CLINICAL INVESTIGATION

Monitoring rules for toxicity in Phase II oncology trials

Phase I oncology trials routinely assess safety and toxicity of investigational agents and/or their combinations. However, given the relatively small number of patients usually accrued in Phase I oncology trials, the maximum tolerated dose and the recommended Phase II dose of the treatment regimen can be imprecisely defined. This may lead to testing treatment regimens in Phase II trials at doses that could lead to excessive toxicity. We review various toxicity monitoring rules that are used to stop a Phase II trial early if the probability of dose-limiting toxicity is higher than what is expected based on the Phase I assessment and provide recommendations on which rules to use.

Keywords: biostatistics • Phase II trials • stopping rules • toxicity

Anastasia Ivanova*[,1],
Guochen Song[2],
Olga Marchenko[2]
& Stergios Moschos[3]
[1]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[2]Quintiles, 5927 South Miami Boulevard, Morrisville, NC 27560, USA
[3]Department of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
*Author for correspondence:
Tel.: +1 919 843 8086
Fax: +1 919 966 3804
aivanova@bios.unc.edu

The goal of the Phase II oncology clinical study is to gain preliminary insights into the clinical activity of an agent or treatment combination and to 'weed out' ineffective drugs from further, more costly, clinical development. Over the last 30 years, a significant number of novel statistical methodologies have been developed to diversify the classical, single-arm, frequentist, single-stage design for a Phase II oncology trial. Despite the development of novel, multiarm [1], or combined Phase I–II trial designs [2], most contemporary Phase II oncology trials remain single-arm in design and rely on information from Phase I studies [3]. Novel designs for such single-arm Phase II trials include various types of stopping rules for lack of efficacy, ranging from multistage enrollments using traditional frequentist approaches [4,5], to Bayesian methodologies that take into account accumulated information from prior experience (prior) as well as data collected (likelihood function) to update and/or adapt the design [6]. Such designs have traditionally maintained a single primary outcome, efficacy. Toxicity has typically been a secondary end point. Because efficacy of a particular treatment regimen can be positively associated with toxicity, it may be equally important within the context of a Phase II oncology trial not only to adopt stopping rules that address lack of efficacy, but also consider stopping rules that detect toxicity early. We believe that lack of prespecified rules for toxicity monitoring in Phase II trials is a major reason that data monitoring committees infrequently stop Phase II trials for safety.

To date, 'unacceptable' toxicity occasionally seen in Phase II trials is commonly managed by frequent (>50%) dose reductions or refusal of further treatment. This approach raises the question whether sufficient statistical power remains to assess if a given dose of the investigational agent for which the trial was specifically designed is actually effective, and whether lack of efficacy is a consequence of frequent dose reductions due to toxicity [7–10]. Frequentist and Bayesian methods have been developed to evaluate both toxicity and efficacy as bivariate (efficacy, safety) variables. Most of the methods are two-stage and range from equal weighing for response and toxicity, to designs with variable trade-offs between these two outcomes [11–15]. Detecting excessive toxicity early is vital, hence evaluating toxicity formally only once during the trial may not be sufficient.

FUTURE SCIENCE part of fsg

On rare occasions, serious adverse events or death (grade ≥4) occur [16,17]. The Phase II study E2205 conducted through the Eastern Cooperative Oncology Group is an interesting example of a trial with high incidence of grade 5 toxicity (i.e., death) observed in a high risk patient population often with poor prognosis [17]. Specifically, E2205 was a study of preoperative administration of systemic anticancer therapy that consisted of oxaliplatin, infusional 5-fluorouracil, and cetuximab concurrent with external beam irradiation prior to definitive esophagectomy in patients with operable esophageal adenocarcinoma. Standard multimodality therapy, namely the regimen without cetuximab, is associated with a known significant adverse event profile. A Simon two-stage design [4] was used with complete pathologic response as the primary end point. Although toxicity was an important secondary end point which was monitored closely in real-time with monthly teleconferences, no particular stopping rules for toxicity were incorporated in the study design. Six treatment-related deaths were observed in the initial cohort of 22 patients and the trial was stopped after stage 1. In this trial, a formal stopping rule for adverse events might have assisted the study team in systematic adverse event monitoring throughout the trial and might have stopped the trial earlier.

