# Group sequential design with maximin efficiency robust test for immunotherapy with generalized delayed treatment effect

# Web Appendix

**Web Appendix A. Performance of** **the alternative boundary methods**

**Web Appendix B. Validity of MERT under strong null hypothesis**

**Web Appendix C. Boundary validity against the mis-specified *S*0(*t*)**

**Web Appendix D. Comparison between MERT and OWLR**

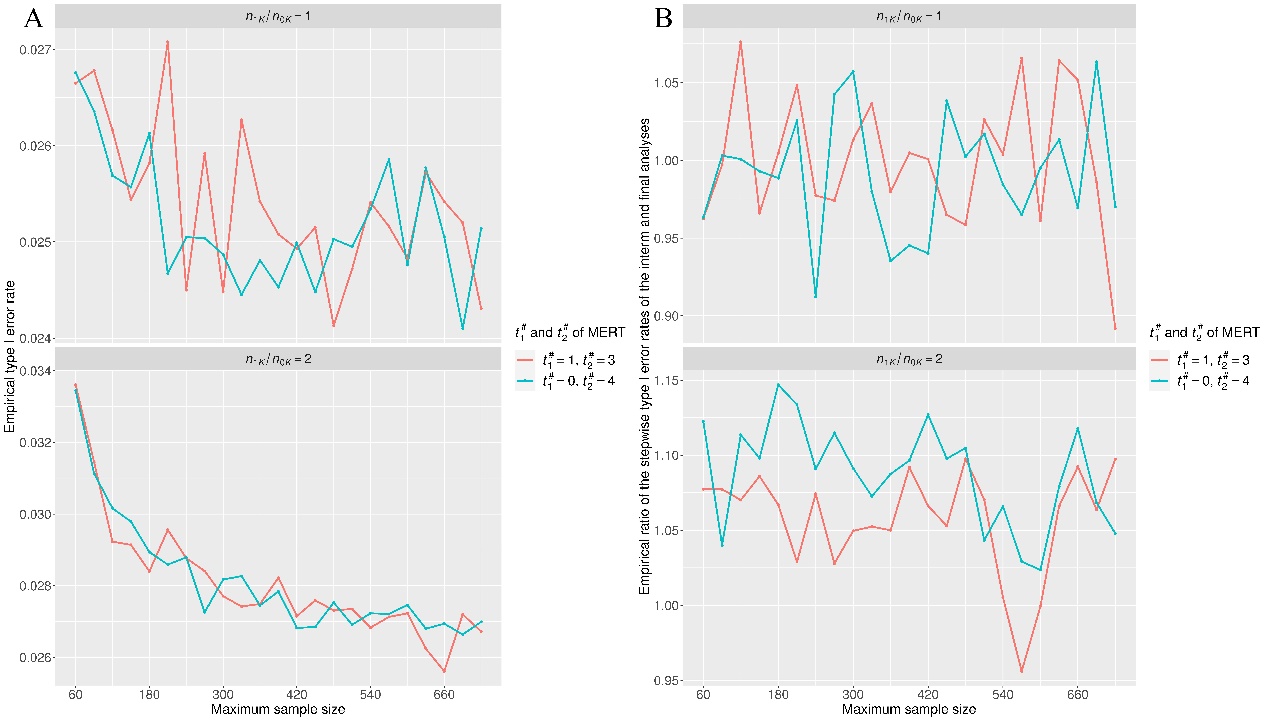
**Web Appendix** **E. Robustness under** **random delay pattern**

**Web Appendix F. Power influences from mis-specified lag functions**

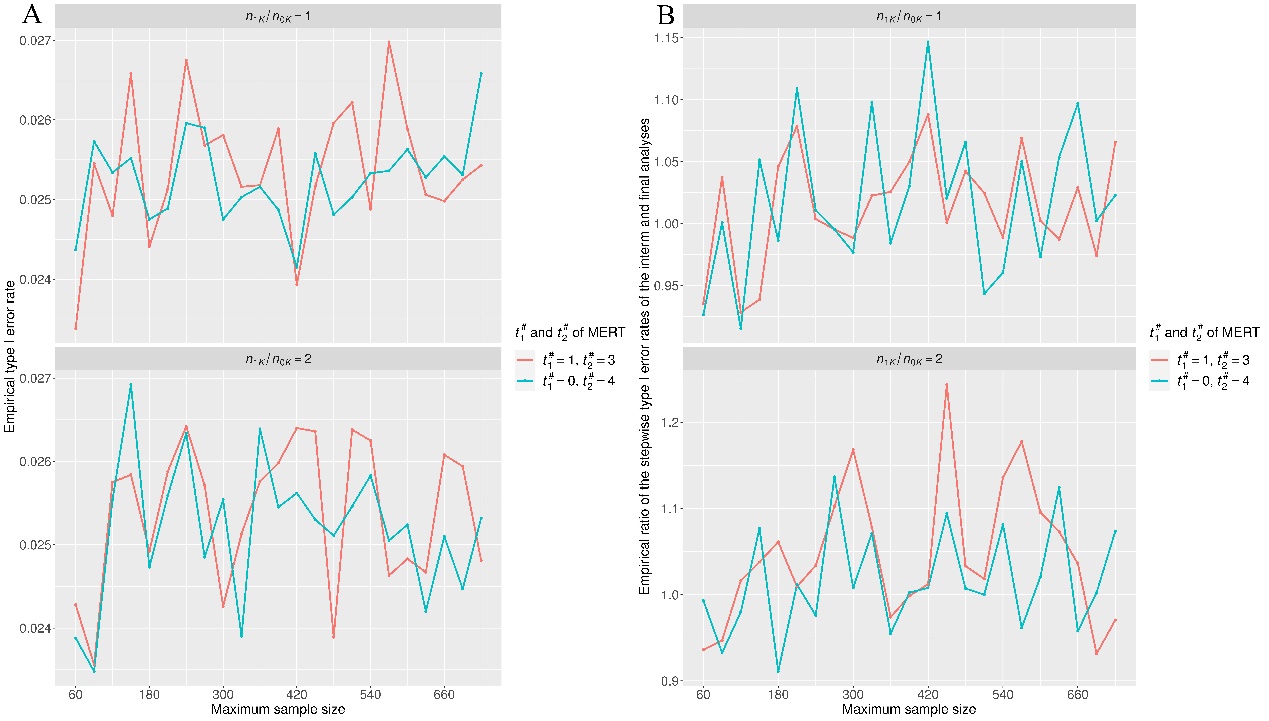
## A Performance of the alternative boundary methods

In this section, we made a supplementary simulation evaluation on the performance of two common alternative group sequential boundary determination methods, such as the integration-based1 and simulation-based methods2,3. Specifically, we investigated the influence of the maximum sample size and the survival function of the control group on the empirical type Ⅰ error rates given the corresponding group sequential boundaries, which were obtained using the above two alternative methods. Besides, the simulation scenarios were specified the same as those in the subsection 4.1 in the main text. Then, we determine the group sequential boundaries for each of the specified scenarios using the integration-based1 and simulation-based methods2,3. And we used the Monte Carlo simulation method to calculate the empirical type Ⅰ error rates under the null hypothesis of no difference, which was presented in Equation (3) in the main text, for each scenario and the corresponding group sequential boundaries.

