Design of treatment trials for functional gastrointestinal disorders

S J O Veldhuyzen van Zanten, N J Talley, P Bytzer, K B Klein, P J Whorwell, A R Zinsmeister

Chair, Committee on Design of Treatment Trials for Functional Gastrointestinal Disorders, Multinational Working Teams to Develop Diagnostic Criteria for Functional Gastrointestinal Disorders (Rome II), Division of Gastroenterology, Dalhousie University, Halifax, Canada
S J O Veldhuyzen van Zanten

Co-Chair, Committee on Design of Treatment Trials for Functional Gastrointestinal Disorders, Multinational Working Teams to Develop Diagnostic Criteria for Functional Gastrointestinal Disorders (Rome II), Department of Medicine, University of Sydney, Nepean Hospital, New South Wales, Australia
N J Talley

Department of Medical Gastroenterology, Glostrup University Hospital, Glostrup, Denmark
P Bytzer

International Drug Development Consulting, Bainbridge Island, Washington, USA
K B Klein

Department of Medicine, University Hospital of South Manchester, Manchester, UK
P J Whorwell

Section of Biostatistics, Mayo Clinic, Rochester, MN, USA
A R Zinsmeister

Correspondence to:
S J O Velduyzen van Zanten, MD, Queen Elizabeth II Health Sciences Center, Victoria General Hospital Site, Room 928, Centennial Building, 1278 Tower Road Halifax, Nova Scotia B3H 2Y9, Canada. Email: zanten@is.dal.ca

Abstract

Until recently many clinical trials of functional gastrointestinal disorders (FGIDs) suffered from important weaknesses in trial design, study execution, and data analysis. This makes it difficult to determine whether truly efficacious therapies exist for these disorders. One of the important methodologic problems is the absence of validated outcome measures and lack of consensus among stakeholders on how to measure outcome. Currently much of the effort is being put into the development of validated outcome measures for several of the FGIDs. The randomized, controlled trial with parallel groups is the design of choice. In this report, guidelines are given for the basic architecture of intervention studies of FGIDs. Further studies on design issues are required to ensure the recommendations will become evidence based in the future.
(*Gut* 1999;**45**(Suppl II):II69–II77)

Keywords: clinical trial; random allocation; functional gastrointestinal disorder(s); dyspepsia; functional dyspepsia; irritable bowel syndrome; evidence based medicine; study design; outcome measures; Rome II

Aim of the working team

This committee was charged with developing guidelines for the design, conduct, and analysis of treatment trials in the functional gastrointestinal disorders (FGIDs). The guidelines should also enable regulatory agencies, health care providers and other interested parties to evaluate better the methodologic quality of published studies. There are few studies which have investigated issues in study design in this area.[1 2] A previous working team report has addressed design issues for the irritable bowel syndrome (IBS).[3] Some of the recommendations from that report have been incorporated into this report and expanded upon. This review is limited to design issues of clinical trials that attempt to evaluate the efficacy of treatment interventions. The report will follow the basic research architecture of randomized clinical trials, and the recommendations are highlighted. There is an urgent need for empirical data to test these recommendations in clinical trials.

The study question

> The study design should reflect the main question that the study proposes to investigate.

The main study question(s) may vary depending on the nature of the intervention being tested, and the interests and goals of the researchers and sponsors carrying out the study. Examples of main objectives of intervention studies are reduction in severity or abolition of symptoms, improvement in quality of life, improved ability to cope with symptoms, and decreased use of health care resources. Given the current lack of effective therapies most treatment trials will ask the question whether the treatment under study is efficacious—that is, leads to an improvement in severity of the symptoms for which the patient sought medical attention.

> The trial must incorporate the principles of best and usual clinical practice as much as possible to ensure that the study results are relevant to the real practice situation.

It is important that investigators keep this important principle in mind while designing the trial. It requires that all patients should be managed according to normal clinical practice after the diagnosis has been made. The patient needs to be given an adequate explanation of the disease, and when necessary, advised about obvious irregularities in the diet and eating habits prior to entering in a trial. Trials do differ from usual practice in several ways including use of placebo, use of admission criteria, frequent follow up visits with extensive data recording, and use of study coordinators. Nevertheless, it is important that standard aspects of diagnosis and management, especially adequate explanation, and reassurance about the disease are not omitted and any intervention be demonstrated to have a benefit over and above standard care.

Patient population

> In general, include a broad spectrum of patients, as defined by the Rome criteria.

The FGIDs are defined by a combination of persistent or recurrent gastrointestinal symptoms. The diagnostic criteria of the Rome working teams of FGIDs have been increasingly adopted by researchers around the world.[4] The Rome II diagnostic criteria themselves will require further validation. It is recommended that the updated diagnostic

Abbreviations used in this paper: FGID, functional gastrointestinal disorder; IBS, irritable bowel syndrome; VAS, visual analogue scales; ITT, intention-to-treat; MCID, minimal clinically important difference.

