Applying the Toxicity Index to Patient-Reported Symptom Data: An Example Using the European Organization for Research and Treatment of Cancer Colorectal Cancer–Specific Quality of Life Questionnaire

Ron D. Hays, PhD¹; Patricia A. Ganz, MD²,³; Karen L. Spritzer, BS¹; and André Rogatko, PhD⁴

¹Department of Medicine, Division of General Internal Medicine & Health Services Research, University of California, Los Angeles, California; ²Department of Health Policy and Management, Fielding School of Public Health, University of California, Los Angeles, California; ³Center for Cancer Prevention and Control Research, Jonsson Comprehensive Cancer Center, University of California, Los Angeles, California; and ⁴Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, California

ABSTRACT

**Purpose:** The toxicity index (TI) is a summary index that accounts for toxicity grades associated with cancer symptoms that is more sensitive than other toxicity systems to treatment differences. The TI can be used with patient-reported symptoms but requires that scores for different items represent equivalent severity. The purpose of this article is to provide an example of scoring patient-reported symptoms that satisfies the requirement of equivalent symptom severity.

**Methods:** A sample of 1232 adults with rectal cancer from a Phase III clinical trial self-reported 18 symptoms on the European Organization for Research and Treatment of Cancer colorectal cancer measure using a 4-category response scale (not at all, a little bit, quite a bit, or very much). The participants were 22 to 85 years of age (mean age, 57 years), 30% were female, 85% were non-Hispanic white, 59% had stage II cancer, and 41% had stage III cancer. A recoded TI was created using item response theory category thresholds.

**Findings:** The recoded TI had larger rank-order correlations than the original TI with Karnofsky performance status index, hemoglobin level, symptom bother, and other aspects of health-related quality of life.

**Implications:** Recoding items based on category thresholds yielded a more valid TI score that can be used to summarize adverse events. (Clin Ther. 2021;XX:XXX–XXX) © 2021 Elsevier HS Journals, Inc. (Clin Ther. 2021;000:1–8.) © 2021 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

**Key words:** adverse events, cancer, symptom scoring, toxicity index.

INTRODUCTION

Toxicity data consist of treatment-attributable adverse events (AEs) graded as 0, 1, 2, 3, 4, or 5 for each of the 790 AE terms, grouped in 26 system organ classes according to Common Terminology Criteria for Adverse Events (CTCAE), version 4.0.¹ Toxicity grading by clinicians is a standard component of cancer clinical trial data collection. Grade 0 AEs represent the absence of toxicity. The toxicity index (TI) was inspired by hash functions and provides a summary of all n observed toxicity grades.² Each of the n toxicity grades X_i(i = 1, ..., n) for an individual is represented in...
descending order: \( X_1 \geq X_2 \geq \cdots \geq X_n \). An individual’s TI score is a function of the ordered toxicity grades:

\[
TI = \sum_{i=1}^{n} \frac{X_i}{\prod_{j=1}^{i-1} (1 + X_j)}
\]

Any TI ≥3 corresponds to a dose-limiting toxicity, and the maximum toxicity grade is the integer part of the final score. For example, a TI of 3.0 indicates a single grade 3 toxic event, whereas a TI of 3.5 means that the patient experienced at least 1 grade 3 toxic event plus additional toxic events. All toxicity grades are represented in the score, although lower grades contribute less to the final score than higher grades.

The TI has potential to be used with patient-reported symptom measures. However, the TI assumes equal levels of impact for the item response categories for different symptoms. CTCAE grades are treated as equivalent across symptoms. It also may be acceptable for patient-reported symptoms measured using a response scale, such as not bothered at all, a little bit bothered, somewhat bothered, bothered quite a bit, and bothered very much. However, the TI approach is not ideal for summarizing patient reports of symptoms when severity is not captured. For example, use of the TI with reports about frequency of symptoms or extent to which symptoms occur may be problematic because severity may differ (eg, runny nose vs vomiting).

Category response curves provide information about item response options in multi-item scales that identifies where they fall on the underlying continuum. Item response theory can be used to get estimates of threshold parameters that represent the underlying trait level necessary to respond above each threshold with 0.50 probability.\(^3\) These thresholds indicate the relative severity by item response options.

This article presents a comparison of scoring the TI for a patient-reported symptom measure scored assuming equal distances between response categories versus scoring based on item thresholds in the National Surgical Adjuvant Breast and Bowel Project R-04 rectal cancer clinical trial.

