Supplemental Data File 1

Evaluation of two Patient-reported Outcome Questionnaires in Ulcerative Colitis

Abdominal Pain and Bowel Urgency:

Psychometric Evaluation Report

Date: 22 December 2021

Version 1.0

Table of Contents

List of Tables

List of Figures

List of Abbreviations and Definitions of Terms

AP Diary Item	Abdominal Pain Diary Item
BU Diary Item	Bowel Urgency Diary Item
CI	confidence interval
CS-AP	cross-sectional analysis population
eCDF	empirical cumulative distribution function
EQ-5D-5L	Five-level EQ-5D
ES	effect size
FDA	Food and Drug Administration
IBDQ	Inflammatory Bowel Disease Questionnaire
ICC	intraclass correlation coefficient
ITT	intent-to-treat
MCID	minimal clinically important difference
MWPC	meaningful within-person change
PDF	probability density function
PGIC	Patient Global Impression of Change
PRO	patient-reported outcome
QD	once per day
ROC	receiver operating characteristic
SAP	statistical analysis plan
SD	standard deviation
SEM	standard error of measurement
SF-36v2®	36-Item Short Form Survey version 2
TRT-AP	test-retest analysis population
UC	ulcerative colitis
UC-SQ	Ulcerative Colitis Symptoms Questionnaire
UPA	upadacitinib
VAS	Visual Analogue Scale
WPAI:UC	Work Productivity and Activity Impairment Questionnaire for Ulcerative Colitis

Executive Summary

Introduction: Ulcerative colitis (UC) is a type of inflammatory bowel disease, a chronic condition that affects the digestive system,[1] which manifests as diffuse mucosal inflammation of the colon and/or rectum. UC is characterized by intermittent flares of symptoms including stool blood, loose and urgent bowel movements, abdominal pain, and fatigue.[2-5] There is no medical cure for UC besides colectomy in more severe cases; however, there are treatments such as aminosalicylic derivatives, immunosuppressants, corticosteroids, biological agents, and anti-tumor necrosis factor therapies. Despite the currently available biologic agents, only 17% to 39% of patients with moderately-to-severely active UC achieve clinical remission.[6-9] As a result, there is still a clear need for additional therapeutic options for patients with moderately-to-severely active UC who respond inadequately or are intolerant to conventional and biologic therapies.

AbbVie initiated a set of clinical trials (a Phase 2b/3 study [M14-234, substudies 1, 2, and 3], a Phase 3 study [M14-675], and a Phase 3 long-term extension study [M14-533]) to evaluate the safety, tolerability, and efficacy of upadacitinib (UPA), a selective and reversible Janus kinase 1 inhibitor designed for oral administration, in adolescents and adults aged 16 years and older with moderately-to-severely active UC. The primary endpoint for all trials is achievement of clinical remission per Adapted Mayo score. In addition, for M14-234 substudies 2 and 3 and M14-675, ranked secondary endpoints to evaluate treatment effects of UPA and to support product labeling using several patient-reported outcome assessments have been identified as important and relevant to the patient experience of UC. This includes the Abdominal Pain (AP) Diary Item (identified as an additional endpoint in the Phase 2b M14-234, Substudy 1 trial), which was utilized to assess abdominal pain associated with UC from the patient perspective, and the Bowel Urgency (BU) Diary Item (also identified as an additional endpoint in the Phase 2b M14-234, Substudy 1 trial), which was utilized to assess bowel urgency associated with UC from the patient perspective. Since the focus of the report and submission is on the pivotal trials (M14-234 substudies 2 and 3, and M14-675), additional details and analyses for M14-533 are not included. This document focuses specifically on the evaluation of the measurement properties and interpretation of the scores produced by the AP Diary Item and the BU Diary Item within the context of these trials.

Goal and objectives: The goals of this report are to (1) summarize the results of the psychometric analysis of the scores produced by the AP Diary Item and BU Diary Item among subjects with UC in Phase 2b/3 and Phase 3 data from M14-234, substudies 1, 2,

and 3, and M14-675, respectively and (2) provide an interpretation of the meaning of the AP Diary Item and BU Diary Item scores.

Methods: Data from two multicentered, randomized, double-blinded, placebo-controlled Phase 2b/3 and 3 clinical trials (M14-234, substudies 1, 2, and 3, and M14-675) were used to evaluate the safety and efficacy of UPA versus placebo during induction/maintenance therapy in subjects (adolescents ages 16–17 years and adults 18–75 years of age) with moderately to severely active UC. However, only data from Phase 2b M14-234, Substudy 1 (Phase 2b) was used to conduct a measurement-focused analysis of the AP and BU Diary Items. For the Phase 2b study, subjects were randomly assigned to one of five groups (UPA 45 mg once per day [QD], UPA 30 mg QD, UPA 15 mg QD, UPA 7.5 mg QD, or placebo) and the AP and BU Daily Diary items were analyzed at three timepoints: Baseline, Week 2, and Week 8. Data from M14-234, Substudy 1 (Phase 2b) were used to evaluate the psychometric performance and interpretation of scores for the AP Diary Item and BU Diary Item. Analyses were executed to evaluate the performance of scores produced by the AP Diary Item and BU Diary Item with respect to distribution, reliability, content validity, and sensitivity to change. Additional analyses were conducted to generate guidelines for interpreting group differences and within-person meaningful change in the AP Diary Item and BU Diary Item scores. For the psychometric analysis, missing data were not imputed for the AP Diary Item and BU Diary Item or any of the supplementary measures.

Results: All randomized subjects who received at least one dose of study drug during the eight-week induction period and who completed the item on the AP Diary Item and BU Diary Item at any of the psychometric analysis timepoints (e.g., Baseline, Week 2, and Week 8) and subjects who achieved a clinical response at Week 8 from M14-234 were included in the study. For M14-234 (N=248), subjects' ages ranged from 15 to 75 (mean=42.3, standard deviation=14.1), and more than half (60.1%) of the sample was female.

Item and total score properties: Quality of completion for the psychometric analysis population was high across the timepoints for the study (91.8%), with the number of participants with missing data ranging from 4 (1.6%) to 19 (8.2%) for both the AP and BU Diary Items. In general, respondents used the entire range of the response scale for the AP and BU Diary Items across assessment timepoints, and item scores trended toward improvement over time.

Score reliability: Reliability results indicate that the AP Diary Item score displayed acceptable test-retest reliability (intraclass correlation coefficient [ICC]=0.804). Reliability results indicate that the BU Diary Item score did not display acceptable test-retest reliability (ICC=0.325).

Construct-related validity: The construct-related validity for the AP and BU Diary Items was evaluated by generating convergent validity estimates, conducting a set of known-groups analyses, and evaluating sensitivity to change over time.

Convergent validity: Scores on the AP Diary Item indicated that it was moderately correlated with scores from the secondary assessments. Scores on the BU Diary Item indicated that it was moderately correlated with scores from the secondary assessments.

Known-groups analysis: Results generated from known-groups analysis also support the construct-related validity of AP and BU Diary Items scores. Specifically, known-groups analysis results for the AP Diary Item showed that participants who were in remission (based on Adapted Mayo Score) had significantly lower scores for the AP Diary Item at Weeks 2 and 8 compared to participants not in remission. Similarly, participants who reported more abdominal pain on both Item 3 of the Ulcerative Colitis Symptoms Questionnaire (UC-SQ) and Item 13 of the Inflammatory Bowel Disease Questionnaire (IBDQ) had significantly higher scores on the AP Diary Item at both timepoints. Furthermore, known-groups analysis results for the BU Diary Item showed that participants who were in remission (based on Adapted Mayo Score) had significantly lower scores for the BU Diary Item at Weeks 2 and 8 compared to participants not in remission. Similarly, participants who reported more bowel urgency on UC-SQ Item 17 had significantly higher scores on the BU Diary Item at both timepoints. Therefore, results presented demonstrate the frequency scores on the AP Diary Item and the BU Diary Item are able to distinguish between clinically distinct groups

Sensitivity to change: In the analysis of score sensitivity, moderate correlations were observed between the AP Diary Item with the conceptually-related supportive questionnaires ($r=0.55$, $p<0.001$). Moderate correlations were observed between the BU Diary Item change score and change scores on the conceptually-related supportive questionnaires ($r=0.53$, $p<0.001$).

Score interpretation: Meaningful within-person change (MWPC) and minimal clinically important difference estimates were generated to help describe the meaning of AP and BU

Diary Item scores when used for within-person change or for group means comparisons, respectively, via employment of distribution-based methods, anchor-based methods, receiver operating characteristic (ROC) curves, empirical cumulative distribution functions (eCDFs), and probability density functions. Specifically, for the AP Diary Item, anchor-based methods, values from eCDFs, and ROC analysis suggested estimates of MWPC of 1 point (decrease in frequency). Supportive distribution-based methods estimated that approximately a 0.3-point difference between group means would be meaningful. For the BU Diary Item, anchor-based methods, values from eCDFs and ROC analysis suggested estimates of MWPC of 1 point (decrease in days with bowel urgency). Supportive distribution-based methods estimated that approximately a 0.5-point difference between group means would be meaningful.

Conclusions: The present findings indicate that scores produced by the AP and BU Diary Items are construct-valid, capable of distinguishing between groups known to be clinically different, and sensitive to change over time. In addition, the AP Diary Item scores showed acceptable test-retest reliability. Thus, the overall pattern of psychometric results presented here support the AP and BU Diary Items as appropriate for use among individuals with moderately-to-severely active UC, and that inferences from the AP and BU Diary Items can be treated as valid and trustworthy. Additionally, this research provides insight into how to interpret changes in the AP and BU Diary Items scores, specifically that a 1-point decrease in the number of days experiencing abdominal pain or bowel urgency may be reflective of meaningful change in this population.

1 Introduction

Ulcerative colitis (UC) is a type of inflammatory bowel disease, a chronic condition that affects the digestive system,[11] which manifests as diffuse mucosal inflammation of the colon and/or rectum. UC is characterized by intermittent flares of symptoms including stool blood, loose and urgent bowel movements, abdominal pain, and fatigue.[2-5] There is no medical cure for UC besides colectomy in more severe cases; however, there are treatments such as aminosalicylic derivatives, immunosuppressants, corticosteroids, biological agents, and anti-tumor necrosis factor therapies. Despite the currently available biologic agents, only 17% to 39% of patients with moderately-to-severely active UC achieve clinical remission.[6-9] As a result, there is still a clear need for additional therapeutic options for patients with moderately-to-severely active UC who respond inadequately or are intolerant to conventional and biologic therapies.

To address the need for additional therapeutic options for patients with moderately-to-severely active UC, AbbVie is investigating the safety, tolerability, and efficacy of upadacitinib (UPA), a selective and reversible Janus kinase 1 inhibitor designed for oral administration, in adolescents and adults aged 16 years and older with moderately-to-severely active UC (clinical trials: a Phase 2b/3 study [M14-234, substudies 1, 2, and 3], a Phase 3 study [M14-675], and a Phase 3 long-term extension study [M14-533]). The primary endpoint for all trials is achievement of clinical remission per Adapted Mayo score. In addition, for M14-234 substudies 2 and 3 and M14-675, ranked secondary endpoints to evaluate treatment effects of UPA and to support product labeling using several patient-reported outcome assessments have been identified as important and relevant to the patient experience of UC. This includes the Abdominal Pain (AP) Diary Item (identified as an additional endpoint in the Phase 2b M14-234, Substudy 1 trial), which was utilized to assess abdominal pain associated with UC from the patient perspective, and the Bowel Urgency (BU) Diary Item (also identified as an additional endpoint in the Phase 2b M14-234, Substudy 1 trial), which was utilized to assess bowel urgency associated with UC from the patient perspective. Since the focus of this report is on the pivotal trials (M14-234 substudies 2 and 3, and M14-675), additional details and analyses for M14-675 and M14-533 are not included. This document focuses specifically on the evaluation of the measurement properties and interpretation of the scores produced by the AP Diary Item and the BU Diary Item within the context of these trials.

The AP Diary Item is a single-item questionnaire that assesses the severity of abdominal pain using a 24-hour recall period and a four-point scale (0=None, 1=Mild, 2=Moderate,

3=Severe). The AP Diary Item has a maximum score of 3 and a minimum score of 0, where a higher score equates to greater severity.

The BU Diary Item is a single-item questionnaire that assesses whether bowel urgency has occurred during the 24-hour recall period and uses a dichotomous scale (Yes or No) to identify whether the respondent has experienced bowel urgency. To aid respondents in selecting a response, a brief definition of bowel urgency is provided.

The goal of the current report is to document that the scores from the diary items completed by participants in AbbVie's Phase 2b/3 trials of UPA for moderately-to-severely active UC are reliable, valid, and interpretable. Therefore, in addition to evaluating safety and efficacy, the M14-234, Substudy 1 clinical trial data were used to conduct an evaluation of the psychometric properties of the AP and BU Diary Items, as well as analyses to evaluate meaningful within-person change (MWPC) that facilitated a deeper understanding of the clinical meaning of observed changes in those scores over time.

2 Goals and Objectives

The goal of the present research was to evaluate the psychometric properties and score interpretation of the AP and BU Diary Items in accordance with good psychometric practice.[10-12] To accomplish this goal and using the AbbVie's Phase 2b trial data from M14-234 sub-study 1, objectives were specified that included evaluating and presenting results for:

- Score variability, distribution, and missingness of scores using descriptive analyses;
- Score reliability (e.g., test-retest reliability);
- Score construct-related validity (e.g., convergent validity and known-groups methods);
- Sensitivity-to-change analyses; and
- Score interpretability, generating MWPC thresholds that are clinically meaningful (i.e., using anchor-based methods, empirical cumulative distribution functions [eCDF], probability density function [PDF] and receiver operator characteristic [ROC] curves) and evaluating supportive between-groups differences that are meaningful (i.e., using distribution-based methods).

3 Trial Methodology

This section outlines the study procedures, population, and questionnaires that were administered in the Phase 2b study (M14-234, Substudy 1) that generated the data for the present analyses.

3.1 Overview of Studies

AbbVie initiated a set of clinical trials (a Phase 2b/3 study [M14-234, substudies 1, 2, and 3], a Phase 3 study [M14-675], and a Phase 3 long-term extension study [M14-533]) to evaluate the safety, tolerability, and efficacy of UPA versus placebo in patients with moderately to severely active UC; however, only data from the M14-234, Substudy 1 clinical trial data were used to conduct a measurement-focused analysis of the AP and BU Diary Items. The M14-234 trial was a multicenter, randomized, double-blinded, placebo-controlled Phase 2b study, in which patients were randomly assigned to one of five groups: UPA 45 mg once per day (QD), UPA 30 mg QD, UPA 15 mg QD, UPA 7.5 mg QD, or placebo. Scores from the AP and BU Diary Items were evaluated across five timepoints: Baseline, Week 2, Week 4, Week 6, and Week 8; however, only the scores for Baseline, Week 2, and Week 8 were utilized for these analyses. Please see Table 1 below for a brief overview of the study.

