

# Exploring differences in adverse symptom event grading thresholds between clinicians and patients in the clinical trial setting

Thomas M. Atkinson<sup>1</sup> · Lauren J. Rogak<sup>2</sup> · Narre Heon<sup>2</sup> · Sean J. Ryan<sup>2,3</sup> · Mary Shaw<sup>1</sup> · Liora P. Stark<sup>1</sup> · Antonia V. Bennett<sup>4</sup> · Ethan Basch<sup>2,4</sup> · Yuelin Li<sup>1</sup>

Received: 15 December 2016 / Accepted: 23 December 2016 / Published online: 16 January 2017  
© Springer-Verlag Berlin Heidelberg 2017

## Abstract

**Purpose** Symptomatic adverse event (AE) monitoring is essential in cancer clinical trials to assess patient safety, as well as inform decisions related to treatment and continued trial participation. As prior research has demonstrated that conventional concordance metrics (e.g., intraclass correlation) may not capture nuanced aspects of the association between clinician and patient-graded AEs, we aimed to characterize differences in AE grading thresholds between doctors (MDs), registered nurses (RNs), and patients using the Bayesian Graded Item Response Model (GRM).

**Methods** From the medical charts of 393 patients aged 26–91 ( $M=62.39$ ; 43% male) receiving chemotherapy, we retrospectively extracted MD, RN and patient AE ratings. Patients reported using previously developed Common Terminology Criteria for Adverse Events (CTCAE) patient-language adaptations called STAR (Symptom Tracking and Reporting). A GRM was fitted to calculate the latent grading thresholds between MDs, RNs and patients.

**Results** Clinicians have overall higher average grading thresholds than patients when assessing diarrhea, dyspnea, nausea and vomiting. However, RNs have lower grading

thresholds than patients and MDs when assessing constipation. The GRM shows higher variability in patients' AE grading thresholds than those obtained from clinicians.

**Conclusions** The present study provides evidence to support the notion that patients report some AEs that clinicians might not consider noteworthy until they are more severe. The availability of GRM methodology could serve to enhance clinical understanding of the patient symptomatic experience and facilitate discussion where AE grading discrepancies exist. Future work should focus on capturing explicit AE grading decision criteria from MDs, RNs, and patients.

**Keywords** Patient-reported outcomes · Adverse events · Clinical trials · Clinician–patient agreement · Item response theory · Neoplasms

## Introduction

Accurate capture and monitoring of symptomatic adverse events (AE) is essential in clinical trials and drug labeling to ensure patient safety and inform treatment-related decision-making (Basch 2010, 2014, 2016). In the United States, the standard approach to collecting this information as part of trials in oncology is clinician reporting using the Common Terminology Criteria for Adverse Events (CTCAE) (National Cancer Institute 2010), which allows licensed clinicians [i.e., medical doctors (MDs) and registered nurses (RNs)] to grade AEs based upon descriptive clinical criteria (e.g., Grade 3 nausea=inadequate oral caloric or fluid intake; tube feeding, TPN, or hospitalization indicated). The assignment of a given AE grade has implications for patient treatment and/or participation in clinical trials.

✉ Thomas M. Atkinson  
atkinsot@mskcc.org

<sup>1</sup> Department of Psychiatry and Behavioral Sciences, Memorial Sloan Kettering Cancer Center, 641 Lexington Ave., 7th Floor, New York, NY 10022, USA

<sup>2</sup> Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, 485 Lexington Ave., 2nd Floor, New York, NY 10017, USA

<sup>3</sup> City University of New York, New York, NY, USA

<sup>4</sup> University of North Carolina-Chapel Hill, Chapel Hill, NC, USA

The CTCAE includes multiple categories of AEs, including lab-based, which are generally sourced directly from lab reports (e.g., neutropenia); clinical measurement-based, which are typically evaluated and reported by clinicians (e.g., hypertension); symptom-based such as fatigue or nausea, which despite being amenable to patient reporting, are still primarily rated by clinicians (Basch et al. 2014).

The increased acceptance of the use of patient-reported outcomes (PROs), defined as the unfiltered direct report of a given symptom by a patient (Basch 2012; Trotti et al. 2007), to characterize the patient symptomatic experience has led to the US National Cancer Institute's initiative to develop a PRO version of the CTCAE (PRO-CTCAE) that will be used in future US-based clinical trials in oncology (Basch et al. 2014; Dueck et al. 2015; Hay et al. 2014). Given that both clinician- and patient-based reporting of symptomatic AEs will be commonplace in US-based oncology clinical trials, it is important to understand how these independent rating sources are associated before this information can be integrated into clinical practice.

As part of our prior work making use of conventional statistical metrics [e.g., intraclass correlation (ICC), Cohen's weighted  $\kappa$ ] to compare clinician and patient reports of AE severity using an ordinal response scale, we have observed that the inter-rater agreement is highly dependent on the prevalence of the AE; a high proportion of "asymptomatic" pairs of ratings (i.e., both ratings are 0, or not present) may lead to an inflated level of agreement, which may not necessarily be an accurate representation of the subset of patients who are experiencing any levels of the symptom (Atkinson et al. 2012, 2016).

