

# Design of Treatment Trials for Functional Gastrointestinal Disorders

Design of Treatment Trials Committee: E. JAN IRVINE,\* WILLIAM E. WHITEHEAD,<sup>†</sup> WILLIAM D. CHEY,<sup>§</sup> KEI MATSUEDA,<sup>||</sup> MICHAEL SHAW,<sup>¶</sup> NICHOLAS J. TALLEY,<sup>#,\*\*</sup> and SANDER J. O. VELDHYZEN VAN ZANTEN<sup>\*\*†</sup>

\*Division of Gastroenterology, St. Michael's Hospital and University of Toronto, Toronto, Ontario, Canada; <sup>†</sup>Division of Gastroenterology, University of North Carolina–Chapel Hill, Chapel Hill, North Carolina; <sup>§</sup>Division of Gastroenterology, University of Michigan, Ann Arbor, Michigan; <sup>||</sup>Division of Gastroenterology, NCNP, Ichakawa City, Japan; <sup>¶</sup>Division of Gastroenterology, Park Nicollet Clinic and University of Minnesota, Minneapolis, Minnesota; <sup>#</sup>Division of Gastroenterology, Mayo Clinic College of Medicine and Division of Gastroenterology and Hepatology, Rochester, Minnesota; <sup>\*\*</sup>Department of Medicine, University of Sydney, Sydney, Australia; and <sup>\*\*†</sup>Division of Gastroenterology Dalhousie University, Halifax, Nova Scotia, Canada

This document addresses the design of trials to assess the efficacy of new treatments for functional gastrointestinal disorders (FGID), emphasizing trials in irritable bowel syndrome and dyspepsia, because most research has been undertaken in these conditions. The double-blind, randomized, placebo-controlled, parallel group trial remains the preferred design. Randomized withdrawal designs, although encouraged by the European Agency for the Evaluation of Medicinal Products, have the same potential disadvantages as a crossover design, including carryover effects, unmasking (unblinding), and overestimation of the potential benefit for clinical practice. Innovative trial designs that evaluate intermittent (on demand) treatment are likely to become more common in the future. Investigators should include as broad a spectrum of patients as possible and should report recruitment strategies, inclusion/exclusion criteria, and attrition data. The primary analysis should be based on the proportion of patients in each treatment arm who satisfy an a priori treatment responder definition, or a prespecified clinically meaningful change in a patient-reported symptom improvement measure. Such measures of improvement are psychometrically validated subjective global assessments or a change from baseline in a validated symptom severity questionnaire. It is unethical to change the responder definition after a trial begins. Data analysis should address all patients enrolled, using an intention-to-treat principle. Reporting of results should follow the Consolidated Standards for Reporting Trials guidelines and include an analysis of harms data and secondary outcome measures to support or explain the primary outcome. Trials should be registered in a public location, prior to initiation, and should be published even if the results are negative or inconclusive.

The committee's aims were to review the literature on trial design for the functional gastrointestinal disorders (FGIDs), to further develop guidelines<sup>1–3</sup> to assist researchers in conducting treatment trials for the FGIDs,

provide standards to help explain the mechanisms of therapeutic success and enable regulatory agencies, researchers, and providers to better evaluate the quality of published studies. This report focuses largely on designs that evaluate treatment efficacy, with emphasis on irritable bowel syndrome (IBS) and functional dyspepsia (FD), because they have been studied most extensively. Studies that address pathophysiology or mechanism of treatment effects are not included in this review because they require quite different and diverse study designs. Recommendations in this article are based largely on consensus of the literature, except where specifically indicated. We refer readers to the corresponding chapter in the Rome III book for a more detailed discussion with evidence-based examples.

## Identifying the Research Question and Hypothesis

The goals of most treatment trials are to ascertain the impact of the intervention(s) on (1) the frequency and severity of symptoms, (2) health status and quality of life, (3) the patient's ability to cope with symptoms, and/or (4) the use of health care resources. Generally, a single trial can answer only 1 or 2 of these questions.<sup>1–3</sup>

*Investigators should select their most important research question(s), pertinent to the specific FGID, develop a hypothesis based on available evidence, and design a study that will most effectively answer the proposed research question.*

---

**Abbreviations used in this paper:** CONSORT, Consolidated Standards for Reporting Trials; EMEA, European Agency for the Evaluation of Medicinal Products; FD, functional dyspepsia; FGID, functional gastrointestinal disorder; ITT, intention to treat; NNH, number needed to harm; NNT, number needed to treat; IBS, irritable bowel syndrome.

© 2006 by the American Gastroenterological Association Institute  
0016-5085/06/\$32.00  
doi:10.1053/j.gastro.2005.11.058

## Patient Population

A broad spectrum of patients should be included to support the generalizability of the trial findings to patients outside of the trial. In pharmaceutical research, particularly, regulatory agencies may limit licensed drug indications to the trial population. The study population should be selected based on the question, treatment (including possible side effects), expected results, and empirical data.<sup>2</sup> A screening log, summarizing the most important demographic variables in patients entered or excluded and the reasons for exclusion, is strongly recommended.

*A screening log provides support for the generalizability of the results.*

Specified inclusion and exclusion criteria are mandatory for all studies and should include the FGID case definition. If enrollment is targeted to a special population to maximize treatment efficacy or minimize side effects, the reasons must be carefully documented.<sup>2,4,5</sup>

*It is advisable to include as broad a spectrum of patients as possible, defined by the ROME-specific FGID criteria. Restricting the study population must be justified and inclusion and exclusion criteria must be specified.*

In clinical practice, many physicians avoid formal investigation in favor of a positive diagnosis, reassurance, and lifestyle modification. However, entry criteria for treatment trials must be more specific. The consensus view is that the minimum evaluation should include a complete blood count, imaging of the relevant part of the gastrointestinal tract within the previous 5 years, and other investigations determined by symptoms and family history.<sup>6</sup> Emerging evidence suggests that screening IBS patients for gluten enteropathy may also be desirable.<sup>7</sup> When testing is required for study inclusion, it should be consistent across all study arms, and the timing should be defined and recorded.

*The minimum screening investigation for eligibility should be specified.*

Most trials of FGIDs have been conducted in academic centers specifically interested in the FGIDs, creating concerns of selection bias that favor inclusion of patients with more severe symptoms and/or with psychosocial issues.<sup>8</sup> Two large trials showed significant differences in treatment response between primary and referred patients with FD.<sup>9</sup> Thus, researchers should consider recruiting broadly, noting if subjects are from primary, secondary, or tertiary care. Differences in baseline severity and treatment response by site or type of recruitment should be examined.

*Patient characteristics should be documented sufficiently to examine the comparability of patients among centers and allow comparisons with other populations.*

Special recruitment strategies such as advertising have been accepted in some countries to accelerate recruitment. A recent IBS study observed that patients recruited by newspaper advertisement, in comparison to patients enrolled by gastroenterologists, were older, more highly educated, more often depressed but less anxious, and had less severe IBS symptoms; primary care patients were also anxious but had symptom severity that was intermediate between patients recruited by advertisement and patients recruited from gastroenterology clinics.<sup>10</sup>

*Recruitment strategies should be clearly identified to allow exploration of different patterns of treatment response.*

## Patient Characteristics

Important patient characteristics to report include age, gender, race, symptom severity, duration of disease, prior treatments for the study condition (and response), and the use of coexisting medications, including over-the-counter drugs and vitamins. These may impact outcomes and should be tested as possible disease or effect modifiers. For example, gender differences in drug response have become evident in clinical trials of certain serotonergic drugs in patients with IBS.<sup>11</sup> Depending on the hypothesis, investigators may choose to enroll only one gender. However, if both women and men are to be included, there should be sufficient numbers of both to allow meaningful subgroup analyses. As data accumulate describing the genetics of the FGIDs in relation to drug responsiveness, it may become relevant to assess these parameters during clinical trials.<sup>11</sup> It is also advisable to assess for psychological distress or prior mental health problems; in trials of psychological interventions or psychoactive drugs, these may be important effect modifiers, and in trials of nonpsychological treatments, they are potential confounders that could influence both baseline symptom severity and response to treatment.<sup>12</sup>

*Potential disease modifiers/confounders that might affect response to therapy should be assessed.*

## Clinical Trial Design

Clinical trials differ from usual practice in several ways, including the application of strict eligibility criteria, use of a placebo, a standardized intervention, frequent follow-up visits with extensive data recording, and the use of study coordinators. Nonetheless, standard aspects of diagnosis and management, especially an adequate explanation and reassurance about the disease, are part of standard care and should be provided to all patients in the trial. Novel interventions must show promise of a benefit over standard care.

