HHS Public Access

Author manuscript

Clin Trials. Author manuscript; available in PMC 2021 February 12.

Published in final edited form as:

Clin Trials. 2021 February ; 18(1): 104–114. doi:10.1177/1740774520975120.

Composite Grading Algorithm for the National Cancer Institute's Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE)

Ethan Basch[1,2], **Claus Becker**[3], **Lauren J. Rogak**[2], **Deborah Schrag**[4], **Bryce B. Reeve**[5], **Patricia Spears**[1], **Mary Lou Smith**[6], **Mrinal M. Gounder**[2], **Michelle R. Mahoney**[7], **Gary K. Schwartz**[8], **Antonia V. Bennett**[1], **Tito R. Mendoza**[9], **Charles S. Cleeland**[9], **Jeff A. Sloan**[7], **Deborah Watkins Bruner**[10], **Gisela Schwab**[11], **Thomas M. Atkinson**[2], **Gita Thanarajasingam**[12], **Monica M. Bertagnolli**[13], **Amylou C. Dueck**[14]

[1]UNC Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC

[2]Memorial Sloan Kettering Cancer Center, New York, NY

[3]Deciphera Pharmaceuticals, Waltham, MA

[4]Dana-Farber/Partners Cancer Care, Harvard Cancer Center, Boston, MA

[5]Duke Cancer Institute, Duke University Medical Center, Durham, NC

[6]Research Advocacy Network, Plano, TX

[7]Alliance Statistics and Data Center, Mayo Clinic, Rochester, MN

[8]Columbia University, New York, NY

[9]University of Texas M. D. Anderson Cancer Center, Houston, TX

[10]Emory University Hospital/Winship Cancer Institute, Atlanta, GA

[11]Exelixis, Inc., South San Francisco, CA

[12]Division of Hematology, Mayo Clinic, Rochester, MN, USA

[13]Alliance Office of the Group Chair, Brigham and Women's Hospital, Boston, MA

[14]Alliance Statistics and Data Center, Mayo Clinic, Scottsdale, AZ

Abstract

Background—Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE) is an item library designed for eliciting patient-reported adverse events in oncology. For each adverse event, up to three individual items are scored for frequency, severity, and interference with daily activities. To align PRO-CTCAE with other standardized tools for adverse event assessment including CTCAE, an algorithm for mapping individual items for any given adverse event to a single composite numerical grade was developed and tested.

Correspondence: Amylou C. Dueck, PhD, Mayo Clinic, 13400 E. Shea Blvd., Scottsdale, AZ 85259, dueck.amylou@mayo.edu, Phone: 480-301-6159.

ClinicalTrials.gov Identifiers: NCT01522443 (COMET-2) and NCT02066181 (Alliance A091105)

Methods—A five-step process was used. (1) All 179 possible PRO-CTCAE score combinations were presented to 20 clinical investigators to subjectively map combinations to single numerical grades ranging from 0-3. (2) Combinations with <75% agreement were presented to investigator committees at a National Clinical Trials Network cooperative group meeting to gain majority consensus via anonymous voting. (3) The resulting algorithm was refined via graphical and tabular approaches to assure directional consistency. (4) Validity, reliability, and sensitivity were assessed in a national study dataset. (5) Accuracy for delineating adverse events between study arms was measured in two phase III clinical trials (NCT02066181 and NCT01522443).

Results—In Step 1, 12/179 score combinations had <75% initial agreement. In Step 2, majority consensus was reached for all combinations. In Step 3, five grades were adjusted to assure directional consistency. In Steps 4 and 5, composite grades performed well and comparably to individual item scores on validity, reliability, sensitivity, and between-arm delineation.

Conclusion—A composite grading algorithm has been developed and yields single numerical grades for adverse events assessed via PRO-CTCAE, and can be useful in analyses and reporting.

Keywords

Adverse event; Patient-reported outcome; Toxicity; Health-related quality of life; Oncology; Symptom; PRO-CTCAE; CTCAE

INTRODUCTION

To improve detection of symptomatic adverse events in cancer clinical trials, the National Cancer Institute (NCI) supported the development of a patient-reported version of its standard adverse event lexicon, the Common Terminology Criteria for Adverse Events (CTCAE), which is called the Patient-Reported Outcomes version of the CTCAE (PRO-CTCAE).[1] An impetus for creating the PRO-CTCAE was the observation that investigators can miss up to half of study participants' symptomatic adverse events in drug development trials (e.g., nausea, sensory neuropathy), leading to potential underestimations of harms.[2] The PRO-CTCAE is currently being adopted across NCI- and industry-funded trials, and the Food and Drug Administration has encouraged its adoption and development of standardized PRO-CTCAE analyses and reporting.[3]

The PRO-CTCAE is a library of 124 items that measure 78 symptomatic adverse events and is publicly available from the NCI (http://healthcaredelivery.cancer.gov/pro-ctcae).[4] For each adverse event, up to three individual items are administered to patients to evaluate the "attributes" of frequency, severity, and interference with daily activities. Therefore, the PRO-CTCAE item library consists of 3-attribute, 2-attribute, and 1-attribute adverse events. For example, PRO-CTCAE for pain (a 3-attribute adverse event) includes three separate questions for pain frequency, pain severity, and pain interference with daily activities. Some adverse events are limited to less than three attributes (e.g., fatigue has separate items for severity and interference, but not frequency) while others are reported as present/not present by patients (e.g., bruising or hives). Terminology for adverse events is phrased in patient-friendly lay language (e.g., "mouth or throat sores" for oral mucositis), and response scores are phrased as simple verbal terms (e.g., "none", "mild", "moderate", "severe", "very

severe"). Measurement properties of PRO-CTCAE items including validity, test-retest reliability, and sensitivity to change over time have previously been published.[5,6,7]

