Annals of Oncology 23: 537–541, 2012
doi:10.1093/annonc/mdr146
Published online 2 May 2011

Toxicity burden score: a novel approach to summarize multiple toxic effects

S. M. Lee[1]*, D. L. Hershman[2], P. Martin[3], J. P. Leonard[3] & Y. K. Cheung[1]

[1]*Department of Biostatistics, Mailman School of Public Health, Columbia University, New York;* [2]*Department of Medicine, College of Physicians and Surgeons, Columbia University, New York;* [3]*Center for Lymphoma and Myeloma, Weill Medical College of Cornell University, New York, USA*

Received 1 October 2010; revised 14 January 2011 & 11 March 2011; accepted 15 March 2011

Background: Toxicity data from cancer trials are summarized into a single outcome, dose-limiting toxicity (DLT), which does not account for multiple lower grade toxic effects nor differentiates between toxicity types and gradations within DLT.

Methods: Toxicity data were summarized into a toxicity burden score (TBS) using a weighted sum. The severity weights were estimated via regression using historical data. We demonstrated the method using historical data from a bortezomib trial and illustrated the advantages of defining DLT based on TBS in a simulated dose-finding trial.

Results: The estimated weights were 0.17, 0.40 and 0.85 for grade 1/2, grade 3 and grade 4 platelets, respectively; 0.19, 0.64, 1.03 and 2.53 for grade 1, 2, 3 and 4 neuropathy, respectively and 0.17 for each grade 3 or higher nonhematologic toxic effects unrelated to treatment. In the simulated trial, the probability of selecting doses above the maximum tolerated dose decreased when using the DLT defined based on TBS.

Conclusions: TBS is a feasible approach to summarize toxicity. It includes information from the grades and types of multiple toxic effects and can be applied in all phases of drug development. Further efforts should focus on validating the method in a large prospective study before applying it in practice.

Key words: adverse event summary, dose-limiting toxicity, toxicity, toxicity types and grades

introduction

The National Cancer Institute—Common Terminology Criteria (NCI–CTC) for Adverse Events [1] is the instrument for documentation and grading of adverse events in cancer trials. For each adverse event, the severity is graded on a scale from 0 to 5, with grade 0 being no toxicity and grade 5 being death. Given the large number of toxic effects observed, a tremendous amount of toxicity data is collected and reported. To summarize these data, early phase clinical trials define a single binary outcome, dose-limiting toxicity (DLT) or not. Middle- and late-stage trials list all toxic effects for which a maximal toxicity of grade 3 or higher is observed. Thus, it is difficult to quantify the toxicity burden that patients experience and to compare the toxicity profile of different therapies. The binary toxicity summaries disregard lower grade toxic effects, which individually are not dose limiting, but in aggregate can be concerning. Binary toxicity summaries also do not differentiate between types of toxic effects. Moreover, the definition of DLT varies by study [2, 3]. The need for including information on various toxicity grades and types was noted over 10 years ago [4]. Thus, there is an urgency to develop methods to summarize and report toxic effects [5].

A method that appropriately summarizes individual toxic effects into an overall toxicity burden score (TBS) will provide a better understanding of the toxicity burden to patients and have an impact in all phases of drug development and cancer care. In the phase I setting, by defining DLT based on TBS, toxicity information that has been overlooked can be incorporated for the determination of the maximum tolerated dose (MTD). In phase II and III settings, investigators can better understand the toxicity impact of the treatments. The method can also be particularly useful in settings involving noncytotoxic therapies and treatments with long-term toxic effects and it can be used in combination with the time-to-event continual reassessment method (TITE-CRM, [6]) to address the issue of late-onset toxic effects.