**Table 1.** Major Sources of Bias

Bias Type	Comments
1. Investigator bias	Conscious or unconscious bias, usually expressed through decisions about eligibility
2. Patient expectancy (placebo)	Especially a problem where end points are subjective
3. Ascertainment bias	
Self-selection for treatment	Patients are more likely to respond positively to treatments they prefer and seek out
Changes in subject pool	Publicity or other factors may influence the subject pool over time
4. Nonspecific effects	
Doctor-patient relationship	Especially important in psychological interventions
Regression to mean	Patients are usually enrolled when most symptomatic and inevitably "improve"
5. Publication bias	Authors are more likely to submit positive results and journals more likely to publish them

*Every trial should incorporate the principles of good clinical practice to ensure that the study results are relevant to real practice situations.*

### Unique Challenges for the Design of Treatment Trials in FGIDs

Study designs for treatment trials in FGID face several important challenges: (1) a high placebo response rate<sup>13,14</sup>; (2) fluctuating symptoms<sup>15</sup>; (3) possible need for multimodal therapy owing to weak effects of available treatments or multiple etiologic mechanisms interacting in the disease process<sup>16</sup>; (4) difficulty of masking (blinding) patients and investigators, particularly in trials of behavioral interventions<sup>17</sup>; (5) contamination from over-the-counter treatments or drugs taken for other conditions (eg, antidepressants); and (6) avoidance of harms in treating non-life-threatening conditions.<sup>18,19</sup>

*Bias*, defined as systematic error that leads to a deviation of the estimated treatment effect from its true value, may enter a clinical trial at any stage from patient enrollment to publication of the results. The major sources of bias are listed in Table 1.<sup>20</sup>

**Masking.** Double masking (of both patients and researchers) to the intervention ensures the validity of the outcome assessment. "Triple masking" is desirable and extends masking to all investigators, including data managers and statisticians.<sup>21</sup> In drug trials, investigators are encouraged to ask both the patient and the interventionist who interacts with the patient at the end of the trial whether they believe the active treatment was administered and to report these data. Certain interventions, such as psychotherapy, hypnosis, sphincterotomy,

or drug trials in which the active drug causes predictable side effects or rapid symptom change, are difficult to mask from patients or investigators, but possible solutions to maintain an investigator-masked outcome assessment include using independent assessors who are unaware of the intervention, using a standardized interviewer or self-administered questionnaire, or performing laboratory tests (eg, anal manometry in fecal incontinence) that are interpreted by individuals not interacting with the patients.

*It is mandatory to undertake the maximum masking possible, determined by the type of intervention and study design.*

**Randomization.** *Randomization* is a process (equivalent to the flip of a coin) used to assign patients to treatment arms in an unbiased fashion. The allocation sequence should be concealed from investigators and research personnel should be unaware of the treatment to which a patient will be assigned until after the patient has been deemed eligible and has consented to participate.<sup>22</sup> *Stratified randomization*, whereby the most important prognostic factors (eg, gender and usual bowel habit) are identified beforehand, uses a separate randomization sequence for each stratum (eg, male versus female or constipation-predominant versus diarrhea-predominant IBS) to balance these factors among treatment groups.<sup>23</sup> Stratification should be limited to 1 or 2 factors.<sup>22,24</sup> Particularly in multicenter trials, in which sites may enroll only a few subjects, randomization can be performed in blocks. A *block* refers to the number of subjects within which the group assignments have to be balanced. A *permuted block design* (variable block size) ensures that the sequence of assignments is unpredictable to the investigator. When reporting the trial, the randomization procedure should be explicitly described because it is a potential source of bias.

*Investigators must include a detailed description of their randomization scheme in the report of the study.*

**Selecting the control group.** A placebo control group is essential to establish the efficacy of a new treatment. When a proven efficacious treatment exists, comparison against this active treatment may be considered,<sup>25</sup> but inclusion of a placebo is still recommended to avoid an inconclusive trial, in which the active treatments are of similar efficacy.

Behavioral therapy trials<sup>17</sup> pose particular challenges to identify inactive comparison treatments that generate expectancy comparable to the active intervention. Untreated patients are poor control subjects because they can experience a "negative expectancy,"<sup>26</sup> which may result in an overestimate of the impact of the intervention. Options to assess the integrity of behavioral trials include (1) testing the credibility of both active and control interventions after initial exposure (eg, by using

the Credibility Scale<sup>27</sup>),<sup>28</sup> or (2) using a process measure to ensure that the active treatment is producing the intended effects on physiology or cognitions while the control treatment does not (eg, does biofeedback for fecal incontinence change anal sphincter squeeze pressure more than the control condition, or does cognitive-behavioral therapy alter the patients dysfunctional attitudes to a greater extent than an education control treatment?).<sup>17</sup> Investigators should discuss residual sources of bias and their potential impact on study findings in the discussion section of the report.

*A placebo control group is essential. In behavioral treatment trials, confirming that the control condition produces a similar expectation of benefit, but does not act on the same physiologic or psychological principles, is recommended.*

### Placebo Interventions

A placebo is an intervention believed to lack any specific effect to change a particular disorder.<sup>29</sup> Placebo effects range from 10% to 70% for FD<sup>14</sup> and 0% to 84% for IBS.<sup>13</sup> This substantial placebo response rate makes it more difficult to demonstrate superior efficacy of new treatments. Of note, a placebo administered by a physician appears to be more powerful than one given by other health professionals.<sup>29</sup> Some treatments also demonstrate an *order effect*, in which an effective drug has a lesser benefit when given after a placebo. This is especially important if a placebo run-in period is implemented to exclude placebo responders or in studies with a crossover design, because approximately half of patients in a crossover study receive placebo first.

External factors may also contribute to changes in health status making it difficult to detect a treatment effect, including (1) a natural variation in symptoms, (2) regression toward the mean, and (3) unidentified or unintended cointerventions. *Regression to the mean* is the likelihood that patients consult when symptoms are particularly severe and improve with time owing to the natural variation in symptom severity and irrespective of trial participation.<sup>30</sup> Important cointerventions such as changes in diet or using over-the-counter remedies could also lead to a false interpretation that an intervention was effective, as could the extra attention given patients by researchers during clinical trials (Hawthorn effect).<sup>20</sup> The magnitude of the placebo response may also be influenced by the wording of the question used to define treatment response or by the use of a compound question; a recently published meta-analysis<sup>31</sup> suggests that the placebo response rate is larger when a responder is defined by a global improvement in IBS symptoms compared to defining a responder by reduction in abdominal pain (average placebo responses of 36% versus 28%).

*The placebo response rate in treatment trials of FGIDs is substantial and largely unavoidable.*

**Baseline observation versus placebo run-in.** A period of prospective baseline measurement before treatment is useful to evaluate patient eligibility. This also limits recall and reporting biases by ensuring that patients are currently symptomatic. It allows comparison of patients in the active and placebo groups, as well as evaluation of a clinically important change in health status.

Older studies have used a placebo run-in period where all patients received placebo for a specified period and their response was assessed, using the study outcome measures. Patients who significantly improved were excluded from further participation to reduce the proportion of placebo responders and to exclude patients with poor adherence. This has been used in several trials of allergic rhinitis and, although acceptable to regulatory agencies, may underestimate the overall effect size.<sup>32,33</sup> It is also difficult to predict whether (1) the placebo response increases, plateaus, or decays after the run-in phase,<sup>34</sup> (2) a differential dropout occurs, and (3) patients removed from a trial have a different response to those who continue. Exclusion of patients for placebo response may also disrupt the doctor-patient relationship for future management.

*The disadvantages of a placebo run-in appear to outweigh the benefits and it is best avoided. However, baseline observations are recommended.*

### Choice of Study Design

The double-masked, randomized, placebo-controlled trial is the gold standard method to test the efficacy of a new treatment. A parallel group study design requires that patients be randomized to receive only one treatment assignment throughout the trial (after a period of baseline assessment without treatment). Dose-ranging studies (different groups receive different doses) and multiple control treatments, with a baseline observation of no treatment or a washout period after treatment, are different variants of a parallel group design.

