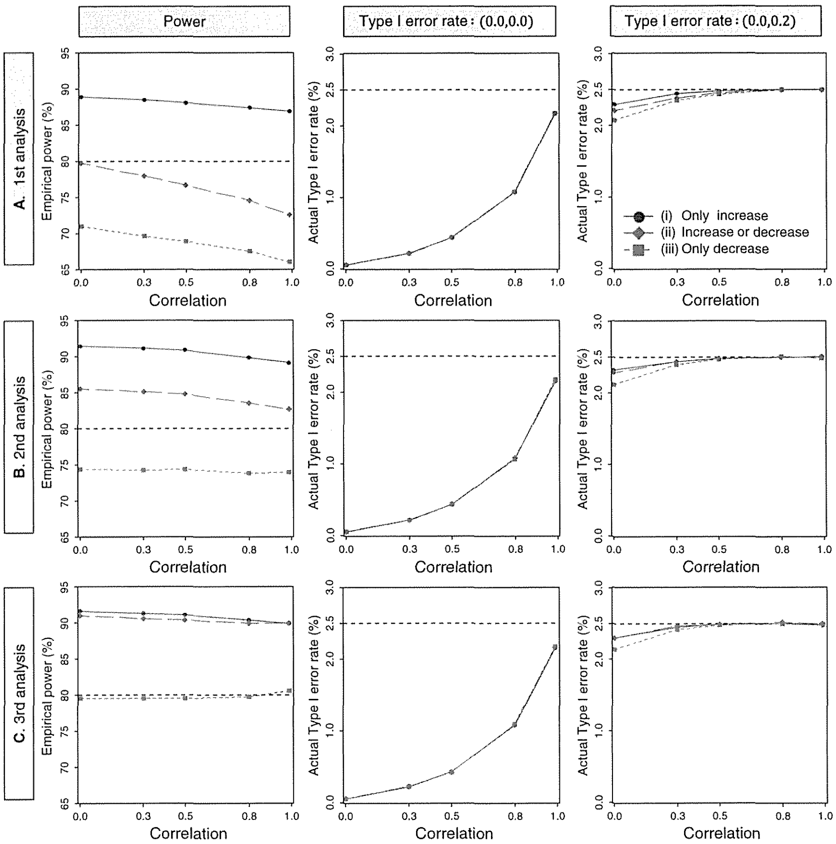

Figure 4. Behavior of the power and Type I error rate as a function of the correlation with sample size recalculation in four-stage group-sequential designs, where the first, second, and third interim points were selected as the timing of the sample size recalculation. The planned MSS per intervention group is calculated to detect the joint difference for two endpoints with the overall power of 80% at the one-sided significance level of 2.5%, where three interims and one final analysis are to be performed. The critical values are determined by the OF boundary for both endpoints, with the LD alpha-spending method. The upper limit of recalculation sample size is $n_2' = \lambda n_2$ with $\lambda = 1.5$. The number of replications for simulation is set to 1,000,000 for evaluation of the Type I error rate and 100,000 replications for the power (DF-1).

These results suggest incorporating the uncertainty of the estimates at the interim into the sample size recalculation is important. The power is much lower than desired power if the sample size recalculation is conducted early in the study, especially when allowing for a decrease in the sample size.

5. Summary and discussion

The determination of sample size and the evaluation of power are fundamental and critical elements in the design of a clinical trial. If a sample size is too small, then important effects may not be detected, while a sample size that is too large is wasteful of resources and unethically puts more participants at risk than necessary. Recently, many clinical trials are designed with more than one endpoint considered as co-primary. As with trials involving a single primary endpoint, designing such trials to include interim analyses (i.e., with repeated testing) may provide efficiencies by detecting trends prior to planned completion of the trial. It may also be prudent to evaluate design assumptions at the interim and potentially make design adjustments (i.e., sample size recalculation) if design assumptions were dramatically inaccurate. However, such design complexities create challenges in the evaluation of power and the calculation of sample size during trial design.

We discuss group-sequential designs with co-primary endpoints. We derive the power and sample size methods under two decision-making frameworks: (i) designing the trial to detect the test intervention's superiority for the two endpoints simultaneously (i.e., at the same interim timepoint of the trial) (DF-1) and (ii) designing the trial to detect superiority for the two endpoints at any interim timepoint (i.e., not necessarily simultaneously) (DF-2). The former is simpler while the latter is more flexible and may be useful when the endpoint is very invasive or expensive, as it allows for stopping the measurement of any endpoint upon which superiority has been demonstrated. We evaluate the behavior of sample size with varying design elements and provide an example to illustrate the methods. We also discuss sample size recalculation using CHW statistics and evaluate the impact on the power and Type I error rate. Although DF-2 will provide a slightly smaller sample size than DF-1, there is modest difference between the two. However, if the endpoint is very invasive and thus stopping measurement may be ethically desirable, there is a benefit of using DF-2 as DF-2 offers the option of stopping measurement of an endpoint for which superiority has been demonstrated. However, stopping measurement on one endpoint could also create operational challenges in study conduct and patient monitoring. The timing of the sample size recalculation should also be carefully considered as the power does not reach desired levels if the sample size recalculation is carried out early in the study when considering a decrease in the sample size.

There are other practical issues and extensions to consider when designing a group-sequential clinical trial with co-primary endpoints. They include the following: how the value of correlation should be selected at the planning and interim, evaluating futility or efficacy and futility simultaneously, other endpoint scales, and other inferential goals. We discuss each of these issues.

There are two important questions regarding the choice of the correlation in sample size calculations. One is whether the observed correlation from external or pilot data should be utilized or whether correlation is assumed to be zero. The other is whether the sample size should be recalculated based on the observed correlation at the interim. Incorporating the observed correlation at the planning or interim may affect the Type I error rate and power. Our experience suggests that when standardized effect sizes are unequal between the endpoints, the power is not improved with higher correlation. With unequal standardized effect sizes, incorporating the correlation into the sample size calculation at planning or interim may have no advantage [25, 26]. Further investigation will be required to assess how the choice of the correlation impacts the operation characteristics of the design.

Because the main objective of the paper is to provide the fundamental foundation in group-sequential designs for co-primary endpoints, our discussion is restricted to a superiority clinical trial comparing two interventions based on two continuous endpoints. The study design allows for early stopping when larger intervention differences are observed, that is, rejecting a null hypothesis only. However, this work provides a foundation for designing clinical trials with other design features. In addition to this fundamental situation, the method discussed here can be straightforwardly extended to other situations such as evaluating futility (rejecting the alternative hypothesis) or evaluating both efficacy and futility.

Time-to-event outcomes are common in oncology, cardiovascular, and infectious disease clinical trials. The method for continuous endpoints described in the paper may not be directly extended to time-to-event endpoints. When considering a trial with two time-to-event outcomes as co-primary with a plan for using the logrank test to compare two interventions in a group-sequential design, information for the two endpoints may accumulate at different rates. This creates challenges when designing trials, that is, the amount of information for the endpoints may be different at any particular interim timepoint of the trial. Further investigation is required to assess this issue.

