Supplemental Materials to "Modified Goldilocks Design with Strict Type I Error Control in Confirmatory Clinical Trials"

February 14, 2020

1 Generalization to Multiple Interim Checkpoints

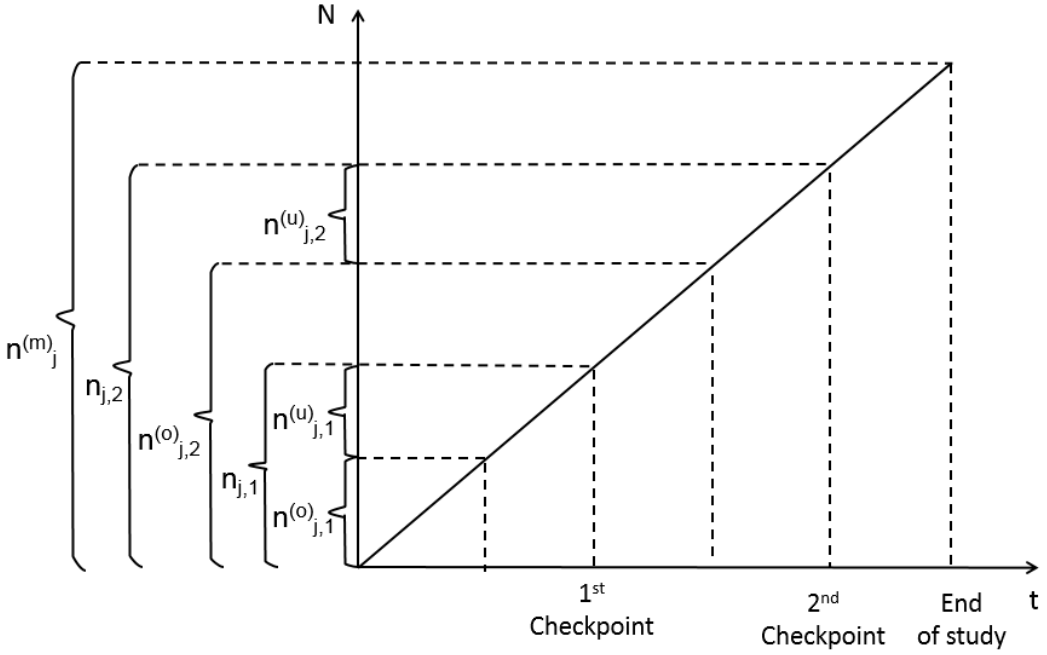
In the main article, we incorporate one ISC to the study and it corresponds to a two-stage design. Here we generalize it to multiple ISCs.

Suppose we have L checkpoints to select a trial's sample size based on predictive probabilities, then each interim look l is performed when $n_{j,l}$ subjects have been recruited to group j. Note that $n_{j,1} < n_{j,2} < ... < n_{j,L} < n_j^{(m)}$, where $n_j^{(m)}$ is the maximum sample size for group j defined earlier. Correspondingly, there will be $n_{j,l}^{(o)}$ subjects that have developed their responses while $n_{j,l}^{(u)}$ have not. Graphical illustrations of population notions in MGD with two ISCs are provided at Figure 1. At interim look l, predictive probabilities $\text{MPP}_{n,l}$ and $\text{MPP}_{max,l}$ are calculated from all currently available data. We ignore the subscripts l for simplicity in the remainder of this section.

For final efficacy analysis, stagewise p-values in MGD are calculated from separate cohorts of subjects as illustrated in the single ISC case in the main article. As a generalization to multiple looks, at checkpoint l, if we observe $\text{MPP}_{max} \geq F_n$ and $\text{MPP}_n > S_n$, which means

that we will stop recruiting subjects, then p_{l+1} is calculated from data $\left[n_{c,l}^{(u)}, n_{t,l}^{(u)}\right]$. On the other hand, if we observe $\text{MPP}_{max} \geq F_n$ and $\text{MPP}_n \leq S_n$, we will continue recruiting subjects to the next checkpoint $l+1$ (when $n_{j,l+1}$ subjects are recruited). The p-value p_{l+1} is instead calculated using data $\left[n_{c,l+1}^{(o)} - n_{c,l}^{(o)}, n_{t,l+1}^{(o)} - n_{t,l}^{(o)}\right)$. If $l+1$ is the end of study enrollment, then p_{l+1} is calculated from $\left[n_c^{(m)} - n_{c,l}^{(o)}, n_t^{(m)} - n_{t,l}^{(o)}\right]$. A rule of thumb is that a cohort of subjects would not contribute to both the predictive probability at one stage and the p-value calculation in the following stage.