Currently, investigators may select a subset of adverse events in the PRO-CTCAE library that are pertinent to a given trial context based on known and anticipated properties of study drugs, and assemble these items into a custom questionnaire.[8,9] PRO-CTCAE items are generally administered to patients pre-treatment, regularly throughout a trial, and during post-treatment (e.g., weekly during active treatment and every three months during follow-up).[10] Currently, each item is analyzed individually; therefore, there can be up to three separate scores for any given symptomatic adverse event (e.g., pain frequency score, pain severity score, pain interference score). Although this approach is useful for understanding the granularity of the patient experience with treatment, it is not consistent with existing standardized metrics for adverse event reporting in clinical research that use a single metric for each adverse event, such as the CTCAE or Medical Dictionary for Regulatory Activities (MedDRA), and may lead to data fatigue for clinicians viewing adverse event tables in publications when a single metric per adverse event may suffice. Therefore, we sought to develop an algorithm to generate a single composite numerical grade for each PRO-CTCAE symptomatic adverse event based on mapping of its individual item scores, and to evaluate the composite grades to assure that their validity, reliability, and sensitivity to change are comparable with individual item scores.

METHODS

Overall framework:

To develop a composite grading algorithm, the research team established the following general principles: First, the algorithm should be based on clinical investigator input and be substantiated using empiric patient data. Second, the algorithm should be developed to generate the same grade for a given score combination regardless of the PRO-CTCAE symptomatic adverse event unless empiric data suggest otherwise. Third, the algorithm should produce composite grades that are analogous to the scale employed by the CTCAE to enable similar reporting, which for symptomatic adverse events generally range from grades 0-3 (Supplemental Table S1). Fourth, the algorithm should apply to only the 59 PRO-CTCAE symptomatic adverse event terms for which magnitude is measured (i.e., not apply to the 21 PRO-CTCAE symptomatic adverse event terms that are scored on the binary present/not present scale).

Procedure for development and testing:

As detailed below, a five-step procedure was implemented including (1) elicitation of an initial composite grading algorithm from clinical investigators via a data collection exercise; (2) refinement of the algorithm through targeted questions administered to a broader audience of clinical investigators; (3) use of graphical approaches to ensure the directional consistency of composite grades; (4) quantitative testing of the algorithm's measurement properties in a national study dataset; and (5) application of the algorithm to data from multi-site randomized cancer clinical trials.

In Step 1 (“Clinical Investigator Grade Assignment”), a data collection form was created with a table showing the 179 possible combinations of PRO-CTCAE item attributes each in a row (e.g., frequency “occasionally”, severity “mild”, interference “somewhat”), with a space for the investigator to select a single composite grade from 0-3 to which he/she felt that combination should map, corresponding to the CTCAE (Table 1, Supplemental Table S2). This form was administered anonymously to 20 clinical investigators with ≥10 years of experience serving as principal or co-investigators for NCI- and industry-sponsored cancer clinical trials. Composite grades assigned by the 20 investigators were tabulated for each combination, with consensus defined *a priori* as endorsement of a specific numerical grade by ≥75% of investigators for a given combination. Combinations that did not reach consensus were further investigated in Step 2.

In Step 2 (“Clinical Investigator Consensus”), input on the algorithm resulting from Step 1 was systematically elicited from clinical investigators attending the Alliance for Clinical Trials in Oncology biannual group meeting in Chicago, IL. During committee meetings for the Breast, Thoracic, Genitourinary, and Gastrointestinal disease committees, clinical scenarios were presented via PowerPoint slides (Supplemental Figure S1) which described patients with adverse events characterized by various individual attribute combinations. Investigators were asked to vote using electronic audience response units on a single composite grade to which they felt each combination should map, consistent with the CTCAE. No identifying information was collected from the investigators, although there was no overlap between investigators involved in Steps 1 and 2. Clinical scenarios included attribute item score combinations that did not reach consensus in Step 1. Clinical scenarios were also presented for 12 randomly selected item attribute combinations that had reached ≥75% consensus in Step 1 to confirm consensus. Agreement across the respondents was tabulated for each combination. The majority grades were used in the composite grading algorithm evaluated in Step 3.

Step 3 (“Directional Consistency Check”) entailed use of graphical and tabular approaches to ensure the directional consistency of composite grades. Contour plots of composite grades by frequency, severity, and interference were created in Matlab (method developed by co-author Claus Becker) and reviewed to identify whether situations existed such that increasing individual PRO-CTCAE attribute scores would lead to decreasing composite grades. Composite grades for adverse events with two-score combinations (e.g., frequency plus severity) were compared to the range of composite grades for adverse events with three-score combinations (frequency plus severity plus interference) to confirm consistency of composite grades. Directionally inconsistent composite scores were modified and the final composite grading algorithm was then quantitatively tested in Steps 4 and 5.

Step 4 (“Validation”) entailed comparing the measurement properties of the composite grades derived from Steps 1-3 to the previously published measurement properties of the individual items, including validity, reliability, and sensitivity to change. We sought to assess whether measurement properties were comparable. The algorithm was applied to PRO-CTCAE data collected at 2-3 visits in the primary validation study of the PRO-CTCAE,[5] which enrolled 940 patients receiving active treatment for a variety of cancer types at 9 US-based centers. Convergent validity was assessed using Pearson correlations between PRO-

CTCAE composite grades and European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire Core 30 (EORTC QLQ-C30) scales,[11] including an overall summary score.[12] EORTC QLQ-C30 health-related quality of life summary and functioning/global scales were reverse scored for analysis (original direction retained for graphical representations) such that higher scores represent inferior outcomes, matching the direction of PRO-CTCAE items. Comparison of mean EORTC QLQ-C30 health-related quality of life summary scores across increasing PRO-CTCAE composite grade groups were carried out using Jonckheere-Terpstra tests, which evaluate for monotonically decreasing health-related quality of life for increased PRO-CTCAE composite grade groups.[13] Known-groups validity was assessed by comparing mean PRO-CTCAE composite grades between 66 previously defined[5] groups of patients on the basis of Eastern Cooperative Oncology Group (ECOG) performance status (0-1 versus 2-4), cancer type, and treatment using two-sample t-tests with effect sizes computed as the differences between group means divided by the pooled standard deviation (i.e., Cohen's *d*). Test-retest reliability was estimated using intraclass correlation coefficients based on a one-way analysis of variance model. Sensitivity to change was assessed by comparing changes from first to last visit between groups of patients reporting worsening, no change, or improvement via global impression of change items at the last visit using Jonckheere-Terpstra tests. Standardized response mean was computed per group as the mean change score divided by the standard deviation of the change scores. All methods and measurement properties of individual PRO-CTCAE items have previously been described.[5]