The three methods proposed for summarizing the individual toxicity grades and types into a score use a weighted sum of individual toxic effects. The severity weight for the toxic effects and the toxic effects to be included differs among these methods. The TAME method [7] assumes equal weights for all toxicity types. This approach dichotomizes toxic effects and does not consider lower grade toxic effects nor differentiate between toxicity types. Rogatko et al. [3] summarize the toxic effects into a toxicity index score defined as the weighted sum of toxicity grades where the weights are the product of the reciprocal grades plus one. The method takes into account all toxicity grades and differentiates between hematologic and nonhematologic toxic effects. However, it does not differentiate

**Correspondence to*: Dr S. M. Lee, Department of Biostatistics, Columbia University, 722 West, 168th Street, Room 645, New York, NY, 10032, USA. Tel: +1-212-342-1266; Fax: 1-212-342-1246; E-mail: sml2114@columbia.edu

© The Author 2011. Published by Oxford University Press on behalf of the European Society for Medical Oncology.
All rights reserved. For permissions, please email: journals.permissions@oup.com

the toxic effects within these categorizations and assumes that the aggregate effect of lower grade toxic effects never amount to that of a higher grade toxicity. Bekele and Thall [8] define toxicity burden as the weighted sum of toxicity grades for a predetermined set of toxic effects based on the drug being studied. The weights are elicited from the physicians based on their impression of the relative impact of the toxicity. The method takes into account the various grades of toxic effects and distinguishes toxicity types, but the weight elicitation process is *ad hoc*. None of the methods has been applied prospectively in the design of a clinical trial. The method by Rogatko et al. [3] has been applied to compare the toxicity burden of different populations [9, 10].

In this paper, we introduce and evaluate the feasibility of a novel approach to summarize toxic effects into a TBS. The method differentiates between the grades and types of treatment-related toxic effects and uses a regression approach to estimate the severity weights through a historical or existing toxicity dataset. We illustrate the approach using historical toxicity data from a trial in patients with lymphoma [11] and the application of TBS by Monte Carlo simulations of a dose-finding clinical trial where DLT is redefined based on TBS. The new definition includes information on toxicity types and grades and takes into account the aggregate effect of multiple lower grade toxic effects. The inclusion of lower grades of toxicity into the primary outcome of dose escalation studies may help decrease the attrition rate of cancer drugs and the rate of nonadherence to drug due to toxicity.

methods

toxicity burden score

The proposed method for obtaining TBS requires the involvement of two or more physicians and the access to the whole databases of previous clinical trials. The toxicity data from the databases are used during the design stage of a trial to build a regression model. For the design of middle- or late-stage clinical trials, the previous clinical trials can be earlier stage studies of the same drug. For the design of an early-stage trial, these can be studies in a drug with the same mechanism of action and similar toxicity profile as the treatment being studied, in a different patient population for whom similar toxic effects are expected or in combination with other drugs.

Assuming that we are designing a dose-finding trial for bortezomib in patients with lymphoma, we illustrate the proposed model building exercise to obtain the severity weights using the toxicity data from a phase I/II trial in patients with previously untreated diffuse large B-cell or mantle cell non-Hodgkin's lymphoma [11]. The trial objective was to determine the MTD of bortezomib and to assess the safety and efficacy bortezomib, when administered in combination with cyclophosphamide, doxorubicine, vincristine, prednisone and rituximab (CHOP-R). The standard dose for CHOP-R was administered every 21 days. There were five dose levels of bortezomib in the dose escalation part of the trial.

A flow chart of the method is detailed in Figure 1. Three physician raters were asked to identify the toxic effects most attributable to the addition of bortezomib. The toxic effects identified by the physicians were neuropathy and low platelet count. For these two treatment-related toxic effects, the physicians were instructed to assign a score of 1 to DLT, 0 to no toxicity and 5 to death and to assign severity scores for the other grades of toxicity relative to these scores. Assigning scores relative to the DLT was previously used by Yuan et al. [12]. These severity scores provided the physicians with a frame of reference for assigning TBS, given a particular patient profile with multiple toxic effects. The three physicians were masked from each other's scores. The severity scores are displayed in Table 1.

The physicians were then given the toxicity data profile for 24 patients from the trial described above. The dataset listed all toxic effects observed and the respective maximal NCI–CTC grade for each toxicity over the course of the treatment of each patient. Physicians were then asked to assign an individual TBS for each patient based on their impression of all the toxic effects experienced, keeping in mind that the TBS should be consistent with the severity scores that they had previously assigned for each toxicity type. If inconsistencies were observed between the TBS assigned to a patient and the severity scores assigned to the toxic effects, the physicians were asked to review the TBS assigned.