A formal adaptive rule should be in place to allow for stopping the trial at any point should the toxicity probability be unacceptably high. Does observing three treatment-related deaths in the first five patients warrant stopping the trial? A question like this might be raised by the data monitoring committee for the study if a formal stopping rule is not prespecified within the protocol. If such a rule is continuous, that is, monitoring occurs throughout the study, it will specify the minimum number of toxicities that warrant stopping the study out of the total number of patients who have already been enrolled and have completed their follow-up for toxicity. This type of rule is also known as the stopping boundary. For example, if the investigators anticipate to 'reasonably' observe approximately 5% of deaths within the study (e.g., allogeneic bone marrow transplantation with high probability of life-threatening graft-versus-host disease), a potential formal stopping rule [18] would recommend trial suspension if two grade five events are observed in the first two to four patients, three events are observed in 5–12 patients, four events in 13–21 patients, five events in 22–31 patients, and if six or more events are observed in more than 31 patients. Stopping rules can also be incorporated into more novel Phase II designs, such as the Phase IB portion of the combined Phase I-II trials to monitor the extended cohort of patients assigned to the estimated maximum tolerated dose (MTD), in which case the stopping rule will be typically based on the tolerable dose-limiting toxicity (DLT) rate of 0.20 or 0.25.

Unfortunately, a rigorous stopping rule for toxicity is not a standard feature of most current Phase II study designs. Even if there is a stopping rule for toxicity in a clinical trial protocol, it is not always mentioned when trial results are published. In this article, we review toxicity stopping rules we have encountered in published Phase II trials and give recommendations on what rules to use.

Bayesian stopping rule

We describe stopping rules based on both the traditional frequentist and the Bayesian statistics. Frequentists use fixed parameters to describe the unknown state of truth. For example, we can use θ to describe the true DLT probability and assume that there is a true value for θ, for example, 0.2 or 0.5. Bayesians, on the other hand, describe the unknowns with a certain degree of uncertainty.

Consider a hypothetical example of a Phase II trial patients with breast cancer with tumor response as a primary outcome. The null tumor response rate is 0.05, the alternative is 0.25, and both type I and type II error rates are 0.1. Simon's two-stage minimax design [4] requires 20 patients for this study with a futility look after 13 patients. To set up a stopping rule for excessive DLT for this study of the maximum of 20 patients, one needs to specify an acceptable probability of DLT, θ_0. Usually, θ_0 is the probability of toxicity that is expected to be seen at the MTD in the corresponding Phase I trial, assuming that each patient develops only a single DLT and that each patient has completed the study. The 3 + 3 design was used in the preceding Phase I study. The 3+3 design [19] is more likely to choose the estimated MTD with the DLT probability of 0.20 or 0.25 [20]. We use the same definition of the DLT in a Phase II trial is we used in preceding Phase I trial. We rely on data from a previous Phase I trial and assume that the probability of DLT is near 0.2 and that there is a 46% probability that the DLT rate is larger than 0.2. Such an assumption is called the 'prior probability distribution,' which in the Bayesian inference field is often represented by β distribution, a statistical distribution defined on (0, 1). In this example it can be expressed as $\beta(4, 16)$, where 4 and 16 are the parameter values needed to define β distribution $\beta(4, 16)$ can be viewed as reflecting the prior information from 20 patients (4 who experienced DLTs, 16 who did not) who were, for example, enrolled in a prior Phase I trial. As data are being collected from the ongoing Phase II trial, the 'posterior distribution' is computed, which combines the prior experience (prior distribution) from Phase I results as well as new Phase II trial data (likelihood function). For example, if five DLTs

are observed and 25 patients have completed the ongoing Phase II trial without a DLT, the posterior distribution of the probability of DLT is $\beta(4 + 5, 16 + 25)$, with a corresponding mean DLT probability of 0.18. The probability that the DLT rate is larger than 0.2 is now 33%. On the other hand, if 10 DLTs are observed among these 30 patients, the posterior distribution of the probability of DLT is $\beta(4 + 10, 16 + 20)$. In this case the DLT probability is estimated as 0.28, and the probability that the DLT rate is larger than 0.2 is now 90%. See [21], for example, for an illustration how the posterior distribution changes as more data become available.

Geller *et al.* proposed a Bayesian stopping rule for continuous monitoring of toxicity [22]. The trial is stopped if the posterior probability of the DLT rate exceeding θ_0 is equal to or higher than a pre-specified value τ. The rule is continuous as it checks whether or not the total number of observed DLTs is too high after DLT information is available on every new patient. In fact, we need only to check after each new DLT is observed. The value of τ is often chosen based on tradition, for example, 0.95 or 0.98 is commonly used. Lines 1 and 2 of Table 1 provide two Bayesian stopping boundaries for a trial of a total of K = 20 patients and τ = 0.98 for θ_0 = 0.2. A stopping boundary is described by a set of integers $b_1,\ldots, b_K$ such that the trial is stopped if there are b_k or more DLTs observed out of first k patients with complete toxicity follow-up. The prior distribution, the value of tolerable DLT probability θ_0, and the value of τ uniquely define the set of integers $b_1,\ldots, b_K$ that can be computed before the trial. To use the Bayesian boundary there is no need to compute the probability that the DLT rate is larger than θ_0 = 0.2 given the current data. Instead one can simply check if the number of observed DLTs in the first k patients is equal to or exceeds b_k. The boundary in line 1 uses the prior $\beta(0.6, 2.4)$, reflecting information from the total of 0.6 + 2.4 = 3 patients. The prior experience might reflect information from a three-patient dose cohort or from a six patient dose cohort of a Phase I trial. In the latter case, the prior experience was down-weighted from 6 to 3 patients, $\beta(0.6, 2.4)$ used instead of $\beta(1.2, 4.8)$, possibly because the Phase I population is different from the Phase II population, or the length of follow-up for toxicity in the Phase II trial is different from Phase I. The overall probability of stopping the trial for the boundary in line 1 of Table 1 when the DLT rate is equal to the acceptable rate of 0.2 is 0.038. The boundary in line 2 of Table 1 uses the prior $\beta(4, 16)$, which reflects information from 20 patients with an observed DLT probability of 4/20 = 0.20. Since we have strong prior information that the DLT probability is close to 0.20, stronger evidence is needed in the Phase II trial that the DLT probability is high to stop the trial compared with the first boundary. This is also reflected in the very small overall probability (0.004) of stopping the trial when the DLT probability is equal to the acceptable DLT rate frequently set to 0.2.

