GASTROENTEROLOGY 2009;137:1944–1953

Psychometric Evaluation of Patient-Reported Outcomes in Irritable Bowel Syndrome Randomized Controlled Trials: A Rome Foundation Report

BRENNAN SPIEGEL,[*,‡,§] MICHAEL CAMILLERI,[‖] ROGER BOLUS,[‡,§] VIOLA ANDRESEN,[¶] WILLIAM D. CHEY,[#] SHERI FEHNEL,[**] ALLEN MANGEL,[**] NICHOLAS J. TALLEY,[‡‡] and WILLIAM E. WHITEHEAD[§§]

[*]Department of Medicine, VA Greater Los Angeles Healthcare System; [‡]Department of Medicine, David Geffen School of Medicine at UCLA; [§]UCLA/VA Center for Outcomes Research and Education (CORE), Los Angeles, California; [‖]Mayo Clinic, Rochester, Minnesota; [¶]Hamburg, Germany; [#]University of Michigan, Ann Arbor, Michigan; [**]RTI Health Solutions, Research Triangle Park, North Carolina; [‡‡]Mayo Clinic, Jacksonville, Florida; [§§]University of North Carolina, Chapel Hill, North Carolina

See related article, Ford A et al, on page 1279 in *CGH*.

BACKGROUND & AIMS: There is debate about how best to measure patient-reported outcomes (PROs) in irritable bowel syndrome (IBS). We pooled data to measure the psychometric properties of IBS end points, including binary responses (eg, "adequate relief") and 50% improvement in symptom severity. **METHODS:** We pooled data from 12 IBS drug trials involving 10,066 participants. We tested the properties of binary response and 50% improvement end points, including the impact of baseline severity on performance, and measured construct validity using clinical anchors. **RESULTS:** There were 9044 evaluable subjects (age, 44 years; 85% female; 58% IBS constipation-prominent [IBS-C]; 31% IBS diarrhea-prominent [IBS-D]). Using the binary end point, the proportion responding in the mild, moderate, and severe groups was 42%, 40%, and 38%, respectively ($P = .0008$). There was no effect of baseline severity on binary response (odds ratio [OR], 0.99; 95% confidence interval [CI], 0.99–1.0; $P = .07$). The proportions reaching 50% improvement in pain were 45%, 41%, and 41%, respectively; there was a small, yet significant, impact of baseline severity (OR, 1.04; 95% CI, 1.03–1.05; $P < .0001$) that did not meet clinical relevance criteria. Both end points revealed strong construct validity and detected "minimally clinically important differences" in symptoms. Both provided better discriminant spread in IBS-D than IBS-C. **CONCLUSIONS: Both the traditional binary and 50% improvement end points are equivalent in their psychometric properties. Neither is impacted by baseline severity, and both demonstrate excellent construct validity. They are optimized for the IBS-D population but also appear valid in IBS-C.**

There is debate about how best to measure patient-reported outcomes (PROs) in irritable bowel syndrome (IBS). This debate is important because IBS remains a patient-reported condition that cannot yet be reliably diagnosed or monitored with biomarkers alone; patient reports are essential. In the absence of valid and reliable biomarkers to accurately substratify patients within an otherwise heterogeneous condition, clinicians and investigators are left interpreting patient-reported symptoms to determine the diagnosis, gauge overall disease severity, develop rational treatment plans, and assess outcomes.

This challenge is now front and center for clinicians, investigators, and regulatory agencies such as the US Food and Drug Administration. The charge for all stakeholders is to identify one or more PRO measures that are sufficiently reliable and valid, both for clinical trials and clinical practice. An optimal PRO measure must be easily administered, able to discriminate between important patient subgroups and disease states in a statistically significant and clinically relevant manner, predictable in behavior when tracked with other indicators of illness severity, not conditional on baseline severity, and readily interpretable.[1]

Most all of the recent high-quality clinical trials in IBS employed a binary PRO end point, such as "adequate relief," "satisfactory relief," or "considerable relief."[2] These end points have 2 levels and, therefore, provide a dichotomous stratification of responder status (yes/no relief). Binary end points are useful because they are easy to administer and straightforward to interpret.[2,3] Moreover, they have been retrospectively shown to have construct validity when compared against other end points in IBS,[4] and they have predicted effectiveness of medications and improvement in quality-of-life measures.[5] On this basis, recent systematic reviews of the published

Abbreviations used in this paper: AR, adequate relief; C, constipation prominent; CI, confidence interval; D, diarrhea prominent; HRQOL, health-related quality of life; IBS, irritable bowel syndrome; IBS-SSS, IBS severity symptom score; LOCF, last observation carried forward; MCID, minimal clinically important difference; OR, odds ratio; PRO, patient reported outcome.

0016-5085/09/$36.00
doi:10.1053/j.gastro.2009.08.047

evidence of IBS end points support the use of binary end points as standard for IBS clinical trials.[2,3]

However, binary end points have been criticized on several grounds. First, Whitehead et al observed that "satisfactory relief" of bowel symptoms—a type of binary end point—appeared to be confounded by baseline IBS severity.[6] The authors found that patients with severe IBS who received usual care in a health maintenance organization were less likely to achieve a response over time compared with those with less severe IBS. In contrast, patients with less severe IBS were most likely to achieve "satisfactory relief" over time but revealed no improvements in symptom severity. Although other investigators have not confirmed these findings in different IBS populations exposed to either pharmacologic or behavioral therapy in clinical trials,[7–9] the results from Whitehead et al[6] suggest that the performance of binary end points might partly depend on baseline severity. In theory, a reliable PRO measure should not be conditional on baseline severity. However, evaluation using the "adequate relief" end point in a randomized, placebo controlled trial did not reveal responses to be sensitive to patient baseline severity.[2] Second, it has been argued that binary end points may not detect minimally clinically important differences (MCIDs) in symptoms or health-related quality of life (HRQOL) and do not provide enough resolution to detect small changes in health status over time. Third, it is claimed that the binary end points lack sufficient capacity to track key illness domains or successfully discriminate between clinical subgroups or disease states. Finally, these end points were not derived from patient focus groups—the "gold standard" approach for developing end points.[1]

The use of a multi-item symptom questionnaire is an alternative to binary end points.[2] However, only 1 IBS-specific symptom severity questionnaire has been shown to be responsive to treatment effects: the Irritable Bowel Syndrome Symptom Severity Scale (IBS-SSS).[10,11] Whitehead et al have shown that the IBS-SSS is not conditional on baseline severity and proposed a "50% improvement" criterion for establishing a responder, in which patients improving by at least 50% from their baseline severity score are considered to have clinically improved.[6] When using the 50% improvement criterion, Whitehead et al found that response to usual care was not dependent on baseline severity and, therefore, proposed that the 50% improvement definition may be superior to binary end points, eg, "adequate relief" or "satisfactory relief."[6]

