Quality of Life Research (2020) 29:1903–1911
https://doi.org/10.1007/s11136-020-02448-y

FDA review summary of patient-reported outcome results for ibrutinib in the treatment of chronic graft versus host disease

Bellinda L. King-Kallimanis[1,3] · Tanya Wroblewski[2] · Virginia Kwitkowski[2] · R. Angelo De Claro[2] · Thomas Gwise[2] · Vishal Bhatnagar[2] · Ann T. Farrell[2] · Paul G. Kluetz[1]

Accepted: 12 February 2020 / Published online: 25 February 2020

Abstract

Purpose On August 2, 2017, the Food and Drug Administration approved ibrutinib (IMBRUVICA) for the treatment of patients with chronic graft versus host disease (cGVHD) after the failure of one or more lines of systemic therapy. The approval was based on results from a single-arm, multicenter trial that enrolled patients with refractory cGVHD. This paper describes the FDA review of patient-reported outcomes (PRO) data from Study PCYC-1129-CA and the decision to incorporate descriptive PRO data in the FDA label to support the primary clinician-reported outcome results.

Methods In this trial, the Lee Chronic GVHD Symptom Scale (LSS) was used to capture patient-reported symptom bother. The 42 patients who received treatment were included in the analysis and completed the PRO tool. Post hoc descriptive analyses were conducted to further understand the measurement properties of the LSS.

Results The analysis submitted to FDA reported that 18 patients had a ≥ 7-point improvement on the LSS overall summary score at any point during the assessment period. For 10 patients, the ≥ 7-point improvement was sustained for ≥ 2 consecutive PRO assessments. An assessment of the responder threshold suggested the threshold submitted to the FDA was reasonable and in line with clinical findings.

Conclusions Overall, study PCYC-1129-CA demonstrated favorable clinician-reported cGVHD efficacy results that were complemented by results from PRO data, supporting the FDA's positive benefit-risk assessment leading to regular approval. Limitations included the single-arm trial design, responder definition, and instrument shortcomings. These limitations were thoroughly explored through additional FDA post hoc analyses.

Keywords FDA · Patient-reported outcomes · Chronic graft versus host disease · Clinical trials

Introduction

Chronic graft versus host disease (cGVHD), a serious and life-threatening condition, occurs in approximately 30–70% of patients who receive allogeneic hematopoietic stem cell transplantation (HSCT) [1–3]. It is characterized by complex allogeneic and autoimmune dysregulation of the immune system. Symptoms may impact multiple organs with a predilection for oral and ocular mucosa, skin, lung, liver, gastrointestinal, and genitourinary tract epithelium. Prior to August 2017, there were no approved second-line treatments and no standard of care. Conducting a randomized, controlled trial in refractory cGVHD is challenging because of its rarity, life-threatening nature, and difficulty identifying a comparator arm.

On August 2, 2017, the Food and Drug Administration (FDA) approved ibrutinib (IMBRUVICA, AbbVie

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11136-020-02448-y) contains supplementary material, which is available to authorized users.

✉ Bellinda L. King-Kallimanis
belinda.kallimanis@fda.hhs.gov

[1] Oncology Center of Excellence, U.S. Food and Drug Administration, Building 22, 10903 New Hampshire Avenue, Silver Spring, MD 20993, USA

[2] Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, USA

[3] U.S. Food and Drug Administration, WO22 Room 2372, 10903 New Hampshire Avenue, Silver Spring, MD 20993, USA

Inc.) for the treatment of patients with cGVHD after failure of one or more lines of systemic therapy. Ibrutinib is the first FDA-approved drug for the treatment of cGVHD. Approval was based on results from study PCYC-1129-CA (NCT02195869), a single-arm trial of 42 patients with cGVHD after progression on first-line corticosteroid therapy and requiring additional therapy. The primary endpoint was a clinician-reported outcome (ClinRO); best overall cGVHD response rate (BORR) per the 2005 National Institutes of Health (NIH) Consensus Panel Response Criteria with modification to align with the updated 2014 NIH Consensus Panel Response Criteria. Using this ClinRO, ibrutinib demonstrated a BORR of 66.7% ($n = 28$, 95% CI 50.5%, 80.4%), which included both partial and complete responders. Median time to response was 12.3 weeks. A sustained response (≥ 20 weeks) was demonstrated in 48% of the patients for whom there was no available therapy. In addition, responses were seen across different organ involvement within the first 3 months of treatment. A sustained response in approximately half of patients with an unmet medical need can be considered clinically meaningful. Importantly, the ClinRO results were supported by favorable patient-reported outcome (PRO) results [4].

