Performance of multinomial designs in comparison with response-based designs in non-randomized phase II trials of targeted cancer agents

R. Jamal1*, R. A. Goodwin2, D. Tu3, W. Walsh3, D. Lacombe4 & E. A. Eisenhauer5

1Department of Hematology-Oncology, Notre-Dame Hospital, CHUM, University of Montreal, Montreal; 2General Division, Ottawa Health Research Institute, Ottawa; 3NCIC Clinical Trials Group, Queen’s University, Kingston, Canada; 4European Organisation for Research and Treatment of Cancer, Brussels, Belgium; 5Department of Oncology, Kingston General Hospital, Queen’s University Cancer Centre of Southeastern Ontario, Kingston, Canada

Received 19 October 2012; revised 25 January 2013; accepted 13 February 2013

Background: In phase II trials of cytotoxic agents, a multinomial phase II design incorporating early progression and response end points was shown to perform more efficiently than designs based only on response. We undertook a study to evaluate the performance of these designs in trials of targeted agents using the actual phase II data.

Patients and methods: Using best response data from sequentially enrolled patients in 15 NCIC Clinical Trials Group and 7 European Organization for Research and Treatment of Cancer trials of targeted agents, we determined that trials would have been stopped at the end of stage I of accrual by applying rules generated by the multinomial and Fleming designs. Two variants of the multinomial design were studied: to stop accrual after stage I of enrolment, Variant A required either response or progression criteria to be met, whereas Variant B required that both response and progression criteria to be met.

Results: Using early progression, null/alternate hypotheses of 60% and 40% (60/40), the multinomial A variant recommended early stopping more often than the Fleming design. In most of the cases, this recommendation was correct given the final trial outcome. In contrast, the multinomial B variant never led to recommendations for early stopping and changing progression hypotheses did not improve the performance of this design.

Conclusions: The multinomial A design using 60/40 hypotheses carried out better than the Fleming design in appropriately stopping trials of inactive targeted agents early. The multinomial B design was not useful for early stopping decisions. The multinomial A design may be favored over response-based designs in phase II trials of targeted agents.

Key words: clinical trials, multinomial design, phase II designs, targeted agents

introduction

Single-agent phase II studies of new cancer agents have traditionally employed objective tumor response as the primary end point to guide decisions on the activity of the agent. Such trials are often non-randomized and use a multistage design such as those of Gehan, Fleming and Simon [1–4] to terminate the trial after the first stage of accrual if too few responses are seen. The stopping rule and sample size of the trial are determined by the null (Ho) and alternate hypotheses (Ha) established for response and the alpha (false-positive) and beta (false-negative) error rates. These parameters will vary based on the tumor type, the patient population and the type of drug being evaluated.

Recently, early disease progression has been suggested to be an additional measure of inactivity in an innovative multinomial design [5]. This design presumes that a truly active treatment may be detected not only by evidence of response, but also by the observation of a low proportion of patients having early disease progression. Thus, information on both the number of observed responses and the number of early progression events is considered in decisions about early stopping and conclusions about activity. Zee et al. [5] defined early progression as a best overall response of progressive disease (PD), meaning that those patients had progressed at, or
before, the first scheduled follow-up evaluation on the trial. Trials would stop early for high numbers of early progressions and too few or no responses. In a comparison of the performance of the multinomial stopping rule developed by Zee et al. [5] with the Fleming and Gehan designs in a series of phase II trials evaluating new cytotoxic agents, the multinomial design was found to be more efficient in recommending the early trial closure of what were eventually found to be ineffective agents [6].

In the last 10 years, the majority of new drugs entering cancer clinical trials have been molecular-targeted agents. Because of the differing mechanism of action of these agents, some argued that stabilization of disease (non-progression) might be a more relevant end point than objective response in phase II designs [7–10].

The multinomial design could theoretically temper the importance given to response in decision-making in such trials. Indeed, the Methodology for the Development of Innovative Cancer Therapies task force recently stated that a multinomial design should be considered for use in phase II studies of targeted drugs, particularly when responses rates are anticipated to be low [11].

