Published in final edited form as:

Head Neck. 2019 January ; 41(1): 37–45. doi:10.1002/hed.25359.

Randomized phase II/III confirmatory treatment selection design with a change of survival endpoints – statistical design of RTOG 1216

Qiang Zhang, MD, PhD^{1,2}, Qian Wu, PhD³, Paul M. Harari, MD⁴, and David I. Rosenthal, MD⁵

¹Thomas Jefferson University, Sidney Kimmel Medical College, 840 Walnut Street, Suite 1536, Philadelphia, PA, 19107, U.S.A

²Department of Biostatistics & Epidemiology, University of Pennsylvania, 423 Guardian Drive, Philadelphia, PA 19104, U.S.A

³Clinical Biostatistics, Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, WA, 98109, U.S.A

⁴University of Wisconsin School of Medicine and Public Health, Madison, WI, 53792, U.S.A

⁵The University of Texas MD Anderson Cancer Center, Houston, TX, 77030, U.S.A

Abstract

Background—To confirm the treatment effects of concurrent Cetuximab plus Docetaxel observed in RTOG 0234 and single out the effect of Cetuximab, we designed RTOG 1216, a randomized phase II/III study, which uses an intermediate endpoint to select the best regimen for definitive testing of survival benefit.

Methods—In phase II the best regimen should demonstrate statistically significant efficacy against the control with predefined advantage over the competing arm regarding disease free survival. We evaluate operating characteristics of the randomized II/III group sequential design through simulations and numerical integrations under the null and various alternative hypotheses.

Results—Results show the randomized II/III design yields substantial savings on sample size and time with well-controlled type I and type II error rates.

Conclusions—Overall, the proposed randomized II/III design has desirable properties that offer cost effectiveness, operational efficiency and most importantly, scientific innovation that can be considered for similar clinical research settings.

Keywords

head and neck; oncology; randomized II/III design; survival; treatment selection

Request addressed to: Qiang Zhang, Ed.Zhang.Jr@gmail.com, Phone: 215-928-3178.

Conflict of interests: The authors declare that they have no competing interests.

Introduction

Approximately 50% of head and neck cancer patients undergo primary surgery in the treatment of their malignancy. For patients with advanced local/regional disease with high-risk features, recurrence rates after surgery alone are high and therefore, postoperative treatment strategies have been actively investigated for several decades. RTOG 0234 enrolled 238 high-risk patients with head and neck cancer in a randomized phase II trial to examine the feasibility and safety of delivering postoperative radiation combined with Cetuximab plus either weekly Cisplatin or Docetaxel chemotherapy. When disease-free survival (DFS) was compared to historical control (the RTOG 9501 chemoradiation arm) the hazard ratio (HR) was 0.76 for Cetuximab-Cisplatin vs. control, while it was 0.69 for Cetuximab-Docetaxel vs. control ($p=0.01$)¹. Therefore, it is important to confirm the treatment effects for the Docetaxel arm in a randomized trial with the same control arm, and single out the effect of Cetuximab with an efficient development plan.

A randomized II/III design offers as a good option for this study because a third concurrent Docetaxel only arm is included and the arm with the best efficacy, regarding DFS if demonstrated, will be chosen to test for the survival benefit in a timely manner. Randomized phase II/III designs, including studies with multiple experimental arms, have been used in selected oncology trials (Korn, et al.²). Advantages of this approach include savings on sample size and overall duration of the treatment evaluation process³. In many cases, the same endpoint is used for treatment selection in phase II and comparison of efficacy in phase III⁴⁻⁷. In practice, phase II decisions are often based on the intermediate endpoints like response and progression-free survival (PFS). Thus endpoints in phase II and III components are likely different in terms of types of measures or definitions. For example, Korn et al.² described the study design for CALGB-80802, which was assessing if doxorubicin plus sorafenib is superior to sorafenib alone in advanced liver cancer. The phase II component involved a comparison of PFS based on the first 170 patients (it targeted a PFS HR of 0.67 with 90% power). The phase III component targets an overall survival (OS) HR of 0.73 with 480 patients.

To date, Hunsberger et al.³, Todd and Stallard⁸, and Royston et al.⁹ proposed phase II/III designs with a change of these survival endpoints from phase II to III. None of these can be applied directly to RTOG 1216. For example, Hunsberger et al.³ explored two-arm designs without treatment selection using PFS and OS in each phase, and suggested to design the phase II study separately based on its own type I error rate and power. To avoid accruing excess patients to a potentially negative phase II trial, they also recommended accrual suspension during the follow-up in phase II. Todd and Stallard⁸ considered the case with more than two arms. They derived one stopping boundary for treatment selection and interim comparisons¹⁰. The arm with the largest treatment effect on an early endpoint is selected at the first interim analysis without confirming statistical significance. However, statistical significance is a requirement for standalone phase II trials to go to phase III. Royston et al.⁹ studied two-stage only procedures using PFS and OS. More than one arm can be chosen from stage one without distinguishing their efficacy. Interim efficacy or futility analyses are not considered for either stage. There are also additional technical issues to consider in the design of RTOG 1216. Bauer and Posch¹¹ pointed out that since the two

outcomes from the same patient are likely to be dependent ¹², thus the type I error rate for the II/III design could be inflated. Moreover, when comparing more than one experimental arm to the control group, we need to adjust for the multiple comparisons in phase II in addition to controlling for the overall error rate for the II/III design. We can use the Dunnett test ¹³ and the Bonferroni method ¹⁴, etc., to protect the error rates in phase II.

