Primary Endpoints for Irritable Bowel Syndrome Trials: A Review of Performance of Endpoints

MICHAEL CAMILLERI,* ALLEN W. MANGEL,† SHERI E. FEHNEL,‡ DOUGLAS A. DROSSMAN,§ EMERAN A. MAYER,¶ and NICHOLAS J. TALLEY*

*CENTER Program, Division of Gastroenterology and Hepatology, Mayo Clinic College of Medicine, Rochester, Minnesota; †Research Triangle Institute, Research Triangle Park, North Carolina; ‡Center for Functional GI and Motility Disorders, Division of Gastroenterology and Hepatology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina; and ¶Center for Neurovisceral Sciences and Women’s Health, Department of Medicine, David Geffen School of Medicine, University of California, Los Angeles, California

The choice of primary endpoint for a clinical trial is one of the most important determinants of the ability of a clinical trial to demonstrate efficacy of therapeutic agents. Although there are still no clear, universally accepted guidelines on the definition of clinical benefit for irritable bowel syndrome (IBS), consensus guidelines stress the importance of using validated endpoints. This article reviews the evidence available in the literature on the psychometric validation and performance of the 3 endpoints recommended by the Rome III Committee for use as primary endpoints in treatment trials of IBS. The Rome III Committee recommends 2 types of measures: binary endpoints addressing the construct of relief (that is, adequate relief and satisfactory relief) and an integrative symptom questionnaire that addresses the change in severity of a representative group of symptoms of IBS (that is, the IBS Severity Scale). The current evidence suggests that at present, adequate relief should be recognized by regulatory authorities as an acceptable primary endpoint in clinical trials. This analysis also suggests that data from individual clinical trials should be pooled and undergo meta-analysis, and that prospective studies should be considered to further characterize the performance of available endpoints as outcome measures in pharmacotherapeutic trials in IBS.

The success of a clinical trial is judged by achieving a statistically significant difference on a clinically relevant primary endpoint. The magnitude of the difference needs to be indicative of a reasonable number needed to treat for the drug being evaluated. Although there are still no clear, universally accepted guidelines on the definition of clinical benefit or the choice of a primary endpoint for irritable bowel syndrome (IBS), there has been significant progress in the design of clinical trials during the last decade. Current trial designs are based on the alosetron and tegaserod clinical development programs, which were considered sufficiently robust and received United States Food and Drug Administration (FDA) approval for the treatment of IBS. However, the appropriateness of using a binary measure as a primary endpoint, such as “adequate relief of abdominal pain and discomfort,” which was used in the alosetron studies, or “satisfactory relief,” which has been used in postapproval tegaserod studies, has recently been questioned.

The Rome III guideline on clinical trials recommends the use of only validated instruments as primary outcome assessment tools in IBS trials. This recommendation is wholly consistent with the draft guidance for industry, “Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims,” which was released by the FDA in February 2006 and describes both the proper development of a patient-reported outcome (PRO) measure and the psychometric properties for which evidence must be presented if the measure is to support regulatory approval or promotional claims.

Psychometric Validation Standards

Although psychometric validation is considered an on-going process of conducting various studies to confirm various hypotheses regarding the internal structure of the construct and its relationships with other variables, validation of a symptom-based measure generally requires evidence for a basic set of properties. In particular, it must be demonstrated that the instrument addresses all symptoms that are indicative of the disorder and importance to patients (content validity); is adequately related to other measures of the same or similar concepts such as symptom improvement and symptom-free days (construct validity); produces similar results when re-administered to patients whose health status has not changed (reliability); detects clinically meaningful change in health status when such a change has occurred (responsiveness); and is associated with clinical indicators that are meaningful to clinicians (predictive or criterion validity).

In addition, the FDA guidance document emphasizes the need for patient involvement in the development of any PRO instrument both to identify constructs of interest, such as those symptoms that are most important to patients, and to minimize measurement error by ensuring that items are easily understood, interpreted similarly across patients, and offer response options that are both comprehensive and mutually exclusive. Qualitative research methods, including in-depth interviews, focus groups, and cognitive interviews, are generally conducted with patients to meet this requirement.