In light of this background, we performed a large pooled analysis of patient-level data from existing clinical trials to evaluate and compare the psychometric properties of end points used in IBS trials, including binary responses and 50% improvement in symptom severity. We retrospectively tested the properties of these end points by pooling patient-level data from 12 trials including over 10,000 patients. The specific aims were to determine whether either type of PRO end point is influenced by baseline IBS severity and whether any of the responder definitions can detect MCIDs in cardinal IBS bowel symptoms, patient reported visceral sensitivity, disease-targeted HRQOL, psychologic distress, and work productivity, all components of the evolving model of IBS illness severity.[12,13]

Patients and Methods

Pooling Patient Level Data From Databases of IBS Clinical Trials

Prior to conducting psychometric analyses of IBS end points, we first sought to pool patient-level data from available trials. This allowed the opportunity to maximize the robustness and explanatory power of its findings and to test whether those findings can be generalized across data sets. The sections below describe the steps followed to systematically acquire, evaluate, and harmonize data from existing clinical trials.

Data Acquisition

We identified pharmaceutical companies that have previously conducted randomized, controlled clinical trials in IBS. We contacted each company and provided documentation describing the study objectives and proposed analyses. Six companies provided evaluable patient-level trial data, including AstraZeneca (1 study), GlaxoSmithKline (2 studies), Ironwood (1 study), Novartis Pharmaceuticals (5 studies), and Solvay (3 studies). Table 1 provides an overview of the data employed for this pooled analysis. There were 10,066 subjects in the combined data set, of which 9044 had minimum required data to be eligible for inclusion in our analyses (ie, baseline and follow-up bowel symptom scores, and end-of-study or "last observation carried forward" [LOCF] binary response, as described further below).

Initial Data Management and Mapping

We developed a list of content areas represented in the included trials that would serve as the focus for data analysis. We then grouped the content areas into major and minor domains, as displayed in Table 2. We created a single harmonized extract data file that transformed the various data structures and variables into a common format across each study. We retained both baseline and study end point variables for the analyses. We relied on precalculated LOCF variables, when available, for patients unable to complete the full study duration.

Standardizing of Variables

There are important challenges and potential barriers to successfully harmonizing data from several studies. These include disparate binary outcome measures (eg, "satisfactory relief" vs "adequate relief" vs "considerable relief"), measures of IBS severity (eg, "pain severity" vs

Table 1. Overview of Studies Included in Endpoints Meta-Analysis

Company	Number of data sets	Number of studies	Study name(s)	Study treatment duration	Study population	No. of cases
Novartis	3	5	A0301	12 Wk	IBS-C M+W,	881
			A0351	12 Wk	IBS-C M+W,	799
			A0358	4 Wk	IBC-C W,	1519
			A2306	4 Wk	IBS-C W,	2660
			A2417	4 Wk	IBS-C/A W	661
Glaxo SmithKline	28	2	SB-22341 2/067	8 Wk	IBS M+W	
			SB-22341 2/068	8 Wk	IBS-C	
Astra Zeneca	11	1	AZ37371	12 Wk	IBS-C, -D, -A M+W	402
Solvay	24	3	S3006	12 Wk	IBS-D	711
			S3009	26 Wk	IBS-D	805
			S3011	12 Wk	IBS-D	193
Ironwood	1	1	103-202	12 Wk	IBS-C	85
Total		12				10,066

A, alternating; M, men; W, women.

IBS-SSS), and scales for all covariates (eg, 5-point vs 7-point Likert scales). To allow for cross-trial analysis, we converted each of the continuous variables to a common scale. This involved the following steps:

Table 2. Variables Included in Data Sets, Grouped by Major and Minor Domains

Major domain	Minor domain
Patient characteristics	Age
	Sex
	IBS subtype
	Treatment status
	IBS duration
	Study length
Bowel symptoms	Bloating
	Pain
	Frequency
	Consistency
	Urgency
	Hard stool
	Incomplete defecation
	Straining
Quality of life	Overall HRQOL
	Activity interference
	Dysphoria
	Food avoidance/diet
	Health worry
	Body image
	Relationship
	Sex
	Social
	Sleep
	Fatigue
Psychologic	Anxiety
	Depression
	Visceral anxiety
Work productivity	Presenteeism
	Absenteeism
	Overall impairment
Binary outcomes	Global adequate relief
	Pain relief
Global improvement	Continuous global improvement

NOTE. The scaling and instrumentation for each minor domain varied considerably from study to study. Refer to the text for the method of harmonization across studies.

1. Within each study, calculation of the mean and standard deviation of each *baseline* measure.
2. Use of these values to transform the variable to a standardized z score (eg, $(z = x - \mu)/\sigma$).
3. For ease of interpretation, recentering of the distribution to a mean of 100 and a standard deviation of 10—a variation of the traditional T scale (eg, T = (z*10) + 100).
4. Application of the same steps that were used to transform the baseline score to the study end points and LOCF versions of the respective variables. By using the baseline values, any absolute changes from baseline to end of study would be retained in the scale transformation.

Calculation of MCID Scores

For binary end points, responses of satisfactory relief, adequate relief, or considerable relief are presumed to reflect clinically significant outcomes. Although some outcome measures have empirically derived MCIDs (eg, the MCID of the IBS-QOL is between 10 and14),[14] most linear end points have no established MCID benchmarks. In the absence of empiric MCID definitions, one established technique recommended by Norman et al is to assume that a half standard deviation (SD) improvement (effect size of 0.5) equates with meaningful change.[15] This approach is based on the "remarkable universality" of a half SD as a surrogate measure for clinical importance and correlates with a "medium" effect size using the traditional rules of Cohen.[16] The half SD technique has been used by previous investigators in IBS, including the developers of the IBS-QOL instrument.[14] The Norman approach allows for standardization of an MCID definition across disparate measures and thus serves as an "exchange currency" to pool various outcome measures and covariates in the same analysis.

We also calculated MCIDs for all linear variables and assigned end-of-study MCID status for each domain in each patient. To do this, we first calculated a "baseline to study end point change score" for each domain, using

subjects with harmonized scale scores at both time points. We evaluated whether the size of the change score *exceeded one-half standard deviation*, which was 5 given a standard deviation of 10 on the harmonized T scale.[15]

Construction of Binary End Point

A version of the binary end point was constructed for each subject using the result reported in the original trial (eg, "adequate relief"). Although the various included end points had different wording, they all shared in common a binary response scale. For purposes of meta-analysis, we assumed that patients understood a similar meaning from the various questions (eg, "adequate relief" vs "considerable relief" [yes, no]) to allow harmonization across binary end points. This assumption, although arguable, allows for construction of a large harmonized database; meta-analysis inevitably requires the assumption that combined data are sufficiently alike to allow for harmonization. If a subject was missing the end of study binary end point data, we substituted the LOCF version of the measure. In this manner, we were able to include patients who did not complete the full study but nonetheless contributed data in the form of an LOCF data point.

