Published in final edited form as:

Am Stat. 2012 ; 66(3): 155–162. doi:10.1080/00031305.2012.720900.

Estimation of Progression-Free Survival for All Treated Patients in the Randomized Discontinuation Trial Design

Theodore G. Karrison [Research Associate (Associate Professor)],

Department of Health Studies, University of Chicago, 5841 S. Maryland Ave., MC2007, Chicago, IL 60637

Mark J. Ratain [Professor],

Department of Medicine, University of Chicago, 5841 S. Maryland Ave., MC2115, Chicago, IL 60637

Walter M. Stadler [Professor], and

Department of Medicine, University of Chicago, 5841 S. Maryland Ave., MC2115, Chicago, IL 60637

Gary L. Rosner [Professor]

Department of Medicine, Johns Hopkins University, 550 N. Broadway, Suite 1103, Baltimore, MD 21205

Theodore G. Karrison: tkarrison@health.bsd.uchicago.edu; Mark J. Ratain: mratain@medicine.bsd.uchicago.edu; Walter M. Stadler: wmstadler@medicine.bsd.uchicago.edu; Gary L. Rosner: grosner1@johnshopkins.edu

Abstract

The randomized discontinuation trial (RDT) design is an enrichment-type design that has been used in a variety of diseases to evaluate the efficacy of new treatments. The RDT design seeks to select a more homogeneous group of patients, consisting of those who are more likely to show a treatment benefit if one exists. In oncology, the RDT design has been applied to evaluate the effects of cytostatic agents, that is, drugs that act primarily by slowing tumor growth rather than shrinking tumors. In the RDT design, all patients receive treatment during an initial, open-label run-in period of duration T . Patients with objective response (substantial tumor shrinkage) remain on therapy while those with early progressive disease are removed from the trial. Patients with stable disease (SD) are then randomized to either continue active treatment or switched to placebo. The main analysis compares outcomes, for example, progression-free survival (PFS), between the two randomized arms. As a secondary objective, investigators may seek to estimate PFS for all treated patients, measured from the time of entry into the study, by combining information from the run-in and post run-in periods. For $t \leq T$, PFS is estimated by the observed proportion of patients who are progression-free among all patients enrolled. For $t > T$, the estimate can be expressed as $\hat{S}(t) = p_{OR} \times \hat{S}_{OR}(t - T) + p_{SD} \times \hat{S}_{SD}(t - T)$, where p_{OR} is the estimated probability of response during the run-in period, p_{SD} is the estimated probability of SD, and $\hat{S}_{OR}(t - T)$ and $\hat{S}_{SD}(t - T)$ are the Kaplan–Meier estimates of subsequent PFS in the responders and patients with SD randomized to continue treatment, respectively. In this article, we derive the variance of $\hat{S}(t)$, enabling the construction of confidence intervals for both $\hat{S}(t)$ and the median survival time. Simulation results indicate that the method provides accurate coverage rates. An interesting aspect of the design is that outcomes during the run-in phase have a negative multinomial distribution, something not frequently encountered in practice.

Keywords

Confidence limits; Enrichment design; Negative multinomial distribution; Phase II clinical trials

1. INTRODUCTION

In phase II oncology clinical trials, the primary endpoint has traditionally been objective response (OR), defined as a 50% or greater, or 30% or greater, reduction in tumor size depending upon whether the sum of bidimensional or unidimensional measurements are used (Therasse et al. 2000). Although this endpoint may be appropriate when the experimental agent is cytotoxic in nature (i.e., kills tumor cells, leading to measurable tumor shrinkage), it is not suitable for cytostatic drugs (i.e., drugs that act primarily by slowing tumor growth rather than shrinking tumors; Korn et al. 2001). While these agents may produce a certain number of objective responders, their primary mode of action is to delay tumor growth. Ratain et al. (2006) reported the results of a randomized discontinuation trial (RDT) of the drug sorafenib in patients with metastatic renal cell cancer. Sorafenib targets both tumor cells and the tumor vasculature and, if active, was believed more likely to delay growth rather than cause significant tumor regression. The data from this trial are analyzed in Section 3. Another important issue in the era of molecularly targeted therapy is the need to identify which patients are most likely to benefit from a given treatment regimen. As noted by Stadler (2007), however, this can be quite difficult, particularly during early-phase drug development. The RDT design (Rosner, Stadler, and Ratain 2002) offers a means to “select” a cohort most likely to benefit from a given treatment and to rigorously evaluate its disease-stabilizing activity.

2. THE RANDOMIZED DISCONTINUATION TRIAL DESIGN IN ONCOLOGY

The RDT is an enrichment design in which all patients receive the experimental treatment during an initial, open-label run-in period of length T . Enrichment designs select subjects for participation in a subsequent randomized comparison phase of a study on the basis of their response during the run-in phase. The objective is to select a subset that is relatively homogeneous and more likely to show a treatment benefit, thereby increasing statistical power. In the RDT design as typically applied in oncology, patients with an OR during the lead-in phase remain on therapy while those with early progressive disease (PD) are removed from the trial. Patients with stable disease (SD) are randomized to either continue active treatment or switched to placebo. The main analysis compares outcomes, for example, progression-free survival (PFS), between the two randomized arms measured from the point of randomization. The design is shown schematically in Figure 1.