© 2007 by the AGA Institute
1542-3565/07/$32.00
doi:10.1016/j.cgh.2007.03.004
Recommendations of Rome III Committee on Selection of Endpoints

After appraising the wide range of endpoints proposed or used in the published literature, the general recommendations included in the Rome III guideline were summarized as follows: “Adequate relief and satisfactory relief are the current standards for primary outcome assessment in treatment trials in FGIDs. Alternative outcome measures such as integrative symptom questionnaires are also acceptable. All of these measures require additional validation.”

The Rome III Committee further commented that the only validated integrative symptom questionnaire was the IBS Severity Scale (IBS-SS).

A meta-analysis of endpoints in IBS trials is desirable but unavailable at present, and it is likely to present significant challenges because most large multicenter trials have not used the same endpoint or the same wording (eg, adequate versus satisfactory relief) to describe a given construct between trials with the same or different medications. Until such a systematic review and meta-analysis are conducted and their conclusions can be appraised in the light of the challenges mentioned, or until the variety of endpoints are prospectively evaluated, it is pertinent to evaluate the psychometric validation and performance of the main endpoints recommended by the Rome III Committee and used to date. These endpoints fall into 2 broad categories, binary endpoints assessing the construct of relief and integrative symptom questionnaires that address the severity of selected relevant symptoms.

Objective of the Review

This review appraises the relative merits and level of validation of the 3 endpoints recommended by the Rome III Committee as primary endpoints in treatment trials of IBS, adequate and satisfactory relief and the IBS-SS.

Binary Endpoints

Adequate Relief

Description. The adequate relief endpoint incorporates 2 unique features; it lets the patient integrate all relevant symptoms and normalizes the assessment to the patient’s own reference system of improvement (ie, adequate is relative to the patient’s own reference system). Both of these features are considered crucial in the case of evaluating treatment responses in IBS, because patients with IBS have multiple symptoms, and the relative contribution of each symptom, including the most bothersome symptom to the overall morbidity, is likely to vary between patients. Moreover, in contrast to biologic endpoints, symptom-based measures are subjective and have to be evaluated within the patient’s own reference system. It might be assumed that the majority of patients will base a judgment of adequate relief on improvements at least in their most bothersome symptom. As for any self-reported judgment, the adequate relief endpoint is subject to variations in patient interpretation. However, the response options (yes/no) are easily defined in comparison to most adjectival response options that might be used in Likert-type endpoints (eg, somewhat improved), and the randomization step provides a means to control for the between-patient variation in interpretation of the adequate relief construct.

The original adequate relief question “In the past 7 days have you had adequate relief of your IBS pain and discomfort” offered a dichotomous yes/no response.

Development of adequate relief endpoint. This endpoint was derived from the program for over-the-counter (OTC) ranitidine, which resulted in approval for the treatment of heartburn. It was piloted in phase 2 trials of the alosetron program and became the primary endpoint in phase 3 trials. It was also used in a phase 2b study in functional dyspepsia. Subsequently, several modifications of the endpoint have been used to accommodate a larger scope of effect than only pain and discomfort. For example, in one program, the endpoint was changed to “In the past seven days have you had overall adequate relief of your IBS symptoms?”

Experience of adequate relief endpoint in clinical trials. Adequate relief as an endpoint is able to distinguish efficacious from nonefficacious drugs. Alosetron, dextoxifosapam, and cilansetron all showed efficacy in the treatment of IBS in studies in which adequate relief was the primary endpoint. For alosetron and dextoxifosapam, the treatment effect for adequate relief was at least as large as for other endpoints collected, and that formed the basis for changing adequate relief from a secondary endpoint in phase 2 trials to the primary endpoint in the phase 3 alosetron program.

The endpoint of adequate relief was accepted by the FDA as a clinically relevant indicator of symptom improvement for the approval of alosetron. The evidence presented above suggests that as an endpoint, adequate relief is responsive, reproducible, and moves in the same direction as other meaningful measures, thus displaying strong reliability and validity.

Satisfactory Relief

Since the initial studies that led to approval of tegaserod, subsequent studies on tegaserod have switched from a 7-point Likert-type item to a binary endpoint assessing satisfactory relief, a very similar construct to adequate relief. Treatment effects with satisfactory relief show greater separation in the effect of tegaserod versus placebo than those obtained with the previous primary endpoint in the tegaserod studies, a subject’s global assessment of relief. Psychometric evidence similar to that described for adequate relief has not been published with satisfactory relief.