Another way to set up a Bayesian boundary is to specify the overall probability of stopping when the probability of toxicity is acceptable, instead of specifying τ. This is the frequentist type I error rate.

Table 1. Stopping boundaries for a trial with 20 patients with acceptable dose-limiting toxicity probability of θ_0 = 0.2†.

Number of patients, *k*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Bayesian boundary with prior $\beta(0.6, 2.4)$, τ = 0.98	–	–	3	4	4	5	5	5	6	6	6	7	7	7	7	8	8	8	9	9
Bayesian boundary with prior $\beta(4, 16)$, τ = 0.98	–	–	–	–	–	6	7	7	7	7	8	8	8	9	9	9	9	10	10	10
Bayesian boundary with prior $\beta(4, 16)$, τ = 0.91	–	–	–	4	5	5	5	5	6	6	6	6	7	7	7	7	8	8	8	8
Pocock boundary, type I error rate is 0.05	–	–	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	9	9
O'Brien–Fleming boundary, type I error rate is 0.05	–	–	–	–	–	6	6	6	6	6	6	7	7	7	7	7	7	8	8	8
Pocock two-stage boundary, type I error rate is 0.05	–	–	–	–	–	–	–	–	–	6	–	–	–	–	–	–	–	–	–	8
Pocock three-stage boundary, type I error rate is 0.05	–	–	–	–	–	5	–	–	–	–	–	–	7	–	–	–	–	–	–	8

†The trial is stopped after *k* patients if the number of observed dose limiting toxicities is equal to or higher than the corresponding value of the boundary.

Line 3 of Table 1 shows a boundary with the prior $\beta(4, 16)$, where the overall probability of stopping is fixed at 0.05 when the true toxicity probability is 0.2. This probability of stopping is achieved when $\tau = 0.911$. Defining the boundary based on the overall probability of stopping when the toxicity rate is acceptable, the type I error rate, is not a Bayesian approach but rather a frequentist concept. Nonetheless, this approach is often used because preserving the type I error rate is usually important irrespective of whether the approach used is Bayesian or frequentist.

The Pocock versus O'Brien–Fleming frequentist stopping boundaries

Over the years several frequentist sequential boundaries have been developed for use in group-sequential trials to stop early for efficacy. The two most frequently used boundaries are the O'Brien–Fleming [23] and the Pocock [18] boundaries (Table 1). The O'Brien–Fleming boundary achieves the higher power compared with the Pocock boundary for a given sample size and type I error rate. That is, when used for sequential monitoring of efficacy, the O'Brien–Fleming boundary yields a higher probability of declaring that the treatment is efficacious when the treatment is indeed effective as compared with the Pocock boundary, which is the reason it is used more often. Note that we are not suggesting that a Phase II study should be powered to detect a certain high DLT rate as resources allocated to a Phase II trials are usually limited. When occasionally used to monitor toxicity, the O'Brien–Fleming boundary yields the higher overall probability of stopping the trial compared with the Pocock boundary when the true DLT probability is higher than tolerable. However, the Pocock boundary allows stopping much earlier than the O'Brien–Fleming boundary, and therefore is used to stop the trial for adverse events or toxicity. For example, as shown in Table 1, if the Pocock boundary is used, the trial will be stopped if 3 DLTs are observed in the first 3 patients. In comparison, if the O'Brien–Fleming boundary is used, the earliest stopping point requires that the first six patients all experience DLTs. Both boundaries shown in Table 1 have the probability of stopping the trial of α. = 0.05, type I error rate, if the toxicity probability is tolerable, $\theta_0 = 0.2$.