The efficiency or power of the RDT design relative to standard, up-front randomization in which patients are assigned to drug or placebo at initial entry into the trial depends on the degree to which enrichment is achieved by eliminating early disease progressors. It must also be assumed that there is no lasting carryover effect from the run-in period for those patients subsequently randomized to placebo. Note that one aspect of the RDT that differs from usual enrichment-type designs is that patients with the most favorable initial outcome, OR, are not randomized, but remain on therapy for ethical reasons. Evaluations of the RDT compared with other designs can be found in Kopec, Abrahamowicz, and Esdaile (1993), Leber and Davis (1998), Capra (2004), Freidlin and Simon (2005), and Fu, Dowlati, and Schluchter (2009). RDT designs are employed much less frequently than standard (“up-front”) randomized designs, but their use is increasing. An Ovid MEDLINE search using the search phrases “Randomized Discontinuation” or “Randomized Withdrawal” found a total of 37 published trials employing the RDT design from 2002 to 2012, eight of which were cancer trials.

As with any enrichment design, a limitation of the RDT is that it does not provide an estimate of the overall efficacy of the agent in the unselected patient population. If an RDT trial is positive, it is likely that any follow-up phase III trial will need to be performed in an

unselected population and will employ up-front randomization, which in turn generally requires assumptions about PFS rates for trial design purposes. While it would appear that the data from an RDT are not helpful in this regard, at least one side of the coin can be recovered, namely, the PFS rate among all *treated* patients measured from the time of enrollment into the study. This, together with historical data on outcomes following standard therapy, could then be used to postulate an effect size (hazard ratio [HR], say) in an unselected population. One can also estimate the HR in an unselected population based on the HR in the RDT and the randomization rate, together with assumptions regarding the HR in the nonrandomized patients. In the phase II RDT trial of sorafenib, Ratain et al. (2006) provided such a PFS estimate as a secondary analysis. The procedure requires combining information from the run-in and post run-in periods. The purpose of this article is to show that the method provides an unbiased estimate and to derive point-wise confidence intervals (CIs) for the PFS curve. Once this is achieved, CIs for the median survival time can be obtained using, for example, the procedure described in Brookmeyer and Crowley (1982).

The next section describes the methodology. The data from the published renal cell cancer trial of Ratain et al. are then used to illustrate the approach. Simulation results are presented in Section 5 to verify the validity of the proposed technique, followed by a summary and discussion.

3. METHODS

As discussed above, following the main analysis from an RDT it may be useful to estimate PFS for all treated patients, measured from the time of entry into the study. This can be performed by combining information from the run-in and post run-in periods. For our purposes, it is useful to conceive of the true PFS curve as $S(t)$ for $t \leq T$ and $S(t) = \Pr(\text{OR}) S_{\text{OR}}(t - T) + \Pr(\text{SD}) S_{\text{SD}}(t - T)$ for $t > T$, where $\Pr(\text{OR})$ is the probability of being a responder at time T , $\Pr(\text{SD})$ is the probability of having SD at time T , and $S_{\text{OR}}(t - T)$ and $S_{\text{SD}}(t - T)$ are the subsequent PFS probabilities in the responders and patients with SD at time T , respectively. More succinctly, for $t > T$, $S(t)$ is the probability of not progressing during the run-in period (duration T) and surviving progression free a further $t - T$ units. We assume that any patients who were eligible for randomization but dropped out did so for reasons unrelated to outcome.

For $t \leq T$, PFS can be estimated simply by the observed proportion of patients progression-free at time t among all patients enrolled, with CIs obtained using the standard normal approximation to the binomial. For $t > T$, an estimate can be obtained as

$$\widehat{S}(t) = \widehat{p}_{\text{OR}} \times \widehat{S}_{\text{OR}}(t - T) + \widehat{p}_{\text{SD}} \times \widehat{S}_{\text{SD}}(t - T), \quad (1)$$

where p_{OR} is the estimated probability of response during the run-in period, p_{SD} is the estimated probability of SD, and $\widehat{S}_{\text{OR}}(t - T)$ and $\widehat{S}_{\text{SD}}(t - T)$ are Kaplan and Meier (1958) estimates of PFS in the responders and patients with SD randomized to continue treatment, respectively. One would expect $\widehat{S}_{\text{OR}}(t - T)$ to exceed $\widehat{S}_{\text{SD}}(t - T)$, that is, patients who achieve an early response should have better subsequent outcomes than those who exhibit only SD during the run-in period.