RTOG 1216 is designed to answer study specific questions, thus it is novel in several aspects: treatment selection is performed in a manner similar to those in usual phase II trials designed with predefined error rate and power and allows only one arm to be carried to the next phase. The Go and No-Go decision is made based on 1). Statistically significant efficacy results on an early endpoint - DFS; and 2). If both arms are statistically significant, we expect better treatment effects over the competing arm for the winner to continue to phase III. We build in separate interim efficacy and futility monitoring rules for each phase.

Methods

In many randomized phase II oncology trials, we use DFS or PFS as the primary endpoint. This is typically defined as time to disease progression or death. Then for the phase III comparison, we consider overall survival, which includes death due to any cause. In general, there are multiple candidate therapies like the setting in this article. The decision regarding experimental arm selection during phase II is made primarily using results from DFS. For example, in phase II we would like to compare the treatment effects between each of the new regimens (arm 2: RT + Chemotherapy B, vs. arm 3: RT + Chemotherapy B + agent) and the standard arm (arm 1: RT + Chemotherapy A) in terms of DFS and select the arm with better efficacy for the phase III testing. For RTOG 1216, Chemotherapy A is Cisplatin, Chemotherapy B is Docetaxel, and the agent is Cetuximab (more details are in the results section).

Regarding the multiplicity issue in the phase II component, we considered a comparison specific type I error rate and we evaluate the overall alpha for all comparisons through either simulations or theoretical calculations. Group sequential methodologies (Armitage et al. ¹⁵; O'Brien and Fleming ¹⁶; Lan and DeMets ¹⁷) are implemented for protecting error rates when there is interim monitoring. These two strategies together will protect the total error rate for the II/III design.

For the proposed design, the type I error rates are α_{II} (e.g., 10%) for each comparison in phase II and α_{III} (e.g., 2.5%) for the comparison in phase III and the study power is $1 - \beta$ (e.g., 80%) in each phase. We calculate the sample size for each component using typical group sequential methods. For efficacy monitoring in both phases, we could consider an O'Brien-Fleming type alpha spending function ¹⁶. This is one of the most used stopping boundaries in clinical trial literature. The futility boundaries are derived using the rho family of spending functions ¹⁷ or the LIB20 method as in Freidlin et al. and Zhang et al. ^{18, 19}, respectively. Therefore, during the study if the interim results are overwhelmingly positive (crossed the upper efficacy boundary) or ineffective (crossed the lower futility boundary), the data monitoring committee may suggest stopping the trial early based on these results and the overall evidence. Details of the interim test statistics are in appendix A. Standard

software such as EAST²⁰ can help with designing the phase II and III trials as built in procedures. Under a proportional hazard assumption, let $\theta_1^{PFS}, \theta_2^{PFS}$ denote the log of hazard ratios between each of the treatment arms and control. In phase II the overall null hypothesis is $H_0^{II}: \theta_1^{PFS} = \theta_2^{PFS} = 0$ (the expected hazard ratios are 1.0 for both comparisons) and the one-sided alternative hypothesis is $H_1^{II}: \theta_2^{PFS} \geq \theta_1^{PFS} + \delta > 0$ (iv). δ is the increase in effect size due to Cetuximab (e.g., the expected hazard ratios are 0.65 and 0.47 for arms 1 and 2.). The power of the study is the probability that arm 2 is selected and the null hypothesis is rejected in favor of a positive treatment difference (θ_2^{OS}) under the alternative hypothesis⁸.

To evaluate the design performance we also consider other configurations such as (i) $\theta_1^{PFS} > 0, \theta_2^{PFS} = 0$ (only arm 2 is effective, combination 3 in Tables 2, 3); (ii) $\theta_1^{PFS} = 0, \theta_2^{PFS} > 0$ (only arm 3 is effective, combination 5 in Tables 2, 3); and (iii) $\theta_2^{PFS} \leq \theta_1^{PFS} + \delta > 0$ (both arms are effective, but the difference is less than, e.g. 5%, combinations 6, 7 in Tables 2, 3). A decision rule based on each of these scenarios is formulated as shown in Table 1 and we investigate the design properties for each case. In comparison to the method of Stallard and Todd⁶, an early endpoint - DFS is used here for treatment selection. One distinction from the method of Todd and Stallard⁸ is that one or both of the experimental arms must improve DFS statistically significantly over the control. If both experimental arms show DFS improvement over control, the winning arm must ultimately demonstrate a 5% advantage over its counterpart. The 5% is chosen to reflect physician's expectation of minimal clinically meaningful difference. Otherwise, an arm is preferred due to less overall toxicity and cost, etc. Thus being the most effective arm with only minimal improvement over the competing options does not satisfy the selection criteria we use in RTOG 1216. Of course, we also need to consider quality of life (QOL) and toxicities for arm selection as explained in the discussion section.

Once an arm is chosen in phase II, we can resume accrual and continue to the phase III. Patients in the selected arm from phase II will be included in the phase III analysis. The phase III null and alternative hypotheses are $H_0^{III}: \theta_{winner}^{OS} = 0$ and $H_1^{III}: \theta_{winner}^{OS} > 0$ (e.g., the expected hazard ratio is 1.0 vs. 0.7). Here, we broaden the usual definitions of error rate and power in standalone phase II, III trials to account for the hybrid nature of the randomized II/III design. According to the phase II decision rules and the phase III hypothesis, the overall error rate is the probability of an arm is chosen using DFS in phase II and then the null hypothesis is rejected again based on OS in phase III. Detailed definitions and expressions of these error rates are in appendix B.

There are two methods to evaluate the performance of the design: numerical integrations and simulations. We will show how to control the type I error rates using the first method in each phase and in the overall II/III design based on theoretical results for the design of RTOG 1216. We first conducted Monte Carlo simulations to investigate the operating characteristics of this type of design and estimate the likely correlations among test statistics in each phase and the change of the overall error rates $\alpha_{II/III}$ from the unadjusted, typical

phase III α_{III} of 0.025 and 0.05. We use the statistical software R²¹ for all simulations and the CRAN library cubature²² for the numerical integrations.