An alternative Bayesian approach to the calculation of concordance known as the Graded Item Response Model (GRM) was recently proposed by Baldwin and colleagues (Baldwin et al. 2009). This approach utilized the underlying principles of the original Samejima GRM (Samejima 1997) in a Bayesian framework. In this example, patient hip fracture radiographs were independently judged by 12 orthopedic surgeons using a four-level classification of severity. Surgeons' hip fracture severity ratings were viewed from an Item Response Theory (IRT) perspective, in which each surgeon's severity rating was modeled as a scale item while radiographs from patients were considered as a sample from a latent continuum of hip fracture severity. This analytic framework allowed an IRT analysis on the raw rectangular dataset from 15 patients evaluated by 12 surgeons (likened to scale items). The item threshold parameters in the fitted Bayesian GRM represented the surgeons' decision cutoffs and the item discrimination represented how sensitive the surgeons' responses were with respect to changes in hip fracture severity. The authors found that the model-predicted decision cutoffs agreed with surgeons' severity ratings reasonably well. This example showed that the Bayesian GRM framework has a potential application for identifying how raters differ in their independent assessments, which may be subtle in the sense that such differences can be nuanced and highly contextual (e.g., concordance at low latent hip fracture severity, with discordance emerging at high latent hip fracture severity).

The present study applied this Bayesian GRM framework to measuring concordance between doctor (MD)-, registered nurse (RNs)-, and patient-based reporting of symptomatic AEs. We sought to model and further

**Table 1** Example of data entry structure

Patient ID	MD 1			MD 2			MD 264		
	MD	RN	Patient	MD	RN	Patient	MD	RN	Patient
1	1	1	2	N/A	N/A	N/A	N/A	N/A	N/A
2	2	1	3	N/A	N/A	N/A	N/A	N/A	N/A
3	1	2	2	N/A	N/A	N/A	N/A	N/A	N/A
4	N/A	N/A	N/A	N/A	N/A	N/A	0	1	2
5	N/A	N/A	N/A	1	1	1	N/A	N/A	N/A
6	N/A	N/A	N/A	0	1	2	N/A	N/A	N/A
7	N/A	N/A	N/A	1	2	2	N/A	N/A	N/A
–									
–									
–									
388	2	1	1	N/A	N/A	N/A	N/A	N/A	N/A
389	N/A	N/A	N/A	2	1	2	N/A	N/A	N/A
390	N/A	N/A	N/A	N/A	N/A	N/A	0	1	1
391	N/A	N/A	N/A	N/A	N/A	N/A	2	2	2
392	N/A	N/A	N/A	N/A	N/A	N/A	1	1	1
393	N/A	N/A	N/A	N/A	N/A	N/A	0	0	0

characterize nuanced differences in AE grading thresholds between MDs, RNs, and patients using this advanced statistical technique, thus providing us with information beyond that which can be obtained through the use of traditional statistical methods such as Cohen’s weighted  $\kappa$  or ICC.

**Methods**

**Patients**

The data sample for this secondary analysis included 393 English-language speaking cancer patients of mixed disease type (i.e., lung, prostate, and gynecologic) who were undergoing chemotherapy regimens as part of an Institutional Review Board approved protocol at Memorial Sloan Kettering Cancer Center (MSK) between March 2005 and August 2009 where informed consent was obtained from all included patients (Basch et al. 2005, 2007a, b, 2009,

2016). Patient records were eligible for inclusion in this analysis if they contained documented independent MD, RN, and patient symptom ratings for a single clinic visit, without any other restrictions (Atkinson et al. 2012).

**Measures**

*Common Terminology Criteria for Adverse Events version 4 (CTCAE)* (National Cancer Institute 2010)—CTCAE consists of a library of over 700 descriptive terms for clinician-based assessment of patient AEs related to cancer treatment. Each CTCAE term is assessed using a 5-point verbal descriptor grading scale, with each grade following a similar grading convention (i.e., 0=not present, 1=mild, 2=moderate, 3=severe and/or requiring medical intervention but not life-threatening, 4=life-threatening consequences, and 5=death).

*Symptom Tracking and Reporting (STAR)* (Basch et al. 2005, 2007a, b, 2009, 2015, 2016)—STAR is a web-based adaptation of CTCAE that was developed and validated to facilitate clinic waiting area and between-visit home-based patient reporting of treatment-related AEs. STAR items are assessed using a 5-point verbal descriptor rating scale similar to CTCAE (i.e., 0=none, 1=mild, 2=moderate, 3=severe, 4=disabling). STAR items assessing constipation, diarrhea, dyspnea, fatigue, nausea, and vomiting were included in the present analysis to correspond with analogous clinician-based CTCAE ratings of patients for these AEs.