*Crossover designs*, in which subjects receive both treatments during distinct time periods, separated by a washout phase have been popular in some FGIDs.<sup>14</sup> Theoretically, lesser variability in outcomes within subjects could require a smaller sample size for the desired statistical power. However, patient dropout rates and missing data have a greater impact than in a parallel design, because patients are omitted from both study arms when data are missing. The greatest disadvantages of crossover designs are (1) the carry-over (period-by-treatment) ef-

facts that occur when the first treatment influences the response to the second treatment or when symptoms change with time,<sup>35</sup> and (2) the high likelihood of unmasking owing to side effects.<sup>35</sup> The European Agency for the Evaluation of Medicinal Products (EMA) may accept a crossover design for a Phase III trial, yet has highlighted problems that could invalidate study results<sup>36</sup> and does not provide guidance for analysis. If period and sequence effects occur, only the first treatment period data should be used to determine efficacy. Although crossover designs are not recommended for treatment trials with subjective end points, they may be used in physiologic studies, where the end points are objectively measured.

A factorial design can be undertaken to evaluate combined treatments.<sup>37</sup> For example, to test the effects of combining two treatments, A and B, subjects are randomly assigned to 4 groups: no A and no B; A and no B; B and no A; or both A and B. Investigators might consider such a design either (1) to save money by testing 2 treatments at once with fewer subjects overall, or (2) to test for synergistic effects of combined treatments. Importantly, the 2 treatments should have distinct mechanisms of action to be able to interpret the simple effects (ie, the comparison of all patients receiving treatment A to all patients not receiving treatment A, and the comparison of all patients receiving treatment B to all patients not receiving treatment B), or to detect whether there is added benefit from combining treatments. Also, a control is required for each intervention. Potential cost savings are frequently offset by the complexity of interpreting the data, except when testing for synergistic treatment effects.

The withdrawal trial is an *enrichment design*, in which all subjects receive the active treatment. At a predefined time point, they are classed as responders or nonresponders and the latter are excluded. Responders are then randomly assigned to receive active treatment or placebo and efficacy is based on the second part of the trial. Potential carry-over effects from the first treatment, however, can prevent an accurate estimate of the drug benefit. The EMA guidelines<sup>38</sup> support this design for testing drugs for short-term efficacy in IBS, and require 2 or more treatment cycles to demonstrate efficacy. The EMA recommends<sup>38</sup> that masked withdrawal (switch to placebo at an unpredictable time) be undertaken after active drug, but does not address how to perform the complex statistical analysis. Like the placebo run-in, this design can overestimate the effect size.<sup>39</sup> One completed study<sup>39</sup> followed the EMA guidelines (with minor variations) and provided data supporting the efficacy of tegaserod for IBS on repeated dosing cycles.

There is a growing interest in developing drugs for intermittent treatment (short-course administration for a predetermined time period after symptom recurrence) or on-demand treatment (medication is taken only during symptoms). These issues have been addressed in gastroesophageal reflux disease.<sup>40</sup> IBS and FD trials have focused on continuous administration of drugs to moderate or prevent attacks.<sup>4,5</sup> However, most patients with FGIDs experience episodic "attacks,"<sup>15,41</sup> and experts believe that patients often take medications only as needed. Trial designs and outcome measures required for testing the efficacy of intermittent therapy differ from those used to test continuously administered treatments. After establishing efficacy during continuous administration, intermittent or on-demand studies can be conducted. Guidelines for intermittent treatment of migraine<sup>42</sup> or gastroesophageal reflux disease<sup>43</sup> may provide a model for FGIDs.

*The parallel group design is the accepted standard for evaluation of efficacy for most treatments and is applicable to most experimental situations. The crossover design is best avoided.*

**Types of trials.** The strongest case supporting the efficacy of a new medication is made by demonstrating clinical and statistical superiority to placebo or an active control treatment. An equivalence study or noninferiority trial can also be considered if (1) a known, effective treatment is available and it would be unethical to administer a placebo (eg, cancer or inflammatory bowel disease, not FGID), or (2) a new treatment might be less costly, safer, or just as good as standard therapy.<sup>25</sup> Such trials are usually more costly than superiority trials, because much larger sample sizes are required. Investigators must first estimate the expected difference between standard treatment and placebo from a meta-analysis or systematic review<sup>25,44</sup> and then define "equivalence margins" that are smaller than the expected difference. The trial is judged to be positive only if the 95% confidence interval for the observed difference between the new and standard treatments falls within the equivalence margins. For a noninferiority trial, only the lower 95% confidence limit must fall within the margins.

In trials that compare the investigational treatment to a different active treatment, the investigator is obliged to show that the treatment arms are in equipoise; it is unethical, for example, to compare the investigational drug to an ineffective dose of an alternative compound.

*Superiority trials (not equivalence or non-inferiority trials) are recommended for FGIDs.*

**Duration of treatment.** Treatment duration for specific FGIDs should be based on natural history data

describing the frequency and duration of episodes. For IBS, this is highly variable,<sup>15</sup> but for the majority of patients both flares and remissions appear to last less than 1 week<sup>15,45</sup>; for dyspepsia, there is a high symptom turnover in the general population.<sup>46</sup> Prior recommendations for trials of 8–12 weeks were based on experience and on concerns for cost and ability to retain patients. EMEA guidelines<sup>38</sup> differentiate trials of short-term efficacy, for which they would accept 4-week trials, from long-term efficacy trials, for which 6-month trials are required. Although both types of trials require patients with active symptoms at randomization, long term-studies could include patients with intermittent symptoms. Further research on the natural history of individual FGIDs should be a high priority, to allow clearer recommendations for trial duration. Extended follow-up should be considered to determine treatment durability and should relate to symptom periodicity and presumed treatment mechanism.

*A minimum treatment duration of 4 weeks that reflects the symptom periodicity and anticipated treatment mechanism is recommended. If chronic use is anticipated, trials of at least 6 months should be undertaken to establish long-term efficacy.*

**Adherence to treatment and study protocol.** Standard methods to assess adherence include interviewing patients, counting unused medication,<sup>47</sup> or measuring blood levels of metabolites<sup>28</sup> and may be especially important when interpreting studies of long duration. The frequency of missed or late appointments and missing data from diaries or questionnaires should be reported for all trials.

*Adherence to the protocol and treatment should be measured.*

## Methods for Collecting Symptoms and Outcome Data

### Accuracy of Symptom Recall

Efficient symptom assessment can be achieved by having patients complete questionnaires before treatment and at follow-up visits. However, concerns about the accuracy of retrospective questionnaires include whether (1) symptoms present on the day they complete the questionnaire influence reporting; (2) poor recall affects the accuracy of a retrospective report; and (3) patients feel pressured to give a more positive report if questionnaires are completed in the presence of the investigator. Although data support the presence of these biases, they do not appear to be substantial.<sup>48</sup> Recall of health-related events appears to be reasonably accurate for up to 3 months.<sup>49</sup>

### Symptom Diaries

Diaries have been used to measure primary or secondary end points and minimize recall bias. Relatively few symptoms are recorded and ratings can be performed at a fixed time (eg, bedtime) or when symptoms actually occur. The former method is simpler for data analysis. A major problem of diaries<sup>50</sup> is poor adherence; patients often complete them retrospectively or just before a visit.<sup>51</sup> Hand-held electronic devices with reminder alarms<sup>51</sup> or collecting information by telephone<sup>52</sup> have been shown to improve adherence to 80%–90%, and patient satisfaction is good.<sup>51,52</sup> Diary symptoms can also be recorded on secure Web sites, which can accurately record the time of completion.

*Retrospective questionnaires are an acceptable method for assessing symptoms provided the recall interval is limited to 3 months. Patients should receive clear instructions on the use of a diary, including the directive to leave it blank if they forget to record information. Electronic diaries are preferred over paper diaries. Methods to ensure adherence to recording methods should be implemented.*

### Outcome Measures

The primary outcome variable(s) provides the basis for judging the success or failure of an intervention. Only 1 or at most 2 variables should be selected and this should be done before the trial begins. The Food and Drug Administration and EMEA have recommended that investigators provide rules, a priori, that allow classification of each participant as a responder or nonresponder for the primary outcome.<sup>38</sup> Most trials also include secondary outcome variables to (1) strengthen the results by showing concordance between individual symptoms and the primary outcome measure, (2) address the mechanism of the intervention, (3) assess the safety or (4) cost effectiveness of the treatment, and (5) identify variables that predict which patients are most or least likely to benefit.

The definition of a responder should reflect a clinically meaningful symptom improvement for each patient. For IBS and other FGIDs, there is no consensus on what constitutes a clinically meaningful improvement. Some studies accept as little as a 10% reduction in a visual analog scale rating of symptom severity<sup>53</sup> or 1 step on a 7-step ordinal scale<sup>39</sup> as clinically meaningful, whereas other studies require a 50% reduction in an aggregate symptom severity index<sup>54</sup> or questionnaire.<sup>55</sup> However, the most commonly employed definition of clinically meaningful improvement in IBS has been a patient's report (yes or no) of "adequate relief of abdominal pain and discomfort"<sup>4,56–58</sup> or "satisfactory relief of IBS symp-

toms.”<sup>59,60</sup> These definitions are assumed to have face validity. However, empirical data are needed for each outcome measure to assess the clinical significance of different degrees of change from both the patient’s and the physician’s perspectives.