METHODOLOGY
Study Design and Sample

Eligible patients were diagnosed with surgically resectable stage II or III rectal adenocarcinoma. A total of 1608 patients participated in the Phase III clinical trial of rectal cancer (NCT00058474) between 2004 and 2010.\(^4\)\(^5\) All patients who spoke English, French, or Spanish were invited to complete a questionnaire at baseline before randomization to treatment. If the patient was not accessible in person, staff were encouraged to mail the questionnaire to the patient or collect responses by telephone.

The trial was approved by the local institutional review boards, and all patients provided written informed consent. The secondary analyses reported here were determined to be exempt by the Cedars Sinai and UCLA institutional review boards. The sample consisted of 1232 adults with complete data for 18 symptom items (see Measures) analyzed. Adults were 22 to 85 years of age (mean age, 57 years), 30% were female, 85% were non-Hispanic white, 59% had stage II cancer, and 41% had stage III cancer (Table I).

Measures

The baseline patient-reported survey included 112 questions. The focus of the analyses are 18 symptoms (items 60 to 77 on the baseline survey) assessed in the European Organization for Research and Treatment of Cancer colorectal cancer–specific quality of life questionnaire (QLQ-CR38).\(^6\) The QLQ CR-38 assesses the extent to which symptoms were experienced in the past week: not at all, a little bit, quite a bit, and very much (Table II). A higher score indicates a greater extent of experiencing symptoms. Also included in the baseline survey was the Functional Assessment of Cancer Therapy–Colorectal Trial Outcomes Index (FACT-C T0I), the Functional Assessment of Cancer Therapy–Gynecologic Oncology Group–Neurotoxicity 13 (FACT-GOG-NTX-13), the 36-Item Short Form Health Survey (SF-36) version 2 vitality scale, and a 17-item symptom checklist (SCL-17)\(^7\)\(^8\)\(^9\)\(^10\) Clinical measures analyzed were the maximum AE grade, hemoglobin level, and the Karnofsky performance status index.

Statistical Analysis

Our primary analysis used baseline survey data, but we looked at consistency of results with 1-year postsurgery survey data. The standard coding of the QLQ-CR38 items is as follows: 0, not at all; 1, a little bit; 2, quite a bit; and 3, very much. We report internal consistency reliability\(^11\) and item-scale correlations for the 18-item QLQ-CR38 symptom score using this scoring. We used categorical confirmatory factor analysis with diagonally weighted least squares to evaluate whether the items were sufficiently unidimensional to
estimate response category thresholds using the item response theory graded response model. Because of content overlap (local dependency) among the QLQ-CR38 symptoms, we included 6 residual correlations (item pairs: 60 and 61, 63 and 64, 72 and 73, 72 and 74, 73 and 74, as well as 76 and 77). We evaluated model fit using the comparative fit index and the root mean square error of approximation. Comparative fit index values >0.95 and root mean square error of approximation values <0.06 are considered good fit.12

Category thresholds for the 18 items (3 thresholds per item) were estimated from the graded response model.13 The SEs around the thresholds were used to create 95% CIs. Overlapping CIs of the 3 thresholds for each item were identified. Threshold estimates were used to adjust the scoring of item responses. The 0 for not at all was preserved, but the distance between scores assigned for other response options were shifted based on differences in item thresholds.

The TI is scored so that a higher score represents a greater toxicity. We hypothesized positive correlations with measures scored so that a higher score is worse (maximum AE grade, SCL-17, and worried about health in the future) and negative correlations with measures scored so that a higher score is better (Karnofsky performance status, hemoglobin, FACT-C TOI, FACT-GOG-NTX-13, and SF-36 version 2 vitality scale). Spearman rank-order correlations of the TI with these variables were estimated.

Confirmatory factor analysis was conducted using Mplus 14 version 7 and all other analyses with SAS software, version 9.4, TSIM3 (SAS Institute, Cary, North Carolina).

## RESULTS

Internal consistency reliability for the 18 symptom item scale was 0.79 and item-scale correlations (corrected for item overlap with the scale total) ranged from 0.26 to 0.49. Sufficient unidimensionality of the 18 QLQ-CR38 symptom items was supported by the fit of the 1-factor confirmatory factor analysis model (comparative fit index = 0.962 and root mean square error of approximation = 0.054). The χ² was 587.385 with 129 df (P < 0.0001).