Figure 1: Population Notations for MGD with Multiple Checkpoints



Having obtained all available stagewise p-values, we could further combine them to test the null hypothesis. Assume a study has L' ($L' \leq L$) checkpoints before stopping enrollment and hence $L'+1$ stagewise p-values. If a study enrolls subjects to the maximum sample size n_j^m, then $L = L'$. A backward recursion algorithm [Brannath et al., 2002] can be applied if there are more than one ISC. A p-value for the combination test is defined by

$$q(p_1, p_2) = \begin{cases} p_1 & \text{if } p_1 \leq \alpha_1 \text{ or } p_1 \geq \alpha_0 \\ \alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_0^1 I[C(x,y) \leq c(\alpha_1, \alpha_0, \alpha) dy dx & \text{otherwise.} \end{cases}$$

In MGD, one would set $\alpha_1 = 0$ and obtain α_0 as described in the previous section. The recursive combination tests operate as following. First calculate combined p-values $q_{L'}(p_{L'}, p_{L'+1})$ for the last two stages, where $p_{L'}$ and $p_{L'+1}$ are p-values from the L'th and $(L'+1)$th stages, respectively. Then one could insert $q_{L'}(p_{L'}, p_{L'+1})$ to $q_{L'-1}[p_{L'-1}, q_{L'}(p_{L'}, p_{L'+1})]$ for the calculation of combined p-value in a previous stage, and repeat these steps up to the first stage. The null hypothesis is rejected if $q_1[p_1, q_2(.)] < \alpha$, where α is the pre-specified type I error rate. If the so-called "p-clud" condition is satisfied for all stage-wise p-values, then the backward recursion algorithm could preserve overall type I error at α [Liu et al., 2002].

2 Additional simulation results

2.1 Binary endpoint with modified predictive probabilities

We provided additional simulations results of using modified predictive probabilities in MGD with binary endpoint in Table 1 and 2 to evaluate type I error control and power performance.

2.2 Binary endpoint in a three-stage design

In this section, we provide a small simulation study for binary endpoint in a three-stage design with two ISCs. We consider a design with no futility stopping ($F_n = 0$) and three varying S_n's at 0.8, 0.85 and 0.9. The number of subjects per group with observed response is $n_1^{(o)} = 100$ by the first ISC, and $n_2^{(o)} = 200$ by the second ISC. The number of subjects with unobserved response is $n_1^{(u)} = n_2^{(u)} = 40$. The maximum sample size per group is $n^{(m)} = 300$. The number of simulation iteration is 10^5, and we use 10^3 samples to evaluated P_n and P_{max} by Monte Carlo integration. Equal weights are utilized in the Inverse Normal combination function.

As shown in Table 3, our MGD accurately controls the type I error at the nominal level $\alpha = 0.025$ at the null scenarios when the treatment effect $\Delta = 0$. The critical value $c_g(\alpha)$ is

Table 1: Type I error rate with modified P_n^F, P_{max}^F, P_n^I, and P_{max}^I

Scenario	θ_0	Sn	Fn	$n^{(o)}$	$n^{(u)}$	MGD^F	MGD^I	GD
S1	0.6	0.8	0.1	80	15	0.025	0.025	0.028
					20	0.025	0.025	0.027
					40	0.025	0.025	0.025
					60	0.025	0.025	0.024
					80	0.025	0.025	0.023
					100	0.025	0.025	0.022
S2	0.4	0.9	0	80	15	0.025	0.025	0.028
					20	0.025	0.025	0.027
					40	0.025	0.025	0.025
					60	0.025	0.025	0.024
					80	0.025	0.025	0.024
					100	0.025	0.025	0.024
S3	0.1	0.8	0.1	80	20	0.024	0.025	0.026
	0.3					0.025	0.025	0.027
	0.4					0.025	0.025	0.027
	0.6					0.025	0.025	0.027
	0.8					0.025	0.025	0.027
	0.9					0.024	0.025	0.026
S4	0.6	0.8	0	20	20	0.025	0.026	0.028
				40		0.025	0.025	0.029
				60		0.025	0.025	0.029
				80		0.025	0.025	0.029
				100		0.025	0.024	0.028
				120		0.025	0.025	0.028

adjusted to be 0.018 for GD to control the error rate at 0.025. In power evaluation, GD is slightly more powerful than MGD when $S_n = 80\%$, but becomes similar for $S_n = 90\%$.

