

Communicating patient-reported outcome scores using graphic formats: results from a mixed-methods evaluation

Michael D. Brundage¹ · Katherine C. Smith^{2,3} · Emily A. Little⁴ · Elissa T. Bantug² · Claire F. Snyder^{2,3,4} · The PRO Data Presentation Stakeholder Advisory Board

Accepted: 24 March 2015 / Published online: 27 May 2015
© Springer International Publishing Switzerland 2015

Abstract

Background Patient-reported outcomes (PROs) promote patient-centered care by using PRO research results (“group-level data”) to inform decision making and by monitoring individual patient’s PROs (“individual-level data”) to inform care. We investigated the interpretability of current PRO data presentation formats.

Method This cross-sectional mixed-methods study randomized purposively sampled cancer patients and clinicians to evaluate six group-data or four individual-data formats. A self-directed exercise assessed participants’ interpretation accuracy and ratings of ease-of-understanding and usefulness (0 = least to 10 = most) of each format. Semi-structured qualitative interviews explored helpful and confusing format attributes.

Results We reached thematic saturation with 50 patients (44 % < college graduate) and 20 clinicians. For group-level data, patients rated simple line graphs highest for ease-of-understanding and usefulness (median 8.0; 33 % selected for easiest to understand/most useful) and

clinicians rated simple line graphs highest for ease-of-understanding and usefulness (median 9.0, 8.5) but most often selected line graphs with confidence limits or norms (30 % for each format for easiest to understand/most useful). Qualitative results support that clinicians value confidence intervals, norms, and *p* values, but patients find them confusing. For individual-level data, both patients and clinicians rated line graphs highest for ease-of-understanding (median 8.0 patients, 8.5 clinicians) and usefulness (median 8.0, 9.0) and selected them as easiest to understand (50, 70 %) and most useful (62, 80 %). The qualitative interviews supported highlighting scores requiring clinical attention and providing reference values.

Conclusions This study has identified preferences and opportunities for improving on current formats for PRO presentation and will inform development of best practices for PRO presentation. Both patients and clinicians prefer line graphs across group-level data and individual-level data formats, but clinicians prefer greater detail (e.g., statistical details) for group-level data.

Presented in part at the 2014 ASCO Annual Meeting, 2014 AcademyHealth Annual Research Meeting, and 2014 International Society for Quality of Life Research Annual Meeting.

✉ Michael D. Brundage
michael.brundage@krcc.on.ca

¹ Queen’s Cancer Research Institute, Queen’s University, 10 Stuart St, Level 2, Kingston, ON K7L 3N6, Canada

² Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins, 1650 Orleans St, Baltimore, MD 21287, USA

³ Johns Hopkins Bloomberg School of Public Health, 624 N. Broadway, Baltimore, MD 21205, USA

⁴ Division of General Internal Medicine, Johns Hopkins School of Medicine, 624 N. Broadway, Baltimore, MD 21205, USA

Keywords Patient-reported outcomes · Clinical trials · Clinical practice · Communication

Introduction

With the intensifying emphasis on patient-centered care, patient-reported outcomes (PROs) are playing an increasingly important role in clinical practice. PROs are reports directly from patients about a health condition or its treatment and include an array of outcomes such as symptoms, functional status, and health-related quality of life (HRQOL) [1, 2]. PRO data can promote patient-centered care. PROs are commonly used in comparative

research studies [3–6], and oncologists have also endorsed the value of PROs in clinical trials [7, 8]. PRO findings from these studies can facilitate patients' and clinicians' understanding of how different treatments affect patient functioning and well-being and can inform treatment decision making [8, 9].

A second important application of PROs in clinical practice is to inform individual patient management [10, 11]. In this application, patients complete a PRO questionnaire, the results of which are shared with their clinician (and frequently the patient) to help identify and address concerns. Use of PROs for individual patient management has been shown to improve clinician–patient communication [12–15].

However, a number of barriers limit PRO use in practice by patients and clinicians, including variation in which PRO measures are used, how these are scored, how scores are scaled, and how data are reported. The PRO and Quality of Life Instrument Database describe over 800 PRO measures [16] that vary in how they are scored (e.g., higher scores are better on some measures but worse on others). PRO measures also vary in scaling; for example, 0–100 may represent the worst-to-best score, or scores may be normed (e.g., to a population average of 50). Thus, scores can have widely different meaning depending on the PRO being used. In patient management applications, patients and clinicians consistently report that the variation in PRO scoring and scaling makes score interpretation and application challenging [17–19]. For group-level data from research studies, oncologists strongly endorse the potential of PROs to enhance clinical trial interpretation, but fewer than half felt comfortable interpreting PRO results themselves [8]. Further, some methods for presenting PRO data are more easily understood and more accurately interpreted than others [20, 21].

We undertook this study to (1) evaluate patient and clinician comprehension of PRO data using existing presentation approaches and (2) obtain qualitative feedback on attributes of different presentation formats that are helpful and challenging.

Methods

Study design

This cross-sectional, mixed-methods study used a self-completed exercise followed by semi-structured qualitative interviews to identify facilitators of and barriers to comprehension and application of current PRO data presentation formats. Enrolled patients and clinicians were randomized 1:1 to evaluate either “group-level” PRO data from comparative research studies or “individual-level”

PRO data for patient monitoring. The Johns Hopkins School of Medicine Institutional Review Board approved the study; all participants provided written consent.