Table 1. Overview of Study

Study Number	Study Design	Number of Subjects	Primary Efficacy Assessments	Ranked Secondary Assessments	Supportive Assessments
M14-234, Substudy 1	Multicenter, randomized, double-blind, placebo-controlled induction study	250	• Adapted Mayo scoring system for Assessment of UC activity	• FACIT-Fatigue • AP Diary Item • BU Diary Item	• IBDQ • WPAI:UC • EQ-5D-5L • SF-36v2® • PGIC • Adapted Mayo scoring system for Assessment of UC activity • UC-SQ

Abbreviations: AP=Abdominal Pain; BU=Bowel Urgency; UC= Ulcerative Colitis; EQ-5D-5L=Five-level EQ-5D; FACIT-Fatigue=Functional Assessment of Chronic Illness Therapy – Fatigue; IBDQ=Inflammatory Bowel Disease Questionnaire; PGIC=Patient Global Impression of Change; SF-36v2®=36-Item Short Form Survey version 2; UC-SQ=Ulcerative Colitis Symptom Questionnaire; WPAI:CD=Work Productivity and Activity Impairment Questionnaire for Ulcerative Colitis

Note: Study number is the study protocol name.

3.1.1 Schedule of Events

For the Phase 2b induction portion of the UPA clinical trial (M14-234, Substudy 1), patients completed a daily diary from Baseline to Week 8, and site visits at Baseline and Weeks 2, 4, 6, and 8. The schedule of the primary, secondary, and supportive assessments across in the study are summarized in Table 2. For the psychometric evaluation of the AP and BU Diary Items (from the daily diary), scores associated with the site visit timepoints were used.

Table 2. Schedule of Assessments

Assessments	Timepoints				
	Phase II B Induction Substudy 1 M14-234				
		8-Week Double-blind Induction			
	Baseline	Week 2	Week 4	Week 6	Week 8
Target Outcomes					
FACIT-Fatigue	x	x			x
AP Diary Item*	x	x	x	x	x
BU Diary Item*	x	x	x	x	x
Supportive Outcomes					
Adapted Mayo	x	x	x	x	x
PGIC		x			x
IBDQ	x	x			x
EQ-5D-5L	x	x			x
SF-36v2®	x	x			x
UC-SQ	x	x			x
WPAI:UC	x	x			x

Abbreviations: AP=Abdominal Pain; BU=Bowel Urgency; UC= Ulcerative Colitis; EQ-5D-5L=Five-level EQ-5D; FACIT-Fatigue=Functional Assessment of Chronic Illness Therapy – Fatigue; IBDQ=Inflammatory Bowel Disease Questionnaire; PGIC=Patient Global Impression of Change; SF-36v2®=36-Item Short Form Survey version 2; UC-SQ=Ulcerative Colitis Symptom Questionnaire; WPAI:CD=Work Productivity and Activity Impairment Questionnaire for Ulcerative Colitis

* The AP and BU Diary Items were collected daily

3.2 Definition of Trial Analysis Populations

Section 3.2.1 below describes the trial protocol and briefly summarizes how the intent-to-treat (ITT) analysis populations were defined in the trial.

3.2.1 Analysis Populations: Induction Study (M14-234)

The ITT is defined as all randomized subjects who have received at least one dose of the study drug or placebo at the time of dose selection. ITT sets were as follows:

- ITTIA: All randomized participants who received at least one dose of study drug from Substudy 1.
- ITT1B: All the additional participants who were randomized to UPA 30 mg and 45 mg groups during the dose-selection period.

3.3 Study Assessments

The primary target assessments for psychometric evaluation included the AP and BU Diary Items. The secondary target assessments (covariates) for psychometric evaluation were the Patient Global Impression of Change (PGIC), Inflammatory Bowel Disease Questionnaire (IBDQ), Five-level EQ-5D (EQ-5D-5L), 36-Item Short-Form Survey version 2 (SF-36v2®), Adapted Mayo score, Ulcerative Colitis Symptom Questionnaire (UC-SQ), and Work Productivity and Activity Impairment: Ulcerative Colitis (WPAI:UC). The primary assessments and each of the secondary assessments are summarized below. A schedule of assessments for all of the specified instruments is presented in Section 3.1.1 (Table 2).

3.3.1 Primary Assessments

The AP and BU Diary Items were the primary assessment for psychometric evaluation (i.e., the instruments for which psychometric properties and score interpretations were evaluated).

3.3.1.1 Abdominal Pain (AP) Diary Item

The AP Diary Item, administered daily via an electronic diary, is a single-item measure designed to assess the severity of abdominal pain using a 24-hour recall period and a four-point scale (0=None to 3=Severe). For the purpose of these analyses, the AP Diary Item was scored utilizing the diary entries from the most recent three consecutive days prior to each study visit (e.g., Day 11 through Day 13 for the visit at Week 2 [Day 14]) to

calculate two scores (AP severity and AP frequency). Item severity for a particular visit was computed as the mean of the item scores from the most recent three consecutive days prior to the visit.

As a daily diary, this measure was also used to evaluate symptom frequency when coded as the daily presence/absence of the symptom over a period of time. For computing symptom frequency, daily AP scores were dichotomized such that 0 represented no pain and 1 denoted mild or more severe pain. The sum of these dichotomized values denoted the frequency of abdominal pain that the patient had experienced during the most recent three consecutive days prior to each study visit. As such, the AP frequency score ranged between 0 and 3.

For both AP severity and AP frequency scores, if data were not available for three consecutive days, the average (for AP severity) and sum (for AP frequency) from the most recent three non-consecutive days in the last 10 days were utilized. If AP scores for fewer than three non-consecutive days in the 10 days prior to a visit were available, then the AP severity and AP frequency scores was set to missing for that visit.

Lastly, a dichotomous flag was created that indicates whether a patient's mean AP severity score was derived from three consecutive days prior to each study visit. Specifically,

- A flag of 0 indicated that three non-consecutive days were used, and
- A flag of 1 indicated that three consecutive days were used.

The screenshot for the AP Diary Item that was used in all trials is presented in Figure 1 below.

3.3.1.2 Bowel Urgency (BU) Diary Item

The BU Diary Item, administered daily via an electronic diary, is a single-item measure designed to assess whether an individual had experienced bowel urgency (i.e., "need for a bowel movement"). This measure uses a 24-hour recall period with response options consisting of "Yes" and "No" to identify if the respondent has experienced BU.

For the purpose of this study, the BU Diary Item was scored by summing the diary entries from the most recent three consecutive days prior to each study visit. Scores were coded as their numeric values (0 or 1) and summed. As such, a patient's BU score for a particular visit ranged between 0 and 3. If data were not available for three consecutive

days, the sum of the entries from the most recent three non-consecutive days in the last 10 days was utilized. If BU scores for fewer than three non-consecutive days in the 10 days prior to a visit were available, then the BU score was set to missing for that visit.

A dichotomous flag was created that indicated whether a patient's mean BU score is derived from three consecutive days prior to each study visit. Specifically,

- A flag of 0 indicated that three non-consecutive days were used, and
- A flag of 1 indicated that three consecutive days were used.

The screenshot for the BU Diary Item that was used in all trials is presented in Figure 2 below.

Figure 1. AP Diary Item

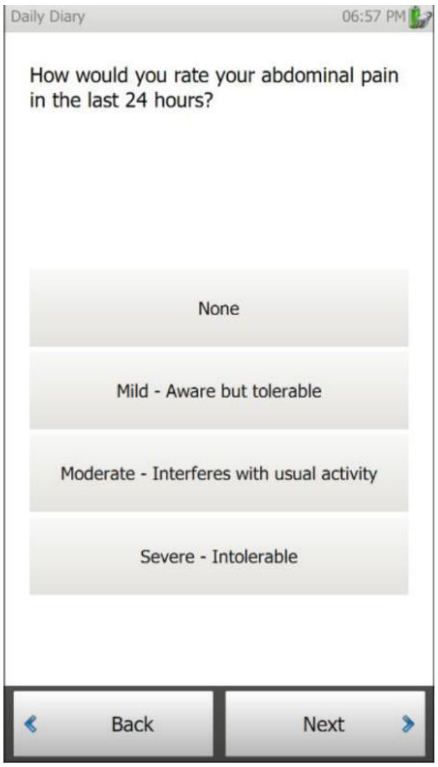
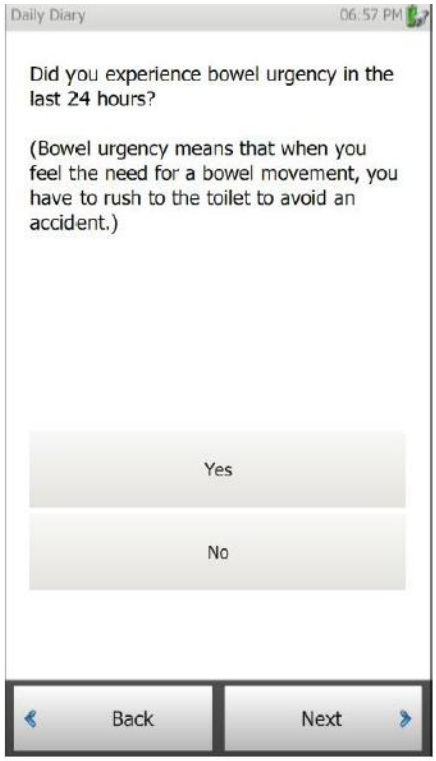


Figure 2. BU Diary Item



3.3.2 Secondary Assessments

The patient-reported outcome (PRO) assessments described in this section were used to support the psychometric and score interpretability evaluation of the primary assessments, the AP and BU Dairy Items.

3.3.2.1 Baseline Demographics

Baseline demographic information and clinical assessment, indicating disease severity, were summarized, including age, gender, duration of disease, and the previous use of immunosuppressants/biologics.

3.3.2.2 Ulcerative Colitis Symptom Questionnaire (UC-SQ)

The UC-SQ is a UC-specific PRO questionnaire consisting of 17 items that assesses UC-related gastrointestinal symptoms (e.g., frequent bowel movements, abdominal pain, cramping) and non-gastrointestinal symptoms (e.g., joint pain and sleep difficulties). Using a recall period of the last week (past seven days), respondents rate the severity of symptoms.

Overall symptom scores are calculated by summing the rating for each item. Higher scores indicate increased severity.

The UC-SQ was electronically administered at Baseline, Week 2, and Week 8.

3.3.2.3 Patient Global Impression of Change (PGIC)

The PGIC is a one-item PRO questionnaire that asks patients to rate the overall change in their UC symptoms since before treatment began on a seven-point verbal rating scale ranging from 1 ("Very much improved") to 7 ("Very much worse"); higher scores indicate worsening disease. Missing PGIC data were not imputed.

The PGIC was electronically administered at Week 2 and Week 8.

3.3.2.4 Adapted Mayo Scoring System for Assessment of Ulcerative Colitis Activity

The Mayo Scoring System for Assessment of Ulcerative Colitis Activity is a clinician-rated measure that consists of four subscores (stool frequency, rectal bleeding, endoscopy results, and clinician's global assessment). This adaptation excludes the clinician global assessment. Each subscore has a range of 0 to 3, with higher scores indicating greater disease severity. The total score ranges between 0 and 9.

The Adapted Mayo score was the primary endpoint for M14-234, Substudy 1 and was used to justify study sample sizes by determining the expected proportion of subjects who achieved clinical remission by Week 8. A subject was defined as having achieved clinical remission if all of the following criteria were satisfied:

1. A stool frequency subscore of 0 or 1 and not greater than the Baseline score;
2. A rectal bleeding subscore of 0; and
3. An endoscopic subscore of 0 or 1.

The Adapted Mayo score was clinically determined biweekly between Baseline and Week 8.

3.3.2.5 Inflammatory Bowel Disease Questionnaire (IBDQ)

Developed as a PRO questionnaire for patients with inflammatory bowel disease, the IBDQ[13] consists of 32 items, and every item has a score range of 1–7, with higher item scores indicating better quality of life. The IBDQ assesses quality of life across four

dimensions: bowel symptoms (10 items, including loose stools, AP), systemic symptoms (five items, including fatigue, altered sleep pattern), social function (five items, including work attendance, need to cancel social events), and emotional function (12 items, including anger, depression, irritability). The analyses for this study focused on the IBDQ total score, which ranges from 32 to 224, with a higher score indicating less disease severity and impact. Each subscale can be calculated with total scores ranging from 10 to 70, 5 to 35, 5 to 35, and 12 to 84, respectively.

The IBDQ was electronically administered at Baseline, Week 2, and Week 8.

3.3.2.6 Five-level EQ-5D (EQ-5D-5L)

The EQ-5D-5L is a generic, non-disease specific instrument for assessing health-related quality of life in clinical and economic evaluations of health care and in population health surveys.[14] It includes the EQ-5D descriptive system and the EQ Visual Analogue Scale (VAS). The descriptive system comprises five dimensions (i.e., mobility, self-care, usual activities, pain/discomfort, and anxiety/depression), and each dimension has five levels with higher scores indicating a more severe condition. Using a recall period of "Today," respondents rate these dimensions by selecting one of five response choices. The EQ VAS reflects a patient's self-evaluated health on a vertical VAS that ranges from 0 (i.e., the worst health you can imagine) to 100 (i.e., the best health you can imagine). The number that a respondent chose on the scale was the VAS score. Analyses focused on item responses to the Usual Activities and Pain/Discomfort dimensions, as well as on the VAS.

The EQ-5D-5L was electronically administered at Baseline, Week 2, and Week 8.

3.3.2.7 36-Item Short Form Survey Version 2 (SF-36v2®)

The SF-36v2® is a 36-item questionnaire that assesses eight health concepts: physical functioning, bodily pain, role limitations due to physical health problems, role limitations due to personal or emotional problems, emotional well-being, social functioning, energy/fatigue, and general health perceptions.[15] These eight health concepts can be aggregated into two summary measures: the Physical and Mental Component Summary scores. Details regarding the scoring algorithm are presented in the SF-36v2® manual.[16]

3.3.2.8 Work Productivity and Activity Impairment Questionnaire for Ulcerative Colitis (WPAI:UC)

Using six items, the WPAI:UC assesses the impact of UC on four domains:

- **Absenteeism** (work time missed) is measured as the number of hours missed from work in the past seven days due to condition-related problems. Scores are expressed as impairment percentages, adjusting for hours actually worked according to the WPAI scoring algorithm.
- **Presenteeism** (impairment at work/reduced on-the-job effectiveness) is measured as the impact of the condition on productivity while at work (e.g., reduced amount or kind of work, or not as focused as usual). Item 1 is scored as either yes or no, Items 2–4 have an open field response, and Item 5 is recorded on a 0–10 numeric rating scale where 0=no effect of UC on work and 10=severe impact of UC while at work.
- **Productivity loss** (overall work impairment) is measured as the sum of hours missed due to condition (i.e., absenteeism) and number of hours worked with impairment (i.e., product of number of hours worked and presenteeism).
- **Activity impairment** (i.e., activities other than paid work, such as work around house, cleaning, shopping, traveling, studying) is recorded and scored in the same way as presenteeism. Higher numbers indicate greater impairment and less productivity.