Psychometrics and Performance of Binary Endpoints

Psychometric evaluation of adequate relief. Although patients were not directly involved in the development of the adequate relief endpoint used in the alosetron studies, “adequate relief of abdominal pain and discomfort,” there is qualitative evidence to support the use of this endpoint. Abdominal pain and discomfort, which are cardinal features of IBS, have also been identified by clinical trial subjects as their most bothersome symptoms, and in one large survey, they represented the main reasons why IBS patients sought treatment.

Evidence in support of the adequate relief endpoint is also documented in the literature. After prospectively defining treatment responders, statistically significant differences (P < .001) were found between responders and nonresponders in measures directly related to pain, such as the percentage of pain-free days and pain severity, as well as improvements in...
other IBS symptoms including urgency ($P < .001$), stool consistency ($P = .002$), and stool frequency ($P = .007$). In the same study, adequate relief was also predictive of improvements in health-related quality of life (HRQOL) and functional status.

Additional trials conducted within the United States and Europe to test a variety of different compounds provide further evidence for the responsiveness and construct validity of this endpoint by demonstrating that adequate relief is consistently able to differentiate effective treatment from placebo and is highly correlated with other standard measures of treatment efficacy used in IBS. Results of a clinical trial in functional dyspepsia provide evidence that the adequate relief endpoint might be suitable for other gastrointestinal disorders, at least across a related condition in the spectrum of functional gastrointestinal disorders (FGIDs).

**Performance of Binary Endpoints**

One theoretical advantage of binary global endpoints is that they allow patients to rank and integrate the various symptoms of importance to them rather than analyze a treatment effect by a single formulaic approach. Improvements seen with therapeutic agents are most significant when they improve individual patients’ most bothersome symptoms.

Theoretical disadvantages associated with a binary endpoint are also described below. Because there is no universal metric to gauge the “clinical relevance” of adequate or satisfactory relief responses, how does one determine whether differences seen over placebo are clinically meaningful? This question is complex because it might rely as much on the properties of a particular therapeutic agent as on the endpoints used. Furthermore, disease states differ with respect to both placebo response rates, which are relatively consistent in IBS patients, and ability of the drug to reduce patients’ symptoms.

Table 1 shows a comparison of the overall treatment effects seen in the alosetron and tegaserod studies compared with findings of registered agents in some acid-related disorders and ulcerative colitis. Data are taken from the respective U.S. Package Inserts for the medications that underwent review and approval by the FDA. Overall, the efficacy of alosetron, as judged by difference over placebo on adequate relief, is comparable to that seen with other therapeutic agents in other disease states.

Some criticisms of the adequate relief endpoint include the lack of validation and sensitivity, inability to detect worsening of symptoms, potential for failure to report adequate relief in subsequent weeks of the trial once this was achieved, the influence of symptom severity at entry on measurement of the response outcome, and variable interpretation of the adequate relief construct and terminology by different patients and in different countries. There are, however, evidence-based responses to these questions or concerns about the adequate relief endpoint, and they are summarized below.

If adequate relief was not optimally sensitive to changes in symptom severity, it would represent a larger, not smaller, hurdle to prove statistical significance and clinical relevance. However, as evidenced in the published clinical trials, statistical and clinical significance is still achieved by using this endpoint. A second criticism of adequate relief is that only binary, yes or no, responses are options, and the endpoint does not allow for detection of worsening of patients’ symptoms. This is correct. However, for regulatory purposes in fields such as asthma, cancer, and IBS, responder definitions are usually required, and by definition, responder versus nonresponder status is binary.

In addition to the primary endpoint, IBS clinical trials include multiple secondary endpoints that provide the ability to document worsening, and these data are always referenced to a placebo group. Once a patient reports adequate relief, reports for alosetron indicate that the cumulative transition probability for adequate relief in 12-week studies is 0.8; they are 80% likely to continue to experience adequate relief during the subsequent weeks in the study. It has to be acknowledged, however, that the transition probability of efficacy will also be a reflection of the specific efficacy profile of a particular drug. Thus, if efficacy wanes over time, the cumulative transition probability will decrease, no matter how precise a measure happens to be.