Population and setting

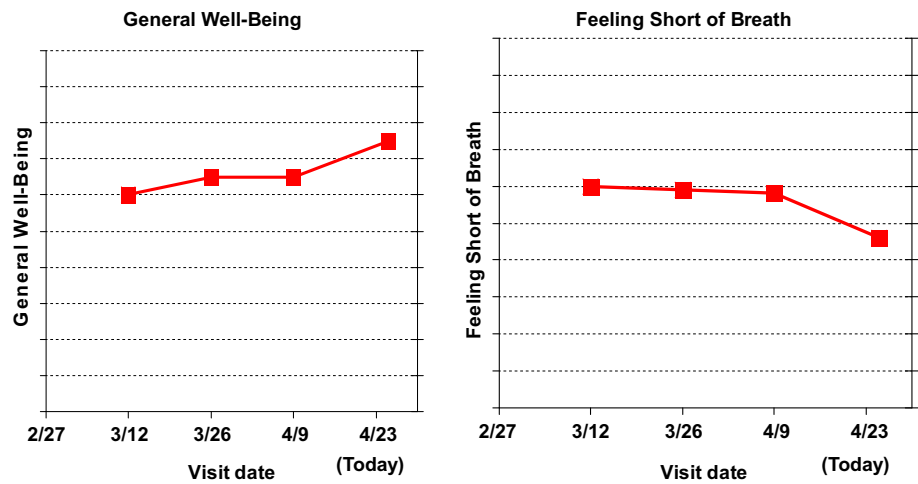
Participants were recruited from eight academic and community institutions representing diverse populations in the Johns Hopkins Clinical Research Network (JHCRN). Eligible patients were ≥ 21 years, diagnosed with any cancer (excluding non-melanoma skin cancer), ≥ 6 months post-diagnosis, not currently receiving acute treatment, and able to communicate in English. We purposively sampled by education attainment, cancer type, and practice setting. Eligible clinicians were in active oncology practice, including medical, radiation, and surgical oncologists, gynecologic oncologists/urologists, oncology nurse practitioners, and oncology fellows. We purposively sampled by practice setting and by clinician specialty.

Study procedures

All interviews had a common structure. Participants began by completing a self-directed paper-based exercise. First, in order to orient participants to PRO questionnaires, each completed six domains from the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire-Core-30 (QLQ-C30) [22]. Second, we assessed participants' intuitive understanding of the meaning of increasing/decreasing PRO scores by asking them to interpret two line graphs: one labeled ‘general well-being’ with a line trending up and one labeled ‘shortness of breath’ with a line trending down (Fig. 1). The nature of the graph was explained (e.g., “each square on the line represents a recent clinic visit”), but no instructions regarding the meaning of direction were provided. For each graph, participants reported whether they thought the graph represented a patient getting “better,” “worse,” or “not sure.”

Third, participants were presented with different approaches for presenting PRO scores using example data from the six QLQ-C30 domains. QLQ-C30 data were generally presented consistent with its scoring conventions (0–100 scale with higher scores representing better function and greater symptom burden), but for some formats, data were presented so that directionality (i.e., whether higher was better/worse) was the same across function and symptom domains. The data presentation formats were selected to illustrate a range of approaches found in the literature. For group-level data, six formats presenting hypothetical clinical trial results were evaluated (Fig. 2): simple line graphs of mean scores over time, line graphs with norms, line graphs with confidence intervals, bar charts of average changes, bar charts based on a responder

Fig. 1 Line graphs used to evaluate participants' intuitive understanding of trending PRO scores for an individual-level, single domain function and symptom. In keeping with the research design, scores on the y-axis are not provided



definition (improved, stable, worsened), and cumulative distribution functions [1]. For individual-level data, four formats presenting a hypothetical patient's scores were evaluated (Fig. 3): line graphs of mean scores [17], tabulated scores [23], bubble plots of scores at a point in time [13], and a heat map [24]. Each format was presented on a separate page with its own explanation and legend, including instructions regarding directionality [e.g., “For function scales, high scores represent high levels of functioning. For symptom scales, high scores represent high symptom levels (greater symptom burden).”]. Participants were instructed to consider only the information on a given page to discourage their applying descriptions of one format to others. To control for potential order effects, formats were randomly presented in four different orders.

For each format, participants responded to two questions that assessed accuracy of interpretation. Finally, for each format, participants rated “How easy is it for you to understand these graphs?” (0 = Very difficult to 10 = Very easy) and “How useful do you find these graphs?” (0 = Not at all to 10 = Very).

Following the self-directed portion, the interviewer conducted a semi-structured debriefing interview. The interviewer assessed participant's reasoning for the intuitive interpretation questions and then reviewed the participants' responses to each format, including the ease-of-understanding and usefulness ratings, and the decision process for one accuracy question. For each format, participants were asked what they liked, did not like/found confusing, and what they would add/remove/change. In the individual-level data interviews, specific probes were used to evaluate particular format attributes (e.g., yellow shading in the tabulated scores). Finally, respondents were asked whether they preferred formats that depict single or multiple time-points and were asked to select the one format that was easiest to understand and most helpful for patients and

clinicians to use in practice. Interviews were audio-recorded and transcribed.

Analyses and sample size

The quantitative data were analyzed descriptively using proportions and medians/ranges. To analyze the qualitative data, we used a “concurrent triangulation” design [25] utilizing an applied “framework approach” [26] that focused on attributes relevant to comprehension, preference, and utility of PRO data. The quantitative data results were considered in the context of the key points identified from the qualitative data to develop overall study findings.

To organize the qualitative data, the research team developed a coding scheme based on the study objectives, interview structure, and content of the initial interviews. Codes related to positive or negative comments made by participants on each format and to emergent themes. After several training rounds performed by the whole team, one team member (E.L.) coded all transcripts using ATLAS.ti [27] and each transcript's coding was reviewed by a second investigator (E.B., M.B., C.S., K.S.). Team members independently identified themes from reports summarizing the coded text (e.g., positive and negative comments on each format), which were then discussed by the group to summarize key points for each format. Selected quotations that illustrate the key findings are included in the results (“[P]” and “[C]” attributions are used to identify patient and clinician respondents, respectively).

Sample size was determined based on thematic saturation for the qualitative objectives. We estimated that 50 patient and 20 clinician interviews would be required (half each addressing group- and individual-level formats). We confirmed thematic saturation by identifying themes from the first 56 interviews and then determining that the final 14 interviews did not identify substantive additions.