Responses from the WPAI items are scored as follows:

- Percentage of work time missed due to UC (Absenteeism): $\frac{\text{Item 2}}{\text{Item 2} + \text{Item 3} + \text{Item 4}}$;
- Percentage of impairment while working due to UC (Presenteeism): Item 5/10;
- Percentage of overall work impairment due to UC (Productivity loss): (Item 2/[Item 2 + Item 4] + [(1-[Item 2]/(Item 2 + Item 4)] x [Item 5/10]); and
- Percentage of activity impairment due to UC (if the answer to the first question reflects that the respondent is currently employed): Item 6/10.

4 Analysis and Results

A summary of the data analytic methods, interpretation strategy, and results is presented in this section. All statistical analyses were performed with SAS® software, version 9.4 or higher (SAS Institute, Cary NC).

Before discussing measurement property and score interpretability results, Sections 4.1, 4.2, 4.3, and 4.4 describe the general data analytic guidelines, handling of missing data, analysis populations specific to the psychometric and score interpretation analyses, and sample demographics of those included in the current analyses, respectively.

4.1 General Guidelines

The following are general guidelines that were applied to proposed statistical or psychometric analyses:

- All analyses were performed using the appropriate analysis population as specified in Section 3.2.
- Data from different timepoints were used to evaluate the measurement properties of scores produced by the AP and BU Diary Items (Baseline and Weeks 2 and 8).
- Continuous variables (e.g., age) are described by number of observations (frequency), mean, standard deviation (SD), median, extreme values (minimum and maximum values), and number of missing values.
- Categorical variables (e.g., gender) are described by frequency and percentage of each response option, with missing data being included in the calculation of percentage.
- The emphasis in a psychometric evaluation study is not on statistical significance testing, but rather on evaluating the magnitude of relationships between variables and overall result patterns. Accordingly, no adjustments were used for multiplicity of tests, though when use of a specific statistical significance test was needed, the threshold for statistical testing was $p<0.05$ for each test.
- Where applicable, proposed specific guidelines for evaluating the results of certain psychometric tests have been noted.
- Timepoints hereafter are simplified and referred to by week number only (i.e., Week 2 and Week 8), except for “Baseline.”

4.2 Handling of Missing Data

If diary entries from the three consecutive days prior to the visit were not available, the three most recent consecutive days in the last 10 days were utilized. If data are not available for three consecutive days, the average of the entries from the most recent three non-consecutive days in the last 10 days were utilized. If fewer than three days of diary data were available, the AP and BU Diary Items were considered missing.

4.3 Analysis Population

Depending on the analysis objective, two psychometric analysis populations originating from the participants who have completed the AP and BU Diary Items at least once in the study M14-234 sub-study 1 were identified: the cross-sectional analysis population (CS-AP) and test-retest analysis population (TRT-AP). It is important to note that since the target measures were implemented into three clinical trials, the timepoints selected for cross-sectional analysis and test-retest analysis were different due to the schedule of assessments of each study. The psychometric analysis populations were generated from the ITT (ITT1A) subjects. See the summary of the CS-AP and TRT-AP in Table 3 below.

Table 3. Psychometric Analysis Populations

Cross-sectional Analysis Population (CS-AP)	
All ITT subjects in the study who have completed at least one item on AP or BU Diary Items at any of the following timepoints: • M14-234: Baseline and Weeks 2, 4, and 8	
Test-retest Analysis Population (TRT-AP)	
Abdominal pain TRT-AP2	• Patients with non-missing mean AP item scores both at Screening and Baseline (assumed *a priori* to be stable).
Bowel urgency TRT-AP3	• Patients with non-missing mean BU item scores both at Screening and Baseline (assumed *a priori* to be stable).

Abbreviations: AP=Abdominal Pain; BU=Bowel Urgency; CS-AP=Cross-sectional analysis population; ITT=intent-to-treat; TRT-AP=Test-retest analysis population

4.4 Population Descriptive Analyses

Descriptive statistics for age, gender, race, and ethnicity were generated to characterize the study sample (M14-234 sub-study 1) upon entry into the study. Specifically, Table 4 summarizes the demographics of the sample; age was presented as continuous variables (frequency, mean, SD, median, and extreme values [minimum and maximum values]). Gender, race, and ethnicity were summarized as categorical variables (frequency and percent).

There were no missing demographic data for this study. There was a total of 248 participants (N=248) whose ages ranged from 18 to 75 (mean=42.3, SD=14.1), and more than a half (60.1%) of the sample was female. The White (73.8%) and Not Hispanic or Latino (95.2%) response options constituted the majority for the racial and ethnic categorization, respectively.

Table 4. Subject Demographic and Health Information (M14-234 Substudy 1 Cross-sectional Analysis Population, N=248)

Characteristic	Overall
Age (years)	
N	248
Mean (SD)	42.3 (14.1)
Median	41
Min–max	18.0–75.0
Missing	0
Gender (n, %)	
Male	99 (39.9%)
Female	149 (60.1%)
Missing	0 (0.0%)
Race	
White	183 (73.8%)
Black	8(3.2%)
Asian	52 (21.0%)
American Indian or Alaska Native	1 (0.4%)
Multiple	4 (1.6%)
Missing	0 (0.0%)
Ethnicity	
Hispanic or Latino	12(4.8%)
Not Hispanic or Latino	236 (95.2%)
Missing	0 (0.0%)

Abbreviations: max=maximum; min=minimum; SD=standard deviation

5 Assessment of Measurement Properties

Consistent with regulatory guidance recommendations, measurement properties[10,17] are specified as reliability, construct-related validity, and ability to detect change. In this section, the measurement properties associated with the scores produced in AbbVie's Phase 2b study are presented.

5.1 Quality of Completion and Distribution of Item Scores

This subsection covers the quality of completion and distribution of both AP and BU Diary Item scores associated with the Baseline, Week 2, and Week 8 timepoints. Quality of completion (e.g., missing data at each timepoint), is an indicator of compliance with the daily diary. Distribution of scores help identify potential floor or ceiling effects for the derived scores for each diary item.

Results for the AP Diary Item are presented in Table 5 . Quality of completion was high, with very few missing scores at Baseline and Week 2 and less than 10% missing at Week 8. Scores ranged from 0 to 3 across the timepoints, and no floor or ceiling effects were noted at Baseline (however, floor effects were seen at Weeks 2 and 8, which likely represents the improvement of abdominal pain during the post-treatment timepoints). The distribution of scores on the AP Diary Item associated with the Baseline, Week 2, and Week 8 visits are also presented in Table 5, and demonstrates that scores range from 0 to 3 at all timepoints.

Table 5. Quality of Completion and Item response distribution of Abdominal Pain Diary Item Frequency Score at Baseline, Week 2, and Week 8 for M14-234, Substudy 1

AP Diary Item Frequency score*	Missing n (%)	Floor† n (%)	Ceiling† n (%)	N	Mean (SD)	Median	Min–Max
Baseline (n=248)	5 (2.0%)	24 (9.7%)	5 (2.0%)	243	1.3 (0.7)	1.0	0.0–3.0
Week 2 (n=248)	4 (1.6%)	51 (20.6%)	--	244	0.9 (0.7)	1.0	0.0–2.7
Week 8 (n=231)	19 (8.2%)	89 (38.5%)	1 (0.4%)	212	0.6 (0.7)	0.3	0.0–3.0

Abbreviations: AP=Abdominal Pain; max=maximum; min=minimum; SD=standard deviation

* Frequency score was calculated as the sum of the most recent consecutive three-day period before each clinic visit. If diary entries from the three consecutive days prior to the visit were not available, the three most recent consecutive days in the last 10 days were utilized. If data were not available for three consecutive days, the average of the entries from the most recent 3 non-consecutive days in the last 10 days was utilized. Therefore, scores range from 0 to 3 but may not be whole integers.

† "Floor" was defined as a score of 0 and "Ceiling" was defined as a score of 3.

Results for the BU Diary Item are presented in Table 6. Quality of completion was high for the M14-234, Substudy 1, with very few missing scores at Baseline and Week 2, and less than 10% missing at Week 8. The distribution of scores on the BU Diary Item associated with the Baseline, Week 2, and Week 8 visits is also summarized and presented in Table 6, and demonstrates that scores range from 0–3 across all timepoints. Ceiling effects were noted at Baseline, but not at Week 8, which likely represents the improvement of bowel urgency during the post-treatment timepoints.

Table 6. Quality of Completion and Item Response Distribution Results of Bowel Urgency Diary Item Score at Baseline, Week 2, and Week 8 for M14-234, Substudy 1

BU Diary Item	Missing n (%)	Floor n (%)	Ceiling n (%)	N	Mean (SD)	Median	Min–Max
Baseline (n=248)	5 (2.0%)	20 (8.1%)	191 (77.0%)	243	2.6 (0.9)	3.0	0.0–3.0
Week 2 (n=248)	4 (1.6%)	42 (16.9%)	147 (59.3%)	244	2.2 (1.2)	3.0	0.0–3.0
Week 8 (n=231)	19 (8.2%)	78 (33.8%)	95 (41.1%)	212	1.6 (1.4)	2.0	0.0–3.0

Abbreviations: BU=Bowel Urgency; max=maximum; min=minimum; SD=standard deviation

* "Floor" was defined as a score of 0 and "Ceiling" was defined as a score of 3.

6 Test-retest Reliability

Reliability estimates characterize consistency and reproducibility of a particular set of scores produced by a questionnaire when administered to a particular target patient population and in a particular context of use.[18] Thus, reliability estimates can and will vary across administrations and, moreover, can be evaluated using various methods, depending on the nature of the assessment and context of administration.

Test-retest reliability measures the degree to which scores are similar at different points in time in a subset of "stable" patients. Test-retest reliability was investigated by calculating the intraclass correlation coefficient (ICC) and its 95% confidence interval (CI).

6.1 Test-retest Reliability for the AP Diary Item

The stability of the AP Diary Item frequency score was assessed in two samples of stable patients between the Baseline and Week 2 timepoints: (1) participants who chose "No Change" on the PGIC at the Week 2 timepoint and (2) participants who chose the same response on Item 3 of the UC-SQ ("During the past week, did you have abdominal pain?") at Baseline and Week 2. The AP Diary Item demonstrated acceptable test-retest reliability, as the ICCs exceed the threshold of 0.70, which was specified in the psychometric SAP as evidence of acceptable test-retest reliability for a scale.[19,20] The results are summarized in Table 7.

Table 7. Test-retest Reliability of Abdominal Pain Diary Item Frequency Score Between Baseline and Week 2 for M14-234, Substudy 1

Score	n	Test-retest reliability (ICC)*	95% CI
PGIC stable†	42	0.804	0.660-0.890
UC-SQ Item 3 stable‡	37	0.850	0.730-0.920

Abbreviations: CI=confidence interval; ICC=intra-class correlation; PGIC=Patient Global Impression of Change; UC-SQ=Ulcerative Colitis Symptoms Questionnaire

* The ICC was computed using the two-way mixed effects model without interaction (ICC[3A,1]).

† Subgroup of participants who selected "No change" on the PGIC at Week 2

‡ Subgroup of participants who selected the same response option on UC-SQ Item 3 (AP) at both Baseline and Week 2

6.2 Test-retest Reliability for the BU Diary Item

Similarly, the stability of the BU Diary Item score was assessed in two samples of stable patients between the Baseline and Week 2 timepoints: (1) participants who chose the "No change" on the PGIC at Week 2 and (2) participants who chose the same response on Item

17 of the UC-SQ ("During the past week, did you experience a sudden or intense need to have a bowel movement?") at Baseline and Week 2. The BU Diary Item did not demonstrate acceptable test-retest reliability, as the ICCs did not exceed the threshold of 0.70 (Table 8).

These low ICC results are difficult to interpret given the small sample size of the stable subgroups. Furthermore, the variability between frequency in the experience of the symptom (as reported in the qualitative interviews summarized in the Bowel Urgency Content Evaluation Report may also contribute to this result. Therefore, these reliability analyses should be considered in the context of this episodic, but distressing and burdensome,[21-23] symptom.

Table 8. Test-retest Reliability of Bowel Urgency Diary Item Score Between Baseline and Week 2 for M14-234, Substudy 1

Score	n	Test-retest reliability (ICC)*	95% CI
PGIC Stable (Week 2)†	42	0.325	0.022, 0.572
UC-SQ Item 17 Stable (Baseline and Week 2)‡	39	0.288	-0.018, 0.547

Abbreviations: CI=confidence interval; ICC=Intra-class correlation; PGIC=Patient Global Impression of Change; UC-SQ=Ulcerative Colitis Symptoms Questionnaire

* The ICC was computed using the two-way mixed effects model without interaction (ICC[3A,1]).

† Subgroup of participants who selected "No change" on the PGIC at Week 2

‡ Subgroup of participants who selected the same response option on the UC-SQ Item 17 (bowel urgency) at both Baseline and Week 2

7 Construct-related Validity

Construct-related validity is defined in the Food and Drug Administration (FDA) PRO Guidance as "evidence that relationships among items, domains, and concepts conform to a priori hypotheses concerning logical relationships that should exist with measures of related concepts or scores produced in similar or diverse patient groups."[10(p.11)] In other words, construct-related validity evaluates the associations between concepts of a specified questionnaire and of other questionnaires (i.e., reasonably strong associations between related concepts/questionnaires and low associations between unrelated concepts/ questionnaires). The construct-related validity for the AP Diary Item and the BU Diary Item frequency scores were evaluated by generating convergent validity estimates (correlations between the scores from the AP and BU Diary Items, and other assessments completed in the clinical trial), conducting a set of known-groups analyses (to evaluate how the scores on the diary items differ between clinically distinct subgroups), and evaluating sensitivity to change over time.

7.1 Convergent Validity

The construct validity of a score is evaluated by examining its relationships with other measures. Stronger correlations are expected with measures of similar constructs (i.e., convergent validity). A strong correlation was defined as ≥0.70 but ≤0.90, moderate correlation as ≥0.30 but <0.70, and a weak correlation as <0.30.[24] Convergent validity was measured by examining the correlations of the AP Diary Item and the BU Diary Item frequency score with other PRO questionnaires that are conceptually linked, including four disease-specific questionnaires (Adapted Mayo Score, IBDQ, UC-SQ, and WPAI:UC) and two generic questionnaires (EQ-5D-5L and SF-36v2®), at Baseline, Week 2, and Week 8. These results are presented in Tables 9 and 10.

Scores on the AP Diary Item were moderately correlated with scores from the disease-specific questionnaires (IBDQ, UC-SQ) at all timepoints, with the exception of the Adapted Mayo Score at Baseline and the WPAI:UC Work time missed score at Week 2, which were weakly correlated. In addition, the AP Diary Item was moderately correlated with the EQ-5D-5L at all timepoints for the Usual Activities and Pain/Discomfort items and the VAS, as well as with the two SF-36v2® component scores.