A recent report on usual medical care suggests that responses to satisfactory relief could possibly be influenced by baseline severity at entry into the study. These results have never been confirmed in either a placebo-controlled drug trial or with adequate relief as the primary endpoint. To properly evaluate responses to an endpoint by entry pain severity, patients need to be receiving the same treatment, and that agent must have similar therapeutic benefit on mild, moderate, and severe pain. The study by Whitehead et al did not fulfill these criteria. In a recent report involving the evaluation of data from more than 1200 IBS patients in the alosetron treatment program, it was concluded that treatment effects based on adequate relief were not significantly different across pain severities at entry (baseline) for either the single therapeutic agent (alosetron) or for placebo.

Figure 1 shows adequate relief response rates stratified by entry baseline pain severity in the dextifosipam trial. For both dextifosipam-treated and placebo-treated patients, adequate relief responder rates were insensitive to baseline entry severity.

| Table 1. Comparative Efficacy of Alosetron and Tegaserod With H2-Blockers and Mesalamine Compounds |
|---------------------------------------------------------------|---------------------------------------------------------------|
| Agent                                      | Difference in active response minus placebo* |
| Alosetron: adequate relief                | 10%–19%                                         |
| Tegaserod: Likert scale                   | 13%–14% month 1                                   |
| Ranitidine: Duodenal ulcer healing        | 5%–11% month 3                                    |
| Gastric ulcer healing                     | 19% week 2                                       |
| Cimetidine: Gastric ulcer healing         | 13%–16% week 4                                    |
| Famotidine: Gastric ulcer healing         | 19%–21% week 6                                    |
| Mesalamine (Asacol): ulcerative colitis disease activity | 13% week 6                              |
| Mesalamine (Pentasa): ulcerative colitis remission | 14%–15%                                     |

*Data are from 2004 Physicians’ Desk Reference.
pain severity. Analogous data are not presently available to us to assess the influence of baseline pain severity on the endpoint of satisfactory relief in response to treatment. This should be addressed by future study of data from placebo-controlled trials that contain this information.

**Summary on Binary Endpoints**

Experience with the use of binary endpoints with several different compounds in thousands of patients in drug trials, as well as the above observations, suggests that the adequate relief endpoint represents an acceptable primary endpoint for use in clinical trials of novel therapeutic agents for IBS. Binary endpoints have been shown to display treatment effects that are as large as or larger than the responses garnered through multiple secondary endpoints in the same studies, thus providing evidence to support their use as acceptable endpoints.

**Irritable Bowel Syndrome Severity Scale**

The Rome III guidance document considers the IBS-SS to be the only IBS symptom severity scale available to assess responsiveness for clinical trials. Its validity and responsiveness to treatment were based on studies of usual medical care or hypnotherapy rather than drug therapeutic studies. Given this endorsement by Rome III, it is important to understand the content, validation, and characteristics of the IBS-SS total score as a trial endpoint.

**Description of Irritable Bowel Syndrome Severity Scale**

The IBS-SS total score is computed by summing patients’ responses to 5 items: abdominal pain (severity and number of days with pain), abdominal distention (severity only), satisfaction with “bowel habit,” and interference with life in general, often referred to as a global HRQOL item. Apart from pain frequency, all questions are answered by using a 0–100 visual analog scale (VAS); item-level scores for the pain frequency item are computed by multiplying the number of days with pain during a typical 10-day period by 10, for a range of 0–100. The maximum achievable score, therefore, is 500 (100 per item), with higher scores indicating greater levels of severity and impact. As defined by Francis et al, mild, moderate, and severe cases generally correspond to scores of 75–175, 175–300, and >300, respectively, on the basis of clinician classification. Controls are scored below 75; patients scoring in this range can be considered to be in remission. The IBS-SS provides a summed scale, similar in principle to the Crohn’s Disease Activity Index and similar instruments. Significant issues need to be overcome before this can be recommended as a primary endpoint in drug trials for IBS.