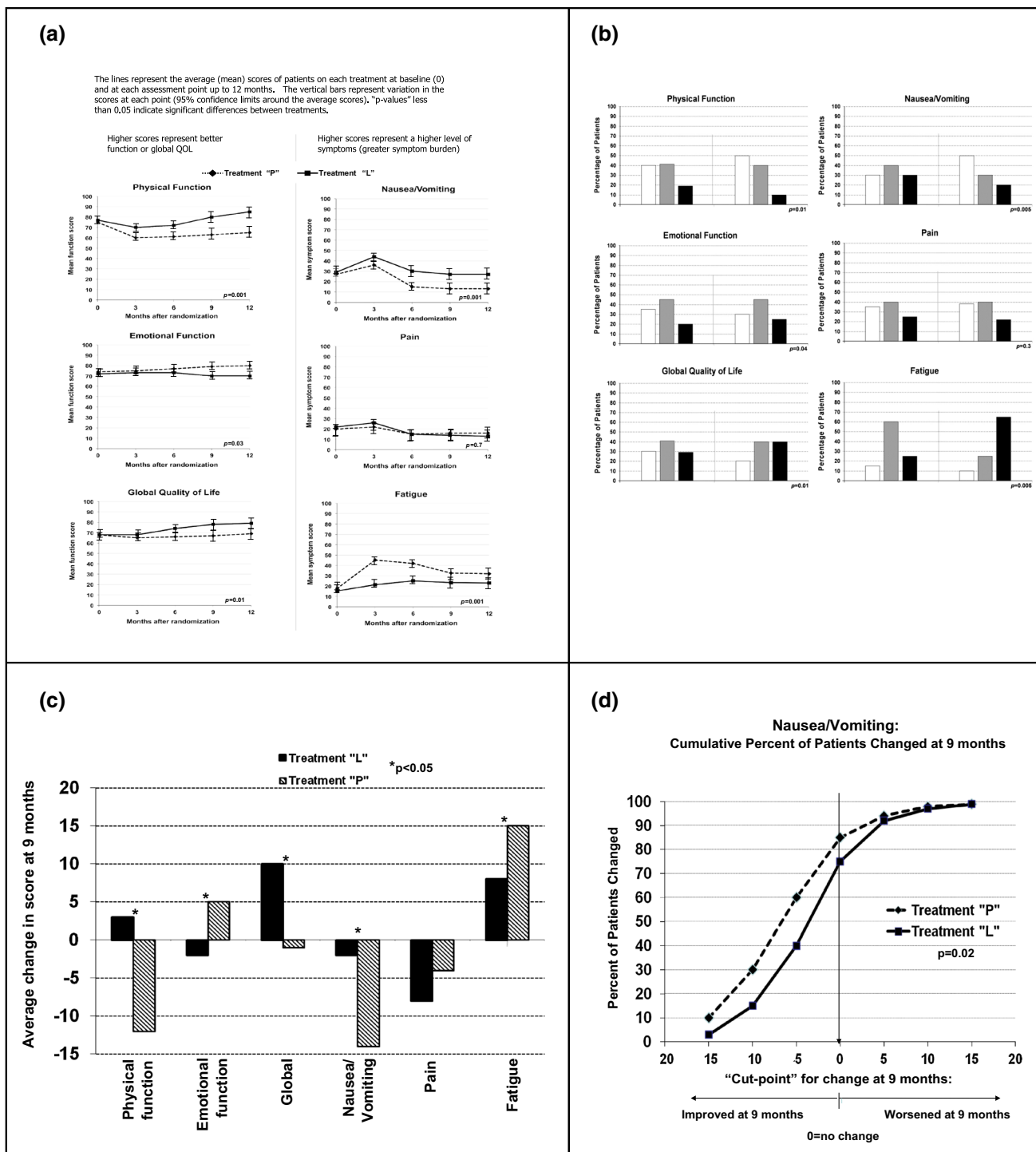


Fig. 2 Group-level data formats included **a** line graphs of mean scores with confidence intervals, **b** proportions responding (improved/same/worsened), **c** bar charts of average changes, **d** cumulative distribution functions. Additional group-level data formats included

simple line graphs of mean scores and line graphs of normed mean scores (not shown). Each format was presented on a separate page with its own explanation and legend (not all of which are shown in the Figure)