**Procedure**

Routinely documented patient electronic medical records were examined using the Health Information System of MSK. Data were abstracted in cases where ratings of constipation, diarrhea, dyspnea, fatigue, nausea, and vomiting were made by an independent MD (via CTCAE), RN (via

**Table 2** Patient characteristics

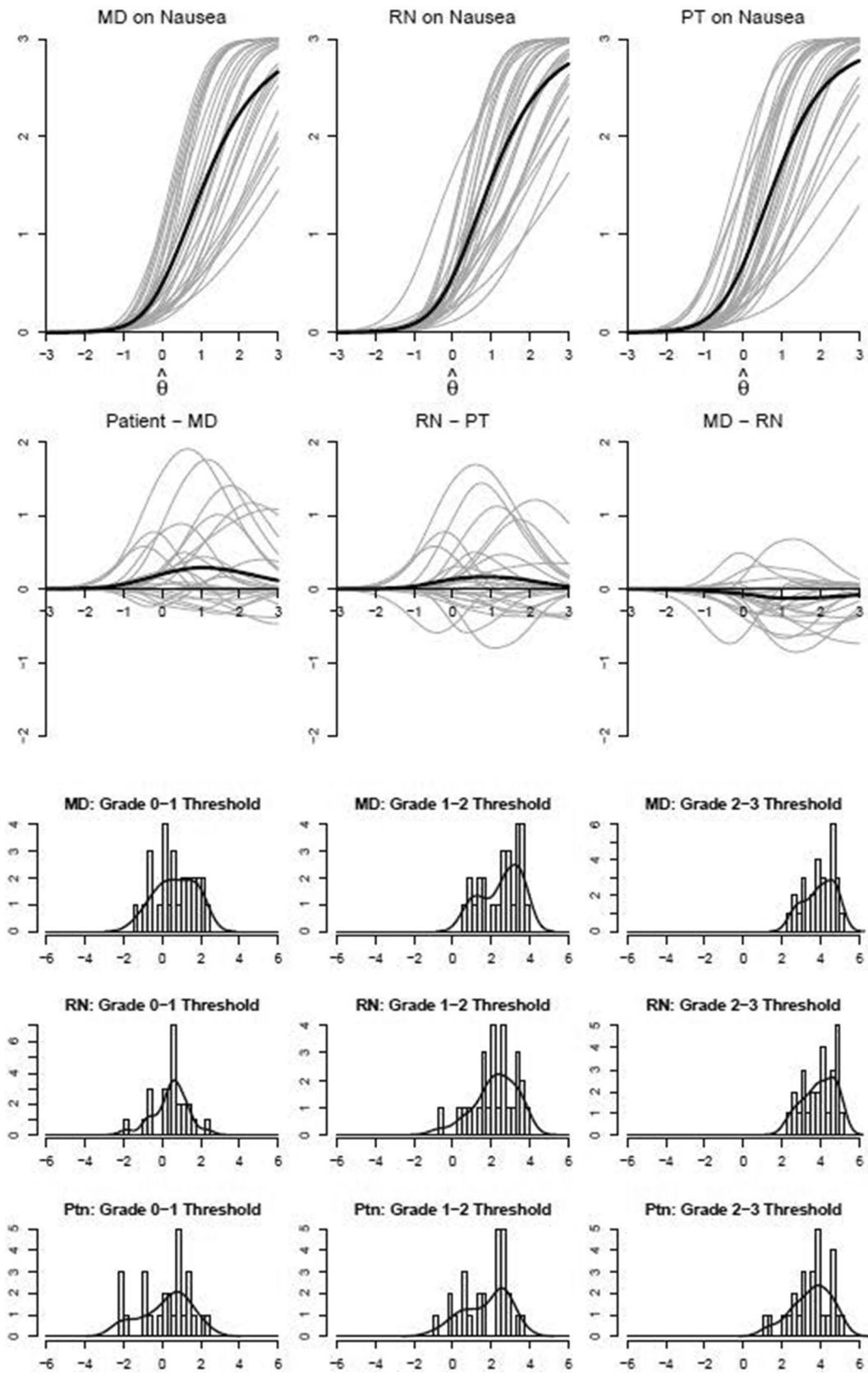
Characteristic	No. of patients (N=393)	%
Age range		
Mean, years	62	
Median, years	63	
Gender		
Female	224	57
Cancer type		
Lung	134	34
Prostate	113	29
Gynecologic	146	37
Race/ethnicity		
African-American	25	7
White Hispanic	11	3
White Non-Hispanic	337	86
Other	13	4

**Table 3** Means, standard deviations, and traditional concordance metrics for adverse event ratings by MDs, RNs, and patients

Adverse event	Mean (standard deviation)			Concordance metric					
				MD versus RN		MD versus patient		RN versus patient	
	MD	RN	Patient	Cohen’s $\kappa$	ICC	Cohen’s $\kappa$	ICC	Cohen’s $\kappa$	ICC
Constipation	0.27 (0.49)	0.48 (0.74)	0.34 (0.62)	0.25*	0.48*	−0.05	0.05	0.04	0.01
Diarrhea	0.09 (0.29)	0.09 (0.32)	0.16 (0.47)	0.40*	0.58*	0.18*	0.36*	0.25*	0.43*
Dyspnea	0.22 (0.46)	0.29 (0.55)	0.23 (0.54)	0.51*	0.69*	0.16*	0.28*	0.14*	0.29*
Fatigue	0.63 (0.64)	0.75 (0.69)	0.66 (0.95)	0.26*	0.50*	0.04	0.24*	0.08*	0.18*
Nausea	0.15 (0.38)	0.22 (0.49)	0.26 (0.63)	0.32*	0.52*	0.19*	0.39*	0.19*	0.36*
Vomiting	0.04 (0.22)	0.03 (0.22)	0.08 (0.37)	0.24*	0.46*	0.23*	0.47*	0.13*	0.23*