The original scoring of the QLQ-CR38 symptom items is given in Table III (0, not at all; 1, a little bit; 2, quite a bit; and 3, very much). Threshold estimates from the graded response model for each item are also given. To determine how to modify the original scoring of the QLQ-CR38 symptoms for the TI summary measure, we compared thresholds across items. We used as many integers as needed and no more than needed to reflect the variation in severity across symptoms indicated by the thresholds. We ended up needing to add 2 integers (4 and 5) to reflect variation in thresholds. For example, we scored responses of very much as 5 for item 60 but 4 for item 63 (bloating feeling in your abdomen) because the threshold between quite a bit and very much for the latter was smaller (mean [SE], 3.12 [0.20]) than the former (mean [SE], 4.43

### Table I. Sample characteristics.

| Characteristic                                      | Finding* (N = 1232) |
|-----------------------------------------------------|----------------------|
| Age, mean (SD) [range]. y                           | 75 (11) [22–85]      |
| Female sex                                          | 30                   |
| Hispanic                                            | 5                    |
| Non-Hispanic black                                  | 5                    |
| Non-Hispanic white                                  | 38                   |
| Non-Hispanic other/unknown race                     | 5                    |
| Body mass index, kg/m²*                              | 21.5 (range)         |
| <18.5 (underweight)                                 | 1                    |
| <25 (normal)                                        | 26                   |
| <30 (overweight)                                    | 36                   |
| ≥30 (obese)                                         | 37                   |
| Clinical stage, %                                   | 11                   |
| II                                                  | 59                   |
| III                                                 | 41                   |
| Karnofsky performance status, %†                    | 85                   |
| Fully active (90–100)                               | 15                   |
| Restricted (70–80) and ambulatory (50–60)           |                      |

* Data are presented as percentage of patients unless otherwise indicated.
† Fully active (90–100) indicates able to perform all pre-disease performance without restriction. Restricted (70–80) indicates restricted in physically strenuous activity but ambulatory. Ambulatory (K50–60) indicates ambulatory and capable of all self-care but unable to perform any work activities.
Clinical Therapeutics

Table II. QLQ-CR38 item frequencies on baseline survey.

| Item | % | | | | | | |
|------|---|---|---|---|---|---|---|
| | Not at All | A Little Bit | Quite a Bit | Very Much |
| During the past week | | | | | | | |
| Did you urinate frequently during the day? (item 60) | 31 | 38 | 26 | 5 |
| Did you urinate frequently during the night? (item 61) | 37 | 47 | 12 | 3 |
| Did you have pain when you urinated? (item 62) | 93 | 5 | 1 | 0.3 |
| Did you have a bloated feeling in your abdomen? (item 63) | 62 | 28 | 8 | 3 |
| Did you have abdominal pain? (item 64) | 72 | 22 | 5 | 1 |
| Did you have pain in your buttocks? (item 65) | 60 | 22 | 10 | 9 |
| Were you bothered by gas (flatulence)? (item 66) | 38 | 41 | 16 | 5 |
| Did you belch? (item 67) | 57 | 36 | 5 | 2 |
| Have you lost weight? (item 68) | 57 | 33 | 7 | 3 |
| Did you have a dry mouth? (item 69) | 73 | 20 | 4 | 2 |
| Have you had thin or lifeless hair as a result of your disease? (item 70) | 76 | 9 | 0.4 | 0.2 |
| Did food and drink taste different than usual? (item 71) | 90 | 8 | 2 | 1 |
| Have you felt physically less attractive as a result of your disease or treatment? (item 72) | 80 | 14 | 3 | 2 |
| Have you been feeling less feminine/masculine as a result of your disease or treatment? (item 73) | 81 | 14 | 3 | 2 |
| Have you been dissatisfied with your body? (item 74) | 66 | 26 | 6 | 2 |
| Were you worried about your health in the future? (item 75) | 19 | 47 | 22 | 13 |
| During the past 4 weeks | | | | | | | |
| To what extent were you interested in sex? (item 76) | 27 | 37 | 23 | 12 |
| To what extent were you sexually active (with or without intercourse)? (item 77) | 39 | 37 | 19 | 6 |

QLQ-CR38 = European Organization for Research and Treatment of Cancer colorectal cancer-specific quality of life questionnaire.

*This question was reversed scored when creating the toxicity index so that a higher score represented greater frequently of symptoms.

The TI index was computed from the original scoring, and then the revised TI was scored based on item category thresholds.

The Spearman rank-order correlations between the TI and revised TI at baseline was 0.65. The revised TI was more strongly associated with other variables than was the TI (Table IV). The revised TI was significantly more highly associated than the TI with 6 variables: (1) Karnofsky performance status index, (2) hemoglobin level, (3) SCL-17 scale, (4) FACT-C TOI, (5) FACT-GOG-NTX-13 score, and (6) SF-36 version 2 vitality scale. We found similar results for survey data collected 1 year after surgery (Table V).