Creation of a Harmonized Pain Severity Scale

Because all the studies included a measure of abdominal pain at baseline and follow-up and because pain is one of the cornerstones of "IBS severity,"[12,13] we adopted pain as our surrogate for IBS illness severity. All studies provided data regarding abdominal pain, thus providing an opportunity to create a harmonized "severity" scale based on pain. As with the other harmonized T scales, we created a harmonized pain severity score, with a mean of 100 and a standard deviation of 10 (higher scores indicated higher severity).

Impact of Baseline Severity on Performance of End Points of Interest

We assessed the relationship between baseline severity and end-of-study response status using 2 competing "responder" definitions: (1) harmonized binary response status and (2) 50% improvement in severity. We defined a "50% responder" as someone who reported at least a 50% reduction in IBS pain severity over time on the harmonized pain scale described above, using the baseline severity score as the reference point.

Using the harmonized pain severity scale, we divided patients at baseline into 3 severity levels, as defined by tertiles from the harmonized baseline severity T scale (ie, mild = bottom tertile, moderate = middle tertile, severe = top tertile). We then measured end-of-study binary response status stratified across the 3 severity groups to determine whether each end point was influenced by baseline IBS severity.

To answer the question whether each end point has utility in assessing treatment response across varying levels of baseline IBS symptom severity, we repeated the analyses within treatment subgroups, first limiting to treatment arms only, then to placebo arms, and then to all data combined. For all analyses, we compared response status across tertiles using χ^2 test and adopted a P value $< .05$ as our definition of statistical significance.

Consideration of Clinical Relevance

Because the sample size is large, we anticipated that some differences might be statistically significant but not clinically relevant. To estimate the clinical relevance of differences across responder groups, we performed a separate set of analyses for each responder definition in which we measured mean baseline severity in responder vs nonresponder groups using t tests. Because the MCID on the harmonized severity scale was set at 5, any difference between groups less than 5 indicated sub-MCID differences and was deemed to be of small to nonexistent clinical significance. For between-group differences in subjects receiving investigational treatment, the MCID was assessed for significance using the rules of Cohen.[17]

Regression Models

We performed a series of multivariable logistic regression models to measure the independent effect of baseline IBS severity on end-of-study response status for each of the 3 response definitions. These models adjusted for IBS subtype, treatment status, age, sex, and disease duration.

Prospective Analysis of Construct Validity of End Points of Interest

We performed a series of prospective construct validity analyses to measure the performance of the competing responder definitions. For these analyses, we measured the ability of each responder definition to track with several clinically important IBS constructs, including the cardinal bowel symptoms (abdominal pain, bloating, stool frequency, stool form, urgency, incomplete evacuation, straining), patient reported visceral sensitivity (using the visceral sensitivity index, a 15-item scale that is a reliable and valid measure of gastrointestinal symptom-specific anxiety[18]), HRQOL, and work productivity.

For each responder definition (eg, binary end point or 50% improvement), we assessed construct validity by conducting a series of t tests to compare *change scores* in each construct stratified by response status. We calculated the P value for each comparison (responder vs nonresponder), adopting $P < .05$ as evidence for statistical significance. However, because of the large sample size, we also adopted a measure of clinical relevance, that is achievement of an MCID using the 0.5 SD definition.[15]

For this set of analyses, we calculated the proportion of patients achieving an MCID over time for each construct stratified by response status. Because the results might vary by IBS subtype, we repeated the analyses in IBS constipation-prominent (IBS-C) and IBS diarrhea-prominent (IBS-D) subgroups and report the results both in combination and separately.

Results

Patient Characteristics

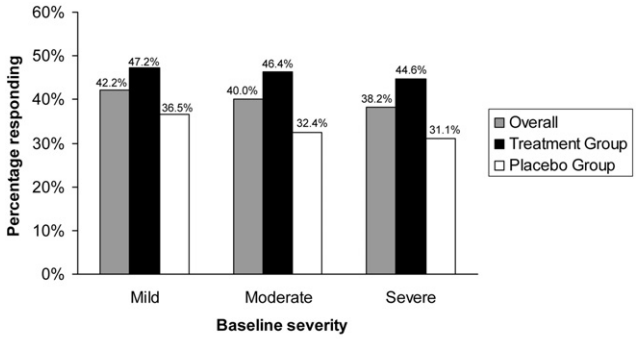
There were 9044 evaluable subjects in the 12 trials. The mean age was 44.3 years, and 85% were female. More than half of the sample had IBS-C (58%) and 31% IBS-D. Fifty-three percent of the cohort had received an investigational IBS treatment, whereas the rest received placebo. Using the binary response definition, 60% of the overall cohort achieved a response at the end of the study follow-up period. Table 3 displays the key descriptive statistics of patients included in the sample.

Effect of Baseline Severity on Responder Status for the Binary Responses

There were 8457 subjects with data included in the analyses with data comparing binary response status with baseline severity. Figure 1 portrays the percentage of patients responding, using the binary definition, stratified by baseline severity tertiles. The data are provided for 3 groups: all subjects combined, subjects receiving investigational treatment, and subjects receiving placebo.

The proportion achieving a response in the mild, moderate, and severe groups was 42%, 40%, and 38%, respectively. Because of the large sample size, these differences were highly significant for both the overall group (P = .0008) and placebo subgroup (P = .009) but not for the investigational treatment group (P = .36) using a χ^2 test. However, multivariable analysis using data from both groups and adjusting for age, IBS subtype, sex, disease duration, and baseline pain as a continuous variable (n = 5510 for model) found a *non*-significant relationship between baseline pain severity and response status (odds ratio [OR], 0.995; 95% confidence interval [CI]: 0.99–1; P = .07).

Table 4 provides the results of the t test comparing mean baseline severity scores in patients with versus without end-of-study (or LOCF) binary response, stratified by treatment status. The data reveal that the absolute differences in baseline severity by binary status were small for each group. For example, the absolute difference in the combined group was 0.6 points, which is below the a priori MCID of 5 points, and indicating that the differ-



P-Values for Above Figure:

Group	Chi-Squared *P* Value
Overall Group	.0008
Placebo Group	.009
Treatment Group	.36

Figure 1. Relationship between trichotomized baseline severity and responder status, defined using harmonized binary end point (aka, "adequate relief"). Data are provided for the overall, treatment, and placebo groups.