Rome criteria, which are published elsewhere in this supplement, are used for selection of patients. Depending on the type of intervention that is being studied, researchers may decide to target enrollment to a special population group which is more narrowly defined than by the Rome criteria or other accepted diagnostic criteria.

Investigators should justify special inclusion criteria.

It is beyond the scope of this article to recommend inclusion and exclusion criteria for all the FGIDs. The inclusion and exclusion criteria need to be specified.

In general, it is recommended not to subcategorize patients with functional gastrointestinal disorders because of symptom instability.

Subcategorization of patients into subgroups has become more common, especially in functional dyspepsia (e.g., ulcer-like, and dysmotility-like) and IBS (diarrhea- or constipation-predominant).[5] These subgroups have also been used to select patients for treatment trials. In dyspepsia there is evidence for considerable overlap among the subgroups.[6] Because of the noticeable instability of subgroups over time, it is preferred that studies include a broad spectrum of patients, as defined by the primary Rome diagnostic criteria. It is recognized that investigators may want to apply special inclusion criteria but then there will be a problem with generalizability. If the study wants to limit enrolment only to certain subgroups of patients (e.g., patients with constipation-predominant IBS) the reasons for this should be explained. The selection criteria need to make clinical sense and should reflect normal clinical practice. It may be impossible to prevent inclusion of some patients with overlap syndromes—for example, amongst patients with gastroesophageal reflux disease and functional dyspepsia.[7] However, it is important that investigators make clear how patients with overlapping syndromes are identified, and whether they are allowed to enrol in the study.

The patient setting needs to be clearly considered as it may affect outcome. Patients from primary, secondary or tertiary level care are likely to have different patterns of response to treatment—for example, more resistant patients are encountered in tertiary care.

Most trials of FGIDs have been conducted in academic centers. For example, in only 8% of studies on functional dyspepsia were patients enrolled at the primary care level.[2] Although empirical data are sparse, there is likely to be a difference in the severity of symptoms and response to treatment between patients from primary, secondary and tertiary care centers.

In multicenter trials patient characteristics need to be measured in detail in order to ensure comparability of patients among centers.

Most larger clinical trials are multicenter. The advantage of such studies is their generalizability, but their possible disadvantage is patient heterogeneity. It is important that information is collected about the setting in which the patients were recruited. In multicenter trials the expected number of patients to be enrolled at each center should be determined prior to the study. The average number and range of entered patients per center should be reported in the study results.

Patient characteristics

A detailed list of patient characteristics that need to be specified in the protocol is beyond the scope of this document. Some of the important patient characteristics that should be recorded include: age, sex, race, severity of disease, duration of disease, previous treatment failures for the condition under study, and coexisting medications. It may also be advisable to document the presence of psychological distress or a history of mental health problems as these factors may influence the response to treatment. (See Psychosocial aspects of functional gastrointestinal disorders in this supplement.)

Architecture of the trial

A placebo control group or adequate control group is an essential requirement.

Although certain treatments look promising (e.g., psychotherapy for IBS,[1]) or have recently been shown to be effective (the use of omeprazole in functional dyspepsia[8]), none of those treatments can be considered the current standard of care. Therefore a placebo control group is an essential requirement for intervention studies of FGIDs. The placebo response can be particularly high in FGIDs making it more difficult to show superiority of a new treatment over placebo. In functional dyspepsia the reported placebo response varied from 13 to 73%,[2] whereas for IBS the reported range has been up to 70%.[1] For certain interventions (i.e., psychotherapy, or sphincterotomy (for biliary dyskinesia)), it is difficult to design a true placebo comparison group and maintain blinding. One solution is the use of independent assessors who were not involved in the delivery of the actual treatment, or the use of sham treatments.

The parallel group design is the study design of choice.

The randomized clinical trial with parallel group design is the accepted standard for evaluation of efficacy for the vast majority of

treatments.[9] Although crossover designs have been popular in treatment trials of certain FGIDs (33% of functional dyspepsia and 50% of IBS studies), this design is not recommended.