The Pocock stopping rule can alternatively be described as repeated testing of the probability of toxicity after each patient completes toxicity follow-up, with the null hypothesis that the DLT probability is equal to $\theta_0 = 0.2$ and a type one error rate α'. The point-wise α-level, α', is much smaller than the overall type I error rate α and can be computed for a given α. In the Pocock stopping boundary in Table 1, point-wise $\alpha' = 0.0196$ corresponding to overall $\alpha = 0.05$. This is also equivalent to using a confidence interval approach [24]. The trial is stopped after enrollment of k patients if the lower bound of the $1 - \alpha'$ level one-sided confidence interval, or equivalently the lower bound of the $1 - 2\alpha'$ level two-sided confidence interval for the DLT probability computed when k patients completed the trial, is above $\theta_0 = 0.2$. Ivanova *et al.* gave a table of values α' for various sample sizes and tolerable DLT probability θ_0 [25]. Free software to generate the Pocock stopping boundary is available at [26]. For given K, θ_0 and α, the software computes the stopping boundary and important quantities that describe the boundary's performance. For several values of the true DLT probabilities the program computes the probability of stopping the trial and declaring that the therapy is too toxic, the average number of DLTs, and the average number of patients in the trial (Table 2). For example, when the probability of DLT is 0.4, about half of the trials will be stopped (probability of stopping is 0.55). The software also provides an example write-up that can be used in clinical trial protocols.

Comparison of Bayesian & frequentist stopping boundaries

In general, and as seen from Table 1, the Bayesian boundary with the β (0.6, 2.4) prior and $\tau = 0.98$ is almost indistinguishable from the Pocock boundary. Less informative priors, priors with low sum of values

Table 2. Operating characteristics of the Pocock Boundary with 20 patients, tolerable dose-limiting toxicity probability of $\theta_0 = 0.2$ and the type I error rate of 0.05.

True DLT probability	The probability of early stopping	Expected number of DLTs	Expected number of enrolled patients
0.2	0.05	3.9	19.5
0.4	0.55	5.8	14.6
0.5	0.83	5.4	10.8
0.6	0.97	4.7	7.8
0.8	1.00	3.6	4.5

DLT: Dose-limiting toxicity.

defining β distribution, for exampl, 0.6 + 2.4 = 3, yield a Bayesian boundary similar to the Pocock boundary as long as the two boundaries yield a similar overall probability of stopping. For a larger sum of values defining β distribution, informative prior, more DLTs are required to occur within the ongoing Phase II trial to recommend trial interruption. We need to observe more DLTs to stop the Phase II trial, because we need to 'override' the prior information that the toxicity rate is tolerable. In the example shown in Table 1, under the Bayesian rule with an informative prior regarding the probability of stopping of 0.05 (τ = 0.911), we stop later than under the Pocock boundary, but stop earlier than under the O'Brien–Fleming boundary. Another method that can be used for stopping due to toxicity is the sequential probability ratio test (SPRT) [27–29]. This method leads to a boundary very similar to the Pocock boundary for given sample size and given actual type I error rate.

If minimal prior information is available, either the Pocock boundary or the Bayesian boundary can be used. Even though they are described by a different statistical language, they are almost identical for a given total sample size and probability of stopping the trial for each value of the true DLT rate. If prior toxicity probability information is available, we recommend using the Bayesian boundary as this prior information can be reflected in the prior distribution for the toxicity rate.

As the Bayesian boundary with a slightly informative prior is very similar to the Pocock boundary when the two boundaries have the same probability of stopping the trial if $\theta = \theta_0$, frequentist software at [26] can be used to find an approximate value τ in the Bayesian boundary. An approximate value τ is equal to 1 minus the step-wise significance level α' corresponding to the probability of stopping, α, under the acceptable DLT probability $\theta = \theta_0$.

Use of stopping boundaries in published Phase II trials

To investigate the rate of reporting stopping rules for toxicity in oncology Phase II trials, we reviewed Phase II trials published in JCO, Annals of Oncology, Cancer, Clinical Cancer Research, and Lancet Oncology. We looked at publications in 2005 and 2010 to see if there is a five-year change in the trend of reporting stopping boundaries in oncology Phase II trials. A total of 291 articles were reviewed. Five trials used toxicity as a primary end point. Out of the remaining 286 trials with nontoxicity primary end point, stopping rules for toxicity were mentioned in 13 out of 286 trials (4.5%). Among 286 trials there was no difference in the frequency of reporting stopping rules for toxicity between year 2005 and 2010 with six out of 139 (4.3%) articles published in 2005 describing safety stopping rules versus seven out of 147 (4.8%) articles in 2010. We postulate that more than 4.5% of the trials have stopping rules for toxicity specified in the protocol but these rules are under-reported in the Patients and Methods section of the published clinical trials. In one trial we reviewed [30] three patients discontinued the study early (during cycle 1) due to treatment-related adverse events. Though a formal stopping rule was not specified in that trial, a high rate of adverse events led to the trial discontinuation after 13 patients were enrolled.