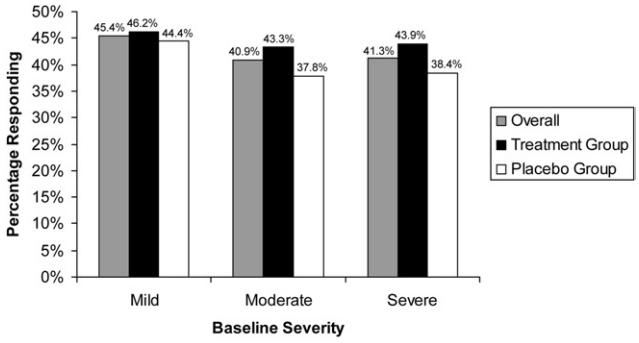
Table 3. Subject Descriptive Statistics From Harmonized Patient-Level Data Set of IBS Clinical Trials

Variable	Aggregate mean	N
Sex (% Female)	84.8	9044
Age, y	44.3 ± 12.7	9044
Duration of IBS, y	11.3 ± 11	6090
IBS subgroups		9044
% IBS-Constipation	57.9	
% IBS-Diarrhea	31.4	
% IBS-Other	10.7	
Treatment groups		9044
% Receiving treatment	53.4	
% Receiving placebo	46.6	
Trichotomized baseline severity		8457
% Mild (bottom tertile)	34.5	
% Moderate (middle tertile)	33.5	
% Severe (top tertile)	31.9	
End-of-study responder status		
% Responder	59.8	9044
% Nonresponder	40.2	

Table 4. *t* Test Comparing Baseline Severity Between Responders vs Nonresponders

	Mean severity score (±SD) in responders	Mean severity score (±SD) in nonresponders	*P* value for *t* test
Binary responder data			
Treatment and placebo groups combined	99.6 ± 10	100.2 ± 9.9	.01
Treatment group alone	99.8 ± 9.9	100.2 ± 10.0	.70
Placebo group alone	99.2 ± 10.2	100.2 ± 9.9	.01
50% Improvement in symptom severity			
Treatment and placebo groups combined	99.5 ± 10.1	100.2 ± 9.7	.03
Treatment group alone	99.4 ± 10.1	97.3 ± 10.0	<.001
Placebo group alone	99.8 ± 9.9	100.1 ± 9.8	.24

NOTE. The top 3 rows provide data stratified by the harmonized binary response end point, and the lower 3 rows provide data stratified by 50% improvement in symptom severity. See text for details.



P-Values for Above Figure:

Group	Chi-Squared *P* Value
Overall Group	.002
Placebo Group	.002
Treatment Group	.27

Figure 2. Relationship between trichotomized baseline severity and responder status, defined using 50% improvement from baseline severity. Data are provided for the overall, treatment, and placebo groups.

ences were not clinically significant for any of the subgroups.

Effect of Baseline Severity on Responder Status 50% Improvement Responder Status

There were 7487 subjects with data comparing 50% improvement in pain severity response status with baseline severity. Figure 2 portrays the percentage of patients responding, using the 50% improvement definition, stratified by baseline severity tertiles. The proportion achieving 50% improvement in the mild, moderate, and severe groups was 45%, 41%, and 41%, respectively. The χ^2 *P* value was highly significant for both the overall group and placebo groups ($P = .002$ for both) but not for the treatment group ($P = .27$). In multivariable logistic regression analysis with baseline pain as a continuous predictor and adjusting for potential confounders, the relationship between baseline severity and 50% improvement status was statistically significant (OR, 1.04; 95% CI: 1.033–1.047; $P < .0001$).

Although baseline severity independently predicted end-of-study 50% improvement in severity, the relationship was numerically small. To test for clinical relevancy, we performed a *t* test comparing mean baseline severity scores in patients with versus without 50% improvement, stratified by treatment status (Table 4). The absolute differences in baseline severity by 50% improvement status were small for each group. For example, the absolute difference in the combined group was 0.7 points, which is below the a priori MCID of 5 points, and indicates that the differences were not clinically significant for any of the subgroups. The between-group difference (ie, patients with vs without 50% improvement) among subjects receiving investigational treatment was 2.1 points or an effect size of 0.21—also below an MCID using the rules of Cohen.[17]

Prospective Construct Validity of Binary Response and 50% Improvement

The prospective construct validity of the binary end point was first tested by performing a series of *t* tests to compare mean *difference in difference* (DID) scores for each symptom or construct in Table 2, stratified by binary response status. The full results of the 23 *t* tests are presented in the Appendix. All of the *t* tests yielded *P* values $< .0005$, indicating that the binary response end point was able to discriminate between groups for all 23 tested constructs. Using the 50% improvement end point, the same analysis found that all *t* tests yielded *P* values $< .007$. Therefore, both the binary and 50% improvement end points were able to discriminate between groups for all tested constructs.

To measure the clinical relevance of these results, we overlaid a 0.5 SD benchmark for each *t* test, as described in the Patients and Methods section. Figure 3 portrays the absolute difference in difference across all 23 constructs, placing the results of the 2 end points side by side. The *dashed line* marked the threshold for clinical relevancy. Visual inspection of Figure 3 reveals that both end points performed almost equivalently in terms of achieving clinically relevant separation between groups. The largest difference in difference scores was achieved in separating patients by abdominal bloating and pain, with the 50% improvement definition achieving numerically higher discriminant validity than the binary response definition. Both end points achieved clinically relevant differences for stool consistency, urgency, hard stool, overall HRQOL, body image, fatigue, visceral sensitivity index scores, presenteeism, and overall work productivity. In contrast, neither end point clinically separated groups by stool frequency, incomplete evacuation, straining, activity interference, diet, relationships, sexual function, social function, sleep, or absenteeism.

In addition, we also measured clinical relevancy by calculating the proportion of patients in each group achieving an MCID, using the 0.5 SD metric for each variable. Figure 4*A* portrays the results in the IBS-C group. In this group, both end points provided maximal discriminant ability for bloating, with lower discriminant ability of both end points for all other symptoms in the IBS-C group, with the smallest values in the stool frequency comparisons. Compared with the binary response end point, the 50% improvement definition achieved a higher numerical spread between groups for bloating, urgency, hard stool, incomplete evacuation, and straining. In contrast, the binary response end point achieved slightly improved discriminant ability for stool frequency and consistency.