*One or at most 2 primary outcome measures should be specified in advance. Investigators should list criteria to classify each patient as a responder or nonresponder based on a clinically meaningful change in symptoms.*

### Choosing a Primary Outcome Variable

In selecting a primary outcome, investigators should examine the trial objectives, population, and mechanism of action of the proposed treatment and should choose either a global measure, which integrates the symptoms into a single numerical index, or the summary score of a validated symptom severity and/or frequency questionnaire.

Attention should be paid to the suitability of the measurement scale used for each outcome measure. A detailed discussion of measurement scales and their properties is beyond the scope of this report, but is more thoroughly addressed in the Rome III book and elsewhere.<sup>2,61</sup>

Physician-reported assessments have been accepted in some studies,<sup>14</sup> but are subject to greater measurement error than patient reports.<sup>62</sup> Therefore, patient-reported measures are endorsed. Only fully validated instruments are recommended as primary outcome assessment tools, and secondary outcome measures should also be assessed for robustness. Psychometric validation requires that (1) the assessment instrument includes symptoms relevant to and fully representative of the disorder (*face validity*); (2) it show a predictable relationship with other measures (*construct validity*); (3) the assessment produces similar results when readministered to patients whose health status has not changed (*reliability*); (4) it can detect clinically meaningful change in health status when such a change has occurred (*responsive*); and (5) changes in score can be related to clinical indicators that are meaningful to clinicians (*criterion validity*).

Validation of a new outcome measure is best established in a separate study.<sup>63</sup> The frequency of data recording for each outcome should also be specified before the trial begins, as should the time frame defining the patient response (whether at the end of the trial, during a prespecified proportion of weeks or months that responder criteria have been fulfilled, or for all time points assessed during the trial).

*A patient-reported outcome assessment is recommended. Psychometric validation of each outcome measure is recommended before it is used in clinical trials.*

**Adequate relief or satisfactory relief as a primary outcome measure.** Since 1999, most published pharmaceutical trials for IBS have used “adequate relief of abdominal pain and discomfort”<sup>4,56–58</sup> or “satisfactory relief of IBS symptoms”<sup>59,60</sup> as their primary outcome measure. Responders were defined as patients who reported “yes” to adequate relief or satisfactory relief on at least half of the weeks in the treatment trial. These studies demonstrated statistically significantly higher responder rates for active drug relative to placebo and led to approvals for alosetron and tegaserod by the Food and Drug Administration.

Mangel et al<sup>64</sup> assessed the validity of the adequate relief measure in diarrhea-predominant IBS patients and showed that responders differed significantly from nonresponders regarding pain-free days, pain severity, urgency, stool frequency, and 6 of 8 SF-36 quality of life subscales plus 8 of 9 scales on a disease-specific quality of life measure. However, correlations among measurements (convergent validity), test–retest reliability, and internal consistency were not reported. Similar validation data have been reported for satisfactory relief.<sup>55,65</sup>

**Integrative symptom questionnaires.** An alternative method for defining a responder in an IBS treatment trial is to ask patients to report the frequency or severity of all (or a representative group) of IBS symptoms prior to and again following treatment, and to define a responder as a patient who reports at least a 50% decrease in IBS symptom severity.<sup>54,55,66</sup> There are several questionnaires that examine the severity of IBS, such as the Gastrointestinal Symptom Rating Scale for IBS<sup>67</sup> and the Functional Bowel Disorder Severity Index.<sup>68</sup> However, the Irritable Bowel Syndrome Symptom Severity Scale<sup>69</sup> is the only IBS symptom severity scale that has been shown to be responsive to treatment effects.<sup>17,69,70</sup>

Whitehead et al<sup>55</sup> compared different outcome measures including satisfactory relief and a 50% reduction in the Irritable Bowel Syndrome Symptom Severity Scale questionnaire, in an observational study of patients’ response to usual medical care for IBS. They reported that the response rate on satisfactory relief was influenced by pretreatment symptom severity: patients with initially mild IBS symptoms showed the highest responder rate but the smallest change in symptom severity, whereas patients with initially severe IBS symptoms showed the lowest responder rate but the largest decrease in severity. In contrast, when defining a responder as a patient who reported at least a 50% decrease in symptom severity, pretreatment symptom severity had no impact on the responder rate. Defining a responder based on a 50% reduction in symptoms has been used in several stud-

ies<sup>54,71</sup> and has been endorsed by 1 expert panel.<sup>72</sup> However, like satisfactory relief and adequate relief, it requires further validation.

Subjects can be classed as responders and nonresponders at different time points during a trial. In published trials, patients were classified as responders if they reported adequate relief or satisfactory relief on at least 50% of weeks over a 1- to 3-month period.<sup>4,39,53,57,59,60</sup> However, this loses important information; the most persuasive evidence for efficacy would be to show that patients in the active treatment had a sustained response once they reported satisfactory or adequate relief. Investigators are encouraged to use more sophisticated statistical models that address the longitudinal trajectory of responder status.<sup>73,74</sup> At a minimum, studies should report the proportion of patients responding at each time point and throughout the trial.

Several well-validated outcome measures have been used in FD trials.<sup>75</sup> These use a single global outcome of a specific symptom (eg, Glasgow Dyspepsia Severity Score<sup>76</sup>), a global overall assessment of dyspepsia symptoms (eg, the Canadian Dyspepsia Score<sup>77</sup>), the LEEDS Dyspepsia Questionnaire,<sup>78</sup> or several questions covering important dyspepsia and quality of life outcomes (eg, the Severity of Dyspepsia Assessment,<sup>79</sup> the Dyspepsia Symptom Questionnaire,<sup>67</sup> the Quality of Life in Reflux and Dyspepsia Questionnaire,<sup>67</sup> and the Nepean Dyspepsia Questionnaire<sup>80</sup>). For some FGIDs, such established outcome measures are yet to be developed.

Pain or discomfort is a key feature of many FGIDs and is typically either the primary outcome variable or an important secondary outcome variable in clinical trials. Pain has 3 dimensions—intensity, duration, and frequency—that can be considered separately or integrated in a global assessment of pain or can be incorporated into a quality of life measure. Different rating scales can be used that are reproducible and sensitive to change.<sup>81</sup> If pain is chosen as the primary outcome, a meaningful clinical response should be defined beforehand, and the proportion of patients reaching this end point reported.

*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.*

**Safety issues and absence of harms.** Every trial should document and report adverse events. Recent attention has focused on the appropriate reporting of harms-related issues in randomized clinical trials.<sup>82</sup> When collecting harms data is a trial objective, it should be reflected in the manuscript reporting the study results and the report should clearly define adverse events and

how they were measured. Investigators should attempt to place benefits and harms for any intervention into perspective.

*Anticipated and unanticipated adverse events should be reported.*

### Choosing Secondary Outcome Variables

The reasons for including each secondary outcome and the plan for analysis should clearly be identified before the trial begins. Health economic outcomes are becoming an important class of secondary outcomes.<sup>83,84</sup>

*Secondary outcomes should be selected based on the study question and should be validated measures that support or explain the results. Integrating health economic outcomes is recommended when feasible.*

**Quality of life assessment.** FGIDs significantly impact quality of life.<sup>85,86</sup> Generic and disease-specific quality of life instruments are available.<sup>63</sup> Generic instruments can assess quality of life in large populations and across a wide spectrum of disorders, but may not reflect all important aspects of health status for specific disorders. They may be less sensitive to detect important treatment effects, but they permit comparisons with other diseases and help to detect unexpected changes in health status after treatment. Examples of validated instruments include the Sickness Impact Profile,<sup>87</sup> the Nottingham Health Profile, the SF-36 (Short Form of General Health Questionnaire),<sup>88</sup> and the Psychological General Well-Being Index.<sup>89</sup> Disease-specific quality of life instruments<sup>75,90,91</sup> examine problems specific to the FGID (eg, the fear of fecal incontinence in IBS). Theoretically, they can detect smaller and more specifically relevant changes in health status, which may be missed by generic instruments. Quality of life measures have not been used as the primary outcomes in pharmaceutical clinical trials because they were believed to be insufficiently responsive to treatment, but have been strongly recommended as secondary outcome variables. One report focusing on the health-related quality of life data from two previously reported trials of alosetron found a significantly greater improvement on active drug compared to placebo,<sup>92</sup> challenging the belief that these measures are not responsive enough to be employed as primary outcome variables.