DISCUSSION

The value of patient-reported symptoms has been documented for >2 decades. The work reported here is consistent with the ongoing efforts to incorporate the patient’s voice into the assessment of AEs in cancer clinical trials. For example, Smith et al noted that some of the AEs in the CTCAE are best assessed by asking the patient.

The revised TI was significantly correlated with several patient-reported measures (SCL-17, FACT-C TOI, FACT-GOG-NTX-13 score, and SF-36 version 2 vitality scale), maximum AE grade, Karnofsky performance status index, and hemoglobin level. The
Table III. Original scoring of items and recoding of each item.

| Item and Threshold | 0 (Not at All) | 1 (A Little Bit) | 2 (Quite a Bit) | 3 (Very Much) |
|-------------------|----------------|------------------|----------------|---------------|
| Item 60           | 0              | 1                | 2              | 4.43^a        |
| Original score    |                | -1.28^e          | 1.21^c,a       |               |
| Threshold         |                |                  |                |               |
| Item 61           | 0              | 1                | 3              | 4.30^a,b      |
| Original score    |                | -0.76^a          | 2.18^a         |               |
| Threshold         |                |                  |                |               |
| Item 62           | 0              | 2                | 4              | 5             |
| Original score    | 2.41^a         | 3.77^a           |               |               |
| Threshold         | 0.47^c         | 1.98^a           |               |               |
| Item 63           | 0              | 1                | 3              | 3.12^b,c      |
| Original score    | 0.85^b         | 2.28^a           |               |               |
| Threshold         | 0.35^c         | 1.50^c           |               |               |
| Item 64           | 0              | 1                | 3              | 3             |
| Original score    | 0.35^c         | 1.50^c           |               |               |
| Threshold         | 0.35^c         | 1.50^c           |               |               |
| Item 65           | 0              | 1                | 2              | 3             |
| Original score    | 0.31^c         | 2.61^a           |               |               |
| Threshold         | 1.00^b         | 2.63^a           |               |               |
| Item 66           | 0              | 1                | 2              | 4             |
| Original score    | 0.31^c         | 2.61^a           |               |               |
| Threshold         | 1.00^b         | 2.63^a           |               |               |
| Item 67           | 0              | 1                | 4              | 5             |
| Original score    | 0.35^c         | 3.37^a           |               |               |
| Threshold         | 0.35^c         | 3.37^a           |               |               |
| Item 68           | 0              | 1                | 4              | 5             |
| Original score    | 0.31^c         | 2.61^a           |               |               |
| Threshold         | 1.00^b         | 2.63^a           |               |               |
| Item 69           | 0              | 2                | 4              | 5             |
| Original score    | 2.84^a         | 4.41^a           |               |               |
| Threshold         | 2.84^a         | 4.41^a           |               |               |
| Item 70           | 0              | 2                | 4              | 5             |
| Original score    | 1.68^c,a       | 2.87^a           |               |               |
| Threshold         | 1.68^c,a       | 2.87^a           |               |               |
| Item 71           | 0              | 2                | 4              | 5             |
| Original score    | 1.09^b         | 2.15^a           |               |               |
| Threshold         | 1.09^b         | 2.15^a           |               |               |
| Item 72           | 0              | 2                | 3              | 4             |
| Original score    | 1.14^b         | 2.18^a           |               |               |
| Threshold         | 1.14^b         | 2.18^a           |               |               |
| Item 73           | 0              | 2                | 3              | 4             |
| Original score    | 1.14^b         | 2.18^a           |               |               |
| Threshold         | 1.14^b         | 2.18^a           |               |               |

(continued on next page)
greater relative validity of the revised TI compared with the TI was supported by consistently larger associations with other variables as hypothesized.