Comments on crossover designs

Crossover designs have the theoretical advantage that by comparing within rather than between patients a smaller sample size is needed to achieve a desired power (over parallel group designs). The reason that use of the crossover design, including the multiple crossover design and N-of-1 (randomized trials in single patient) is not recommended is that the baseline assumptions are not valid in trials of FGIDs. These assumptions are that symptoms (1) remain relatively constant over time, and (2) return to baseline before crossover to the alternate treatment.[10] A washout period is usually applied to achieve the latter. A well recognized problem in the two period, two treatment crossover trials is that the response in the second period is in part dictated by the response in the first period. This is an order or sequence effect, often referred to as a period-by-treatment interaction.[10] Given the known variability in severity of symptoms, it is unlikely that a return to baseline takes place in most cases. Furthermore, in many FGID studies a continuing improvement in symptoms has been observed in the placebo-treated controls, thereby creating a period effect.[1 2]

Given the special methodologic challenges that FGID trials pose, novel designs may be worth developing. One design which might be considered is the double blind, randomized withdrawal design. This design has been successfully used in inflammatory bowel disease.[11 12] In this design all participants are started on active treatment to see whether they improve. If improvement occurs, responders are randomized to continue with either placebo or active medication. If the active medication is indeed effective, then those randomized to placebo should have a more frequent relapse of their symptoms than those remaining on active treatment.

Adequate blinding of patients and research personnel is essential.

The blinded randomized controlled clinical trial is the gold standard for conduct of studies that evaluate the efficacy of new treatments. Blinding of patients and research personnel is of vital importance due to the subjective nature of the measured responses. For some interventions it may be difficult or impossible to keep patients or investigators blinded. One solution to overcome the problem of blinding is the use of independent assessors as it ensures that the caregivers, who administered the intervention, cannot bias the recording of the outcome measures.

The randomization method needs to be adequate.

Randomization gives the highest likelihood of balancing treatment groups for factors influencing the response, including unknown prognostic factors.[13] Randomization can be stratified for factors that are known to be important predictors of an outcome to ensure they are balanced among the treatment groups. The protocol should clearly describe the method of randomization.

A placebo run-in should be avoided as it may introduce bias.

A placebo run-in phase has been used frequently in FGID trials. One theoretical advantage for its use is that it may reduce the high placebo response. The expectation is that by eliminating placebo responders, the chance of detecting a true effect in favor of the intervention may be increased. Another proposed benefit of the run-in phase is the elimination of patients with poor compliance.

There are potential disadvantages in the use of a placebo run-in phase. It is unknown whether the placebo response is sustained after the run-in phase is over. There may be a natural variation in symptoms that coincides with the placebo run-in phase. During the placebo run-in differential dropout may occur—that is, people who would have had a different response rate relative to those who stay in the trial. Indeed, one may also end up enrolling patients with more resistant symptoms who are less likely to respond. Finally, it is possible that the placebo response will level off if the trial is of sufficient duration (8–12 weeks). In functional dyspepsia there is some empirical evidence for such an effect.[14] Given these drawbacks, use of a placebo run-in phase is not recommended.

A period of baseline observation without treatment is recommended.

A placebo run-in period should be distinguished from obtaining baseline observations prior to the start of the intervention. The assessment of the severity of symptoms at baseline documents that patients in the active and control group are comparable. It can also be used for comparison to determine the efficacy of the intervention. The optimal length of time for baseline observation is uncertain and will depend on the type of disorder and type of intervention. The assessment can be retrospective but ideally is done prospectively. Some study protocols focus only on the question whether the intervention has resulted in an improvement irrespective of the baseline situation. Although one can document change (hopefully improvement) that way, measurement of the baseline situation has the benefit of documenting the state (such as severity of the disease) of the patients from which the change takes place.

Timing of the intervention

The timing of the intervention should be carefully considered.

Investigators should be aware that there might be spontaneous improvement after recent investigations. Patients will often be reassured by diagnostic tests when they do not reveal a serious underlying disease. Fear of cancer has been shown to be a frequent reason of concern in patients with functional dyspepsia who underwent gastroscopy.[15] The improvement may coincide with the baseline observation period, or may occur later during the period in which the intervention is administered. No data are available to provide guidance on the optimal timing of diagnostic studies but it may be worthwhile to avoid testing immediately before randomization. The protocol should define the timing of the diagnostic tests as part of the entry criteria.

Duration of treatment and follow up

The duration of treatment and the expected time of response should be specified in the protocol.

The duration of treatment and the expected time of response depend on the type of intervention and will vary among the different disorders. The expected time of response may be short if one is testing drugs with a known short acting effect (e.g., antispasmodic agents for IBS) or may be longer (e.g., psychotherapy in IBS).

Generally a minimum treatment duration of 8–12 weeks is recommended.

Many studies of FGIDs have been of surprisingly short duration given the chronic nature of these conditions.[1 2] The trial must be long enough to determine whether any response will be sustained. For IBS trials there was evidence that the response in studies with longer duration was less than in shorter studies.[1] Given the chronicity of the FGIDs, we recommend that generally the minimum duration of treatment should be 8–12 weeks. If investigators expect that a treatment response will occur over a very short time period, investigators may decide that a shorter duration of treatment is appropriate.

Follow up after treatment is recommended.