Figure 4*B* portrays the results in the IBS-D group. The data reveal that both end points provided maximal discriminant ability for bloating and urgency. Unlike in the IBS-C subgroup, the discriminant ability of both end points remained sizeable (20% spread or higher) for all

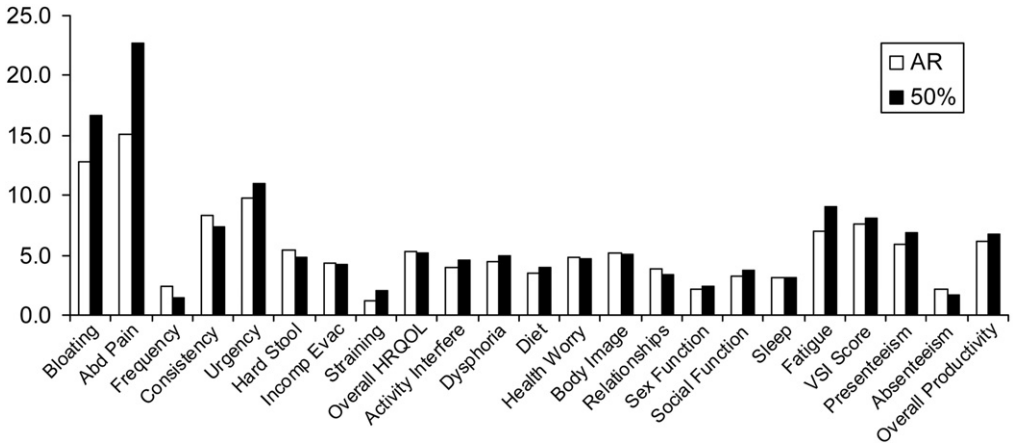


Figure 3. "Difference in difference" (DID) scores stratified by response status. The Figure depicts 2 series of data: 1 for adequate relief (*AR*) and 1 for 50% improvement in severity. The data are stratified by 23 variables, including bowel symptoms, HRQOL domains, and work productivity domains. Each *bar* reveals the mean DID scores for each variable between responders and nonresponders, stratified by responder definition. The higher the DID the better the discriminant validity. For example, using the AR response definition, the mean DID abdominal pain score between responder groups was 15. In contrast, using the 50% improvement definition, the mean DID score between groups was 22.5. Both *bars* exceed the threshold for "clinical significance" depicted by the *dashed line* at 5 points (ie, half standard deviation MCID definition).

symptoms except straining, which itself is not a cardinal symptom of IBS-D and therefore of unclear significance in this patient subgroup.

Discussion

This pooled analysis was motivated by questions regarding the validity of traditional IBS end points, with particular focus on binary end points.[6] The Rome Foundation Outcomes and Endpoints Committee combined data from over 9000 patients from 12 randomized controlled drug trials involving 5 separate investigational treatments with different mechanisms of action. Our goal was to leverage the power of this harmonized database to explicitly test key psychometric properties of binary end points and to compare the performance of binary end points with the "50% improvement" criterion suggested as an alternative metric.[6]

Regarding the impact of baseline severity on end point performance, we found that the relationship between severity and binary responder status was not statistically significant in the adjusted analysis, and the relationship did not meet criteria for clinical significance, regardless of the resulting *P* values. In contrast, we found that the "50% improvement" criterion for pain severity was significantly associated with baseline severity in the adjusted analysis, particularly for patients receiving investigational treatment. However, the clinical relevance of the relationship with baseline severity was minimal across all treatment groups. The observation that there is *not* an important relationship between baseline pain severity and response status, either defined with a binary end point or "50% improvement" criterion, is consistent with previous studies[7–9] and contrary to the results of Whitehead et al.[6] It is possible that the analyses reported in a community sample by Whitehead et al included subjects from a health maintenance organization with a broader range of IBS symptom severity than the subjects included in the clinical trials summarized in this pooled analysis.

We further tested the construct validity of both end points against a range of IBS illness domains. In short, we found that both the binary response and the 50% improvement end points reveal excellent construct validity across a wide range of variables (Figure 3). Both end points are able to detect MCIDs in key bowel symptoms, including bloating, abdominal pain, consistency, urgency, and hard stool. They are also able to detect MCIDs in worker productivity, visceral hypersensitivity scores, and fatigue scores. Whereas both end points are able to detect MCIDs for overall HRQOL, they are less capable of detecting MCIDs for the individual HRQOL components. Thus, both end points track with key components of IBS illness severity, neither is clearly superior over the other, and both work as expected.

We found that both the binary response and the 50% improvement end points performed similarly in discriminating between MCID responders and nonresponders for bowel symptoms in IBS subgroups. Of note, both end points appear to provide better discrimination in the IBS-D than IBS-C subgroups (Figure 4*A* and *B*). This has potential implications for studies that seek to establish differences in response rates between treatment and placebo groups. The data suggest that both end points may be better suited for the IBS-D population. In IBS-C patients, the percentage achieving an MCID is numerically smaller, suggesting that more sensitive end points might be necessary for the IBS-C groups. A corollary is that drugs failing to show large effect sizes in this population might have been hampered by the psychometric properties of the binary end points. Further research should aim to test current and future end points in both IBS-C and

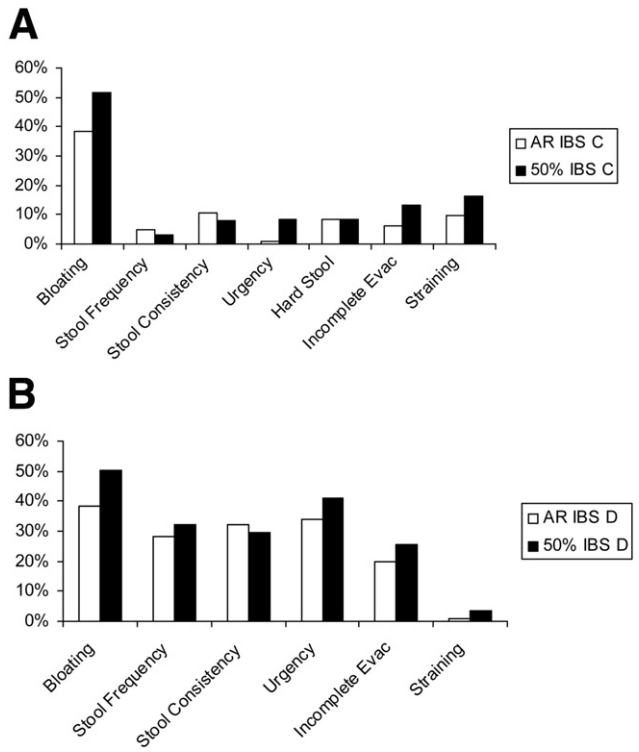


Figure 4. Proportion of patients achieving an MCID for individual bowel symptoms: AR vs 50%. *Panel A* provides data in the IBS-C subpopulation, and *panel B* provides data in the IBD-D subpopulation. Each *bar* represents the results of an individual 2 × 2 table and depicts the difference in MCID achievement between responders and nonresponders. For example, 38% more IBS-C patients in the binary response group achieved an MCID in bloating vs those not achieving a binary response. In contrast, 52% more IBS-C patients in the 50% improvement group achieved an MCID in bloating vs those not achieving 50% improvement.