**Development of the Irritable Bowel Syndrome Severity Scale**

Francis et al reported that the constructs indicative of IBS severity were identified by one of the authors on the basis of previous experience. It is unclear whether the construct validity of this PRO instrument was confirmed with patient involvement to optimize the question wording and response scales to enhance the consistent interpretation across respondents and minimize measurement error.

Another concern with the IBS-SS is the construction of the items, eg, the use of VAS response scales. There is literature indicating that both physicians and patients prefer numeric rating scales in a variety of contexts. In the IBS-SS, respondents indicate their answers with an “X,” which might introduce error because a cross might intersect a line at 2 rather than 1 point. The IBS-SS uses the word currently to define the reference period, whereas the instructions section provides an inconsistent instruction because the respondents are asked to consider “the last 10 days or so” when formulating their answers. Such inconsistency might introduce measurement error.

**Psychometric Evaluation of the Irritable Bowel Syndrome Severity Scale**

Published evidence for a number of key psychometric properties (eg, factor structure, internal consistency, and construct validity) of the IBS-SS is not available, and there is no psychometric support to justify the IBS-SS scoring algorithm, which assumes that all items should be weighted equally, eg, each symptom’s severity is given the same weight as the HRQOL item, and all symptoms are equally important to patients.

With regard to test-retest reliability, Francis et al collected retest data from 32 patients and 15 controls who were instructed to complete the IBS-SS a second time within 6–24 hours and to mail back their responses. This short test-retest period is unusual and raises the risk that subjects might have remembered their original responses. Reliability data were interpreted qualitatively rather than as an intraclass correlation coefficient (ICC). IBS-SS scores (repeated within 24 hours) were considered to be reproducible (Figure 2). However, within the moderate severity subgroup (n = 11) the range of differences of 243 (~97 to +146) was nearly half (250) the value of the highest possible total score (500).
Responsiveness was assessed in patients undergoing conventional therapy and hypnotherapy for IBS. Patients were stratified on the basis of physician rating into 2 groups, “little changed” (n = 19) and “considerably changed” (n = 17), after 3 months of treatment. The average change in IBS-SS total score between these 2 groups was significantly different (P < .001), suggesting that the IBS-SS was able to detect a clinically relevant change. IBS-SS demonstrates discriminant validity and responsiveness for patients undergoing hypnosis trials.

**Summary of Irritable Bowel Syndrome Severity Scale**

Whereas IBS-SS has reasonable discriminant validity and responsiveness, current data on internal consistency, test-retest reliability, content and construct validity require further validation, and the instrument has not been tested in randomized controlled drug trials.

**Comparison of Performance of Satisfactory Relief and an Integrated Symptom Questionnaire**

Whitehead et al\(^2\) have reported that the endpoint of satisfactory relief, when used by IBS patients receiving usual medical care without a placebo control group, is confounded by baseline symptom severity and does not accurately reflect symptom improvement. They compared a binary satisfactory relief measure to the IBS-SS, which assessed symptom severity as mild, moderate, or severe, in 350 patients (81% female; average age, 50 years). This study was performed in a health maintenance organization (HMO) setting. Patients were reappraised 6 months after receiving usual medical care, which was not standardized.

They reported that initial severity of IBS significantly affected the proportion who reported satisfactory relief (mild, 72%; moderate, 53%; severe, 44%). Initial severity also influenced the proportion who were somewhat or markedly better (mild, 62%; moderate, 44%; severe, 38%). On the other hand, initial severity of IBS did not affect the proportion of patients who reported a 50% reduction in symptoms per the IBS-SS (mild, 26%; moderate, 25%; severe, 23%). Although patients with mild IBS severity at baseline were the most likely to report satisfactory relief, they showed no significant average decrease in symptom severity or improvement in HRQOL. On the basis of this information, Whitehead et al\(^2\) suggested that the severity scale performed in a superior manner to the endpoint of satisfactory relief.

This article raises issues that deserve further study. Certainly the binary endpoints are advantaged by the fact that at least 15,000 patients have participated in at least 10 large multicenter clinical trials\(^3\)–\(^11,13,18,19,23,29\)–\(^33\) and show responsiveness. Nevertheless, formal and direct prospective studies are needed to ascertain the performance characteristics of the 2 types of endpoints.