One limitation of the TI is that it requires rank-based analysis because it does not follow any well-known probability distribution, such as the normal distribution. However, it contains more information than other toxicity analysis methods by accounting for both the multiplicity and severity of toxic effects, without losing the natural interpretability of the maximum grade approach. This added information yields greater power in detecting treatment differ-

Table III. (continued)

| Item and Threshold | 0 (Not at All) | 1 (A Little Bit) | 2 (Quite a Bit) | 3 (Very Much) |
|--------------------|---------------|-----------------|---------------|--------------|
| Item 74            |               |                 |               |              |
| Original score     |               |                 |               |              |
| Threshold          | 0.61<sup>c</sup> | 2.26<sup>a</sup> | 3.29<sup>a,b</sup> |              |
| Item 75            |               |                 |               |              |
| Original score     |               |                 |               |              |
| Threshold          | -2.03<sup>e</sup> | 0.98<sup>c</sup> | 2.80<sup>b,c</sup> |              |
| Item 76            |               |                 |               |              |
| Original score     |               |                 |               |              |
| Threshold          | -3.96<sup>f</sup> | -1.21<sup>e</sup> | 1.95<sup>c</sup> |              |
| Item 77            |               |                 |               |              |
| Original score     |               |                 |               |              |
| Threshold          | -5.59<sup>f</sup> | -2.21<sup>e</sup> | 0.97<sup>d</sup> |              |

* See Table I for item wording. Thresholds in rows that share a superscript letter do not differ significantly from one another. For example, the mean (SE) threshold between quite a bit and very much for item 60 (urinate frequently during the day) was 4.43 (0.41), whereas the mean (SE) threshold between not at all and a little bit was 1.28 (0.14), and these thresholds were significantly different.

Table IV. Spearman correlations at baseline for 2 TI variants created from the QLQ-CR38.

| Variable                                | Spearman Correlation (P) |
|-----------------------------------------|--------------------------|
|                                         | Original TI              | Recoded TI<sup>*</sup> |
| Maximum adverse event grade             | 0.06 (0.0376)            | 0.08 (0.0036)            |
| Karnofsky performance status index      | -0.07 (0.0091)           | -0.13 (<0.0001)          |
| Hemoglobin level                        | -0.08 (0.0049)           | -0.15 (<0.0001)          |
| SCL-17                                  | -0.02 (0.4814)           | 0.16 (<0.0001)           |
| FACT-C TOI                              | 0.01 (0.6021)            | -0.17 (<0.0001)          |
| FACT-GOG-NTX-13 score                   | -0.03 (0.3840)           | -0.18 (<0.0001)          |
| SF-36 version 2 vitality scale (4 items)| -0.05 (0.0918)           | -0.21 (<0.0001)          |

FACT-C TOI = Functional Assessment of Cancer Therapy–Colorectal Trial Outcomes Index; FACT-GOG-NTX-13 = Functional Assessment of Cancer Therapy–Gynecologic Oncology Group–Neurotoxicity 13; SCL-17 = 17-item symptom checklist; SF-36 = 36-item Short Form Health Survey; QLQ-CR38 = European Organization for Research and Treatment of Cancer colorectal cancer-specific quality of life questionnaire.

* Recoded TI is based on recoding of item scores using graded response model threshold estimates.
ences than maximum grade and average toxicity approaches.17,18

CONCLUSIONS
This article provides a prototype of how the TI can be applied to patient-reported symptom measures and illustrates the value of adjusting item scoring to account for different levels of underlying symptom severity. The method used to adjust scores is not the only or necessarily the best approach. Future research and applications are needed to evaluate similar and different strategies to adjust category scoring of polytomous symptom items to satisfy the underlying assumption of equivalence across items implicit in the scoring of the TI.

ACKNOWLEDGMENTS
Author contributions are as follows: Ron D. Hays: conceptualization, formal analysis, methodology, and writing original draft; Patricia A. Ganz: data curation, funding acquisition, project administration, resources, supervision, and review and editing; Karen L. Spritzer: formal analysis, software supervision, and review and editing of manuscript; and André Rogatko: funding acquisition, project administration, resources, review and editing of manuscript.

FUNDING SOURCES
This work was supported in part by grant 1U01CA232859-01 from the National Cancer Institute, National Institutes of Health. The sponsor did not have a role in the study design, collection, analysis, interpretation of data, or writing of the manuscript.

REFERENCES
1. Department of Health and Human Services. Common Terminology Criteria for Adverse Events (CTCAE). Version 4.0. National Institutes of Health. National Cancer Institute; 2009.
2. Rogatko A, Babb JS, Wang H, Slifker MJ, Hudes GR. Patient characteristics compete with dose as predictors of acute treatment toxicity in early phase clinical trials. Clin Cancer Res. 2004;10:4645–4651.
3. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care. 2000;38(Suppl) II-28–II-42.
4. Allegra CJ, Yothers G, O’Connell MJ, et al. Neoadjuvant 5-FU or capcitabine plus radiation with or without oxaliplatin in rectal cancer patients: a phase III randomized clinical trial. J Natl Cancer Inst. 2015;107:djv248.
5. Russell MM, Ganz PA, Lopa S, et al. Comparative effectiveness of sphincter-sparing surgery versus abdominoperineal resection in rectal cancer: patient-reported outcomes in the National Surgical