IBS-D subgroups and to establish whether the psychometric properties are similar or different in these phenotypically distinct populations. It is possible that a "one size fits all" approach to end points may not apply in IBS: different subgroups may be better captured with tailored end points. This finding raises the question of whether clinical trialists should employ different end points for IBS-C vs IBS-D. This would represent a notable change in our approach to end point measurement in IBS. Our study is unable to determine why end points may behave differently by subgroup; instead, it merely raises the question. Future research should aim to understand why this might be. In the meantime, our finding suggests that further research should carefully evaluate end point performance in both groups separately.

These data add to previous conclusions that global binary end points are useful in IBS,[2] based on the collective clinical trial experience in almost 20,000 IBS patients with at least 5 different medications (alosetron, cilansetron, tegaserod, lubiprostone, dextofisopam) tested with binary end points. Binary end points have been devalued given the relative lack of psychometric validation until now. However, even before this pooled analysis, previous investigators demonstrated that binary end points were acceptable to patients and that binary responses were driven by the patients' most bothersome symptom.[19,20] Based on a systematic review of 12 prespecified criteria, Bijkerk et al concluded that the weight of evidence was in favor of using "adequate relief"—a binary end point—among the different available end points used in IBS trials.[3] Drugs that are effective, based on the binary response end points, were also found to improve general or disease-specific quality of life.[5] Based on these collective data, the Rome III guidance on IBS clinical trials endorsed using a global measure that integrates the symptom data into a single numerical index, measured either as a binary end point or a continuous integrative symptom questionnaire such as the IBS-SSS.[11] We have now expanded and confirmed these collective results and conclusions by demonstrating excellent construct validity of the binary end points with a wide range of patient symptoms, psychosocial illness experiences, visceral sensitivity reporting, HRQOL, and even work productivity. Moreover, we have found that the performance of a binary end point is psychometrically equivalent to monitoring pain severity on a continuous scale and adopting the "50% improvement" criterion recommended by Whitehead et al.[6]

Based on our data, coupled with extensive preexisting data supporting the validity of the binary end points, it is reasonable to conclude that use of binary end points in IBS clinical trials is rational and valid. No end point can be fully validated; establishing the validity of a PRO is an ongoing and iterative effort. However, our results add to this effort and further confirm that binary end points get the job done—they work as expected. This is an important conclusion because it supports the validity of existing studies, highlights the efficacy of therapies originally tested in trials employing binary end points, and indicates that future studies could also use these end points without undue concern.

Our study has several strengths. First, the sample size of this analysis is large, and the use of pooled patient-level data is a more powerful method of synthesizing multiple studies than conventional meta-analysis. This provides considerable power to investigate the psychometric properties of IBS end points. Second, because we are cognizant that large sample sizes can yield statistically significant relationships that are not clinically relevant, we overlaid a priori criteria for clinical relevance and reported results that were both statistically significant and clinically relevant. Third, we conducted subanalyses across key groups, including IBS subgroups (ie, IBS-C vs IBS-D) and treatment groups (active vs placebo). This allows us to generalize our results across different populations. Finally, we measured a range of key psychometric properties using multiple clinical anchors. This allows

us to triangulate the validity of the end points from several perspectives.

Our study has limitations. First, as with any meta-analysis, we were faced with combining disparate data from different studies, each with unique inclusion and exclusion criteria, disease characteristics, and end point evaluations. However, we have been careful to acknowledge these variations, as described in our Patients and Methods section, and have attempted to balance the power of harmonizing large data sets with the inevitable methodologic shortcomings of combining disparate data. Second, it is possible that patients in randomized controlled trials are systematically different from other populations of IBS patients. However, this is precisely the population in question because the current main use of PRO measures is for clinical trials to test the effect of pharmacologic interventions in IBS. As PROs continue to penetrate into everyday clinical practice, further validation studies will be necessary in nonclinical trial populations. Third, our measure of "IBS severity" was limited to "pain severity." We were unable to employ multiattribute severity scales like the IBS-SSS because there were inadequate data for this purpose. However, pain is a cardinal symptom of IBS,[12,13] and it drives overall illness severity more, on average, than any other symptom. In short, there is sufficient rationale and precedent to use pain severity as a surrogate for overall IBS illness severity, as we have done here.

In conclusion, this large patient-level meta-analysis reveals that both the binary and 50% improvement end points are equivalent in their psychometric properties. Neither is impacted by baseline severity, and both demonstrate excellent construct validity. They appear optimized for the IBS-D population but are also valid in IBS-C.

Supplementary Data

Note: To access the supplementary material accompanying this article, visit the online version of *Gastroenterology* at www.gastrojournal.org, and at doi: 10.1053/j.gastro.2009.08.047.