ICC Intraclass correlation coefficient, MD medical doctor, RN registered nurse

\*Indicates significant  $p < 0.05$



**Fig. 1** Graded response model estimates and histograms for MDs, RNs, patients and the difference between MD, RN, and clinician thresholds for rating nausea. For the *top two rows*, each trace line represents the expected a posteriori (EAP) AE ratings made by each individual over a range of latent AE values. For the histograms, the thick Gaussian kernel density trace line estimates represent the smooth version of responses

CTCAE), and patient (via STAR) triplets during the same clinic visit.

### Statistical analysis

A Bayesian GRM was fitted to calculate the latent grading thresholds between clinics and patients (Baldwin et al. 2009). In this analysis we focused on the model-based expected item responses between MDs, RNs, and patients. This model-based approach is advantageous in that it facilitates extraction of core information from data that contain multiple sources of variability. The resulting model-estimated responses then represent the most likely AE ratings from MDs, RNs, and patients with random error variabilities parsed out. The set of six individual AEs were treated as unidimensional, given that each AE was probed using a single item and independently rated by MDs, RNs, and patients.

It was necessary to code the data in a manner that was amenable to the GRM framework. Whereas the Baldwin example contained single ratings for each observation, the present dataset contains as many as three patient ratings and a clinic rating for each symptom. Table 1 represents a single symptom example of the data structure in our analysis, with the columns representing scale items fitted. For each column, GRM item discrimination and thresholds were calculated. The posterior mean values of the model-fitted item responses were calculated to represent model-based AE grades obtained from MDs, RNs, and patients independently. Since all MDs and RNs did not assess AEs in all patients, instances where a given MD did not make a rating were treated as missing (noted by “N/A”). For example, MD 264 may have rated patients 004 and 390–393 but no other patients in the dataset. The Bayesian GRM approach updates the parameter estimates based on available data only; therefore, missing data provides no information with respect to the posterior distributions of the parameters. This permitted the modeling of decision thresholds across the aggregated clinic clusters in an actual clinical encounter, without the need to compel a rectangular data structure.

With respect to the prior distributions,  $\alpha$ s follow a Gaussian distribution with a mean of 1.0 and a standard deviation of 2.5, truncated at a value  $>0.0$ . The threshold parameter  $\kappa$  values follow a Gaussian distribution of a mean of zero and a standard deviation of 2.5. The  $\theta$  values were constrained to have a standard normal distribution.