Table 9. Pearson Correlation Coefficients Between Abdominal Pain Diary Item Frequency Score and Other Assessments at Baseline, Week 2, and Week 8 for M14-234, Substudy 1

Assessment/Score	AP Diary Item Frequency Score			
	Hypothesized Relationship with the AP Diary Item	Baseline (N=248)	Week 2 (N=248)	Week 8 (N=231)
Disease-specific questionnaires				
Adapted Mayo Score*	+	0.265	N/A	0.388
IBDQ Bowel Symptom Domain	-	-0.640	-0.592	-0.644
IBDQ Systemic Symptoms Domain	-	-0.582	-0.582	-0.539
IBDQ Emotional Function Domain	-	-0.467	-0.506	-0.445
IBDQ Social Function Domain	-	-0.513	-0.473	-0.499
UC-SQ Total score	+	0.609	0.566	0.598
WPAI:UC Activity impairment	+	0.459	0.485	0.475
WPAI:UC Impairment while working	+	0.434	0.570	0.411
WPAI:UC Overall work impairment	+	0.477	0.538	0.439
WPAI:UC Work time missed	+	0.358	0.291	0.373
Generic questionnaires				
EQ-5D-5L Mobility	+	0.344	0.291	0.363
EQ-5D-5L Self care	+	0.213	0.141	0.324
EQ-5D-5L Usual activities	+	0.420	0.451	0.491
EQ-5D-5L Pain/Discomfort	++	0.622	0.593	0.625
EQ-5D-5L Anxiety/depression	+	0.306	0.281	0.346
EQ-5D-5L Visual Analogue Scale	-	-0.496	-0.488	-0.486
SF-36v2® Physical Component Summary	-	-0.573	-0.520	-0.582
SF-36v2® Mental Component Summary	-	-0.352	-0.331	-0.334

Abbreviations: AP=Abdominal Pain; EQ-5D-5L=Five-level EQ-5D; IBDQ=Inflammatory Bowel Disease Questionnaire; SF-36v2®=36-Item Short-Form Survey version 2; UC-SQ=Ulcerative Colitis Symptoms Questionnaire; WPAI:UC=Work Productivity and Activity Impairment Questionnaire for Ulcerative Colitis

Note: ++=moderate positive relationship ($0.30<|r|\leq 0.70$); +/-=weak positive/negative relationship ($0.00\leq|r|\leq 0.30$)

* Adapted Mayo Score calculated at Baseline and Week 8 only

Scores on the BU Diary Item were more strongly correlated with scores from the disease-specific questionnaires, the IBDQ and UC-SQ, compared to the generic questionnaires. At Baseline, correlations were weak with the exception moderate correlations with the UC-SQ Total score and two of the WPAI:UC scores (Impairment while working and Overall work impairment). This might be due to restricted range on the questionnaires due to the severity of participants' UC at the start of the clinical trial. At Week 8 correlations are moderate for all disease-specific scores, with the exception of a weak correlation with WPAI:UC Work time missed. In addition, Week 8 scores on the BU Diary Item are moderately corelated with the following score associated with the generic questionnaires:

EQ-5D-5L Usual Activities, Pain/Discomfort, and VAS for overall health, and the SF-36v2® Physical Component Summary score.

Table 10. Pearson Correlation Coefficients Between Bowel Urgency Diary Item Score and Other Assessments at Baseline, Week 2, and Week 8 for M14-234, Substudy 1

Assessment/Score	BU Diary Item Total Score			
	Hypothesized Relationship with BU Diary Item	Baseline (N=248)	Week 2 (N=248)	Week 8 (N=231)
Disease-Specific Questionnaires				
Adapted Mayo Score*	+	0.195	N/A	0.579
IBDQ Bowel Symptom Domain	-	-0.291	-0.459	-0.510
IBDQ Systemic Symptoms Domain	-	-0.286	-0.308	-0.428
IBDQ Emotional Function Domain	-	-0.224	-0.367	-0.353
IBDQ Social Function Domain	-	-0.252	-0.276	-0.396
UC-SQ Total score	+	0.462	0.456	0.515
WPAI:UC Activity impairment	+	0.208	0.280	0.434
WPAI:UC Impairment while working	+	0.340	0.376	0.507
WPAI:UC Overall work impairment	+	0.326	0.260	0.398
WPAI:UC Work Time missed	+	0.171	0.021	0.128
Generic Questionnaires				
EQ-5D-5L Mobility	+	0.076	0.165	0.180
EQ-5D-5L Self care	+	-0.027	0.095	0.183
EQ-5D-5L Usual activities	+	0.171	0.241	0.329
EQ-5D-5L Pain/Discomfort	+	0.228	0.289	0.380
EQ-5D-5L Anxiety/depression	+	0.076	0.109	0.278
EQ-5D-5L Visual Analogue Scale	-	-0.117	-0.217	-0.383
SF-36v2 Physical Component Summary	-	-0.275	-0.281	-0.405
SF-36v2 Mental Component Summary	-	-0.059	-0.130	-0.270

Abbreviations: BU=Bowel Urgency; EQ-5D-5L=Five-level EQ-5D; IBDQ=Inflammatory Bowel Disease Questionnaire; SF-36v2=36-Item Short Form Survey version 2; UC-SQ=Ulcerative Colitis Symptoms Questionnaire; WPAI:UC=Work Productivity and Activity Impairment Questionnaire: Ulcerative Colitis

Note: +/-=weak positive/negative relationship ($0.00 \leq |r| \leq 0.30$)

* Adapted Mayo Score was calculated at Baseline and Week 8 only

7.2 Known-groups Analysis

Known-groups methods characterize the degree to which a PRO questionnaire generates scores capable of distinguishing among groups hypothesized *a priori* to be clinically distinct.[10]

For the AP Diary Item total scores, known-groups analysis was conducted using (1) the Adapted Mayo Score, (2) UC-SQ Item 3 (abdominal pain), and (3) IBDQ Item 13 (troubled by pain in the abdomen). Using the scores at Week 2 and Week 8, patients were classified based on these three assessments. These results are presented in Table 11 and show that participants who were in remission (based on Adapted Mayo Score) had significantly lower scores for the AP Diary Item at Weeks 2 and 8 compared to participants not in remission. Similarly, participants who reported more abdominal pain on both UC-SQ Item 3 and IBDQ Item 13 had significantly higher scores on the AP Diary Item at both timepoints.

Table 11. Known-groups Comparisons for Abdominal Pain Diary Item Frequency Score at Week 2 and Week 8 for M14-234, Substudy 1

Comparison Group	n	Mean (SD)	Median	P-value*
Week 2 (N=248)				
Adapted Mayo Score				
Clinical remission	28	0.7 (0.6)	0.7	0.029
Non-remission	176	0.9 (0.7)	1.0	
UC-SQ Item 3 (abdominal pain)				
Not at all	21	0.3 (0.6)	0.0	<0.001
A little bit	38	0.6 (0.5)	0.5	
Somewhat	35	1.3 (0.5)	1.0	
Quite a bit	16	1.5 (0.5)	1.3	
Very much	4	1.9 (0.4)	2.0	
IBDQ Item 13 (troubled by pain in the abdomen)				
All of the time	4	1.8 (0.3)	1.8	<0.001
Most of the time	24	1.6 (0.5)	1.7	
A good bit of the time	40	1.5 (0.6)	1.7	
Some of the time	53	1.0 (0.5)	1.0	
A little of the time	45	0.7 (0.4)	0.7	
Hardly any of the time	44	0.4 (0.5)	0.3	
None of the time	24	0.1 (0.4)	0.0	
Week 8 (N=248)				
Adapted Mayo Score				
Clinical remission	28	0.2 (0.5)	0.0	<0.001
Non-remission	175	0.7 (0.7)	0.7	

Table 11. Known-groups Comparisons for Abdominal Pain Diary Item Frequency Score at Week 2 and Week 8 for M14-234, Substudy 1

Comparison Group	n	Mean (SD)	Median	P-value*
UC-SQ Item 3 (abdominal pain)				
Not at all	45	0.2 (0.4)	0.0	<0.001
A little bit	40	0.6 (0.5)	0.7	
Somewhat	17	1.1 (0.4)	1.0	
Quite a bit	10	1.4 (0.6)	1.2	
Very much	3	1.6 (1.0)	1.0	
IBDQ Item 13 (troubled by pain in the abdomen)				
All of the time	3	1.9 (0.8)	2.0	<0.001
Most of the time	12	1.5 (0.4)	1.5	
A good bit of the time	13	1.2 (0.7)	1.0	
Some of the time	33	1.1 (0.5)	1.0	
A little of the time	30	0.8 (0.6)	1.0	
Hardly any of the time	54	0.4 (0.5)	0.0	
None of the time	51	0.1 (0.2)	0.0	

Abbreviations: IBDQ=Inflammatory Bowel Disease Questionnaire; SD=standard deviation; UC-SQ=Ulcerative Colitis Symptoms Questionnaire

Note: Clinical remission on the Adapted Mayo is defined as having a stool frequency subscore of 0 or 1 and not greater than the Baseline score AND rectal bleeding subscore of 0 AND endoscopic subscore of 0 or 1; non-remission is defined as individuals not in the remission state

* p-values are from a Mann–Whitney U test for comparisons between mean from two groups and a Kruskal–Wallis test for more than two groups.

The known-groups analysis for the BU Diary Item scores was conducted using (1) the Adapted Mayo Score and (2) UC-SQ Item 17 (“Did you experience a sudden or intense need to have a bowel movement”). Using the scores at Week 2 and Week 8, patients were classified based on these two assessments.

Results are presented in Table 12 and show that participants who were in remission (based on Adapted Mayo Score) had significantly lower scores for the BU Diary Item at Weeks 2 and 8, compared to participants not in remission. Similarly, participants who reported more bowel urgency on UC-SQ Item 17 had significantly higher scores on the BU Diary Item at both timepoints.

Therefore, results presented demonstrate the frequency scores on both the AP and BU Diary Items are able to distinguish between clinically distinct groups.

Table 12. Known-Groups Comparisons for BU Diary Total Score at Week 2 and Week 8 for M14-234, Substudy 1

Comparison Group	n	Mean (SD)	Median	P-value*
Week 2 (N=248)				
Adapted Mayo Score				
Clinical remission†	28	1.5 (1.3)	1.0	<0.001
Non-remission: Individuals not in the remission state described above	176	2.2 (1.1)	3.0	
UC-SQ Item 17 (experience sudden or intense need to have a bowel movement)				
Not at all	9	0.0 (0.0)	0.0	<0.001
A little bit	16	1.9 (1.3)	2.5	
Somewhat	46	2.2 (0.9)	2.5	
Quite a bit	35	2.8 (0.5)	3.0	
Very much	8	3.0 (0.0)	3.0	
Week 8 (N=248)				
Adapted Mayo Score				
Clinical remission†	28	0.4 (0.9)	0.0	<0.001
Non-remission: Individuals not in the remission state described above	175	1.8 (1.3)	3.0	
UC-SQ Item 17 (experience sudden or intense need to have a bowel movement)				
Not at all	23	0.3 (0.9)	0.0	<0.001
A little bit	26	1.3 (1.3)	1.0	
Somewhat	43	2.0 (1.2)	3.0	
Quite a bit	15	2.8 (0.6)	3.0	
Very much	8	3.0 (0.0)	3.0	

Abbreviations: SD=standard deviation; UC-SQ=Ulcerative Colitis Symptoms Questionnaire

* p-values are from a Mann–Whitney U test for comparisons between means from two groups and a Kruskal–Wallis test for more than two groups.

† Clinical remission on the Adapted Mayo is defined as having a stool frequency subscore of 0 or 1 and not greater than the Baseline score AND rectal bleeding subscore of 0 AND endoscopic subscore of 0 or 1; non-remission is defined as individuals not in the remission state

7.3 Sensitivity to Change

A score that fluctuates in accordance with true changes in the construct it is designed to measure is said to be sensitive to change. Therefore, sensitivity-to-change analyses focus on change scores over time and, for example, are specified to show that observed improvements or reductions in those scores correspond to improvements or reductions in external criteria also related to the construct.

Sensitivity to change was assessed by correlating change scores from Baseline to Week 8 for both the AP and BU Diary Item scores and the change scores for the EQ-5D-5L,

IBDQ, PGIC, SF-36v2®, UC-SQ, and WPAI:UC. These results are presented in Tables 13 and 14, for AP Diary Item and BU Diary Item, respectively.

Change scores on the AP Diary Item were moderately correlated with change scores for conceptually related supportive questionnaires (Table 13). Specifically, EQ-5D-5L Usual Activities, Pain/Discomfort, and VAS scores; IBDQ total score; PGIC at Week 8; both component scores of the SF-36v2®; UC-SQ Total score; and WPAI:UC Impairment while working, Overall work impairment, and Activity impairment were more strongly correlated.

Similarly, change scores on the BU Diary Item were moderately correlated with change scores for conceptually-related supportive questionnaires, specifically EQ-5D-5L Usual Activities, Pain/Discomfort, and VAS scores; IBDQ total score; PGIC at Week 8; both component scores of the SF-36v2®; UC-SQ Total score; and WPAI:UC Impairment while working, Overall work impairment, and Activity impairment (Table 14). Weak correlations, but in the hypothesized direction, were noted for EQ-5D-5L Mobility, Self-care, and Anxiety/Depression domains, and the WPAI:UC Work time missed score.