Rule 1 below had appeared four times and rule 2 three times in published Phase II trials we reviewed:

- 1. Two-stage design in which a trial can be stopped after stage 1 because of low antitumor response or high probability of toxicity [14].
- 2. Two-stage design where a trial can be stopped after stage 1 because of high toxicity [31–33].

The following rules appeared twice:

- 1. Stop the trial as soon as a certain number of DLTs are observed [34,35].
- 2. A multistage design with interim analysis for toxicity after every 10 patients (or after every 20 patients in the other published trial) [36,37].

Most of the trials we reviewed were single arm trials. There were some randomized noncomparative trials and a few randomized comparative trials among those reviewed. In multiarm trials toxicity monitoring can be performed in each arm separately as we do not expect that there is enough power to compare DLT probabilities across arms in a randomized Phase II trial. In one six-arm trial [38], the probability of DLT during the first cycle was compared across arms during an interim analysis to detect and possibly drop arms with high probability of DLT. In another trial [39], a combined primary end point, 'therapeutic success,' was used to take into account activity, toxicity, and compliance.

We prefer continuous stopping boundaries to two-stage boundaries or to boundaries with interim after, every 20 patients, for example. Consider a two-stage Pocock boundary with a probability of stopping of at most 0.05 when toxicity probability is 0.2. According to the two-stage boundary the trial is stopped if six or more DLTs are observed in the first 10 patients and if eight or more DLTs are observed in 20 patients, or equivalently point-wise α'= 0.0325 is used (line six, Table 2). This boundary is a two-stage counterpart of the Pocock continuous boundary (line 4 in Table 1). Due to discreteness of the binomial distribution the decision rules for a two-stage boundary in this example

are very similar to the corresponding rules in the continuous boundary: ≥6/10 and ≥8/20 for the two-stage and ≥6/10 and ≥9/20 for the continuous boundary, however, the continuous boundary allows stopping at many other points of the trial. The expected number of DLTs before the trial is stopped if the two-stage boundary is used for true DLT probabilities of 0.4, 0.5, 0.6 and 0.8 are 7.3, 8.1, 8.3 and 8.3 compared with 5.8, 5.4, 4.7 and 3.6 for the continuous boundary (Table 2). That is, three more DLTs are observed on average if the two-stage boundary is used. The expected number of enrolled patients for the two-stage boundary is 18.3, 16.2, 13.7 and 10.3 compared with 14.6, 10.8, 7.8 and 4.5 for the continuous boundary (Table 2). For a three-stage boundary the decision rules are ≥5/7, ≥7/14 and ≥8/20 (line 7, Table 1). The rules co-inside with the continuous boundary decision rules except at the last look. Multistage boundary allows exhausting the type I error rate much better and hence reducing the number of looks, as in two- or three-stage boundary, does not bring increased efficiency.

One boundary among those we found in published Phase II trials that we do not recommend is the rule when the trial is stopped as soon as *n* DLTs are observed [34,35]. Despite its simplicity, this rule does not take into account the denominator, that is, the number of patients enrolled in the study at the time of analysis. For example, for a trial with 20 patients and an acceptable DLT probability of 0.20, the two rules with the type I error rate closest to 0.05 are 'stop the trial when eight DLTs are observed' with the type I error rate of 0.032 and 'stop the trial when seven DLTs are observed' with the type I error rate of 0.087. Therefore, at least seven or eight DLTs are to be observed before the trial can be stopped. In comparison, the maximum expected number of DLTs under the Pocock boundary is 5.8 when the true DLT probability is 0.4 and less for lower or higher DLT probabilities.

The 'no denominator' boundary is often used when the acceptable DLT probability is low, for example, $\theta_0 = 0.05$. According to the Pocock boundary for a study of up to 20 patients, $\theta_0 = 0.05$ and $\alpha = 0.05$, the trial will be stopped if two DLTs are observed in the first 2–5 patients, or three DLTs are observed in first 6–14 patients, or four DLTs are observed in more than 14 patients. This boundary yields fewer expected DLTs than a constant boundary, in which case the trial is stopped as soon as three DLTs are observed. Therefore, we recommend using the Pocock or the Bayesian boundary, but not the constant boundary.

Conclusions & recommendations for the clinical trialist

We have reviewed several stopping rules for toxicity that have been infrequently used in Phase II trials. We propose to keep the probability of stopping the trial when the DLT probability is equal to the acceptable DLT rate at 0.05 or lower. In such cases, the trial can be stopped early for toxicity only if there is strong evidence that the DLT probability is high. The goal is to stop the trial if there is 'strong' evidence of a high DLT rate. The term 'strong' implies that stopping rules for toxicity should by no means be extremely conservative to the point that they overshadow the main purpose of the Phase II study, namely the efficacy assessment. Under this concept we anticipate that stopping rules for toxicity would be enabled before efficacy rules only infrequently. In addition, we argue that a continuous stopping rule for toxicity should be used, that is, a rule that allows stopping the trial at any point. Between the two stopping boundaries most commonly used in clinical trials, the O'Brien–Fleming boundary and the Pocock boundary, we recommend the Pocock boundary since it allows stopping for toxicity as early as possible.