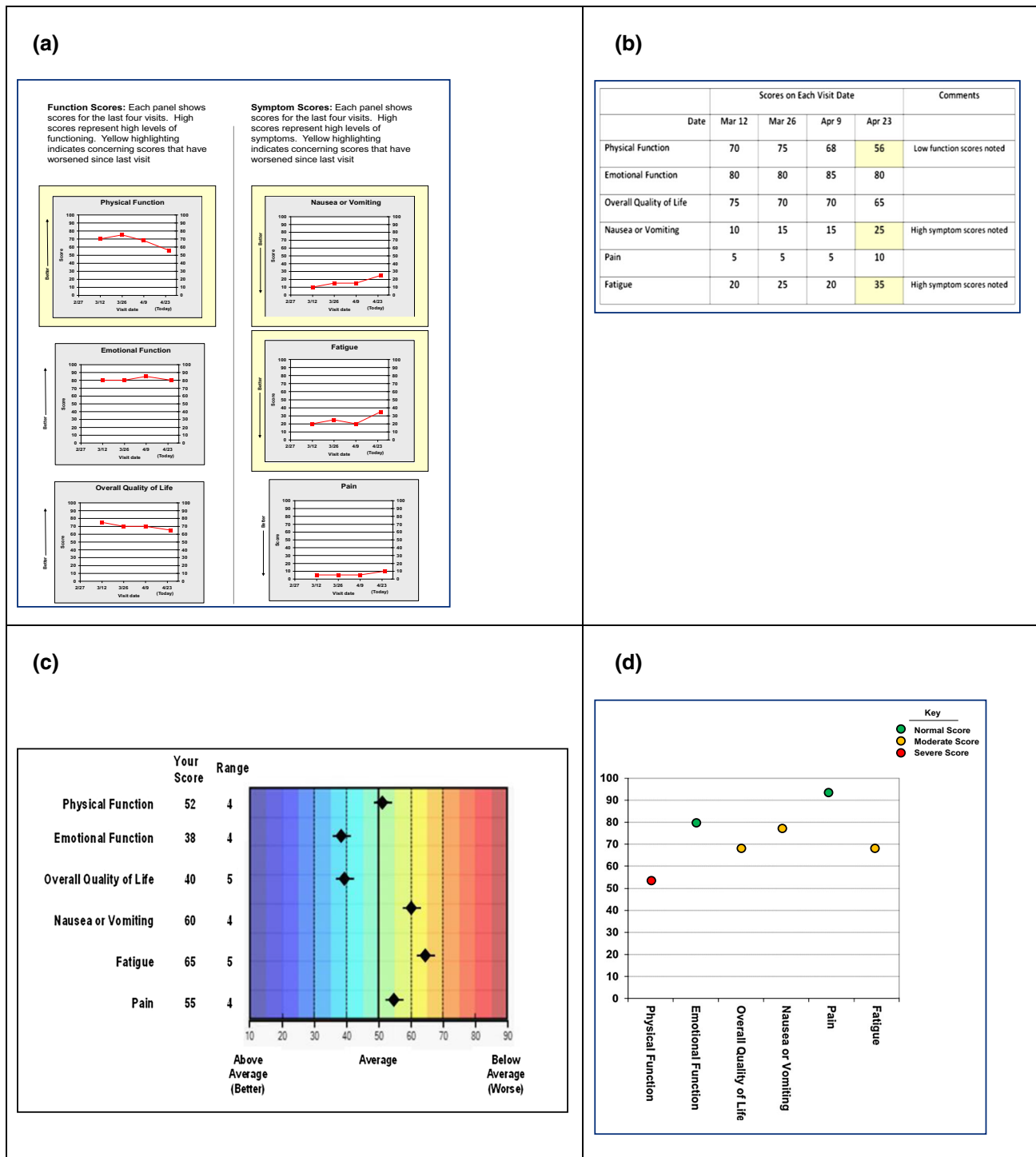


Fig. 3 Individual-level data formats included **a** line graphs of scores over time, **b** tabulated scores, **c** heat map of normed scores, **d** bubble plot of scores. Each format was presented on a separate page with its own explanation and legend (not all of which are shown in the Figure)

Results

Study sample

To recruit 50 patients, we assessed the eligibility of 147, of whom 25 were not eligible, 10 declined participation, and 62 were registered but not randomized (e.g., not meeting purposive sample needs). The participants' characteristics are listed in Table 1. The median age was 65.5 years, 54 % were female, and 78 % were white. Per our purposive sampling strategy, participants had a range of education levels, and the most common cancers represented were

breast (32 %) and prostate (28 %). The median time from first diagnosis was 5 years (range 0–32), and median time from most recent diagnosis was 3 years (range 0–7). Patients were recruited from 6 JHCRN practice locations.

To recruit 20 clinicians, we assessed the eligibility of 37, of whom 3 were not eligible, 1 declined participation, and 13 were registered but not randomized. The clinician participants' characteristics are also shown in Table 1. The sample included clinicians from a variety of specialties, with a median age of 42, 40 % of whom were female. Clinicians were recruited from 6 JHCRN practice locations.

Table 1 Participant demographics

Group	Characteristic	
Patients		(<i>N</i> = 50)
	Age: median (range)	65.5 (25–89)
	Sex: [<i>N</i> (%)] female	27 (54)
	Years from most recent diagnosis: median (range)	3 (0–7)
	Ongoing treatment [<i>N</i> (%)]	25 (50)
	Race [<i>N</i> (%)]	
	White	39 (78)
	Black	10 (20)
	Other	1 (2)
	Education [<i>N</i> (%)]	
	High school graduate	10 (20)
	Some college	12 (24)
	College graduate	8 (16)
	Postgraduate	20 (40)
	Cancer type [<i>N</i> (%)]	
	Breast	16 (32)
	Prostate	14 (28)
	Lung	4 (8)
	Lymphoma	4 (8)
	Other	12 (24)
Clinicians		(<i>N</i> = 20)
	Age: median (range)	42 (27–69)
	Sex: [<i>N</i> (%)] female	8 (40)
	Race [<i>N</i> (%)]	
	White	17 (85)
	Black	0 (0)
	Other	3 (15)
	Specialty [<i>N</i> (%)]	
	Medical oncology	5 (25)
	Radiation oncology	3 (15)
	Surgical oncology	5 (25)
	Gyneoncology/urology	2 (10)
	Oncology NP/PA	3 (15)
	Oncology fellow	2 (10)
Years in practice:		
Median (range)	16.5 (4–42)	

Intuitive interpretation findings

The vast majority of patients and clinicians interpreted the upward-trending “general well-being” line as getting better (patients: 96 % better, 2 % worse, 2 % not sure; clinicians: 80 % better, 5 % worse, 15 % not sure). In the interviews, participants reported that uptrending lines intuitively signify improvement. Participants reporting “not sure” were often unwilling to guess direction in the absence of labeled axes or scale descriptors. One reason for selecting “getting worse” is that the QLQ-C30 response options (1 = no problem at all and 4 = very much problem) inferred that higher numbers indicated worse quality of life.

In contrast, participants’ interpretation of the downward-trending “shortness of breath” line showed greater variation (patients: 60 % worse, 34 % better, 6 % not sure; clinicians: 10 % worse, 70 % better, 20 % not sure). Many who chose “worse” commented on the intuitive understanding of increasing lines being better. Some interpreted the downward trend as reflecting less symptom burden (“This was going down, so that meant the person was not feeling as short of breath” [P007]), whereas several participants were not sure, some indicating their unwillingness to guess direction in the absence of labeled axes or scale descriptors.

Group-level data formats

Table 2 summarizes respondents’ ratings, overall preferences, and feedback on the group-level data formats. Both patients and clinicians rated simple line graphs highest for ease-of-understanding and usefulness. Across formats, clinicians’ accuracy of interpretation ranged from 56 % (a Cumulative Distribution Function question) to 100 % (a Normed Line Graph question), whereas patients’ accuracy ranged from 36 % (a Cumulative Distribution Function question) to 100 % (a Line Graphs with Confidence Intervals question) (data not shown).

Line graphs were generally described as “straightforward” and “clear.” Patients tended to find normed scores, *p* values, and confidence intervals confusing. Clinicians appreciated that normed scores provided a basis for comparison (beyond the two study arms) and valued *p* values and confidence intervals as particularly important for publication. Across the line graph formats, the inconsistency in scoring direction was often reported as unclear: “for function it’s high levels of functioning, for symptoms it’s greater burden, which if you get confused between the two that throws you off there a little bit” [P002].

There was also some support for bar chart formats. Respondents noted that it was easy to compare treatments using bar charts of average changes. Many clinicians

thought that the bar charts of proportions responding to treatment were clinically meaningful (“you’d rather have things that are stable, worse or improved and that’s usually how we judge things” [C014]). Both patients and clinicians found the cumulative distribution function confusing and difficult to interpret (“it stinks and I kept liking it worse!” [C012]), though the ability to compare treatments on proportions responding across multiple cut-points was appreciated among the few participants who grasped the concept.

In terms of overall preferences, 90 % of clinicians and 87 % of patients preferred formats displaying multiple time-points. Patients most often selected simple line graphs for overall ease-of-understanding and usefulness, whereas clinicians selected either line graphs of normed scores or with confidence intervals (Table 2).