Fourth, IBS-SS was compared with the one-time assessment of satisfactory relief, which in every other trial was used as a weekly question during a 3-month trial. It is also unclear whether one can equate absence of symptoms with adequate or satisfactory relief in a symptomatic study population undergoing a treatment trial, because such a study has not been done. There is concern that there might be floor and ceiling effects that will affect the response in the outcome measured in the usual care cohort receiving different treatments,\(^2\) because those with mild severity cannot get much better, whereas those starting with greater severity can only get better.

Finally, with only one evaluation time point after the first assessment in a usual care setting, the response rates (irrespective of patient severity assessed by the IBS-SS)\(^2\) are low, indicating that the IBS-SS might not be a sensitive instrument in the usual care setting. Because of these and other previously described methodologic issues with the study by Whitehead et al\(^2\), definitive conclusions on the relationship of satisfactory relief responses to severity at baseline will require further analysis in a standard treatment trial situation.

**Conclusion**

There is a need for greater validation of all IBS endpoints. Trials are required to characterize the responsiveness of the IBS-SS to drug therapy. This analysis suggests that although further validations of binary endpoints are desirable,
the adequate relief endpoint is currently supported by psychometric data based on published studies and is easily understood by IBS patients.35 Those trials have demonstrated the responsiveness of the measure to a variety of orally administered medications in the appropriate subgroups of patients with IBS. 

Future studies are needed to determine (1) the role of pre-treatment severity in the various outcome measures; (2) the psychometric properties of the IBS-SS and its utility in pharmaceutical trials; (3) the performance of various outcome measures by pooling available data from existing treatment trials (eg, via a meta-analysis); and (4) the relative effect size produced by dichotomous endpoints as compared with Likert-type endpoints. However, this review of existing data also suggests that at present, adequate relief should be recognized by regulatory authorities as a standard endpoint, and it should be acceptable as a primary endpoint in trials in FGIDs until future research demonstrates it should be replaced by a better endpoint.

References


Address requests for reprints to: Michael Camilleri, MD, Mayo Clinic, Charlton 8-110, 200 First St SW, Rochester, MN 55905. e-mail: camilleri.michael@mayo.edu.

Drs Camilleri (R01 DK54681, K24 DK02638), Drossman (R24 DK067674), Mayer (P50 DK64539, R24 AT002681, R01 DK58173, R01 DK48351), and Talley (U01 DK 65713) receive grants for studies in the field of functional gastrointestinal disorders from the National Institutes of Health. The authors received support from RTI, a not-for-profit research institute, for travel to a face-to-face meeting. Dr Camilleri receives current research support for single center pharmacodynamic studies from GlaxoSmithKline, Johnson and Johnson, and Bristol-Myers Squibb and serves as a consultant for Astellas, Dynoegen, GlaxoSmithKline, Novartis, Theravance, and Zeria. Dr Drossman receives research support from Novartis Pharmaceuticals, Microbia, and Procter & Gamble pharmaceuticals and is a consultant for Novartis, Procter & Gamble, Microbia, Astellas, and Tioga. Dr Fehnel is a consultant for Microbia, Novartis, and Tioga. Dr Mangel is a consultant for Tioga, Vela, Bristol-Myers Squibb, Astellas, GlaxoSmithKline, Trine, Napo, Microbia, Novartis, Boehringer Ingelheim, Pharmos, and Theravance. Dr Mayer receives research support for single center pharmacodynamic studies from Avera, GlaxoSmithKline, Johnson & Johnson, Lilly, and Novartis and serves as a consultant for GlaxoSmithKline, Novartis, Avera, Allergan, AstraZeneca, and Sanofi. Dr Talley is a consultant for AstraZeneca, Axcan, Chugai, EBMeds, Giaccone, GlaxoSmithKline, Kosan, KV Pharmaceuticals, Medscape, ProEd Communication, Renovis, Inc, Solvay, Strategic Consultants International, Takeda Pharmaceuticals, Inc, TAP Pharmaceutical Products, Inc, Therapeutic Gastrointestinal Group, Theravance, Yamanouchi and receives research support from Merck KGaA, Novartis, TAP Pharmaceuticals, Axcan, Boehringer-Ingelehim, and Forest.