Recent legislation, including the Twenty-first Century Cures Act, has highlighted the importance of capturing patient input to inform medical product development. As the symptom burden associated with cGVHD is high, PRO measures are especially useful for capturing relevant symptoms and describing patients' experience with treatments and interventions. One commonly used PRO strategy in clinical trials to capture patient experience is the inclusion of fit-for-purpose PRO measures. The FDA defines "fit-for-purpose" as "a conclusion that the level of validation associated with a medical product development tool is sufficient to support its context of use" [5]. The Lee Chronic GVHD Symptom Scale (hereafter referred to as the LSS) is a PRO measure developed in 2002 to assess the heterogeneous symptom bother and impacts of cGVHD [6] and was included in the registration trial for ibrutinib.

In previous research, patients with cGVHD who completed PRO measures were found to experience detrimental effects to their physical functioning and other symptoms when disease symptoms increase in severity [7]. In another study, newly diagnosed cGVHD patients who met clinical criteria for response, had greater reduction in symptom burden [8].

Despite the advantages of collecting PRO data in rare disease trials such as those conducted in cGVHD, one challenge is the frequent absence of a control arm, leading to concern that patients may overestimate benefit when aware of treatment assignment [9]. In this trial, additional limitations of the PRO results included identification of a responder definition, and other instrument shortcomings. In this manuscript, we report the FDA review of PRO data from Study PCYC-1129-CA, and the decision to incorporate descriptive data from a PRO measure in the FDA label to support the primary clinical results.

Materials and methods

Study participants

Study PCYC-1129-CA was a multicenter, single-arm, open-label phase 1b/2 trial of ibrutinib in patients with steroid dependent or refractory cGVHD after allogeneic HSCT. Patients were required to have received ≤ 3 prior therapies for cGVHD. Patients were also required to have either > 25% body surface area erythematous rash or NIH mouth score > 4 [10]. Patients in the trial were asked to complete screening, treatment and follow-up assessments. A total of 45 patients were enrolled, and 43 treated. One patient who received ibrutinib was excluded due to relapse of underlying disease at baseline, resulting in a population of 42 patients. The design of this trial has been described in further detail elsewhere [11].

PRO measures

The LSS consists of 30 items that are used to create 7 subscales: Skin (5 items), Eye (3 items), Mouth (2 items), Lung (5 items), Nutrition (5 items), Energy (7 items), and Psychological (3 items) [6]. For each item, patients rate how bothered they were by symptoms (e.g., mouth ulcers), impacts (e.g., avoiding certain foods) and medical interventions (e.g., use of eye drops) over the past month using a 5-point response scale with the options: 0 = Not at all, 1 = Slightly, 2 = Moderately, 3 = Quite a bit and 4 = Extremely. Items are summed to generate subscale scores, and the LSS total score is calculated as the average of the subscale scores. All calculated scores are linearly transformed to a 0–100 scale (per scoring algorithm). A higher score indicates more bother from cGVHD symptoms. A decrease or improvement of ≥ 7-points on the LSS total score has been published as a clinically meaningful difference. This > 7-point threshold was calculated using distribution methods (i.e., half a standard deviation of the baseline LSS total score for the population) [6].

In addition to the LSS, the drug sponsor collected the Patient Self-Report section (Form B) of the NIH cGVHD Response Assessment Form [12]. For this study, FDA focused on two global items:

- Item 1. “Overall, do you think that your chronic graft versus host disease is mild, moderate, or severe?” (0 = None; 1 = Mild; 2 = Moderate; 3 = Severe)
- Item 3. “Compared to a month ago, overall would you say your chronic GVHD symptoms are” (3 = Very much better, 2 = Moderately better, 1 = A little better, 0 = About the same, − 1 = A little worse, − 2 = Moderately worse, − 3 = Very much worse).

Data were collected at week 1 (baseline), week 13, and every 12 weeks thereafter, with additional assessments at the progressive disease visit (if applicable), end of treatment (EoT) visit and response follow-up visits. A late protocol amendment was implemented to capture an additional PRO assessment at week 5.

Statistical analysis for PRO

The PRO measure was used to assess the secondary endpoint: “Change in symptom burden measured by the Lee cGVHD Symptom Scale,” with no adjustment for Type I error. The analyses presented by the applicant were replicated by the FDA and will be presented in the results section. Summary statistics were used to describe the LSS total and subscale scores over study visits. A responder analysis was conducted using a ≥ 7-point improvement on the LSS total score as the response threshold based on the previous literature. Patients who experienced a ≥ 7-point improvement are referred to as patients with a PRO response in this paper. A sub-group analysis of PRO responders by the clinical outcome, best overall response was also investigated. Finally, mean change from baseline on the LSS total score was assessed.

