Psychometric Evaluation of the Advanced Systemic Mastocytosis Symptom Assessment Form (AdvSM-SAF) in patients with Advanced Systemic Mastocytosis

Fiona Taylor,[1] Alan L. Shields,[1] Shirley Li,[1] Christine Yip,[1] Brad Padilla,[1] Tanya Green,[2] Deepti Radia,[3] Michael Deininger,[4] Jason R. Gotlib,[5] Prithviraj Bose,[6] Mark W. Drummond,[7] Elizabeth Hexner,[8] William Robinson,[9] Albert Quiery, Jr.,[10] Elliott F. Winton,[11] Daniel J. DeAngelo,[12] Brenton Mar[2]

[1]Adelphi Values, [2]Blueprint Medicines Corporation, [3]Guy and St. Thomas' NHS Foundation Trust, [4]University of Utah, [5]Stanford Cancer Institute, [6]The University of Texas MD Anderson Cancer Center, [7]Beatson West of Scotland Cancer Centre, [8]Abramson Cancer Center at the University of Pennsylvania, [9]University of Colorado Hospital (UCH), [10]University of Michigan Rogel Cancer Center, [11]Winship Cancer Institute of Emory University, [12]Dana Farber Cancer Institute

INTRODUCTION

- The Advanced Systemic Mastocytosis Symptom Assessment Form (AdvSM-SAF) was developed to assess the signs and symptoms experienced by subjects with advanced systemic mastocytosis (AdvSM),[1,2] a rare condition characterized by neoplastic mast cell infiltration of tissues and shortened survival.[3]
- As a patient-reported outcome (PRO) questionnaire intended for use in clinical trials to evaluate treatment efficacy hypotheses, the tool was developed in a manner consistent with guidance provided by the Food and Drug Administration (FDA)[4,5] and best practices in questionnaire development.[6,7]

OBJECTIVE

- The objectives of this study are: (1) to present preliminary psychometric performance results related to the scores produced by the AdvSM-SAF and (2) to provide evidence to inform conclusions regarding how AdvSM-SAF scores may be interpreted in future studies.

METHODS

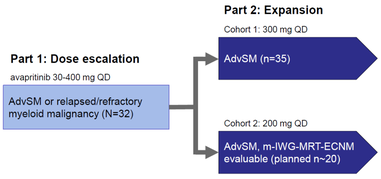
Study design

- The AdvSM-SAF was administered daily using an electronic PRO device in the expansion stage of BLU-285-2101 (NCT02561988), an open-label Phase I trial designed to evaluate the effect of the KIT inhibitor avapritinib in subjects with AdvSM **(Figure 1)**.
- The AdvSM-SAF is a 10-item diary that assesses eight symptoms of AdvSM including abdominal pain, nausea, vomiting, diarrhea, spots, itching, flushing, and fatigue. Using a 24-hour recall period, eight items assess symptom severity with an 11-point numerical rating scale, where 0=No [symptom] and 10=Worst imaginable [symptom], and two items (vomiting and diarrhea) assess symptom frequency by asking subjects to enter a discrete numerical value.
- The AdvSM-SAF is scored as a seven-day average (i.e., a "weekly score") and only derived if at least four completed daily scores are available within the pre-specified seven-day period. AdvSM-SAF severity item scores are summed to create a Total Symptom Score (TSS; range 0-80), Gastrointestinal Symptom Score (GSS; range 0-40), and Skin Symptom Score (SSS; range 0-30). All contributing items need to be completed to calculate a daily score.
- Psychometric evaluation of the AdvSM-SAF is supported by other clinical and PRO assessments in BLU-285-2101 **(Table 1)**.
- Two analysis populations were used for the study:
  - Cross-sectional analysis population (CS-AP): All subjects with AdvSM-SAF scores at Baseline (C1D-7 to C1-1; i.e., 1 to 7 days before Cycle 1) and at least one follow-up visit at C3D1, C7D1, or C11D1 (n=31).
  - Test-retest analysis population (TRT-AP): Subjects who exhibited no change in ECOG-PS score from Baseline to C1D8 (n=21).