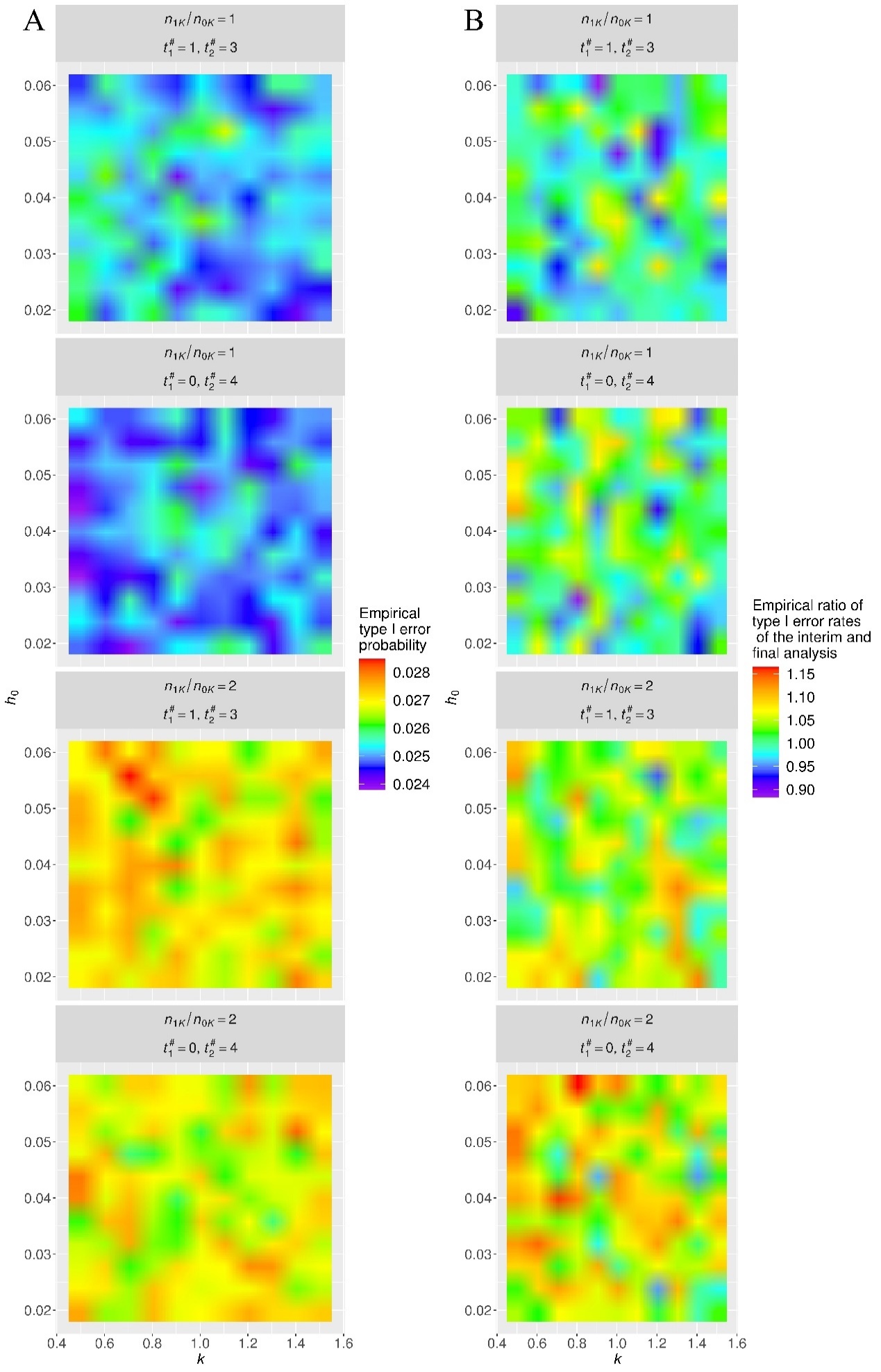
The simulation results reflecting the influence of the maximum sample size on the empirical type Ⅰ error rates were shown in Figures S1 and S2 that corresponded to the integration-based and simulation-based methods, respectively. Figure S1A and S2A depicted the trend between the empirical type I error rate and maximum sample size, whereas Figures S1B and S2B shows a relationship between the maximum sample size and the empirical ratio of the stepwise empirical type I error rate of the interim analyses to that of the ﬁnal analysis. Additionally, the simulation results that reflected the influence of the survival function of the control group on the empirical type Ⅰ error rates were presented in Figures S3 and S4, which corresponded to the integration-based and simulation-based methods, respectively. Figure S3A and S4A depicted the trend between the empirical type I error rate and survival function of the control group, whereas Figures S3B and S4B shows a relationship between the survival function of the control group and the empirical ratio of the stepwise empirical type I error rate of the interim analyses to that of the ﬁnal analysis.



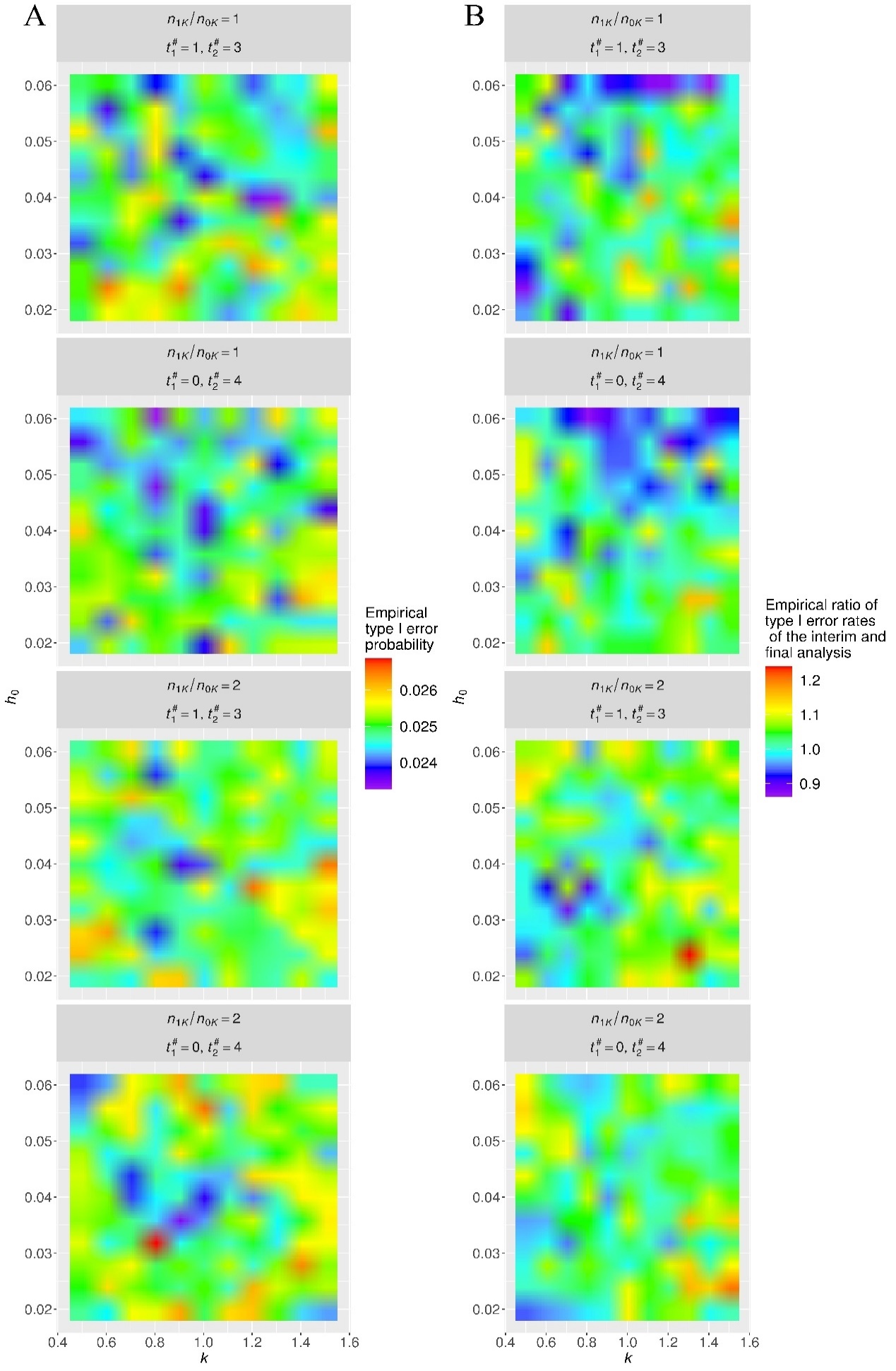
**Figure S1. The overall value and allocation ratio between the interim and ﬁnal analyses of the empirical type I error rates versus the maximum sample size when the integration method adopted.**



**Figure S2. The overall value and allocation ratio between the interim and ﬁnal analyses of the empirical type I error rates versus the maximum sample size when the simulation method adopted.**



**Figure S3. The overall value and allocation ratio between the interim and ﬁnal analyses of the empirical type I error rates versus the survival function of the control group when the integration method adopted.**



**Figure S4. The overall value and allocation ratio between the interim and ﬁnal analyses of the empirical type I error rates versus the survival function of the control group when the simulation method adopted.**

According to Figure S1, we found that the maximum sample size *nK* had a significance influence on the overall type I error rate and the actual spending fractions of the type I error rate when the integration method was adopted to determine the group sequential boundaries. First, the overall type I error rate was well-controlled at the signiﬁcance level under a balanced design, except slight inflations for extremely small *nK*. However, under the unbalanced design, the overall type I error rate notably inﬂated for the extremely small *nK*, whereas the inﬂation remained within an acceptable range when *nK* was large enough. Additionally, the type I error rate was spent in each analysis, almost in accordance with the pre-speciﬁed spending fractions under the balanced design. However, the type I error rate was often spent more in the interim analysis than in the ﬁnal analysis under the unbalanced design, compared with the specified fractions. The trends were similar with those when the Markov-chain method was used to determine the group sequential boundaries, as reflected by Figure 1 in the main text.

According to Figure S2, we found that the maximum sample size *nK* had no notable influence on the he overall type I error rate and the actual spending fractions of the type I error rate when the simulation method was adopted to determine the group sequential boundaries. First, the overall type I error rate fluctuated around the significance level under both the balanced design and unbalanced design with *nK* increasing. Second, the ratio of the empirical type I error rates between the interim and final analyses fluctuated around the pre-specified ratio both the balanced design and unbalanced design with the increase of the *nK*. These two phenomena testified the potential ability of the simulation method to determine the accurate group sequential boundaries under both balanced and unbalanced design. However, the fluctuation ranges were moderately wide, implying that some determined group sequential boundaries may not be accurate enough. In our views, the reason for this situation was that the randomness of the simulation method may have relatively notable influences on accuracy of the determined group sequential boundaries when the number of repetitions was not large enough. In other words, the considerable increase in the number of repetitions in the simulation-based procedure may be able to improve the accuracies of the group sequential boundaries.

According to Figure S3, we found that the survival function of the control group had no remarkable influences on the overall type I error rates and spending fractions of the type I error rates between the interim and final analysis when the group sequential boundaries are determined under each scenario by the integration method. However, Figure S3 also reflected some consistent trends of the type I error rates for the determined group sequential boundaries. First, the overall empirical type I error rate was usually around the significance level under the balanced design, whereas the overall empirical type I error rate inflated slightly under unbalanced design. Second, the empirical type I error rate was spent almost in accordance with the pre-specified fractions under a balanced design, whereas the empirical type I error rate was usually more spent in the interim analysis than the final analysis under an unbalanced design. These trends were similar with that when the Markov-chain method was used to determine the group sequential boundaries, as reflected by Figure 1 in the main text, as reflected by Figure 2 in the main text.