The severity scores represented the physician's TBS assignment for a patient with a single treatment-related toxicity. Therefore, it was important to incorporate this information in the model building. The approach taken to include the information into the model was to add to the dataset one hypothetical patient with a single type and grade of toxicity and the corresponding severity score assigned by each rater as the TBS. The hypothetical patients were created based on the physician's severity score in Table 1. The covariates of interest were the two toxic effects related to

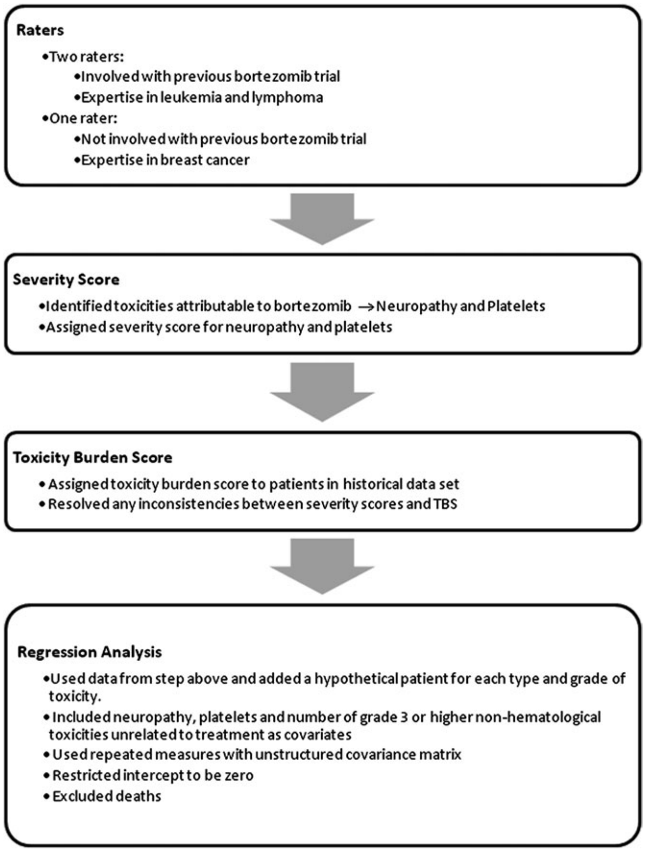
Raters
•Two raters:
•Involved with previous bortezomib trial
•Expertise in leukemia and lymphoma
•One rater:
•Not involved with previous bortezomib trial
•Expertise in breast cancer
Severity Score
•Identified toxicities attributable to bortezomib → Neuropathy and Platelets
•Assigned severity score for neuropathy and platelets
Toxicity Burden Score
• Assigned toxicity burden score to patients in historical data set
• Resolved any inconsistencies between severity scores and TBS
Regression Analysis
•Used data from step above and added a hypothetical patient for each type and grade of toxicity.
• Included neuropathy, platelets and number of grade 3 or higher non-hematological toxicities unrelated to treatment as covariates
• Used repeated measures with unstructured covariance matrix
• Restricted intercept to be zero
• Excluded deaths

Figure 1. Flow chart with the steps for implementing the method.

Table 1. Severity scores of neuropathy and low platelets

NCI toxicity grade	Low platelets			Neuropathy		
	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3
Grade 1	0	0	0.10	0.25	0.30	0.20
Grade 2	0	0	0.30	0.50	0.70	0.50
Grade 3	0.33	0.30	0.60	1.00	1.00	1.00
Grade 4	1.00	1.00	0.80	2.00	3.00	2.00

treatment and the number of grade 3 or higher nonhematologic toxic effects that are unrelated to treatment. For the toxic effects related to treatment, dummy variables were created for each grade of toxicity since severity score was not expected to be linear with respect to toxicity grade. The number of grade 3 or higher nonhematologic toxic effects unrelated to treatment was included as a covariate because the physicians thought it was important for the determination of overall patient toxicity burden. The intercept of the model was restricted to be zero, as a patient experiencing no toxic effects should have a zero TBS. To account for multiple TBS values for each patient from multiple raters, the TBS data were modeled as repeated measurements with an unstructured covariance matrix. Death was not included in the model fitting and it was assigned a TBS of 5.