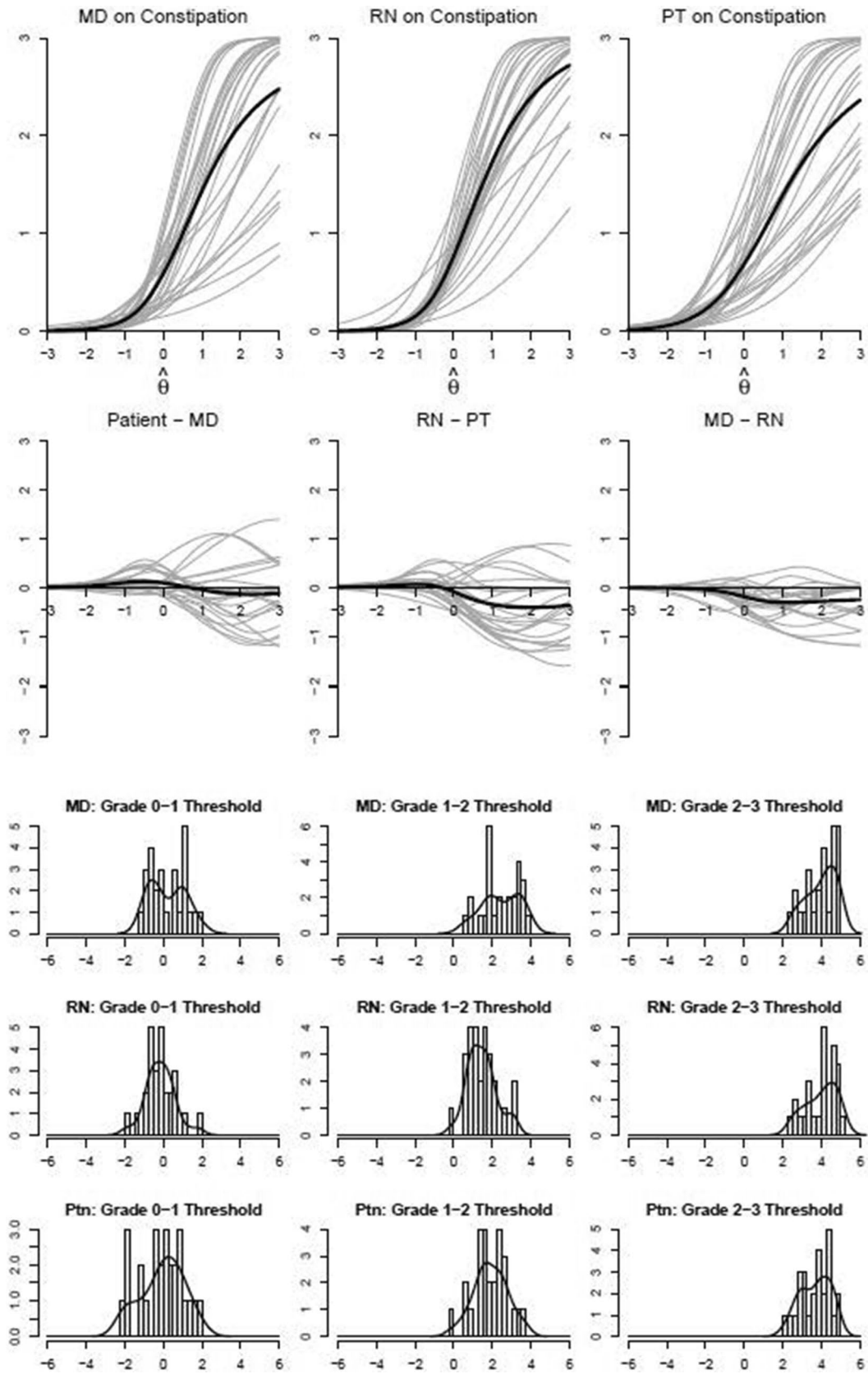
A total of 86,000 iterative burn-in simulations were completed, with the first 6000 iterations discarded, and thinning every ten simulations for the remaining 80,000 iterations. Other specific details on the Bayesian computation are explained elsewhere (Baldwin et al. 2009). Local independence among MD, RN, and patient ratings was assumed to simplify the illustrative examples. All analyses were completed using R version 3.2.3 (R Development Core Team 2016) and Just Another Gibbs Sampler (JAGS) version 4.1 (Plummer 2016). The RJAGS package was used in R as a conduit to send the data to JAGS for the actual simulations.

### Results

Table 2 includes characteristics of the included patients ( $N=393$ ). Patients (median age = 63, range = 26–91 years) were diagnosed with lung (34%), prostate (29%), or gynecologic (37%) malignancies; the majority of patients (85%) were high functioning (i.e., score  $\geq 80/100$ ), as captured by the clinician-reported Karnofsky Performance Status (Karnofsky and Burchenal 1949) measure. These patients were independently rated by 1 of 26 attending oncologists and corresponding 26 RNs, without having access to each other’s assessments, as part of their routine clinic visit. The average amount of time between MD and RN ratings was 68.04 min (Atkinson et al. 2012).

Table 3 displays the means, standard deviations, and traditional concordance metrics for patient AE ratings, separated by comparisons of MDs and RNs, MDs and patients, and RNs and patients. ICCs less than 0.40 indicate poor agreement, values between 0.40 and 0.75 are indicative of moderate agreement, with values of 0.75 or higher indicating excellent agreement (Rosner 2005). Cohen’s  $\kappa$  estimates follow a similar convention, with values from 0.00 to 0.40 representing poor concordance, 0.41–0.75 indicating fair to good agreement, and values over 0.75 indicative of excellent agreement (Shrout and Fleiss 1979). For the current sample, Cohen’s  $\kappa$  and ICC estimates were poor to moderate at best when comparing any of these rating sources for each of the AEs. Additionally, the Cohen’s  $\kappa$  for the comparison between MD and patient ratings of constipation was  $-0.05$ , which cannot be meaningfully interpreted.

Figure 1 represents GRM estimates for MDs, RNs, patients, and the resulting difference between MD and patient, RN and patient, and MD and RN ratings for nausea. Each trace line represents the expected a posteriori (EAP) AE ratings made by each individual over a range of latent toxicity values. The upper left subplot of Fig. 1 displays the EAP MD CTCAE ratings for nausea. The upper center subplot of Fig. 1 displays the EAP RN CTCAE ratings for nausea. The upper right subplot displays the EAP STAR ratings for all patients for nausea.



**Fig. 2** Graded response model estimates and histograms for MDs, RNs, patients and the difference between MD, RN, and clinician thresholds for rating constipation. For the *top two rows*, each trace line represents the expected a posteriori (EAP) AE ratings made by each individual over a range of latent AE values. For the histograms, the thick Gaussian kernel density trace line estimates represent the smooth version of responses

The second row of subplots of Fig. 1 displays the differences between model-estimated patient and MD, RN and patient, and MD and RN EAP AE ratings for nausea. To interpret these subplots, a difference of zero would represent perfect concordance between raters, with positive and negative values indicative of underestimation or overestimation of relative AE ratings, respectively. Here, MDs and RNs were observed to underestimate patient-reported nausea, with a slight RN overestimation of nausea when compared to MD ratings. Further, the Bayesian GRM shows higher variability in patients' thresholds in assessing nausea than those obtained from MDs and RN, as indicated by the more extreme trace lines for the patient versus MD and patient versus RN subplots.