Table 13. Spearman Correlation Coefficients Between Changes from Baseline to Week 8 on the Abdominal Pain Diary Item Frequency Score and on Other Assessments for M14-234, Substudy 1

Assessment/Score	AP Diary Item Change Score			
	Hypothesized Direction of Correlation of Change Scores*	N	Correlation	P-value
EQ-5D-5L Mobility	+	189	0.22	0.002
EQ-5D-5L Self-care	+	189	0.11	0.138
EQ-5D-5L Usual activities	+	189	0.37	<0.001
EQ-5D-5L Pain/discomfort	+	189	0.50	<0.001
EQ-5D-5L Anxiety/depression	+	190	0.28	<0.001
EQ-5D-5L Visual analogue scale	-	190	-0.45	<0.001
IBDQ total score	-	191	-0.58	<0.001
PGIC†	+	196	0.43	<0.001
SF-36v2® Physical Component Summary	-	190	-0.44	<0.001
SF-36v2® Mental Component Summary	-	190	-0.33	<0.001
UC-SQ Total score	+	91	0.55	<0.001
WPAI:UC Work time missed	+	117	0.14	0.134
WPAI:UC Impairment while working	+	107	0.37	<0.001
WPAI:UC Overall work impairment	+	117	0.32	<0.001
WPAI:UC Activity impairment	+	190	0.40	<0.001

Abbreviations: AP=Abdominal Pain; EQ-5D-5L=Five-level EQ-5D; IBDQ=Inflammatory Bowel Disease Questionnaire; PGIC=Patient Global Impression of Change; SF-36v2®=36-Item Short Form Survey version 2; UC-SQ=Ulcerative Colitis Symptoms Questionnaire; WPAI:UC=Work Productivity and Activity Impairment Questionnaire for Ulcerative Colitis

* Hypothesized direction of correlation was based on the change score ranges in comparison to the AP Diary Item total score (where lower scores mean fewer days with abdominal pain, therefore change scores that decrease mean improvement in abdominal pain)

† PGIC scores from Week 8 were correlated with change scores between Baseline and Week 8 on the AP Diary Item

Table 14. Spearman Correlation Coefficients Between Changes from Baseline to Week 8 on the Bowel Urgency Diary Item Score and on Other Assessments for M14-234, Substudy 1

Assessment/score	BU Diary Item Change Score			
	Hypothesized Direction of Correlation of Change Scores*	N	Correlation	p-value
EQ-5D-5L Mobility	+	189	0.08	0.284
EQ-5D-5L Self-care	+	189	0.07	0.325
EQ-5D-5L Usual activities	+	189	0.30	<0.001
EQ-5D-5L Pain/discomfort	+	189	0.34	<0.001
EQ-5D-5L Anxiety/depression	+	189	0.29	<0.001
EQ-5D-5L Visual analogue scale	-	190	-0.48	<0.001
IBDQ total score	-	191	-0.47	<0.001
PGIC†	+	196	0.47	<0.001
SF-36v2® Physical Component Summary	-	190	-0.37	<0.001
SF-36v2® Mental Component Summary	-	190	-0.33	<0.001
UC-SQ Total score	+	91	0.53	<0.001
WPAI:UC Work time missed	+	117	0.05	0.557
WPAI:UC Impairment while working	+	107	0.41	<0.001
WPAI:UC Overall work impairment	+	117	0.34	<0.001
WPAI:UC Activity impairment	+	190	0.40	<0.001

Abbreviations: BU=Bowel Urgency; EQ-5D-5L=Five-level EQ-5D; IBDQ=Inflammatory Bowel Disease Questionnaire; PGIC=Patient Global Impression of Change; SF-36v2®=36-Item Short Form Survey version 2 ; UC-SQ=Ulcerative Colitis Symptoms Questionnaire; WPAI:UC=Work Productivity and Activity Impairment Questionnaire for Ulcerative Colitis

* Hypothesized direction of correlation was based on the change score ranges in comparison to the BU Diary Item score (where lower scores mean fewer days with bowel urgency, therefore change scores that decrease mean improvement in bowel urgency)

† PGIC scores from Week 8 were correlated with change scores for between Baseline and Week 8 on the BU Diary Item

In addition to the correlations of change scores between the Diary Item scores and supportive assessments presented above, the magnitude of the change for the AP and BU Diary Item frequency scores were also evaluated. Specifically, Cohen's d effect size (ES) was computed for the change between Baseline and Week 8, and the following definitions were used for interpretation of the magnitude: ES ≤ 0.20 (no change), ES > 0.20 to ≤ 0.50 (small change), ES > 0.50 to ≤ 0.80 (moderate change), and ES > 0.80 (large change).[25] For the AP Diary Item score, an average decrease of -0.6 (SD=0.8) points was observed between Baseline and Week 8, which corresponds to a moderate change (ES=-0.79). For the BU Diary Item score, an average decrease of 0.9 (SD=1.4) points was observed between Baseline and Week 8, which corresponds to a strong change (effect size=-0.95)

8 Interpretation of Scores

While sensitivity to change (sometimes referred to as score responsiveness) characterizes how well scores produced by a given questionnaire can detect or otherwise adhere to actual change in the concept of measurement, score interpretation analysis allows researchers to attribute meaning to that change beyond what can be inferred from "statistically significant" results. While the resolution of abdominal pain is the ultimate goal of the AP Diary Item and the BU Diary Item frequency scores, and inherently meaningful to patients, this is the most conservative endpoint for this questionnaire. Therefore, this section summarizes the methods used to inform how observed differences and changes in the AP Diary Item and the BU Diary Item frequency scores other than complete resolution can be interpreted (at both the group and individual level) via employment of distribution-based methods, anchor-based methods, ROC curves, eCDFs, and PDFs.

There are two primary ways in which scores on a PRO questionnaire can be interpreted; one is at the group level, while the other is at the individual level. The point at which an observed difference between group mean scores can be concluded to be meaningful is referred to here as a minimal clinically important difference (MCID), while the point at which an observed within-person change can be concluded to be meaningful is referred to as a meaningful within-person change (MWPC).[26]

8.1 Anchor-based Methods

Anchor-based methods were employed to generate results that could be used to deepen the understanding of observed within-person change in the AP Diary Item and the BU Diary Item frequency scores. Specifically, when achievement of 0 pain days (e.g., complete resolution of abdominal pain or complete resolution of bowel urgency) is not possible, the results presented below can be used to inform conclusions on what amount of improvement might represent a reasonable change that is still meaningful and can be applied in additional exploratory endpoints.

8.1.1 Anchor-based Analysis Results

The suitability of the PGIC as anchor was first determined by examining the correlation of the PGIC with the AP and BU Diary Item frequency scores. As reported in Tables 13 and 14 above, the correlation between the change score of each diary item and the PGIC score

at Week 8 is moderate (r=0.43 for AP Diary Item, and r=0.47 for BU Diary Item); therefore, the PGIC is an acceptable anchor.

8.1.1.1 Meaningful Within-person Change for AP Diary Item

Changes on the frequency score from Baseline to Week 8 on the AP Diary Item were summarized using the PGIC, along with the display of eCDFs and PDFs by each of the PGIC response options. In addition, ROC curves were used to determine the threshold values for improvement (or deterioration) for the AP Diary Item score.[27]

Table 15 shows the change scores on the AP Diary Item between Baseline and Week 8 for each PGIC change category, and we note some issues with the ordering of the change scores for the categories (e.g., similar change scores for improved and no change response options). Of note, a reduction of 1 on the AP Diary Item frequency score (which ranges from 0 to 3), is associated with patient-reported improvement (specifically "much improved" or "very much improved") on the PGIC.

Table 15. Mean Change from Baseline to Week 8 of the Abdominal Pain Diary Item Frequency Score for Patient Global Impression of Change Anchor for M14-234, Substudy 1

PGIC Response Options	n	Baseline Mean (SD)	Week 8 Mean (SD)	Mean Change (SD)	Within-group P-value*	Between-groups P-value†
Very much improved (PGIC=1)	56	1.20 (0.69)	0.28 (0.49)	-0.92 (0.68)	<0.001	<0.001
Much improved (PGIC=2)	58	1.24 (0.84)	0.44 (0.52)	-0.80 (0.88)	<0.001	
Improved (PGIC=3)	36	1.32 (0.77)	1.05 (0.79)	-0.28 (0.50)	0.001	
No change (PGIC=4)	28	1.10 (0.63)	0.83 (0.65)	-0.26 (0.40)	0.001	
Worsened (PGIC=5)	7	1.10 (0.16)	1.19 (0.33)	0.10 (0.16)	0.500	
Much worsened (PGIC=6)	6	1.39 (1.02)	1.22 (0.75)	-0.17 (0.94)	1.000	
Very much worsened (PGIC=7)	0	N.A. (N.A.)	N.A. (N.A.)	N.A. (N.A.)	N.A.	

Abbreviations: PGIC=Patient Global Impression Change; SD=standard deviation

* The within-group p-value is from a Wilcoxon Signed Rank test on change scores at each level of PGIC response.

† The between-groups p-value is from a Kruskal-Wallis testing distributional shift in change scores between PGIC response groups

Exploratory proxy anchors were created using select items from the IBDQ and UC-SQ questionnaires that assess abdominal pain. Change scores between Baseline and Week 8 for IBDQ Item 13 (troubled by pain in abdomen), and UC-SQ Item 3 (abdominal pain) were calculated for each of the PGIC response options at Week 8. Mean change scores associated with PGIC categories were evaluated to defined improvement/no change/ worsening categories for each proxy anchor. Table 16 demonstrates that the ordering of

the change scores associated with the proxy anchor groups supports a 1-point reduction and is as expected (in comparison to the PGIC above). Additionally, the results for the exploratory anchors show similar change scores on the AP Diary Item associated with improvement on IBDQ Item 13 and UC-SQ Item 3. Therefore a 1-point reduction on the AP Diary Item can be considered a meaningful improvement for the frequency of abdominal pain for the target patient population.

Table 16. Mean Change from Baseline to Week 8 of the Abdominal Pain Diary Item Frequency Score for M14-234, Substudy 1

Anchors	n	Baseline Mean (SD)	Week 8 Mean (SD)	Mean Change (SD)	Within-group P-value	Between-groups P-value
IBDQ Item 13. How often during the last 2 weeks have you been troubled by pain in the abdomen?*						
Very much improved (>2-point change)	54	1.59 (0.72)	0.33 (0.48)	-1.25 (0.75)	<0.001	<0.001
Much improved (2-point change)	32	1.21 (0.60)	0.57 (0.65)	-0.64 (0.61)	<0.001	
Improved (1-point change)	34	1.03 (0.65)	0.53 (0.66)	-0.50 (0.56)	<0.001	
No change/Worsened (≤0-point change)	65	1.02 (0.79)	0.89 (0.73)	-0.12 (0.49)	0.022	
UC-SQ Item 3. During the past week, did you have abdominal pain?†						
Improved (≤-1-point change)	56	1.33 (0.67)	0.48 (0.57)	-0.85 (0.77)	<0.001	<0.001
No change/Worsened (>-1-point change)	29	1.00 (0.91)	0.84 (0.80)	-0.16 (0.61)	0.195	

Abbreviations: IBDQ=Inflammatory Bowel Disease Questionnaire; SD=standard deviation; UC-SQ=Ulcerative Colitis Symptoms Questionnaire

* The within-group p-value is from a Wilcoxon Signed Rank test on change scores at each level of IBDQ response, and the between-groups p-value is from a Kruskal-Wallis testing distributional shift in change scores between IBDQ response groups.

† The within-group p-value is from a Wilcoxon Signed Rank test on change scores at each level of UC-SQ response, and the between-groups p-value is from a Kruskal-Wallis testing distributional shift in change scores between UC-SQ response groups.

8.1.1.2 Meaningful Within-person Change for BU Diary Item

Table 17 shows the change scores on the BU Diary Item between Baseline and Week 8 for each PGIC change category and, based on this information, a reduction of 1 to 2 days on the BU Diary Item (which ranges from 0 to 3) is associated with patient-reported improvement on the PGIC (Much or Very Much improved).

An exploratory proxy anchor was created using Item 17 of the UC-SQ questionnaire, which also assesses bowel urgency. Change scores between Baseline and Week 8 for UC-SQ Item 17 (bowel urgency) was calculated for each of the PGIC response options at Week 8. Mean change scores associated with PGIC categories were evaluated against

defined improvement/no change/worsening categories for the proxy anchor. Change scores on the BU Diary Item associated with improvement on the UC-SQ Item 17 (Table 18), were similar to the results presented above for the PGIC. Specifically, the change score on the BU Diary Item associated with the "much improved" category is -1.62, which supports a 2-point reduction in days with bowel urgency as a meaningful improvement for the target patient population.

Table 17. Mean Change from Baseline to Week 8 of the Bowel Urgency Diary Item Score for Patient Global Impression of Change Anchor for M14-234, Substudy 1

PGIC Response Options	n	Baseline Mean (SD)	Week 8 Mean (SD)	Mean Change (SD)	Within-Group P-value*	Between-Groups P-value†
Very Much Improved (PGIC=1)	56	2.45 (1.09)	0.88 (1.24)	-1.57 (1.43)	<0.001	<0.001
Much Improved (PGIC=2)	58	2.59 (0.82)	1.17 (1.30)	-1.41 (1.34)	<0.001	
Improved (PGIC=3)	36	2.64 (0.80)	2.28 (1.11)	-0.36 (0.96)	0.034	
No Change (PGIC=4)	28	2.57 (1.00)	2.46 (1.00)	-0.11 (0.69)	0.625	
Worsened (PGIC=5)	7	2.00 (1.41)	2.71 (0.49)	0.71 (0.95)	0.250	
Much Worsened (PGIC=6)	6	2.00 (1.55)	2.50 (1.22)	0.50 (1.22)	1.000	
Very Much Worsened (PGIC=7)	0	N.A. (N.A.)	N.A. (N.A.)	N.A. (N.A.)	N.A.	

Abbreviations: PGIC=Patient Global Impression Change; SD=standard deviation

* The within-group p-value is from a Wilcoxon Signed Rank test on change scores at each level of PGIC response.

† The between-groups p-value is from a Kruskal-Wallis testing distributional shift in change scores between PGIC response groups

Table 18. Mean Change from Baseline to Week 8 of the Bowel Urgency Diary Item Score for the Proxy Anchor M14-234, Substudy 1

Anchors	n	Baseline Mean (SD)	Week 8 Mean (SD)	Mean Change (SD)	Within-group p-value	Between-groups p-value
UC-SQ Item 17 (During the past week, did you experience sudden or intense need to have a bowel movement)*						
Very much improved <-2-point change)	9	2.89 (0.33)	0.11 (0.33)	-2.78 (0.44)	0.004	<0.001
Much improved (-2-point change)	21	2.95 (0.22)	1.33 (1.43)	-1.62 (1.40)	<0.001	
Improved (-1-point change)	30	2.20 (1.24)	1.73 (1.31)	-0.47 (1.41)	0.103	
No change/Worsened ≥0-point change)	25	2.64 (0.81)	2.56 (0.77)	-0.08 (0.76)	0.813	

Abbreviations: SD=standard deviation; UC-SQ=Ulcerative Colitis Symptoms Questionnaire

* Response options: Never (0), Rarely (1), Sometimes (2), Often (3), Always (4); change scores were calculated between Baseline and Week 8, for each PGIC category at Week 8 The within-group p-value is from a Wilcoxon Signed Rank test on change scores at each level of UC-SQ response, and the between-groups p-value is from a Kruskal-Wallis testing distributional shift in change scores between UC-SQ response groups.

8.1.2 Empirical Cumulative Distribution Functions

The FDA PRO Guidance, rather than emphasizing a precise amount of change that can be considered clinically meaningful, emphasizes the presentation of a plot of the eCDF for endpoints to support parametric treatment effect analysis or non-parametric responder analysis (based typically on a single criterion to define "responder").[10] The eCDF allows for an understanding of the change in an endpoint at all points on the continuum, and thus allows for a more comprehensive understanding of the questionnaire results. The eCDF also allows an evaluation of the consistency of effects across the entire distribution and is not impacted by outliers. The eCDF is recommended as supportive to parametric analyses because it provides more information than does a single-point estimate of the difference between group mean changes.[28] The eCDF allows all estimated MWPCs for an anchor to be evaluated simultaneously. Notably, the aim of eCDF plots is not to estimate MWPCs but to evaluate the performance of each.

The eCDF plots display a continuous AP Diary Item and BU Diary Item score changes from Baseline on the x-axis for each of the scores. The axis is ordered from the best possible decreases in frequency of abdominal pain on the left to the worst possible increases in abdominal pain frequency on the right. The cumulative percentage of patients experiencing that change is displayed on the y-axis, plotted as an empirical (i.e., step-wise) eCDF. The eCDF curves are split by anchor group so the separation of the curves can be visually compared across the range of potential thresholds identified across the different anchor methods.