To validate the model, two of the physicians were asked to assign TBS to an additional set of 17 patients from the same phase I/II trial in lymphoma patients [11]. Assuming the estimated TBS from the original fitted model to be a rater, the ratings were compared using intraclass correlation as well as repeated measures analysis with rater as the covariate of interest. The covariance matrix was assumed to be unstructured. The analyses were done using SAS 9.0, SAS Institute Inc., Cary, NC. [13].

simulation study

We used a simulation study to illustrate the application of TBS in a dose-finding trial of bortezomib in lymphoma patients. We assumed five dose levels of bortezomib, with dose level three being the starting dose. The sample size was 18 and the target probability of toxicity was 0.25. Dose escalation was conducted according to the CRM [14]. For the CRM, the dose-toxicity model was assumed to be empiric, $P(\text{Toxicity at dose d}) = d^{\exp(\beta)}$, with the prior distribution of β being normal with mean 0 and a variance of 1.34. The initial guesses of the probabilities of DLT were 0.011, 0.082, 0.25, 0.464 and 0.654. These were obtained using the algorithm specified in Lee and Cheung [15].

DLT was defined as a platelet count <10 000/mm^3, a grade 3 or greater nonhematologic toxicity or a grade 4 or greater hematologic toxicity. To compare the results using the DLT definition versus the DLT defined based on TBS as outcome, we simulated three population dose-toxicity scenarios. For the DLT defined based on TBS, the model was used to summarize the NCI–CTC toxicity grades for individual patients into a TBS, which was then dichotomized in order to use existing methods for dose-finding trials. Both the CRM and traditional 3+3 design require a binary outcome. A natural dichotomization was based on a TBS of 1, since for an individual toxicity a score of 1 was assigned to the cut-off for the DLT definition. Since all DLTs had a TBS ≥1, by definition, the probability of DLT was always less than or equal to the probability of observing TBS ≥1. Defining DLT as TBS ≥1 allowed for the inclusion of several lower grade toxic effects as a DLT.

The results using the CRM were also compared with those obtained using the 3+3 design with the original DLT definition, as this is the most frequently applied method. For each method and under each scenario, we carried out 1000 simulations. The simulations were carried out using R[16].

results

We obtained the severity scores from the physicians for the grades of neuropathy and low platelets. Once the severity scores were assigned, the physicians were asked to assign TBS scores to 24 patients from the phase I/II trial in lymphoma. A sample of 4 patients from the list of the 24 patients is displayed in Table 2.

Table 2. Sample list of patient toxicity used for elicitation process

Patient	Toxic effects	TBS rater1	TBS rater2	TBS rater3
1	**1 neuropathy, 1 platelets-low,** 2 hemoglobin-anemia, 1 nausea	0.25	0.30	0.30
2	**4 platelets-low,** 1 neutrophils (ANC), 3 WBC-leukocytes	1.0	1.0	0.8
3	**3 neuropathy,** 2 hemoglobin-anemia, 3 anorexia, 3 fatigue, 1 fever, 4 neutrophils (ANC), 2 atrial fib, 2 edema, 1 diarrhea, 1 nausea, 1 constipation, 1 dyspnea, 2 insomnia, 1 hypertension	1.25	1.5	1.5
4	**1 neuropathy**	0.25	0.30	0.20

TBS, toxicity burden score.

To incorporate the information regarding the severity scores, eight hypothetical patients were added to the dataset since there are four grades of toxicity for both neuropathy and platelets. For example, to incorporate the physicians' severity score for a grade 1 neuropathy, one patient was added with neuropathy as the only toxicity and TBS scores of 0.25, 0.30 and 0.20 for raters 1, 2 and 3, respectively. Thus, 32 patients were included for the training dataset. The fixed effect for grade 1 and grade 2 platelet were not significantly different and thus the categories were combined. Figure 2 (training data) suggests that the fitted model is in accordance with the raters. The significant predictors of TBS are listed in Table 3. The coefficient for a grade 4 low platelet is 0.85. This reflects the fact that some physicians consider a platelet count <25 000/mm^3 (grade 4 low platelets) less severe than a DLT, which in some studies is defined as a platelet count <10 000/mm^3. Based on the fixed effects, the estimated TBS is defined as:

$$\begin{aligned}\text{Estimated TBS} = {} & 0.17\ \text{platelet}_{1,2} \\ & + 0.40\ \text{platelet}_3 + 0.85\ \text{platelet}_4 + 0.19\ \text{neuropathy}_1 \\ & + 0.64\ \text{neuropathy}_2 + 1.03\ \text{neuropathy}_3 \\ & + 2.53\ \text{neuropathy}_4 \\ & + 0.17\ \text{number of grade 3 or higher} \\ & \quad \text{non hematologic toxic effects,}\end{aligned}$$

where platelet$_j$ indicates a grade j platelet and neuropathy$_j$ indicates a grade j neuropathy.