The bottom subplots of Fig. 1 represent histograms of the GRM-estimated rating scale thresholds for nausea (i.e., Grade 0–1, 1–2, or 2–3), separated by MDs, RNs, and patients. For example, the subplot labeled “MD: Grade 0–1 Threshold” is the plot of the estimated thresholds for all 26 MDs, with the *x*-axis representing the latent implicit decision thresholds, in terms of standard deviations above or below the norm, and the *y*-axis representing the frequency count (i.e., the tallest bar represents four MDs with a latent implicit decision threshold near 0 standard deviations). The thick Gaussian kernel density trace line estimates represent the smooth versions of the histograms (Silverman 1986).

The MD and RN latent implicit decision threshold peak is represented by approximately five standard deviations above the norm, whereas the patient latent implicit decision threshold peak occurs at approximately four standard deviations above the norm. This implies that differences in AE grading between patients and MDs or RNs are more likely to occur at these higher levels of nausea toxicity.

Figure 2 represents constipation and follows the same general format as Fig. 1. Here, concordance between patients and MDs appears to be fairly high, with subtle MD underestimation at lower grading thresholds and overestimation at higher thresholds. RNs overestimate higher grading thresholds of patient- and MD-rated constipation. The frequency distribution subplots of Fig. 2 indicate that MD, RN, and patient latent implicit thresholds are relatively similar for the Grade 2–3 threshold, but that differences as large as 1 or 2 grades occur when RNs rate constipation at the Grade 1–2 threshold, as compared to MDs. Appendix II includes similar figures for the remaining four AEs (i.e., diarrhea, dyspnea, fatigue, vomiting).

## Discussion

Traditional methods of calculating concordance have been well established to characterize the relationship between two independent sources of information. However, when applying these methods to AE reporting, where there is likely to be a significant number of instances where MDs, RNs and patients agree due to a symptom not being present, the resulting coefficients may not be an accurate representation of the actual level of agreement. Additionally, a single coefficient does not provide us with a complete story of the relatedness of clinician- and patient-based AE ratings, particularly with respect to the direction and magnitude of the discrepancies. In the oncology clinical trial setting, where a difference as small as 1 CTCAE grade can determine whether a patient continues their participation in the trial, it is crucial to accurately identify and understand any sources of discrepancy in AE ratings. In this study, we used a Bayesian Graded Item Response Model to model concordance between MD-, RN-, and patient-based AE reporting, as well as characterize potentially nuanced differences in AE grading thresholds between these three rating sources.

We found that on average, the disagreements between MDs, RNs, and patients were generally less than one grade, but in some instances, these discrepancies can vary by up to two grades. Overall, MDs and RNs underestimate patient-reported diarrhea, dyspnea, nausea, vomiting, and fatigue. The Bayesian GRM analysis also demonstrated that RNs overestimate higher levels (i.e., Grade 1–2) of constipation when compared to patient or MD ratings, which is consistent with previous findings from a study of patients undergoing chemotherapy (Cirillo et al. 2009).

Additionally, the Bayesian GRM indicated the presence of higher variability in the latent patient AE rating thresholds versus those obtained from MDs or RNs. This finding is consistent with our previous work that indicates clinician-based toxicity reports underestimate the frequency and severity of AEs when compared to patient reports of these AEs (Basch et al. 2009). Patient variability in their AE-reporting thresholds is likely due to the highly subjective and contextual nature of AE self-reporting, where a given patient's rating of a severe AE could potentially be analogous to that same AE being rated as mild for another patient. Patients also may not be aware that important decisions related to their treatment and continued participation in a clinical trial may be impacted by their AE levels. As patient reporting of AEs becomes commonplace in oncology clinical trials, it may be important to provide patients with additional context with respect to the treatment-related implications of reporting a higher grade of a given AE.

The Bayesian GRM analysis begins to provide evidence to support the notion that patients report some symptoms

that MDs and/or RNs might not consider to be important until the AE has reached a more elevated level of severity. This is important to understand as the inclusion of patient-reported AEs nears standardization in US-based clinical trials in oncology. Clinical trial participation can impact clinician AE grading. In such a case, assigning a higher AE grade for a particular symptom may result in that patient being removed from the trial, despite any other evidence of therapeutic benefit. Utilization of the Bayesian GRM visualization of differences in AE grading thresholds could be a potentially useful tool that would allow MDs and RNs to communicate and acknowledge differences between clinician and patient AE reports while explaining the implications of assigning a higher AE grade.

Given that patient reports of AEs are becoming increasingly accepted for inclusion in clinical trials, an outstanding issue could be related to which source of AE reports should be considered to be the definitive “gold standard” indicator of AE levels. In the present study, MD ratings were compared with RN ratings and patient ratings of AEs. While patients were mentioned as the reference category when compared to MDs or RNs, this was only for the point of illustrating differences between sources of AE ratings. Unfortunately, in the absence of standardized AE grading decision criteria for MDs, RNs, and patients, there may be no definitive “gold standard” source of AE information. Nevertheless, these multiple AE rating sources should be used as complementary pieces of information that can provide clinicians with a more complete picture of the patient symptomatic experience.