Results

To see the overall error rates for the randomized II/III design relative to the phase III levels of 0.025 and 0.05, we have conducted simulations with 10^4 runs using parameters as specified in appendix C. In Tables 2–3, we present the probabilities of making a Go decision at the phase II final analysis, and the observed type I error rates/powers for the whole trial as defined in (1) in the appendix. For each combination of treatment effects (e.g., no difference in either arm, only one arm is effective, or both arms are effective, etc., as in Table 1.), the upper/lower rows are results for 80%/90% power and each cell shows rejection probabilities for the phase II and the overall II/III trial. As can be seen in Table 2, we have included eight combinations of hazard ratios, each representing different treatment effects on DFS and OS by arm. The correlations are for between provisional DFS (or progression) and OS based on the bivariate exponential distributions.

Under the global null hypothesis (combination 1), there is 24% probability to continue to phase III if the correlation is 0.5, it is similar in the same rows in Table 3. The probability is 91%/96% (combination 6) under each power when the two experimental arms are equally effective. With this design, we have an overall error rate of 0.02/0.02 (Table 2) and 0.042/0.037 (Table 3) as comparing to the phase III design levels of 0.025 and 0.05 for each power. Notice that when the correlation is 1.0, the overall error rates are 0.0234/0.0274 under the null in Table 2, the latter is slightly inflated numerically. These overall error rates are 0.054/0.055 for the same cases in Table 3. This is less of a concern since the correlation between the two endpoints is usually less than 1.0. Results are similar with 10^5 simulations. Therefore, the error rates are well controlled for the phase II and the II/III design.

When the two experimental arms are equally effective (combination 6), the overall study power would be 78–79%/90% with the same correlation (0.5) in each table. So when only arm 3 is effective (combination 3), 79%/88% of the trials will continue to the phase III. Under the alternative hypothesis (combination 8) and each design power, the observed study power for the overall trial is 95/98% in both Tables 2 and 3. The statistical power for other combinations of treatment effects are in tables 2 and 3. These show the performance of the design if only DFS is improved (combinations 2 and 4), only one arm is statistically significant for both endpoints (combinations 3 and 5) or if the treatment is slightly less effective for OS in arm 2 (combination 7). Overall, the statistical power is satisfactory under the alternative hypothesis. Please note that the power for each of the comparisons in phase II and phase III alone is 80% per the separate design.

In phase II, when the correlation decreases from 1 to 0.25 between the two endpoints, the correlations between the test statistics are also reduced from 0.8 to 0.5 on average within each arm and 0.4 to 0.25 on average between the two experimental arms in Tables 2 and 3 under the null and each power. This builds the basis for the numerical integrations and simulations in the design of RTOG 1216. In addition, when the correlation decreases, the overall error rates and the statistical power become lower in general but remain satisfactory

for each table. For completeness, supplemental Tables 1 and 2 show the percentage of rejections for each of the decision rules in phase II and the overall II/III trial for combinations 1 and 6.

RTOG 1216 is a randomized phase II/III trial of surgery and postoperative radiation delivered with concurrent Cisplatin versus Docetaxel versus Docetaxel and Cetuximab for high-risk squamous cell cancer of the head and neck¹. The goal of the phase II component is to select the better experimental arm to improve DFS over the control arm of Cisplatin plus radiation. In addition, this design aims to single out the effect of Cetuximab. When the design concept was reviewed, we evaluated the operating characteristics through simulations using asymptotic joint distributions among the log-rank tests of DFS and OS. This is different from the two methods we consider in this paper. Here, we directly simulated survival times based on bivariate exponential distributions for provisional DFS and OS and carried out log-rank tests at each stage as described in the methods section. The schema and detailed design parameters for RTOG 1216 are provided in Appendix D¹. The correlations among the test statistics for numerical integrations are based on the estimated averages over all simulations unlike the assumed values in the protocol. Under the null in phase II, for a correlation of 0.5 and estimated test correlations of about 0.45 and 0.23 between DFS, OS within each arm and with the other arm, the probability of rejecting the null hypothesis is 9% and 9.7%. If only arm 2 or 3 is statistically significant, it is 4.4% and 1.5% for scenarios 4 and 5. When the two experimental arms are equally effective, these are 13.6%, 13.8%, 44.7% and 16.9% based on 10^4 simulations. When only arm 3 is statistically significant for DFS, the probabilities are 0.9%, 62%, 4.2% and 9.1%. The overall error rate is 0.027 from these simulations. Under the alternative hypothesis, the power is 94%, 80% and 86% in phase II, III and II/III. We can also successfully design these II/III trials and maintain error rates for phase II, III and the overall study directly through numerical integrations without the complicated simulations, details can be found in Appendix B.

According to (3) and (4) in the appendix the expected sample sizes for RTOG 1216 are 254 and 449 under each hypothesis. The expected trial durations are 4.5 and 7 years, respectively. However, if we run the three arm phase II then a separate phase III, it will take at least 12 years (if the transition from II to III takes about 3 years) and 588 patients. If we run two phase II studies sequentially, then the phase III trial, the total study duration will be at least 18 years with sample size of 768 patients.

Discussion

The proposed RTOG 1216 design yields savings on sample size and time with well-controlled type I and type II error rates. Our design considers an early endpoint and it has clearly defined phase II and phase III type I, II error rates that are similar to those in separate trials. In addition, many of the recently proposed designs would pick the arm with the largest observed treatment effect regardless of the actual difference relative to other competing arms. This selection method does not guarantee it is statistically significantly better than the control. These designs cannot answer the questions we posed in RTOG 1216, in which we would like to know if both experimental arms are better than the Cisplatin arm, and if so, is Cetuximab providing any additional benefit. Unlike the two stage procedures, according to

our design, we are able to monitor for early efficacy and futility in each component. In addition, the phase II portion allows sufficient follow up with or without accrual suspension for time to event data and thus protects statistical power for the phase II efficacy testing.