To investigate the performance of the multinomial design in phase II trials of targeted agents in comparison with a response-based design, we applied both types of design to actual data from a series of phase II studies.

materials and methods

trial data

Individual patient response results were obtained from 15 completed NCIC Clinical Trials Group (CTG) phase II trials of targeted agents (Table 1). All 15 trials were conducted using a standard two-stage Fleming design (Ho: proportion responding \( \leq 5\% \) versus Ha: \( \geq 20\% \)). For trials I and O, the hypotheses were Ho: \( \geq 10\% \) versus Ha: \( \geq 30\% \). In addition, through a collaboration with the European Organization for Research and Treatment of Cancer (EORTC), we obtained data from seven phase II EORTC trials evaluating targeted agents (Table 2). Three of these trials used an objective response end point in the original design of the study and the rest of the trials used progression-free survival, but all had collected objective response information. Because several EORTC trials evaluated the same targeted agent in different tumor types, we compiled 13 treatment arms that were analyzed as individual trials (considered as trials 1–13 in Table 2).

For each study, we obtained the date of patient enrolment and their best objective response to treatment on study.

designs evaluated

multinomial and Fleming

The multinomial two-stage design has been previously described in detail [5]. It is based on generating a combination of hypotheses for response and progression identifying the values for the null and alternate hypotheses for both variables based on the rates of tumor response and progression deemed to be of interest in the particular tumor setting. Two variants are possible: in one (multinomial A), stopping at the end of stage I of accrual occurs if either response or progression criteria are met. In multinomial B, stopping at the end of stage I requires that both the response and progression criteria for stopping are met. In the Fleming two-stage design [2], null and alternate hypotheses for response are set and generate a stopping rule based only on response observations in the prespecified number of first stage patients. Because of our familiarity with this design, it was selected for evaluation in this project. The Simon two-stage design is based on similar principles and yields similar decision rules [3].

design parameters: hypotheses and error rates

As summarized in Table 3, similar response null and alternate hypotheses were employed for all designs tested. The multinomial designs tested incorporated null and alternate hypotheses for progression that included a variety of paired values as presented in Table 3. The alpha and beta used for the derivation of the stopping rule were 0.1 for all cases considered. The primary comparisons were between Fleming design and the multinomial A and B designs using progression hypotheses of Ho 60% and Ha 40%. Following the results of these comparisons, we explored the impact on the conclusions of the study using the other progression hypotheses summarized in Table 3. Each design generated rules for stopping (accepting the null hypotheses) or acceptance of agent as interesting (accepting the alternate hypotheses) based on the observations of the number of responses and early progression events observed in a minimum number of accruals.

application of design rules to actual clinical trial data

We applied the rules for stopping trials at the accrual point representing stage I of accrual and at the end of study (if data were available) using Fleming, multinomial A and B designs to the NCIC CTG and EORTC single-agent phase II trials. The best overall response to therapy (complete response, partial response, stable disease or PD) was compiled. For the purpose of the multinomial designs, early progression was considered to have occurred if the best response to therapy was PD. Sequential patient data were reviewed in each trial to tabulate a STOP/GO decision at the end of stage I of accrual and a REJECT/ACCEPT

### Table 1. NCIC CTG phase II trials of single-agent targeted therapies

<table>
<thead>
<tr>
<th>Trials</th>
<th>NCIC CTG phase II single-agent targeted therapy trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>ZD6474 in recurrent/relapsed multiple myeloma</td>
</tr>
<tr>
<td>B</td>
<td>MG98 in renal cell carcinoma</td>
</tr>
<tr>
<td>C</td>
<td>Flavopiridol in untreated metastatic malignant melanoma</td>
</tr>
<tr>
<td>D</td>
<td>Flavopiridol in untreated metastatic or locally advanced soft tissue sarcoma</td>
</tr>
<tr>
<td>E</td>
<td>Sunitinib in locally advanced or metastatic cervical carcinoma</td>
</tr>
<tr>
<td>F</td>
<td>Sunitinib in relapsed or refractory diffuse or mediastinal large B-cell lymphoma</td>
</tr>
<tr>
<td>G</td>
<td>CGP 64128A (ISIS 3521) in locally advanced or metastatic colorectal cancer</td>
</tr>
<tr>
<td>H</td>
<td>CGP 69846A (ISIS 5132) in locally advanced or metastatic colorectal cancer</td>
</tr>
<tr>
<td>I</td>
<td>CCI-779 in metastatic and/or locally advanced recurrent endometrial cancer</td>
</tr>
<tr>
<td>J</td>
<td>BB-2516 in malignant melanoma and cutaneous metastases</td>
</tr>
<tr>
<td>K</td>
<td>Sunitinib in advanced ovarian, primary fallopian tube or peritoneal cancer</td>
</tr>
<tr>
<td>L</td>
<td>BAY 43-9006 in hormone refractory prostate cancer</td>
</tr>
<tr>
<td>M</td>
<td>Flavopiridol in untreated or relapsed mantle cell lymphoma</td>
</tr>
<tr>
<td>N</td>
<td>PS-341(bortezomib) in mantle cell lymphoma</td>
</tr>
<tr>
<td>O</td>
<td>PS-341 (bortezomib) in Waldenström's macroglobulinemia</td>
</tr>
</tbody>
</table>
decision on the regimen’s activity at the end of the study based on each design. A comparison of the performance of the designs was based on whether the recommendations to close accrual early were correct as determined by the final trial observations (when available; see example of process in Table 4 using trial J).