The intraclass correlation for the three raters was 0.83.

Figure 3 compares the estimated TBS from the model to the TBS assigned by the two raters as well as the TBS from the two raters. The line indicates equality. There is a disagreement for high values of TBS between rater 2 and rater 3 and also between rater 2 and the estimated TBS from the model. The intraclass correlation between the three ratings was 0.67. The intraclass correlation between each pair of measures was 0.54, 0.78 and 0.68 for the two raters, for rater 2 and the fitted model and for rater 3 and the fitted model, respectively.

application of TBS

To use TBS for the simulated dose-finding trial of bortezomib, we summarized each individual patient's NCI–CTC toxic effects into a TBS value using the model. For example, a patient with a grade 2 neuropathy, a grade 3 platelet (platelet count >25 000/mm^3 and <50 000/mm^3) and no grade 3 or higher

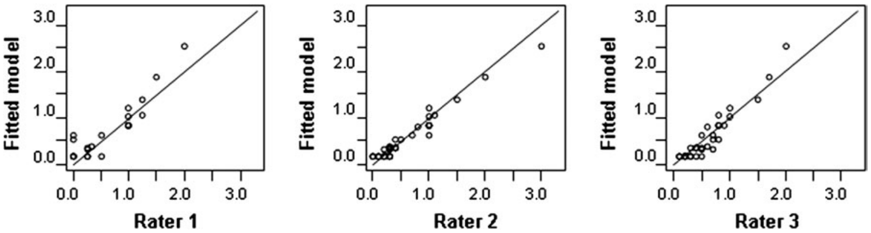

Figure 2. Scatter plots of the toxicity burden scores (TBSs) assigned by raters versus the estimated TBS from the fitted model for the training data (N = 32 for rater 2 and 3, N = 21 for rater 1).

Table 3. Significant predictors of toxicity burden score

Significant predictors	Coefficient (SE)	*P*value
Low platelet count		
Grade 1 or 2	0.17 (0.03)	<0.001
Grade 3	0.40 (0.09)	<0.001
Grade 4	0.85 (0.04)	<0.001
Neuropathy		
Grade 1	0.19 (0.03)	<0.001
Grade 2	0.64 (0.06)	<0.001
Grade 3	1.03 (0.06)	<0.001
Grade 4	2.53 (0.09)	<0.001
Number of grade 3 or higher nonhematologic toxicity unrelated to treatment	0.17 (0.03)	<0.001

SE, standard error.

nonhematologic toxicity unrelated to treatment has a TBS of 0.64 + 0.40 = 1.04. This patient would not have a DLT if the toxic effects are examined individually since neither a grade 3 platelet nor a grade 2 neuropathy constitutes a DLT. However, by accommodating for multiple toxic effects, we took into account the cumulative effects for multiple non-DLTs.

We compared the definitions of DLT in the simulated trial assuming three population scenarios. The true probabilities of toxic effects and the percentage with which each dose level is recommended using CRM with the DLT definition used in the trial [CRM(DLT)] versus TBS ≥1 [CRM(TBS)] as the outcome and the 3+3 are displayed in Figure 4. For example, in the first scenario, the probability of DLT as defined in the trial is 0.10, 0.25, 0.40, 0.45 and 0.55 for each dose level respectively, while the probability of TBS ≥1 is 0.25, 0.35, 0.55, 0.60 and 0.75. For dose level 1 of this scenario, the probability that TBS ≥1 and individually the toxic effects do not constitute a DLT is 0.15. Consistent with previous publications, CRM(DLT) recommended doses above the MTD more frequently than the 3+3 design, and the recommendation percentages of the 3+3 design were more spread out to the extreme doses than the CRM. In all scenarios, CRM(TBS) was less likely to recommend an overdose than CRM(DLT). It also selected the extreme doses less often than the 3+3 design.