Although our primary interest is *co-primary* endpoints, these results provide a fundamental foundation to other inferential goals, for example, designing a trial to detect an effect on *at least one* endpoint. Many authors have proposed methods for the *at least one* endpoint goal in fixed sample size designs, for example, a weighted Bonferroni procedure, the prospective alpha allocation scheme method, the adaptive alpha allocation approach, the Bonferroni-type parametric procedure, and the fallback-type parametric procedure (e.g., see [4, 27, 28]). In addition, several authors have discussed an extension of methods to the group-sequential designs with an inferential goal of *at least one* endpoint [29–32]. For example, Tang and Geller [30] discuss a method based on closed testing procedures, and Tamhane *et al.* [31, 32] discuss sample size methods in two-stage group-sequential designs based on the gatekeeping procedures with hierarchically ordered multiple endpoints.

Appendix

A.1. Power calculation

The power (1) for DF-1 can be calculated by partitioning the set in (1) into mutually exclusive subsets and taking the sum of their probabilities as follows:

$$\begin{aligned}1-\beta &= \Pr\left[\bigcup\nolimits_{l=1}^{L}\{A_{1l}\cap A_{2l}\}\,|\mathrm{H}_1\right]\\ &= \Pr\left[A_{11}\cap A_{21}\,|\mathrm{H}_1\right]+\sum_{l=2}^{L}\Pr\left[\bigcap\nolimits_{l'=1}^{l-1}\{\bar{A}_{1l'}\cup\bar{A}_{2l'}\}\cap\{A_{1l}\cap A_{2l}\}\middle|\,\mathrm{H}_1\right],\end{aligned}\tag{A1}$$

where $A_{kl} = \{Z_{kl} > c_{kl}\}$ and $\bar{A}_{kl} = \{Z_{kl} \leqslant c_{kl}\}(k = 1,2;\ l = 1,\ldots,L)$. The probability of $\{\bar{A}_{1l'} \cup \bar{A}_{2l'}\}$ can be written as $\Pr[\bar{A}_{1l'} \cup \bar{A}_{2l'}] = \Pr[\tilde{A}^1_{l'}] + \Pr[\tilde{A}^2_{l'}] + \Pr[\tilde{A}^3_{l'}]$, where $\tilde{A}^1_{l'} = \{\bar{A}_{1l'} \cap A_{2l'}\}$, $\tilde{A}^2_{l'} = \{A_{1l'} \cap \bar{A}_{2l'}\}$ and $\tilde{A}^3_{l'} = \{\bar{A}_{1l'} \cap \bar{A}_{2l'}\}(l' = 1,\ldots,L-1)$. Similarly, the probability of the union of $\{\bar{A}_{1l'} \cup \bar{A}_{2l'}\}$ can be written by the sum of the probabilities of the unions composed of $\tilde{A}^1_{l'}$, $\tilde{A}^2_{l'}$ and $\tilde{A}^3_{l'}$. Then, the second term of the right-hand side in (A.1) can be rewritten as

$$\begin{aligned}&\sum_{l=2}^{L}\Pr\left[\bigcap\nolimits_{l'=1}^{l-1}\{\bar{A}_{1l'}\cup\bar{A}_{2l'}\}\cap\{A_{1l}\cap A_{2l}\}\middle|\,\mathrm{H}_1\right]\\ &\quad=\sum\nolimits_{l=2}^{L}\left(\sum\nolimits_{h_1=1}^{3}\cdots\sum\nolimits_{h_{l-1}=1}^{3}\Pr\left[\left\{\bigcap\nolimits_{l'=1}^{l-1}\tilde{A}^{h_{l'}}_{l'}\right\}\cap\{A_{1l}\cap A_{2l}\}\middle|\,\mathrm{H}_1\right]\right).\end{aligned}$$

The probability of $\tilde{A}^1_{l'}$ is calculated by a bivariate normal integral as follows:

$$\Pr\left[\tilde{A}^1_{l'}\right]=\int_{-\infty}^{c_{1l'}}\int_{c_{2l'}}^{\infty}f_2\left(z_{1l'},z_{2l'}\right)dz_{2l'}dz_{1l'},$$

where $f_2(z_{1l'}, z_{2l'})$ is the density function of the joint distribution of $(Z_{1l'}, Z_{2l'})$ with the means and the covariance matrix given in Section 2.1. The probabilities of $\tilde{A}^2_{l'}$, $\tilde{A}^3_{l'}$ and $\{A_{1l'} \cap A_{2l'}\}$ are calculated similarly. Then, the probability of the union composed of $\tilde{A}^1_{l'}$, $\tilde{A}^2_{l'}$, $\tilde{A}^3_{l'}$ and $\{A_{1l'} \cap A_{2l'}\}$ is calculated by a multivariate normal integral and the power is the sum of $(3^L - 1)/2$ multivariate normal integrals. For details of the computation related to multivariate normal, please see [33].

For illustration, we provide the case of $L = 2$ and $r = r_1 = r_2$. In this case, the power can be rewritten as

$$\begin{aligned}1-\beta&=\Pr\left[A_{11}\cap A_{21}\,|\mathrm{H}_1\right]+\sum\nolimits_{h_1=1}^{3}\Pr\left[\tilde{A}^{h_1}_1\cap\{A_{12}\cap A_{22}\}\,|\mathrm{H}_1\right]\\ &=\int_{c_{11}}^{\infty}\int_{c_{21}}^{\infty}f_2(z_{11},z_{21})dz_{21}dz_{11}+\int_{-\infty}^{c_{11}}\int_{c_{21}}^{\infty}\int_{c_{12}}^{\infty}\int_{c_{22}}^{\infty}f_4\left(z_{11},z_{21},z_{12},z_{22}\right)dz_{22}dz_{12}dz_{21}dz_{11}\\ &\quad+\int_{c_{11}}^{\infty}\int_{-\infty}^{c_{21}}\int_{c_{12}}^{\infty}\int_{c_{22}}^{\infty}f_4(z_{11},z_{21},z_{12},z_{22})dz_{22}dz_{12}dz_{21}dz_{11}\\ &\quad+\int_{-\infty}^{c_{11}}\int_{-\infty}^{c_{21}}\int_{c_{12}}^{\infty}\int_{c_{22}}^{\infty}f_4\left(z_{11},z_{21},z_{12},z_{22}\right)dz_{22}dz_{12}dz_{21}dz_{11},\end{aligned}$$

where $f_2(z_{11}, z_{21})$ is the density function of the bivariate normal distribution of $\boldsymbol{Z}_2 = (Z_{11}, Z_{21})^{\mathrm{T}}$, which is given by