results

NCIC CTG trials: performance of the Fleming versus multinomial A and B stopping rules

When the stopping rules generated by the various designs were applied to 15 NCIC CTG phase II trials of targeted agents, the Fleming design recommended continuing seven trials (I–O) to stage 2 of accrual, while the multinomial A 60/40 design would have stopped all but three of these (M–O, Table 5). The recommendation at the end of the first stage was concordant in 11 of the 15 trials (in eight (A–H), recommendation was to stop and in three (M–O), recommendation was to continue accrual; Table 5). In the four trials with discordant recommendations (I–L), all would have been continued using the Fleming design and all stopped using multinomial A. Because these trials were actually conducted using a Fleming design, all seven trials also had second stage accrual data. Of these, only two were concluded to be positive by the Fleming design (N, O), whereas those same two trials plus an additional three studies (total 5, K–O) met criteria to consider the agent active by the multinomial A design—interestingly, this included two studies that the multinomial A had recommended for early stopping.

In contrast, when the multinomial B 60/40 design parameters were evaluated, none of the 15 trials met criteria for stopping after the first stage of accrual. As summarized in Table 5, changing progression hypotheses to 50/30 and 40/20 had the effect of increasing the probability of early stopping for both multinomial designs such that, for the 40/20 hypotheses, all trials were stopped early in multinomial A and for multinomial B, 4 of the 15 trials were stopped early (B, F–H), all of which were in the same subset of trials stopped early by the original Fleming design.
EORTC trials: performance of the Fleming stopping rule versus the multinomial A and B

The seven EORTC trials included 13 different arms of treatment, which were treated as individual trials for the purpose of analysis. Overall, the Fleming design recommended continuing seven trials (trials 7–13) to stage 2 of accrual, while the multinomial A 60/40 recommended only two of the trials continue (trials 12 and 13, Table 6). In comparing the early stopping recommendation between the two designs, the recommendation at the end of the first stage was concordant in 8 of the 13 trials: in six (trials 1–6), the recommendation was to stop and in two (trials 12 and 13), the recommendation was to continue accrual (Table 6). Similar to the NCIC CTG trial experience, the five discordant recommendations (trials 7–11) were all to continue accrual using the Fleming design parameters, but to stop accrual using the multinomial A 60/40 design. Because the EORTC trials actually did continue to accrue, we can compare the final interpretation of results in those studies using both designs. Of the seven trials, moving that continued to stage 2 on the basis of Fleming design parameters, but to stop accrual using the multinomial A 60/40 criteria to be considered positive. This included two studies (10 and 11) that the multinomial A design would have stopped early.
Again, as in the NCIC CTG cohort, the stopping rule derived from the multinomial B 60/40 design never led to early trial termination. As summarized in Table 6, changing progression hypotheses to 50/30 and 40/20 had very little effect on increasing the probability of early stopping using multinomial A, while it increased the number of trials stopping early using multinomial B.

**discussion**

Single-agent phase II trials in oncology have historically used a response-based primary end point. With the introduction of targeted molecular agents into the clinic, several reviews have reported on the need to consider the use of other primary end points in these trials, such as progression-free survival or the multinomial end point [7–13]. Limitations of the response rate as an endpoint have been well described in hepatocellular and renal cell carcinomas, where sorafenib, despite response rates of <10%, has led to significant prolongation of both progression-free survival and overall survival [8, 9, 14, 15].

A recent review of 19 targeted drugs being evaluated in a series of 89 single-agent phase II trials showed that objective response was the primary end point used to design 57% of the trials, while a multinomial end point or progression-free survival end point were each used in ~10% of trials [13]. In an another review of 351 phase II studies on targeted agents, 7.4% and 75.9% of the phase II studies were conducted with a progressive disease and response end point, respectively [12]. Certainly, there remains controversy about what the primary end point should be in phase II trials evaluating targeted agents.