Finally, in Step 5 ("Clinical Trial Evaluation"), the grading algorithm was applied to PRO-CTCAE data collected in two double-blind placebo-controlled phase III trials: sorafenib vs. placebo in patients with advanced desmoid tumors (Alliance A091105, NCT02066181)[14]; and cabozantinib vs. mitoxantrone plus prednisone in patients with metastatic castration-resistant prostate cancer (Exelixis COMET-2, NCT01522443).[15] Details of patient consent are previously reported.[14,15] To assess the utility and meaningfulness of the composite grading approach, individual adverse event rates were compared between arms using individual PRO-CTCAE item scores vs. composite grades. Post-baseline PRO-CTCAE rates were tabulated and compared with Fisher's exact tests using a previously described method for baseline adjustment,[8] following application of the composite grading algorithm to scores at each assessment time point.

RESULTS

In Step 1, ≥75% of 20 participating clinical investigators assigned the same CTCAE grades for 175/179 (98%) PRO-CTCAE score combinations. Table 2 shows the 12 score combinations for which less than 75% of investigators assigned the same grade, with the distributions of assigned grades for those items. Most disagreements were within one CTCAE grade, and 4/12 (33%) were considered "unlikely" or "impossible" combinations (e.g., "frequently" occurring pain of no severity which interferes "very much" with daily activities).

In Step 2, consensus was reached with ≥60% of clinical investigators assigning the same composite grade for 11/12 PRO-CTCAE score combinations which failed to reach

consensus in Step 1 (Table 2). The one PRO-CTCAE score combination not reaching consensus had 36% of investigators finding the combination to be impossible or uncertain. For the additional randomly selected combinations which had reached consensus in Step 1, voting confirmed the selections from Step 1 in all cases (data not shown).

In Step 3, contour plots were reviewed and 2/179 PRO-CTCAE score combinations were identified for which changes in an attribute score led to inconsistent changes in composite grade (Figure 1). These directional inconsistencies, identified through this visualization technique, would not have otherwise been found. The first of these combinations had a frequency score 3 (“frequently”), severity score 0 (“none”), and interference score 4 (“very much”) with composite grade of 3. Increasing the severity score to 1 (“mild”) led to a decreased composite grade of 2, which was directionally inconsistent. Similarly, increasing the frequency score to 4 (“almost constantly”) also led to an inconsistent decreased composite grade of 2. The second identified score combination was inconsistent in only one direction. Specifically, a PRO-CTCAE score combination with a frequency score of 2 (“occasionally”), severity score of 4 (“very severe”), and interference score of 1 (“a little bit”) yielded a composite grade of 3. However, increasing frequency to a score of 3 (“frequently”) resulted in a composite grade of 2, which was directionally inconsistent. The algorithm was revised to a grade of 2 for both score combinations. Next, in the tabular comparison of two-score vs. three-score combinations, the two-score combinations of “frequently” plus “severe” or “very severe”, and “frequently” plus “quite a bit”, mapped to grade 2, whereas related three-score combinations mapped to grade 3. Therefore, composite grades for these two-score combinations were modified to grade 3.

In Step 4, we compared results of validity and reliability testing of the individual PRO-CTCAE items, as previously published, with validity and reliability testing of composite grades. In the original validation study of individual items, we found 122 of 124 individual PRO-CTCAE items were associated in the expected direction with the EORTC QLQ-C30 health-related quality of life summary score (107/124 with meaningful correlation [i.e., Pearson $r \geq 0.1$]; 102/124 $p < 0.05$; 87/124 $p < 0.001$). Individual items with meaningful correlation represented 54 of the 59 adverse events for which magnitude was measured; and with statistical significance represented 51/59 adverse events (45/59 adverse events with $p < 0.001$). In the analysis of composite grades, 53/59 adverse events were meaningfully correlated (i.e., Pearson $r \geq 0.1$) with the EORTC QLQ-C30 health-related quality of life summary score (Pearson correlation test: 51/59 $p < 0.05$, 44/59 $p < 0.001$; Jonckheere-Terpstra test: 49/59 $p < 0.05$, 43/59 $p < 0.001$; Figure 2 and Supplemental Figure S2) at the first visit.

In the prior analysis, scores for 94 of 124 individual PRO-CTCAE items (representing 47/59 adverse events for which magnitude was measured) were higher in the ECOG performance status 2 to 4 vs 0 to 1 group (58/124 $p < 0.05$ with magnitude items representing 28/59 adverse events; 37/124 $p < 0.001$ with magnitude items representing 19/59 adverse events). Similar to the prior analysis, composite grades for 46/59 (78%) adverse events were higher (effect size [Cohen’s d]>0) in patients with ECOG PS 2-4 vs 0-1 (median effect size 0.23 [range −0.49-0.73]; 32/59 effect size ≥0.2; 25/59 $p < 0.05$; 15/59 $p < 0.001$). Differences in the number of statistically significant adverse events are likely due to a smaller number of tested comparisons, and due to compression in scale (from a 0-4 to a 0-3 range). In the additional

known-groups comparisons by cancer type/treatment, 56/66 (85%) composite grades were higher in patients expected to have worse symptom experience (median effect size 0.40 [range −0.14-1.46]; 49/66 effect size ≥0.2; 47/66 $p<0.05$; 32/66 $p<0.001$).