Extended follow up of patients after the intervention should be considered to determine the long term efficacy of treatment. Given the chronic nature of FGIDs, it is surprising how few studies follow patients after the intervention is stopped.[2] Long follow up is mainly relevant if the benefit of treatment is expected to last and thus may not be necessary for short-acting drugs.

Once the efficacy of an intervention during short term treatment (e.g., 8–12 weeks) is documented it is essential that longer term studies are conducted to document the long term efficacy of the intervention both on and off therapy.

For treatments that are proved to be efficacious during short term treatment, additional long term studies (the suggested length is 6–12 months) are essential to document the continuous efficacy and safety of the intervention. It may also be appropriate to consider evaluation of "on demand therapy" in such studies, especially if periods of remission and exacerbation of symptoms are expected to occur.

Adherence to treatment and study protocol

Compliance with treatment should be measured.

Compliance with the intervention should be assessed. Standard methods are available for this and include interviewing the patient, requesting the return of unused medication and counts of returned tablets.

Compliance with the study protocol should be measured. Documentation that the study protocol is adhered to is important. The protocol should clearly indicate how the data will be recorded. It is unclear how the study data can best be recorded and how frequently this should be done. Both questionnaires administered at regular intervals and diary cards are frequently used. Questionnaires generally provide cross-sectional information during the course of treatment—that is, the response at a given point in time. Diary cards provide longitudinal information (i.e., the response over a period of time), thereby theoretically reducing the problem of recall. There are methodologic problems with diary cards, including retrospective completion. These limitations lend themselves to technological innovation, such as automated daily telephone recording or the use of hand held electronic data collection devices.

Outcome measures

The most important outcomes in the treatment of FGIDs are those that reflect the patient's symptoms. Since individual symptoms can vary from patient to patient and from time to time, a measure of overall change in symptoms should be the primary outcome criterion.

The trial's main result should be based on the primary outcome measure.

The most important problem in study design of clinical trials of FGIDs is the lack of consensus on how to measure outcome or how to define efficacy of an intervention. Generally accepted validated outcome measures are lacking and as a result there is notable variation in the outcome measures that have been used.

An assessment which integrates the symptoms of the particular functional gastrointestinal disorder is recommended as the primary outcome measure.

The primary outcome measure(s) should focus on the severity of symptoms relevant for the particular disorder and/or the impact on quality of life. Until recently little attention has been paid to the importance of validated outcome measures.[1–3] There are currently studies ongoing that try to validate primary outcome measures for several FGIDs. However, *currently no measures for the FGIDs are sufficiently validated to be recommended unequivocally as the primary outcome measure*.

As for what the *primary outcome measure* should consist of, it is the opinion of this working team that an *assessment that* **integrates** *the key symptoms of the particular FGID be used as the primary outcome measure. The symptoms that result in a diagnosis of FGID are varied, and interact in complex ways. Thus, there is a strong argument for a primary outcome measure that allows the patient to integrate the contribution of a disparate group of symptoms into a single global clinical rating. Alternatively the primary outcome measure can be the summary score of a validated disease-specific questionnaire that evaluates the relevant aspects of a patient's symptoms and disease-related quality of life.*

It is preferable that the main assessment of outcome be done by the patient.

It is doubtful that physician assessment is more accurate or reliable than assessment by the patient, and there may be substantial inter- and intra-individual variation in physician recording of symptom severity. Therefore, physician assessment should not be used as a main outcome measure.

The study protocol should define a priori the definition of a responder or a response, namely the change in the outcome measure which is considered clinically meaningful.

The protocol should give a clear definition of a responder or response. This definition of a responder or response should incorporate the main outcome measure on which the success or failure of the trial depends. Rather than designing the trial to detect a mean change in a score among treatment groups, it is recommended that the study compares the proportion of patients who achieve the stipulated amount of improvement necessary to be qualified as a responder (e.g., the proportion of patients who become symptom-free).

The study should include measures of change for each of the symptoms which were part of the entry criteria.

Although individual symptoms may not be the primary outcome measure, the study should include measures of change for each of the symptoms that were part of the entry criteria (e.g., the Rome diagnostic criteria). Measurement of change should include deterioration and not just improvement. The expectation is that the change (and direction of change) of the cardinal symptoms of a disorder will be similar to the changes that take place for the primary outcome measure.

Measurement of pain

The previous working team report on design of IBS trials discussed measurement of pain.[3] The presence of pain or discomfort is one of the key features of many of the FGID and often is one of the outcome measures in clinical trials. Pain or discomfort is an unpleasant sensation that can be assessed by verbal and non-verbal observation, and in this context, clearly requires the cooperation of the patient. Discomfort is best described as a less severe form of pain. Pain can be considered to have three dimensions: intensity, duration and frequency. All these aspects of pain can be measured separately. Alternatively, an integrated measure of pain, such as a global assessment of severity of pain, can be used as an outcome measure. Whether such a global assessment indeed incorporates ("weighs") all the relevant aspects of pain, especially in the context of the FGIDs, is unclear. Apart from the above dimensions, one can also measure the impact that pain has on a person's ability to carry out normal daily activities or work.