2.3 Operating characteristics of MGD^F in time-to-event endpoint

The operating characteristics of MGD^F in the time-to-event case study at Section 5 are at Table 4.

References

[Brannath et al., 2002] Brannath, W., Posch, M., and Bauer, P. (2002). Recursive combination tests. *Journal of the American Statistical Association*, 97(457):236–244.

[Liu et al., 2002] Liu, Q., Proschan, M. A., and Pledger, G. W. (2002). A unified theory of two-stage adaptive designs. *Journal of the American Statistical Association*, 97(460):1034–1041.

Table 2: Power with modified P_n^F, P_{max}^F, P_n^I, and P_{max}^I

Scenario	θ_c	Δ	Sn	Fn	$n^{(o)}$	$n^{(u)}$	MGD^F	MGD^I	GD
S1	0.1	0	0.8	0.1	80	15	0.024	0.025	0.027
	0.3						0.025	0.025	0.028
	0.5						0.025	0.025	0.027
	0.7						0.025	0.025	0.028
	0.9						0.024	0.025	0.027
S1	0.3	0.143	0.8	0.1	80	15	0.811	0.788	0.804
		0.17					0.919	0.885	0.910
		0.188					0.958	0.922	0.950
	0.7	0.127					0.819	0.810	0.811
		0.145					0.913	0.895	0.903
		0.16					0.959	0.937	0.949
S2	0.1	0	0.8	0.1	80	40	0.024	0.025	0.024
	0.3						0.025	0.025	0.025
	0.5						0.025	0.025	0.023
	0.7						0.025	0.025	0.025
	0.9						0.024	0.025	0.024
S2	0.4	0.15	0.8	0.1	80	40	0.822	0.810	0.807
		0.175					0.919	0.905	0.907
		0.195					0.962	0.949	0.953
	0.6	0.14					0.813	0.801	0.805
		0.16					0.907	0.892	0.898
		0.18					0.960	0.947	0.953
S3	0.1	0	0.8	0.1	80	80	0.024	0.025	0.022
	0.3						0.025	0.025	0.022
	0.5						0.025	0.025	0.021
	0.7						0.025	0.025	0.023
	0.9						0.025	0.025	0.022
S3	0.4	0.149	0.8	0.1	80	80	0.819	0.814	0.806
		0.172					0.912	0.907	0.902
		0.192					0.958	0.954	0.952
	0.6	0.137					0.797	0.797	0.792
		0.159					0.905	0.903	0.902
		0.178					0.958	0.956	0.956

Table 3: Type I error and Power in a three-stage design

Scenario	θ_c	Δ	Sn	MGD^F	MGD^I	GD
S1	0.2	0	0.8	0.025	0.025	0.025
	0.4			0.025	0.025	0.024
	0.6			0.024	0.025	0.024
	0.8			0.025	0.025	0.025
	0.6	0.1		0.640	0.653	0.667
		0.15		0.933	0.921	0.941
S2	0.2	0	0.85	0.025	0.025	0.024
	0.4			0.025	0.026	0.024
	0.6			0.025	0.025	0.024
	0.8			0.025	0.025	0.024
	0.6	0.1		0.649	0.664	0.672
		0.15		0.943	0.928	0.946
S3	0.2	0	0.9	0.025	0.026	0.023
	0.4			0.026	0.026	0.023
	0.6			0.025	0.026	0.022
	0.8			0.025	0.025	0.022
	0.6	0.1		0.665	0.678	0.678
		0.15		0.955	0.940	0.955

Table 4: Operating characteristics of MGD^F in time-to-event endpoint

Enrollment rate (per month)	Control 1-year survival	Hazard ratio	Early stopping for futility	Early stopping and predicting success	Type I error / power	ASN
10	0.20	1.00	0.564	0.006	0.021	286
	0.30	1.00	0.542	0.006	0.020	290
	0.50	1.00	0.502	0.007	0.020	298
	0.60	1.00	0.487	0.008	0.021	301
	0.30	0.72	0.081	0.141	0.669	355
	0.30	0.66	0.039	0.235	0.844	345
	0.30	0.61	0.016	0.355	0.943	326
5	0.20	1.00	0.624	0.006	0.024	274
	0.30	1.00	0.610	0.006	0.023	277
	0.50	1.00	0.597	0.006	0.023	279
	0.60	1.00	0.587	0.006	0.024	281
	0.30	0.74	0.097	0.160	0.632	348
	0.30	0.68	0.045	0.277	0.824	336
	0.30	0.63	0.017	0.422	0.936	312