8.1.2.1 Empirical Cumulative Distribution Function for AP Diary Item

Figure 3 presents the eCDF curve for the AP Diary Item change scores between Baseline and Week 8, by the PGIC anchor. There is a separation between the no change/worsening groups compared to the improvement groups for change scores greater than a 1-point *decrease* in abdominal pain frequency. Table 19 precents the percentiles of the change-score distributions for each of the PGIC anchor groups. Because negative change scores denote *decreases* in frequency of abdominal pain, the most-improved patients are summarized in the *lowest* percentiles (i.e., the left end) of the score distribution. Given the sample sizes of the anchor groups from the PGIC (See Table 15), an exploratory eCDF figure with collapsed anchor categories (Very Much Improved, Much Improved, Minimally improved/No Change, and Worsened [Minimally, Much and Very Much]) was examined and results align with a 1-point decrease on the AP Diary Item.

Similar to the results presented in Figure 3, there is a separation between placebo and treatment groups for change scores of the AP Diary Item greater than a 1-point *decrease* in abdominal pain frequency.

Figure 3. Empirical Cumulative Distribution Function for Change in Abdominal Pain Diary Item Frequency Score Between Baseline and Week 8, by Patient Global Impression of Change Response Option Categories at Week 8 (M14-234, Substudy 1)

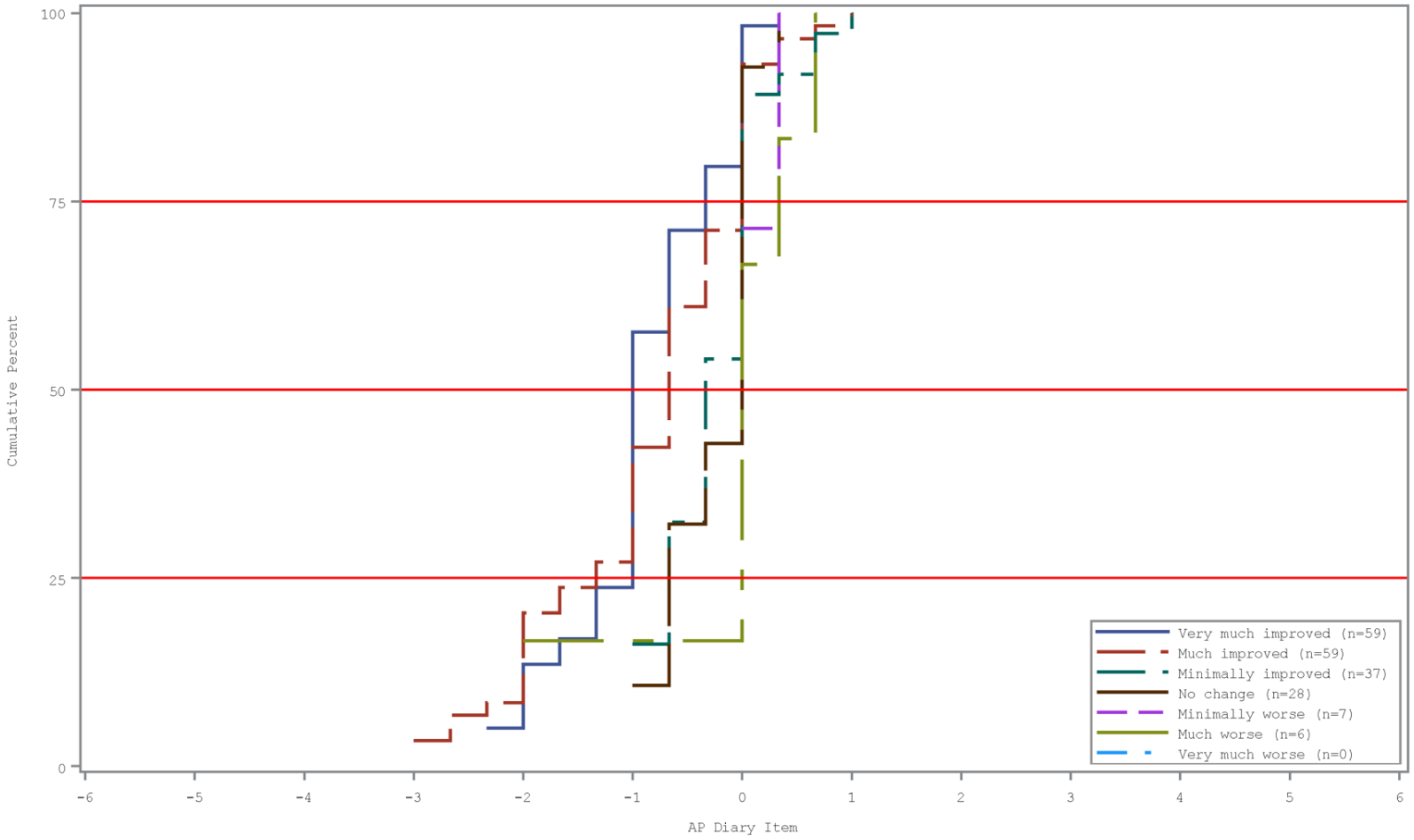


Table 19. Percentile Change in Abdominal Pain Diary Item Frequency Score from Baseline to Week 8 by Patient Global Impression of Change Response Groups per Empirical Cumulative Distribution Function Curve (M14-234, Substudy 1)

Change in AP Diary Item frequency score	PGIC anchor category							
	Very much improved	Much improved	Minimally improved	No change	Minimally worse	Much worse	Very much worse	Total
N	59	59	37	28	7	6	0	196
Mean (SD)*	-0.9 (0.7)	-0.8 (0.9)	-0.3 (0.5)	-0.3 (0.4)	0.1 (0.2)	-0.2 (0.9)	--	-0.6 (0.8)
10th percentile†	-2.0	-2.0	-1.0	-1.0	0.0	-2.0	--	-2.0
25th percentile†	-1.0	-1.3	-0.7	-0.7	0.0	0.0	--	-1.0
Median (50th percentile)†	-1.0	-0.7	-0.3	0.0	0.0	0.0	--	-0.7
75th percentile†	-0.3	0.0	0.0	0.0	0.3	0.3	--	0.0
90th percentile†	0.0	0.0	0.3	0.0	0.3	0.7	--	0.0

Abbreviations: AP=Abdominal Pain; PGIC=Patient Global Impression of Change; SD=standard deviation

* Mean (SD) for the change score on the AP Diary Item between Baseline and Week 8 for each anchor category

† Change score for AP Diary Item is presented associated with each percentile group.-

8.1.2.2 Empirical Cumulative Distribution Function for BU Diary Item

Similar to the results presented above for the AP Diary Item, Figure 4 presents the eCDF curve for the BU Diary Item change scores between Baseline and Week 8, by the PGIC anchor. There is a separation between the no change/worsening groups compared to the improvement groups for change scores greater than a 1-point *decrease* in bowel urgency. Table 20 presents the percentiles of the change-score distributions for each of the PGIC anchor groups and estimates of greater than 1.4 points are associated with improvement. Because negative change denotes *decreases* in the number of days with bowel urgency, the most-improved patients are summarized in the *lowest* percentiles (i.e., left end) of the score distribution. Given the sample sizes of the anchor groups from the PGIC (See Table 17), an exploratory eCDF figure with collapsed anchor categories (Very Much Improved, Much Improved, Minimally improved/No Change, and Worsened [Minimally, Much and Very Much]) was examined and results align with a 1-point decrease on the BU Diary Item.

Similar to the results presented in Figure 4, there is a separation between placebo and treatment groups for change scores of the BU Diary Item greater than a 1-point *decrease* in bowel urgency.

Figure 4. Empirical Cumulative Distribution Function for Change in BU Diary Item Score Between Baseline and Week 8, by Patient Global Impression of Change Response Option Categories at Week 8 (M14-234, Substudy 1)

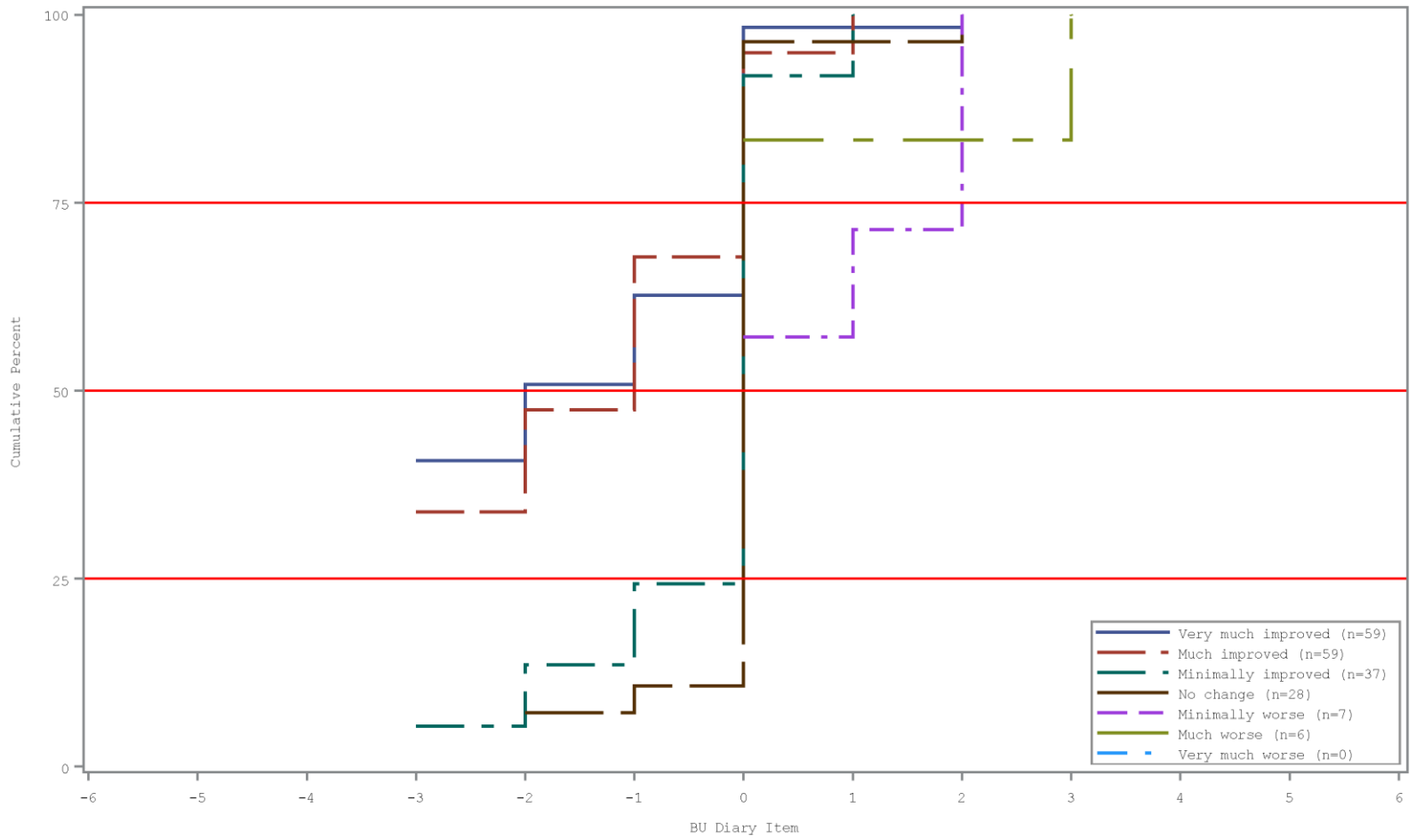


Table 20. Percentile Change in Bowel Urgency Diary Item from Baseline to Week 8 by Patient Global Impression of Change Response Groups per Empirical Cumulative Distribution Function Curve (M14-234, Substudy 1)

Change in BU Diary Item Score	PGIC anchor category							
	Very Much Improved	Much Improved	Minimally Improved	No Change	Minimally Worse	Much Worse	Very Much Worse	Total
N	59	59	37	28	7	6	0	196
Mean (SD)*	-1.5 (1.4)	-1.4 (1.3)	-0.4 (0.9)	-0.1 (0.7)	0.7 (1.0)	0.5 (1.2)	---	-0.9 (1.4)
10th percentile†	-3.0	-3.0	-2.0	-1.0	0.0	0.0	---	-3.0
25th percentile†	-3.0	-3.0	0.0	0.0	0.0	0.0	---	-2.0
Median (50th percentile)†	-2.0	-1.0	0.0	0.0	0.0	0.0	---	0.0
75th percentile†	0.0	0.0	0.0	0.0	2.0	0.0	---	0.0
90th percentile†	0.0	0.0	0.0	0.0	2.0	3.0	---	0.0

Abbreviations: BU=Bowel Urgency; PGIC=Patient Global Impression Change; SD=standard deviation

* Mean (SD) for the change score on the BU Diary Item between Baseline and Week 8 for each anchor category

† Change score for BU Diary Item is presented associated with each percentile group.

8.1.3 Probability Density Function Curves

PDFs were also plotted with continuous AP Diary Item score changes from Baseline on the x-axis for each of the scores, from the worst possible deterioration on the right to the best possible improvement on the left (given improvement is a *decrease in frequency* of abdominal pain or bowel urgency). The PDF curves can be especially informative for diagnosis purposes when there is not a clear, consistent separation between the eCDF curves. The percentage of patients experiencing that change are displayed on the y-axis, plotted as a kernel-smoothed PDF, and split by PGIC anchor groups, so the separation of the curves can be visually compared across the range of potential thresholds identified across the different anchor methods.