References

1. Burke LB, Kennedy DL, Miskala PH, et al. The use of patient-reported outcome measures in the evaluation of medical products for regulatory approval. Clin Pharmacol Ther 2008;84:281–283.
2. Camilleri M, Mangel AW, Fehnel SE, et al. Primary end points for irritable bowel syndrome trials: a review of performance of endpoints. Clin Gastroenterol Hepatol 2007;5:534–540.
3. Bijkerk CJ, de Wit NJ, Muris JW, et al. Outcome measures in irritable bowel syndrome: comparison of psychometric and methodological characteristics. Am J Gastroenterol 2003;98:122–127.
4. Mangel AW, Hahn BA, Heath AT, et al. Adequate relief as an endpoint in clinical trials in irritable bowel syndrome. J Int Med Res 1998;26:76–81.
5. El-Serag HB, Olden K, Bjorkman D. Health-related quality of life among persons with irritable bowel syndrome: a systematic review. Aliment Pharmacol Ther 2002;16:1171–1185.
6. Whitehead WE, Palsson OS, Levy RL, et al. Reports of "satisfactory relief" by IBS patients receiving usual medical care are confounded by baseline symptom severity and do not accurately reflect symptom improvement. Am J Gastroenterol 2006;101:1057–1065.
7. Leventer SM, Raudibaugh K, Frissora CL, et al. Clinical trial: dextofisopam in the treatment of patients with diarrhoea-predominant or alternating irritable bowel syndrome. Aliment Pharmacol Ther 2008;27:197–206.
8. Lackner JM, Jaccard J, Krasner SS, et al. Self-administered cognitive behavior therapy for moderate to severe irritable bowel syndrome: clinical efficacy, tolerability, feasibility. Clin Gastroenterol Hepatol 2008;6:899–906.
9. Ameen VZ, Health AT, McSorley D, et al. Global measure of adequate relief predicts clinically important difference in pain and is independent of baseline pain severity in irritable bowel syndrome (abstract 941). Gastroenterology 2007;132:A140.
10. Francis CY, Morris J, Whorwell PJ. The irritable bowel severity scoring system: a simple method of monitoring irritable bowel syndrome and its progress. Aliment Pharmacol Ther 1997;11:395–402.
11. Irvine EJ, Whitehead WE, Chey WD, et al. Design of treatment trials for functional gastrointestinal disorders. Gastroenterology 2006;130:1538–1551.
12. Lembo A, Ameen VZ, Drossman DA. Irritable bowel syndrome: toward an understanding of severity. Clin Gastroenterol Hepatol 2005;3:717–725.
13. Spiegel B, Strickland A, Naliboff BD, et al. Predictors of patient-assessed illness severity in irritable bowel syndrome. Am J Gastroenterol 2008;103:2536–2543.
14. Drossman D, Morris CB, Hu Y, et al. Characterization of health related quality of life (HRQOL) for patients with functional bowel disorder (FBD) and its response to treatment. Am J Gastroenterol 2007;102:1442–1453.
15. 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;41:582–592.
16. Cohen J. Statistical power analysis for the behavioral sciences. London: Academic Press, 1969.
17. Cohen J. Statistical power analysis for the behavioral sciences. Lawrence Erlbaum Associates, Inc, 1988.
18. Labus JS, Mayer EA, Chang L, et al. The central role of gastrointestinal-specific anxiety in irritable bowel syndrome: further validation of the visceral sensitivity index. Psychosom Med 2007;69:89–98.
19. Fehnel S JJ, Kurtz C, Mangel A. Assessing global change and symptom severity in subjects with IBS: qualitative item testing. Am J Gastroenterol 2006;101:S483.
20. Gordon S, Ameen V, Bagby B, et al. Validation of irritable bowel syndrome Global Improvement Scale: an integrated symptom end point for assessing treatment efficacy. Dig Dis Sci 2003;48:1317–1323.

Received April 29, 2009. Accepted August 12, 2009.

Reprint requests

Address requests for reprints to: Brennan M.R. Spiegel, MD, MSHS, 11301 Wilshire Boulevard, Building 115, Room 215, Los Angeles, California 90073. e-mail: bspiegel@mednet.ulca.edu; fax: (310) 268-4510.

Acknowledgments

The Rome Foundation thanks the participating pharmaceutical companies for their donated time and willingness to contribute data to this effort.

The authors thank Carlar Blackman of the Rome Foundation for her administrative and logistical support of this research project and Dr Douglas Drossman for his stewardship of the Rome Foundation and the pharmaceutical companies that donated their time and data to assist the Working Group in successfully completing this project.

Drs Emeran Mayer and Hashem El-Serag are nonauthor members of the Rome Foundation Endpoints Working Group.

Participating organizations: The paper is the result of the Rome Foundation Outcomes and Endpoints Working Group (chair, M. Camilleri, MD) whose members (listed as authors) are responsible for the conduct, analysis, and interpretation of the data. The following companies provided data in kind for this meta-analysis: AstraZeneca, GlaxoSmithKline, Ironwood, Novartis Pharmaceuticals, Rotta Pharmaceuticals, and Solvay. The data from Rotta Pharmaceuticals lacked necessary elements for purposes of our analysis and was therefore unable to be employed in this study. The authors and Rome Foundation maintained complete control of all data and analyses.

The opinions and assertions contained herein are the sole views of the authors and are not to be construed as official or as reflecting the views of the Department of Veteran Affairs.

Conflicts of interest

The authors disclose the following: Dr Spiegel has served as a consultant for AstraZeneca, McNeail Consumer, Novartis, Prometheus, Takeda Pharmaceuticals, and TAP Pharmaceuticals and has received grant support from Amgen, AstraZeneca, Bristol Myers Squibb, Novartis, Salix, and Takeda. Dr Camilleri has served as a consult for GlaxoSmithKline and has received research support from Ironwood Pharmaceuticals and Novartis. Drs Fehnel and Mangel are employees of RTI Health Solutions. Dr Chey is a consultant for Novartis, GlaxoSmithKline, Solvay, and Ironwood and is on the speaker's bureau of Novartis. Dr Talley is a consultant for Astellas Pharma Inc US, AstraZeneca, Centocor, Eisai, Elsevier, Ferring Pharmaceuticals, Focus, Gilead, In2MedEd, Ironwood Pharmaceuticals, McNeil Consumer, Medscape, Meritage Pharma, Metabolic Pharma, Microbia Inc, Novartis, Optum HC, Salix, SK Life Sciences, Steigerwald, *The Journal of Medicine*, Therevance, and Wyeth and received grant support from GlaxoSmithKline, Dynogen, and Tioga. The remaining authors disclose no conflicts.

Funding

Supported by the Rome Foundation and by a Veteran's Affairs Health Services Research and Development (HSR&D) Career Development Transition Award (RCD 03-179-2; to B.S.), the CURE Digestive Disease Research Center (NIH 2P30 DK 041301-17; to B.S.), and NIH Center Grant 1 R24 AT002681-NCCAM (to B.S.).

Appendix. Adequate Relief Longitudinal Construct Validity: *t*-Tests by Covariates