$$f_2(\boldsymbol{Z}_2)=\frac{1}{2\pi\left|\boldsymbol{\Sigma}_2\right|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{Z}_2-\boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{Z}_2-\boldsymbol{\mu}_2)\right],\ -\infty<z_{11},z_{21}<\infty$$

with mean vector $\boldsymbol{\mu}_2 = \sqrt{rn_1/(1+r)}(\delta_1/\sigma_1, \delta_2/\sigma_2)^{\mathrm{T}}$ and correlation matrix

$$\boldsymbol{\Sigma}_2=\begin{pmatrix}1^2 & (r\rho_{\mathrm{T}}+\rho_{\mathrm{C}})/(1+r)\\ (r\rho_{\mathrm{T}}+\rho_{\mathrm{C}})/(1+r) & 1^2\end{pmatrix}.$$

and $f_4(z_{11}, z_{21}, z_{12}, z_{22})$ is the density function of the tetra-variate normal distribution of $\boldsymbol{Z}_4 = (Z_{11}, Z_{21}, Z_{12}, Z_{22})^{\mathrm{T}}$ given by

$$f_4(\boldsymbol{Z}_4)=\frac{1}{(2\pi)^2\left|\boldsymbol{\Sigma}_4\right|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{Z}_4-\boldsymbol{\mu}_4)^{\mathrm{T}}\boldsymbol{\Sigma}_4^{-1}(\boldsymbol{Z}_4-\boldsymbol{\mu}_4)\right],\ -\infty<z_{11},z_{21},z_{12},z_{22}<\infty$$

with mean vector $\boldsymbol{\mu}_4 = \sqrt{(1+r)/r}\left(\sqrt{n_1}\delta_1/\sigma_1, \sqrt{n_1}\delta_2/\sigma_2, \sqrt{n_2}\delta_1/\sigma_1, \sqrt{n_2}\delta_2/\sigma_2\right)^{\mathrm{T}}$ and correlation matrix

$$\boldsymbol{\Sigma}_4 = \begin{pmatrix} \boldsymbol{\Sigma}_2 & \sqrt{n_1/n_2}\boldsymbol{\Sigma}_2 \\ \sqrt{n_1/n_2}\boldsymbol{\Sigma}_2 & \boldsymbol{\Sigma}_2 \end{pmatrix},$$

where $\boldsymbol{\Sigma}_4$ is positive definite matrix under $|\rho_{\mathrm{T}}|, |\rho_{\mathrm{C}}| < 1$ and $n_1 \neq n_2$ as $|\boldsymbol{\Sigma}_4| = |\boldsymbol{\Sigma}_2|^2(1-n_1/n_2)^2$.

The power (2) for DF-2 can be calculated from two L-variate normal integrals and a $2L$-variate normal integral.

$$\begin{aligned} 1-\beta &= \Pr\left[\left\{\bigcup_{l=1}^{L} A_{1l}\right\} \cap \left\{\bigcup_{l=1}^{L} A_{2l}\right\} \middle| \mathrm{H}_1\right] \\ &= 1-\left(\Pr\left[\bigcap_{l=1}^{L} \bar{A}_{1l} \middle| \mathrm{H}_1\right] + \Pr\left[\bigcap_{l=1}^{L} \bar{A}_{2l} \middle| \mathrm{H}_1\right] - \Pr\left[\bigcap_{l=1}^{L} \{\bar{A}_{1l} \cap \bar{A}_{2l}\} \middle| \mathrm{H}_1\right]\right). \end{aligned}$$

The power can be calculated similarly as discussed in the power (1) for DF-1.

A.2. ASN calculation

The ASN (3) for DF-1 can be calculated by the sum of multivariate normal integrals

$$\begin{aligned} \mathrm{ASN} &= n_L\left(1+\sum_{l=1}^{L-1} \Pr\left[\{\bar{A}_{11} \cup \bar{A}_{21}\} \cap \cdots \cap \{\bar{A}_{1l} \cup \bar{A}_{2l}\}\right]\right) \Big/ L \\ &= n_L\left\{1+\sum_{l=1}^{L-1}\left(\sum_{h_1=1}^{3} \cdots \sum_{h_l=1}^{3} \Pr\left[\bigcap_{l'=1}^{l} \tilde{A}_{l'}^{h_{l'}}\right]\right)\right\} \Big/ L. \end{aligned}$$

Similarly, the ASN (4) for DF-2 can be calculated by

$$\begin{aligned} \mathrm{ASN} &= n_L\left(1+\sum_{l=1}^{L-1} \Pr\left[\{\bar{A}_{11} \cap \cdots \cap \bar{A}_{1l}\} \cup \{\bar{A}_{21} \cap \cdots \cap \bar{A}_{2l}\}\right]\right) \Big/ L \\ &= n_L\left\{1+\sum_{l=1}^{L-1}\left(\Pr\left[\bigcap_{l'=1}^{l} \bar{A}_{1l'}\right] + \Pr\left[\bigcap_{l'=1}^{l} \bar{A}_{2l'}\right] - \Pr\left[\bigcap_{l'=1}^{l} \{\bar{A}_{1l'} \cap \bar{A}_{2l'}\}\right]\right)\right\} \Big/ L. \end{aligned}$$

A.3. Conditional power

The conditional power (5) for DF-1 is described by

$$\begin{aligned} CP &= \Pr\left[\bigcup_{m=R+1}^{L} \{A_{1m} \cap A_{2m}\} \,|a_{1R}, a_{2R}\right] \\ &= \Pr\left[A_{1,R+1} \cap A_{2,R+1} \,|a_{1R}, a_{2R}\right] \qquad \text{(A2)} \\ &\quad + \sum_{m=R+2}^{L} \Pr\left[\bigcap_{m'=R+1}^{m-1} \{\bar{A}_{1m'} \cup \bar{A}_{2m'}\} \cap \{A_{1m} \cap A_{2m}\} \,|a_{1R}, a_{2R}\right], \end{aligned}$$

if $Z_{1l} \leqslant c_{1l}$ or $Z_{2l} \leqslant c_{2l}$ for all $l = 1, \ldots, R$, where $A_{km} = \{Z_{km} > c_{km}\}$, $\bar{A}_{km} = \{Z_{km} \leqslant c_{km}\}$ $(k = 1, 2;\ m = R+1, \ldots, L)$ and (a_{1R}, a_{2R}) is a given observed value of (Z_{1R}, Z_{2R}). The second term of the right-hand side in (A2) can be calculated in a similar way to that for the power calculation (Appendix A.1). The conditional distribution of $(Z_{1,R+1}, Z_{2,R+1}, \ldots, Z_{1L}, Z_{2L}|a_{1R}, a_{2R}.)$ is a multivariate normal with their means $\mathrm{E}[Z_{km}|a_{1R}, a_{2R}] = \sqrt{n_m/2}\delta_k + \sqrt{n_R/n_m}(a_{kR} - \sqrt{n_R/2}\delta_k)$ and covariance given by $\mathrm{cov}[Z_{km}, Z_{k'm'}\,|a_{1R}, a_{2R}] = (n_{m'} - n_R)/\sqrt{n_m n_{m'}}$ if $k = k'$;