This design is flexible and can certainly be extended to cases when there are multiple agents to be added to the same or different backbone regimens one at the time or through other combinations, for example, having increasing number of agents. We can develop decision rules for selecting one or multiple arms with design treatment differences among these candidate arms and against the control. The estimation of the correlation between PFS and OS in clinical trials is critical for these designs. A recent proposal in Li and Zhang (2015) considered the more flexible Weibull distribution ²³.

As mentioned in the methods section, quality of life and toxicity outcomes are also important in selecting the best arm ¹. In RTOG 1216, we will also assess grade 3–5 side effects and QOL among the three arms and identify if there is a statistically significant difference. If the Cetuximab arm is not at least 5% better than arm 2 when both are statistically significant and the toxicities are similar, we will consider QOL differences. Otherwise, if arm 3 is more toxic than arm 2 regarding grade 3–5 side effects including mucositis, dysphagia, dermatitis, etc., then we will choose arm 2 as the winner. For scenario 5 in Table 1, if arm 3 also causes substantially increased toxicities, we will consider very carefully (including QOL) in selecting the preferred arm for phase III testing.

One potential drawback of the proposed design is that the overall planning time could be longer for these randomized II/III designs, but for the case of RTOG 1216, from concept review to study pre-activation, it took 299 days meeting the NCI OEWG deadline for single-phase III trials. With more emerging choices for early efficacy, such as tumor size as measured by imaging, we can also utilize different endpoints (continuous, binary) in the phase II component. A combination chemotherapy regimen consisting of oxaliplatin, irinotecan, fluorouracil, and leucovorin (FOLFIRINOX) versus gemcitabine for metastatic pancreatic cancer.^{2, 24} This trial considered response rate among the first 44 patient on the experimental arm and both the response rate and the overall survival benefits were confirmed in each phase. Overall, the proposed randomized II/III design has desirable properties which offer cost effectiveness, operational efficiency and most importantly, scientific innovation that may prove valuable to consider for similar clinical research settings.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding source:

The project was supported by RTOG grant U10 CA21661 and CCOP grant U10 CA37422 from the NCI.

The first author would like to thank Dr. Robert Gray for his support and communication regarding approximating the correlations between the test of an early binary endpoint and overall survival under the asymptotic joint

distribution in an earlier version of ECOG 1912. The first author also appreciates input from Dr. Boris Freidlin from NCI for the helpful discussions at the design stage when he worked at RTOG and during revision of the paper.

References

1. Randomized phase II/III trial of surgery and postoperative radiation delivered with concurrent Cisplatin versus Docetaxel versus Docetaxel and Cetuximab for high-risk squamous cell cancer of the Head and Neck, RTOG 1216 protocol.
2. Korn EL, Freidlin B, Abrams JS, Halabi S. Design Issues in Randomized Phase II/III Trials. *Journal of Clinical Oncology*. 2012; 30:667–671. [PubMed: 22271475]
3. Hunsberger S, Zhao Y, Simon R. A comparison of phase II study strategies. *Clinical Cancer Research*. 2009; 15:5950–5955. [PubMed: 19789306]
4. Thall PF, Simon R, Ellenberg SS. Two-stage selection and testing designs for comparative clinical trials. *Biometrika*. 1988; 75(2):303–310.
5. Schaid DJ, Wieand S, Therneau TM. Optimal two stage screening designs for survival comparisons. *Biometrika*. 1990; 77:507–13.
6. Stallard N, Todd S. Sequential designs for phase III clinical trials incorporating treatment selection. *Statistics in Medicine*. 2003; 22:689–703. [PubMed: 12587100]
7. Stallard N, Friede T. A group-sequential design for clinical trials with treatment selection. *Statistics in Medicine*. 2008; 27:6209–6227. [PubMed: 18792085]
8. Jennison, C, Turnbull, BW. Group sequential methods with applications to clinical trials. Chapman & Hall/CRC; 2000.
9. Todd S, Stallard N. A new clinical trial design combining phases II and III: Sequential designs with treatment selection and a change of endpoint. *Drug Information Journal*. 2005; 39:109–18.
10. Royston P, Mahesh KB, Qian W. Novel designs for multi-arm clinical trials with survival outcomes with an application in ovarian cancer. *Statistics in Medicine*. 2003; 22:2239–2256. [PubMed: 12854091]
11. Bauer P, Posch M. Letter to the editor: modification of the sample size and the schedule of interim analyses in survival trials based on data inspections. *Statistics in Medicine*. 2004; 23:1333–1335. [PubMed: 15083486]
12. Redman MW, Goldman BH, LeBlanc M, Schott A, Baker LH. Modeling the relationship between progression-free survival and overall survival: the phase II/III trial. *Clinical Cancer Research*. 2013; 19:2646–2656. [PubMed: 23669424]
13. Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*. 1955; 50:1096–1121.
14. Dunn OJ. Multiple comparisons among means. *Journal of the American Statistical Association*. 1961; 56:52–64.
15. Armitage P, McPherson CK, Rowe BC. Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society, A*. 1969; 132:235–44.
16. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics*. 1979; 35:459–56.
17. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika*. 1983; 70(3): 659–63.
18. Gail, MH, DeMets, DL, Slud, EV. Simulation studies on increments of the two-sample log-rank score test for survival time data with application to group sequential boundaries. In: Crowley, J, Johnson, RA, editors. *Survival analysis*. Hayward, CA: Institute for Mathematical Statistics; 1982. 287–301.
19. Freidlin B, Korn KL, Gray R. A general inefficacy interim monitoring rule for randomized clinical trials. *Clinical Trials*. 2010; 7:197–208. [PubMed: 20423925]
20. Zhang Q, Freidlin B, Korn KL, Halabi S, Mandrekas S, Dignam J. Comparison of futility monitoring guidelines using completed phase III oncology trials. *Clinical Trials*. 2017; 14(1):48–58. [PubMed: 27590208]
21. East® 5 Manual, A.2.4, Spending function boundaries. Cytel Inc; Cambridge, MA: 2007.

22. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2017. URL <https://www.R-project.org/>
23. Narasimhan, B; Johnson, SG. cubature: Adaptive Multivariate Integration over Hypercubes. 2016. R package version 1.3–6. <https://CRAN.R-project.org/package=cubature>
24. Li Y, Zhang Q. A Weibull multi-state model for the dependence of progression-free survival and overall survival. *Statistics in Medicine*. 2015; 34:2497–2513. [PubMed: 25865438]
25. Conroy T, Desseigne F, Ychou M, et al. FOLFIRINOX versus gemcitabine for metastatic pancreatic cancer. *N Engl J Med*. 2011; 364:1817–1825. [PubMed: 21561347]

Appendix

A

The distributions of the interim statistics are as follows: Let θ denote the measure of the difference of efficacy between one of the experimental arms and the control. In oncology trials, this is usually log-hazard ratios (for the time to event endpoints like DFS and OS). We want to test the null hypothesis $H_0: \theta = 0$ vs. the one-sided alternative $\theta > 0$, indicating the experimental arm is superior to the control. We plan to conduct a total of J analyses and let $\hat{\theta}_j$ be the estimate of θ based on a log-rank test at analysis j . So the interim Z statistics have an asymptotic multivariate joint normal distribution²⁵: $\hat{\theta}_j \sim N(\theta\sqrt{\frac{d_j}{4}}, 1)$, $j=1, \dots, J$ with variance covariance matrix of: $cov(\hat{\theta}_j, \hat{\theta}_{j'}) = var(\hat{\theta}_{j'}) = \frac{d_{j'}}{4}$, $1 \leq j \leq j' \leq J$, where d_j is the number of events observed at analysis j . When compared with the upper (efficacy) and lower (futility) critical values, $u_1^{II}, \dots, u_{J_{II}}^{II}$ and $l_1^{II}, \dots, l_{J_{II}}^{II}$ in phase II and $u_1^{III}, \dots, u_{J_{III}}^{III}$ and $l_1^{III}, \dots, l_{J_{III}}^{III}$ in phase III, if $Z_j > u_j$ then we stop and reject the null hypothesis. If however, $Z_j < l_j$ we recommend stop and reject the alternative hypothesis. Note that we are monitoring one of the two endpoints alone in each phase, so the stopping boundaries can be derived using comparison specific error rates.

B

As discussed earlier that the standardized log-rank test statistics (normalized treatment difference): $Y_{DFS(2 \text{ vs. } 1)}$, $Y_{DFS(3 \text{ vs. } 1)}$ in phase II and $Z_{OS(2 \text{ vs. } 1)}$, $Z_{OS(3 \text{ vs. } 1)}$ in phase III are correlated, so the joint distribution g_1 of, for example, $Y_{DFS(2 \text{ vs. } 1)}$, $Y_{DFS(3 \text{ vs. } 1)}$ and $Z_{OS(2 \text{ vs. } 1)}$ are asymptotically three dimensional normal. In phase II, the total error rate accounting for multiple comparisons is:

$$\begin{aligned}
\alpha_{II\text{total}} = & \int_{-\infty}^{c_{II}} \int_{c_{II}}^{\infty} f_1 dy_{DFS(2\text{ vs. }1)} dy_{DFS(3\text{ vs. }1)} + \int_{c_{II}}^{\infty} \int_{-\infty}^{c_{II}} f_2 dy_{DFS(2\text{ vs. }1)} dy_{DFS(3\text{ vs. }1)} \quad (1) \\
& + \int_{c_{DFS_3 - DFS_2 \geq 0.05}}^{\infty} \int_{c_{II}}^{\infty} f_2 dy_{DFS(2\text{ vs. }1)} dy_{DFS(3\text{ vs. }1)} \\
& + \int_{c_{II}}^{c_{DFS_3 - DFS_2 < 0.05}} \int_{c_{II}}^{\infty} f_1 dy_{DFS(2\text{ vs. }1)} dy_{DFS(3\text{ vs. }1)},
\end{aligned}$$

Then the total error rate of the II/III design according to the phase II, III final analyses is:

$$\begin{aligned}
\alpha_{II/III} = & \int_{c_{III}}^{\infty} \int_{-\infty}^{c_{II}} \int_{c_{II}}^{\infty} g_1 dy_{DFS(2\text{ vs. }1)} dy_{DFS(3\text{ vs. }1)} dz_{OS(2\text{ vs. }1)} \quad (2) \\
& + \int_{c_{III}}^{\infty} \int_{c_{II}}^{\infty} \int_{-\infty}^{c_{II}} g_2 dy_{DFS(2\text{ vs. }1)} dy_{DFS(3\text{ vs. }1)} dz_{OS(3\text{ vs. }1)} \\
& + \int_{c_{III}}^{\infty} \int_{c_{DFS_3 - DFS_2 < 0.05}}^{\infty} \int_{c_{11}}^{\infty} g_2 dy_{DFS(2\text{ vs. }1)} dy_{DFS(3\text{ vs. }1)} dz_{OS(2\text{ vs. }1)} \\
& + \int_{c_{III}}^{\infty} \int_{c_{II}}^{c_{DFS_3 - DFS_2 \geq 0.05}} \int_{c_{II}}^{\infty} g_1 dy_{DFS(2\text{ vs. }1)} dy_{DFS(3\text{ vs. }1)} dz_{OS(3\text{ vs. }1)},
\end{aligned}$$