This study is not without several limitations. Our sample was collected in a single, tertiary cancer center and was limited in diversity with respect to race, ethnicity, and disease type; only three cancer-type populations were included (i.e., lung, prostate, and gynecologic). Additionally, while the Bayesian GRM model is helpful in depicting underlying patterns of concordance between clinician- and patient-based AE ratings, this statistical method does not explain sources of discordance between raters. The STAR measure has been previously validated as a tool to capture patient-reported AEs (Basch et al. 2005); however, this instrument assesses a limited number of patient AEs. With the recent development of PRO-CTCAE (Basch et al. 2014; Dueck et al. 2015; Hay et al. 2014), it follows that this Bayesian GRM analysis be used in a multicenter prospective study of patients across multiple disease types to assess a wide range of treatment-related AEs, as assessed by CTCAE and PRO-CTCAE. Finally, in this context the GRM operates under the assumption that MD, RN, and patient ratings are locally independent given the model. As such, the results should be interpreted with caution, as non-independence may exist between these ratings. A formal investigation of this potential statistical codependence is beyond the scope

of this article. Future applications of this analysis should accommodate the multi-level data structure (i.e., patients nested within RNs, who are nested within MDs) and potentially assess the utility of employing alternative models to accommodate such a structure, such as the Rasch testlet model (Wang and Wilson 2005).

The Bayesian GRM can be a potentially useful descriptive tool for understanding and visualizing the nuanced differences between MD-, RN-, and patient AE-reporting thresholds. For instances where MDs and RNs may rate the same patient or set of patients, the Bayesian GRM can display subtle patterns of discrepancies between such ratings and show where any potential large, 1–2 grade differences may exist for a given AE. This information can help to assist MDs and RNs in the standardization of AE grading. Similarly, as patient reports of treatment-related AEs become commonplace in oncology clinical trials, their ratings can be included in a Bayesian GRM framework to be displayed relative to their respective clinician ratings for a given AE. Such information can serve to enhance communication between patient and provider and potentially help patients understand the importance of accurate AE reporting, toward ultimately improving decisions related to treatment and long-term patient health outcomes.

#### Compliance with ethical standards

**Funding** This project was supported by a National Institutes of Health Support Grant (NCI 2 P30 CA08748-48), which provides partial support for the Behavioral Research Methods Core Facility used in conducting this investigation. This study was also supported by a grant from the Society of Memorial Sloan Kettering.

**Conflict of interest** The authors declare that there is no conflict of interest.

**Ethical approval** All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

**Informed consent** Informed consent was obtained from all individual participants included in the study.

#### References

- Atkinson TM, Li Y, Coffey CW, Sit L, Shaw M, Lavene D, Basch E (2012) Reliability of adverse symptom event reporting by clinicians. *Qual Life Res* 21(7):1159–1164. doi:[10.1007/s11136-011-0031-4](https://doi.org/10.1007/s11136-011-0031-4)
- Atkinson TM, Ryan SJ, Bennett AV, Stover AM, Saracino RM, Rogak LJ, Jewell ST, Matsoukas K, Li Y, Basch E (2016) The association between clinician-based common terminology criteria for adverse events (CTCAE) and patient-reported outcomes