Variable	Adequate relief responder status	N	Mean	Standard deviation	*P* value
Bloating	Non-responder	4334	−4.675	10.608	<.0001
	Responder	3176	−17.49	13.707	
	Difference		12.819	12.016	
Abdominal pain	Non-responder	4326	−6.642	11.744	<.0001
	Responder	3171	−21.68	13.909	
	Difference		15.043	12.705	
Stool frequency	Non-responder	4334	1.8786	9.8759	<.0001
	Responder	3170	4.2653	13.005	
	Difference		−2.387	11.304	
Constipation	Non-responder	1413	−3.418	11.739	<.0001
	Responder	1010	−11.71	16.786	
	Difference		8.2953	14.064	
Urgency	Non-responder	1296	−4.201	10.481	<.0001
	Responder	920	−13.95	13.067	
	Difference		9.7497	11.625	
Hard stool	Non-responder	2804	4.9333	11.387	<.0001
	Responder	2107	10.315	11.014	
	Difference		−5.382	11.228	
Incomplete evacuation	Non-responder	1284	−1.614	10.648	<.0001
	Responder	920	−5.948	13.732	
	Difference		4.3341	12.032	
Straining	Non-responder	1347	−0.005	9.9331	.007
	Responder	950	−1.208	11.428	
	Difference		1.2032	10.577	
Overall health-related quality of life	Non-responder	2775	3.1989	6.9662	<.0001
	Responder	2023	7.7736	9.3314	
	Difference		−4.575	8.0485	
Activity interference	Non-responder	3615	3.5167	8.1522	<.0001
	Responder	2371	6.9216	9.8737	
	Difference		−3.405	8.874	
Dysphoria	Non-responder	3785	4.2787	8.1574	<.0001
	Responder	2472	8.2383	9.3926	
	Difference		−3.96	8.6664	
Food avoidance	Non-responder	3796	3.1517	8.1169	<.0001
	Responder	2479	5.9935	9.6438	
	Difference		−2.842	8.752	
Health worry	Non-responder	2775	2.9355	8.2654	<.0001
	Responder	2019	7.1689	9.8771	
	Difference		−4.233	8.9794	
Body image	Non-responder	2786	2.3543	7.498	<.0001
	Responder	2030	6.8419	9.6556	
	Difference		−4.488	8.4746	
Relationships	Non-responder	2771	2.2696	7.3864	<.0001
	Responder	2016	5.5129	8.6728	
	Difference		−3.243	7.9535	
Sexual function	Non-responder	3242	2.0386	7.3094	<.0001
	Responder	2192	3.8568	8.2903	
	Difference		−1.818	7.72	
Social functioning	Non-responder	3773	3.4002	7.8744	<.0001
	Responder	2465	6.1659	8.9809	
	Difference		−2.766	8.3292	
Sleep	Non-responder	1011	5.2631	8.99	<.0001
	Responder	452	8.2965	10.399	
	Difference		−3.033	9.4473	
Fatigue	Non-responder	158	4.7025	8.7116	<.0001
	Responder	98	11.286	9.7874	
	Difference		−6.583	9.1374	
Visceral sensitivity	Non-responder	164	2.4573	8.1567	<.0001
	Responder	99	8.5859	10.204	
	Difference		−6.129	8.9803	
Presenteeism	Non-responder	597	−2.881	9.4459	<.0001
	Responder	659	−8.83	9.6824	
	Difference		5.949	9.5708	

Appendix. Continued

Variable	Adequate relief responder status	N	Mean	Standard deviation	*P* value
Abstenteeism	Non-responder	551	0.3793	10.09	.0005
	Responder	590	−1.769	10.549	
	Difference		2.1488	10.33	
Overall work productivity	Non-responder	548	−2.464	9.1216	<.0001
	Responder	589	−8.577	9.9271	
	Difference		6.1137	9.5474	

50% improvement longitudinal construct validity: paired *t*-tests by covariates

Variable	50% improvement responder status	N	Mean	Std dev	*P* value
Bloating	Non-responder	4305	−3.01	9.0346	<.0001
	Responder	3191	−19.67	12.806	
	Difference (1–2)		16.663	10.802	
Abdominal pain	Non-responder	4305	−3.348	8.607	<.0001
	Responder	3192	−26.03	10.653	
	Difference (1–2)		22.68	9.5319	
Stool frequency	Non-responder	4301	2.243	10.049	<.0001
	Responder	3189	3.746	12.852	
	Difference (1–2)		−1.503	11.328	
Constipation	Non-responder	1359	−3.772	12.264	<.0001
	Responder	1050	−11.1	16.289	
	Difference (1–2)		7.33	14.16	
Urgency	Non-responder	1210	−3.283	9.8167	<.0001
	Responder	1006	−14.22	12.933	
	Difference (1–2)		10.938	11.338	
Hard stool	Non-responder	2827	5.2101	11.518	<.0001
	Responder	2084	9.999	10.985	
	Difference (1–2)		−4.789	11.295	
Incomplete evacuation	Non-responder	1200	−1.48	9.3918	<.0001
	Responder	1004	−5.745	14.578	
	Difference (1–2)		4.265	12.034	
Straining	Non-responder	1263	0.5154	9.9692	<.0001
	Responder	1020	−1.593	11.066	
	Difference (1–2)		2.1086	10.474	
Overall health-related quality of life	Non-responder	2444	3.8691	7.442	<.0001
	Responder	1428	9.1169	9.2618	
	Difference (1–2)		−5.248	8.1604	
Activity interference	Non-responder	3010	3.4887	8.0691	<.0001
	Responder	1882	8.0499	10.078	
	Difference (1–2)		−4.561	8.8955	
Dysphoria	Non-responder	3154	4.5672	8.1038	<.0001
	Responder	1947	9.5485	9.3654	
	Difference (1–2)		−4.981	8.6071	
Food avoidance	Non-responder	3168	3.1146	8.1299	<.0001
	Responder	1954	7.0507	9.7815	
	Difference (1–2)		−3.936	8.7965	
Health worry	Non-responder	2445	3.6908	8.5519	<.0001
	Responder	1427	8.3994	9.8451	
	Difference (1–2)		−4.709	9.0499	
Body image	Non-responder	2459	3.0378	8.0239	<.0001
	Responder	1431	8.0727	9.827	
	Difference (1–2)		−5.035	8.7304	
Relationships	Non-responder	2442	2.8771	7.8049	<.0001
	Responder	1426	6.2756	8.6758	
	Difference (1–2)		−3.398	8.1368	
Sexual function	Non-responder	2752	2.1592	7.4818	<.0001
	Responder	1681	4.6181	8.2922	
	Difference (1–2)		−2.459	7.799	
Social functioning	Non-responder	3144	3.4892	7.8522	<.0001
	Responder	1942	7.2595	9.1168	
	Difference (1–2)		−3.77	8.3576	
Sleep	Non-responder	709	4.9958	9.0361	<.0001
	Responder	522	8.1341	10.049	
	Difference (1–2)		−3.138	9.4789	

Appendix. Continued

50% improvement longitudinal construct validity: paired *t*-tests by covariates					
Variable	50% improvement responder status	N	Mean	Std dev	*P* value
Fatigue	Non-responder	130	4.9	7.7977	<.0001
	Responder	63	13.937	9.9578	
	Difference (1–2)		−9.037	8.5588	
Visceral sensitivity	Non-responder	132	2.1742	7.9199	<.0001
	Responder	64	10.313	10.593	
	Difference (1–2)		−8.138	8.8768	
Presenteeism	Non-responder	691	−2.931	9.5614	<.0001
	Responder	565	−9.759	9.2613	
	Difference (1–2)		6.8288	9.4276	
Abstenteeism	Non-responder	628	0.0207	9.4579	.007
	Responder	513	−1.653	11.352	
	Difference (1–2)		1.6737	10.352	
Overall work productivity	Non-responder	624	−2.571	9.214	<.0001
	Responder	513	−9.353	9.7053	
	Difference (1–2)		6.7823	9.4388	