in which, c_{II} and c_{III} are the phase II and III critical values. $c_{DFS_3 - DFS_2 \geq 0.05}$ is the critical value derived using the difference of DFS between two experimental arms based on the asymptotic results in appendix A. Improvement in DFS (5%) can be translated to a further reduction of failure rate. The corresponding hazard ratio leads to the higher rejection value. f_1 , f_2 and g_1 , g_2 are the asymptotic joint multivariate normal distributions. For example, the mean and covariance for g_1 are $(\theta_{DFS(2\text{ vs. }1)}, \theta_{DFS(3\text{ vs. }1)}, \theta_{OS(2\text{ vs. }1)})'$ and

$$\begin{pmatrix} \sigma_{DFS(2\text{ vs. }1)}^2 & \sigma_{DFS\text{ tests}} & \sigma_{DFS\text{ OS tests}} \\ \sigma_{DFS\text{ tests}} & \sigma_{DFS(3\text{ vs. }1)}^2 & \sigma_{DFS\text{ OS tests}} \\ \sigma_{OS\text{ DFS tests}} & \sigma_{OS\text{ DFS tests}} & \sigma_{OS(2\text{ vs. }1)}^2 \end{pmatrix}.$$

Numerical integrations of the four multiple integrals

quantifies the decision rules under the global null and specific alternative hypotheses. To do this we either set the expected error rate $\alpha_{II/III}$ for the randomized II/III trial then derive the phase III critical value/alpha (α_{III}) or use the preset phase III critical value/alpha to obtain the overall error rates through numerical integrations using the theoretical results in (1) and (2) above. For example in phase II (supplemental Figure 1, Panel 1), when $\alpha_{II\text{total}} = 0.15$, the probabilities for choosing arm 2 is 12% (include 2.7% for scenario 4 in table 1) and 10%

(include 0.8% for scenario 5 in table 1) for arm 3. Panel 2 in supplemental Figure 1 shows the phase III final alphas if we expect the overall error rates $\alpha_{II/III} = 0.01, 0.025$ and 0.05 for the whole study with given design correlations among test statistics. For example, we need a phase III final alpha of 0.035 to obtain an overall error rate of 0.025 ($\rho=0.6$ between DFS and OS tests in phase II). When the correlation increases, the phase III design alphas decrease. We can see that the overall alpha is less than the design level in phase III for these cases. On the other hand, we can also tell how much the error rate changes given a phase III design type I error rate α_{III} (Panel 3). If we want to control the conditional probabilities of rejecting the null hypothesis given that any arm is chosen to go to the phase III (9.6% from simulations) $\alpha_{II/III} / \alpha_{II/III}^{total}$, we can find the phase III design alpha and overall error rate using similar algorithms (Panel 4). For a design similar to that of RTOG 1216 with correlations of 0.4 and 0.2 among test statistics within and between arms at final analyses, the overall adjusted error rate after three interim analyses for the II/III design is 0.034 using numerical integration. Alternatively, we can verify these results through simulations with various parameter combinations as in Tables 2 and 3.

The expected sample sizes and trial time of the phase II, III component are calculated as in a typical group sequential design. However, the expected sample size for the II/III design is³:

$$E(N) = 3n_1 + (n_2 - 2n_1)P(y_{DFS(2 \text{ vs. } 1)}^{winner}) + (n_2 - 2n_1)P(y_{DFS(3 \text{ vs. } 1)}^{winner}), \quad (3)$$

where n_1 is the sample size for one arm in phase II and n_2 is the total sample size for phase III.

The expected trial time is:

$$E(T) = t_1 + f_1 + (t_2 - t_1 - f_1 + f_2)P(y_{DFS(2 \text{ vs. } 1)}^{winner}) + (t_2 - t_1 - f_1 + f_2)P(y_{DFS(3 \text{ vs. } 1)}^{winner}), \quad (4)$$

where t_1, f_1 are the accrual and follow up time in phase II, t_2, f_2 are the accrual and follow up times for phase III. $P(y_{DFS(2 \text{ vs. } 1)}^{winner})$ and $P(y_{DFS(3 \text{ vs. } 1)}^{winner})$ are calculated from (1) above. These calculations will be illustrated in the design of RTOG 1216.

C

For all simulations, provisional DFS (or progression) and OS are generated from bivariate exponential distributions with correlation parameters $\rho = (0, 0.25, 0.5, 0.75, 0.85 \text{ and } 1)$. The actual DFS time was set as minimum of (t_{DFS}, t_{OS}) as in Hunsberger et al.³ based on semi-competing risk model. We derive the total number of events and patients separately for each phase using the expected hazard ratios, type I and type II error rates. These and the trial duration are calculated by considering interim analyses at each stage. Annual accrual rate is 180 patients/year. A patient is censored if the survival time is less than the trial duration in each phase similar to that in RTOG 0234. Hazard ratios are 1.0 under the null and 0.65 for DFS (control failure rate: 0.35) and 0.70 for OS (control failure rate: 0.20) under the

alternative hypotheses. For illustration purpose, we will investigate design properties with type I error rates of 0.025, 0.05 and power of 80% and 90%. Supplemental Tables 1 and 2 show the percentage of rejections for each of the decision rules in phase II and the overall II/III trial for combinations 1 and 6. For example in phase II, under the null (supplemental Table 1) and correlation of 0.5, the probability that only one arm is better than the control is about 18–19%, this is similar in both tables. Under the null for scenario 4 in Table 1, the probability is 3.7%/4.4% under each power, but it is 1.3%/1% in each table according to scenario 5 for DFS at 1.5 years. Therefore, we see that when there is no treatment effect, the false positive rate is more likely due to one of the arms being false positive. When the two experimental arms are equally effective (supplemental Table 2), for scenarios 2, 3 together, the probability is about 25%/17%. For scenario 4, the probability is 51%/63%, but it is 14%/17% in each table according to scenario 5. For this case, it is more likely arm 2 will be selected if both are equally effective. However under the alternative hypothesis (not shown in supplemental Tables 1 or 2), these are 0.8%, 21.8%, 32% and 44% with 80% power. So the probability that arm 3 is selected is much higher now.