*Quality of life assessments are important secondary outcomes. Investigators are encouraged to include both a baseline generic and a pre-post disease-specific quality of life instrument.*

### Analysis and Data Reporting

The type of statistical analysis is determined by the particular study design and primary outcome mea-

sure(s). The Consolidated Standards for Reporting Trials (CONSORT) statement was developed by scientists and editors to improve the quality of reporting parallel group, randomized, controlled trials.<sup>93</sup> It emphasizes the importance of transparently reporting the study objective and how the study was conducted and analyzed. Evidence supports improved quality of methodology and data reporting<sup>93</sup> when CONSORT guidelines are used. Many journals now require that manuscripts describing clinical trials conform to the CONSORT guidelines, found on the web at [www.consort-statement.org](http://www.consort-statement.org). Recent publications have made similar recommendations for studies evaluating diagnostic testing (Standards for Reporting of Diagnostic Accuracy initiative<sup>94</sup>) and reporting meta-analyses (Quality of Reports of Meta-Analyses statement<sup>95</sup>).

*Investigators should adhere to the CONSORT statement on reporting of clinical trials.*

The main analysis for FGID trials should focus on the primary outcome measure(s) to determine whether or not the study results support a new treatment. Although the main outcome often compares the end of treatment and baseline observations, data should also describe how patients changed during the study; the results of a trial are far more compelling if patients have had a sustained response to the intervention. When 2 primary outcome variables are included in a trial, investigators should specify in advance whether the trial will be considered positive if only 1 outcome measure is significant, or if both are required. If significance on any primary outcome suffices, the analysis should adjust for multiple comparisons, for example, using the Bonferroni correction.<sup>96</sup> The committee suggests that the EMEA recommendation requiring 2 positive primary outcomes for trials in IBS may be overly conservative. The primary outcome results should be stated in absolute numbers to include both a numerator and denominator; it is not sufficient to list only percentages of (non)-responders (eg, not 20% but rather 10/50, 20%). For all outcome measures, the estimated effect of the intervention (difference between active and placebo treatment) and a 95% 2-sided confidence interval should also be included.<sup>97</sup>

*The main result of the study must be based on the evaluation of the primary outcome measure as stated in the protocol before the study begins. The primary outcome should be stated in absolute numbers and should include a 95% confidence interval.*

Statistically significant differences between study groups can also be expressed using a *P* value. Actual values and not thresholds (ie, not  $P < .05$ ) should be provided and should be complementary to confidence intervals. The reciprocal of the absolute risk reduction, in a risk reduction trial, or therapeutic gain, in a treatment

efficacy trial, can also allow computation of the number of patients who need to be treated (NNT) to encounter a patient who will experience a clinical benefit. Although the NNT is reported infrequently in randomized, controlled trials, its inclusion can convey the clinical importance of a study result.<sup>98</sup> Similarly, harms data can be used to estimate the number of patients that would need to be treated with a drug to see an adverse event (number needed to harm [NNH]). Calculation of the NNT and NNH allows the researcher or clinician to more quantitatively assess the benefits and risks of any given therapy.

*When reporting *P* values, actual values and not thresholds should be provided. An NNH can be calculated based on the risk of adverse effects and can be weighed against the NNT.*

The statistical analysis should be based on an intention-to-treat (ITT) principle<sup>99</sup> with a plan for handling dropouts. The trial can either be analyzed as the proportion of responders in each group, treating all dropouts as nonresponders, or by carrying forward the last observation available for the primary outcome. A dual analysis, examining for differences in results using the 2 different methods should be performed. Many studies also report a per protocol (all patients who followed the protocol) or an all-patients-treated (all patients who received treatment following randomization) analysis. These analyses may provide insight as to whether a treatment works under optimal conditions, but cannot replace the ITT analysis. When there is a discrepancy between the ITT (negative) and per protocol (positive) analyses, the results should be interpreted as inconclusive. The effect of potential modifiers such as gender, age, duration or severity of disease, and presence of psychological stress can be assessed using a logistic regression analysis, where the binary dependent variable represents the a priori specified definition of a responder.<sup>100</sup> Such covariates should also be prespecified.

*The primary analysis should be the ITT analysis and must include all patients randomized.*

## Secondary Outcome Measures and Subgroups

Results should be reported for all prespecified outcomes. Score changes should be reported for each cardinal symptom of the entry criteria. Secondary outcomes that are used to support or refute the primary analysis should be analyzed by ITT and not per protocol. Adjustment for multiple comparisons is generally unnecessary when analyzing secondary outcome measures because the efficacy of the treatment is judged on the basis of the analysis of the primary outcome variable, not the secondary outcomes. Sec-

ondary outcome measures are examined to support the primary outcome analysis. When many secondary variables are included to identify predictors of response or explore for other benefits, the type I error rate may be inflated and can be adjusted.<sup>96</sup> However, the Bonferroni correction may be too conservative<sup>96</sup> and can increase the likelihood of a type II statistical error, rendering truly important differences nonsignificant. Using descriptive rather than inferential statistics (eg, means and confidence intervals) or reporting actual *P* values is a possible solution.

Specific plans to present and analyze harms data should be clearly described and withdrawals from each arm of the trial should be detailed. ITT is the preferred analysis for harms data.<sup>82</sup>

Exploratory subgroup analyses are commonly performed in trials of FGIDs, although their validity is controversial.<sup>101</sup> The recommended test of interaction<sup>101</sup> evaluates for differences in treatment effects between complementary subgroups (eg, older and younger subjects), rather than simply comparing *P* values for each subgroup, thereby maintaining statistical power.

*Secondary analyses used to support an efficacy claim should be ITT analyses. Harms data should be analyzed by ITT when possible, but absolute incidence rates and 95% confidence intervals should also be provided.*

### Sample Size and Power

The protocol should present and clearly specify the assumptions underlying the sample size calculation. These elements include the minimum effect size (difference in primary outcome between groups) that the trial is designed to detect, the  $\alpha$  (type I) error level, the statistical power or  $\beta$  (type II) error level, and when evaluating continuous outcomes (eg, difference in severity scores), the standard deviation of the difference. Recent trials have been powered to detect differences as small as 10%,<sup>58</sup> 12%,<sup>59,60</sup> or 15%.<sup>4,57</sup> Often, a power of 80% is used ( $\beta$  error or type II error of 20%) and  $\alpha$  (type I) error of 5% using a 2-sided test. An allowance for dropouts should also be made in determining the appropriate sample size, but efforts should be made to keep the dropout rate below 10%–20%. It is inappropriate for an investigator to conclude, from an inadequately powered study that fails to find a statistically significant difference between interventions, that the 2 interventions are equivalent.<sup>44</sup>

*A sample size calculation should be routinely performed and should be based on the expected behavior of the primary outcome measure.*

### Interim Analysis and Stopping Rules

There is no compelling reason to incorporate interim analyses in trials to determine efficacy because FGIDs are not life threatening. Moreover, because the incidence of serious adverse events is expected to be low, any occurrence of a serious adverse event is likely to prompt the safety committee to reevaluate the trial without carrying out an interim analysis. Thus, interim analyses in trials of FGIDs are normally only done to assess the futility of continuing the trial. Plans for interim analyses should be clearly prespecified in the study protocol and appropriate statistical methods to adjust for multiple comparisons are necessary.<sup>101,102</sup> The most common method is to partition the  $\alpha$  level for the trial by subtracting the  $\alpha$  level for the interim analysis from the  $\alpha$  level intended for the final analysis. Consequently, most investigators use a conservative  $\alpha$  level, such as .001, for the interim analysis so that sufficient power is reserved for the final analysis. If an interim analysis is preplanned,  $\alpha$  sharing can be incorporated when calculating the sample size. Unplanned preliminary analyses should be avoided; premature presentation of results may affect the further conduct of the trial and can lead to the reporting of inaccurate observations.

There are few guidelines for conducting interim analyses to assess the futility of continuing a trial. However, to preserve the credibility of the investigators (a) such analyses should be overseen by a Data and Safety Monitoring Board that is independent from the investigators, (b) the analysis should test for equivalence rather than superiority of 1 treatment relative to the other, and (c) liberal equivalence margins for the effect size should be defined a priori and will likely be wider than those applied to serious harms.