discussion

We have proposed a simple regression approach to obtain the severity weights and to summarize patient toxic effects into a TBS. By redefining a DLT as a TBS ≥1 instead of using specific cut-off grades for the maximal toxicity, the approach incorporates information from the grades and types of multiple toxic effects for the determination of the MTD. In the single toxicity case or when lower grades of toxicity are not of interest, the method is equivalent to using the traditional DLT definition as outcome. In the multiple toxicity case where it is important to account for multiple lower grades of toxic effects, the method offers advantages over traditional approaches since it includes the aggregate effects of lower grades of toxicity. While we illustrate the application of TBS in a dose-finding trial with a binary outcome (DLT), statistical methods that can incorporate a continuous or ordinal outcome such as the TBS have been developed for dose-finding trials [8, 17].

In this paper, the method is illustrated using the toxicity data from a single previous clinical trial, showing the feasibility of the method for designing early-stage trials, when limited data is available. These data, although limited, still provide for a more rigorous approach to estimate the severity weights of the various types and grades of toxic effects compared with other previously proposed methods [3, 7, 8]. It also allows for the inclusion of information on lower grade toxic effects. While the method cannot be applied for early-stage trials when a previous study is not available and the study is the first in human, often early-stage studies are carried out for the same drug in different populations, for different indications or in different combinations. In those settings, the data from previous experiences can be used to build the model. Further research is warranted to obtain data to build models for particular types of compounds. For late-stage studies, the method benefits from toxicity data from preexisting early-stage trials and other late-stage studies. The availability of these data implies more effort to build a model, but in exchange for better reliability. Using the model, a TBS can be calculated for each patient and the TBS of the treatments can be compared. The comparison of TBS complements the comparison of individual symptoms currently carried out in late-stage clinical trials for the drug approval process.

While we have illustrated the feasibility of the approach, the study is limited by its evaluation of one clinical trial with a limited number of toxic effects and physicians. Further efforts should focus on validating these methods using large-scale prospective studies from multiple tumor types with a large number of physicians to examine the effect of study drug, rare adverse events and tumor types on severity scores and TBS. This validation is necessary before the method is applied in practice. The method can also be extended in the future to include patients' self report of toxicity burden. Since

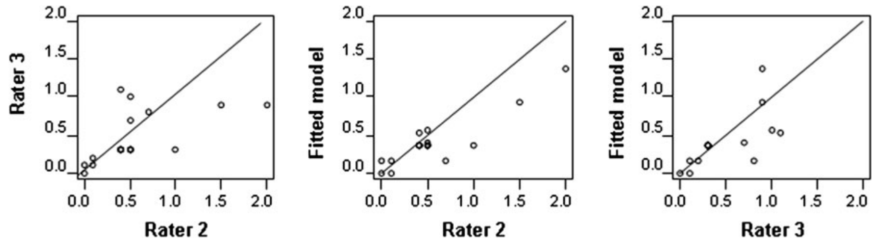

Figure 3. Scatter plots of the toxicity burden score (TBSs) assigned by raters versus the estimated TBS from the fitted model for validation data ($N = 17$).