Table V. Spearman correlations at 1 year after surgery for 2 TI variants created from the QLQ-CR38.

| Variable | Spearman correlation (P) |
|----------|--------------------------|
|          | Original TI | Recoded TI  |
| Maximum adverse event grade | 0.06 (0.0608) | 0.06 (0.1200) |
| Karnofsky performance status index | 0.00 (0.9492) | −0.12 (0.0008) |
| Hemoglobin level | −0.09 (0.0113) | −0.14 (<0.0001) |
| SCL-17 | −0.07 (0.0395) | 0.09 (0.0096) |
| FACT-C TOI | 0.00 (0.9499) | −0.14 (<0.0001) |
| FACT-GOG-NTX-13 score | 0.04 (0.2114) | −0.09 (0.0187) |
| SF-36 version 2 vitality scale (4 items) | −0.05 (0.1130) | −0.17 (<0.0001) |

FACT-C TOI = Functional Assessment of Cancer Therapy–Colorectal Trial Outcomes Index; FACT-GOG-NTX-13 = Functional Assessment of Cancer Therapy–Gynecologic Oncology Group–Neurotoxicity 13; SCL-17 = 17-item symptom checklist; SF-36 = 36-item Short Form Health Survey; QLQ-CR38 = European Organization for Research and Treatment of Cancer colorectal cancer–specific quality of life questionnaire.

* Recoded TI is based on recoding of item scores using graded response model threshold estimates.
Clinical Therapeutics

Adjuvant Breast and Bowel Project randomized trial R04. 
*Ann Surg.* 2015;261:144–148.

6. Sprangers MA, te Velde A, Aaronson NK. The construction and testing of the EORTC colorectal cancer-specific quality of life questionnaire module (QLQ-CR38). European Organization for Research and Treatment of Cancer Study Group on Quality of Life. *Eur J Cancer.* 1999;35:238–247.

7. Kopec JA, Yothers G, Ganz PA, et al. Quality of life in operable colon cancer patients receiving oral compared with intravenous chemotherapy: results from National Surgical Adjuvant Breast and Bowel Project Trial C-06. *J Clin Oncol.* 2007;25:424–430.

8. Maglinte G, Hays RD, Kaplan RM. U.S. general population norms for telephone administration of the SF-36v2. *J Clin Epidemiol.* 2012;65:497–450.

9. Pearman TP, Beaumont JL, Mroczek D, O’Connor M, Cella D. Validity and usefulness of a single-item measure of patient-reported bother from side effects of cancer therapy. *Cancer.* 2018;12:991–997.

10. Ward WL, Hahn EA, Mo F, et al. Reliability and validity of the Functional Assessment of Cancer Therapy-Colorectal (FACT-C) quality of life instrument. *Qual Life Res.* 1999;8:181–195.

11. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika.* 1951;16:297–334.

12. Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling.* 1999;6:1–55.

13. Samejima F. *Handbook of Modern Item Response Theory.* Graded response model. New York, NY: Springer; 1996.

14. Muthén L, Muthén B. *Mplus User’s Guide.* CA: Los Angeles; 2001.

15. Justice AC, Rabeneck L, Hays RD, Wu AW, Bozzette SA. The sensitivity, reliability, and clinical validity of provider-reported symptoms: a comparison with self-report. *J AIDS.* 1999;21:126–133.

16. Smith AW, Mitchell SA, De Aguilar CK, et al. News from the NIH: Person-centered outcomes measurement: NIH-supported measurement systems to evaluate self-assessed health, functional performance, and symptom toxicity. *Transl Behav Med.* 2016;6:470–474.

17. Gresham G, Diniz MA, Raznee ZS, et al. Evaluating treatment tolerability in cancer clinical trials using the toxicity index. *J Natl Cancer Inst.* 2020;11:1266–1274.

18. Razae ZS, Amini AA, Diniz MS, et al. On the properties of the toxicity index and its statistical efficiency. *Stat Med.* 2021;40:1535–1552.

Address correspondence to: Ron D. Hays, PhD, Department of Medicine, Division of General Internal Medicine & Health Services Research, University of California, Los Angeles, 1100 Glendon Avenue, Suite 850, Los Angeles, CA 90024.E-mail: drhays@ucla.edu.