$(n_{m'} - n_R)\rho / \sqrt{n_m n_{m'}}$ if $k \neq k'$, where $m' \leqslant m = R+1, \ldots, L$. For DF-2, the conditional power (6) can be described as

$$CP = \begin{cases} \Pr\left[\bigcup_{m=R+1}^{L} A_{1m} \, | a_{1R}, a_{2l'}\right] = 1 - \Pr\left[\bigcap_{m=R+1}^{L} \bar{A}_{1m} \, | a_{1R}, a_{2l'}\right] \\ \quad \text{if } Z_{1l} \leqslant c_{1l} \text{ for all } l = 1, \ldots, R \text{ and } Z_{2l'} > c_{2l'} \text{ for some } l' = 1, \ldots, R, \\ \Pr\left[\bigcup_{m=R+1}^{L} A_{2m} \, | a_{2R}, a_{1l'}\right] = 1 - \Pr\left[\bigcap_{m=R+1}^{L} \bar{A}_{2m} \, | a_{2R}, a_{1l'}\right] \\ \quad \text{if } Z_{2l} \leqslant c_{2l} \text{ for all } l = 1, \ldots, R \text{ and } Z_{1l'} > c_{1l'} \text{ for some } l' = 1, \ldots, R, \\ \Pr\left[\left\{\bigcup_{m=R+1}^{L} A_{1m}\right\} \cap \left\{\bigcup_{m=R+1}^{L} A_{2m}\right\} \, | a_{1R}, a_{2R}\right] \\ \quad = 1 - \Pr\left[\bigcap_{m=R+1}^{L} \bar{A}_{1m} \, | a_{1R}, a_{2R}\right] - \Pr\left[\bigcap_{m=R+1}^{L} \bar{A}_{2m} \, | a_{1R}, a_{2R}\right] \\ \qquad + \Pr\left[\bigcap_{m=R+1}^{L} \left\{\bar{A}_{1m} \cap \bar{A}_{2m}\right\} | a_{1R}, a_{2R}\right] \\ \quad \text{if } Z_{1l} \leqslant c_{1l} \text{ and } Z_{2l} \leqslant c_{2l} \text{ for all } l = 1, \ldots, R, \end{cases}$$

and calculated similarly as discussed in the power for DF-2 (Appendix A.1).

When $R = L - 1$, the conditional power for DF-1 can be rewritten as

$$CP = \Pr[A_{1L} \cap A_{2L} \, | a_{1R}, a_{2R}] = \Phi_2\left(-c_1^*, -c_2^* | \rho\right),$$

where $\Phi_2(\cdot, \cdot|\rho)$ is the cumulative distribution function of the standard bivariate normal distribution with the correlation ρ, and $c_1^* = \left(c_{1L} - a_{1R}\sqrt{t}\right) / \sqrt{1-t} - \delta_1 \sqrt{n_L - n_R} / \sqrt{2}$ and $c_2^* = \left(c_{2L} - a_{2R}\sqrt{t}\right) / \sqrt{1-t} - \delta_2 \sqrt{n_L - n_R} / \sqrt{2}$ with $t = n_R / n_L$. For DF-2, the conditional power can be rewritten as

$$CP = \begin{cases} \Pr\left[A_{1L} \, | a_{1R}, a_{2l'}\right] = 1 - \Phi\left(c_1^*\right) \\ \quad \text{if } Z_{1l} \leqslant c_{1l} \text{ for all } l = 1, \ldots, R \text{ and } Z_{2l'} > c_{2l'} \text{ for some } l' = 1, \ldots, R, \\ \Pr\left[A_{2L} \, | a_{2R}, a_{1l'}\right] = 1 - \Phi\left(c_2^*\right) \\ \quad \text{if } Z_{2l} \leqslant c_{2l} \text{ for all } l = 1, \ldots, R \text{ and } Z_{1l'} > c_{1l'} \text{ for some } l' = 1, \ldots, R, \\ \Pr\left[A_{1L} \cap A_{2L} \, | a_{1R}, a_{2R}\right] = \Phi_2\left(-c_1^*, -c_2^* \, | \rho\right) \\ \quad \text{if } Z_{1l} \leqslant c_{1l} \text{ and } Z_{2l} \leqslant c_{2l} \text{ for all } l = 1, \ldots, R, \end{cases}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standardized normal distribution.

Acknowledgements

The authors are grateful to the two anonymous referees and the associate editor for their valuable suggestions and helpful comments that improved the content and presentation of the paper. We thank Dr. H. M. James Hung and Dr. Sue-Jane Wang for encouraging us in this research with their helpful advice. Research reported in this publication was supported by JSPS KAKENHI under Grant Number 23500348 and the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Numbers UM1AI104681 and UM1AI068634. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Committee for Medicinal Products for Human Use (CHMP). Guideline on Medicinal Products for the Treatment Alzheimer's Disease and Other Dementias (CPMP/EWP/553/95 Rev.1). EMEA: London, 2008.
2. Offen W, Chuang-Stein C, Dmitrienko A, Littman G, Maca J, Meyerson L, Muirhead R, Stryszak P, Boddy A, Chen K, Copley-Merriman K, Dere W, Givens S, Hall D, Henry D, Jackson JD, Krishen A, Liu T, Ryder S, Sankoh AJ, Wang J, Yeh CH. Multiple co-primary endpoints: medical and statistical solutions. *Drug Information Journal* 2007; **41**:31–46. DOI: 10.1177/009286150704100105.
3. Hung HMJ, Wang SJ. Some controversial multiple testing problems in regulatory applications. *Journal of Biopharmaceutical Statistics* 2009; **19**:1–11. DOI: 10.1080/10543400802541693.
4. Dmitrienko A, Tamhane AC, Bretz F. *Multiple Testing Problems in Pharmaceutical Statistics*. Chapman & Hall: Boca Raton, FL., 2010.
5. Xiong C, Yu K, Gao F, Yan Y, Zhang Z. Power and sample size for clinical trials when efficacy is required in multiple endpoints: application to an Alzheimer's treatment trial. *Clinical Trials* 2005; **2**:387–393. DOI: 10.1191/1740774505cn112oa.
6. Sozu T, Kanou T, Hamada C, Yoshimura I. Power and sample size calculations in clinical trials with multiple primary variables. *Japanese Journal of Biometrics* 2006; **27**:83–96. DOI: 10.5691/jjb.27.83.