- (PRO): a systematic review. *Supp Care Cancer* 24:3669–3676. doi:[10.1007/s00520-016-3297-9](https://doi.org/10.1007/s00520-016-3297-9)
- Baldwin P, Bernstein J, Wainer H (2009) Hip psychometrics. *Stat Med* 28(17):2277–2292. doi:[10.1002/sim.3616](https://doi.org/10.1002/sim.3616)
- Basch E (2010) The missing voice of patients in drug-safety reporting. *N Engl J Med* 362(10):865–869. doi:[10.1056/NEJMp0911494](https://doi.org/10.1056/NEJMp0911494)
- Basch E (2012) Beyond the FDA PRO guidance: steps toward integrating meaningful patient-reported outcomes into regulatory trials and US drug labels. *Value Health* 15(3):401–403. doi:[10.1016/j.jval.2012.03.1385](https://doi.org/10.1016/j.jval.2012.03.1385)
- Basch E (2014) New frontiers in patient-reported outcomes: adverse event reporting, comparative effectiveness, and quality assessment. *Annu Rev Med* 65:307–317. doi:[10.1146/annurev-med-010713-141500](https://doi.org/10.1146/annurev-med-010713-141500)
- Basch E (2016) Missing patients' symptoms in cancer care delivery—the importance of patient-reported outcomes. *JAMA Oncol* 2(4):433–434. doi:[10.1001/jamaoncol.2015.4719](https://doi.org/10.1001/jamaoncol.2015.4719)
- Basch E, Artz D, Dulko D, Scher K, Sabbatini P, Hensley M et al (2005) Patient online self-reporting of toxicity symptoms during chemotherapy. *J Clin Oncol* 23(15):3552–3561. doi:[10.1200/JCO.2005.04.275](https://doi.org/10.1200/JCO.2005.04.275)
- Basch E, Artz D, Iasonos A, Speakman J, Shannon K, Lin K et al (2007a) Evaluation of an online platform for cancer patient self-reporting of chemotherapy toxicities. *J Am Med Inform Assoc* 14(3):264–268. doi:[10.1197/jamia.M2177](https://doi.org/10.1197/jamia.M2177)
- Basch E, Iasonos A, Barz A, Culkun A, Kris MG, Artz D, ... Schrag, D (2007b) Long-term toxicity monitoring via electronic patient-reported outcomes in patients receiving chemotherapy. *J Clin Oncol* 25(34):5374–5380. doi:[10.1200/JCO.2007.11.2243](https://doi.org/10.1200/JCO.2007.11.2243)
- Basch E, Jia X, Heller G, Barz A, Sit L, Fruscione M et al (2009) Adverse symptom event reporting by patients vs clinicians: relationships with clinical outcomes. *J Natl Cancer Inst* 101(23):1624–1632. doi:[10.1093/jnci/djp386](https://doi.org/10.1093/jnci/djp386)
- Basch E, Reeve BB, Mitchell SA, Clauser SB, Minasian LM, Dueck AC et al (2014) 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*, 106(9). doi:[10.1093/jnci/dju244](https://doi.org/10.1093/jnci/dju244)
- Basch E, Wood WA, Schrag D, Sima CS, Shaw M, Rogak LJ et al (2015) Feasibility and clinical impact of sharing patient-reported symptom toxicities and performance status with clinical investigators during a phase 2 cancer treatment trial. *Clin Trials*. doi:[10.1177/1740774515615540](https://doi.org/10.1177/1740774515615540)
- Basch E, Deal AM, Kris MG, Scher HI, Hudis CA, Sabbatini P et al (2016) Symptom monitoring with patient-reported outcomes during routine cancer treatment: a randomized controlled trial. *J Clin Oncol* 34(6):557–565. doi:[10.1200/JCO.2015.63.0830](https://doi.org/10.1200/JCO.2015.63.0830)
- Cirillo M, Venturini M, Ciccarelli L, Coati F, Bortolami O, Verlato G (2009) Clinician versus nurse symptom reporting using the National Cancer Institute-Common Terminology Criteria for Adverse Events during chemotherapy: results of a comparison based on patient's self-reported questionnaire. *Ann Oncol* 20(12):1929–1935. doi:[10.1093/annonc/mdp287](https://doi.org/10.1093/annonc/mdp287)
- Dueck AC, Mendoza TR, Mitchell SA, Reeve BB, Castro KM, Rogak LJ et al (2015) Validity and reliability of the U.S. National Cancer Institute's patient-reported outcomes version of the common terminology criteria for adverse events (PRO-CTCAE). *JAMA Oncol*. doi:[10.1001/jamaoncol.2015.2639](https://doi.org/10.1001/jamaoncol.2015.2639)
- Hay JL, Atkinson TM, Reeve BB, Mitchell SA, Mendoza TR, Willis G et al (2014) 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* 23(1):257–269. doi:[10.1007/s11136-013-0470-1](https://doi.org/10.1007/s11136-013-0470-1)
- Karnofsky DA, Burchenal JH (1949) The clinical evaluation of chemotherapeutic agents in cancer. In: Macleod CM (ed) *Evaluation of chemotherapeutic agents*. Columbia University Press, New York, pp 199–205
- National Cancer Institute, National Institutes of Health, U.S. Department of Health and Human Services. Common Terminology Criteria for Adverse Events (CTCAE) Version 4.0. Published May 28, 2009; Revised Version 4.03 June 14, 2010. [http://evs.nci.nih.gov/ftp1/CTCAE/CTCAE\\_4.03\\_2010-06-14\\_QuickReference\\_8.5x11.pdf](http://evs.nci.nih.gov/ftp1/CTCAE/CTCAE_4.03_2010-06-14_QuickReference_8.5x11.pdf). Accessed 9 Dec 2016
- Plummer M (2016) JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. <http://mcmc-jags.sourceforge.net>
- R Development Core Team (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. <http://www.R-project.org>. Accessed 9 Dec 2016
- Rosner B (2005) *Fundamentals of biostatistics*. Duxbury, Belmont
- Shrout PE, Fleiss JL (1979) Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull* 86:420–428. doi:[10.1037//0033-2909.86.2.420](https://doi.org/10.1037//0033-2909.86.2.420)
- Silverman BW (1986) *Density estimation*. Chapman and Hall, London
- Trotti A, Colevas AD, Setser A, Basch E (2007) Patient-reported outcomes and the evolution of adverse event reporting in oncology. *J Clin Oncol* 25:5121–5127
- Wang WC, Wilson M (2005) The Rasch testlet model. *Appl Psychol Meas* 29(2):126–149. doi:[10.1177/0146621604271053](https://doi.org/10.1177/0146621604271053)
- Samejima F (1997) *Graded response model Handbook of modern item response theory*. Springer, New York, pp 85–100