Post hoc *FDA analysis*

Completion rate for the LSS was calculated as the number of patients completing > 50% of items at each PRO assessment divided by the number of patients expected to complete the LSS at that assessment (i.e., patients still on treatment). The denominator did not include patients who had progressed or died [13].

The sensitivity analyses outlined below addressed two limitations: (1) the ≥ 7-point threshold may not be meaningful and (2) open-label bias may have overestimated the treatment benefit.

First, the relationship between the PRO response was compared to clinical response using descriptive statistics. Next, the threshold for meaningful change for the LSS total score was assessed using anchor-based methods supplemented with cumulative distribution function (CDF) curves as is suggested in the FDA Guidance to Industry for PROs [9]. Here, the PRO measurement results are defined (i.e., anchored) in terms of change external to, in this case, the LSS. The anchors were: patient-reported change from baseline on global cGVHD severity (item 1) and change in overall cGVHD symptoms (item 3) from the NIH Response Assessment. Due to small sample size, adjacent response options on global change (item 3) were collapsed (e.g., Better = Very much better, Moderately better, A little better). This resulted in three categories; Worse, About the same and Better. Change from baseline and week 13 was used for global severity (item 3) where > 0 = Better, 0 = No change and < 0 = Worse. Week 13 was used due to reduced sample size at subsequent assessments. The mean score of the improvement group was considered as the threshold for meaningful change.

Baseline differences on the LSS total score, subscales and psychological items were explored between patients with and without a PRO response. This was presented under the assumption that certain subscales may have been more sensitive due to this being an open-label study.

Finally, we looked at floor and ceiling effects for each item. These effects were considered present if more than 20% of baseline responses were in the highest (ceiling) or lowest (floor) response categories. For example, a floor effect for an item would exist if more than 20% of patients responded “Not at all” to being bothered, whereas a ceiling effect would be present if > 20% of patients responded as being “extremely” bothered.

Analyses were performed on the pooled phase 1b/2 data (i.e., all-treated population). All analyses were completed using SAS software (release 9.4, SAS Institute, Inc., Cary, NC).

Results

Forty-two patients were included in the all-treated analysis population. Median age was 56 years (range 19, 74 years) and 52.4% were male. Median duration of time on treatment was 4.4 months (range 7 days, 24.9 months), with 12 responding patients still on treatment at end of study.

PRO: completion rates

All 42 patients completed the baseline assessment. Post-baseline completion for the LSS was > 83% at all other designated clinic visits, except for the week 5 visit, which was added as a late protocol amendment (Table 1 in Online Appendix). Ten patients completed a week 5 assessment, however, the completion rate is unclear as the denominator (number of patients eligible) after the protocol amendment was not adequately described in the submission.

PRO: mean change from baseline

Mean change from baseline for the LSS total and subscale scores was reported for each study visit for patients who completed an assessment. Over the first 12 months of treatment, the mean change from baseline for LSS total score monotonically improved from − 1 (standard deviation (SD) = 10, *N* = 10) at week 5 to − 9 (SD = 12, *N* = 15) at week 49 (Table 1). The largest changes were observed for the Skin and Eye subscales (Figs. 1 and 2 in Online Appendix). Item-level change for these two subscales indicated that no single item was responsible for the change.

Patients with ≥ 7-point improvement on LSS total score

Analyses submitted to FDA reported that 18 (42.9%) patients had a ≥ 7-point improvement on the LSS total score at any point during the assessment period (Table 2). Seventeen of these 18 patients were classified by the investigator as experiencing a clinical partial response or better.

Post hoc *FDA analysis*

Table 1 Mean change from baseline for LSS total score

	Mean LSS Total score (SD)	Mean change from baseline
Baseline (*n* = 42)	34 (13)	NA
Week 5 (*n* = 10)[a]	35 (16)	− 1 (10)
Week 13 (*n* = 32)	30 (14)	− 4 (10)
Week 25 (*n* = 18)	30 (15)	− 4 (17)
Week 37 (*n* = 16)	28 (17)	− 5 (12)
Week 49 (*n* = 15)	26 (14)	− 9 (12)
Week 61 (*n* = 10)	29 (15)	− 5 (18)
Week 73 (*n* = 8)	25 (12)	− 7 (15)
Week 85 (*n* = 3)	30 (16)	− 8 (5)
Week 97 (*n* = 3)	25 (12)	− 14 (9)
Week 109 (*n* = 2)	19 (2)	− 8 (9)