8.1.3.1 Probability Density Function Curves for AP Diary Item

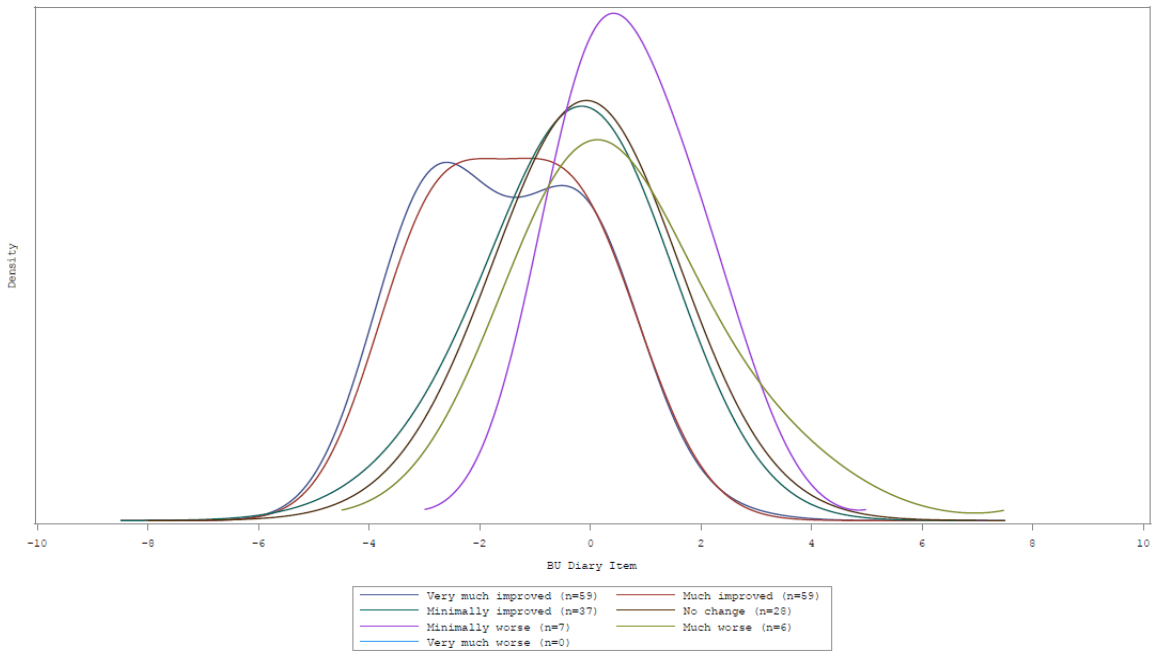
Similar to the results noted for the anchor-based analyses (Table 15), the anchor categories for the PDF curve were not ordered as expected, which could be a function of the small sample sizes for the anchor groups and outliers. Therefore, the seven-category PGIC anchor may be too granular and collapsing categories may produce more interpretable results. Collapsed anchor groups for the PGIC were created and Exploratory Figure 4.4 shows the PDF curve for the AP Diary Item for each of four anchor groups (Very Much Improved, Much Improved, Minimally Improved/No change, and Worse [Minimally, Much, and Very Much]). Similar to the results for the seven-category PGIC, the collapsed four-category anchor still shows considerable overlap for the groups.

8.1.3.2 Probability Density Function Curves for BU Diary Item

PDF curves were also plotted with continuous BU Diary Item score change from Baseline on the x-axis for each of the scores, from the worst possible deterioration on the right to the best possible improvement on the left (given improvement is a decrease in bowel urgency). The interpretation of the PDF (Figure 5) is challenging due to the fact that 52% and 43% of patients who reported, respectively, “Very much improved” and “Much improved” on the PGIC had BU change scores of -3 (complete resolution of bowel urgency) between Baseline and Week 8. Because the PDFs were derived by a kernel-smoothing process and were not empirical, the domain of the PDFs extended beyond the range of possible score changes. These effects are common when there are significant floor or ceiling effects in the change score distribution. Similar to the exploratory results presented for eCDF in Section 8.1.2, PGIC anchor groups were collapsed into four groups

(Very much improved, Much Improved, Minimally improved/No change, and Worsening [Minimally, Much, and Very Much]), and align with the results above, supporting a 1- to 2-point decrease on the BU Diary Item as associated with improvement on the PGIC.

Figure 5. Probability Density Function for Change in Bowel Urgency Diary Item Score by Patient Global Impression of Change Response Groups from Baseline to Week 8 for M14-234, Substudy 1



8.1.4 Receiver Operating Characteristic Curves

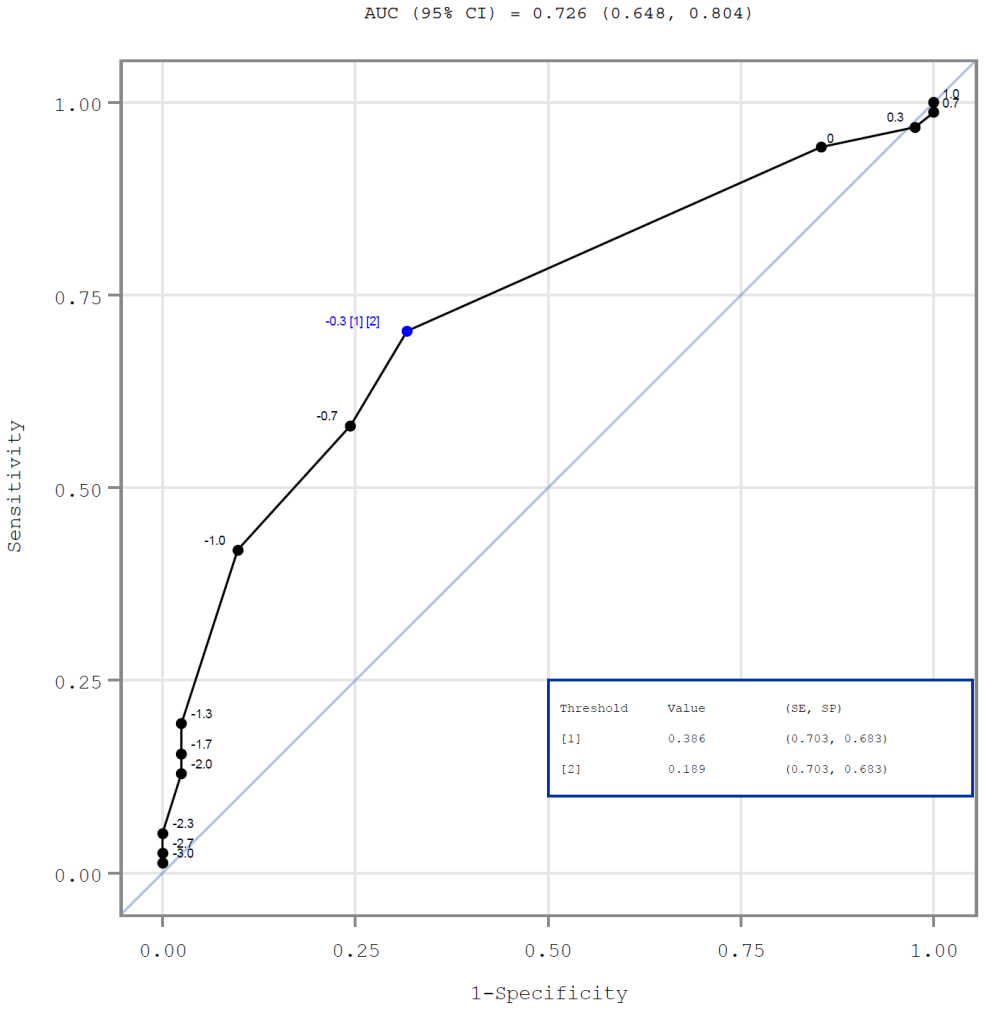
ROC curve analysis was used to specify the optimal cut-point of the change score on the target measure that optimally discriminates between each anchor group defined by the anchors. To determine the optimal threshold values for the AP Diary Item and the BU Diary Item frequency scores, two different methods were considered:

- A: Maximum[sensitivity+specificity-1], i.e., maximizes the sum of sensitivity and specificity, also referred to as Youden's J index[29]; and
- B: Minimum[(1 – sensitivity)^2+(1 – specificity)],[10] i.e., the point in the ROC space that minimizes the sum of squares, which is equivalent to choosing the cut-point closest to the top left corner of the ROC curve.

The threshold values determined by using methods A and B are presented in Figures 6 and 7 (see below) for AP Diary Item and Figures 8 and 9 for BU Diary Item, below.

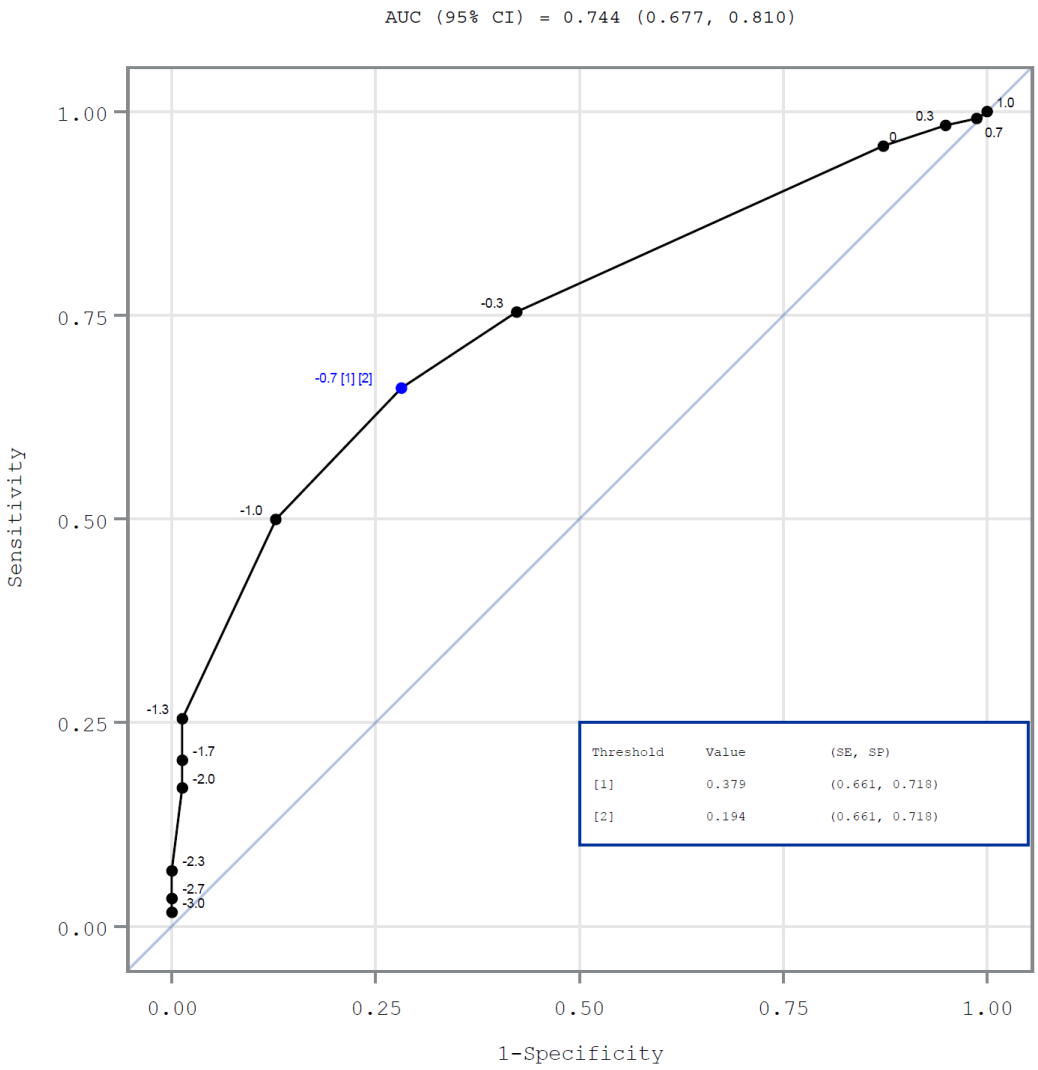
Results support an improvement/decrease of between 0.3 and 0.7 points in abdominal pain frequency, and 1 point in bowel urgency frequency.

Figure 6. Received Operating Characteristic Curve For Abdominal Pain Diary Item change scores between Baseline and Week 8, by Patient Global Impression of Change ≤ "Minimally improved" at Week 8 (M14-234, Substudy 1)



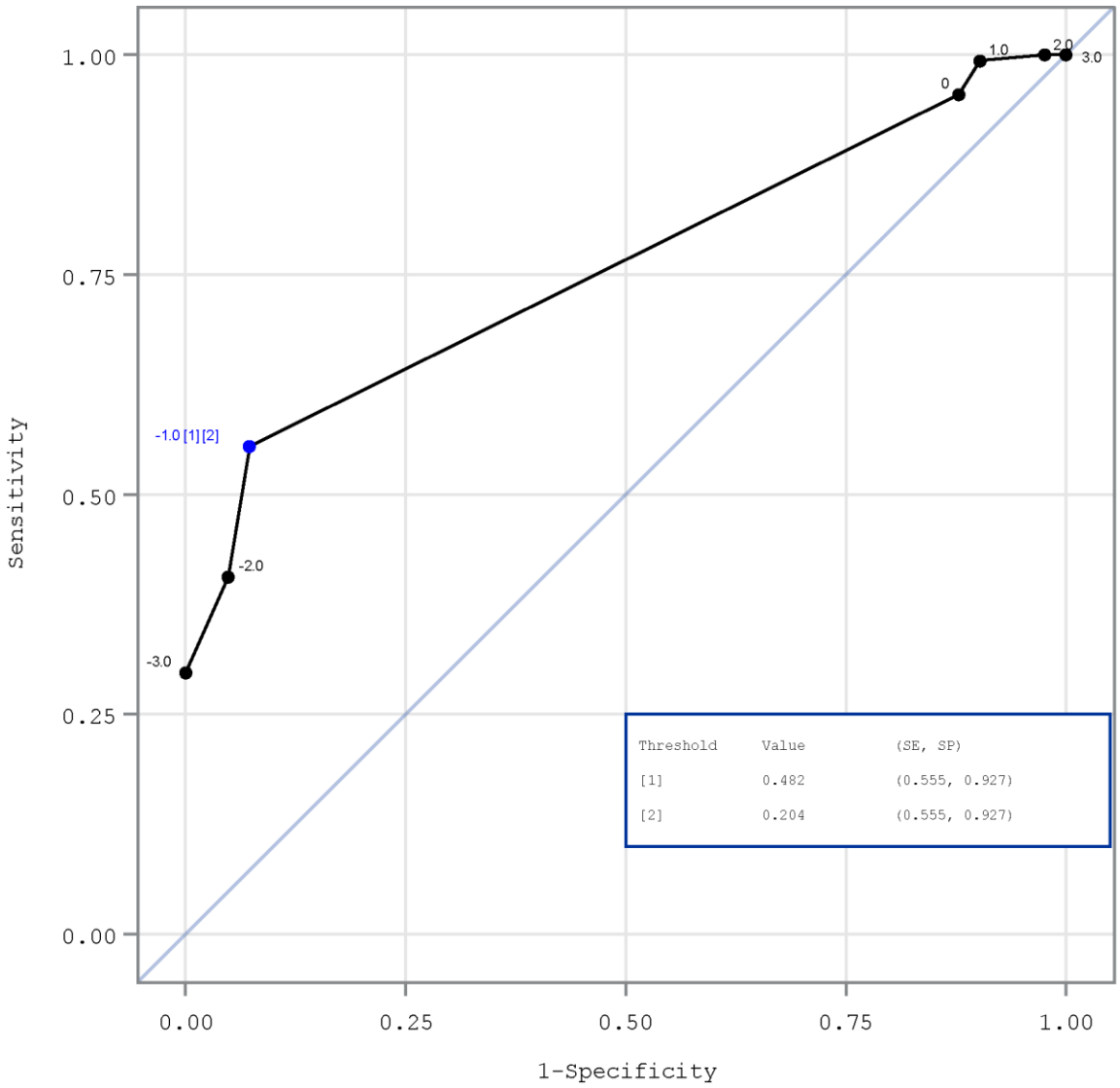
Abbreviations: AUC=area under the curve; CI=confidence interval; SE=sensitivity; SP=specificity

Figure 7. Received Operating Characteristic Curve For Abdominal Pain Diary Item change scores between Baseline and Week 8, by Patient Global Impression of Change ≤ "Much improved" at Week 8 (M14-234, Substudy 1)