Overall in the previously published analysis, 119 of 124 individual items (representing 57/59 adverse events for which magnitude was measured) met the validity criterion defined in Dueck et al. as statistical significance ($p<0.05$) along with a meaningful effect size (Pearson $r\geq0.1$ or group difference effect size ≥0.2) observed for at least 1 convergent or known-groups validity analysis. Identical to the analysis of individual items, 57 of 59 composite grades met validity criterion with only “nosebleeds” and “pain during vaginal sex” failing to meet the minimum statistical significance and meaningful effect size requirements. All conducted construct validity analyses appear in Supplemental Table S3.

Test-retest reliability was 0.7 or greater for 36 of 49 prespecified individual items (median [range] intraclass correlation coefficient, 0.76 [0.53-.96]). The test-retest reliability for the corresponding 24 selected composite grades completed by 80 subjects on consecutive business days (median 1 day, range 1-3 days) ranged from 0.57-0.96 (median intraclass correlation coefficient 0.77) with 18/24 (75%) grades having an intraclass correlation coefficient ≥0.7, which was comparable to previously published reliability findings for individual PRO-CTCAE items (Supplemental Table S4).

Correlations between PRO-CTCAE item changes from first to last visit and corresponding EORTC QLQ-C30 scale changes were statistically significant for 27 prespecified individual items in 835 subjects (median [range] $r=0.43$ [0.10-0.56]; all $p\leq0.006$). For 13 corresponding adverse events, 11/14 (79%) correlations between PRO-CTCAE grade changes from first to last visit and corresponding EORTC QLQ-C30 scale changes were statistically significant (median [range] $r=0.49$ [0.11-0.58]; all $p<0.001$). Changes from first to last visit were also statistically significantly monotonically decreasing (Jonckheere-Terpstra test $p<0.05$) for 12/14 (86%) pre-specified PRO-CTCAE composite grades across subjects reporting worsening versus no change versus improvement. The median (range) standardized response mean in patients reporting worsening, no change, and improvement were 0.20 (0.03-0.34), −0.06 (−0.20-0.03), and −0.12 (−0.32-0.06) -- also similar to previously published findings for PRO-CTCAE individual items.[5]

In Step 5, standard toxicity tables were generated for two phase III trials showing rates of adverse events >0 and ≥3, separately for individual PRO-CTCAE items and for composite grades. Rates of adverse events in the A091105 trial (Table 3) are comparable between the individual item scores (top) and composite grades (bottom). Specifically, the pattern and directionality of between-arm differences was preserved with composite grades, and adverse events with statistically significant differences detected between arms were identical (nausea, diarrhea, rash, and hand/foot syndrome). This comparison is shown graphically for hand-foot syndrome at each PRO-CTCAE assessment time point (Figure 3). Findings were similar in the comparison of PRO-CTCAE items scores and composite grades in the COMET-2 trial (Supplemental Table S5).

DISCUSSION

The PRO-CTCAE was developed to improve the accuracy and patient-centeredness of adverse event evaluation in oncology trials. By design, the PRO-CTCAE separately elicits discrete information about frequency, severity, and interference of many included adverse events, which facilitates granular analyses of the patient experience. Yielding a single composite grade for each adverse event further mproves the utility of the PRO-CTCAE by aligning its output with other common metrics like CTCAE and MedDRA, and enabling more efficient analysis, interpretation, and reporting by investigators, regulators, and patients. Use of composite grades may be particularly helpful in tables for succinct listing of adverse events in publications or in drug labels. However, wherever more granular understanding of the patient experience is warranted for analyses, individual item scores should be analyzed.

PRO-CTCAE data, whether analyzed as individual item scores or composite grades, are distinct and discrepant from CTCAE data, and provide a more direct and precise measurement of the patient experience.[5] The US FDA has clarified that PRO-CTCAE is not considered safety data, and there is no expectation that PRO-CTCAE data be reported to the FDA directly as safety data in cancer trials.[16]

This paper describes a multistep development and testing approach. This involved two steps of feedback from clinical investigators to establish an initial algorithm; graphical analysis to assure directional consistency; assessment of measurement properties to demonstrate comparability with individual PRO-CTCAE item scores; and tabulation in two clinical trials to show ability to delineate between arms comparably with individual PRO-CTCAE item scores. Such an approach provides confidence in both the clinical and methodological meaningfulness of the algorithm. Patient input was included throughout these steps, and was extensively included in development of the PRO-CTCAE itself.[1] Particular confidence in the finalized algorithm (Supplemental Table S2) is provided by the high levels of clinical investigator consensus for PRO-CTCAE score combinations. Among the small number of combinations for which there was not consensus, one-third were deemed impossible combinations (e.g., high frequency plus no severity), and for the others, consensus was able to be reached. Further confidence is provided because the final algorithm performed well in testing of validity, reliability, and sensitivity to change similar to previously published results for individual PRO-CTCAE items,[5] and was able to delineate between study arms in clinical trials without observed loss of information compared to analysis of individual PRO-CTCAE items in terms of pattern and directionality of adverse events and detection of adverse events with statistically significant between-arm differences.

There are limitations of this approach. First, although there were high levels of agreement on the algorithm among clinical investigators overall, consensus was not perfect. This serves as a reminder that clinical judgment varies, and no algorithm will yield universal consensus. Notably, this is the case for most summary metrics in clinical research spanning radiographic and biomarker assessment. Second, quantitative evaluations used existing clinical study datasets in which PRO-CTCAE frequency, severity, and interference scores were already known to perform well on validity, reliability, sensitivity, and between-arm

delineations. Notably, this provided an opportunity to assure comparability of performance of the composite grades with individual item scores. Third, the same grading algorithm was applied across all adverse events, rather than tailoring the algorithm for each adverse event. An alternative approach would vary the algorithm between different adverse events. However, use of a single algorithm across adverse events is supported by the similar monotonically decreasing health-related quality of life summary scores observed for increasing composite grades across adverse events (Figure 2). Moreover, varying the algorithm would add substantial complexity and risk of errors in analyses, and would likely be infeasible to evaluate quantitatively given the large amounts of necessary data to do so. Fourth, there may be alternative methods to derive/test a composite grading algorithm for the PRO-CTCAE. However, there is no existing standard methodology, and a strength of the current approach was the combination of qualitative, graphical, and quantitative assessments. As more data become available from ongoing and future studies administering the PRO-CTCAE, further refinements or confirmations of the algorithm may become possible. Fifth, this analysis was conducted using data in the context of cancer care. The algorithm is likely translatable to other disease contexts, which could be empirically evaluated, given that there is increasing interest to use PRO-CTCAE in clinical trials outside oncology.