Let X_1 denote the number of objective responders during the run-in period, X_2 the number of patients with disease progression, and n_{SD} the number with SD. In planning RDT studies, the primary focus is on the comparison of the two randomized groups; therefore, n_{SD} is generally *fixed* by design, chosen to provide a desired level of power to detect a true difference between the two arms. Consequently, $X = (X_1, X_2)$ has a negative multinomial distribution with parameters $k = n_{\text{SD}}$, m_1 , and m_2 (see, e.g., Bishop, Fienberg, and Holland

1975, BFH). Using Equation (13.8-3) of BFH for the mixed factorial moments of the negative multinomial distribution, we have

$$\begin{aligned} E(X_i) &= m_i \\ V(X_i) &= \frac{m_i(m_i+k)}{k} \\ \text{cov}(X_1, X_2) &= \frac{m_1 m_2}{k}. \end{aligned}$$

Unbiased estimates of p_{OR} , the probability of response during the run-in period, and p_{SD} , the probability of SD, are $p_{\text{OR}} = X_1/(n-1)$ and $p_{\text{SD}} = (n_{\text{SD}}-1)/(n-1)$, where $n = X_1 + X_2 + n_{\text{SD}}$, with variances

$$V(\widehat{p}_{\text{OR}}) = \frac{X_1(n-X_1)}{n^2(n-1)} \quad (2)$$

and

$$V(\widehat{p}_{\text{SD}}) = \frac{n_{\text{SD}}(n-n_{\text{SD}})}{n^2(n-1)}, \quad (3)$$

respectively (Haldane 1945). Note that replacing $(n-1)$ by n in Equations (2) and (3) results in the familiar variance estimates obtained in the multinomial case when n , rather than n_{SD} , is fixed.

Letting $f_1(X_1, X_2) = p_{\text{OR}}$, $f_2(X_1, X_2) = p_{\text{SD}}$, and $\sigma_{ij} = \text{cov}(X_i, X_j)$, the Δ -method gives

$$\begin{aligned} \text{cov}(\widehat{p}_{\text{OR}}, \widehat{p}_{\text{SD}}) &\cong \\ &\cong \sum_{i,j=1}^2 \left(\frac{\partial f_1}{\partial X_i} \right) \left(\frac{\partial f_2}{\partial X_j} \right) \sigma_{ij} \\ &= \left(\frac{n-1-X_1}{(n-1)^2} \right) \left(-\frac{(n_{\text{SD}}-1)}{(n-1)^2} \right) \frac{m_1(m_1+n_{\text{SD}})}{n_{\text{SD}}} \\ &\quad + \left(\frac{n-1-X_1}{(n-1)^2} \right) \left(-\frac{(n_{\text{SD}}-1)}{(n-1)^2} \right) \frac{m_1 m_2}{n_{\text{SD}}} \\ &\quad + \left(\frac{-X_1}{(n-1)^2} \right) \left(-\frac{(n_{\text{SD}}-1)}{(n-1)^2} \right) \frac{m_1 m_2}{n_{\text{SD}}} \\ &\quad + \left(\frac{-X_1}{(n-1)^2} \right) \left(-\frac{(n_{\text{SD}}-1)}{(n-1)^2} \right) \frac{m_2(m_2+n_{\text{SD}})}{n_{\text{SD}}}, \end{aligned}$$

which after reduction and substitution of X_j for m_j becomes

$$\text{cov}(\widehat{p}_{\text{OR}}, \widehat{p}_{\text{SD}}) \cong \frac{-X_1 \cdot n_{\text{SD}}}{n^3}, \quad (4)$$

again a familiar expression. The expectation and variance of $\widehat{S}(t)$ can now be obtained straightforwardly by first conditioning on (X_1, X_2) . For the mean,

$$\begin{aligned} E[\widehat{S}(t)] &= E[E(\widehat{S}(t)|X_1, X_2)] \\ &= E[\widehat{p}_{\text{OR}} S_{\text{OR}}(t-T) + \widehat{p}_{\text{SD}} S_{\text{SD}}(t-T)] \\ &= p_{\text{OR}} S_{\text{OR}}(t-T) + p_{\text{SD}} S_{\text{SD}}(t-T) \\ &= S(t). \end{aligned}$$

And for the variance,

$$\begin{aligned}
V(\widehat{S}(t)) &= E\{V(\widehat{S}(t)|X_1, X_2)\} + V\{E(\widehat{S}(t)|X_1, X_2)\} \\
&= E\{\widehat{p}_{OR}^2 V(\widehat{S}_{OR}(t-T)) + \widehat{p}_{SD}^2 V(\widehat{S}_{SD}(t-T))\} \\
&\quad + V\{\widehat{p}_{OR} S_{OR}(t-T) + \widehat{p}_{SD} S_{SD}(t-T)\}, \\
&= V(\widehat{S}_{OR}(t-T)) E(\widehat{p}_{OR}^2) + V(\widehat{S}_{SD}(t-T)) E(\widehat{p}_{SD}^2) \\
&\quad + (S_{OR}(t-T))^2 V(\widehat{p}_{OR}) + (S_{SD}(t-T))^2 V(\widehat{p}_{SD}) \\
&\quad + 2S_{OR}(t-T)S_{SD}(t-T)\text{cov}(\widehat{p}_{OR}, \widehat{p}_{SD}) \\
&= V(\widehat{S}_{OR}(t-T)) [V(\widehat{p}_{OR}) + p_{OR}^2] + V(\widehat{S}_{SD}(t-T)) \\
&\quad \times [V(\widehat{p}_{SD}) + p_{SD}^2] + (S_{OR}(t-T))^2 V(\widehat{p}_{OR}) \\
&\quad + (S_{SD}(t-T))^2 V(\widehat{p}_{SD}) + 2S_{OR}(t-T)S_{SD}(t-T) \\
&\quad \times \text{cov}(\widehat{p}_{OR}, \widehat{p}_{SD}),
\end{aligned} \tag{5}$$

where we have used the fact that the Kaplan–Meier estimators are unbiased.