Scores are transformed on a 0–100 scale and higher score indicate greater bother

[a] Week 5 added as a later protocol amendment

Table 2 Number of patients by clinical response and PRO LSS total score responders

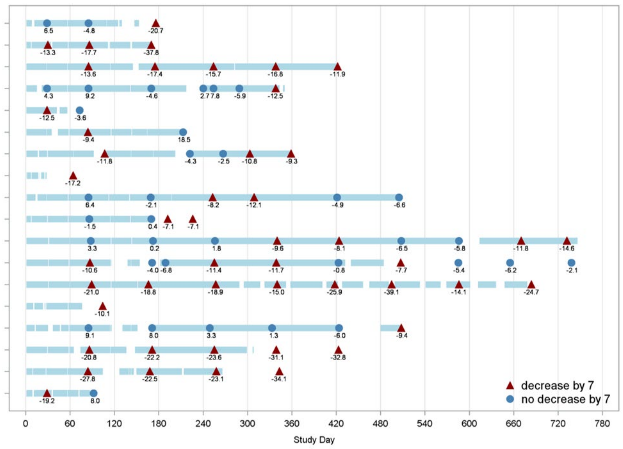
	PR or CR (*n* = 28)	No CR/PR (*n* = 14)	All patients (*n* = 42)
No PRO LSS total score response	9	7	16
PRO LSS total score response	17	1	18
Missing PRO[a]	2	6	8

PR partial response, *CR* complete response, *PRO* patient-reported outcome, *LSS* Lee Symptom Scale

[a] Missing is due to patient only having a baseline PRO assessment and no follow-up assessments after starting therapy

FDA further explored duration of PRO response using the ≥ 7-point threshold. Ten out of 42 (23.8%) patients had a ≥ 7-point improvement on the LSS total score at any point that was maintained for ≥ 2 consecutive visits. Of these ten sustained responses, 1 patient had an initial response that was captured at the EoT visit and was sustained at a follow-up visit. Of the patients who were considered PRO responders, their mean change from baseline was − 14.2 (SD = 5.7, range − 7.1, − 27.7), and the median time to an improvement of ≥ 7-points was 2.9 months (range 0.9, 16.69, Fig. 1).

FDA assessed the meaningfulness of the 7-point change threshold. The Spearman correlation between change from baseline at week 13 on the global severity of cGVHD item and the LSS total score was 0.5 and − 0.3 between patient global impression of change and change from baseline at week 13 for the LSS total score. This suggests these anchors were appropriate [14]. At week 13, nine (29%) patients reported less (better) severity of their GVHD symptoms and 13 (42%) patients reported an improved (better) change in GVHD symptoms over the past month (Table 3). No patients had worse severity when comparing their week 13 score to baseline, but three patients reported their symptoms had gotten worse over the past month based on the global impression of change. The patient global severity item possibly overestimated the effect (effect size > 0.5, Table 3), therefore, we focus on the LSS threshold using global impression of change as the anchor (item 3). This analysis suggested a meaningful threshold to be 6.4 or greater. The CDF curves (Fig. 8a, b in Online Appendix) revealed a separation



Solid bars represent days on ibrutinib treatment.
Displayed values represent change from baseline.

Fig. 1 Swimmer's plot for PRO LSS total score responders (N = 18)

Table 3 Anchor-based analysis of Lee Chronic GVHD Scale total score

	LSS total score			
	n	SD at baseline	Mean change	Effect size
Change in global severity (*n* = 31)				
Better	9	13.73	− 8.95	− 0.65
No change	22		− 1.41	0.10
Global impression of change (*n* = 31)				
Better	13		− 6.35	− 0.46
Worse	3	13.73	− 2.17	− 0.16
No change	15		− 1.5	− 0.11

Change in global severity was calculated as the difference between the patient-reported score from baseline and week 13. The global impression of change was patients self-report of change in their overall symptoms at week 13 over the past month. One patient had missing responses to the anchors, therefore, the SD of the LSS total score at baseline is calculated for *n* = 31

between the stable and better group. There were too few patients who reported worsening symptoms to interpret that curve, and worsening was not included.

Further LSS analyses

FDA analysis noted floor effects at baseline for 20 of the 30 items. For 10 of these items, more than 50% of patients endorsed the lowest response category (0 = Not at All). Three items had ceiling effects (Table 4).