According to Figure S4, we found that the survival function of the control group had no remarkable influences on the overall type I error rates and spending fractions of the type I error rates between the interim and final analysis when the group sequential boundaries are determined under each scenario using the simulation method. Additionally, Figure S4 also reflected that the group sequential boundaries determined by the simulation method was accurate in most cases, which was drawn from the following two facts. First, the overall empirical type I error rate was usually very close to the significance level when the balanced or unbalanced design was considered. Second, the empirical type I error rate was in common spent almost in accordance with the pre-specified fractions under when the balanced or unbalanced design was considered. However, the randomness of the simulation method rarely produced inaccurate group sequential boundaries, which led to accidental deviations of the overall empirical type I error rates and the empirical ratio of the type I error rates of the interim and final analyses.

## B Validity of MERT under strong null hypothesis

In this section, we made some supplementary evaluations of the group sequential boundaries under a certain scenario which satisfies the following strong null hypothesis: . First, we determine the group sequential boundaries using the proposed method under the corresponding scenario which satisfies the weak null hypothesis, i.e., . Second, we calculate the empirical type Ⅰ error rate under the specified scenario satisfying the above strong null hypothesis using the Monte Carlo simulation procedure. Besides, various specifications of the maximum sample size, allocation ratio and of MERT statistics are considered in the simulation study of this section.

The simulation scenarios are specified as follows: The recruitment of patients was planned to last for 24 months, while one interim analysis and the final analysis were scheduled at the 36th and 60th month, respectively. The parameters of the Markov chain model were set as *ϵ*=0.1 and *b*=10. The rate of loss to follow-up during each month was specified as 0.74%. The survival function of the control group is specified by and . The survival function of the immunotherapy group is set by , and where , And the survival functions of two groups are shown in Figure S5. The specifications of in MERT were (0,3), (3,6) and (6,9). The allocation ratio of the immunotherapy and control groups was specified as 1:1 or 2:1. The true maximum sample size increased from 300 to 1200 with an increment of 30. Under the null hypothesis of no difference, the significance level is set to 0.025, which is allocated equally to the interim analysis and the final analysis. Besides, the number of repetitions in the Monte Carlo simulation procedure is 100000.

The results are presented in Figure S6. First, the empirical type Ⅰ error rate is nearly always zero when the interval is on the left side of the interval . Second, the empirical type Ⅰ error inflates moderately and increases with the maximum sample size when the intervals and are identical. Third, the empirical type Ⅰ error inflates considerably and increases notably with the maximum sample size when the interval is on the right side of the interval .

图表, 折线图, 直方图

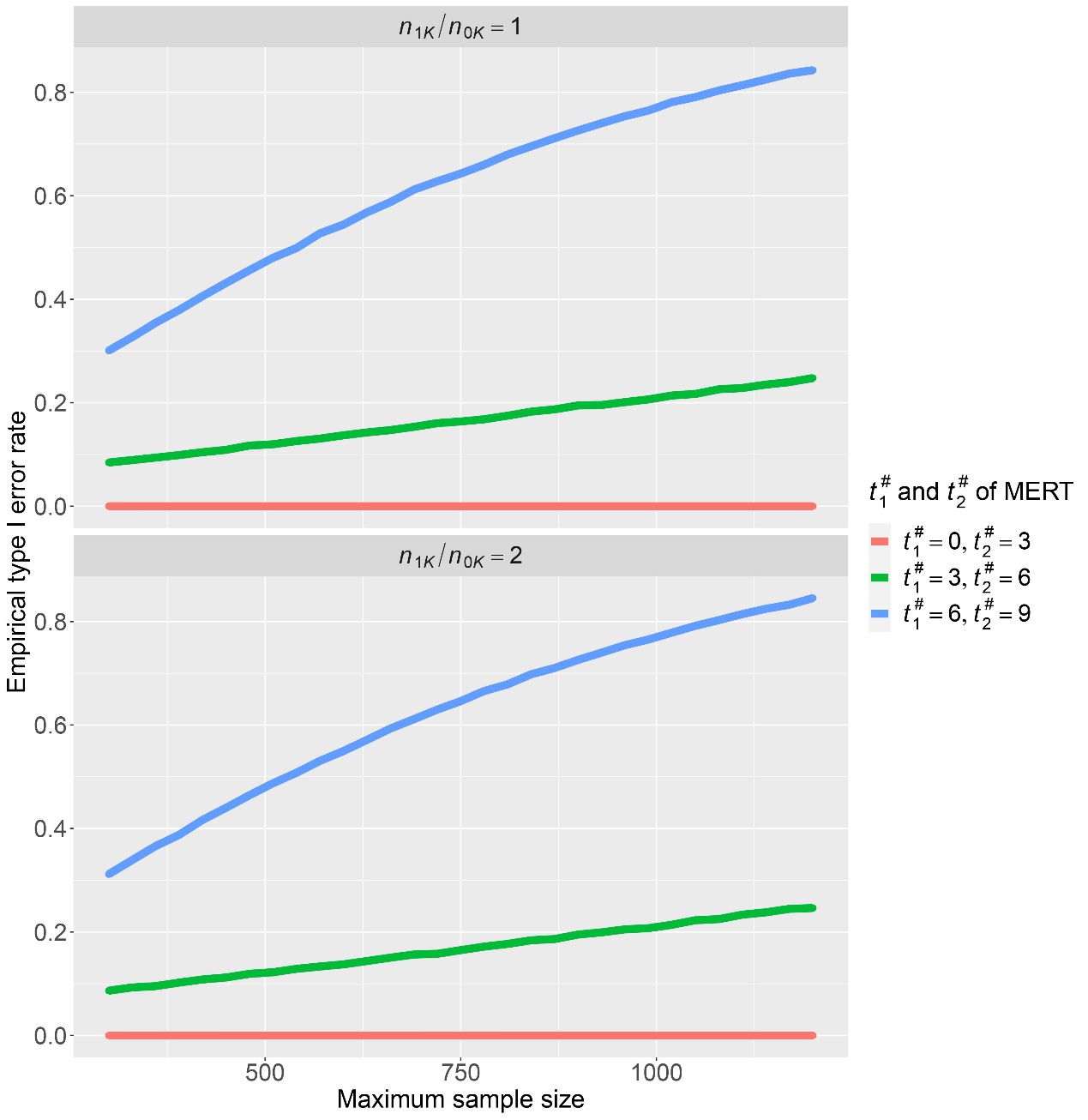
描述已自动生成

**Figure S5. The survival functions of the control and immunotherapy group under one scenario satisfying the strong null hypothesis.**

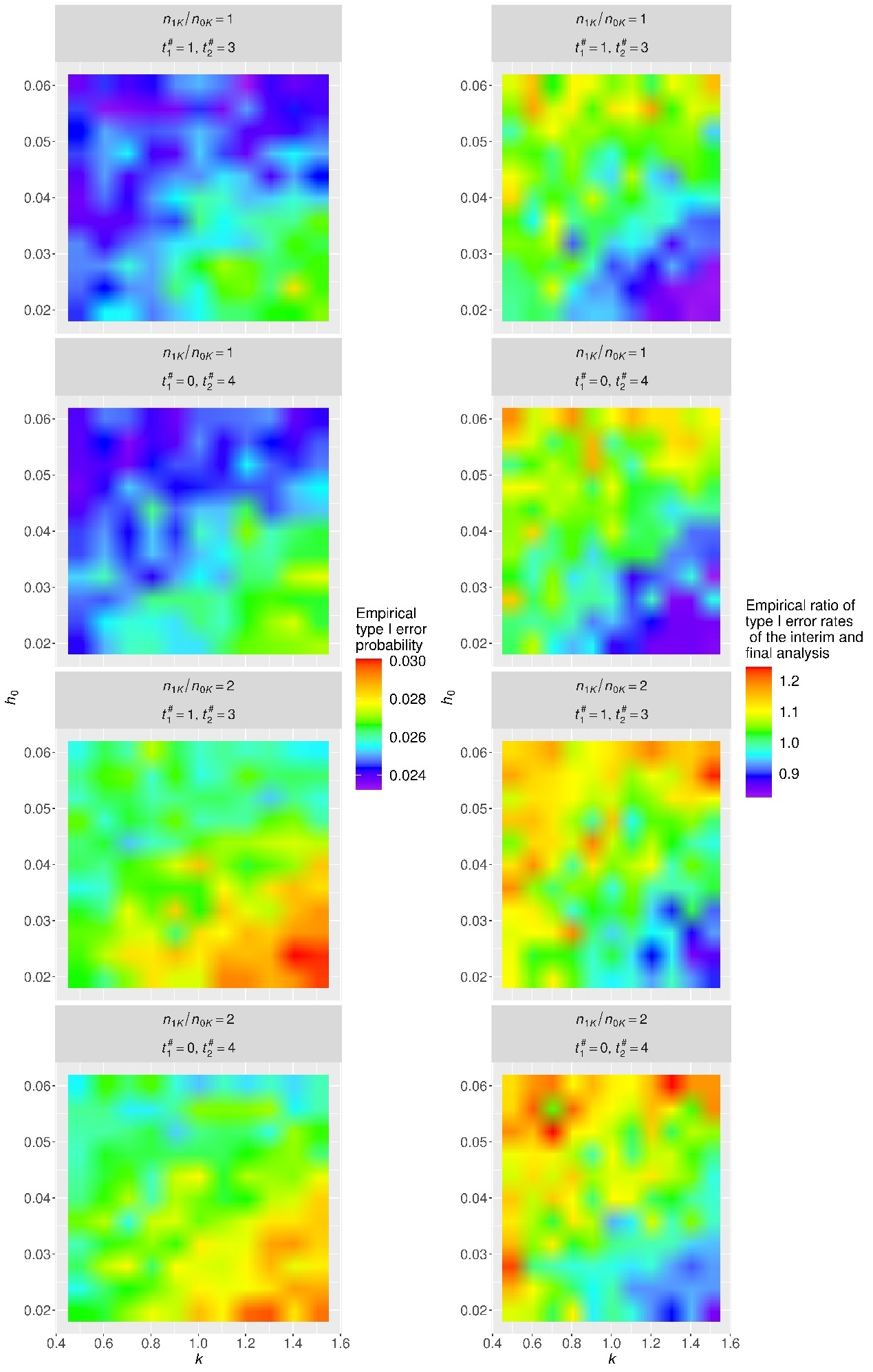
## C. Boundary validity against the mis-specified *S*0(*t*)