The variance can be calculated from (5) by replacing true values with their respective estimates, using Greenwood’s formula for $V(\widehat{S}_{OR}(t-T))$ and $V(\widehat{S}_{SD}(t-T))$, and applying Equations (2)–(4).

As an aside, an alternative way of estimating $S(t)$ is to multiply the probability of remaining progression-free during the run-in period by the weighted average of the PFS curves for the two subgroups treated beyond the run-in period. Specifically, using the above notation,

$$\begin{aligned}
\widehat{S}(t) &= \frac{X_1 + n_{SD}}{X_1 + X_2 + n_{SD}} \\
&\quad \times \left[\frac{X_1}{X_1 + n_{SD}} \widehat{S}_{OR}(t-T) + \frac{n_{SD}}{X_1 + n_{SD}} \widehat{S}_{SD}(t-T) \right]. \tag{6}
\end{aligned}$$

This was the approach taken by Ratain et al. (2006) in analyzing the data from the renal cell cancer study (although no standard error was provided). Equation (6) is equivalent to Equation (1) except for minor adjustments.

4. EXAMPLE

Ratain et al. (2006) reported the results of an RDT in renal cell carcinoma. A total of 202 patients were enrolled in the trial: 14 patients discontinued treatment before the 12 week assessment (12 due to adverse events, 1 patient withdrew consent, and 1 was lost to follow-up), therefore the analysis is based on $n = 188$ patients. Seventy-three patients had tumor shrinkage of 25% or more during the 12-week run-in phase and, per protocol, remained on open-label sorafenib, while 6 other patients also were continued on open-label drug at the discretion of the treating physician; 44 patients progressed or died during the run-in phase; 65 patients had SD at 12 weeks (with the exception of one patient with PD who was inadvertently randomized), of whom 32 were randomly assigned to continue sorafenib and 33 to receive placebo. Thus, using the above notation, $n_{SD} = 65$, $X_1 = 79$, and $X_2 = 44$. Of note, the dataset analyzed here is not identical to the dataset analyzed in the 2006 publication—the latter includes updated follow-up in a handful of cases. PFS from week 12 in the two randomized arms was significantly different in the two groups ($p = 0.0087$) with a median of 24 weeks in the sorafenib arm compared with 6 weeks in the placebo group.

Our interest is in estimating PFS following treatment with sorafenib, measured from the date of *initial* entry into the trial. Figure 2 shows the estimated curve along with 95% CIs using the method described in Section 3. Six-month and one-year PFS rates were 49.0% (95% CI: 40.9%–58.7%) and 11.7% (95% CI: 5.5%–25.0%), respectively. The median PFS time was 176 days.

A good way to see how the procedure works graphically is to consider the alternative form of the estimator shown in Equation (6). The proportion of patients alive and without

progression at the end of the run-in period is $\frac{X_1 + n_{SD}}{X_1 + X_2 + n_{SD}} = \frac{79+65}{69+44+65} = 0.766$. This quantity, multiplied by the weighted average of the curves for the OR and SD subgroups, yields the overall estimate for $t > T$. The two components and the final estimate are shown in Figure 3.

From the pointwise confidence limits for the survival curve, a CI for the median PFS time can be obtained as the set of all time points t such that 0.50 lies within the 95% CI generated at time t (Brookmeyer and Crowley 1982). This gives an asymmetrical interval ranging from 163 to 244 days. Of interest, the RDT trial was followed by a phase III, up-front randomized clinical trial (Escudier et al. 2007) that also found a statistically significant benefit for sorafenib ($p < 0.001$) with a median PFS of 5.5 months (168 days) in the sorafenib arm, consistent with that calculated above from the RDT.