*Interim analyses are not recommended because they may jeopardize the trial integrity unless there is reason to believe participation in the trial (either in the active treatment or control group) places the patient at risk.*

### Ethical Issues

The main result of a trial must be presented according to the predetermined primary outcome measure(s). Selecting a primary outcome measure after the trial is concluded inflates the type I error rate and is misleading. Unexpected results that were not part of the original hypothesis<sup>103</sup> should be considered as purely exploratory, for testing in future studies. Adherence to study goals is strengthened when an independent advisory group is assembled.

*Changing the primary outcome measure(s) in the analysis phase of a study should not be done; it invalidates the statistical*

**Table 2.** Recommendations for Future Research

1. Examine the periodicity and severity of symptoms in natural history studies.
2. Evaluate the multidimensional construct of symptom severity (eg, frequency, number present, clustering, severity, contribution to "global severity," and changes in primary symptoms over time).
3. Examine the influence of disease modifiers (predictors) such as disease duration, baseline severity, psychological status, comorbidity, surgeries, and response to prior treatments.
4. Investigate what contributes to the placebo response in different FGIDs and how to minimize its impact on efficacy assessment.
5. Evaluate the impact of baseline observations and diagnostic testing on symptoms, data quality, and treatment response.
6. Further validate adequate and satisfactory relief during clinical trials.
7. Develop, validate fully, and determine minimal clinically important differences for new outcome measures and disease-specific quality of life instruments. Catalog and critically appraise them.
8. Further evaluate and validate definitions of the treatment responder measure(s) including a 50% reduction in symptom severity and ensure that the definitions are clinically relevant.
9. Develop and validate trial designs for testing on-demand treatments for intermittent symptoms.
10. Examine the impact of CONSORT, EMEA, and Food and Drug Administration guidelines on study quality.

*analysis and renders the conclusions of uncertain value by inflating the chances of a type I error.*

Concern has been raised that several negative FGID treatment trials have not been published, overestimating the efficacy of some treatments and/or diminishing safety concerns. Investigators are ethically obliged to publish the results of all completed studies, and journal editors should publish methodologically sound studies, whether results are negative or positive. Some journals now require investigators to register clinical trials before initiation, and failure to do so bars their publication by subscribing journals.<sup>104</sup> The Cochrane Collaboration systematic reviews also underscore the need for publication of all relevant studies.<sup>105</sup>

*It is unethical to withhold publishing the results of a completed trial.*

In reviewing the relevant literature for this report, the committee identified a number of areas that require additional evaluation. These recommendations for future research are listed in Table 2.

## References

1. Irvine EJ, Whitehead WE, Chey WD, Matsueda K, Talley NJ, Shaw M, Veldhuyzen van Zanten SJO. Design of treatment trials for functional gastrointestinal disorders. In: Drossman DA, Corazzari E, Delvaux M, Talley NJ, Thompson WG, Spiller RC, Whitehead WE, eds. The functional gastrointestinal disorders: diagnosis, pathophysiology and treatment. A multinational consensus. 3rd ed. McLean, VA: Degnon Associates, 2006.
2. Veldhuyzen van Zanten SJ, Talley NJ, Bytzer P, Klein KB, Whorwell PJ, Zinsmeister AR. Design of treatment trials for functional gastrointestinal disorders. *Gut* 1999;45(Suppl 2):II69-II77.
3. Talley NJ, Nyren O, Drossman DA, Heaton KW, Veldhuyzen van Zanten SJO, Koch MM, Ransohoff DF. The irritable bowel syndrome: toward optimal design of controlled treatment trials. *Gastroenterology International* 1993;189-211.
4. Camilleri M, Northcutt AR, Kong S, Dukes GE, McSorley D, Mangel AW. Efficacy and safety of alosetron in women with irritable bowel syndrome: a randomised, placebo-controlled trial. *Lancet* 2000;355:1035-1040.
5. Muller-Lissner SA, Fumagalli I, Bardhan KD, Pace F, Pecher E, Nault B, Ruegg P. Tegaserod, a 5-HT(4) receptor partial agonist, relieves symptoms in irritable bowel syndrome patients with abdominal pain, bloating and constipation. *Aliment Pharmacol Ther* 2001;15:1655-1666.
6. Fass R, Longstreth GF, Pimentel M, Fullerton S, Russak SM, Chiou CF, Reyes E, Crane P, Eisen G, McCarberg B, Ofman J. Evidence- and consensus-based practice guidelines for the diagnosis of irritable bowel syndrome. *Arch Intern Med* 2001;161:2081-2088.
7. Cash BD, Schoenfeld P, Chey WD. The utility of diagnostic tests in irritable bowel syndrome patients: a systematic review. *Am J Gastroenterol* 2002;97:2812-2819.
8. Jones R. Likely impacts of recruitment site and methodology on characteristics of enrolled patient population: irritable bowel syndrome clinical trial design. *Am J Med* 1999;107:85S-90S.
9. Talley NJ, Meineche-Schmidt V, Pare P, Duckworth M, Raisanen P, Pap A, Kordecki H, Schmid V. Efficacy of omeprazole in functional dyspepsia: double-blind, randomized, placebo-controlled trials (the Bond and Opera studies). *Aliment Pharmacol Ther* 1998;12:1055-1065.
10. Longstreth GF, Hawkey CJ, Mayer EA, Jones RH, Naesdal J, Wilson IK, Peacock RA, Wiklund IK. Characteristics of patients with irritable bowel syndrome recruited from three sources: implications for clinical trials. *Aliment Pharmacol Ther* 2001;15:959-964.
11. Camilleri M, Atanasova E, Carlson PJ, Ahmad U, Kim HJ, Viamontes BE, McKinzie S, Urrutia R. Serotonin-transporter polymorphism pharmacogenetics in diarrhea-predominant irritable bowel syndrome. *Gastroenterology* 2002;123:425-432.
12. Guthrie E, Barlow J, Fernandes L, Ratcliffe J, Read N, Thompson DG, Tomenson B, Creed F. Changes in tolerance to rectal distension correlate with changes in psychological state in patients with severe irritable bowel syndrome. *Psychosom Med* 2004;66:578-582.
13. Spiller RC. Problems and challenges in the design of irritable bowel syndrome clinical trials: experience from published trials. *Am J Med* 1999;107:91S-97S.
14. Veldhuyzen van Zanten SJ, Cleary C, Talley NJ, Peterson TC, Nyren O, Bradley LA, Verlinden M, Tytgat GN. Drug treatment of functional dyspepsia: a systematic analysis of trial methodology with recommendations for design of future trials. *Am J Gastroenterol* 1996;91:660-673.
15. Hahn B, Watson M, Yan S, Gunput D, Heuierjans J. Irritable bowel syndrome symptom patterns: frequency, duration, and severity. *Dig Dis Sci* 1998;43:2715-2718.
16. Drossman DA, Thompson WG. The irritable bowel syndrome: review and a graduated multicomponent treatment approach. *Ann Intern Med* 1992;116:1009-1016.
17. Whitehead WE. Control groups appropriate for behavioral interventions. *Gastroenterol* 2004;126:S159-S163.
18. FDA updates warnings for cisapride. FDA Talk Paper T00-6. Rockville, MD, 2000.
19. Camilleri M. Safety concerns about alosetron. *Arch Intern Med* 2002;162:100-101.