For medical intervention items, no patients reported bother with the intravenous line/feeding tube item, and only 2 patients reported slight or moderate bother on the use of oxygen item throughout the study. For the item assessing bother associated with eye drop use, at any time during the study, more than 50% of patients reported being bothered by the frequent use of eye drops.

Baseline characteristics of patients with ≥ 7-point improvement on the LSS

To understand whether patients with a clinical response reported different responses to the LSS items at baseline, we looked at descriptive statistics for the baseline assessment by clinical response. Overall, the patients with a PRO response had, on average, higher scores on the LSS total score and 6 of 7 of the subscales at baseline (Table 5).

Table 4 Floor and ceiling effects for the LSS items

Item	Floor	Ceiling
Abnormal skin color	9 (21%)	NA
Rashes	15 (36%)	NA
Thickened skin	16 (38%)	NA
Sores on skin	17 (40%)	NA
Itchy skin	NA	NA
Shortness of breath	13 (31%)	NA
Joint and muscle aches	NA	NA
Limited joint movement	12 (29%)	NA
Muscle cramps	12 (29%)	NA
Weak muscles	NA	NA
Loss of energy	NA	NA
Need to sleep more	NA	NA
Frequent cough	30 (71%)	NA
Colored sputum	36 (88%)	NA
Short of breath at rest	33 (79%)	NA
Need to use oxygen	41 (98%)	NA
Fevers	41 (98%)	NA
Dry eyes	NA	13 (31%)
Use eye drops freq	NA	14 (33%)
Difficulty seeing clearly	NA	NA
Feeding tube	42 (100%)	NA
Difficulty swallowing solid food	27 (64%)	NA
Difficulty swallowing liquid	33 (79%)	NA
Vomiting	40 (95%)	NA
Weight loss	30 (71%)	NA
Avoid certain foods due to mouth pain	NA	14 (33%)
Ulcers in mouth	17 (40%)	NA
Depression	18 (43%)	NA
Anxiety	17 (41%)	NA
Difficulty sleeping	NA	NA

Ceiling and floor effects we considered present when either the highest or lowest response option had more than 20% of the sample endorse this option

Discussion

Results from our FDA analyses suggest that patients who experienced a clinical response while on ibrutinib were also likely to self-report reduced bother in their symptoms using the LSS. This review of the PRO data was challenging due to several important limitations of the trial design, assessment tool and analysis. Our review focused on three main issues, (1) responder definition/duration of response on the LSS total score, (2) appropriateness of LSS instrument, and (3) study design (single-arm trial/concern for bias).

Table 5 Baseline descriptive statistics for the LSS subscales and total scores by LSS PRO responder sub-group

	PRO responder ($n = 18$)			PRO non-responder ($n = 16$)		
	Mean	Median	Range	Mean	Median	Range
Total score	36	36	8–65	30	29	17–49
Skin	42	42	0–75	35	30	0–85
Energy	49	50	0–86	42	39	21–79
Lung	6	0	0–20	3	0	0–15
Eye	61	62	0–100	62	75	17–92
Nutrition	10	0	0–50	6	0	0–35
Mouth	46	44	0–100	39	31	0–100
Psychological	37	29	0–83	25	25	0–58

All scores were transformed into a scale of 0–100

Denominators do not add to 42 because 8 patients did not have an evaluable follow-up that included the PRO assessments

Responder definition: clinically meaningful change threshold for the LSS

Clinically meaningful change for the LSS has been proposed as a decrease of ≥ 7-points on the LSS total score [6]. This threshold was arrived at using a distribution-based method [15]. These methods are data driven, do not reflect the patient's assessment, and are often considered to be supportive to anchor-based methods. However, including anchors in trials may not always be feasible, and there is some evidence that there may be more commonality than difference [16]. Using the patient-rated global items, an anchor-based calculation was applied and supplemented with CDF curves. The threshold estimated was a change of 6.4-points on the LSS total score, which corresponds closely to the 7-point threshold estimated by Lee et al. [6]. The CDF plots also suggested threshold values close to this magnitude to be reasonable. The majority of patients (78%) had a change that was at least 11 points (4 points greater than 7), suggesting that even if a slightly more conservative threshold had been chosen, few patients would have been reclassified.

These estimations are limited by a trial not designed for these analyses, e.g., the anchor item has a 1-month recall period and was used to assess mean change from baseline at 13 weeks, and a very small sample size. Despite these limitations there was concordance observed with the literature. Given this and the results presented, the FDA review concluded that the magnitude of the patient-reported responses were supportive of meaningful clinical benefit.