5. SIMULATION STUDY

A simulation study was undertaken to confirm the accuracy of (5) and to assess coverage rates of CIs. We simulated a typical phase II trial in which patients are accrued over an interval $[0, a]$ and follow-up is continued beyond the accrual period for an additional time f . The number of patients entering the randomized phase, n_{SD} , was fixed and outcomes (OR, SD, PR) during the run-in period were drawn randomly until n_{SD} SDs occurred. More specifically, the entry time for the i th patient, e_i , was drawn from a uniform distribution over $[0, a]$. The run-in period, T , was added to the entry time (thus, it was assumed that no patients would dropout or be lost to follow-up during the run-in phase) and for patients with outcomes of OR or SD, the subsequent time to disease progression/death, u_i , was drawn from one of two Weibull distributions:

$$\begin{aligned} S_{OR}(u) &= \exp(-(\lambda_1 u)^{p_1}) \\ S_{SD}(u) &= \exp(-(\lambda_2 u)^{p_2}). \end{aligned}$$

An equal number of patients, $n_{SD}/2$, were allocated to the treatment and placebo arms in the SD subgroup and only the former entered into the subsequent calculations. If, for the i th patient, $e_i + T + u_i \leq a + f$, the event was observed, whereas if $e_i + T + u_i > a + f$, the patient was censored at $a + f$. An additional random censoring mechanism following an exponential distribution with rate parameter c was also incorporated to allow for losses to follow-up postrandomization. All calculations were performed in SAS, version 9.2 (Cary, NC). After generation of the data for a simulated trial, Equation (1) was used to estimate the overall PFS curve and the variance was calculated via (5). Pointwise 95% CIs were then determined based on four different transformations g (Klein and Moeschberger 1997):

$$g(\hat{S}(t_j)) \pm 1.96 \sqrt{(g'(\hat{S}(t_j)))^2 V(\hat{S}(t_j))}, \quad j=1, 2, \dots, k. \quad (7)$$

Transformations examined were identity (untransformed), log, complementary log–log, and logit. Note that if the Kaplan–Meier estimate of the survival rate at some time point becomes zero, the standard error from Greenwood's formula is also zero. In these cases, the standard error from the previous time point was used in Equation (5) when calculating $V(\hat{S}(t))$, and for the log, complementary log–log, and logit based CIs, the interval obtained from the identity transformation was employed instead. Finally, when the Kaplan–Meier estimate was undefined due to censoring (last observation censored prior to the time point in question), the replicate was excluded from the calculations for that time point.

Tables 1-3 show results based on 10,000 replications for three different scenarios and sample sizes $n_{SD} = 60, 80, \text{ and } 100$. The three scenarios correspond to SD rates of 60%, 70%, and 45% and response rates during the run-in period of 15%, 10%, and 5%, respectively. In all three scenarios, the accrual period, a , was set to 48 weeks, the follow-up period, f , to 36 weeks, and the run-period, T , to 12 weeks. The true overall survival rate, $S(t)$, for $t = 18, 24, \dots, 48$ weeks is displayed in column 4 of the tables.

Overall the simulation results support the validity of the approach. The PFS estimator (1) is seen to have minimal bias. The estimated standard deviation based on Equation (5) was quite close on average to the empirical standard deviation in all cases. The complementary log-log and logit transformations appear to provide the best coverage rates, although some have observed that the latter may be slightly conservative (Link 1984).

6. DISCUSSION

The RDT is conducted in two stages. The first stage treats all patients with the new agent and potentially identifies a sensitive subpopulation. This subpopulation is then evaluated in the second stage in a randomized manner. The efficiency of the RDT relative to a standard, two-arm-randomized trial depends on its ability to identify a sensitive subpopulation (if one exists) and this, in turn, depends on the total sample size and the duration of the run-in period. Stadler (2007) pointed out, for example, that too short a run-in period may reduce the ability to detect an active agent, due to lack of enrichment.

The primary focus of the RDT is on the randomized comparison. Following this analysis, interest may turn to estimating survival in the population of all treated patients measured from the time of initial entry into the trial. This estimate, together with historical control data, would be useful for planning a follow-on phase III trial. At first glance, it would appear that no such estimate is afforded by the RDT design; however, as described above, an estimator can be recovered from the data by piecing together information from the two phases of the trial. While we have focused on estimation of PFS, the method is also applicable to the estimation of overall survival. This would be accomplished by estimating $S(t)$ for $t \leq T$ by the proportion of survivors at time t (here $S(t)$ denotes overall survival) and extending Equation (1) to

$$\begin{aligned} \widehat{S}(t) = & \widehat{p}_{OR} \times \widehat{S}_{OR}(t-T) + \widehat{p}_{SD} \times \widehat{S}_{SD}(t-T) \\ & + \widehat{p}_{PD} \times \widehat{S}_{PD}(t-T), t > T, \end{aligned}$$

where p_{PD} is the estimated probability of PD during the run period and $\widehat{S}_{PD}(t-T)$ is the Kaplan-Meier estimate of subsequent survival in patients with disease progression during the run-in. A caveat in oncology trials, however, is that once a patient's disease progresses they are usually administered additional therapies that may affect their subsequent survival course.