Individual-level data formats

Table 3 summarizes respondents’ ratings, preferences, and feedback on the individual-level data formats. In the self-directed exercise, ratings for ease-of-understanding and usefulness were highest for line graphs for both patients and clinicians. Clinicians’ accuracy of interpretation was high across formats (90–100 % on both questions on all formats), and patients’ ranged from 64 % on one Line Graphs question to 96 % on the other Line Graphs question and on one Heat Map question (data not shown).

In the interviews, participants noted that line graphs “...gave a lot of information in a visually easy way to look at ...you could just glance at it” [C022]. Some clinicians drew analogies to other clinical information: “Looks like a blood pressure graph or sodium graph” [C066]. The importance of highlighting concerning scores was noted, but the approach used in the sample format (highlighting the graph in yellow) was not always effective. Several respondents expressed a desire for reference values/scale: “It would be nice to know what the normal range is,” [P001] whereas others said, “I don’t really care how I compare to others” [P054]. There was also a question of who the normative population should be (e.g., US average, other cancer patients).

As with the group-level formats, the inconsistency between whether higher scores were better or worse (indicated with an arrow on the *y*-axis pointing which direction was better) posed a barrier to interpretation, contributing to several participants’ incorrect responses to accuracy questions. As noted above, the two accuracy questions for line graphs represented both the highest and lowest accuracy rates. A “Physical Function” graph trending down with the “Better” arrow indicating that higher scores are better was correctly interpreted as worsening by 96 % of patients and 100 % of clinicians. However, a “Fatigue” graph trending up with the “Better” arrow indicating that lower scores are

Table 2 Summary of results for group-level data formats

Format	Positive comments	Negative comments	Suggestions for change	Patient ratings ^a				Clinician ratings ^a			
				Median ease-of-understanding (range)	Median usefulness (range)	% Selecting as easiest to understand format	% Selecting as most useful format	Median ease-of-understanding (range)	Median usefulness (range)	% Selecting as easiest to understand format	% Selecting as most useful format
Line graphs of mean scores	Generally perceived as straightforward by both clinicians and patients Gives a quick assessment of treatment impact Common representation of data in journals	Better/worse reversal was confusing to some Some clinicians preferred confidence limits Terms confusing to some patients (e.g., “mean”, “baseline”) Some patients did not understand what the <i>p</i> values indicated Some patients had trouble discerning solid versus dotted lines	Some clinicians felt that direction should be consistent or that clearer separation of function from symptom scores was needed Some patients wanted <i>p</i> values dropped or replaced by an asterisk Better differentiation between dashed and solid lines was recommended	8.0 (3–10)	8.0 (3–10)	33 %	33 %	9.0 (3–10)	8.5 (1–10)	0 %	0 %
Line graphs of normed mean scores	Many found format easy to read and interpret Many clinicians liked having a reference to compare scores against Simplicity appreciated by many	Similar to line graphs (direction changing, line format, font size) Some patients had difficulty or lacked confidence interpreting the format Some clinicians felt this format was less helpful than actual scores Need to clarify where the norms come from	Annotate norm score better (e.g., put number next to population average, use color) Consider highlighting differences in direction of “better/worse” Consider increasing clarity of line annotations and font size Clarify where the norms come from	7.0 (1–10)	7.0 (2–10)	8 %	12 %	7.0 (6–10)	8.0 (3–10)	30 %	30 %

Table 2 continued

Format	Positive comments	Negative comments	Suggestions for change	Patient ratings ^a			Clinician ratings ^a		
				Median ease-of-understanding (range)	Median usefulness (range)	% Selecting as easiest to understand format	Median ease-of-understanding (range)	Median usefulness (range)	% Selecting as easiest to understand format
Line graphs with confidence intervals (CIs)	Positive comments similar to line graphs Many clinicians preferred CIs be included Some patients found CIs were helpful	Many patients felt CIs confused interpretability and reduced clarity Some patients did not understand CIs, or could not interpret them Some clinicians did not like averages (wanted individual change information)	Similar recommendations as for line graphs Some felt better explanations of CIs would help Some wanted <i>p</i> values dropped or replaced by an asterisk	7.0 (0–10)	7.0 (2–10)	8 %	7.5 (6–10)	8.0 (2–10)	30 %
Bar chart of average changes	Many patients and clinicians found the format intuitive Some found it easier to compare treatments Many clinicians used to seeing this format	Some confusion with direction of bars “up and down” versus improved or not Displays only single point in time Some patients did not understand statistical significance	Some clinicians wanted <i>p</i> values and/or error bars indicated Some clinicians and patients wanted color coding Some felt improved should be above 0	7.0 (1–10)	6.0 (2–10)	17 %	8.0 (4–10)	8.0 (1–10)	20 %

Table 2 continued

Format	Positive comments	Negative comments	Suggestions for change	Patient ratings ^a			Clinician ratings ^a		
				Median ease-of-understanding (range)	Median usefulness (range)	% Selecting as easiest to understand format	Median ease-of-understanding (range)	Median usefulness (range)	% Selecting as easiest to understand format
Proportions responding	Some reported the concept of proportions easier to understand Some clinicians felt the response categories to be intuitive and clinically meaningful Bars were easily distinguished from each other	Some patients reported not understanding the construct of proportions Some patients didn't like black shading or diagonal lines One clinician had to "keep checking legend"	Some patients wished to remove <i>p</i> values Some suggested color coding; one suggested pie charts Some clinicians suggested: legend with each graph; clustering response arms; adding percentages to bars; making <i>p</i> values more prominent or replacing with an asterisk	7.0 (0–10)	7.0 (2–10)	29 %	8.0 (6–10)	8.0 (2–10)	20 %
Cumulative distribution of proportions	One clinician preferred multiple cut-points One patient considered it a good idea, once understood	Many found the format to be unfamiliar, non-intuitive, difficult to understand Some found the middle cut-point confusing	Many comments reflected a lack of understanding of the graph Requires better explanation, clarifying meaning of <i>x</i> - and <i>y</i> -axes Consider reversing the direction of improved/worsening	5.0 (0–10)	5.0 (0–8)	4 %	3.0 (0–9)	4.0 (0–9)	0 %