If the investigator does not wish to use prior information about the toxicity probability in Phase II (e.g., the demographics, comorbid factors, and/or cancer type of subjects treated as part of the Phase I study are significantly different from those of patients treated under the Phase II), the Pocock or the Bayesian boundary with a noninformative prior can be used, as they are virtually identical under these conditions. If there is reliable prior information about the toxicity probability to use in the stopping rule, we recommend using the Bayesian boundary since it is the only boundary that takes into account prior information about toxicity. On the other hand, prior information must be used with caution as various factors including potential differences in the patient population currently under study compared with previous cohorts might affect the DLT probability of the investigational treatment.

All methodologies pertaining to continuous toxicity monitoring that were discussed here have limitations. For example, the boundaries we mentioned here require complete-mature follow-up DLT data for all patients enrolled. Therefore, these methods might not be appropriate for trials that require long-term follow-up to observe DLTs or when patients are still under study and/or early in their treatment, in which case insufficient information is available. Several methods have been developed for trials with long follow-up for toxicity [40–41]. In addition, for trials with an unrestricted accrual rate, a significant number of patients could be potentially enrolled within a short period of time, which might result in a significant number of DLTs before discussion by the Data Monitoring Committee (DMC) for the study. In that case a

stopping rule alone might not prevent observing excessive toxicity. For clinical trials in oncology where toxicities can be both life-threatening and/or take time to develop (e.g., subacute encephalopathy in clinical trials using investigational agents plus whole brain radiation therapy in patients with active brain metastases), we recommend statistical designs that consider both stopping and enrollment rules [42]. An enrollment rule guides the accrual rate of patients, not allowing the enrollment of many patients at once when not much is known about the DLT probability in the trial or when there is evidence that the DLT probability can be high.

All Phase II clinical trials are monitored by the clinical team and many trials are also monitored by a DMC. For example, if a trial is initiated by an investigator at an academic cancer center, the trial is monitored by the center's DMC. Many DMCs review aggregate toxicity data. However, if aggregate data are not provided, a pattern of increased toxicity may not emerge. This is why it is important to have a stopping rule for toxicity that requires aggregate data to apply the rule. A formal stopping rule for toxicity can serve is a useful reference for a DMC when reviewing the totality of toxicity data. Having a formal stopping rule might not increase the frequency of stopping a trial when toxicity is high but such trials might be stopped earlier by the DMC or by the investigator.

Future perspective

Continuous boundary to monitor toxicity in a Phase II oncology trial was described in [22] and [25] 10 years ago. Since then, more and more Phase II oncology trials utilize stopping rules for toxicity. Now that free and easy to use software to generate the boundary is available more researchers are implementing stopping for toxicity in thier protocols. We believe that adoption of stopping rules for toxicity will continue to accelerate.

Acknowledgements

The authors thank M Gibson (University of Pittsburgh Medical Center), P Catalano (ECOG-ACRIN Gastrointestinal Committee Statistician) and L Kleinberg (Johns Hopkins University) for discussing the results of the E2205 study. The authors also thank S Middleton for editorial assistance and A Snavely for helpful comments.

Financial & competing interests disclosure

S Moschos grant support is University Cancer Research Funds. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Executive summary

Background

- Due to small sample sizes in oncology Phase I trials, the toxicity profile of the dose used in a Phase II trial might not be well established. A rigorous stopping rule for toxicity should be used in a Phase II trial to stop early in case the number of observed toxicities is higher than expected. The stopping rule is continuous if it checks whether or not the total number of observed toxicities is too high every time a new toxicity is observed.

Bayesian stopping rule

- A Bayesian stopping rule requires specifying a prior distribution of the probability of toxicity. The prior might be obtained from a preceding Phase I trial. Geller *et al.* [22] proposed a continuous Bayesian stopping boundary for toxicity.

The Pocock versus O'Brien–Fleming frequentist stopping boundaries

- The Pocock and O'Brien–Fleming are the two most frequently used sequential boundaries. The Pocock boundary allows stopping as early as possible and the O'Brien–Fleming yields the highest probability of stopping. Due to the desire to stop the trial as early as possible when the toxicity probability is high, the Pocock boundary is recommended for use in the stopping rule for excessive toxicity.

Comparison of Bayesian and frequentist stopping boundaries

- Even though Bayesian and frequentist boundaries are described by a different statistical language, they often result in almost identical or very similar stopping rules given the total sample size and the overall probability of stopping the trial for a given value of the true DLT probability.