7. Chuang-Stein C, Stryszak P, Dmitrienko A, Offen W. Challenge of multiple co-primary endpoints: a new approach. *Statistics in Medicine* 2007; **26**:1181–1192. DOI: 10.1002/sim.2604.
8. Eaton ML, Muirhead RJ. On multiple endpoints testing problem. *Journal of Statistical Planning & Inference* 2007; **137**:3416–3429. DOI: 10.1016/j.jspi.2007.03.021.
9. Senn S, Bretz F. Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics* 2007; **6**:161–170. DOI: 10.1002/pst.301.
10. Kordzakhia G, Siddiqui O, Huque MF. Method of balanced adjustment in testing co-primary endpoints. *Statistics in Medicine* 2010; **29**:2055–2066. DOI: 10.1002/sim.3950.
11. Sozu T, Sugimoto T, Hamasaki T. Sample size determination in clinical trials with multiple co-primary binary endpoints. *Statistics in Medicine* 2010; **29**:2169–2179. DOI: 10.1002/sim.3972.
12. Sozu T, Sugimoto T, Hamasaki T. Sample size determination in superiority clinical trials with multiple co-primary correlated endpoints. *Journal of Biopharmaceutical Statistics* 2011; **21**:650–668. DOI: 10.1080/10543406.2011.551329.
13. Julious S, Mclntyre NE. Sample sizes for trials involving multiple correlated must-win comparisons. *Pharmaceutical Statistics* 2012; **11**:177–185. DOI: 10.1002/pst.515.
14. Sugimoto T, Sozu T, Hamasaki T. A convenient formula for sample size calculations in clinical trials with multiple co-primary continuous endpoints. *Pharmaceutical Statistics* 2012; **11**:118–128. DOI: 10.1002/pst.505.
15. Hamasaki T, Sugimoto T, Evans SR, Sozu T. Sample size determination for clinical trials with co-primary outcomes: exponential event times. *Pharmaceutical Statistics* 2013; **12**:28–34. DOI: 10.1002/pst.1545.
16. Sugimoto T, Sozu T, Hamasaki T, Evans SR. A logrank test-based method for sizing clinical trials with two co-primary time-to-event endpoints. *Biostatistics* 2013; **14**:409–421. DOI: 10.1093/biostatistics/kxs057.
17. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**:659–663. DOI: 10.1093/biomet/70.3.659.
18. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979; **35**:549–556. DOI: 10.2307/2530245.
19. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 1977; **64**:191–199. DOI: 10.1093/biomet/64.2.191.
20. Glimm E, Maurer W, Bretz F. Hierarchical testing of multiple endpoints in group-sequential trials. *Statistics in Medicine* 2010; **29**:219–228. DOI: 10.1002/sim.3748.
21. Tamhane AC, Mehta CR, Liu L. Testing a primary and secondary endpoint in a group sequential design. *Biometrics* 2010; **66**:1174–1184. DOI: 10.1111/j.1541-0420.2010.01402.x.
22. Green RC, Schneider LS, Amato DA, Beelen AP, Wilcock G, Swabb EA, Zavitz KH, for the Tarenflurbil Phase 3 Study Group. Effect of tarenflurbil on cognitive decline and activities of daily living in patients with mild Alzheimer disease: a randomized controlled trial. *Journal of the American Medical Association* 2009; **302**:2557–2564. DOI: 10.1001/jama.2009.1866.
23. Doraiswamy PM, Bieber F, Kaiser L, Krishnan KR, Reuning-Scherer J, Gulanski B. The Alzheimer's disease assessment scale: patterns and predictors of baseline cognitive performance in multicenter Alzheimer's disease trials. *Neurology* 1997; **48**:1511–1517. DOI: 10.1212/WNL.48.6.1511.
24. Cui L, Hung HMJ, Wang SJ. Modification of sample size in group sequential clinical trials. *Biometrics* 1999; **55**:853–857. DOI: 10.1111/j.0006-341X.1999.00853.x.
25. Asakura K, Hayashi K, Sugimoto T, Sozu T, Hamasaki T. Sample size evaluation in group sequential designs for clinical trials with two continuous endpoints as co-primary contrasts. *Joint Statistical Meetings 2013*, Montreal, Quebec, Canada, August 3-8, 2013.
26. Hamasaki T, Asakura K, Sugimoto T, Evans SR. Sample size modification in group-sequential clinical trials with two co-primary endpoints. *Proceedings of Joint Meeting of the IASC Satellite Conference and 8th Conference of the Asian Regional Section of the IASC*, Seoul, Korea, August 21-14, 2013; 311–317.
27. Moyé LA. *Multiple Analyses in Clinical Trials*. Springer: New York, NY, 2003.
28. Moyé LA, Baraniuk S. Dependence, hyper-dependence and hypothesis testing in clinical trials. *Contemporary Clinical Trials* 2013; **28**:68–78. DOI: 10.1016/j.cct.2006.05.010.
29. Jennison C, Turnbull BW. Group sequential tests for bivariate response: interim analyses of clinical trials with both efficacy and safety. *Biometrics* 1993; **49**:741–752.
30. Tang DI, Geller NL. Closed testing procedures for group sequential clinical trials with multiple endpoints. *Biometrics* 1999; **55**:1188–1192. DOI: 10.1111/j.0006-341X.1999.01188.x.
31. Tamhane AC, Wu Y, Mehta C. Adaptive extensions of a two-stage group sequential procedure for testing primary and secondary endpoints (I): unknown correlation between the endpoints. *Statistics in Medicine* 2012; **31**:2027–2040. DOI: 10.1002/sim.5372.
32. Tamhane AC, Wu Y, Mehta C. Adaptive extensions of a two-stage group sequential procedure for testing primary and secondary endpoints (II): sample size re-estimation. *Statistics in Medicine* 2012; **31**:2041–2054. DOI: 10.1002/sim.5377.
33. Genz A, Bretz F. *Computation of Multivariate Normal and t Probabilities*. Springer Verlag: Berlin, 2009.

Supporting information

Additional supporting information may be found in online version of this article at the publisher's web site.

胎児鏡下手術

左合治彦*

胎児鏡を羊水腔に挿入して行う胎児鏡下手術には，双胎間輸血症候群に対する胎児鏡下レーザー凝固術，先天性横隔膜ヘルニアに対する胎児鏡下気管閉塞術，無心体に対する臍帯血流遮断術，羊膜索症候群に対する羊膜索離断術，下部尿路閉鎖症に対する尿路閉鎖解除術，脊髄髄膜瘤に対する修復術などがある。胎児鏡下手術は侵襲性が比較的低く，現在，胎児治療の主体となっているが，そのほとんどは胎児鏡下レーザー凝固術であり，妊娠26週未満の双胎間輸血症候群の第一選択治療法である。また胎児鏡下気管閉塞術も日本において早期安全性試験が始まった。