^a Responses to questions: "How easy is it for you to understand these graphs" (from self-directed portion); "How useful do you find these graphs" (from self-directed portion); "Overall, which format was easiest for you to understand?" (from interview); "Which format do you think would be most helpful to you for working together with your doctor or nurse?/Which format would be most useful for you to understand your patient's situation?" (from interview)

Table 3 Summary of results for individual-level data formats

Format	Positive comments	Negative comments	Suggestions for change	Patient ratings ^a				Clinician ratings ^a			
				Median ease-of-understanding (range)	Median usefulness (range)	% Selecting as easiest to understand format	% Selecting as most useful format	Median ease-of-understanding (range)	Median usefulness (range)	% Selecting as easiest to understand format	% Selecting as most useful format
Line graphs of scores over time	Many thought format was very clear “Better” directional arrows helped many Many accustomed to seeing data this way Many found that yellow highlighting was helpful	Yellow highlighting: Not noticed by everyone; Meaning was unclear to some; Highlighting of frame unclear; Some found confusing/redundant Some thought all worsening scores should be highlighted Some confused by “better” arrows switching direction Some felt “too much information”	Consider making higher better/worse all same direction, or, make “better” arrow more prominent Consider alternative ways of labeling y-axis or present one graph for function, one for symptoms Consider highlighting a specific point rather than frame Improve description without use of the word “concerning” Use color sparingly Consider highlighting improving domains	8.0 (2–10)	8.0 (2–10)	50 %	62 %	8.5 (2–10)	9.0 (2–10)	70 %	80 %
Tabulated scores	Many found the format concise, clear, and easy to understand Some found highlighting helpful (particularly highlighting the specific score) Comments helped focus attention to specific scores Considered useful when small differences in numbers are clinically important	Some felt lots of numbers; too much data Some felt scores are meaningless—lacking a scale Some wanted to know the normal range Similar to line graphs; issues with inconsistency in direction Some felt it took too long to understand what was shown owing to non-intuitive nature Some felt comments add clutter	Some suggested keeping direction consistent or provide greater separation between function and symptoms Clarify that all domains are on the same scale; some suggested use of 1–10 for all Consider providing range for what is normal Simplify comments (e.g., “worse” instead of “high symptom scores noted”)	7.0 (1–10)	7.0 (0–10)	12 %	8 %	5.0 (1–10)	5.0 (1–10)	0 %	0 %

Table 3 continued

Format	Positive comments	Negative comments	Suggestions for change	Patient ratings ^a			Clinician ratings ^a		
				Median ease-of-understanding (range)	Median usefulness (range)	% Selecting as easiest to understand format	Median ease-of-understanding (range)	Median usefulness (range)	% Selecting as easiest to understand format
Bubble plot of scores	Many found easy to use Many liked the colors and found them appropriate and intuitive Many found clear, simple to look at	Some missed trends over time Hard to read sideways labels Some constantly checking key, which was far away Some concerned with colors for color blind; not everyone understood them Hash marks too busy Lack of reference problematic for some (e.g., what does “normal” and “moderate” mean?) Unfamiliar/less intuitive to some Confused with pain being 92 but “normal”	Consider changing orientation of labels Consider different scaling or separate symptoms from function Provide a range and/or reference for score Make date labels clearer Simplify (e.g., consider no y-axis, no hash marks) Better connect label with corresponding “bubble” Consider finding ways of adding trends over time Consider use of symbols with or instead of color	7.0 (2–10)	7.0 (0–10)	25 %	6.0 (0–10)	7.0 (0–10)	10 %
						17 %			10 %

Table 3 continued

Format	Positive comments	Negative comments	Suggestions for change	Patient ratings ^a			Clinician ratings ^a				
				Median ease-of-understanding (range)	Median usefulness (range)	% Selecting as easiest to understand format	% Selecting as most useful format	Median ease-of-understanding (range)	Median usefulness (range)	% Selecting as easiest to understand format	
Heat map of normed scores	Some found color to be helpful and made numbers less important Some found they understood quickly and could “put in perspective” Use of red intuitive for alarm Easy to see all the data in one place Some found reference score helpful	Some found colors too busy; issue for color blind; not clear whether blue or red is better or worse; black text does not show up Some missed trends over time Unsure who the “average” score represents Some felt better/worse right/left should be switched Some found range confusing, not helpful Some found it took a long time to orient to the graph Some found the description unclear Labels not well aligned with points	Consider simplified color scheme Consider finding ways of adding trends over time Consider switching right/left better/worse Consider removing range or do a better job of explaining it Suggest removing dashed lines Consider name other than “heat” map and improve descriptions Consider removing numbers Improve alignment between point and label	6.0 (0–10)	5.0 (0–9)	12 %	12 %	7.0 (1–10)	7.0 (1–10)	20 %	10 %

^a Responses to questions: “How easy is it for you to understand these graphs” (from self-directed portion); “How useful do you find these graphs” (from self-directed portion); “Overall, which format was easiest for you to understand?” (from interview); “Which format do you think would be most helpful to you for working together with your doctor or nurse?” (from interview); “Which format would be most useful for you to understand your patient’s situation?” (from interview)

better was correctly interpreted by only 64 % of patients and 90 % of clinicians. When asked about the responses to the “Fatigue” question in the debriefing interview, some participants did not notice that lower scores were better and reported intuitively interpreting higher scores as better (“I imagine if it’s going up, it’s getting better” [P029]). Others noticed only on further examination that lower scores were better for Fatigue (“...I didn’t notice [indications of direction]” [P051] and “For some of them going up is better and so...my first instinct was incorrect”) [C009].

Among other individual-level formats, tabulated scores were considered “boring but straightforward,” [P054] with one respondent noting the value if small differences in numbers are clinically important. The highlighting of the particular score considered potentially concerning was appreciated (as compared to highlighting the entire graph for line graphs). However, others thought that the format presented too much data and that the scores were difficult to interpret without a reference scale. One respondent thought that the comments were an “apology for a confusing table” [C016]. The inconsistency in scoring directionality was noted here, too. On the heat map, color was considered helpful, but several respondents recommended simplifying the color scheme. There was also some debate as to whether red should indicate good or bad, and some thought that better scores should be on the right. Most patients did not understand the concept of “range.” There was also an issue with having 50 as the norm: “Statistically it’s correct, but people don’t want to be 50. If you...take a test, you don’t want to get a C” [C026]. While bubble plots were considered “easy on the eye,” [C005], the trends over time were missed and the scaling was not clear to everyone. Notably, on this format, scores were depicted consistently (higher always better), but one respondent noted that it was confusing that pain was 92 but “normal.”