In this section, we evaluated the validity of the group sequential boundaries against the mis-specification of the survival function of the control group, namely *S*0(*t*), when the calendar-driven and event-driven approaches were applied, respectively. Concretely, we first determined the group sequential boundaries using the calendar-driven approach proposed in Section 3.2 under a given specification of *S*0(*t*) along with the numbers of events occurring by the end of every analysis. Then, we calculated the empirical type Ⅰ error rates under the weak null hypothesis using the Monte Carlo simulation procedures that used calendar-driven and event-driven approaches, respectively, when parameters *k* and *h*0 of *S*0(*t*) were mis-specified. Finally, the validity evaluation when *S*0(*t*) was mis-specified was conducted by comparing the empirical overall type Ⅰ error rates against the significance level and comparing the ratio of the stepwise type Ⅰ error rates in the interim and final analyses against the specified spending ratio of the significance levels.

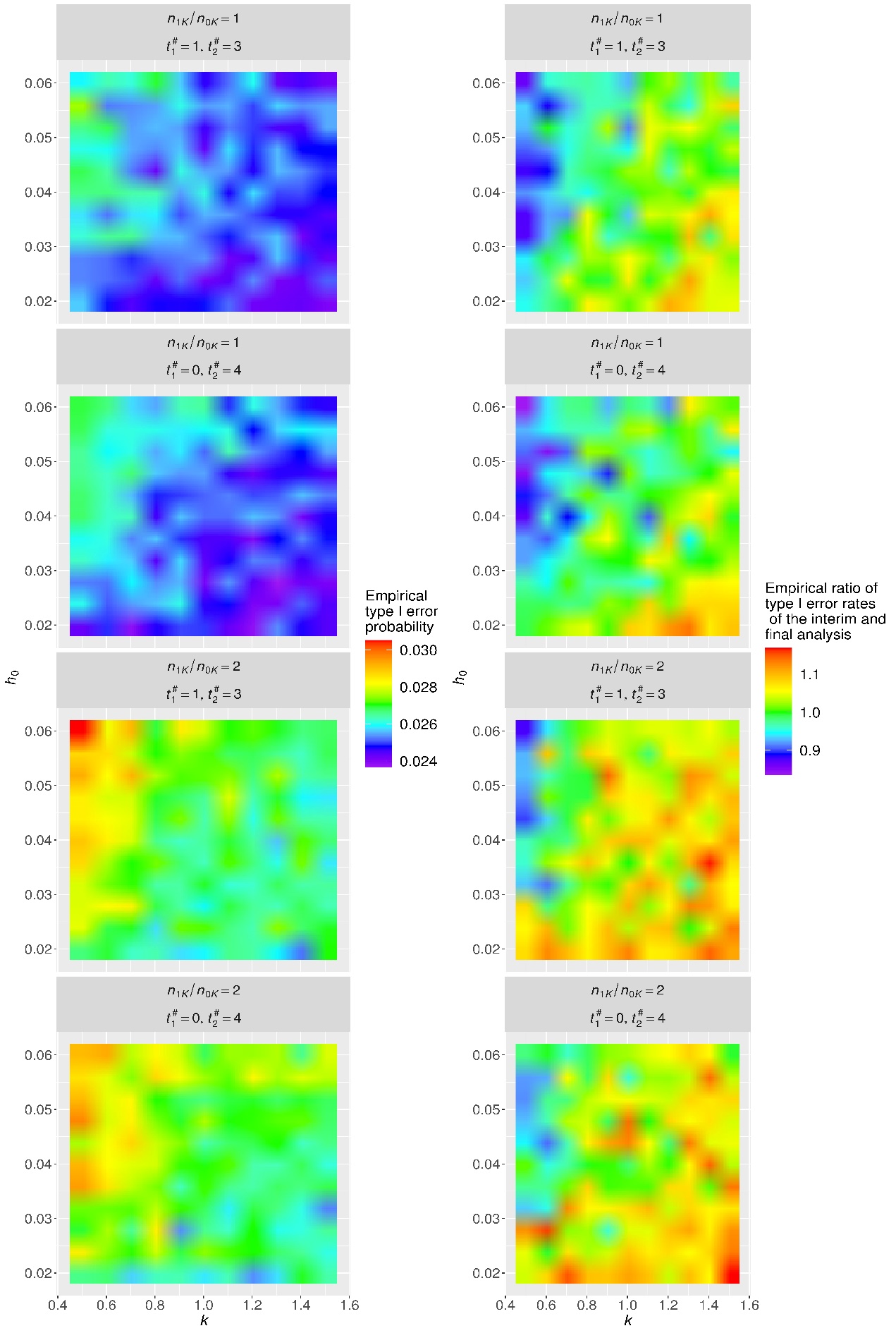
The related simulation scenarios were set as follows. The recruitment of patients was planned to last for 24 months, while the interim and ﬁnal analyses were scheduled at the 36th and 60 th months, respectively. The Markov chain parameters were set as and . The loss-to-follow-up rate during each month was speciﬁed as 0.74%. The signiﬁcance level, at 0.0125, was the same for the interim and ﬁnal analyses; thus, the overall signiﬁcance level for the one-sided test was 0.025. The speciﬁcations of in MERT were (1,3) and (0,4), respectively. The allocation ratios of the immunotherapy and control groups were speciﬁed as 1:1 and 2:1, respectively, whereas the maximum sample size was speciﬁed as 600. The survival function of the control group was specified by the parameters of and to determine the group sequential boundaries. Each determination of the group sequential boundaries corresponded to the combination of the allocation ratios of two groups and speciﬁcations of in MERT. Additionally, for each allocation ratio, we calculated the number of events which occurred by the end of the analyses with calendar time points of 36 and 60 months under the null hypothesis assuming that in which and in order to conduct an event-driven group sequential monitoring strategy. However, the true survival function of the control group was specified by the parameters of *k*=0.5, 0.6, ⋯, 1.4 or 1.5 and *h*0=0.02, 0.024, ⋯ ,0.056 or 0.06, respectively, in the Monte-Carlo simulation procedures. In the end, we used the Monte Carlo simulation procedure to calculate the empirical type Ⅰ error rates for each true specification of the survival function of the control group. And the number of repetitions was specified as 100,000. Additionally, the calendar-driven and event-driven methods were used in the Monte Carlo simulation procedures, respectively. The obtained empirical overall type Ⅰ error rates and ratio of the stepwise type Ⅰ error rates in the interim and final analyses were presented in Figure S7 and S8, respectively, which corresponded to the calendar--driven and event-driven approaches, respectively.



**Figure S6. The empirical type Ⅰ error rate versus the maximum sample size under the scenario that satisfies the strong null hypothesis**

****

**Figure S7: The overall value and allocation ratio between the interim and ﬁnal analyses of the empirical type I error rates versus the mis-specified survival function of the control group when the calendar-driven approach was adopted**

****

**Figure S8: The overall value and allocation ratio between the interim and ﬁnal analyses of the empirical type I error rates versus the mis-specified survival function of the control group when the event-driven approach was adopted**

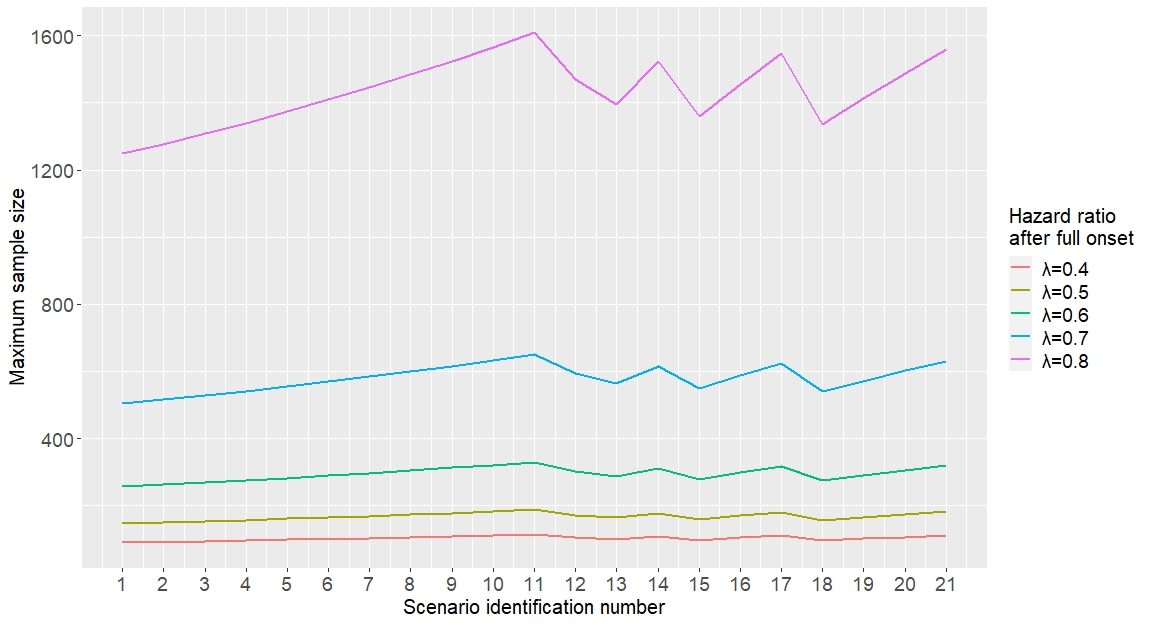
According to Figure S7, we have the following discoveries on the type Ⅰ error rates when the calendar-driven approach was adopted and *S*0(*t*) was mis-specified. First, the overall type Ⅰ error rates inflated and ratios of the stepwise type Ⅰ error rates in the interim and final analyses decreased when *k* was mis-specified larger and *h*0 was mis-specified smaller, corresponding to a small early-phase hazard and a large late-phase hazard. The behaviors were more obvious in the unbalanced design than those in the balanced design. On the contrary, the overall type Ⅰ error rates decreased slightly and the ratios of the stepwise type Ⅰ error rates in the interim and final analyses increased slightly when *k* was mis-specified smaller and *h*0 was mis-specified larger, corresponding to a large early-phase hazard and a small late-phase hazard.