For measuring pain intensity or overall severity rating scales are usually used, either ordinal (Likert) scales or visual analogue scales (VAS). Both methods have been shown to be reproducible and sensitive to change.[3] Measurement of pain can be chosen as (one of) the primary outcome(s) in clinical trials. Multidimensional scales have also been validated for measurement of pain. An example of this is the clinical McGill Pain Questionnaire which assesses three principal dimensions of pain. This includes sensory–discriminative, multinational–affective and cognitive–evaluative aspects of pain. This scale has been extensively used in clinical trials.[16]

Definition of symptoms

It is recommended that investigators comply with the definitions of symptoms suggested by the Rome I working team.[3 5] If investigators, for reasons of preference or culture, decide to measure different symptoms, they should state clearly what was meant by the terms they used. Investigators need to decide over what time periods symptoms are reported—for example, during a clinic visit, and over what time period (days, weeks or longer) are patients asked to rate the severity of their symptoms. Diary cards and daily automatic recording of symptoms by telephone can overcome the problem of recall but have other problems, which include validity and logistics.

Use of validated outcome measures (instruments) is recommended.

Validation of outcome measures is necessary before they can be accepted for use in clinical trials. Validation of a primary outcome measure or disease-specific quality of life instrument requires that: (1) it includes symptoms that are relevant and representative of the disorder; (2) the measure is reproducible, namely produces a similar result when administered to patients whose health status has not changed; (3) it is able to detect change (responsive) in relevant clinical symptoms assuming a change took place; (4) a change in outcome measures should reflect a real change in general health status.[17 18] The latter is an especially difficult methodologic problem for most FGIDs as no other hard outcome measures are available to compare (or anchor) with a new measure.

If sufficiently validated outcome measures for a particular disorder are not available investigators must be very careful in their choice of outcome measures. It is important that the chosen definition of response makes clinical sense, is rigorous, and can be easily replicated by other investigators.

Measurement scales

A detailed discussion of measurement scales is beyond the scope of this report.[19] Their use in FGIDs has been reviewed.[3] The two most popular scales to assess severity of symptoms are categorical scales (often referred to as Likert scales) and VAS. Categorical scales seem to be more commonly used than VAS but both scales perform well and the choice therefore depends on the preference of the investigators.[20] There is no conclusive evidence to recommend an optimal number of categories for categorical scales to measure change induced by the intervention. As meta-analysis is becoming so popular it would be helpful if an international consensus could be reached on the optimal number of categories for categorical scales; such a consensus would allow for comparisons and pooling of the results of different studies. We suggest the use of seven point scales because these are used in many quality of life instruments and because they are able to detect small but potentially relevant differences. Guyatt's group, using summary scores of disease-specific quality of life questionnaires, has shown (in asthma/chronic obstructive pulmonary disease and heart failure) that a 0.5 change per question (measured on seven point scale) equates to the minimal clinically important difference.[21]

Use of validated quality of life instruments measuring impact of symptoms on normal daily life is recommended, especially when such instruments have been shown to be able to detect change (responsive).

The outcome assessment of intervention studies should capture multidimensional aspects of the disease under study. This may include assessments of the severity or frequency of symptoms, measurement of health-related quality of life, evaluation of functional limitations as a result of the disease (disability), and cost benefit or cost effectiveness measures. Utility measures will not be discussed here, but they are becoming more popular as they make cost effectiveness analyses possible.

Quality of life instruments can be divided into two categories: generic and disease-specific. Generic instruments are able to assess quality of life of large populations and do not focus on any disorder in particular.[15] For many conditions disease-specific quality of life questionnaires have been developed that incorporate several dimensions, such as disease-related symptoms, general well being, functional capacity, and psychosocial functions.[15] The advantage of disease-specific instruments is that they focus on the disease of interest. For both functional dyspepsia and IBS validated instruments have become available,[22–27] but for most the responsiveness to change still needs to be tested. Further validation is clearly required before they can be recommended as main outcome measures.

Measurement of psychological status is recommended as this may be an important variable modifying outcome.

In some patients psychological factors are present that may influence symptom severity. Measurement of psychological well being at least at baseline is recommended, as it may be a determinant of the success of the intervention.

Physiological parameters should not be used as primary outcome measures, but may be used as secondary outcome measures to help explain the therapeutic effect of the intervention.

At this time no pathophysiologic parameters explain the symptoms of FGIDs so they cannot be used as primary outcome measures. Investigators may want to measure certain parameters based on the expected mode of action of an intervention, as it may help explain the presumed physiological basis of a treatment effect. Treatment that "normalizes" such parameters, without improving symptoms, is unlikely to be of value to the patient.