In terms of overall preferences, 90 % of clinicians and 92 % of patients preferred formats displaying multiple time-points versus single time-points. Both patients and clinicians most often selected line graphs for overall ease-of-understanding and usefulness (Table 3).

Discussion

For PRO data to contribute meaningfully to patient-centered care, patients and clinicians must be able to understand and interpret these data. Previous research has demonstrated that variation in PRO scoring, scaling, and presentation poses obstacles to interpretation and application. This study aimed to elicit patient and clinician input on formats for PRO data presentation and their respective attributes that would help to promote PRO data understanding and clinical use.

For group-level data, our findings suggest the need to use different approaches when presenting data to patients (e.g., in educational materials or decision aids) versus clinicians (e.g., in peer-reviewed publications). Patients tended to prefer simple line graphs, whereas clinicians valued greater detail (e.g., normed scores, or inclusion of confidence intervals or *p* values). Because the appropriate format also depends on how the study endpoint is conceptualized (e.g., mean scores over time vs. proportions responding to treatment), it is also necessary to consider best practices for presenting bar charts of group-level data. For individual-level data, line graphs were the consensus favorite for patients and clinicians.

The findings of this study should be considered in the context of its design. All candidate formats displayed data from the QLQ-C30, an instrument with distinct subscales that may be easier to graph. For the most part, the data were reported based on the QLQ-C30’s scoring conventions (0–100, with higher scores representing better function and greater symptom burden). In some cases, we reported the data normed or with consistent scoring directionality—such as is used by other PRO instruments. Additional research should explore whether the findings from this study are replicated when the results of different PRO measures are displayed.

Our purposive sampling of patients ensured representation across cancer types, care settings, and education levels. We recruited patients who were at least 6 months post-diagnosis and who had completed acute treatment, in order not to burden patients who were under active treatment. Thus, our sample may not reflect the perspectives of patients with short-term survival (e.g., pancreatic cancer) or who are currently dealing with acute side effects. Clinicians were sampled to represent a variety of specialties and care settings. Our purposive sampling strategy ensured inclusion of patient participants with only high school education (20 %). We note, however, that 40 % of patients had postgraduate education; this distribution of patient education experiences may have influenced how patients responded to different format options. An additional strength of the study design was that we controlled for potential learning or other order effects by randomly assigning format presentation order across interviews. Finally, we expect that these results from a cancer setting will be generalizable to other disease settings, although this requires confirmation.

This mixed-methods study provides quantitative data (ratings, % preferring overall, accuracy questions) to complement the qualitative feedback. We applied a relatively simple qualitative content analysis to meet the aims of this study; a more in-depth exploration of qualitative data from this and future studies could contribute further insights regarding particular issues, such as

why respondents had difficulty understanding the opposite scoring of symptom and function domains, but that was beyond the scope of the current study. While qualitative thematic saturation was confirmed by analyzing the last 14 interviews separately from the first 56, the sample size of 50 patients, and especially 20 clinicians, limited precision for the quantitative measures. The accuracy of interpretation questions was primarily directed at engaging respondents in active interpretation of the graphs. We interpret those findings with caution, as questions asked across formats may not have been equally difficult, and participants might have guessed the correct answers.

Nonetheless, the accuracy scores, when interpreted in the context of the data regarding intuitive understanding of directionality, do illustrate substantial room for improvement in communicating PROs effectively. The physical function line graph was consistent with participants' intuitive interpretation of whether higher scores on a function domain would be better or worse and were accurately interpreted by all clinicians and all but one patient. There was no clear consensus on the intuitive meaning of a symptom graph trending up, and this was reflected in the comparatively low accuracy rates for interpreting the fatigue line graph. Approaches to address this inconsistency in directionality are needed to improve PRO data interpretation.

We noted that for many formats, median ease-of-understanding and usefulness scores were 8/10 or higher. The wide range of responses on virtually all formats, however, emphasizes the fact that format preferences vary substantially, even within patient and clinician groups. Finally, the patients' and clinicians' preferences and interpretation reflect current levels of familiarity. While more complex formats (e.g., cumulative distribution function) were rated poorly, it is possible that with time and greater exposure, the information provided by these more complex approaches would be valued.

In the next steps of this research, we are engaging patients and clinicians in an iterative Work Group process to develop approaches to address the issues identified from this study. Work groups will develop potential approaches to address inconsistent directionality of scoring, to effectively highlight statistical significance and clinical meaningfulness (group-level), and to highlight clinically important scores and to provide reference scores (individual-level). Approaches developed by these Work Groups will then be tested in additional one-on-one interviews and subsequently evaluated in quantitative studies powered to appropriately test their interpretation. Our findings are expected to provide best practices for presenting PRO data to patients and clinicians to facilitate better understanding and application of PRO data in practice.

Acknowledgments This analysis was (partially) supported by a Patient-Centered Outcomes Research Institute (PCORI) Award (R-1410-24904). All statements in this report, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors or Methodology Committee. Drs. Snyder and Smith are members of the Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins (NCI P30 CA 006973). The funding sources had no role in study design, data collection, analysis, interpretation, writing, or decision to submit the manuscript for publication. The PRO Data Presentation Stakeholder Advisory Board includes Neil K. Aaronson, PhD (Netherlands Cancer Institute); Patricia A. Ganz, MD (University of California-Los Angeles and Jonsson Comprehensive Cancer Center); Ravin Garg, MD (Anne Arundel Medical Center); Michael Fisch, MD (M.D. Anderson Cancer Center); Bryce B. Reeve, PhD (University of North Carolina at Chapel Hill and Lineberger Comprehensive Cancer Center); Marc Silver (author of *Breast Cancer Husband—How to Help Your Wife (and Yourself) Through Diagnosis, Treatment, and Beyond*); Eden Stotsky-Himelfarb (Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins); Ellen Stovall (National Coalition for Cancer Survivorship); Matthew Zachary (Stupid Cancer). The Johns Hopkins Clinical Research Network (JHCRN) site investigators and staff include: Ravin Garg, MD, and Steven P. DeMartino, CCRC, CRT, RPFT (Anne Arundel Medical Center), Melissa Gerstenhaber, MAS, MSN, RN, CCRN (JHCRN/Anne Arundel Medical Center); Gary Cohen, MD, and Cynthia MacInnis, BS, CCRP (Greater Baltimore Medical Center); James Zabora, ScD, MSW (Inova Health System), and Sandra Schaefer, BSN, RN, OCN (JHCRN/Inova Health System); Paul Zorsky, MD, Lynne Armiger, MSN, CRNP, ANP-C, Sandra L. Heineken, BS, RN, OCN, and Nancy J. Mayonado, MS (Peninsula Regional Medical Center); Michael Carducci, MD (Johns Hopkins Sibley Memorial Hospital); Carolyn Hendricks, MD, Melissa Hyman, RN, BSN, OCN, and Barbara Squiller, MSN, MPH, CRNP (Suburban Hospital). Finally, we are most grateful to the patients and clinicians who participated in this study.