Limitations of the LSS

The LSS was developed in 2002 to evaluate how bothersome patients find their symptoms [6]. This questionnaire has been well established for use in clinical practice, however, it was not designed as a standalone outcome measure for clinical trials to support regulatory action. Despite this, there are benefit of using patient-reported symptoms which can include bother and impact, which compliments clinician-assessed symptoms. Further study should be done on whether PRO measures are being routinely employed in clinical practice for the assessment of cGVHD. Currently, care guidelines, such as the guidelines published by the National Comprehensive Cancer Network do not include recommendations to capture patient reports of symptoms [17].

The LSS only measures symptom bother; other important aspects of symptoms such as severity and interference with function are not captured. Symptom bother can be a challenging concept to measure and can vary as a function of disease stage and individual tolerance. For example, patients may report being bothered by a symptom that is not very severe, or, a patient may come to tolerate a symptom and report less "bother" even though the symptom remains severe. Because of these challenges, FDA generally recommends measuring symptom severity or frequency, where appropriate, as these concepts might be more sensitive to the treatment effect. However, FDA recognizes that bother, burden, or interference can provide additional important information once severity or frequency is established.

In this trial, patient-reported global severity of cGVHD and impression of change in cGVHD symptoms were assessed. FDA analyzed the relationship between these global items and the LSS total score and found moderate correlations between global severity and LSS total score (baseline and week 13). Evidence of agreement for patients reporting improvement at the same PRO assessment on both a global item and the LSS total score was weak (Table 2 in Online Appendix). Poor agreement could be due to the threshold, or differences between a single-item measure versus a multi-item measure, or because each instrument measures a slightly different concept, or finally some combination of these issues. In general, because this trial was not designed to optimally carry out additional assessments of the measurement properties of the LSS, some findings are difficult to interpret. At best, we found moderate evidence that in addition to improvements in bother, patients were reporting decreases in overall disease and symptom severity.

Another challenge with the LSS concerns the subscales and item content. For example, it is unclear why items measuring bother by *joint and muscle aches,* are scored as part of the Energy subscale. As our primary focus was on the LSS total score, the mapping of items to subscales was not further explored. If future studies focus on change in the subscales, additional work will be required to determine whether the subscales can be considered fit-for-purpose. Additionally, as an exploratory trial objective was to evaluate the effect of ibrutinib on cGVHD symptom bother, the inclusion of items measuring bother due to other non-investigational medical treatments on the LSS was problematic. The impact of these items was considered small as only one of the medical treatments was prevalent (use of eye drops) in the trial population.

Finally, many items were found to have either floor or ceiling effects. This remains a limitation of LSS, although these effects may be difficult to avoid entirely given the clinical reality that cGVHD symptomatology is heterogeneous. For instance, while it is important to cover all symptoms/outcomes, not all symptoms and functional impacts will occur for an individual patient. This was observed in this trial and can result in large proportions of patients responding with the lowest (floor) response options, making it difficult to distinguish between patients due to a lack of sensitivity. Despite the large proportion of items with floor effects in this trial, we still observed change.

The decision to incorporate LSS results in FDA labeling of ibrutinib should not be construed as endorsement of the current LSS questionnaire as "fit-for-purpose" for a clinical outcome assessment to quantify benefit for cGVHD in registration trials. We support efforts to further modify and improve the LSS and other measures of cGVHD symptoms and impacts for use in future cGVHD trials. Efforts to improve the measurement of cGVHD symptoms could consider the limitations outlined in this manuscript, particularly around alignment of domains and item content. Instrument developers are encouraged to meet with FDA to obtain specific recommendations on how to adapt existing PRO instruments for regulatory purposes early in their drug development program.

Limitations of the trial design: open-label trials and concern for bias

Non-randomized studies are susceptible to bias through knowledge of treatment assignment which may lead to expectation of treatment benefit. Additionally, concomitant treatments that would be expected to affect patients' reports of their symptoms (e.g., topical treatments) should be standardized, recorded and analyzed. This was not the case in this trial and this information could not be incorporated into our analysis.

The degree to which PRO results are influenced by response bias in open-label trials is poorly understood and there are no agreed-upon methods to account for this potential effect. FDA hypothesized that knowledge of treatment assignment may provide emotional benefit, and that the psychological subscale may be more susceptible to overestimation of treatment benefit. However; our analysis of the psychological subscale did not find evidence to suggest that PRO responders were overly influenced by improvements on this subscale. Skin and Eye subscales were most improved; however, skin and eye are hallmark symptoms of cGVHD, and patients reported high skin and eye bother at baseline, in part due to the inclusion criteria requiring patients have either > 25% erythematous rash or > 4 total mouth score per NIH criteria. This requirement may have influenced the dominance of these domains. Alternatively, this could indicate that the other cGVHD symptoms were not as relevant to the study population.