Finally, although a formal step of systematic patient review of the algorithm was not included, patients were extensively involved in development of the PRO-CTCAE and design of the validity and clinical trial assessments. Also, two patient co-investigators are part of the current research team and are authors on this paper, and reported results throughout this paper were based on analyses of patient-reported datasets. Notably, this grading algorithm does not apply to the pediatric PRO-CTCAE[17] due to its different scoring approach, and a composite scoring algorithm for the pediatric PRO-CTCAE is in development separately.

CONCLUSION

A grading algorithm to yield single numerical grades for adverse events based on multiple PRO-CTCAE items has been derived from direct clinical investigator input and quantitative validation. This algorithm can be useful when analyzing and reporting PRO-CTCAE results.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding: Data analyses for this study were supported by Cancer Moonshot funding (U.S. National Cancer Institute U01 CA233046); data collection for the PRO-CTCAE validation study was conducted under the U.S. National Cancer Institute (contract HHSN261200800043C); data collection for the COMET-2 trial was conducted by Exelixis; data collection for A091105 was conducted by the Alliance for Clinical Trials in Oncology with support from Bayer and U.S. National Cancer Institute funding (grants UG1CA189823 [NCORP], U10CA180821, U10CA180882, U10CA180790, U10CA180791, U10CA180838, U10CA180857, U10CA180858, U10CA180867, UG1CA189960). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

1. Basch E, Reeve BB, Mitchell SA, et al. Development of the National Cancer Institute's patient-reported outcomes version of the common terminology criteria for adverse events (PRO-CTCAE). J Natl Cancer Inst. 2014 9 29;106(9).
2. Basch E, Jia X, Heller G, et al. Adverse symptom event reporting by patients vs clinicians: relationships with clinical outcomes. J Natl Cancer Inst 2009;101(23):1624–32. [PubMed: 19920223]
3. Kluetz PG, Papadopoulos EJ, Johnson LL, et al. Focusing on Core Patient-Reported Outcomes in Cancer Clinical Trials-Response. Clin Cancer Res. 2016 11 15;22(22):5618. [PubMed: 28151715]
4. National Cancer Institute. Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events. Available at: https://healthcaredelivery.cancer.gov/pro-ctcae. Last accessed June 20, 2017.
5. Dueck AC, Mendoza TR, Mitchell SA, et al. Validity and Reliability of the US National Cancer Institute's Patient-Reported Outcomes Version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE). JAMA Oncol. 2015 11; 1(8):1051–9. [PubMed: 26270597]
6. Hay JL, Atkinson TM, Reeve BB, et al. Cognitive interviewing of the US National Cancer Institute's Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE). Qual Life Res. 2014 2; 23(1):257–69. [PubMed: 23868457]
7. Bennett AV, Dueck AC, Mitchell SA, et al. Mode equivalence and acceptability of tablet computer-, interactive voice response system-, and paper-based administration of the U.S. National Cancer Institute's Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE). Health Qual Life Outcomes. 2016 2 19; 14:24. doi: 10.1186/s12955-016-0426-6. [PubMed: 26892667]
8. Basch E, Rogak LJ, Dueck AC. Methods for Implementing and Reporting Patient-reported Outcome (PRO) Measures of Symptomatic Adverse Events in Cancer Clinical Trials. Clin Ther. 2016 4;38(4):821–30. [PubMed: 27045992]
9. Basch E, Pugh SL, Dueck AC, et al. Feasibility of Patient Reporting of Symptomatic Adverse Events via the Patient-Reported Outcomes Version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE) in a Chemoradiotherapy Cooperative Group Multicenter Clinical Trial. Int J Radiat Oncol Biol Phys. 2017 6 1; 98(2):409–418. [PubMed: 28463161]
10. Basch E, Dueck AC. Patient-reported outcome measurement in drug discovery: a tool to improve accuracy and completeness of efficacy and safety data. Expert Opin Drug Discov. 2016 8;11(8):753–8. [PubMed: 27310432]
11. Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst. 1993;85(5):365–376. [PubMed: 8433390]
12. Gundy CM, Fayers PM, Groenvold M,et al. Comparing higher order models for the EORTC QLQ-C30. Qual Life Res. 2012 11;21(9):1607–17. [PubMed: 22187352]
13. Jonckheere AR. A distribution-free k-sample test against ordered alternatives. Biometrika 1954; 41:133–45.
14. Gounder MM, Mahoney MR, Van Tine BA, et al. Sorafenib for Advanced and Refractory Desmoid Tumors. N Engl J Med. 2018 12 20;379(25):2417–2428. [PubMed: 30575484]
15. Basch EM, Scholz M, de Bono JS, et al. Cabozantinib Versus Mitoxantrone-prednisone in Symptomatic Metastatic Castration-resistant Prostate Cancer: A Randomized Phase 3 Trial with a Primary Pain Endpoint. Eur Urol. 2018 12 4. pii: S0302-2838(18)30932-1.
16. Kim J, Singh H, Ayalew K, et al. Use of PRO Measures to Inform Tolerability in Oncology Trials: Implications for Clinical Review, IND Safety Reporting, and Clinical Site Inspections. Clin Cancer Res. 2018;24(8):1780–1784. [PubMed: 29237718]
17. McFatrich M, Brondon J, Lucas NR, et al. Mapping child and adolescent self-reported symptom data to clinician-reported adverse event grading to improve pediatric oncology care and research. Cancer. 2019 9 25.
18. Reeve BB, Mitchell SA, Dueck AC, et al. Recommended patient-reported core set of symptoms to measure in adult cancer treatment trials. J Natl Cancer Inst. 2014 7 8;106(7). pii: dju129.