はじめに

胎児鏡の歴史は古く，超音波診断が普及する以前の1970年代に観察用として導入された。しかし，当時は内視鏡の径は太く侵襲性の問題があり，超音波診断技術の発達とともに胎児鏡の使用は衰退したが，内視鏡技術の進歩に伴い1990年代には胎児鏡による胎児手術が行われるようになった。胎児鏡を用いた胎盤・臍帯・羊膜に対する手術には，双胎間輸血症候群(twin-twin transfusion syndrome；TTTS)に対する胎児鏡下レーザー凝固術(fetoscopic laser photocoagulation；FLP)，無心体に対する臍帯血流遮断術，羊膜索症候群に対する羊膜索離断術などがある[1]。胎児に対する手術には，先天性横隔膜ヘルニア(congenital diaphragmatic hernia；CDH)に対する胎児鏡下気管閉塞術(fetal endoscopic tracheal occlusion/fetoscopic endoluminal tracheal occlusion；FETO)，下部尿路閉鎖症に対する尿路閉塞解除術，脊髄髄膜瘤に対する修復術などがある[1]。TTTSに対するFLPは有用性が証明された数少ない胎児治療法の1つで，現在行われている胎児治療法の大半を占める。CDHに対するFETOも有用性が期待されているが，その他の胎児鏡下胎児治療法は施行例も少なく有用性は不明である。本稿では，胎児鏡下手術の特徴，FLP，FETOについて概説する。また『日本胎児治療グループ(Japan Fetal Therapy Group)』のホームページでは，胎児治療についてわかりやすく解説してあるので参照されたい[2]。

1．胎児鏡下手術の特徴

胎児鏡下手術は，子宮内の羊水腔内へ内視鏡を挿入して行う手術であり，他の領域の内視鏡手術とは種々の面で異なる[1]。

1 次の対応策がない

通常の内視鏡手術では，手術困難な場合は開腹手術などで対応できるが，胎児鏡下手術は次の対応策がないため極めて慎重に行うことが求

* Haruhiko Sago 国立成育医療研究センター周産期・母性診療センター

められる。

2 超音波診断装置の併用

操作時に超音波診断装置を併用するため，内視鏡と超音波診断装置の２つの画像を見ながら手術を行う。超音波により胎盤や胎児の位置を確認し，安全に胎児鏡を挿入するとともに，術中の胎児心拍のモニタリングや出血や羊膜剝離など合併症の観察などに用いる。

3 羊水還流

通常の腹腔鏡下手術は腹腔内にCO_2ガスを挿入して行うが，子宮内の内視鏡は液体環境で行われる。ガス環境のほうが，明瞭な視野が得られる，電気メス・CO_2レーザーが使えるなどの利点があるが，CO_2ガスを用いると胎児がアシドーシスになり，また超音波が使えないなどの欠点がある。視野の確保やレーザーファイバー先端の保護のために人工羊水を注入する。

2．胎児鏡下レーザー凝固術（FLP）

1 TTTSの病態と治療原理

一絨毛膜双胎では，１つの胎盤を双胎間で共有しており，双胎間に胎盤吻合血管が存在する。TTTSは，一絨毛膜双胎において胎盤の吻合血管を介して双胎間に慢性の血流不均衡が起こり生じる病態で，供血児は羊水過少，受血児は羊水過多を認める。一絨毛膜双胎の約10％に発症するといわれており，児の発育不全，心不全，脳神経障害，早産，子宮内死亡などを併発し，妊娠中期に発症した場合の予後は極めて不良である。

FLPはTTTSの病因と考えられる両児間の胎盤血管吻合を遮断する治療法で，両児間の血流不均衡を是正する根治療法である。また一児死亡した場合も健児から死児への急性血液移行を防ぐことができ，一児死亡による健児への影響を回避できる。

2 FLPの適応

FLP手術の適応は，①TTTSである。すなわちMD双胎で一児に羊水過少（最大羊水深度2 cm以下）を認め，かつ，もう一児に羊水過多（最大羊水深度8 cm以上）を認める。②妊娠16週以上，26週未満，である。

FLP手術の要約は，1）未破水，2）羊膜穿破・羊膜剝離がない，3）明らかな切迫流早産徴候がない（頸管長20 mm以上を原則とする，10 mm以下は禁忌），4）母体に大きなリスクがない，5）母体感染症がない（HBV，HCV感染がないことを原則とする，HIV感染は禁忌），である。

3 FLPの手術方法

手術方法の模式図を図1に示す[3)4)]。胎盤吻合血管の観察がしやすい穿刺部位を選定し，超音波ガイド下で経皮的にトロッカー（約4 mm弱）を羊水過多の羊膜腔（受血児側）に挿入する。操作用のチャネルを有したシース（約3 mmの外筒）のなかに胎児鏡（2 mm：ドイツ・カールストルツ社製）を装着して，トロッカーを通して子宮内へ挿入して，胎盤表面の血管を観察する。通常，視野の良い硬性内視鏡を用いているが，前壁胎盤などで弯曲したシースを用いるときは半軟性内視鏡を用いている。胎盤の端から端まで観察して，動脈-静脈吻合，動脈-動脈吻合，静脈-静脈吻合など双胎間の吻合血管をすべて見出して，操作用のチャネルからから挿入したYAGレーザーファイバーを用いて凝固する。羊水を除去して手術を終了する。

4 FLPの治療成績

FLPは1990年代から欧米で積極的に行われ，2004年には，Eurofoetusによるランダム化比較対照試験で，26週未満のTTTSにおいて羊水吸引術に比べFLPが，より有効な治療法であることが立証された[5)]。わが国で2002年7月から2006年12月までにFLPを施行した181例（362児）の治療成績は欧米に劣らぬ良好な成績であった[6)]。手術施行妊娠週数の平均は21週で，術後の分娩週数の中間値は33週であった。2児生存率は約60％，一児生存率は約30％で，少なくとも1児生存率は90％であった。生後6カ月で神経後遺症を認めた児は5％であった。また2012年4月には胎児手術として初めて保険診療に収載された。現在，日本では年間約150例が施行され，少なくとも1児生存率は95％に達している。26週未満のTTTSに対してFLPは第