For the overall II/III trial with each power, the probabilities for scenarios 2 & 3, 4, 5 are 1.6%/1.53%, 0.31%/0.36% and 0.10%/0.08% in supplemental Table 1 and these are 3.31%/2.85%, 0.63%/0.68% and 0.21%/0.16% in supplemental Table 2 under the null. When the two experimental arms are equally effective, the probabilities for scenarios 2 & 3, 4, 5 are 22.04%/15.72%, 44.14%/58.16% and 11.87%/15.76% in supplemental Table 1 and these are 22.29%/15.79%, 44.47%/58.40% and 11.96%/15.83% in supplemental Table 2. The total for each set of these, for example in supplemental table 2, equals to 0.0415/0.0369 and 0.79/0.90, which are the overall error rates and power in Table 3 (combinations 1 and 6). Under the alternative hypothesis, these are 0.78%, 20.9%, 30.7% and 42.2% with 80% power. The patterns regarding change of probability for picking the winner for each scenario from the null to the alternative hypothesis are similar to those in phase II as described earlier.

D: RTOG 1216

Randomized Phase II/III Trial of Surgery and Postoperative Radiation Delivered with Concurrent Cisplatin versus Docetaxel versus Docetaxel and Cetuximab for High-Risk Squamous Cell Cancer of the Head and Neck

SCHEMA (2/20/14)

			Zubrod Performance		
S			Status	S	
T	For all	S	1. 0	T	Arm 1: IMRT 60 Gy in 6 weeks and
E	patients:	T	2. 1	E	cisplatin 40 mg/m ² weekly × 6 doses
P	Mandatory	R		P	
I	submission	A	Primary Tumor Site	2	
	of tissue for	T	1. Oral Cavity		Arm 2: IMRT 60 Gy in 6 weeks and
R	EGFR	I	2. Larynx	R	weekly docetaxel (15 mg/m ²)

			Zubrod Performance		
E		F	3. Hypopharynx	A	x 6 doses
G	For	Y	4. p16-negative	N	
I	oropharyngeal		oropharynx	D	Arm 3: IMRT 60 Gy in 6 weeks and
S	Cancer			O	Cetuximab (loading 400 mg/m ² , then
T	patients:		EGFR Expression	M	250 mg/m ² weekly × 6 doses)
E	Mandatory		1. High	I	and docetaxel (15 mg/m ²) weekly
R	p16 analysis		2. Low	Z	x 6 doses
			3. Inevaluable	E	

NOTE: If the trial proceeds to the phase III component, Arm 2 or Arm 3 will be chosen as the experimental arm. Patients accrued in the phase II component of the trial will complete the treatment to which they are randomized (Arm 1, 2, or 3) and will be followed as specified in the protocol.

In RTOG 1216, we assume the 3 year DFS is 35% based on RTOG 0234, with one-sided type I error rate of 15% and 80% power and a hazard ratio of 0.6, we need 60 analyzable patients per arm in the phase II study. The accrual will take 2.2 years and total study duration of 3.5 years. We conduct one interim futility analysis when there is 50% of the required information in phase II. We will suspend the study accrual during the follow up phase and compare the arms using DFS, while considering effects on quality of life, toxicities, etc. Table 1 shows the decision rule for selecting a better arm, in particular, if both arms are statistically significant and the Cetuximab arm is at least 5% better in terms of DFS, then it will go to the phase III. The dropped arm will not be included in the phase III final analysis. For the design of phase III, the 120 patients from phase II will be included. We assume that the 3 year overall survival is 45% from data of RTOG 9501 and 0234. As mentioned in the introduction, since DFS and OS are correlated the overall error rate could be inflated due to the phase II arm selection and comparison. So we used one-sided phase III error rate of 0.033 (personal communications, Boris Freidlin, PhD, NCI)⁴ to prevent potential inflation over 0.05. For this component, a group sequential design with three interim analyses and a final analysis based on O'Brien-Fleming^{16, 17} and LIB20^{18, 19} boundaries are used. With 80% power and hazard ratio of 0.67, we need 408 analyzable patients. Final analysis will be conducted in 7.2 years. This study was approved by site IRBs and all participants signed the written informed consent.

Table 1

Phase II decision rule

Scenarios	Arm 2 vs. Arm 1	Arm 3 vs. Arm 1	Decision
1. H_0^{II}	Not significant	Not significant	Stop trial; report phase II results
2. (i)	Not significant	Significant	Arm 3 to phase III
3. (ii)	Significant	Not significant	Arm 2 to phase III
4. (iii)	Significant	Significant, but < 5% better than Arm 2	Arm 2 to phase III
5. H_1^{II} (iv)	Significant	Significant, but 5% better than Arm 2	Arm 3 to phase III

Table 2

Phase II / II/III error rates and power for various global hypotheses - phase III design alpha 2.5% and 80%/90% power