Frequency of data recording

Recording of symptoms at regular intervals is recommended, as it will allow for documentation of treatment response over time.

A successful treatment would be expected to demonstrate a gradual and sustained improvement from baseline rather than extensive fluctuations over the course of the study. Thus data on fluctuations should be summarized and reported. The study results will be more convincing if an improvement is detected compared with the baseline, along with observation

of a gradual and sustained response over the course of the study.

An a priori specification of the time interval over which a responder or response occurs should be included. This should usually be toward the end of the treatment trial. Investigators should specify the time interval over which they expect the effect to occur. In many instances this will be toward the end of the intervention, but sometimes the final effect of the treatment (e.g., with psychotherapy) may occur later.

Statistical analysis and data reporting

> In reporting the results of the study, investigators should adhere to the CONSORT guidelines on reporting of clinical trials.

The CONSORT guidelines have made recommendations about the necessity of a detailed flowchart that describes how patients progressed through the study. It is recommended that the CONSORT guidelines, now adopted by several leading medical journals, are followed in reporting trial results.[28] It requires that data are provided on the number of patients: those enrolled in the study, those randomized, those lost to follow up or withdrawn (with reasons for dropout provided), and those who completed the trial. CONSORT also gives guidelines on the main components of reporting of the study results.

> The main result of the study must be based on evaluation of the primary outcome measure, which is stated in the protocol before the study begins.

The main analysis should focus on the chosen primary outcome measure on which the overall conclusion of the study is based. This should determine whether the study has a positive or a negative result. Although the main outcome often will be reported as a comparison between the end of treatment result and baseline, it is also important that data are provided on how patients changed throughout the course of the study. A situation in which patients are classified as responders some of the time and as failures at other times is far less convincing than results indicating patients have a sustained response after the intervention is started.

> The main analysis will depend on the definition of the primary outcome measure.

A detailed discussion of data analysis cannot be provided as the type of statistical analysis will depend on the study design and definition of the primary outcome measure(s). The statistical analysis should be based primarily on an intention-to-treat (ITT) analysis. 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 into how a treatment might work under optimal conditions, but it generally cannot replace the ITT analysis. The study report should clearly state the numbers of patients for all analyses and how these numbers were arrived at. The number of patients who were lost to follow up needs to be stated and when this occurred. **For all outcome measures the results should state the estimated effect of the intervention (difference between active and placebo treatment) and a 95% confidence interval**.[29] The results should be stated in absolute numbers. It is not sufficient just to list percentages (e.g., not 20%, but 10/50, 20%).

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. The CONSORT statement makes recommendations about key elements of statistical reporting, and investigators should adhere to them.

It is recommended that changes in all the symptoms that are part of the entry criteria be reported. Most likely, the different symptoms of a FGID are not independent of each other, and hence they may respond in a similar fashion. Reporting on these secondary outcome measures will help to support the direction and magnitude of the effect of the intervention on the primary outcome measure. Analysis of numerous outcome measures post hoc will inflate the overall type I error rate due to multiple comparisons[30] and are one reason that many FGID treatment trials came to the conclusion that the intervention was superior to placebo. One approach to deal with this problem is to adjust for the number of comparisons—for example, use of the Bonferroni correction. (This correction divides the α-level, usually set at 0.05, by the number of comparisons.) It has been argued that the Bonferroni correction may be too conservative.[31] It also may increase the likelihood of type II errors, so that truly important differences are deemed non-significant. Because the potential problem of multiple comparisons often affects the interpretation of results, it is recommended that secondary outcome measures be mainly analyzed using descriptive statistics (e.g., summarizing the response outcomes using 95% confidence intervals).[32] If investigators a priori specify a small number of key symptoms as important outcome measures, correction for multiple comparisons of each specified outcome measure is not necessarily required.

> The sample size calculation should be based on the expected behavior of the primary outcome measure.

The protocol should specify the assumptions on which the sample size calculation was based. This should include what is considered (a priori) to be the minimal clinically important difference (MCID) in the proportion of responders between active treatment and placebo. The study must have sufficient power

to detect the MCID.[33] Often a power of 80% is used (type II error of 20%) and type I error of 5% using a two sided test. An allowance for dropouts should be incorporated and efforts should be made to keep the dropout rate below 10–20%. The number and timing of the dropouts should be reported. For dropouts occurring in studies of longer duration it is reasonable to use the last observation of the patient while in the study in order not to loose all the information gathered from this individual.

Ethical issues

> It is unethical to change the primary outcome measure(s) in the analysis phase of the study.