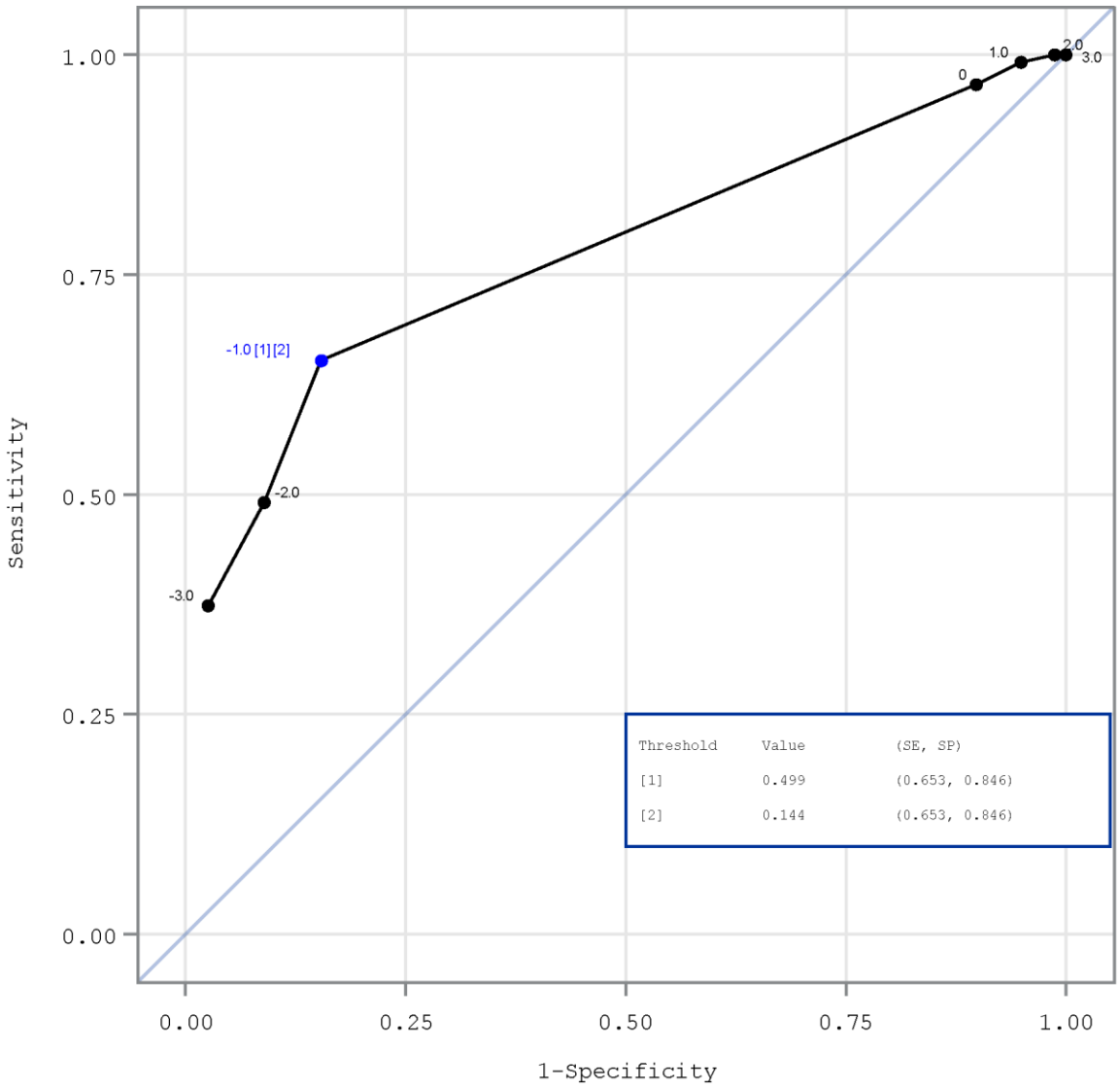
Abbreviations: AUC=area under the curve; CI=confidence interval; SE=sensitivity; SP=specificity

Figure 8. Receiver Operating Characteristic Curve for Bowel Urgency Diary Item Change Scores Between Baseline and Week 8, by Patient Global Impression of Change ≤ "Minimally Improved" at Week 8 (M14-234, Substudy 1)



Abbreviations: SE=sensitivity; SP=specificity

Figure 9. Receiver Operating Characteristic Curve for Bowel Urgency Diary Item Change Scores Between Baseline and Week 8, by Patient Global Impression of Change ≤ "Much Improved" at Week 8 (M14-234, Substudy 1)



Abbreviations: SE=sensitivity; SP=specificity

8.2 Distribution-based Methods

The observed distribution of the data can be used to generate estimates of between-groups MCID by identifying the amount of change that exceeds measurement error on a group level. The distribution-based methods consist of the following:

- MCID1: This distribution based approach involved calculating 0.5 of the SD of the AP or BU Diary Item frequency score at Baseline[30]; and
- MCID2: Standard error of measurement (SEM), which was calculated as follows:

$$SEM = SD\ at\ Baseline * \sqrt{1 - reliability}$$

For the AP Diary Item the MCID2 was assessed using test-retest reliability of the AP Diary Item frequency score from both the PGIC-defined stable subgroup and the stable subgroup defined by UC-SQ Item 3, between Baseline and Week 2. Results presented in Table 21 suggest that a meaningful between-groups difference on the AP Diary Item frequency score is between 0.3 and 0.4 points.

Table 21. Distribution-based Statistics for Abdominal Pain Diary Item Frequency Score at Baseline (M14-243, Substudy 1)

Scale	n	MCID1 0.5 SD	Reliability	MCID2 SEM
AP Diary Item frequency score	243	0.37	0.80*	0.33
			0.85†	0.29

Abbreviations: AP=Abdominal Pain; ICC=intraclass correlation coefficient; MCID=minimal clinically important difference; SD=standard deviation; SEM=standard error of measurement;

Note: SEM is calculated as Baseline [SD×(1-r)$^{1/2}$], where r is the reliability of the score

* Reliability for AP Diary Item is based on the test-retest reliability (ICC) in PGIC stable subjects (who selected "no change" at Week 2.

† Reliability for AP Diary Item is based on the test-retest reliability (ICC) in UC-SQ Item 3 stable subjects between Baseline and Week 2.

Similarly for the BU Diary Item, the MCID2 was assessed using test-retest reliability from both the PGIC-defined stable subgroup (participants who reported "no change" on PGIC at Week 2), and the UC-SQ Item 17-defined stability (participants who endorsed the same response at Baseline and Week 2). Results presented in Table 22 suggest that a meaningful between-groups difference (MCID) on the BU Diary Item is between 0.46 and 0.77 points.

Table 22. Distribution-Based Statistics for Bowel Urgency Diary Item Score at Baseline

Scale	n	MCID1 0.5 SD	Reliability	MCID2 SEM
BU Diary Item score	243	0.46	0.33*	0.75
			0.29†	0.77

Abbreviations: BU=Bowel Urgency; MCID=minimal clinically important difference; SD=standard deviation; SEM=standard error of measurement

* Reliability for BU Diary Items are based on the test-retest reliability in PGIC-defined stable subjects between Baseline and Week 2.

† Reliability for BU Diary Items are based on the test-retest reliability in UC-SQ Item 17-defined stable subjects between Baseline and Week 2.

9 Conclusion

The results presented herein support the psychometric performance of the score generated by the AP and BU Diary Items, and provided guidance for interpretation of both meaningful within-person change and between-group differences.

AP Diary Item

Specifically for the AP Diary, quality of completion for the psychometric analysis population was high, and scores captured the full range of the scale across all timepoints (Baseline, Week 2, and Week 8). In addition, scores demonstrated adequate test-retest reliability between Baseline and Week 2, convergent validity (correlated more strongly with scores from disease-specific questionnaires [specific to UC or inflammatory bowel disease] than scores from generic questionnaires), known-groups analysis results (scores differed between clinically distinct groups, as expected), and sensitivity to change. Overall, the AP Diary Item demonstrated strong psychometric properties for evaluating frequency of abdominal pain among participants of the Phase 2b clinical trial.

While the complete resolution of abdominal pain (e.g., having 0 days with abdominal pain) is inherently meaningful to patients, additional responder definitions that are less stringent were also evaluated for their meaningfulness to patients. Specifically, anchor-based methods, values from cumulative distribution functions, and receiver operating characteristic analysis suggested estimates of MWPC of 1 point (decrease in frequency). Supportive distribution-based methods estimated approximately a 0.3-point difference between group means would be meaningful.

BU Diary Item

For the BU Diary, qualitative of completion was high with very minimal missing data associated with the timepoints of the analyses (Baseline, Week 2, and Week 8). In addition, the scores used the full range of the scale at these timepoints. To evaluate test-retest reliability in the absence of two timepoints before initiating treatment, two stable sub-groups were defined: (1) participants who selected “no change” on the Patient Global Impression of Change (PGIC) at Week 2, and (2) participants who selected the same response on the US-SQ Item 17 (bowel urgency) at Baseline and Week 2. Scores demonstrated low test-retest reliability between Baseline and Week 2 for both stable sub-groups; however, the small sample sizes for the subgroups impacts the interpretability of the results. Score on the BU Diary Item demonstrated acceptable convergent validity

(correlated more strongly with scores from disease-specific questionnaires [specific to UC or inflammatory bowel disease], compared to generic questionnaires), adequate known-groups results (scores differed between clinically distinct groups, as expected) and sensitivity to change. Overall, the BU Diary Item demonstrated acceptable validity, in spite of the low test-retest reliability for evaluating the frequency of bowel urgency among participants of the Phase 2b clinical trial.

Similar to abdominal pain, while the complete resolution of bowel urgency (e.g., having 0 days with bowel urgency), is inherently meaningful to patients, additional responder definitions that are less stringent were also evaluated for their meaningfulness to patients. Specifically, anchor-based methods, values from eCDFs and ROC analysis suggested estimates of MWPC of 1-point (decrease in days with bowel urgency). Supportive distribution-based methods estimated approximately a 0.5-point difference between group means would be meaningful.

10 References

1. Gohil K, Carramusa B. Ulcerative colitis and Crohn's disease. *P T.* 2014;39(8):576-577.
2. Langan RC, Gotsch PB, Krafczyk MA, Skillinge DD. Ulcerative colitis: diagnosis and treatment. *Am FamPhysician.* 2007;76(9):1323-1330.
3. National Institute of Diabetes and Digestive and Kidney Diseases Ulcerative Colitis. 09/2014; https://www.niddk.nih.gov/health-information/digestive-diseases/ulcerative-colitis. Accessed 05/02/2018.
4. Danese S, Fiocchi C. Ulcerative colitis. *N Engl J Med.* 2011;365(18):1713-1725.
5. Panaccione R. Mechanisms of inflammatory bowel disease. *Gastroenterology & hepatology.* 2013;9(8):529-532.
6. Rutgeerts P, Sandborn WJ, Feagan BG, et al. Infliximab for induction and maintenance therapy for ulcerative colitis. *N Engl J Med.* 2005;353(23):2462-2476.
7. Sandborn WJ, van Assche G, Reinisch W, et al. Adalimumab induces and maintains clinical remission in patients with moderate-to-severe ulcerative colitis. *Gastroenterology.* 2012;142(2):257-265.e251-253.
8. Feagan BG, Greenberg GR, Wild G, et al. Treatment of ulcerative colitis with a humanized antibody to the alpha4beta7 integrin. *N Engl J Med.* 2005;352(24):2499-2507.
9. Sandborn WJ, Feagan BG, Marano C, et al. Subcutaneous golimumab induces clinical response and remission in patients with moderate-to-severe ulcerative colitis. *Gastroenterology.* 2014;146(1):85-95; quiz e14-85.
10. US Department of Health and Human Services, Food and Drug Administration, 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.* Silver Spring, MD: Office of Communications, Division of Drug Information;2009.
11. Patrick DL, Burke LB, Gwaltney CJ, 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.
12. Patrick DL, Burke LB, Gwaltney CJ, 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.

13. Guyatt G, Mitchell A, Irvine EJ, et al. A new measure of health status for clinical trials in inflammatory bowel disease. *Gastroenterology.* 1989;96(3):804-810.
14. Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res.* 2011;20(10):1727-1736.
15. Ware JE, Jr., Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care.* 1992;30(6):473-483.
16. Maruish ME, ed *User's Manual for the SF-36v2®.* 3rd ed. Lincoln, RI: QualityMetric Inc; 2011.
17. European Medicines Agency. *Reflection paper on the Regulatory Guidance for the Use of Health-related Quality of Life (HRQL) measures in the evaluation of medicinal products. EMEA/CHMP/EWP/139391/2004.* 7/27/2005 2005.
18. Thompson B, Vacha-Haase T. Psychometrics is Datametrics: the Test is not Reliable. *Educational and Psychological Measurement.* 2000;60(2):174-195.
19. Bland JM, Altman DG. Cronbach's alpha. *BMJ.* 1997;314(7080):572.
20. Weir J. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strenth and Conditioning Research.* 2005;19(1):231-240.
21. Hibi T, Ishibashi T, Ikenoue Y, Yoshihara R, Nihei A, Kobayashi T. Ulcerative Colitis: Disease Burden, Impact on Daily Life, and Reluctance to Consult Medical Professionals: Results from a Japanese Internet Survey. *Inflamm Intest Dis.* 2020;5(1):27-35.
22. Joyce JC, Waljee AK, Khan T, et al. Identification of symptom domains in ulcerative colitis that occur frequently during flares and are responsive to changes in disease activity. *Health Qual Life Outcomes.* 2008;6:69.
23. Louis E, Ramos-Goñi JM, Cuervo J, et al. A Qualitative Research for Defining Meaningful Attributes for the Treatment of Inflammatory Bowel Disease from the Patient Perspective. *The patient.* 2020;13(3):317-325.
24. Hinkle DE, Jurs SG, Wiersma W. *Applied statistics for the behavioral sciences.* 2nd ed. Boston: Houghton Mifflin; 2003.
25. Maher JM, Markey JC, Ebert-May D. The other half of the story: effect size analysis in quantitative research. *CBE life sciences education.* 2013;12(3):345-351.

26. Cappelleri JC, Zou KH, Bushmakin AG, Alvir JMJ, Alemayehu D, Symonds T. *Patient-Reported Outcomes: Measurement, Implementation and Interpretation.* Boca Raton, FL: CRC Press; 2013.

27. Coon CD, Cappelleri JC. Interpreting Change in Scores on Patient-Reported Outcome Instruments. *Therapeutic Innovation & Regulatory Science.* 2016;50(1):22-29.

28. Patrick DL, Burke LB, Powers JH, et al. Patient-reported outcomes to support medical product labeling claims: FDA perspective. *Value in Health.* 2007;10 Suppl 2:S125-S137.

29. Youden WJ. Index for rating diagnostic tests. *Cancer.* 1950;3(1):32-35.

30. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care.* 2003;41(5):582-592.