Psychometric analyses

- **Internal consistency reliability** reflects the extent to which individual items from a scale are measuring the same general concept[8] and is investigated by calculating Cronbach's alpha coefficient (α, range 0 to 1).[9,10] Alpha was calculated for the weekly TSS, GSS, and SSS using the CS-AP at Baseline and C3D1, and again with each individual item within a domain removed.
- **Test-retest reliability** assesses if items in an instrument produce stable, reliable scores under similar conditions, at different assessment points during which no change (or minimal change) in the patient's condition is expected to occur.[11] Test-retest reliability was evaluated among the TRT-AP using AdvSM-SAF weekly scores.
- **Construct-related validity** evaluates the associations between concepts of a specified assessment and of other assessments (i.e., reasonably strong associations should exist between related concepts and low associations between unrelated concepts).[4] The construct-related validity for the weekly AdvSM-SAF was evaluated by generating correlation coefficients between its scores and other clinical and PRO assessments at Baseline and C3D1.
- **Known-groups methods** characterize the degree to which a PRO questionnaire generates scores capable of distinguishing among subject groups hypothesized to be clinically distinct.[4] This analysis was conducted using the PGIS and ECOG-PS to categorize subjects into "known groups" at Baseline, and AdvSM-SAF weekly scores were described across patient severity groups.
- **Sensitivity-to-change analyses** focus on the evaluation of change scores in a target assessment over time to show that improvements (or worsening) seen in those scores correspond to improvements (or worsening) in other areas expected to change.[12] This was addressed by examining the mean change and associated effect size of weekly AdvSM-SAF scores, as well as the correlations between the AdvSM-SAF change scores and change scores of other measures.

Figure 1. BLU-285-2101 study design



RESULTS

Study sample

- In the total CS-AP of 31 patients, mean age was 63.7 years (SD=10.3), 51.6% were female, and 90.3% were white.
- Baseline ECOG-PS ranged from 0 to 3, with 35.5% ECOG-PS 1.

Psychometric properties

- **Item distribution**
  - The mean severity item scores ranged from 0.8 to 2.6, except fatigue (mean=5.8). The mean TSS, GSS, and SSS were 18.9, 7.3, and 5.5, respectively, at Baseline.
  - The range of scores was restricted (i.e., the full range of response options was not used) for most of the items, especially for the itching (0-6.1), vomit (0-5.8), and diarrhea (0-5.3) severity items.
- **Internal consistency reliability (Table 2)**
  - The weekly GSS, SSS, and TSS all met criteria pre-specified for internal consistency (α>0.70).
  - Removal of any individual item within a domain did not improve the internal consistency of TSS.
- **Test-retest reliability (Table 3)**
  - The weekly item, domain, and total AdvSM-SAF scores were all reliable (>0.7), except the vomiting frequency item.
  - The gastrointestinal (GI) frequency items (not included in domain or total scores) were less reliable than the severity items.
- **Construct-related validity (Table 4a)**
  - Weekly AdvSM-SAF scores more strongly (r≥0.60) correlated to EORTC QLQ-C30 symptom items than to more distal concepts.
- **Known-groups analysis**
  - Weekly AdvSM-SAF scores were able to distinguish among clinically unique groups specified by PGIS **(Table 5)** and ECOG-PS at Baseline (p<0.05; exclusive of weekly SSS).
- **Sensitivity to change (Table 4b)**
  - The GSS change score was moderately to strongly correlated with the change scores in PGIS, serum tryptase, and the EORTC QLQ-C30 items and domains (r=0.240-0.697) and weakly correlated with the change score of ECOG (r=0.168). The greatest correlations for the GSS were observed with the EORTC QLQ-C30 GI items.
  - The SSS change score was moderately correlated with the change scores in ECOG, EORTC QLQ-C30 items of dyspnea, insomnia, and fatigue (r=0.341-0.433), and weakly correlated with the change scores of PGIS, serum tryptase, and most of the EORTC QLQ-C30 items and domains (r<0.3).
  - The TSS change score was moderately to strongly correlated with the change scores in PGIS, ECOG and the EORTC QLQ-C30 items and domains (r=0.306-0.812), with the exception of constipation (r=-0.191) and serum tryptase (r=0.266). The highest correlation for the TSS was with the EORTC QLQ-C30 fatigue item (0.812).