The main results of a trial must be presented according to the predetermined primary outcome measure(s). It is misleading if the original stated primary outcome measures are exchanged for other outcome measures, which supported the active treatment. A secondary analysis may lead to the discovery of positive results that were not part of the original hypotheses that the study was designed to answer. Such data should be regarded as hypothesis generating rather than hypothesis testing. The positive results of such secondary analyses should be further tested in subsequent studies.

It is unethical not to publish the results of a completed trial. Investigators have an ethical obligation to publish the results of all completed studies regardless of whether the results were positive or negative. There is a concern that negative FGID studies have not always been published. Now that there is a concerted effort underway to review systematically the entire medical literature through the Cochrane collaboration[34] it is important that all studies are published. This also means that journal editors have an obligation to publish methodologically sound studies with negative results.

Recommendations for future research

- Validated outcome measures and disease-specific quality of life instruments should be developed for FGIDs. Better validated outcome measures are needed for treatment trials of FGIDs. Innovative methods to determine what amount of clinical improvement is clinically relevant (i.e., determination of the MCID) are also needed.
- Studies are needed on the associations between physiological and psychological measures and symptom outcome. At the moment there are no physiologic parameters suitable for use as primary outcome measures in treatment trials of FGIDs. More basic research is necessary to help unravel the pathophysiologic abnormalities which cause the symptoms.
- The natural history of the functional gastrointestinal disorders should be studied. Studies are also needed that investigate the natural fluctuation in symptoms to help identify the optimal duration for studies. There are surprisingly few data on the natural history and the natural fluctuation in symptoms of FGIDs. Such data are necessary to make firm recommendations on optimal duration of studies.
- The effect of recent diagnostic tests on treatment response should be studied. The magnitude, frequency and duration of the reassurance effect of diagnostic studies is unclear and warrants further study.
- Studies on the placebo response in FGIDs, including studies of the patient–doctor relationship as an explanation for different response rates, are needed. The placebo response is poorly understood and may have many etiologies, including recent diagnostic tests, increased medical attention and possibly certain characteristics of the patient–doctor relationship.
- Studies of patient characteristics that predict response to treatment should be done. Studies are required which help explain why some patients do well long term and others poorly.
- Studies on potential differences in disease severity and treatment response in primary care versus secondary or tertiary care need to be done. There are few studies which have addressed whether there are differences in treatment response related to the clinical setting.

Conclusion

Until recently many treatment trials of FGIDs suffered from important weaknesses in study design and/or analysis. Fortunately, this problem has been increasingly recognized, and the more recent studies have paid greater attention to these issues. Methodologically sound clinical trials are possible in the FGIDs.

A major methodological problem that has not yet been resolved is the lack of validated outcome measures. Despite progress in this area there is at the moment no one generally accepted outcome measure, as all current instruments still require further validation.

It is important that the newly developed disease-specific outcome tools are tested in various populations using appropriate study designs to determine which perform the best. It will be important that investigators, representatives of regulatory and funding agencies, and patients reach consensus on the best outcome measures for future studies. For this a regular exchange of data and consensus meetings about design issues among all stakeholders are encouraged. It will also be helpful if agreement is reached on the scales that are used to record outcome measures, as it will allow for comparisons among studies and possible statistical pooling of the results in the future.

Treatment trials of the FGIDs pose special methodologic challenges but it is our committee's belief that they can be overcome. The recommendations in this report hopefully will help this process. Clearly there is still a need for more empirical data that evaluate study design in this field.

Dr Veldhuyzen van Zanten is the recipient of a Nova Scotia Clinical Scholar Award.

The committee wishes to thank Drs D Drossman, J Everhard, D Patrick, R C Spiller, G K Turnbull, and W E Whitehead for their helpful suggestions for the manuscript.