According to Figure S8, we have the following discoveries on the type Ⅰ error rates when the event-driven approach was adopted and *S*0(*t*) was mis-specified. Firstly, the overall type Ⅰ error rates inflated and ratios of the stepwise type Ⅰ error rates in the interim and final analyses decreased when *k* was mis-specified smaller and *h*0 was mis-specified larger, corresponding to a large early-phase hazard and a small late-phase hazard. The behaviors were more obvious in the unbalanced design than those in the balanced design. On the contrary, the overall type Ⅰ error rates decreased slightly and the ratios of the stepwise type Ⅰ error rates in the interim and final analyses increased moderately when *k* was mis-specified larger and *h*0 was mis-specified smaller, corresponding to a small early-phase hazard and a large late-phase hazard.

## D. Comparison between MERT and OWLR

In this section, we carried out a comparison between the maximin efficient robust test (MERT) and the optimum weighted log-rank test (OWLR) mainly in terms of power. OWLR is the weighted log-rank test of which the weights are equal to the logarithm of the hazard ratio of the control and immunotherapy groups, according to the researches of Schoenfeld4 and Xu et al.5 Firstly, we estimated the maximum sample sizes required by the group sequential trials with OWLR under various scenarios. Then, we conducted the Monte Carlo simulation to obtain the empirical power of the group sequential trials with OWLR and MERT, respectively, under every scenario given the maximum sample size estimated before. Then, the comparison between the empirical power values of the group sequential trials with OWLR and MERT evaluate the power robustness of MERT. Additionally, we also calculated the empirical type Ⅰ error rates of the group sequential trials using MERT and OWLR as a supplementary reference.

The scenario settings were specified as follows. The recruitment was uniform and lasted for 24 months, whereas the interim and final analyses were arranged in the 36-th and 60-th months after the clinical trial started. Randomly censoring time was assumed to follow an exponential distribution, and the loss-to-follow-up rate during each month was 0.74%. The survival function of the control group was specified by *k*=1.1 and . The hazard ratio after the full onset of the immunotherapy λ was set to 0.4, 0.5, 0.6, 0.7 and 0.8, respectively. The true delayed treatment effect was modeled by the lag functions *l*(t) presented in Table S1 when comparing OWLR and MERT. Therefore, the weighted function of OWLR was specified as (referring to Equation (1) in the main text) where *l*(*t*) was specified referring to Table S1, and the weight functions of MERT were specified by parameters and . The randomization ratio of the immunotherapy and control groups is 1:1. The significance level for the considered single-side test hypothesis was 0.025. Then, given the nominal power of 0.9, we estimated the required maximum sample size of the group sequential trial using OWLR using the similar approach proposed in the main text under each specified scenario. The obtained maximum sample sizes under each scenario were shown in Figure S9. Given the estimated maximum sample size, we used the Monte Carlo simulation procedure to obtain the empirical power of the group sequential trials with OWLR and MERT. The obtained empirical power values under each scenario were presented in Figure S10. Then, we also used the Monte Carlo simulation procedure to calculate the empirical type Ⅰ error rates under the null hypothesis for each specified scenario and obtained maximum sample size. The obtained empirical type Ⅰ error rates under the null hypothesis were presented in Figure S11 for each specified scenario and obtained maximum sample size.

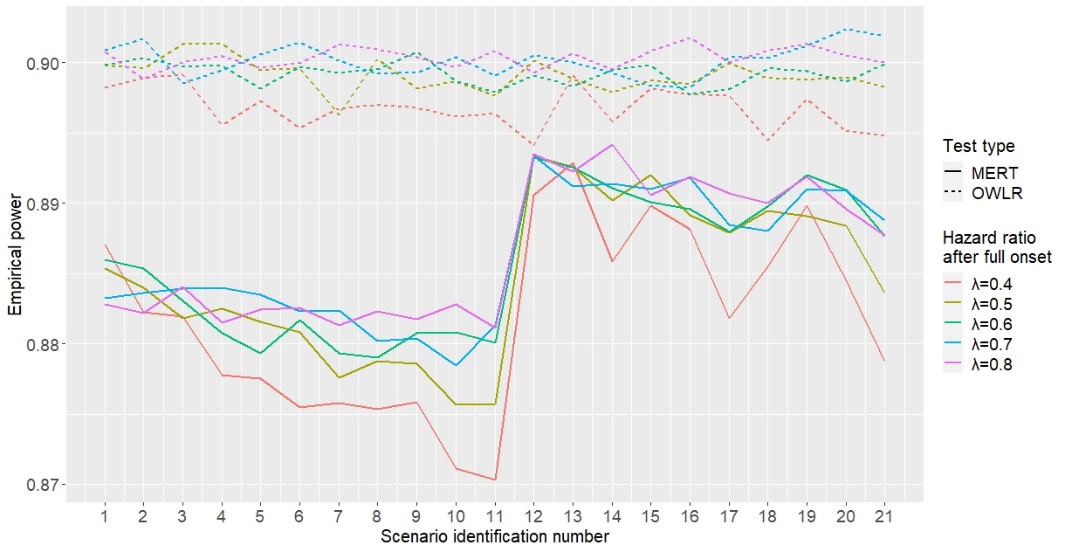


**Figure S9. The maximum sample size estimations for the group sequential trials using OWLR under each specified scenario setting.**

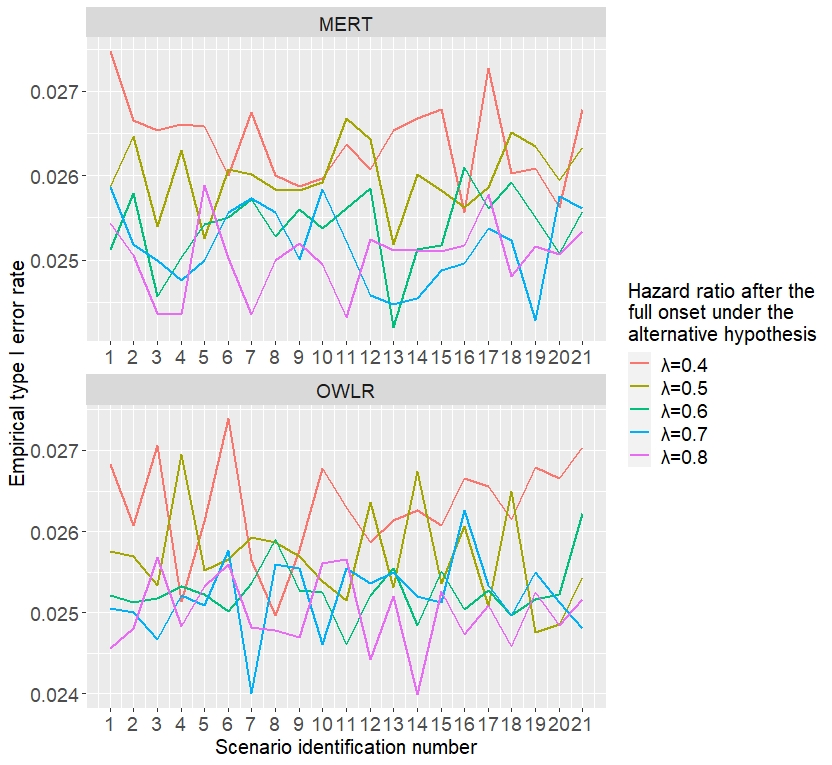
Referring to Figure S10, the empirical power of the group sequential trial using MERT was only slightly smaller than that of the group sequential trial using OWLR. Their difference was often in the range of 0.35% and 2.6%, which was really small. The finding further verified the power robustness of MERT. More interestingly, we discovered frequent, slightly larger losses of power of MERT against OWLR (1.1-2.6%) when threshold lags occurred (Scenarios 1-11 in Table S1) compared with the losses (0.35-1.6%) when the generalized lags occurred (Scenario 12-21 in S1). The generalized lag implies that the lag is a gradual process happening during one certain period. Additionally, we referred to Figure S11 and found that the empirical type Ⅰ error rates under the null hypothesis usually fluctuated around the specified significance level for the group sequential trials using MERT or OWLR. However, we also found a very slight inflation when the survival difference between two groups was large, corresponding to a small hazard ratio λ along with a small maximum sample size.