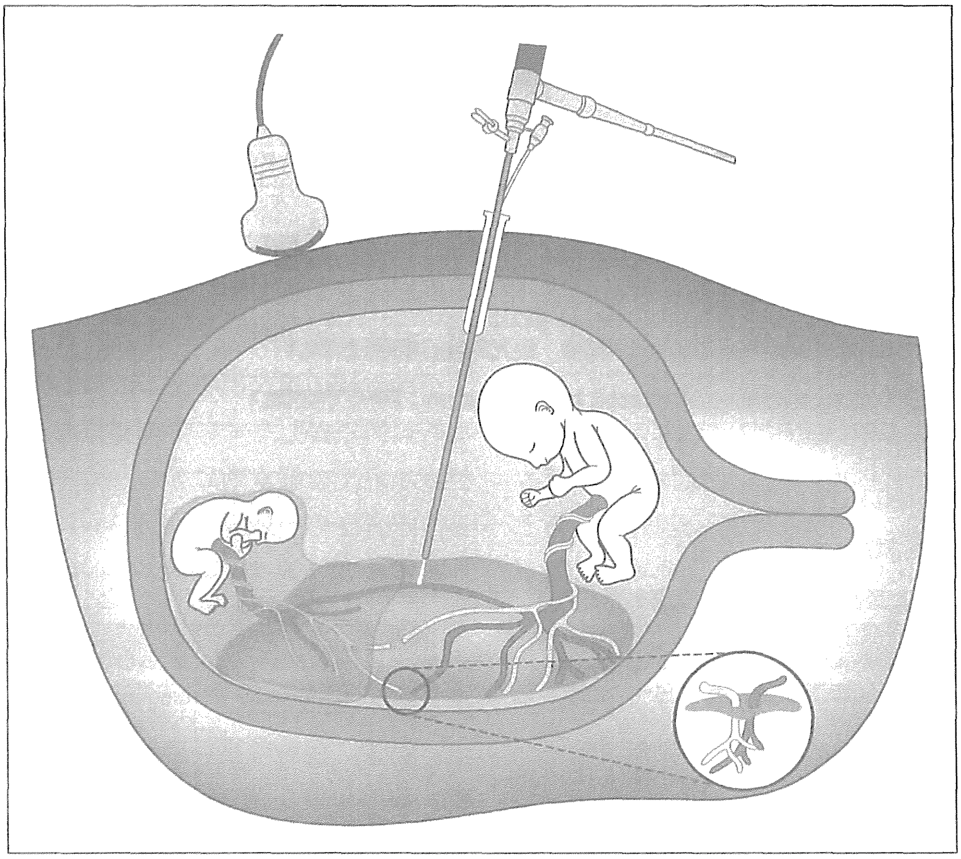

図1　胎児鏡下レーザー凝固術（FLP）の模式図

一選択治療法である。

5　FLPの適応拡大

FLPは胎盤吻合血管を凝固・遮断する治療であり，TTTSと同様に胎盤吻合血管により引き起こされる病態はFLPにより治療可能であると考えられる。FLPの有用性が証明されているのは妊娠26週未満のTTTSに対してのみであるが，FLPの治療成績は良好で施行例も多く，現在，FLPの適応拡大を試みている[2)7)]。

FLPの適応拡大対象と考えられ，臨床試験として行っているのは，以下の2つである。

① 妊娠適応期間の延長（28週未満）

妊娠26週以降のTTTSはFLPの適応外であるが，妊娠26～28週の治療成績も26週未満と変わらないとする報告もある[8)]。妊娠26週0日から27週6日のTTTS（ただし受血児の羊水過多10 cm以上）に対してFLPを施行することを臨床試験として実施している。

② selective IUGR（sIUGR）

Selective IUGR（sIUGR）は羊水過少・過多のTTTSの診断基準を満たさないが，羊水量の異常や胎児発育不全を認め，予後が不良でTTTS類似の病態と考えられる。一児の胎児推定体重が−1.5 SD以下で，小さい児の羊水過少（最大羊水深度1 cm以下）と血流異常（臍帯動脈拡張期途絶・逆流）を認める例の予後は極めて不良である[9)]。これらのsIUGRに対するFLPの有用性を確認するために臨床試験を実施している。

3．胎児鏡下気管閉塞術（FETO）

1　CDHの病態と治療原理

CDHは，横隔膜の先天的な欠損により腹腔臓器が胸腔内脱出する疾患で，正常肺の発育阻害から肺低形成となり，出生直後から呼吸障害と肺高血圧をきたす重篤な疾患である。生後に胸腔内の腹腔臓器を腹腔内に還納して横隔膜の欠損を修復するという手術が行われるが，死因の多くは肺低形成による呼吸不全である。腹腔臓器の嵌入が少ないものは肺低形成が少なく予後は良いが，肝臓が嵌入しているもので肺低形

表1 日本におけるFETOの早期安全性試験の治療適応と実施手順

治療適応基準
1）妊娠27週0日～31週6日
2）妊婦は16歳以上45歳未満
3）胎児は左側CDHと出生前診断されている単胎である
4）胎児は肝臓脱出型(liver up)のCDHであり，胃泡の半分以上が右胸腔内に脱出している(Kitanoの分類 Grade 3)
5）当該疾患以外の重篤な胎児奇形(染色体異常，致死的な心疾患)がない
6）妊娠高血圧症候群(pregnancy-induced hypertension；PIH)ではない
7）性器出血がない
8）破水していない
9）子宮頸管長20 mm以上である
10）患者本人と患者の配偶者から同意を得られている
実施手順
1）O/E LHR 25%未満は27週0日～29週6日で，O/E LHR 25%以上45%未満は30週0日～31週6日でFETOを施行
2）34週0日～34週6日でバルーン除去術の施行
3）出生後，標準化された積極的治療を施行

成が高度なものの予後は極めて悪い。

病因の主体は肺低形成と考えられ，FETOは胎児の気管を閉塞すると肺分泌液が貯留し肺が拡張して肺の成長が促されることを応用した治療法で，胎児鏡下で胎児の気管に着脱式バルーン（フランス・バルト社製）を挿入し，一時的気管閉塞を行い，低形成肺の発育を期待するものである。

2 FETOの適応

欧州では2008年よりTOTAL trialというFETOのランダム化比較臨床試験が行われている。日本では当センターにおいて2013年10月よりFETOの早期安全性試験が開始されたので，その適応を記載する[1]。

われわれは日本における胎児左CDH 109例の解析から，肝臓脱出型（liver up）で胃泡の半分以上が右胸腔内に脱出（Kitano分類Grade 3）している左CDHの予後が極めて不良であることを明らかにした[10]。また肺低形成の評価法であるobserved expected Lung Head Ratio（o/eLHR）と予後は良く相関するといわれており，TOTAL trialの適応基準に用いられている。そこで両者を組み合わせて適応基準と実施手順を作定した（表1）。要点は，①他の合併奇形のない左CDH，②肝臓脱出型で胃泡の半分以上が右胸腔内に脱出（Kitano分類Grade 3），③o/eLHR25%未満は妊娠27週0日から29週6日に，o/eLHR25%以上45%未満は30週0日から31週6日にFETOを施行すること，である。

3 FETOの手術方法

手術方法の模式図を図2に示す。穿刺部位を選定し，超音波ガイド下で経皮的にトロッカー（約4 mm弱）を子宮内の羊膜腔に挿入する。操作用のチャネルを有したFETO専用のシース（約3 mmの外筒）のなかに胎児鏡（1.3 mm：ドイツ・カールストルツ社製）を装着して，トロッカーを通して子宮内へ挿入する。胎児鏡を胎児の口へ挿入し，喉頭蓋の見える位置まで進め，喉頭蓋の裏側の気管内へ挿入する。気管分岐部を確認し，その少し口側で着脱型のバルーン