1 Klein KB. Controlled treatment trials in the irritable bowel syndrome. a critique. *Gastroentrology* 1988;**95**:232–41.
2 Veldhuyzen van Zanten SJO, Cleary C, Talley NJ, *et al.* Drug treatment of functional dyspepsia. A systematic analysis of trial methodology with recommendations for design of future trials. *Am J Gastroenterol* 1996;**91**:660–71.
3 Talley NJ, Nyren O, Drossman DA, *et al.* The irritable bowel syndrome.Toward optimal design of controlled treatment trials. *Gastroenterol Int* 1993;**6**:189–211.
4 Drossman DA, Richter JE, Talley NJ, *et al* (eds). The *functional gastrointestinal disorders: diagnosis, pathophysiology, and treatment*. McLean, VA: Degnon Associates, 1994.
5 Drossman DA, Thompson WG, Talley NJ, *et al.* Identification of subgroups of functional gastrointestinal disorders. *Gastroenterol Int* 1990;**3**:159–72.
6 Talley NJ, Weaver AL, Tesmer DL, *et al.* Lack of discriminant value of dyspepsia subgroups in patients referred for upper endoscopy. *Gastroenterology* 1993;**105**: 1378–86.
7 Klauser A, Voderholzer WA, Knesewitsch PA, *et al.* What is behind dyspepsia? *Dig Dis Sci* 1993;**38**:147–54.
8 Talley NJ, Meineche-Schmidt V, Pare P, *et al.* Omeprazole is efficacious in non-ulcer dyspepsia. *Can J Gastroenterol* 1998;**12**(suppl):70–71A.
9 Lavori PW, Louis TA, Bailar JC III, *et al.* Designs for experiments. Parallel comparisons of treatment. *N Engl J Med* 1983;**309**:1291–8.
10 Woods JR, Williams JG, Tavel M. The two-period crossover design in medical research. *Ann Intern Med* 1989;**110**:560–6.
11 O'Donoghue DP, Dawson AM, Powell-Tuck Brown RL, *et al.* Double blind withdrawal trial of azathioprine as maintenance treatment for Crohn's disease. *Lancet* 1978;ii:955–7.
12 Hawthorne AB, Logan RFA, Hawkey CJ, *et al.* Randomized controlled trial of azathioprine withdrawal in ulcerative colitis. *BMJ* 1992;**305**:20–2.
13 Altman DG. Randomisation. *BMJ* 1991;**302**:1481–2.
14 Rösch W. Cisapride in non-ulcer dyspepsia. Results of a placebo-controlled trial. *Scand J Gastroenterol* 1987;**22**: 161–4.
15 Talley NJ, Silverstein MC, Agreus L, *et al.* AGA Technical review. Evaluation of dyspepsia. *Gastroenterology* 1988; **114**(suppl A):582–95.
16 Melzack R. The McGill Pain Questionnaire. Major properties and scoring methods. *Pain* 1975;**1**:277–99.
17 Guyatt GH, Veldhuyzen van Zanten SJO, Feeney DH, *et al.* Measuring quality of life in clinical trials. A taxonomy and review. *Can Med J Assoc* 1989;**140**:1441–8.
18 Guyatt GH, Feeney DH, Patrick DL. Measuring health-related quality of life. *Ann Intern Med* 1993:**118**:622–9.
19 MacKenzie CR, Charlson ME. Standards for the use of ordinal scales in clinical trials. *BMJ* 1986;**292**:40–3.
20 Jaeschke R, Singer J, Guyatt GH. A comparison of seven-point and visual analogue scales. *Control Clin Trials* 1990;**11**:43–51.
21 Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* 1989;**10**:407–15.
22 Dimenas E. Glise H, Ballerback B, *et al.* Well-being and gastrointestinal symptoms among patients referred to endoscopy due to suspected duodenal ulcer. *Scand J Gastroenterol* 1995;**30**:1046–52.
23 El-Omar EM, Banerjee S, Wirz A, *et al.* The Glasgow Dyspepsia Severity Score. A tool for the global measurement of dyspepsia. *Eur J Gastroenterol Hepatol* 1996;**8**:967–71.
24 Veldhuyzen van Zanten SJO, Tygat KMAJ, Pollak PT, *et al.* Can severity of symptoms be used as outcome measures in trials of non-ulcer dyspepsia and Helicobacter pylori. *J Clin Epidemiol* 1993;**46**:273–9.
25 Houghton LA, Heyman DJ, Whorwell PJ. Symptomatology, quality of life and economic features of irritable bowel syndrome. The effect of hypnotherapy. *Aliment Pharmacol Ther* 1996;**10**:91–5.
26 Hahn BA, Kirchdoerfer LJ, Fullerton S, *et al.* Evaluation of a new quality of life questionnaire for patients with irritable bowel syndrome. *Aliment Pharmacol Ther* 1997;**11**:547–52.
27 Patrick DL, Drossman DA, Frederick IO, *et al.* Quality of life questionnaire for patients with irritable bowel syndrome. Development and validation of a new measure. *Dig Dis Sci* 1998:**43**:400–11.
28 Altman DG. Better reporting of randomized controlled trials. The CONSORT statement. *JAMA* 1996;**276**:637–9.
29 Simon R. Confidence intervals for reporting results of clinical trials. *Ann Intern Med* 1986;**105**:429–35.
30 Smith DG. Clemens J, Crede W, *et al.* Impact of multiple comparisons in randomized clinical trials. *Am J Med* 1987; **83**:545–50.
31 Perniger TV. What is wrong with Bonferroni adjustments. *BMJ* 1998;**316**:1236–8.
32 Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. *Ann Intern Med* 1992;**116**:78–84.
33 Freiman JA, Chalmers TC, Smith H, *et al.* The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. *N Engl J Med* 1978;**299**:690–4.
34 Bero L, Rennie C. The Cochrane Collaboration. Preparing, maintaining and disseminating systematic reviews of the effects of health care. *JAMA* 1995;**274**:1935–8.