The RDT design is generally set up by fixing the number of patients randomized to obtain enough power for the comparison of the randomized arms. Thus, the method described here assumes that the number of patients with SD, n_{SD} , is fixed by design. This leads to the negative multinomial distribution for the number of responders and early progressors during the run-in period. The differences, though, are minor relative to the results that would be obtained if the total sample size, n , were treated as fixed. As discussed in Section 3, fixing the total sample size instead of the number of SDs changes the variances of the respective probability estimates by a factor of $(n-1)/n$. That is, the variances for p_{OR} and p_{SD} in Equations (2) and (3) have n in the denominator instead of $(n-1)$. One could then carry that

through in the calculations of the final variance in Equation (5). In fact, another approach to designing RDTs considers n , T , and T_2 , where T_2 is the duration of the postrandomization follow-up period, as “tuning parameters” (Trippa, Rosner, and Muller 2012). Trippa et al. described a Bayesian decision-theoretic approach for choosing these parameters in an optimal manner. For practical sample sizes, the methods presented here should provide appropriate estimates under either set of assumptions.

The method should work well with reasonable sample sizes. Additional simulations run under the first scenario with n_{SD} equal to 40 and 20 (mean $n = 67$ and 33) continued to have minimal bias and good coverage rates. We did employ a γ -method approximation to obtain the covariance of p_{OR} and p_{SD} in Equation (4). And Greenwood’s formula is used in Equation (5) to obtain the variance of the Kaplan–Meier estimators. It is well known that when censoring is heavy, Greenwood’s formula can underestimate the variance in the tails of the survival distribution (Peto et al. 1977).

We have not addressed other approaches to predict phase III results from a positive RDT. It would also be desirable to estimate the HR for a standard randomized controlled trial, on the basis of an RDT. This cannot be precisely estimated, since the HR for nonrandomized patients is unknown. However, given the increasing use of the RDT design in oncology, additional statistical work will be important to maximize the value of the ensuing data.

References

- Bishop, YMM.; Fienberg, SE.; Holland, PW. Discrete Multivariate Analysis. Cambridge: MIT Press; 1975.
- Brookmeyer R, Crowley J. A Confidence Interval for the Median Survival Time. *Biometrics*. 1982; 38:29–41.
- Capra W. Comparing the Power of the Discontinuation Design to That of the Classic Randomized Design on Time-to-Event Endpoints. *Controlled Clinical Trials*. 2004; 25:168–177. [PubMed: 15020035]
- Escudier B, Eisen T, Stadler WM, Szczylik C, Oudard S, Siebels M, Negrier S, Chevreau C, Solska E, Desai AA, Rolland F, Demkow T, Hutson TE, Gore M, Freeman S, Schwartz B, Shan M, Simantov R, Bukowski RM. TARGET Study Group. Sorafenib in Advanced Clear-Cell Renal-Cell Carcinoma. *New England Journal of Medicine*. 2007; 356:125–134. [PubMed: 17215530]
- Freidlin B, Simon R. Evaluation of Randomized Discontinuation Design. *Journal of Clinical Oncology*. 2005; 23:5094–5098. [PubMed: 15983399]
- Fu P, Dowlati A, Schluchter M. Comparison of Power Between Randomized Discontinuation Design and Upfront Randomization Design on Progression-Free Survival. *Journal of Clinical Oncology*. 2009; 27:4135–4141. [PubMed: 19636018]
- Haldane JBS. On a Method of Estimating Frequencies. *Biometrika*. 1945; 33:222–225. [PubMed: 21006837]
- Kaplan EL, Meier P. Nonparametric Estimation From Incomplete Observations. *Journal of the American Statistical Association*. 1958; 53:457–481.
- Klein, JP.; Moeschberger, ML. *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer-Verlag; 1997.
- Kopec JA, Abrahamowicz M, Esdaile JM. Randomized Discontinuation Trials: Utility and Efficiency. *Journal of Clinical Epidemiology*. 1993; 46:959–971. [PubMed: 8263581]
- Korn EL, Arbuck SG, Pluda JM, Simon R, Kaplan RS, Christian MC. Clinical Trial Designs for Cytostatic Agents: Are New Approaches Needed? *Journal of Clinical Oncology*. 2001; 19:265–272. [PubMed: 11134222]
- Leber PD, Davis CS. Threats to the Validity of Clinical Trials Employing Enrichment Strategies for Sample Selection. *Controlled Clinical Trials*. 1998; 19:178–187. [PubMed: 9551282]
- Link C. Confidence Intervals for the Survival Function using Cox’s Proportional-Hazard Model With Covariates. *Biometrics*. 1984; 40:601–610. [PubMed: 6394063]

- Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. Design and Analysis of Randomized Clinical Trials Requiring Prolonged Observation of Each Patient, II: Analysis and Examples. *British Journal of Cancer*. 1977; 35:1–39. Statistical Note 6. [PubMed: 831755]
- Ratain MJ, Eisen T, Stadler WM, Flaherty KT, Kaye SB, Rosner GL, Gore M, Desai AA, Patnaik A, Xiong HQ, Rowinsky E, Abbruzzese JL, Xia C, Simantov R, Schwartz B, O'Dwyer PJ. Phase II Placebo-Controlled Randomized Discontinuation Trial of Sorafenib in Patients With Metastatic Renal Cell Carcinoma. *Journal of Clinical Oncology*. 2006; 24:2505–2512. [PubMed: 16636341]
- Rosner GL, Stadler W, Ratain MJ. Randomized Discontinuation Design: Application to Cytostatic Antineoplastic Agents. *Journal of Clinical Oncology*. 2002; 20:4478–4484. [PubMed: 12431972]
- Stadler WS. The Randomized Discontinuation Trial: A Phase II Design to Assess Growth-Inhibitory Agents. *Molecular Cancer Therapeutics*. 2007; 6:1180–1184. [PubMed: 17431101]
- Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, Verweij J, Van Glabbeke M, van Oosterom AT, Christian MC, Gwyther SG. New Guidelines to Evaluate the Response to Treatment in Solid Tumors. *Journal of the National Cancer Institute*. 2000; 92:205–216. [PubMed: 10655437]
- Trippa L, Rosner GL, Muller P. Bayesian Enrichment Strategies for Randomized Discontinuation Trials. *Biometrics*. 2012; 68:203–212. [PubMed: 21714780]

Figure 1.
Randomized discontinuation trial design. The online version of this figure is in color.

Figure 2. Estimated progression-free survival (solid line) and pointwise 95% confidence intervals (dotted lines) following treatment with sorafenib, measured from start of treatment. Confidence intervals are based on the log transformation. The online version of this figure is in color.

Figure 3. Estimated progression-free survival (solid line) and OR and SD components (dashed lines). For $T > 84$ days, the solid line is the weighted average of the dashed lines (weights = 0.549 and 0.451, respectively).

Table 1

Simulation results, first scenario^a. $R = 10,000$ replications, except as indicated in the footnotes

n_{SD}	Mean n	t	$S(t)$	Mean $\hat{S}(t)$	Empirical Std Dev	Mean $\sqrt{V(\hat{S}(t))}$	Coverage rates (%)			
							Un	Log	CLL	Logit
60	99.9	18	0.5852	0.5849	0.0608	0.0599	94.2	94.1	95.0	95.0
		24	0.4298	0.4295	0.0644	0.0631	93.9	94.7	94.6	94.9
		30	0.3114	0.3109	0.0560	0.0591	93.8	94.9	94.7	95.2
		36	0.2259	0.2262 ^b	0.0531	0.0526	93.5	95.0	94.9	95.3
		42	0.1652	0.1661 ^c	0.0471	0.0467	94.0	96.1	96.0	96.5
80	133.4	48	0.1221	0.1232 ^d	0.0423	0.0427	95.0	96.0	97.7	96.8
		18	0.5852	0.5858	0.0524	0.0519	94.3	94.4	95.0	95.2
		24	0.4298	0.4301	0.0555	0.0547	94.4	94.7	94.9	95.0
		30	0.3114	0.3120	0.0525	0.0513	94.0	94.7	94.7	95.1
		36	0.2259	0.2270	0.0467	0.0456	93.5	94.4	94.7	94.8
100	166.7	42	0.1652	0.1667 ^e	0.0414	0.0403	93.6	94.9	94.8	95.2
		48	0.1221	0.1244 ^f	0.0372	0.0364	94.2	95.4	96.6	96.3
		18	0.5852	0.5848	0.0465	0.0465	94.6	95.0	94.9	95.1
		24	0.4298	0.4296	0.0490	0.0490	94.6	94.8	95.2	95.1
		30	0.3114	0.3113	0.0457	0.0459	94.5	95.0	95.0	95.1
		36	0.2259	0.2261	0.0409	0.0409	94.1	95.2	94.8	95.3
		42	0.1652	0.1655 ^g	0.0368	0.0359	93.6	94.9	94.7	95.1
		48	0.1221	0.1229 ^h	0.0326	0.0322	94.2	95.8	95.9	96.3

Means are mean values and Empirical Std Dev is Monte Carlo standard deviation over 10,000 replications. Un: untransformed scale, CLL: complementary log-log.

^a $POR = 0.15$, $PSD = 0.60$, $1 = 0.003$, $p1 = 1.7$, $2 = 0.009$, $p2 = 1.2$, $c = 0.0005$.

^b $R = 9999$.

^c $R = 9877$.

^d $R = 9405$.

^e $R = 9948$.

$$\begin{aligned} f^*R &= 9600. \\ g^*R &= 9987. \\ h^*R &= 9754. \end{aligned}$$

Table 2

Simulation results, second scenario^a. $R = 10,000$ replications, except as indicated in the footnotes