20. Sackett DL. Bias in analytic research. *J Chronic Dis* 1979;32:51–63.
21. Spilker B. Choosing and validating the clinical trial's blind. *Guide to clinical trials*. New York: Raven Press, 1991:15–20.
22. Altman DG. Randomisation. *Br Med J* 1991;302:1481–1482.
23. Altman DG. Comparability of randomised groups. *The Statistician* 1985;34:125–136.
24. Spilker BI. Randomization procedures. *Guide to clinical trials*. New York: Raven Press, 1991:69–73.
25. Temple RJ. When are clinical trials of a given agent vs. placebo no longer appropriate or feasible? *Control Clin Trials* 1997;18:613–620.
26. Guthrie E, Creed F, Dawson D, Tomenson B. A controlled trial of psychological treatment for the irritable bowel syndrome. *Gastroenterology* 1991;100:450–457.
27. Borkovec TD, Nau SD. Credibility of analogue therapy rationales. *J Behav Ther Exp Psychiatry* 1972;3:257–260.
28. Drossman DA, Toner BB, Whitehead WE, Diamant NE, Dalton CB, Duncan S, Emmott S, Proffitt V, Akman D, Frusciantie K, Le T, Meyer K, Bradshaw B, Mikula K, Morris CB, Blackman CJ, Hu Y, Jia H, Li JZ, Koch GG, Bangdiwala SI. Cognitive-behavioral therapy versus education and desipramine versus placebo for moderate to severe functional bowel disorders. *Gastroenterology* 2003;125:19–31.
29. Thompson WG. Placebos: a review of the placebo response. *Am J Gastroenterol* 2000;95:1637–1643.
30. Bland JM, Altman DG. Some examples of regression towards the mean. *BMJ* 1994;309:780.
31. Pitz M, Cheang M, Bernstein CN. Defining the predictors of the placebo response in irritable bowel syndrome. *Clin Gastroenterol Hepatol* 2005;3:237–247.
32. Howarth PH, Stern MA, Roi L, Reynolds R, Bousquet J. Double-blind, placebo-controlled study comparing the efficacy and safety of fexofenadine hydrochloride (120 and 180 mg once daily) and cetirizine in seasonal allergic rhinitis. *J Allergy Clin Immunol* 1999;104:927–933.
33. Bachert C, Brostoff J, Scadding GK, Tasman J, Stalla-Bourdillon A, Murrieta M. Mizolastine therapy also has an effect on nasal blockade in perennial allergic rhinoconjunctivitis. RIPERAN Study Group. *Allergy* 1998;53:969–975.
34. Berger VW, Rezvani A, Makarewicz VA. Direct effect on validity of response run-in selection in clinical trials. *Control Clin Trials* 2003;24:156–166.
35. Hills M, Armitage P. The two-period cross-over clinical trial. *Br J Clin Pharmacol* 2004;58:S703–S716.
36. Committee for Proprietary Medicinal Products (CPMP). Notes for guidance on statistical principles for clinical trials. ICH/363/96. London, UK: European Agency for Evaluation of Medicinal Products, 1998.
37. Cleophas TJ, Zwinderman AH. Limitations of randomized clinical trials. Proposed alternative designs. *Clin Chem Lab Med* 2000;38:1217–1223.
38. Committee for Proprietary Medicinal Products (CPMP). CPMP/EWP/785/97. Points to consider on the evaluation of medicinal products for the treatment of IBS. 785/97. European Agency for the Evaluation of Medicinal Products, London, England 2003.
39. Tack J, Muller-Lissner S, Bytzer P, Corinaldesi R, Chang L, Viegas A, Schnekenbuehl S, Dunger-Baldauf C, Rueegg P. A randomised controlled trial assessing the efficacy and safety of repeated tegaserod therapy in women with irritable bowel syndrome with constipation (IBS-C). *Gut* 2005;54:1707–1713.
40. Bardhan KD. Intermittent and on-demand use of proton pump inhibitors in the management of symptomatic gastroesophageal reflux disease. *Am J Gastroenterol* 2003;98:S40–S48.
41. Thompson WG, Longstreth G, Drossman DA, Heaton K, Irvine EJ, Muller-Lissner S. Functional bowel disorders and functional abdominal pain. In: Drossman DA, Corazziari E, Talley NJ, Thompson WG, Whitehead WE, eds. *Rome II: the functional gastrointestinal disorders*. 2nd ed. McLean, VA: Degnon Associates, 2000:351–432.
42. Tfelt-Hansen P, Block G, Dahlof C, Diener HC, Ferrari MD, Goadsby PJ, Guidetti V, Jones B, Lipton RB, Massiou H, Meinert C, Sandrini G, Steiner T, Winter PB. Guidelines for controlled trials of drugs in migraine. 2nd ed. *Cephalalgia* 2000;20:765–786.
43. Tytgat GN, Heading RC, Muller-Lissner S, Kamm MA, Scholmerich J, Berstad A, Fried M, Chaussade S, Jewell D, Briggs A. Contemporary understanding and management of reflux and constipation in the general population and pregnancy: a consensus meeting. *Aliment Pharmacol Ther* 2003;18:291–301.
44. Tinmouth JM, Steele LS, Tomlinson G, Glazier RH. Are claims of equivalency in digestive diseases trials supported by the evidence? *Gastroenterology* 2004;126:1700–1710.
45. Tillisch K, Labus JS, Naliboff BD, Bolus R, Shetzline M, Mayer EA, Chang L. Characterization of the alternating bowel habit subtype in patients with irritable bowel syndrome. *Am J Gastroenterol* 2005;100:896–904.
46. Talley NJ, Weaver AL, Zinsmeister AR, Melton LJ III. Onset and disappearance of gastrointestinal symptoms and functional gastrointestinal disorders. *Am J Epidemiol* 1992;136:165–177.
47. Compliance in health care. Baltimore, MD: The Johns Hopkins University Press, 1979.
48. Von KM, Moore JC. Stepped care for back pain: activating approaches for primary care. *Ann Intern Med* 2001;134:911–917.
49. Means B, Nigam A, Zarrow M, Loftus EF, Donaldson MS. *Autobiographical memory for health-related events*. DHHS Publication No. PHS 89-1077. *Vital and Health Statistics Series 6. Cognitive and Survey Measurement*. Washington, DC: US Government Printing Office, 1989.
50. Sandha GS, Hunt RH, Veldhuyzen van Zanten SJ. A systematic overview of the use of diary cards, quality-of-life questionnaires, and psychometric tests in treatment trials of *Helicobacter pylori*-positive and -negative non-ulcer dyspepsia. *Scand J Gastroenterol* 1999;34:244–249.
51. Stone AA, Shiffman S, Schwartz JE, Broderick JE, Hufford MR. Patient non-compliance with paper diaries. *BMJ* 2002;324:1193–1194.
52. Harding JP, Hamm LR, Ehsanullah RS, Heath AT, Sorrells SC, Haw J, Dukes GE, Wolfe SG, Mangel AW, Northcutt AR. Use of a novel electronic data collection system in multicenter studies of irritable bowel syndrome. *Aliment Pharmacol Ther* 1997;11:1073–1076.
53. Bardhan KD, Bodemar G, Geldof H, Schutz E, Heath A, Mills JG, Jacques LA. A double-blind, randomized, placebo-controlled dose-ranging study to evaluate the efficacy of alosetron in the treatment of irritable bowel syndrome. *Aliment Pharmacol Ther* 2000;14:23–34.
54. Payne A, Blanchard EB. A controlled comparison of cognitive therapy and self-help support groups in the treatment of irritable bowel syndrome. *J Consult Clin Psychol* 1995;63:779–786.
55. Whitehead WE, Palsson OS, Levy RL, Feld AD, Von Korff M, Turner M. 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* (In press).
56. Camilleri M, Mayer EA, Drossman DA, Heath A, Dukes GE, McSorley D, Kong S, Mangel AW, Northcutt AR. Improvement in pain and bowel function in female irritable bowel patients with alosetron, a 5-HT<sub>3</sub> receptor antagonist. *Aliment Pharmacol Ther* 1999;13:1149–1159.
57. Camilleri M, Chey WY, Mayer EA, Northcutt AR, Heath A, Dukes GE, McSorley D, Mangel AM. A randomized controlled clinical trial of the serotonin type 3 receptor antagonist alosetron in