Table 1. Supplementary clinical and PRO assessments in BLU-285-2101

Assessment	Concepts assessed	Response scale	Recall period
European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire Core 30 (EORTC QLQ-C30)	Symptoms, impacts, and overall health	4-point scale for symptoms and impacts, 7-point scale for the overall health	Past week
Patient Global Impression of Severity (PGIS)	Symptom severity	5-point scale	At present
Eastern Cooperative Oncology Group Performance Status (ECOG-PS)	Level of functioning	6-point scale	At present

Table 2. Internal consistency reliability estimates (α) at Baseline (N=31) of weekly AdvSM-SAF domain and total symptom scores

Domain/Total score	Cronbach's Alpha
Weekly AdvSM-SAF Gastrointestinal Symptom Score	
Overall internal consistency:	**0.801**
Weekly AdvSM-SAF Weekly Skin Symptom Score	
Overall internal consistency:	**0.789**
Weekly AdvSM-SAF Weekly Total Symptom Score	
Overall internal consistency:	**0.844**
Score if variable deleted:	
Q1: Weekly Abdominal Pain Severity	0.820
Q2: Weekly Nausea Severity	0.800
Q3: Weekly Spots Severity	0.838
Q4: Weekly Itching Severity	0.827
Q5: Weekly Flushing Severity	0.837
Q6: Weekly Fatigue Severity	0.804
Q8: Weekly Vomit Severity	0.834
Q10: Weekly Diarrhea Severity	0.838

Table 3. Test-Retest Reliability for AdvSM-SAF between C1D1 and C1D8

Weekly AdvSM-SAF Items/Domains	n	Reliability
Gastrointestinal Symptom Score	**21**	**0.883 (0.716, 0.952)**
Skin Symptom Score	**20**	**0.955 (0.888, 0.982)**
Total Symptom Score	**20**	**0.945 (0.863, 0.978)**
Q1: Abdominal Pain Severity	21	0.858 (0.655, 0.942)
Q2: Nausea Severity	21	0.940 (0.852, 0.975)
Q3: Spots Severity	21	0.952 (0.881, 0.981)
Q4: Itching Severity	20	0.937 (0.843, 0.975)
Q5: Flushing Severity	20	0.956 (0.888, 0.983)
Q6: Fatigue Severity	21	0.957 (0.895, 0.982)
Q7: Vomit frequency count	21	0.020 (-1.376, 0.600)
Q8: Vomit Severity	21	0.878 (0.706, 0.950)
Q9: Diarrhea frequency count	21	0.728 (0.341, 0.889)
Q10: Diarrhea Severity	21	0.856 (0.651, 0.941)

Table 5. Known-groups analysis at Baseline for the weekly AdvSM-SAF domain and total symptom scores

AdvSM-SAF domain	Known group (PGIS)	n	Mean (SD)	Median	ANOVA p-value
Gastrointestinal Symptom Score (0-40)	Absent/ Minimal	8	4.1 (8.7)	0.1	0.016
	Moderate	9	4.1 (4.1)	2.6	
	Severe/Very Severe	9	13.3 (7.7)	16.2	
Skin Symptom Score (0-30)	Absent/ Minimal	8	4.2 (5.7)	2.2	0.688
	Moderate	8	5.1 (6.1)	3.0	
	Severe/Very Severe	9	6.9 (7.4)	4.1	
Total Symptom Score (0-80)	Absent/ Minimal	8	11.8 (16.4)	6.2	0.035
	Moderate	8	15.2 (11.4)	9.9	
	Severe/Very Severe	9	28.3 (10.5)	29.0	

Table 4. Spearman correlation coefficients between concurrent measures and (a) AdvSM-SAF weekly domain/total scores at Baseline; and (b) AdvSM-SAF weekly domain/total change from Baseline to C3D1 scores (CS-AP, N=31)

≥0.6=green
<0.2=red

Concurrent scores	(a) Construct-related validity: Correlation of Baseline scores						(b) Sensitivity to change: Correlation of change from Baseline to C3D1		
	N	GSS	N	SSS	N	TSS	GSS	SSS	TSS
QLQ-C30: Global health status	27	-0.562	26	-0.206	26	-0.534	-0.455	-0.057	-0.391
QLQ-C30: Physical functioning	27	-0.384	26	-0.123	26	-0.422	-0.250	-0.222	-0.501
QLQ-C30: Role functioning	27	-0.426	26	-0.162	26	-0.413	-0.521	-0.102	-0.437
QLQ-C30: Emotional functioning	27	-0.641	26	-0.069	26	-0.566	-0.289	-0.281	-0.444
QLQ-C30: Cognitive functioning	27	-0.481	26	-0.216	26	-0.515	-0.468	-0.228	-0.507
QLQ-C30: Social functioning	27	-0.491	26	-0.217	26	-0.544	-0.458	-0.222	-0.366
QLQ-C30: Fatigue	27	0.591	26	0.366	26	0.710	0.619	0.341	0.812
QLQ-C30: Nausea and vomiting	27	0.850	26	0.441	26	0.811	0.697	0.023	0.511
QLQ-C30: Pain	27	0.801	26	0.412	26	0.752	0.615	0.052	0.415
QLQ-C30: Dyspnea	27	0.241	26	0.017	26	0.250	0.269	0.433	0.562
QLQ-C30: Insomnia	27	0.288	26	0.329	26	0.373	0.627	0.380	0.736
QLQ-C30: Appetite loss	27	0.420	26	0.161	26	0.438	0.240	0.083	0.350
QLQ-C30: Constipation	27	0.231	26	0.529	26	0.356	-0.349	0.137	-0.191
QLQ-C30: Diarrhea	27	0.608	26	0.194	26	0.465	0.483	0.042	0.380
QLQ-C30: Financial difficulties	27	0.503	26	0.142	26	0.387	0.477	0.146	0.481
PGIS	26	0.543	25	0.238	25	0.614	0.451	-0.105	0.306
ECOG-PS	28	0.418	27	0.251	27	0.579	0.168	0.370	0.472
Serum tryptase (ng/mL)	25	0.245	24	0.043	24	0.132	0.347	0.076	0.266