There has been little in the way of evaluation of the performance of single-arm phase II designs using response or a multinomial end point using real data generated from actual phase II studies. We previously reported on this comparison within a series of cytotoxic agent phase II trials and have now applied the same approach to a series of targeted single-agent NCIC CTG and EORTC studies. We have shown that a subset of nine trials continued using the Fleming design would have been stopped had the multinomial A design with progression hypotheses of 60% (null) and 40% (alternate), been used to conduct the studies. This is consistent with the fact that multinomial A is designed in such a way that either failure to achieve the response criteria or the progression criteria for continuing can lead to early termination. Thus, it is very sensitive to any evidence of the agent showing signals of failure.

In the NCIC CTG data, the final outcome of the four discordant trials supported the multinomial A 60/40 design recommendations, since none showed particularly interesting drug activity at the end of accrual (trials I–L). Paradoxically, two of these studies 'technically' met the multinomial A level of interest at the end of accrual through the relatively low rates of early PD seen in the final dataset (trials K and L). Indeed, in the NCIC CTG cohort, all the trials that would have stopped at the first stage of accrual by multinomial A 60/40 were found to have no, limited, or at best modest activity at the end of the trial (Table 5).

Observations in the EORTC studies paralleled those of the NCIC CTG studies, i.e. multinomial A 60/40 would have stopped early five trials that would have been continued by applying the Fleming derived stopping rules (trials 7–11, Table 6). In three of these trials, the final outcome of the trials suggested no activity and thus supported the multinomial A 60/40 design recommendations for early stopping (trials 7–9, Table 6). In the remaining two discordant trials, the published data concluded that there was some modest but interesting activity of the agent (trials 10 and 11). Trials 10 and 11 showed interesting activity of pazopanib in some types of soft tissue sarcoma (synovial sarcoma and other types of soft tissue sarcoma) at the end of the study, but according to the multinomial A design, both of these trials would have been stopped at the first stage of accrual [16]. Based on this interesting activity at the end of the phase II EORTC trial,
Pazopanib was evaluated recently in a randomized, double-blind, placebo-controlled phase III trial that included leiomyosarcoma, other sarcoma and synovial sarcoma [16]. Pazopanib significantly increased progression-free survival compared with placebo, but no significant benefit in terms of overall survival was seen. The progression rate in this phase III trial was lower than the one documented in the EORTC phase II trial, but obviously the population and eligibility criteria were not the same making comparisons difficult [16, 17]. Had the EORTC phase II trial been conducted with a multinomial A design, both trials 10 and 11 would have closed after stage I of accrual and no further phase III trial would have been probably conducted. The only two trials given the go-ahead at the first stage of accrual in the EORTC cohort by the multinomial A design were trials 12 and 13 [18, 19]. Both of these trials were enriched for a molecular biomarker predicting response and thus, it is not surprising to document low progression rates and high response rates in trials 12 and 13.

Imatinib has been approved in c-Kit-mutated GIST and DermatoFibroSarcoma Protuberans (DFSP) since then.

In both NCIC CTG and EORTC series, the performance of multinomial A using lower progression hypotheses (50/30) was similar to multinomial A 60/40. By lowering the levels of early progression rates further to 40/20, multinomial A stopped all NCIC CTG trials early and most EORTC trials early. Thus, it seemed that multinomial A 40/20 could have the potential to miss active agents and is not recommended further.

In contrast to the data from multinomial A 60/40, the multinomial B variant, which would continue accrual if either response or progression criteria showed favorable results, did not appear very useful as an approach to early termination of the trials of inactive agents, since all but one trial continued to stage 2 under its stopping parameters. Reducing progression hypotheses to 50/30 and 40/20 improved the performance of the multinomial B variant somewhat but not enough to consider it a useful approach.

The findings of this study of trials of targeted agents are similar to those described in a comparison of the performance of the multinomial stopping rule (variant A) versus the Fleming rule in 23 NCIC CTG and EORTC studies of cytotoxic studies of cytotoxic and molecular markers. In a target-enriched population where the multinomial A design carried out better in identifying inactive agents early and the performance of the multinomial B variant somewhat but not enough to consider it a useful approach.