**Table S1: The lag functions *l*(*t*) under true delayed treatment effect scenarios comparing MERT and OWLR.**

|  |  |
| --- | --- |
| Scenario Identiﬁcation number | The speciﬁcation of *l*(*t*) |
| 1 |  |
| 2 | ) |
| 3 |  |
| 4 |  |
| 5 |  |
| 6 |  |
| 7 |  |
| 8 |  |
| 9 |  |
| 10 |  |
| 11 |  |
| 12 |  |
| 13 |  |
| 14 |  |
| 15 |  |
| 16 |  |
| 17 |  |
| 18 |  |
| 19 |  |
| 20 |  |
| 21 |  |

****

**Figure S10. The empirical power of the group sequential trials using MERT and OWLR under each specified scenario.**



**Figure S11. The empirical type Ⅰ error rates of the group sequential trials with MERT and OWLR corresponding each specified scenario.**

Additionally, we noticed that Ding an Wu6 made the similar power comparisons in the fixed sample trials between MERT and OWLR and put the relevant results in their Table 2. And their results also indicated the power of MERT was only slightly smaller than that of OWLR under the threshold, linear and generalized linear lag scenarios6. In addition, their results implied that the difference between the power of MERT and that of OWLR is moderately larger (up to 8%) when the lags occurred extremely later. But the power difference is very small in practice because the immunotherapy lags usually occur within six months, much earlier than the extreme cases assumed by Ding an Wu6.

## E. Robustness under random delay mechanism

In practice, a generalized delayed treatment effect mechanism may not be consistent with the reality, and some other delay mechanism might occur. For example, some researches often considered random delay mechanism which assumed that the delay was instantaneous for each patient and that delay time had the individual heterogeneity and should be treated as a random variable7. And the individual delay time was recommended to be assumed to follow the below distribution: 7,8. Thereby, we intended to evaluate the robustness of the proposed group sequential designs using MERT and a survival model of the generalized delayed treatment effect, proposed in Section 2 of the main text, under the random delay mechanism where individual delay time was assumed to follow the above distribution.

The considered simulation scenario settings were identical to those in the subsection 4.2 of the main text, except the specifications on the delayed treatment effect. Here, we assumed that the delay of the treatment effect was instantaneous for each patient and that the delay time was assumed to follow in which were set to (0,4) or (1,3 ) and (*a*, *b*) were set to (1,2), (1,1) and (2,1). Thereby, the specified lag scenarios corresponded to Scenarios 2,3 and 4 in Table 1 in the main text, respectively, given the values of . Thus, the specified lag scenarios were denoted by Scenario 2’, 3’ and 4’, respectively. Then, we directly adopted the maximum sample size estimations in the Scenarios 2, 3 and 4 in Table 1 in the main text, and then conducted the new Monte Carlo simulations to calculate the empirical power under each of the specified random delay scenarios with the other scenario settings kept the same with those in the subsection 4.2 of the main text. The obtained results were presented in Table S2. We were able to find that the maximum sample sizes previously estimated under the generalized delay treatment effect patterns, defined in the section 2 of the main text, were also accurate under the corresponding random delay scenarios. This finding implied that the generalized delayed treatment effect mechanism proposed in section 2 of the main text was compatible with the random delay mechanism that a treatment effect delay was instantaneous for each patient and that delay time was assumed to follow the distribution of . Thus, the survival model of the generalized delayed treatment effect (in Section 2) was also robust under the above random delay mechanism.

## F. Power influences from mis-specified lag scenarios

In this appendix, we investigated the influences of the misspecifications of the lag scenario on the power. Specifically, we adopted the maximum sample size estimations for the balanced design, which were presented in Table 2 of the main text, and calculated the empirical power under the true and mis-specified lag scenarios of which the lag functions were summarized in Table 1 of the main text with the other scenario settings kept identical to those specified in the subsection 4.2. Finally, the comparison between the empirical and nominal power can reflect the influences of misspecifications of the lag scenarios on the statistical power of MERT, which also further evaluated the power robustness of MERT from another perspective.

**Table S2. Accuracies of the maximum sample size estimations under various scenarios satisfying the random delay mechanism**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | *λ* |  |  | Identification number of the forms of lag function l(t) | | |
| Scenario 2’ | Scenario 3’ | Scenario 4’ |
| 1 | 0.4 | (0,4) | [0,4] | 94(0.89788) | 98(0.89918) | 101(0.89632) |
| 1 | 0.5 | (0,4) | [0,4] | 155(0.89986) | 160(0.89964) | 166(0.89872) |
| 1 | 0.6 | (0,4) | [0,4] | 271(0.89858) | 281(0.89987) | 290(0.90117) |
| 1 | 0.7 | (0,4) | [0,4] | 535(0.90023) | 553(0.89949) | 572(0.90129) |
| 1 | 0.8 | (0,4) | [0,4] | 1326(0.90066) | 1370(0.90051) | 1415(0.90047) |
| 2 | 0.4 | (0,4) | [0,4] | 109(0.91741) | 113(0.91986) | 117(0.91796) |
| 2 | 0.5 | (0,4) | [0,4] | 178(0.91582) | 185(0.91694) | 191(0.91627) |
| 2 | 0.6 | (0,4) | [0,4] | 312(0.91209) | 323(0.91263) | 334(0.91308) |
| 2 | 0.7 | (0,4) | [0,4] | 613(0.91018) | 633(0.90982) | 655(0.90933) |
| 2 | 0.8 | (0,4) | [0,4] | 1510(0.90696) | 1559(0.90654) | 1610(0.90572) |
| 1 | 0.4 | (0,4) | [1,3] | 96(0.89872) | 98(0.89981) | 99(0.89646) |
| 1 | 0.5 | (0,4) | [1,3] | 157(0.89875) | 160(0.89718) | 163(0.89995) |
| 1 | 0.6 | (0,4) | [1,3] | 276(0.90072) | 280(0.89918) | 285(0.89966) |
| 1 | 0.7 | (0,4) | [1,3] | 544(0.90052) | 553(0.89912) | 562(0.90102) |
| 1 | 0.8 | (0,4) | [1,3] | 1346(0.89935) | 1368(0.9) | 1391(0.90114) |
| 2 | 0.4 | (0,4) | [1,3] | 111(0.91862) | 113(0.91931) | 115(0.91714) |
| 2 | 0.5 | (0,4) | [1,3] | 181(0.91451) | 184(0.91321) | 188(0.91674) |
| 2 | 0.6 | (0,4) | [1,3] | 317(0.91221) | 322(0.91215) | 328(0.91341) |
| 2 | 0.7 | (0,4) | [1,3] | 622(0.91097) | 632(0.91107) | 643(0.91034) |
| 2 | 0.8 | (0,4) | [1,3] | 1532(0.90474) | 1557(0.90583) | 1583(0.90706) |
| 1 | 0.4 | (1,3) | [0,4] | 93(0.89827) | 97(0.89953) | 101(0.89919) |
| 1 | 0.5 | (1,3) | [0,4] | 153(0.89925) | 159(0.89913) | 165(0.90041) |
| 1 | 0.6 | (1,3) | [0,4] | 268(0.8995) | 278(0.9002) | 288(0.89823) |
| 1 | 0.7 | (1,3) | [0,4] | 529(0.90154) | 548(0.90155) | 567(0.89931) |
| 1 | 0.8 | (1,3) | [0,4] | 1312(0.89949) | 1356(0.89899) | 1402(0.90097) |
| 2 | 0.4 | (1,3) | [0,4] | 107(0.91824) | 112(0.91748) | 116(0.91784) |
| 2 | 0.5 | (1,3) | [0,4] | 176(0.91509) | 183(0.91574) | 190(0.91489) |
| 2 | 0.6 | (1,3) | [0,4] | 308(0.91188) | 319(0.91169) | 331(0.91196) |
| 2 | 0.7 | (1,3) | [0,4] | 606(0.90922) | 627(0.91) | 649(0.9099) |
| 2 | 0.8 | (1,3) | [0,4] | 1493(0.90584) | 1543(0.90736) | 1596(0.9073) |
| 1 | 0.4 | (1,3) | [1,3] | 94(0.89879) | 96(0.90075) | 97(0.89709) |
| 1 | 0.5 | (1,3) | [1,3] | 154(0.89878) | 157(0.90186) | 159(0.89769) |
| 1 | 0.6 | (1,3) | [1,3] | 270(0.89997) | 274(0.90165) | 279(0.90115) |
| 1 | 0.7 | (1,3) | [1,3] | 532(0.90003) | 541(0.89728) | 550(0.90135) |
| 1 | 0.8 | (1,3) | [1,3] | 1318(0.89785) | 1339(0.90152) | 1361(0.90095) |
| 2 | 0.4 | (1,3) | [1,3] | 108(0.91664) | 110(0.91721) | 112(0.91702) |
| 2 | 0.5 | (1,3) | [1,3] | 177(0.9151) | 180(0.91407) | 183(0.91558) |
| 2 | 0.6 | (1,3) | [1,3] | 310(0.91171) | 315(0.9135) | 320(0.91252) |
| 2 | 0.7 | (1,3) | [1,3] | 608(0.9093) | 619(0.91009) | 629(0.91053) |
| 2 | 0.8 | (1,3) | [1,3] | 1499(0.9055) | 1524(0.90494) | 1549(0.90639) |

The scenario settings were specified as below. The uniform recruitment of patients was planned to last for 24 months, and the recruited patients were equally randomized to the control and immunotherapy groups. The interim and ﬁnal analyses were scheduled at the 36th and 60th months, respectively. The Markov chain parameters were set as and . The loss-to-follow-up rate during each month was set as 0.74%. The signiﬁcance level, at 0.0125, was the same for the interim and ﬁnal analyses; thus, the overall signiﬁcance level for the one-sided test was 0.025. The nominal power was 0.9. The specifications of MERT, namely , were specified as (1,3) or (0,4). The survival function of the control group wassetbytheparametersofand.Thehazardratioof the immunotherapy and control groups after a full onset of effect (*λ*) was set to 0.4, 0.5, 0.6, 0.7 and 0.8, respectively. The range of possible delays, i.e., , was specified as [1, 3] and [0, 4], respectively, whereas the delayed treatment effect was modeled by the assumed lag functions *l*(*t*) (namely Scenarios 1-5) that were shown in Table 1 of the main text. For every assumed lag function, we estimated the maximum sample size, and we calculated the empirical power using the Monte-Carlo simulation procedure under the true, i.e., the assumed or mis-specified, lag scenarios of which the lag functions were also those presented in Table 1 of the main text. Additionally, the number of repetitions was 100,000 in the Monte-Carlo simulation procedure. The obtained results were presented in Table S3-S6, respectively, that corresponded to each of the following four settings: (a) ; (b) ; (c) ; (d) .