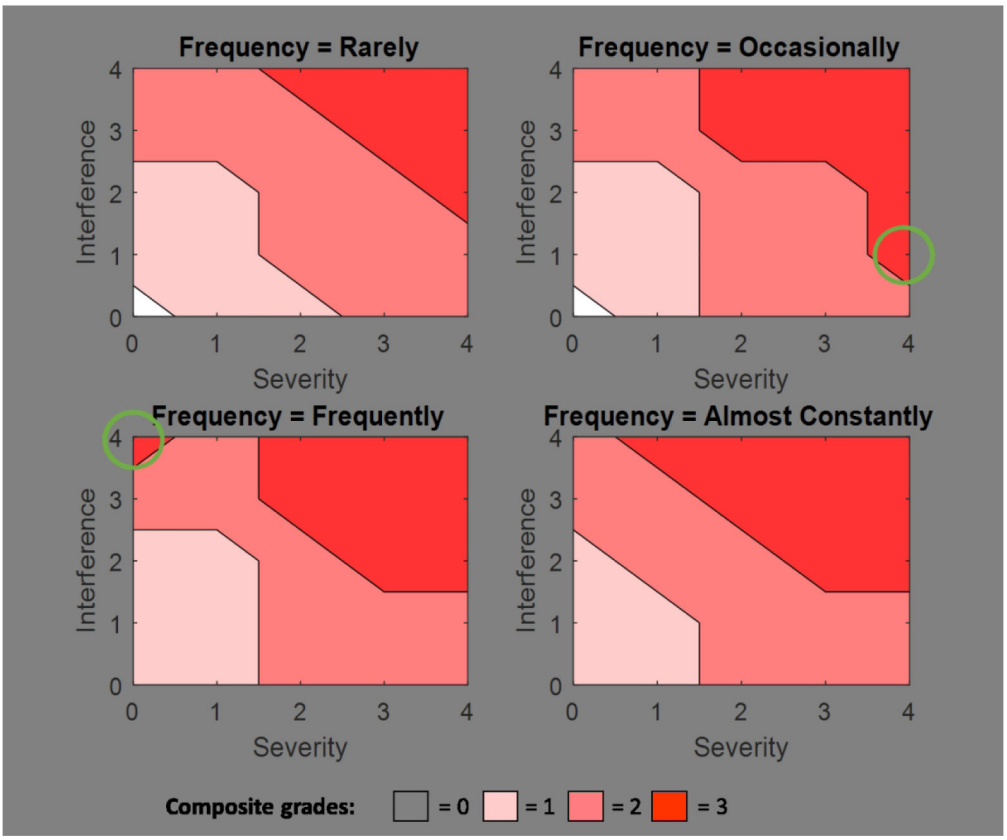

Figure 1.

Contour plots of PRO-CTCAE composite grades by frequency, severity, and interference scores. Circles show score combinations that have composite grades that were directionally inconsistent with the composite grades of neighboring score combinations.

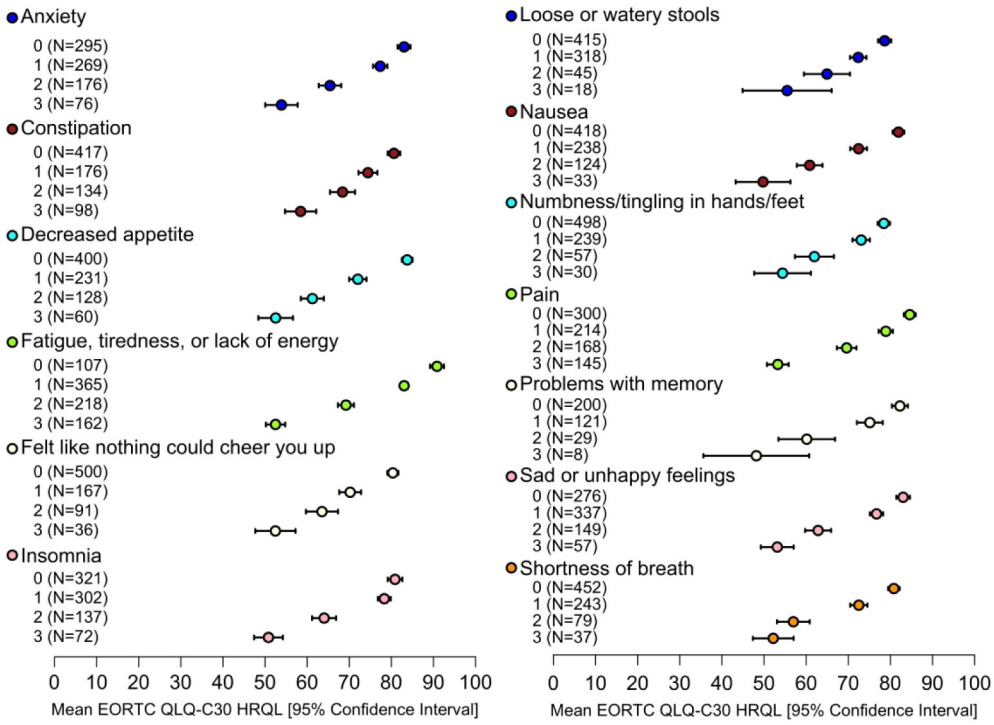

Figure 2.

Mean EORTC QLQ-C30 Health-Related Quality of Life (HRQL) Summary Scores with 95% confidence intervals by PRO-CTCAE composite grades for PRO-CTCAE symptomatic adverse events matching the recommended core symptoms of Reeve et al.[18] Higher EORTC QLQ-C30 scores indicate better health-related quality of life.