Use of stopping boundaries in published Phase II trials

- Stopping boundaries for toxicity are rarely reported when Phase II trials are published. Most reported boundaries were two-stage. Due to discreetness of binomial distribution, a two- or three-stage testing does not exhaust the allowable type I error and therefore multistage (or continuous) boundaries are more efficient overall.

Conclusions and recommendations for the clinical trialist

- We recommend using a continuous stopping boundary to monitor toxicity in a Phase II oncology trial. A Bayesian continuous boundary can be used if there is a need to utilize prior information, otherwise a frequentist boundary based on the Pocock method is recommended.

References

Papers of special note have been highlighted as:
• of interest; •• of considerable interest

1 Rubinstein L, Crowley J, Ivy P *et al.* Randomized Phase II designs. *Clin. Cancer Res.* 15, 1883–1890 (2009).

2 Hoering A, LeBlanc M, Crowley J. Seamless Phase I-II trial design for assessing toxicity and efficacy for targeted agents. *Clin. Cancer Res.* 17, 640–646 (2001).

3 El-Maraghi RH, Eisenhauer EA. Review of Phase II trial designs used in studies of molecular targeted agents: outcomes and predictors of success in Phase III. *J. Clin. Oncol.* 26, 1346–1354 (2008).

4 Simon R. Optimal two-stage designs for Phase II clinical trials. *Control. Clin. Trials* 10, 1–10 (1989).

• **Most frequently used design to monitor efficacy in Phase II oncology trials.**

5 Ensign LG, Gehan EA, Kamen DS *et al.* An optimal three-stage design for Phase II clinical trials. *Stat. Med.* 13, 1727–1736 (1994).

6 Berry DA. Adaptive clinical trials in oncology. *Nat. Rev. Clin. Oncol.* 9, 199–207 (2012).

7 Maki RG, D'Adamo DR, Keohan ML *et al.* Phase II study of sorafenib in patients with metastatic or recurrent sarcomas. *J. Clin. Oncol.* 27, 3133–3140 (2009).

8 Alberts SR, Fitch TR, Kim GP *et al.* Cediranib (AZD2171) in patients with advanced hepatocellular carcinoma: a Phase II North Central Cancer Treatment Group Clinical Trial. *Am. J. Clin. Oncol.* 35, 329–333 (2012).

9 D'Adamo DR, Anderson SE, Albritton K *et al.* Phase II study of doxorubicin and bevacizumab for patients with metastatic soft-tissue sarcomas. *J. Clin. Oncol.* 23, 7135–7142 (2005).

10 Wyman K, Atkins MB, Prieto V *et al.* Multicenter Phase II trial of high-dose imatinib mesylate in metastatic melanoma: significant toxicity with no clinical efficacy. *Cancer* 106, 2005–2011 (2006).

11 Jin H. Alternative designs of Phase II trials considering response and toxicity. *Contemp. Clin. Trials* 28, 525–531 (2007).

12 Thall PF, Cheng SC. Treatment comparisons based on two-dimensional safety and efficacy alternatives in oncology trials. *Biometrics* 55, 746–753 (1999).

13 Conaway MR, Petroni GR. Bivariate sequential designs for Phase II trials. *Biometrics* 51, 656–664 (1995).

14 Bryant J, Day R. Incorporating toxicity considerations into the design of two-stage Phase II clinical trials. *Biometrics* 51, 1372–1383 (1995).

15 Tournoux C, De Rycke Y, Medioni J *et al.* Methods of joint evaluation of efficacy and toxicity in Phase II clinical trials. *Contemp. Clin. Trials* 28, 514–524 (2007).

16 Dawson NA, Halabi S, Ou SS *et al.* A Phase II study of estramustine, docetaxel, and exisulind in patients with hormone- refractory prostate cancer: results of cancer and leukemia group B trial 90004. *Clin. Genitourin. Cancer* 6, 110–116 (2008).

17 Gibson MK, Catalano PJ, Kleinberg L *et al.* E2205: A Phase II study to measure response rate and toxicity of neoadjuvant chemoradiotherapy (CRT) with oxaliplatin (OX) and infusional 5-fluorouracil (5-FU) plus cetuximab (C) followed by postoperative docetaxel (DT) and C in patients with operable adenocarcinoma of the esophagus (abstr 4064). *J. Clin. Oncol.* 28(Suppl. 15), Abstract 4064 (2010).

18 Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 64, 191–199 (1977).

19 Storer BE. Design and analysis of Phase I clinical trials. *Biometrics* 45, 925–937 (1989).

20 Reiner E, Paoletti X, O'Quigley J. Operating characteristics of the standard Phase I clinical trial design. *Comput. Statist. Data Anal.* 30, 303–315 (1999).