n_{SD}	Mean	n	t	$S(t)$	Mean $\hat{S}(t)$	Empirical Std Dev	Mean $\sqrt{V(\hat{S}(t))}$	Coverage rates (%)			
								Un	Log	CLL	Logit
60	85.7	18	0.6699		0.6693 ^b	0.0618	0.0614	93.8	93.9	95.0	95.2
			0.5200		0.5190 ^b	0.0707	0.0699	94.0	94.5	95.1	95.4
			0.3913		0.3901 ^c	0.0708	0.0696	93.7	94.6	94.8	95.2
			0.2906		0.2896 ^d	0.0652	0.0647	93.7	95.2	95.1	95.7
			0.2155		0.2145 ^e	0.0589	0.0585	93.2	95.6	95.5	96.3
80	114.3	18	0.1612		0.1616 ^f	0.0542	0.0542	93.3	96.2	97.1	97.1
			0.6699		0.6701	0.0534	0.0532	94.0	94.2	94.8	95.0
			0.5200		0.5204	0.0614	0.0606	94.3	94.4	94.9	95.2
			0.3913		0.3922	0.0614	0.0605	94.2	94.8	95.0	95.2
			0.2906		0.2911	0.0568	0.0563	93.6	94.8	94.8	95.2
100	142.8	18	0.2155		0.2158 ^g	0.0519	0.0510	93.2	94.7	94.6	95.1
			0.1612		0.1620 ^h	0.0472	0.0462	93.1	95.5	95.4	96.0
			0.6699		0.6707	0.0484	0.0477	94.2	93.9	94.8	95.0
			0.5200		0.5202	0.0546	0.0543	94.5	94.5	95.0	95.3
			0.3913		0.3913	0.0542	0.0542	94.7	95.1	95.3	95.5
		36	0.2906		0.2902	0.0508	0.0504	94.3	95.0	95.0	95.3
			0.2155		0.2149 ⁱ	0.0456	0.0456	93.8	95.2	94.9	95.3
			0.1612		0.1611 ^j	0.0422	0.0412	92.9	95.1	94.8	95.3

Means are mean values and Empirical Std Dev is Monte Carlo standard deviation over 10,000 replications. Un: untransformed scale, CLL: complementary log-log.

^a $pOR = 0.10$, $pSD = 0.70$, $1 = 0.002$, $p1 = 1.8$, $2 = 0.007$, $p2 = 1.3$, $c = 0.0005$.

^b $R = 9996$.

^c $R = 9993$.

^d $R = 9991$.

^e $R = 9948$.

$$\begin{aligned} fR &= 9696, \\ gR &= 9989, \\ hR &= 9874, \\ iR &= 9998, \\ jR &= 9952. \end{aligned}$$

Table 3

Simulation results, third scenario.^a $R = 10,000$ replications, except as indicated in the footnotes.

n_{SD}	Mean n	t	$S(t)$	Mean $\hat{S}(t)$	Empirical Std Dev	Mean $\sqrt{V(\hat{S}(t))}$	Coverage rates (%)			
							Un	Log	CLL	Logit
60	133.4	18	0.4391	0.4385 ^b	0.0482	0.0474	94.5	94.7	95.0	95.1
		24	0.3563	0.3559 ^c	0.0510	0.0497	94.4	94.7	94.8	94.9
		30	0.2779	0.2776 ^c	0.0496	0.0490	94.1	95.4	95.0	95.5
	177.9	36	0.2120	0.2118 ^d	0.0464	0.0457	93.6	95.0	94.8	95.4
		42	0.1605	0.1607 ^e	0.0427	0.0418	93.3	95.0	94.8	95.2
80	133.4	48	0.1221	0.1230 ^f	0.0388	0.0381	93.1	95.7	95.5	96.0
		18	0.4391	0.4392 ^g	0.0409	0.0411	94.9	95.0	95.2	95.4
		24	0.3563	0.3566 ^g	0.0429	0.0432	95.0	94.9	95.4	95.4
	177.9	30	0.2779	0.2784 ^h	0.0423	0.0424	94.7	95.0	95.2	95.3
		36	0.2120	0.2125 ^h	0.0399	0.0397	94.0	94.9	94.8	95.1
100	133.4	42	0.1605	0.1611 ⁱ	0.0368	0.0363	93.7	94.8	94.7	95.0
		48	0.1221	0.1227 ^j	0.0338	0.0330	92.7	95.2	94.8	95.2
	222.3	18	0.4391	0.4384	0.0370	0.0367	94.6	94.8	94.8	94.9
		24	0.3563	0.3560	0.0389	0.0386	94.6	94.8	95.0	94.9
		30	0.2779	0.2775	0.0383	0.0379	94.5	95.2	95.0	95.3
100	222.3	36	0.2120	0.2116	0.0357	0.0355	94.4	95.1	95.1	95.3
		42	0.1605	0.1602 ^k	0.0328	0.0325	93.7	95.2	94.9	95.3
		48	0.1221	0.1218 ^l	0.0303	0.0296	92.8	95.1	94.5	95.2

Means are mean values and Empirical Std Dev is Monte Carlo standard deviation over 10,000 replications. Un: untransformed scale, CLL: complementary log-log.

^a $POR = 0.05$, $PSD = 0.45$, $1 = 0.001$, $p_1 = 1.9$, $2 = 0.006$, $p_2 = 1.4$, $c = 0.0005$.

^b $R = 9983$.

^c $R = 9980$.

^d $R = 9975$.

$^c R = 9931.$
 $^f R = 9771.$
 $^g R = 9997.$
 $^h R = 9996.$
 $^i R = 9988.$
 $^j R = 9934.$
 $^k R = 9998.$
 $^l R = 9979.$