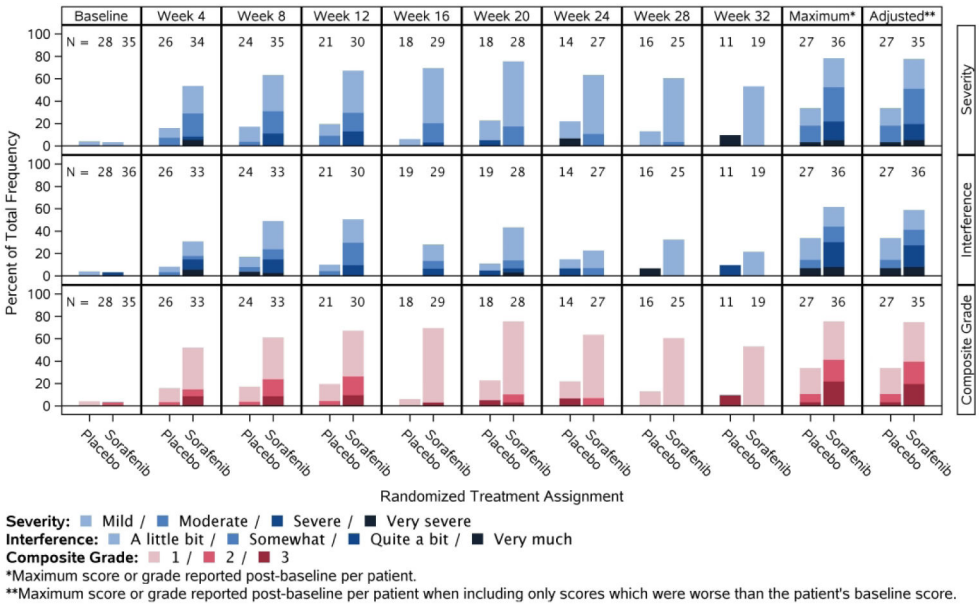

Figure 3.

Distribution of PRO-CTCAE data (individual severity and interference scores and composite grades) for hand-foot syndrome in A091105 at each assessment time point.

Table 1.

Portion of the Data Collection Form in Step 1 showing a subset of 10 combinations of threeway combinations of PRO-CTCAE Frequency, Severity, and Interference scores. See Supplemental Table S2 for all 179 combinations across various combinations of 1, 2, and 3 attributes of the PRO-CTCAE.

PRO-CTCAE FREQUENCY	PRO-CTCAE SEVERITY	PRO-CTCAE INTERFERENCE	ASSIGN CTCAE GRADE
Occasionally (2)	None (0)	Not at all (0)	
Occasionally (2)	None (0)	A little bit (1)	
Occasionally (2)	None (0)	Somewhat (2)	
Occasionally (2)	None (0)	Quite a bit (3)	
Occasionally (2)	None (0)	Very much (4)	
Occasionally (2)	Mild (1)	Not at all (0)	
Occasionally (2)	Mild (1)	A little bit (1)	
Occasionally (2)	Mild (1)	Somewhat (2)	
Occasionally (2)	Mild (1)	Quite a bit (3)	
Occasionally (2)	Mild (1)	Very much (4)	

Table 2.

PRO-CTCAE score combinations with <75% consensus on mapping to composite grades by 20 clinical investigators in Step 1, and results of anonymous voting on presented composite grades to clinical investigators in Alliance for Clinical Trials in Oncology national meetings in Step 2.

PRO-CTCAE individual verbal response scores			Composite CTCAE grades selected by clinical investigators in Step 1 with <75% consensus (N)	Composite CTCAE grades selected by clinical investigators in Step 2 (N)[†]	Clinical plausibility of score combination
Frequency	**Severity**	**Interference**			
Rarely (1)	None (0)	Not at all (0)	0 (N=8/20 [40%]) 1 (N=12/20 [60%])	0 (N=32/46 [70%])[‡] 1 (N=14/46 [30%])	Impossible
Rarely (1)	Moderate (2)	Not at all (0)	1 (N=8/20 [40%]) 2 (N=12/20 [60%])	1 (N=49/61 [80%])[‡] (N=12/61 [20%])	Possible
Occasionally (2)	None (0)	Not at all (0)	0 (N=6/20 [30%]) 1 (N=14/20 [70%])	0 (N=37/61 [61%])[‡] 1 (N=23/61 [38%]) 3 (N=1/61 [2%])	Impossible
Frequently (3)	None (0)	Very Much (4)	0 (N=1/20 [5%]) 2 (N=5/20 [25%]) 3 (N=14/20 [70%])	1 (N=4/42 [10%]) 2 (N=8/42 [19%]) 3 (N=15/42 [36%]) [*‡] Impossible (N=5/42 [12%]) Unsure (N=10/42 [24%])	Impossible
Almost Constantly (4)	Mild (1)	Somewhat (2)	1 (N=6/20 [30%]) 2 (N=14/20 [70%])	1 (N=10/61 [16%]) 2 (N=44/61 [72%])[‡] 3 (N=7/61 [11%])	Possible
Almost Constantly (4)	Mild (1)	Very Much (4)	2 (N=6/20 [30%]) 3 (N=14/20 [70%])	1 (N=3/53 [6%] 2 (N=17/53 [32%]) 3 (N=32/53 [60%])[‡] Impossible (N=1/53 [2%])	Unlikely
Occasionally (2)	Moderate (2)	–	1 (N=6/20 [30%) 2 (N=14/20 [70%])	1 (N=12/39 [31%]) 2 (N=25/39 [64%])[‡] Unsure (N=2/39 [5%])	Possible
-	Moderate (2)	Somewhat (2)	1 (N=7/20 [35%]) 2 (N=13/20 [65%])	1 (N=2/25 [8%]) 2 (N=23/25 [92%])[‡]	Possible
-	Moderate (2)	Very Much (4)	2 (N=7/20 [35%]) 3 (N=13/20 [65%])	2 (N=8/26 [31%]) 3 (N=18/26 [69%])[‡]	Possible
-	Severe (3)	Quite a bit (3)	2 (N=5/20 [25%]) 3 (N=13/20 [65%]) No response (N=2/20 [10%])	2 (N=3/27 [11%]) 3 (N=24/27 [89%])[‡]	Possible
–	Moderate (2)	–	1 (N=5/20 [25%]) 2 (N=14/20 [70%]) No response (N=1/20 [5%])	1 (N=8/35 [23%]) 2 (N=26/35 [74%])[‡] Unsure (N=1/35 [3%])	Possible