21 Berry D. Bayesian clinical trials. *Nat. Rev. Drug Discov.* 5(1), 27–36, (2006).

22 Geller NL, Follmann DF, Leifer ES *et al.* Design of early trials in peripheral blood stem cell transplantation: a hybrid frequentist-Bayesian approach. In: *Advances in Clinical Trial Biostatistics.* Geller NL (Ed.). Marcel Dekker, New York and Basel, 40–52 (2005).

•• **Description of Bayesian stopping boundaries for toxicity in Phase II oncology trials.**

23 O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 35, 549–556 (1979).

24 Clopper C, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26, 404–413 (1934).

25 Ivanova A, Qaqish BF, Schell MJ. Continuous toxicity monitoring in Phase II trials in oncology. *Biometrics* 61, 540–546 (2005).

•• **Description of frequentist stopping boundaries for toxicity in Phase II oncology trials.**

26 Software: Continuous monitoring for toxicity using Pocock-type boundary. http://cancer.unc.edu

27 Wald A. Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics* 16, 117–186 (1945).

28 Armitage P. *Sequential Medical Trials* (2nd Edition). John Wiley and Sons, NY, USA (1975).

29 Goldman AI. Issues in designing sequential stopping rules for monitoring side effects in clinical trials. *Control. Clin. Trials* 8, 327–337 (1987).

30 Haura EB, Ricart AD, Larson TG *et al.* A Phase II study of PD-0325901, anoral MEK inhibitor, in previously treated patients with advanced non-small cell lung cancer. *Clin. Cancer Res.* 16(8), 2450–2457 (2010).

31 Penson RT, Dizon DS, Cannistra SA *et al.* Phase II study of carboplatin, paclitaxel, and bevacizumab with maintenance bevacizumab as first-line chemotherapy for advanced mullerian tumors. *J. Clin. Oncol.* 28(1), 154–159, (2010).

32 Krug LM, Miller VA, Patel J *et al.* Randomized Phase II study of weekly docetaxel plus trastuzumab versus weekly paclitaxel plus trastuzumab in patients with previously untreated advanced nonsmall cell lung carcinoma. *Cancer* 104(10), 2149–2155 (2005).

33 Elter T, Borchmann P, Schulz H *et al.* Fludarabine in combination with alemtuzumab is effective and feasible

in patients with relapsed or refractory B-cell chronic lymphocytic leukemia: results of a Phase II trial. *J. Clin. Oncol.* 23(28), 7024–7031 (2005).

34 Shusterman S, London WB, Gillies SD *et al.* Antitumor activity of hu14.18-IL2 in patients with relapsed/refractory neuroblastoma: a Children's Oncology Group (COG) Phase II study. *J. Clin. Oncol.* 28(33), 4969–4975 (2010).

35 Li J, Juliar B, Yiannoutsos C *et al.* Weekly paclitaxel and gemcitabine in advanced transitional-cell carcinoma of the urothelium: a Phase II Hoosier Oncology Group study. *J. Clin. Oncol.* 23(6), 1185–1191 (2005).

36 Ross HJ, Blumenschein GR Jr, Aisner J *et al.* Randomized Phase II multicenter trial of two schedules of lapatinib as first- or second-line monotherapy in patients with advanced or metastatic non-small cell lung cancer. *Clin. Cancer Res.* 16(6), 1938–1949 (2010).

37 Brada M, Ashley S, Dowe A *et al.* Neoadjuvant Phase II multicentre study of new agents in patients with malignant glioma after minimal surgery. Report of a cohort of 187 patients treated with temozolomide. *Ann. Oncol.* 16(6), 942–949 (2005).

38 Jacobs S, Fox E, Krailo M *et al.* Phase II trial of ixabepilone administered daily for five days in children and young adults with refractory solid tumors: a report from the children's oncology group. *Clin. Cancer Res.* 16(2), 750–754 (2010).

39 Ardizzoni A, Favaretto A, Boni L *et al.* Platinum-etoposide chemotherapy in elderly patients with small-cell lung cancer: results of a randomized multicenter Phase II study assessing attenuated-dose or full-dose with lenograstim prophylaxis--a Forza Operativa Nazionale Italiana Carcinoma Polmonare and Gruppo Studio Tumori Polmonari Veneto (FONICAP-GSTPV) study. *J. Clin. Oncol.* 23(3), 569–575 (2005).

40 Follmann D, Albert PS. Bayesian monitoring of event rates with censored data. *Biometrics* 55, 603–607 (1999).

41 Rosner GL. Bayesian monitoring of clinical trials with failure-time end points. *Biometrics* 61, 239–245 (2005).

42 Song G, Ivanova A. Frequentist enrollment and stopping rules for managing toxicity requiring long follow-up in Phase II oncology trials. *J. Biopharm. Statistics.* (2015) (In press).

- **Description of enrollment rules as well as frequentist stopping boundaries for toxicity in Phase II oncology trials.**