Conflict of interest The authors report no conflict of interest.

References

1. U.S. Food and Drug Administration. (2009). Guidance for industry. Patient reported outcome measures: Use in medical product development to support labelling claims. *Federal Register*, 74, 65132–65133.
2. Acquadro, C., Berzon, R., Dubois, D., et al. (2003). Incorporating the patient's perspective into drug development and communication: An ad hoc task force report of the patient-reported outcomes (PRO) harmonization group meeting at the food and drug administration, February 16, 2001. *Value in Health*, 6, 522–531.
3. Lipscomb, J., Gotay, C., & Snyder, C. (2005). *Outcomes assessment in cancer: Measures, methods, and applications*. Cambridge: Cambridge University Press.
4. Bruner, D. W., Bryan, C. J., Aaronson, N., et al. (2007). Issues and challenges with integrating patient-reported outcomes in clinical trials supported by the National Cancer Institute—sponsored clinical trials networks. *Journal of Clinical Oncology*, 25, 5051–5057.
5. Till, J. E., Osoba, D., Pater, J. L., et al. (1994). Research on health-related quality of life: Dissemination into practical applications. *Quality of Life Research*, 3, 279–283.
6. Au, H.-J., Ringash, J., Brundage, M., et al. (2010). Added value of health-related quality of life measurement in cancer clinical trials: The experience of the NCIC CTG. *Expert Review of Pharmacoeconomics and Outcomes Research*, 10(2), 119–128.

7. Bezjak, A., Ng, P., Skeel, R., et al. (2001). Oncologists' use of quality of life information: Results of a survey of eastern cooperative oncology group physicians. *Quality of Life Research*, *10*(1), 1–13.
8. Brundage, M., Bass, B., Jolie, R, et al. (2011). A knowledge translation challenge: Clinical use of quality of life data from cancer clinical trials. *Quality of Life Research*, *20*(7), 979–985.
9. Brundage MD, Feldman-Stewart D, Bezjak A, et al. (2005). *The value of quality of life information in a cancer treatment decision*. ISOQOL 11th annual conference, San Francisco.
10. Snyder, C. F., & Aaronson, N. K. (2009). Use of patient-reported outcomes in clinical practice. *Lancet*, *374*, 369–370.
11. Greenhalgh, J. (2009). The applications of PROs in clinical practice: What are they, do they work, and why? *Quality of Life Research*, *18*, 115–123.
12. Velikova, G., Booth, L., Smith, A. B., et al. (2004). Measuring quality of life in routine oncology practice improves communication and patient well-being: A randomized controlled trial. *Journal of Clinical Oncology*, *22*, 714–724.
13. Santana, M. J., Feeny, D., Johnson, J. A., et al. (2010). Assessing the use of health-related quality of life measures in the routine clinical care of lung-transplant patients. *Quality of Life Research*, *19*, 371–379.
14. Detmar, S. B., Muller, M. J., Schornagel, J. H., et al. (2002). Health related quality of life assessments and patient physician communication. *Journal of the American Medical Association*, *288*, 3027–3034.
15. Berry, D. L., Blumenstein, B. A., Halpenny, B., et al. (2011). Enhancing patient-provider communication with the electronic self-report assessment for cancer: A randomized trial. *Journal of Clinical Oncology*, *29*, 1029–1035.
16. PROQOLID (2012). The patient-reported outcome and quality of life instruments database. <http://www.proquolid.org>.
17. Snyder, C. F., Jensen, R., Courtin, S. O., et al. (2009). Patientviewpoint: A website for patient-reported outcomes assessment. *Quality of Life Research*, *18*(7), 793–800.
18. Jones, J. B., Snyder, C. F., & Wu, A. W. (2007). Issues in the design of internet-based systems for collecting patient-reported outcomes. *Quality of Life Research*, *16*, 1407–1417.
19. Snyder, C. F., Blackford, A. L., Wolff, A. C., et al. (2013). Feasibility and value of Patientviewpoint: A web system for patient-reported outcomes assessment in clinical practice. *Psycho-Oncology*, *22*, 895–901.
20. Brundage, M., Feldman-Stewart, D., Leis, A., et al. (2005). Communicating quality of life information to cancer patients: A study of six presentation formats. *Journal of Clinical Oncology*, *23*, 6949–6956.
21. McNair, A. G., Brookes, S. T., Davis, C. R., et al. (2010). Communicating the results of randomized clinical trials: Do patients understand multidimensional patient-reported outcomes? *Journal of Clinical Oncology*, *28*, 738–743.
22. Aaronson, N. K., Ahmedzai, S., Bergman, B., et al. (1993). The european organization for research and treatment of cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. *Journal of the National Cancer Institute*, *85*, 365–376.
23. Abernethy, A. P., Wheeler, J. L., & Zafar, S. Y. (2010). Management of gastrointestinal symptoms in advanced cancer patients: The rapid learning cancer clinic model. *Current Opinion in Supportive and Palliative Care*, *4*, 36–45.
24. PROMIS Software Demonstration (2012). <http://nihpromis.org/software/demonstration>.
25. Creswell, J. W. (2007). *Qualitative inquiry and research design: Choosing among five approaches* (2nd ed.). Thousand Oaks: Sage Publications.
26. Pope, C., Ziebland, S., & Mays, N. (1999). Analysing qualitative data. In C. Pope & N. Mays (Eds.), *Qualitative research in health care* (2nd ed., pp. 75–88). London: BMJ Publishing Group.
27. (2014). AtlasTi, in, ATLAS.ti Scientific Software Development GmbH.