This study has some limitations. Certainly, when assessing if the final outcome of the trial supported rejection of the drug by the multinomial A design in the first stage of accrual, we drew conclusions about the ‘appropriateness’ of the recommendations based on the conclusions of the published trial—which were often based on other measures (response or PFS). The Fleming design was chosen as a comparator in our series, because in the NCIC CTG cohort all 15 trials were initially designed using the Fleming two-stage design. Because the principles behind the Fleming and the Simon designs are similar, using the Fleming design, as a comparator made it is unlikely that other two-stage designs using response as an end point would have offered substantially different conclusions when compared with the multinomial design. Finally, our analysis assumes that one of either the Fleming or the multinomial will provide ‘correct’ recommendations for drug development and in fact neither may do so based on subsequent phase III results.

In this review, we evaluated two variants of the multinomial design, each with different hypotheses we thought to be reasonable in the clinical setting. However, in using the multinomial design in phase II trials, the hypotheses would need to be adapted to each clinical setting and should be based on the level of activity of the compound in that setting which would be deemed to be of interest. Related to this, most of the trials in both series were not enriched for a molecular target—only two EORTC trials had populations determined by molecular markers. In a target-enriched population where the target is confirmed to be relevant to activity by earlier studies, one would expect the bar to be set higher for a new drug than the typical null hypothesis level of a 5% response rate.

The strength of this review is that it is focused on actual trial data from a range of targeted agents, which more closely reflect how various designs perform in reality. The failure of multinomial B to ever stop a trial when used with original 60/40 hypotheses suggests that it does not function as intended as a two-stage design.

In conclusion, in both NCIC CTG and EORTC studies, the multinomial A and Fleming designs carried out well in rejecting obviously inactive agents. Certainly, response as end point is still widely used in both studies evaluating cytotoxic and targeted therapies but as shown in the NCIC CTG series, there were also clear cases where the multinomial A design carried out better in identifying inactive agents early in the study as supported by the final outcome of those trials. In the EORTC set, Fleming was again shown to be permissive compared with multinomial A. Interestingly, in both series of trials, the multinomial A design generally carried out best with the original hypotheses of $H_0: p \geq 60\%$ versus $H_a: p \leq 40\%$ for progression, while the multinomial B design was ineffective in identifying active agents early. Thus, in choosing between the Fleming versus multinomial A, one must choose the nature of the error one wants to make at the first stage of accrual. The multinomial A 60/40 design might lead to false-negative signals in activity (type II error). Because there is little chance that these agents will be further pursued in drug development, this could result in potentially active agents being permanently disregarded. However, in the context of restricted resources, stopping trials earlier when the signal for efficacy is very low could also be a reason to choose a design using dual end points instead of a pure response end point.

**funding**

Drs Jamal and Goodwin carried out this work as part of a Drug Development Fellowship at the NCIC Clinical Trials Group funded by the Queen’s University Terry Fox Foundation Training Program in Transdisciplinary Cancer Research in partnership with the Canadian Institutes of Health Research (no grant number).

**disclosure**

The authors have declared no conflicts of interest.
Non-inferiority cancer clinical trials: scope and purposes underlying their design

R. P. Riechelmann*, A. Alex, L. Cruz, G. M. Bariani & P. M. Hoff

Discipline of Radiology and Oncology, Universidade de Sao Paulo, Instituto do Câncer de Estado de São Paulo, Sao Paulo, Brazil

Received 7 December 2012; revised 26 January 2013; accepted 28 January 2013

Background: Non-inferiority clinical trials (NIFCTs) aim to demonstrate the experimental therapy has advantages over the standard of care, with acceptable loss of efficacy. We evaluated the purposes underlying the selection of a non-inferiority design in oncology and the size of their non-inferiority margins (NIFm’s).

Patients and methods: All NIFCTs of cancer-directed therapies and supportive care agents published in a 10-year period were eligible. Two investigators extracted the data and independently classified the trials by their purpose to choose a non-inferiority design.

Results: Seventy-five were included: 43% received funds from industry, overall survival was the most common primary end point and 73% reported positive results. The most frequent purposes underlying the selection of a non-inferiority design were to test more conveniently administered schedules and/or less toxic treatments. In 13 (17%) trials, a clear purpose was not identified. Among the trials that reported a pre-specified NIFm, the median value was 12.5% (range 4%–25%) for trials with binary primary end points and Hazard Ratio of 1.25 (range 1.10–1.50) for trials that used time-dependent methodology for the development of innovative cancer therapies (MDICT). Eur J Cancer 2008; 44: 25–28.


11. Booth CM, Calvert AH, Giaccone G et al. Design and conduct of phase II studies of targeted anticancer therapy: recommendations from the task force on references