Another assumption explored was the notion that if a response was driven by being in an open-label trial (perception of symptom benefit in the absence of true therapeutic efficacy), the observed PRO benefit would occur early, and be less durable as treatment side effects and untreated disease symptoms would overcome this response bias. In this trial, assessment of early improvement was limited because time between baseline and first on-treatment assessment was 13 weeks for 3 quarters of the patients enrolled and this corresponds to the median time to response on the LSS total score of 2.9 months (range 0.9, 16.7). While an amendment was made to include a week 5 assessment, only 10 patients completed this assessment. The fact that half the responses occurred or were still present after 3 months of treatment, and more than half of the patients had a response that lasted 2 or more visits suggested that responses were not all early and of short duration.

Finally, results were consistent with previous research identifying an association between PR/CR (i.e., the clinical response) and an LSS improvement [18]. In another study of imatinib or rituximab to improve cutaneous sclerosis in patients with cGVHD, the authors observed a significant decrease in the Skin subscale for patients in the imatinib arm. This was generally in line with clinical findings [19].

Based on FDA sensitivity analyses, it was felt to be unlikely that the PRO results were heavily influenced by response bias due to being an open-label trial. However, patients may still have perceived a larger magnitude of benefit knowing they were on an investigational agent, and this remains a significant limitation of the study design.

Importantly, the PRO results were not the primary endpoint of the study and were providing supportive evidence of treatment efficacy demonstrated by clinical evaluation of both signs and symptoms of cGVHD.

Our focus was to describe analyses of the Lee Symptom Scale and its use in the regulatory context of the ibrutinib approval. We recognize an important area of future research is to understand the relationship between PROs and clinical adverse events (e.g., infections such as pneumonia).

Conclusion

Study PCYC-1129-CA demonstrated favorable clinician-reported cGVHD efficacy results that were complemented by results from PRO data, supporting the FDA's positive benefit–risk assessment leading to regular approval. Limitations of the PRO results include single-arm trial design, responder definition, and instrument shortcomings. These limitations were thoroughly explored through additional FDA post hoc analyses. Despite the limitations identified with the LSS as a clinical outcome assessment for regulatory use, the tool is familiar to physicians treating cGVHD and the FDA review concluded these results were important to convey to treating physicians in the product label. Modification of the LSS to improve its use as a clinical outcome assessment tool for regulatory decision-making should be considered for future trials.

Author contributions BKK, TG, ADC, VK and PK designed the research, BKK and ADC conducted analyses, TW, TG, AF and PK provided oversight for the analyses, BKK, TW, VK, ADC, VB, AF and PK wrote the manuscript.

Funding None.

Compliance with ethical standards

Conflict of interest The authors declare no competing financial interests.

Ethical approval IRB: This is US Government work and is a secondary analysis of data that was submitted to the Food and Drug Administration, therefore, IRB was not applicable. The trial identifier is NCT02195869.