**Table S3. Evaluation on the influence from the mis-specification of lag scenarios on power ()**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Hazard ratio *λ* | Assumed lag scenario | Sample size | Empirical power under the true lag scenario | | | | |
| Scenario 1 | Scenario 2 | Scenario 3 | Scenario 4 | Scenario 5 |
| 0.4 | Scenario 1 | 88 | 0.89755 | 0.87664 | 0.86557 | 0.85164 | 0.82264 |
| 0.4 | Scenario 2 | 94 | 0.91663 | 0.89811 | 0.88617 | 0.87542 | 0.84592 |
| 0.4 | Scenario 3 | 98 | 0.92923 | 0.90896 | 0.89948 | 0.88893 | 0.86265 |
| 0.4 | Scenario 4 | 101 | 0.93375 | 0.91858 | 0.90857 | 0.89668 | 0.87192 |
| 0.4 | Scenario 5 | 109 | 0.94991 | 0.93503 | 0.92905 | 0.91834 | 0.89866 |
| 0.5 | Scenario 1 | 145 | 0.90117 | 0.87958 | 0.86921 | 0.85702 | 0.82782 |
| 0.5 | Scenario 2 | 155 | 0.91753 | 0.89951 | 0.88862 | 0.87744 | 0.85215 |
| 0.5 | Scenario 3 | 160 | 0.92589 | 0.91121 | 0.89958 | 0.88794 | 0.86363 |
| 0.5 | Scenario 4 | 166 | 0.93425 | 0.91782 | 0.91071 | 0.89909 | 0.87674 |
| 0.5 | Scenario 5 | 179 | 0.95026 | 0.93774 | 0.92875 | 0.92014 | 0.90053 |
| 0.6 | Scenario 1 | 255 | 0.89803 | 0.87955 | 0.87028 | 0.85619 | 0.83432 |
| 0.6 | Scenario 2 | 271 | 0.91732 | 0.89865 | 0.88979 | 0.87917 | 0.85238 |
| 0.6 | Scenario 3 | 281 | 0.92559 | 0.90958 | 0.90053 | 0.89158 | 0.86758 |
| 0.6 | Scenario 4 | 290 | 0.9342 | 0.919 | 0.90738 | 0.89888 | 0.87765 |
| 0.6 | Scenario 5 | 312 | 0.94912 | 0.93593 | 0.92763 | 0.91833 | 0.89939 |
| 0.7 | Scenario 1 | 504 | 0.90105 | 0.88296 | 0.87064 | 0.86135 | 0.83562 |
| 0.7 | Scenario 2 | 535 | 0.91649 | 0.90068 | 0.89126 | 0.88221 | 0.85682 |
| 0.7 | Scenario 3 | 553 | 0.92536 | 0.91211 | 0.90037 | 0.89207 | 0.86811 |
| 0.7 | Scenario 4 | 572 | 0.93374 | 0.91732 | 0.90918 | 0.9029 | 0.87937 |
| 0.7 | Scenario 5 | 614 | 0.94903 | 0.93454 | 0.92772 | 0.9194 | 0.90076 |
| 0.8 | Scenario 1 | 1251 | 0.90048 | 0.88238 | 0.87234 | 0.86191 | 0.83606 |
| 0.8 | Scenario 2 | 1326 | 0.91362 | 0.90102 | 0.89085 | 0.881 | 0.85879 |
| 0.8 | Scenario 3 | 1370 | 0.92502 | 0.90965 | 0.90047 | 0.89249 | 0.86885 |
| 0.8 | Scenario 4 | 1415 | 0.93175 | 0.91852 | 0.91017 | 0.89858 | 0.8789 |
| 0.8 | Scenario 5 | 1515 | 0.94603 | 0.9333 | 0.92662 | 0.918 | 0.90052 |

**Table S4. Evaluation on the influence from the mis-specification of lag scenarios on power ()**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Hazard ratio *λ* | Assumed lag scenario | Sample size | Empirical power under the true lag scenario | | | | |
| Scenario 1 | Scenario 2 | Scenario 3 | Scenario 4 | Scenario 5 |
| 0.4 | Scenario 1 | 93 | 0.89952 | 0.88916 | 0.88196 | 0.87813 | 0.86442 |
| 0.4 | Scenario 2 | 96 | 0.91033 | 0.897 | 0.89204 | 0.88748 | 0.87472 |
| 0.4 | Scenario 3 | 98 | 0.91568 | 0.9035 | 0.89898 | 0.89356 | 0.88082 |
| 0.4 | Scenario 4 | 99 | 0.91577 | 0.90621 | 0.90294 | 0.89788 | 0.88503 |
| 0.4 | Scenario 5 | 103 | 0.92694 | 0.91847 | 0.91326 | 0.90761 | 0.89895 |
| 0.5 | Scenario 1 | 152 | 0.90181 | 0.88943 | 0.88341 | 0.87716 | 0.86547 |
| 0.5 | Scenario 2 | 157 | 0.91024 | 0.90079 | 0.89283 | 0.88739 | 0.87724 |
| 0.5 | Scenario 3 | 160 | 0.91213 | 0.90533 | 0.89993 | 0.89367 | 0.88352 |
| 0.5 | Scenario 4 | 163 | 0.92046 | 0.9117 | 0.90486 | 0.90027 | 0.88774 |
| 0.5 | Scenario 5 | 169 | 0.92803 | 0.91965 | 0.91478 | 0.91113 | 0.89807 |
| 0.6 | Scenario 1 | 267 | 0.90044 | 0.89244 | 0.88591 | 0.87884 | 0.86791 |
| 0.6 | Scenario 2 | 276 | 0.91037 | 0.89957 | 0.89557 | 0.88884 | 0.87883 |
| 0.6 | Scenario 3 | 280 | 0.91399 | 0.9038 | 0.90007 | 0.89502 | 0.88316 |
| 0.6 | Scenario 4 | 285 | 0.91838 | 0.90865 | 0.90564 | 0.89997 | 0.88933 |
| 0.6 | Scenario 5 | 295 | 0.92764 | 0.91874 | 0.913 | 0.90749 | 0.89944 |
| 0.7 | Scenario 1 | 526 | 0.90062 | 0.89034 | 0.88707 | 0.87982 | 0.86987 |
| 0.7 | Scenario 2 | 544 | 0.90925 | 0.90161 | 0.89525 | 0.89045 | 0.87969 |
| 0.7 | Scenario 3 | 553 | 0.9143 | 0.90589 | 0.90109 | 0.8961 | 0.88528 |
| 0.7 | Scenario 4 | 562 | 0.91723 | 0.91018 | 0.90559 | 0.90151 | 0.89235 |
| 0.7 | Scenario 5 | 582 | 0.92671 | 0.91976 | 0.91346 | 0.90952 | 0.90061 |
| 0.8 | Scenario 1 | 1304 | 0.89895 | 0.89236 | 0.88607 | 0.8814 | 0.87001 |
| 0.8 | Scenario 2 | 1346 | 0.90867 | 0.89957 | 0.89597 | 0.8904 | 0.8804 |
| 0.8 | Scenario 3 | 1368 | 0.91224 | 0.90493 | 0.90109 | 0.89667 | 0.88504 |
| 0.8 | Scenario 4 | 1391 | 0.91902 | 0.91006 | 0.90625 | 0.8999 | 0.89028 |
| 0.8 | Scenario 5 | 1438 | 0.92554 | 0.91982 | 0.91336 | 0.90919 | 0.89979 |

According to Tables S3-S6, we concluded that the deviation of empirical power against the nominal power was relatively small (-8.5%~5.6%) when a lag scenario specified in Table 1 of the main text was mis-specified by another lag function in same table. Additionally, we summarized the following patterns on this deviation when the lag function is mis-specified:

1. The deviation of the empirical power against the nominal power decreases slightly with an increase of the hazard ratio of the immunotherapy and control groups after a full onset of effect (*λ*).
2. Mis-specifications of the lag functions *l*(*t*) have more obvious impacts on the deviations of the empirical power against the nominal power when the true range of possible delays is wider.
3. Given the true range of possible delays, namely , the deviations of the empirical power against the nominal power are more obvious when is included in () compared with the corresponding cases when is equal to if the mis-specified or assumed lag function is ; otherwise, the deviations are comparable.
4. Given the true range of possible delays, namely , the deviations of the empirical power against the nominal power are comparable when is included in compared with the corresponding cases when is equal to .