- women with diarrhea-predominant irritable bowel syndrome. *Arch Intern Med* 2001;161:1733–1740.
58. Chey WD, Chey WY, Heath AT, Dukes GE, Carter EG, Northcutt A, Ameen VZ. Long-term safety and efficacy of alosetron in women with severe diarrhea-predominant irritable bowel syndrome. *Am J Gastroenterol* 2004;99:2195–2203.
  59. Kellow J, Lee OY, Chang FY, Thongsawat S, Mazlam MZ, Yuen H, Gwee KA, Bak YT, Jones J, Wagner A. An Asia-Pacific, double blind, placebo controlled, randomised study to evaluate the efficacy, safety, and tolerability of tegaserod in patients with irritable bowel syndrome. *Gut* 2003;52:671–676.
  60. Nyhlin H, Bang C, Elsborg L, Silvennoinen J, Holme I, Ruegg P, Jones J, Wagner A. A double-blind, placebo-controlled, randomized study to evaluate the efficacy, safety and tolerability of tegaserod in patients with irritable bowel syndrome. *Scand J Gastroenterol* 2004;39:119–126.
  61. Wyrwich KW, Tardino VM. A blueprint for symptom scales and responses: measurement and reporting. *Gut* 2004;53(Suppl 4):iv45–iv48.
  62. Fallone CA, Guyatt GH, Armstrong D, Wiklund I, Degl'Innocenti A, Heels-Ansdell D, Barkun AN, Chiba N, Zanten SJ, El Dika S, Austin P, Tanser L, Schunemann HJ. Do physicians correctly assess patient symptom severity in gastro-oesophageal reflux disease? *Aliment Pharmacol Ther* 2004;20:1161–1169.
  63. Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. *Ann Intern Med* 1993;118:622–629.
  64. Mangel AW, Hahn BA, Heath AT, Northcutt AR, Kong S, Dukes GE, McSorley D. Adequate relief as an endpoint in clinical trials in irritable bowel syndrome. *J Int Med Res* 1998;26:76–81.
  65. Dunger-Baldauf C, Nyhlin H, Rueegg P, Wagner A. Subject's global assessment of satisfactory relief as a measure to assess treatment effect in clinical trials in irritable bowel syndrome (IBS). *Am J Gastroenterol* 2003;98(Suppl 1):S269.
  66. Blanchard EB, Scharff L, Payne A, Schwarz SP, Suls JM, Malamood H. Prediction of outcome from cognitive-behavioral treatment of irritable bowel syndrome. *Behav Res Ther* 1992;30:647–650.
  67. Wiklund IK, Junghard O, Grace E, Talley NJ, Kamm M, Veldhuyzen van Santen SJ, Pare P, Chiba N, Leddin DS, Bigard MA, Colin R, Schoenfeld P. Quality of Life in Reflux and Dyspepsia patients. Psychometric documentation of a new disease-specific questionnaire (QOLRAD) *Eur J Surg Suppl* 1998;583:41–49.
  68. Drossman DA, Li Z, Toner BB, Diamant NE, Creed FH, Thompson D, Read NW, Babbs C, Barreiro M, Bank L. Functional bowel disorders. A multicenter comparison of health status and development of illness severity index. *Dig Dis Sci* 1995;40:986–995.
  69. 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.
  70. Gonsalkorale WM, Miller V, Afzal A, Whorwell PJ. Long term benefits of hypnotherapy for irritable bowel syndrome. *Gut* 2003;52:1623–1629.
  71. Whitehead WE, Levy RL, Von Korff M, Feld AD, Palsson OS, Turner MJ, Drossman DA. Usual medical care for irritable bowel syndrome. *Aliment Pharmacol Ther* 2004;20:1305–1315.
  72. Corazziari E, Bytzer P, Delvaux M, Holtmann G, Malagelada JR, Morris J, Muller-Lissner S, Spiller RC, Tack J, Whorwell PJ. Clinical trial guidelines for pharmacological treatment of irritable bowel syndrome. *Aliment Pharmacol Ther* 2003;18:569–580.
  73. Twisk JWR. *Applied longitudinal data analysis for epidemiology*. Cambridge, UK: Cambridge University Press, 2003.
  74. Snijders TAB, Bosler RJ. *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. London: Sage, 1999.
  75. Fraser A, Delaney B, Moayyedi P. Symptom-based outcome measures for dyspepsia and GERD trials: a systematic review. *Am J Gastroenterol* 2005;100:442–452.
  76. El-Omar EM, Banerjee S, Wirz A, McColl KE. The Glasgow Dyspepsia Severity Score—a tool for the global measurement of dyspepsia. *Eur J Gastroenterol Hepatol* 1996;8:967–971.
  77. Veldhuyzen van Zanten SJ, Tytgat KM, Pollak PT, Goldie J, Goodacre RL, Riddell RH, Hunt RH. Can severity of symptoms be used as an outcome measure in trials of non-ulcer dyspepsia and *Helicobacter pylori* associated gastritis? *J Clin Epidemiol* 1993;46:273–279.
  78. Moayyedi P, Duffett S, Brauholtz D, Mason S, Richards ID, Dowell AC, Axon AT. The Leeds Dyspepsia Questionnaire: a valid tool for measuring the presence and severity of dyspepsia. *Aliment Pharmacol Ther* 1998;12:1257–1262.
  79. Rabeneck L, Wristers K, Goldstein JL, Eisen G, Dedhiya SD, Burke TA. Reliability, validity, and responsiveness of severity of dyspepsia assessment (SODA) in a randomized clinical trial of a COX-2-specific inhibitor and traditional NSAID therapy. *Am J Gastroenterol* 2002;97:32–39.
  80. Talley NJ, Verlinden M, Jones M. Validity of a new quality of life scale for functional dyspepsia: a United States multicenter trial of the Nepean Dyspepsia Index. *Am J Gastroenterol* 1999;94:2390–2397.
  81. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. *J Chronic Dis* 1987;40:171–178.
  82. Ioannidis JP, Evans SJ, Gotzsche PC, O'Neill RT, Altman DG, Schulz K, Moher D. Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Ann Intern Med* 2004;141:781–788.
  83. Creed F, Fernandes L, Guthrie E, Palmer S, Ratcliffe J, Read N, Rigby C, Thompson D, Tomenson B. The cost-effectiveness of psychotherapy and paroxetine for severe irritable bowel syndrome. *Gastroenterology* 2003;124:303–317.
  84. Calvert EL, Houghton LA, Cooper P, Morris J, Whorwell PJ. Long-term improvement in functional dyspepsia using hypnotherapy. *Gastroenterology* 2002;123:1778–1785.
  85. Whitehead WE, Burnett CK, Cook EW III, Taub E. Impact of irritable bowel syndrome on quality of life. *Dig Dis Sci* 1996;41:2248–2253.
  86. Talley NJ, Weaver AL, Zinsmeister AR. Impact of functional dyspepsia on quality of life. *Dig Dis Sci* 1995;40:584–589.
  87. Bergner M, Bobbitt RA, Carter WB, Gilson BS. The Sickness Impact Profile: development and final revision of a health status measure. *Med Care* 1981;19:787–805.
  88. Stewart AL, Hays RD, Ware JE Jr. The MOS short-form general health survey. Reliability and validity in a patient population. *Med Care* 1988;26:724–735.
  89. Dimenas E, Glise H, Hallerback B, Hernqvist H, Svedlund J, Wiklund I. Well-being and gastrointestinal symptoms among patients referred to endoscopy owing to suspected duodenal ulcer. *Scand J Gastroenterol* 1995;30:1046–1052.
  90. Patrick DL, Drossman DA, Frederick IO, Dicesare J, Puder KL. Quality of life in persons with irritable bowel syndrome: development and validation of a new measure. *Dig Dis Sci* 1998;43:400–411.
  91. Borgaonkar MR, Irvine EJ. Quality of life measurement in gastrointestinal and liver disorders. *Gut* 2000;47:444–454.
  92. Watson ME, Lacey L, Kong S, Northcutt AR, McSorley D, Hahn B, Mangel AW. Alosetron improves quality of life in women with diarrhea-predominant irritable bowel syndrome. *Am J Gastroenterol* 2001;96:455–459.
  93. Moher D, Schulz KF, Altman D. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA* 2001;285:1987–1991.

94. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HC. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ* 2003;326:41–44.
95. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. *Quality of Reporting of Meta-analyses*. *Lancet* 1999;354:1896–1900.
96. Perneger TV. What's wrong with Bonferroni adjustments. *BMJ* 1998;316:1236–1238.
97. Guyatt G, Jaeschke R, Heddle N, Cook D, Shannon H, Walter S. Basic statistics for clinicians: 2. Interpreting study results: confidence intervals. *CMAJ* 1995;152:169–173.
98. Nuovo J, Melnikow J, Chang D. Reporting number needed to treat and absolute risk reduction in randomized controlled trials. *JAMA* 2002;287:2813–2814.
99. Gore SM, Jones G, Thompson SG. The Lancet's statistical review process: areas for improvement by authors. *Lancet* 1992;340:100–102.
100. Katz MH. Multivariable analysis: a primer for readers of medical research. *Ann Intern Med* 2003;138:644–650.
101. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, Gotzsche PC, Lang T. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001;134:663–694.
102. DeMets DL, Pocock SJ, Julian DG. The agonising negative trend in monitoring of clinical trials. *Lancet* 1999;354:1983–1988.
103. Campbell MJ, Julious SA, Altman DG. Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. *BMJ* 1995;311:1145–1148.
104. De Angelis C, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, Kotzin S, Laine C, Marusic A, Overbeke AJ, Schroeder TV, Sox HC, Van Der Weyden MB. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. *N Engl J Med* 2004;351:1250–1251.
105. Bero L, Rennie D. The Cochrane Collaboration. Preparing, maintaining, and disseminating systematic reviews of the effects of health care. *JAMA* 1995;274:1935–1938.

---

Received March 2, 2005. Accepted November 3, 2005.

Address requests for reprints to: E. Jan Irvine, MD, Professor of Medicine, University of Toronto, Head, Division of Gastroenterology, 16-054 CC Wing, Saint Michael's Hospital, 30 Bond Street, Toronto, Ontario, Canada M5B 1W8.