PRO-CTCAE individual verbal response scores			Composite CTCAE grades selected by clinical investigators in Step 1 with <75% consensus (N)	Composite CTCAE grades selected by clinical investigators in Step 2 (N)†	Clinical plausibility of score combination
Frequency	**Severity**	**Interference**			
–	Severe (3)	–	1 (N=5/20 [25%]) 3 (N=14/20 [70%]) No response (N=1/20 [5%])	2 (N=7/44 [16%]) 3 (N=31/44 [70%])‡ Unsure (N=6/44 [14%])	Possible

Blank cells indicate a missing attribute (i.e., two- and one-attribute combinations).

*Subsequently modified to a composite grade of 2 in Step 3 to ensure directional consistency of grading.

†Denominators varied depending on the number of investigators attending presentations of various scenarios.

‡Composite grades receiving the majority of votes.

Table 3.

Toxicity table for phase III trial Alliance A091105 (sorafenib vs. placebo in patients with advanced desmoid tumors) showing rates of symptomatic adverse events >0 and ≥3 for PRO-CTCAE individual items (A) and for composite grades (B).

	N		With baseline adjustment >0			With baseline adjustment ≥3		
	Sorafenib	Placebo	Sorafenib N %	Placebo N %	Fisher P	Sorafenib N %	Placebo N %	Fisher P
(A) PRO-CTCAE Individual Item Analysis								
Constipation severity	35	26	12 (34%)	13 (50%)	0.29	5 (14%)	5 (19%)	0.73
Decreased appetite severity	36	27	22 (61%)	10 (37%)	0.08	4 (11%)	2 (7%)	0.69
Decreased appetite interference	36	27	17 (47%)	8 (30%)	0.20	2 (6%)	1 (4%)	1.00
Diarrhea frequency	35	27	23 (66%)	10 (37%)	0.04	9 (26%)	3 (11%)	0.20
Fatigue severity	36	27	21 (58%)	17 (63%)	0.80	12 (33%)	10 (37%)	0.79
Fatigue interference	36	26	17 (47%)	13 (50%)	1.00	10 (28%)	7 (27%)	1.00
Hand/foot syndrome severity	35	27	27 (77%)	9 (33%)	<0.001	7 (20%)	1 (4%)	0.12
Hand/foot syndrome interference	36	27	21 (58%)	9 (33%)	0.07	10 (28%)	2 (7%)	0.05
Insomnia severity	35	27	18 (51%)	9 (33%)	0.20	6 (17%)	2 (7%)	0.45
Insomnia interference	35	26	15 (43%)	11 (42%)	1.00	5 (14%)	4 (15%)	1.00
Mouth or throat sores severity	36	27	17 (47%)	9 (33%)	0.31	5 (14%)	0 (0%)	0.07
Nausea frequency	36	27	21 (58%)	6 (22%)	0.005	5 (14%)	2 (7%)	0.69
Nausea severity	36	27	22 (61%)	8 (30%)	0.02	4 (11%)	4 (15%)	0.72
Pain frequency	35	26	16 (46%)	10 (38%)	0.61	13 (37%)	6 (23%)	0.28
Pain severity	35	26	16 (46%)	11 (42%)	1.00	9 (26%)	8 (31%)	0.78
Pain interference	36	26	19 (53%)	12 (46%)	0.80	13 (36%)	5 (19%)	0.17
Rash presence	35	27	24 (69%)	8 (30%)	0.004	–	–	–
Vomiting frequency	35	27	11 (31%)	6 (22%)	0.57	0 (0%)	1 (4%)	0.44
Vomiting severity	35	27	10 (29%)	5 (19%)	0.39	2 (6%)	1 (4%)	1.00
(B) PRO-CTCAE Composite Grade Analysis								
Constipation	35	26	11 (31%)	12 (46%)	0.29	4 (11%)	4 (15%)	0.71
Decreased appetite	36	27	18 (50%)	11 (41%)	0.61	1 (3%)	1 (4%)	1.00
Diarrhea	35	27	20 (57%)	8 (30%)	0.04	0 (0%)	1 (4%)	0.44
Fatigue	36	26	18 (50%)	12 (46%)	0.80	8 (22%)	6 (23%)	1.00
Hand/foot syndrome	35	27	26 (74%)	9 (33%)	0.002	7 (20%)	1 (4%)	0.12
Insomnia	35	26	16 (46%)	9 (35%)	0.44	5 (14%)	2 (8%)	0.69
Mouth or throat sores	36	27	17 (47%)	9 (33%)	0.31	5 (14%)	0 (0%)	0.07
Nausea	36	27	21 (58%)	7 (26%)	0.01	2 (6%)	3 (11%)	0.64
Pain	35	26	14 (40%)	10 (38%)	1.00	8 (23%)	6 (23%)	1.00
Rash	35	27	24 (69%)	8 (30%)	0.004	-	-	-

	N		With baseline adjustment >0			With baseline adjustment ≥3		
	Sorafenib	**Placebo**	**Sorafenib N %**	**Placebo N %**	**Fisher P**	**Sorafenib N %**	**Placebo N %**	**Fisher P**
Vomiting	35	27	10 (29%)	6 (22%)	0.77	0 (0%)	1 (4%)	0.44

Blank cells indicate that numbers and percentages are not applicable for Rash, a binary present/not present item.