**Table S5. Evaluation on the influence from the mis-specification of lag scenarios on power ()**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Hazard ratio *λ* | Assumed lag scenario | Sample size | Empirical power under the true lag scenario | | | | |
| Scenario 1 | Scenario 2 | Scenario 3 | Scenario 4 | Scenario 5 |
| 0.4 | Scenario 1 | 91 | 0.89914 | 0.88824 | 0.88328 | 0.875 | 0.862 |
| 0.4 | Scenario 2 | 94 | 0.90991 | 0.89889 | 0.8915 | 0.88578 | 0.87482 |
| 0.4 | Scenario 3 | 96 | 0.91541 | 0.90575 | 0.90028 | 0.89364 | 0.88315 |
| 0.4 | Scenario 4 | 97 | 0.91915 | 0.90753 | 0.90316 | 0.8967 | 0.88622 |
| 0.4 | Scenario 5 | 101 | 0.928 | 0.91913 | 0.91415 | 0.90884 | 0.8975 |
| 0.5 | Scenario 1 | 149 | 0.90142 | 0.89058 | 0.88389 | 0.87894 | 0.86578 |
| 0.5 | Scenario 2 | 154 | 0.91018 | 0.90092 | 0.89282 | 0.89135 | 0.87689 |
| 0.5 | Scenario 3 | 157 | 0.91648 | 0.9048 | 0.90075 | 0.89568 | 0.88343 |
| 0.5 | Scenario 4 | 159 | 0.91938 | 0.90956 | 0.90548 | 0.89928 | 0.88659 |
| 0.5 | Scenario 5 | 165 | 0.9255 | 0.91952 | 0.91353 | 0.90903 | 0.89908 |
| 0.6 | Scenario 1 | 261 | 0.90145 | 0.88955 | 0.88451 | 0.88023 | 0.86926 |
| 0.6 | Scenario 2 | 270 | 0.9101 | 0.90136 | 0.89354 | 0.88903 | 0.87966 |
| 0.6 | Scenario 3 | 274 | 0.91379 | 0.90377 | 0.89808 | 0.89468 | 0.88226 |
| 0.6 | Scenario 4 | 279 | 0.9177 | 0.90994 | 0.90553 | 0.89752 | 0.88995 |
| 0.6 | Scenario 5 | 289 | 0.92687 | 0.92137 | 0.91501 | 0.90996 | 0.89972 |
| 0.7 | Scenario 1 | 515 | 0.89926 | 0.89108 | 0.88636 | 0.88189 | 0.86896 |
| 0.7 | Scenario 2 | 532 | 0.90813 | 0.89969 | 0.89441 | 0.89076 | 0.87907 |
| 0.7 | Scenario 3 | 541 | 0.91518 | 0.90523 | 0.90214 | 0.89316 | 0.88511 |
| 0.7 | Scenario 4 | 550 | 0.91902 | 0.90852 | 0.90406 | 0.90014 | 0.89074 |
| 0.7 | Scenario 5 | 569 | 0.92621 | 0.91675 | 0.91309 | 0.90983 | 0.89875 |
| 0.8 | Scenario 1 | 1277 | 0.90211 | 0.89088 | 0.88532 | 0.88087 | 0.87181 |
| 0.8 | Scenario 2 | 1318 | 0.90903 | 0.89967 | 0.89589 | 0.88886 | 0.88064 |
| 0.8 | Scenario 3 | 1339 | 0.91296 | 0.90356 | 0.89995 | 0.89548 | 0.88516 |
| 0.8 | Scenario 4 | 1361 | 0.91798 | 0.90825 | 0.90554 | 0.90026 | 0.89034 |
| 0.8 | Scenario 5 | 1408 | 0.92579 | 0.9179 | 0.9152 | 0.90849 | 0.90094 |

**Table S6. Evaluation on the influence from the mis-specification of lag scenarios on power ()**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Hazard ratio *λ* | Assumed lag scenario | Sample size | Empirical power under the true lag scenario | | | | |
| Scenario 1 | Scenario 2 | Scenario 3 | Scenario 4 | Scenario 5 |
| 0.4 | Scenario 1 | 90 | 0.89973 | 0.88792 | 0.87318 | 0.86231 | 0.81535 |
| 0.4 | Scenario 2 | 93 | 0.9106 | 0.89966 | 0.88526 | 0.87408 | 0.82873 |
| 0.4 | Scenario 3 | 97 | 0.92078 | 0.91092 | 0.89966 | 0.88609 | 0.8445 |
| 0.4 | Scenario 4 | 101 | 0.93129 | 0.92063 | 0.91181 | 0.8999 | 0.85949 |
| 0.4 | Scenario 5 | 114 | 0.95595 | 0.94899 | 0.94044 | 0.93178 | 0.89734 |
| 0.5 | Scenario 1 | 148 | 0.89986 | 0.89005 | 0.8772 | 0.86451 | 0.82174 |
| 0.5 | Scenario 2 | 153 | 0.90951 | 0.89949 | 0.88805 | 0.87821 | 0.83552 |
| 0.5 | Scenario 3 | 159 | 0.91927 | 0.91166 | 0.89871 | 0.8896 | 0.85122 |
| 0.5 | Scenario 4 | 165 | 0.93056 | 0.92029 | 0.90928 | 0.9018 | 0.86133 |
| 0.5 | Scenario 5 | 185 | 0.95378 | 0.94651 | 0.94015 | 0.92971 | 0.89928 |
| 0.6 | Scenario 1 | 259 | 0.90148 | 0.88837 | 0.88057 | 0.86605 | 0.82576 |
| 0.6 | Scenario 2 | 268 | 0.91108 | 0.90234 | 0.8886 | 0.87768 | 0.83558 |
| 0.6 | Scenario 3 | 278 | 0.91855 | 0.91132 | 0.90239 | 0.88976 | 0.8528 |
| 0.6 | Scenario 4 | 288 | 0.92691 | 0.91979 | 0.90955 | 0.89818 | 0.86536 |
| 0.6 | Scenario 5 | 322 | 0.95102 | 0.94546 | 0.93746 | 0.92922 | 0.89768 |
| 0.7 | Scenario 1 | 513 | 0.90067 | 0.89144 | 0.88039 | 0.86876 | 0.82968 |
| 0.7 | Scenario 2 | 529 | 0.90918 | 0.9005 | 0.88891 | 0.87965 | 0.84222 |
| 0.7 | Scenario 3 | 548 | 0.91913 | 0.91057 | 0.90242 | 0.89027 | 0.85395 |
| 0.7 | Scenario 4 | 567 | 0.92621 | 0.91862 | 0.91134 | 0.90071 | 0.86556 |
| 0.7 | Scenario 5 | 633 | 0.95051 | 0.9445 | 0.93653 | 0.92961 | 0.90049 |
| 0.8 | Scenario 1 | 1272 | 0.89996 | 0.89244 | 0.87917 | 0.86767 | 0.83254 |
| 0.8 | Scenario 2 | 1312 | 0.90999 | 0.90134 | 0.89076 | 0.8814 | 0.84293 |
| 0.8 | Scenario 3 | 1356 | 0.91879 | 0.9088 | 0.89973 | 0.891 | 0.85537 |
| 0.8 | Scenario 4 | 1402 | 0.92657 | 0.9187 | 0.91291 | 0.90027 | 0.86591 |
| 0.8 | Scenario 5 | 1561 | 0.94797 | 0.94415 | 0.93597 | 0.9301 | 0.90028 |

**Reference**

1. Sugimoto T, Hamasaki T, Evans SR, Halabi S. Group-sequential logrank methods for trial designs using bivariate non-competing event-time outcomes. *Lifetime Data Analysis*. Apr 2020;26(2):266-291. doi:10.1007/s10985-019-09470-4

2. Li B, Su L, Ye Y, Yan F. M&M: A maximum duration design with the Maxcombo test for a group sequential trial of an immunotherapy with a random delayed treatment effect. *Statistics in Medicine*. 2022-02-20 2022;41(4):815-830. doi:<https://doi.org/10.1002/sim.9251>

3. Phadnis MA, Mayo MS. Group sequential design for time-to-event data using the concept of proportional time. *Stat Methods Med Res*. Jul 2020;29(7):1867-1890. doi:10.1177/0962280219876313

4. Schoenfeld DA. Sample-Size Formula for the Proportional-Hazards Regression Model. *Biometrics*. 1983;39(2):499-503.

5. Xu Z, Zhen B, Park Y, Zhu B. Designing therapeutic cancer vaccine trials with delayed treatment effect. *Statistics in Medicine*. Feb 2017;36(4):592-605. doi:10.1002/sim.7157

6. Ding X, Wu J. Designing cancer immunotherapy trials with delayed treatment effect using maximin efficiency robust statistics. *Pharmaceutical Statistics*. Jul 2020;19(4):424-435. doi:10.1002/pst.2003

7. Xu Z, Park Y, Zhen B, Zhu B. Designing cancer immunotherapy trials with random treatment time-lag effect. *Statistics in Medicine*. Dec 2018;37(30):4589-4609. doi:10.1002/sim.7937

8. Li B, Su L, Gao J, Jiang L, Yan F. A group sequential design and sample size estimation for an immunotherapy trial with a delayed treatment effect. *Stat Methods Med Res*. 2021;30(3):904-915. doi:10.1177/0962280220980780