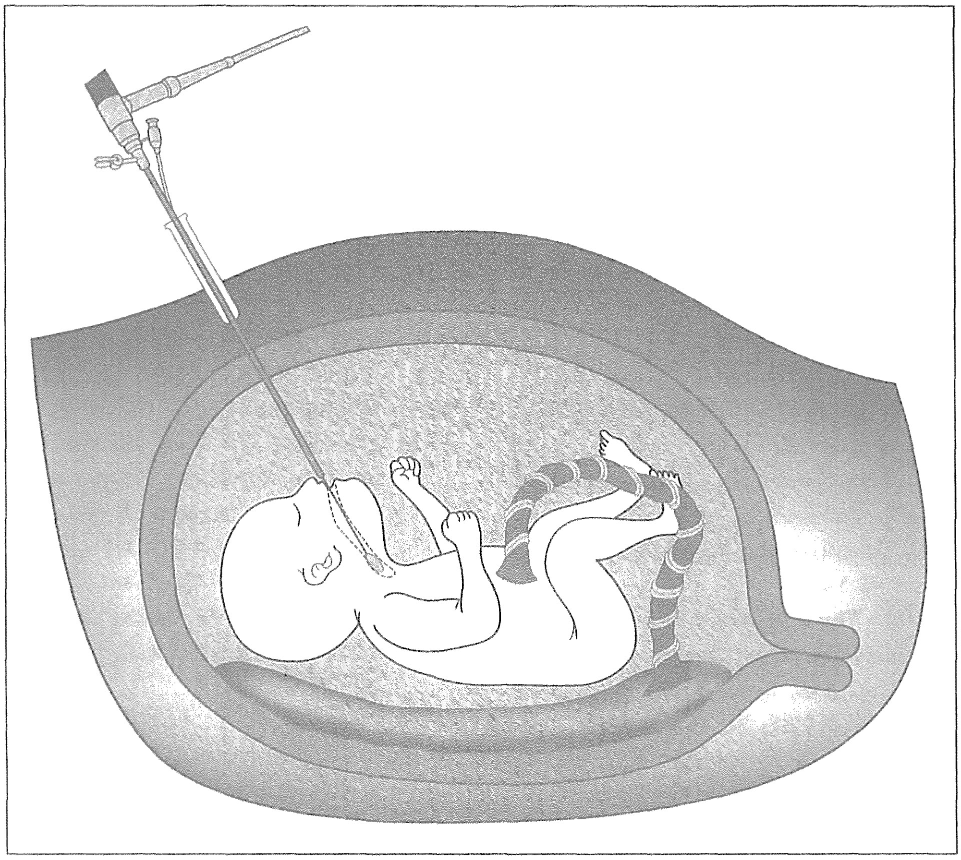

図2　胎児鏡下気管閉塞術（FETO）の模式図

（フランス・バルト社製）を膨らませて留置する。

4　バルーン抜去術

バルーン抜去術は，妊娠34週0日から34週6日に行う。超音波下で穿刺針によるバルーンの穿破が可能な場合は，超音波下で行う。不可能な場合は，バルーン挿入と同じ要領で胎児鏡を胎児の気管内へ挿入し，操作用チャネルから穿刺針を挿入してバルーンを穿破する。34週未満に分娩となる場合は，出生直後に気管支鏡下でバルーンを抜去する。出生直後にバルーン抜去が必要となる場合があり，緊急で対応できる体制が重要である。

5　FETOの治療成績

米国におけるFETOのランダム化比較試験（計24例）では，対照群（生後治療）の成績が予想外に高く，有用性を示すことはできなかった[11]。これは胎児治療の適応基準に問題があると考えられた。欧州では適応基準を見直し，また，より低侵襲な手術方法を用いて2001～2008年に210例のFETOを施行した。左CDHのFETOの生存率は49％で，同じ重症度に相当する生後治療の生存率は24％であった[12]。期待できる結果であったが，あくまで後ろ向き研究であり，2008年からTOTAL trialというランダム化比較試験が開始された。2012年のブラジルのランダム化比較試験の結果は，生存率はFETO群（20例）50％，対照群（21例）5％と有意な結果であったが，バルーン抜去はex utero intrapartum treatment（EXIT）で行っていた[13]。FETOの生存率は50％前後と推定され，適応基準が重要となる。日本の早期安全性試験，TOTAL trialの成果を期待したい。

おわりに

TTTSに対するFLPは確立された胎児治療法で保険適用となり，日常産科臨床において行われている。CDHに対するFETOは有用性が期待されており，日本においても早期安全性試験が始まった。胎児鏡下手術は比較的侵襲度が低く，今後期待される胎児治療法である。

文　献

1) 左合治彦：内視鏡による胎児手術．医学のあゆみ 207：409-413，2003

2) 日本胎児治療グループ(Japan Fetal Therapy Group)
http://fetusjapan.jp/

3) 左合治彦：一絨毛膜双胎 基本から update まで—双胎間輸血症候群に対する治療．p135-157，メディカルビュー社，2007

4) 左合治彦：胎児手術—双胎間輸血症候群．OGS now No. 15 妊娠中の手術・胎児手術，p126-133，メジカルビュー社，2013

5) Senat MV et al : Endoscopic laser surgery versus serial amniorreduction for severe twin-to-twin transfusion syndrome. N Engl J Med 351 : 136-144, 2004

6) Sago H et al : The outcome and prognostic factors of twin-twin transfusion syndrome following fetoscopic laser surgery. Prenat Diagn 30 : 1185-1191, 2010

7) 左合治彦：クリニカルディベート 周産期——絨毛膜双胎娩出のタイミング—胎児鏡下レーザー凝固術積極的適応の立場に立って．日産科婦人科会誌 63：N-191-195，2011

8) Baud D et al : Fetoscopic laser therapy for twin-twin transfusion syndrome before 17 andafter 26 weeks' gestation. Am J Obstet Gynecol 208 : 197. e1-7, 2013

9) Ishii K et al : Ultrasound and Doppler predictors of mortality in monochorionic twins with selective intrauterine growth restriction. Ultrasound Obstet Gynecol. 37 : 22-6, 2011

10) Kitano Y et al : Reevaluation of stomach position as a simple prognostic factor in fetal left congenital diaphragmatic hernia : a multicenter survey in Japan. Ultrasound Obstet Gynecol 37 : 277-282, 2011

11) Harrison MR et al : A randomized trial of fetal endoscopic tracheal occlusion for severe fetal congenital diaphragmatic hernia. N Engl J Med 349 : 1916-1924, 2003

12) Jani JC et al : Severe diaphragmatic hernia treated by fetal endoscopic tracheal occlusion. Ultrasound Obstet Gynecol. 34 : 304-31, 2009

13) Ruano R et al : A randomized controlled trial of fetal endoscopic tracheal occlusion versus postnatal management of severe isolated congenital diaphragmatic hernia. Ultrasound Obstet Gynecol 39 : 20-27, 2012