CONCLUSIONS

- The AdvSM-SAF produced reliable, construct-valid, and sensitive scores when administered in the target patient population.
- These results, along with its strong development history and evidence of content validity, support its future use in evaluating the signs and symptoms of AdvSM and assessing treatment benefit in AdvSM clinical studies.

FUNDING STATEMENT/DISCLOSURES

- This study was funded by Blueprint Medicines Corporation. Blueprint Medicines Corporation participated in the interpretation of data, review, and approval of the publication.
- BM and TG are employees of Blueprint Medicines Corporation. ALS, FT, SL, CY, and BP are employed by Adelphi Values, which received payment from Blueprint Medicines Corporation for participation in this research. MWD and EH received consulting fees from Blueprint Medicines Corporation. PB and MD served on paid advisory boards and consulted for Blueprint Medicines Corporation.

REFERENCES

1. Taylor F et al. Cognitive debriefing of the Advanced Systemic Mastocytosis Symptom Assessment Form (AdvSM-SAF). 22nd Annual International Meeting of the International Society for Pharmacoeconomics and Outcomes Research (ISPOR), 20-24 May 2017; Boston, MA USA.
2. Mazar I et al. Development and content validity of the advanced systemic mastocytosis symptom assessment form (AdvSM-SAF). ISPOR 19th Annual European Congress, 29 October-2 November 2016; Vienna, Austria.
3. Valent P et al. Mastocytosis: 2016 updated WHO classification and novel emerging treatment concepts. Blood. 2017;129(11):1420-1427.
4. US Department of Health and Human Services, FDA, Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research, Center for Devices and Radiological Health. Guidance for Industry: Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. 2009.
5. US FDA. Methods to Identify What is Important to Patients & Select, Develop or Modify Fit-for-Purpose Clinical Outcomes Assessments. 2018. https://www.fda.gov/downloads/Drugs/ NewsEvents/UCM620708.pdf. Accessed 11/29/2018.
6. Patrick DL et al. Content Validity-Establishing and Reporting the Evidence in Newly Developed Patient-Reported Outcomes (PRO) Instruments for Medical Product Evaluation: ISPOR PRO Good Research Practices Task Force Report: Part 1-Eliciting Concepts for a New PRO Instrument. Value in Health. 2011;14(8):967-977.
7. Patrick DL et al. Content Validity-Establishing and Reporting the Evidence in Newly Developed Patient-Reported Outcomes (PRO) Instruments for Medical Product Evaluation: ISPOR PRO Good Research Practices Task Force Report: Part 2-Assessing Respondent Understanding. Value in Health. 2011;14(8):978-988.
8. Thompson B. Understanding reliability and coefficient alpha, really. In: Thompson B, ed. Score reliability: contemporary thinking on reliability issues. Thousand Oaks, CA: Sage Publications, Inc.; 2003:3-21.
9. Bland JM, Altman DG. Cronbach's alpha. BMJ. 1997;314(7080):572.
10. Cronbach LJ. Coefficient Alpha and the Internal Structure of Tests. Psychometrika. 1951;16(3):297-334.
11. Guttman L. A basis for analyzing test-retest reliability. Psychometrika. 1945;10:255-282.
12. Stratford PW, Riddle DL. Assessing sensitivity to change: choosing the appropriate change coefficient. Health Qual Life Outcomes. 2005;3:23.

ADELPHI VALUES