True HRs for DFS			True HRs for OS		Correlation between DFS and OS					
Arm 2/1	Arm 3/1		Arm 2/1	Arm 3/1	1	0.85	0.75	0.5	0.25	0
1	1		1	1	0.2434/0.0234	0.2401/0.0236	0.2465/0.0232	0.2436/0.0202	0.2458/0.0150	0.2463/0.0118
					0.2431/0.0274	0.2445/0.0257	0.2510/0.0263	0.2360/0.0198	0.2423/0.0140	0.2405/0.0124
1	0.65		1	1	0.8088/0.0300	0.7467/0.0342	0.6823/0.0260	0.6019/0.0263	0.5688/0.0252	0.5542/0.0191
					0.9064/0.0287	0.8518/0.0333	0.7905/0.0280	0.6963/0.0286	0.6643/0.0243	0.6506/0.0232
1	0.65		1	0.7	0.8075/0.6537	0.8056/0.6489	0.8002/0.6483	0.7850/0.6182	0.7756/0.6081	0.7746/0.5828
					0.9033/0.7978	0.8994/0.7932	0.9009/0.7956	0.8823/0.7652	0.8783/0.7628	0.8720/0.7480
0.65	1		1	1	0.7997/0.0250	0.7348/0.0262	0.6739/0.0253	0.6056/0.0241	0.5778/0.0230	0.5515/0.0208
					0.8953/0.0235	0.8481/0.0251	0.7846/0.0243	0.7044/0.0268	0.6692/0.0234	0.6455/0.0198
0.65	1		0.7	1	0.8148/0.6990	0.7996/0.6827	0.7939/0.6740	0.7815/0.6552	0.7747/0.6321	0.7722/0.6218
					0.9016/0.8377	0.9006/0.8308	0.8918/0.8229	0.8851/0.8129	0.8783/0.7945	0.8706/0.7830
0.65	0.65		0.7	0.7	0.9147/0.8021	0.9167/0.8024	0.9151/0.7958	0.9052/0.7805	0.8977/0.7603	0.9011/0.7504
					0.9660/0.9058	0.9674/0.9055	0.9698/0.9083	0.9635/0.8964	0.9576/0.8835	0.9579/0.8739
0.65	0.65		0.75	0.7	0.9230/0.6980	0.9163/0.6984	0.9057/0.6915	0.8929/0.6758	0.8975/0.6684	0.8880/0.6394
					0.9693/0.8082	0.9704/0.8141	0.9664/0.8144	0.9532/0.8045	0.9542/0.7970	0.9489/0.7731
0.65	0.47		0.7	0.53	0.9934/0.9547	0.9913/0.9575	0.9900/0.9506	0.9866/0.9452	0.9888/0.9459	0.9878/0.9414
					0.9997/0.9782	0.9995/0.9822	0.9990/0.9795	0.9987/0.9773	0.9990/0.9789	0.9990/0.9860

Table 3

Phase II / II/III error rates and power for various global hypotheses - phase III design alpha 5% and 80%/90% power

True HRs for DFS			True HRs for OS		Correlation between DFS and OS					
Arm 2/1	Arm 3/1	Arm 2/1	Arm 3/1	1	0.85	0.75	0.5	0.25	0	
1	1	1	1	0.2434/0.0535	0.2400/0.0507	0.2461/0.0468	0.2428/0.0415	0.2456/0.0298	0.2471/0.0247	
				0.2431/0.0545	0.2445/0.0528	0.2510/0.0461	0.2360/0.0369	0.2423/0.0295	0.2405/0.0258	
1	0.65	1	1	0.8080/0.0592	0.7477/0.0653	0.6816/0.0543	0.6005/0.0508	0.5688/0.0512	0.5551/0.0390	
				0.9064/0.0618	0.8518/0.0630	0.7905/0.0558	0.6963/0.0535	0.6643/0.0515	0.6506/0.0459	
1	0.65	1	0.7	0.8074/0.6698	0.8051/0.6628	0.8006/0.6646	0.7850/0.6297	0.7762/0.6095	0.7738/0.5897	
				0.9033/0.8083	0.8994/0.8025	0.9009/0.8069	0.8823/0.7719	0.8783/0.7688	0.8720/0.7520	
0.65	1	1	1	0.7989/0.0516	0.7355/0.0483	0.6733/0.0501	0.6062/0.0478	0.5777/0.0475	0.5517/0.0393	
				0.8953/0.0501	0.8481/0.0509	0.7846/0.0492	0.7044/0.0546	0.6692/0.0467	0.6455/0.0417	
0.65	1	0.7	1	0.8152/0.7102	0.7989/0.6877	0.7939/0.6803	0.7814/0.6592	0.7759/0.6387	0.7723/0.6283	
				0.9016/0.8415	0.9006/0.8339	0.8918/0.8293	0.8851/0.8140	0.8783/0.7988	0.8706/0.7897	
0.65	0.65	0.7	0.7	0.9140/0.8141	0.9172/0.8090	0.9157/0.8095	0.9052/0.7871	0.8968/0.7597	0.9019/0.7549	
				0.9660/0.9123	0.9674/0.9106	0.9698/0.9105	0.9635/0.9001	0.9576/0.8858	0.9579/0.8751	
0.65	0.65	0.75	0.7	0.9229/0.7207	0.9167/0.7189	0.9056/0.7130	0.8925/0.6978	0.8978/0.6895	0.8879/0.6469	
				0.9693/0.8276	0.9704/0.8332	0.9664/0.8297	0.9532/0.8157	0.9542/0.8137	0.9489/0.7853	
0.65	0.47	0.7	0.53	0.9934/0.9578	0.9913/0.9597	0.9900/0.9544	0.9866/0.9502	0.9888/0.9456	0.9878/0.9399	
				0.9997/0.9801	0.9995/0.9841	0.9990/0.9800	0.9987/0.9788	0.9990/0.9787	0.9990/0.9763	