References

1. Flowers, M. E. D., & Martin, P. J. (2014). How we treat chronic graft-versus-host disease. *Blood*. https://doi.org/10.1182/blood-2014-08-551994.
2. Martin, P. J., Counts, G. W., Appelbaum, F. R., Lee, S. J., Sanders, J. E., Deeg, H. J., et al. (2010). Life expectancy in patients surviving more than 5 years after hematopoietic cell transplantation. *Journal of Clinical Oncology, 28*(6), 1011–1016. https://doi.org/10.1200/JCO.2009.25.6693.
3. Wingard, J. R., Majhail, N. S., Brazauskas, R., Wang, Z., Sobocinski, K. A., Jacobsohn, D., et al. (2011). Long-term survival and late deaths after allogeneic hematopoietic cell transplantation. *Journal of Clinical Oncology, 29*(16), 2230–2239. https://doi.org/10.1200/JCO.2010.33.7212.
4. Drugs @ FDA Full Prescribing Information—Imbruvica. Retrieved January 4, 2018, from https://www.accessdata.fda.gov/drugsatfda_docs/label/2018/210563s000lbl.pdf.
5. FDA-NIH Biomarker Working Group. (2016). BEST (Biomarkers, EndpointS, and other Tools) Resource.
6. Lee, S. J., Cook, E. F., Soiffer, R., & Antin, J. H. (2002). Development and validation of a scale to measure symptoms of chronic graft-versus-host disease. *Biology of Blood and Marrow Transplantation, 8*(8), 444–452. https://doi.org/10.1053/bbmt.2002.v8.pm12234170.
7. Pidala, J., Kurland, B. F., Chai, X., Vogelsang, G., Weisdorf, D. J., Pavletic, S., et al. (2011). Sensitivity of changes in chronic graft-versus-host disease activity to changes in patient-reported quality of life: Results from the Chronic Graft-versus-Host Disease Consortium. *Haematologica, 96*(10), 1528–1535.
8. Inamoto, Y., Martin, P. J., Chai, X., Jagasia, M., Palmer, J., Pidala, J., et al. (2012). Clinical benefit of response in chronic graft-versus-host disease. *Biology of Blood and Marrow Transplantation, 18*(10), 1517–1524. https://doi.org/10.1016/j.bbmt.2012.05.016.
9. US Department of Health and Human Services; US Food and Drug Administration; Center for Drug Evaluation and Research (CDER); Center for Biologics Evaluation and Research (CBER); Center for Devices and Radiological Health (CDRH). (2009). Guidance for industry. Patient-reported outcome measures: Use in medical product development to support labeling claims. Silver Spring, MD.
10. Pavletic, S. Z., Martin, P., Lee, S. J., Mitchell, S., Jacobsohn, D., Cowen, E. W., et al. (2006). Measuring therapeutic response in chronic graft-versus-host disease: National Institutes of Health consensus development project on criteria for clinical trials in chronic graft-versus-host disease: IV. Response criteria working group report. *Biology of Blood and Marrow Transplantation, 12*(3), 252–266.
11. Miklos, D., Cutler, C. S., Arora, M., Waller, E. K., Jagasia, M., Pusic, I., et al. (2017). Ibrutinib for chronic graft-versus-host disease after failure of prior therapy. *Blood*. https://doi.org/10.1182/blood-2017-07-793786.
12. Lee, S. J., Wolff, D., Kitko, C., Koreth, J., Inamoto, Y., Jagasia, M., et al. (2015). Measuring therapeutic response in chronic graft-versus-host disease: National Institutes of Health consensus development project on criteria for clinical trials in chronic graft-versus-host disease: IV. The 2014 Response Criteria Working Group report. *Biology of Blood and Marrow Transplantation, 21*(6), 984–999. https://doi.org/10.1016/j.bbmt.2015.02.025.
13. Osoba, D., Bezjak, A., Brundage, M., Zee, B., Tu, D., Pater, J., et al. (2005). Analysis and interpretation of health-related quality-of-life data from clinical trials: Basic approach of The National Cancer Institute of Canada Clinical Trials Group. *European Journal of Cancer, 41*(2), 280–287. https://doi.org/10.1016/j.ejca.2004.10.017.
14. Yost, K. J., Cella, D., Chawla, A., Holmgren, E., Eton, D. T., Ayanian, J. Z., et al. (2005). Minimally important differences were estimated for the Functional Assessment of Cancer Therapy-Colorectal (FACT-C) instrument using a combination of distribution- and anchor-based approaches. *Journal of Clinical Epidemiology, 58*(12), 1241–1251. https://doi.org/10.1016/j.jclinepi.2005.07.008.
15. Osoba, D., Rodrigues, G., Myles, J., Zee, B., & Pater, J. (1998). Interpreting the significance of changes in health-related

quality-of-life scores. *Journal of Clinical Oncology, 16*(1), 139–144. https://doi.org/10.1200/JCO.1998.16.1.139.

16. Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2003). Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. *Medical Care, 41*(5), 582–592.
17. National Comprehensive Cancer Network. Hematopoietic Cell Transplantation: Pre-Transplantation Recipient Evaluation and Management of Graft-Versus-Host Disease (Version 1.2020). Retrieved November 22, 2019, from https://www.nccn.org/professionals/physician_gls/pdf/hct.pdf.
18. Martin, P. J., Storer, B. E., Inamoto, Y., Flowers, M. E. D., Carpenter, P. A., Pidala, J., et al. (2017). An endpoint associated with clinical benefit after initial treatment of chronic graft-versus-host disease. *Blood, 130*(3), 360–367. https://doi.org/10.1182/blood-2017-03-775767.
19. Arai, S., Pidala, J., Pusic, I., Chai, X., Jaglowski, S., Khera, N., et al. (2016). A randomized phase II crossover study of imatinib or rituximab for cutaneous sclerosis after hematopoietic cell transplantation. *Clinical Cancer Research, 22*(2), 319–327. https://doi.org/10.1158/1078-0432.CCR-15-1443.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.