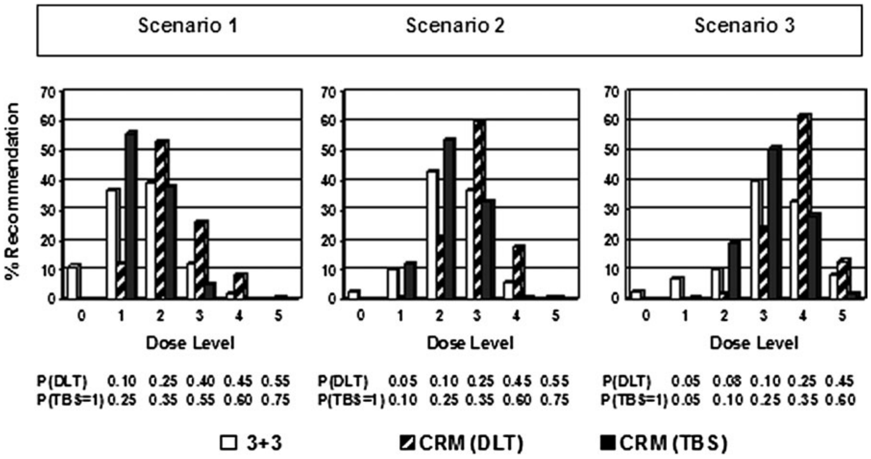

Figure 4. Recommendation percentages under three different scenarios of toxicity probabilities. Toxicity probabilities are displayed under each dose level. Open bar: 3+3, 3+3 dose escalation design; hatched bar: CRM(DLT), continual reassessment method with DLT definition as specified in the trial as outcome; DLT, dose-limiting toxicity as defined in the trial; solid bar: CRM(TBS), continual reassessment method with TBS as outcome. TBS, toxicity burden score.

physicians' and patients' report can differ significantly, including patients' self report will improve current clinical designs [18].

funding

This work was supported by National Institutes of Health (R01 NS055809).

disclosure

The authors declare no conflicts of interest.

references

1. National Cancer Institute: CTEP. Common Terminology Criteria for Adverse Events (CTCAE) Version 3.0. Bethesda, MD: National Institutes of Health 2003.
2. Dent SF, Eisenhauer EA. Phase I trial design: are new methodologies being put into practice? Ann Oncol 1996; 7: 561–566.
3. Rogatko A, Babb JS, Wang H et al. Patient characteristics compete with dose as predictors of acute treatment toxicity in early phase clinical trials. Clin Cancer Res 2004; 10: 4645–4651.
4. Eisenhauer EA, O'Dwyer PJ, Christian M, Humphrey JS. Phase I clinical trial design in cancer drug development. J Clin Oncol 2000; 18: 684–692.
5. Trotti A, Bentzen SM. The need for adverse events reporting standards in oncology clinical trials. J Clin Oncol 2004; 22(1): 19–22.
6. Cheung YK, Chappell R. Sequential designs for phase I clinical trials with late-onset-toxicities. Biometrics 2000; 56: 1177–1182.
7. Trotti A, Pajak T, Gwede C et al. TAME: development of a new method for summarising adverse events of cancer treatment by the Radiation Therapy Oncology Group. Lancet Oncol 2007; 8(7): 613–624.
8. Bekele B, Thall P. Dose-finding based on multiple toxicities in a soft tissue sarcoma trial. J Am Stat Assoc 2004; 60: 684–693.
9. Townsley CA, Pond GR, Oza AM et al. Evaluation of adverse events experienced by older patients participating in studies of molecularly targeted agents alone or in combination. Clin Cancer Res 2006; 12: 2141–2149.
10. Pond GR, Siu LL, Moore M et al. Nomograms to predict serious adverse events in phase II clinical trials of molecularly targeted agents. J Clin Oncol 2008; 26(8): 1324–1330.
11. Leornard JP, Furman RR, Cheung YKK et al. Phase I/II trial of botezomib plus CHOP-Rituximab in diffuse large B cell (DLBCL) and mantle cell lymphona (MCL): phase I results. Blood 2005; 106: (Abstr 491).
12. Yuan Z, Chappell R, Bailey H. The continual reassessment method for multiple toxicity grades: a bayesian quasi-likelihood approach. Biometrics 2007; 63: 173–179.
13. SAS Institute Inc.. SAS Version 9.0. Cary, NC: SAS Institute Inc. 2004.
14. O'Quigley J, Pepe M, Fisher L. Continual reassessment method: a practical design for phase I clinical trials in cancer. Biometrics 1990; 46: 33–48.
15. Lee SM, Cheung YK. Model calibration in the continual reassessment method. Clin Trials 2009; 6: 227–238.
16. R Development Core Team. R: A language and environment for statistical computing., Vienna, Austria: R Foundation for Statistical Computing 2008.
17. Lee SM, Cheng B, Cheung YK. Continual reassessment method with multiple toxicity constraints. Biostatistics 2011; 12: 386–398.
18. Basch E. The missing voice of patients in drug-safety reporting. N Engl J Med 2010; 362(10): 865–869.