Alimentary Pharmacology & Therapeutics

WILEY

AP&T Alimentary Pharmacology & Therapeutics

REVIEW ARTICLE

Rome Foundation Working Team Report: Consensus Statement on the Design and Conduct of Behavioural Clinical Trials for Disorders of Gut–Brain Interaction

Helen Burton-Murray[1,2] | Livia Guadagnoli[3,4] | Kendra Kamp[5] | Inês A. Trindade[6] | Lynda H. Powell[7] | Magnus Simrén[8,9] | Brjánn Ljótsson[10] | Laurie Keefer[11]

[1]Division of Gastroenterology, Massachusetts General Hospital, Boston, Massachusetts, USA | [2]Harvard Medical School, Boston, Massachusetts, USA | [3]Division of Gastroenterology and Hepatology, Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA | [4]Laboratory for Brain-Gut Axis Studies (LaBGAS), Translational Research in Gastrointestinal Disorders (TARGID), Department of Chronic Diseases and Metabolism (CHROMETA), KU Leuven, Leuven, Belgium | [5]Department of Biobehavioural Nursing and Health Informatics, School of Nursing, University of Washington, Seattle, Washington, USA | [6]EMBRACE Lab, School of Behavioural, Social and Legal Sciences, University of Örebro, Örebro, Sweden | [7]Department of Family and Preventive Medicine, Rush University Medical College, Chicago, Illinois, USA | [8]Department of Molecular and Clinical Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden | [9]Center for Functional GI and Motility Disorders, University of North Carolina-Chapel Hill, Chapel Hill, North Carolina, USA | [10]Department of Clinical Neuroscience, Division of Psychology, Karolinska Institutet, Stockholm, Sweden | [11]Division of Gastroenterology, Icahn School of Medicine at Mount Sinai, New York, New York, USA

Correspondence: Helen Burton-Murray (hbmurray@mgh.harvard.edu)

Received: 3 September 2024 | **Revised:** 17 September 2024 | **Accepted:** 23 December 2024

Handling Editor: Alexander Ford

Funding: This work was supported by Fonds Wetenschappelijk Onderzoek (12A7822N) and National Institute of Diabetes and Digestive and Kidney Diseases (K23 DK131334 and K23 NR020044).

Keywords: brain–gut behaviour therapy | clinical trials | disorders of gut–brain interaction | efficacy | minimum clinically important difference | ORBIT model

ABSTRACT

Background: Brain–gut behaviour therapies (BGBT) have gained widespread acceptance as therapeutic modalities for the management of disorders of gut–brain interaction (DGBI). However, existing treatment evaluation methods in the medical field fail to capture the specific elements of scientific rigour unique to behavioural trial evaluation.

Aims: To offer the first consensus on the development and testing of BGBT in DGBI.

Methods: An international, interdisciplinary team of experts developed a consensus statement heavily informed by best practice recommendations for behavioural clinical trials for chronic diseases, organised by a selected treatment development model.

Results: We suggest an existing behavioural treatment development model that has an iterative progression aligned with the drug development model with nuances specific to BGBT. We describe the iterative phases through initial discovery and experimental work, assembly of a mechanistic pathway and candidate treatment components, treatment refinement and optimisation, initial proof-of-concept, feasibility of clinical trials and, finally, confirmatory efficacy and effectiveness testing. We delineate

Helen Burton-Murray, Livia Guadagnoli, Kendra Kamp, and Inês A. Trindade are co-first authors.

Brjánn Ljótsson and Laurie Keefer are co-senior authors.

recommendations for and provide examples that lend themselves to gastroenterologists planning to develop or test BGBT, reviewing proposals for or results from BGBT studies or writing management guidelines for DGBI.

Conclusions: This working team report facilitates a shared understanding of the elements of scientific rigour necessary for BGBT development and could support future standards on which BGBT are evaluated in gastroenterology.

There is widespread acceptance of brain-gut behavior therapies (BGBT) in the clinical care of disorders of gut-brain interaction (DGBI) [1] with multiple positive efficacy trials and meta-analyses (for irritable bowel syndrome) [2–5]. However, support by gastroenterology societies globally relegate BGBT to have 'low-quality evidence' [6–13]. Most guidelines provide 'weak' recommendations [6, 8, 9, 14] for BGBT in irritable bowel syndrome management, with rare 'strong' recommendations [10]. For DGBI beyond irritable bowel syndrome, confirmatory efficacy data is lacking, thus some gastroenterology societies explicitly do *not* endorse BGBT [8, 15] and others provide 'weak'/'very weak' recommendations for BGBT [11–13]. We propose that the existing approaches to evaluating quality and rigour in pharmacological trials for DGBI, including the Grading of Recommendations Assessment, Development and Evaluation (GRADE) methodology [16, 17] (e.g., double-masking, placebo-control), shortchange the proper evaluation of BGBT, failing to capture the specific elements of scientific rigour that uniquely define the conduct of behavioural trials as has been well-defined within the psychological sciences. Further, many BGBT have jumped to demonstrating either efficacy or effectiveness prior to understanding for whom, when and why an intervention is best suited—given extremely limited evidence on moderators and mediators of BGBT [18], our field would benefit from further tinkering with existing BGBT.

The aim of this Working Team Report is therefore to generate a consensus statement on the best practices for the future development and testing of BGBT and associated techniques. Our intent is for this consensus statement to inform both investigators in the planning and execution of BGBT trials, as well as future standards on which BGBT are evaluated in DGBI. In doing so, we aim to ensure that there is a shared understanding of the elements of scientific rigour necessary to: (a) support quality rating metrics for BGBT and (b) provide stronger rationale supporting the need for pre-clinical and early phase intervention studies as part of a more efficient approach to the development, evaluation and implementation of personalised therapies for patients living with DGBI. We also chose to limit discussion in a few key areas given other considerations for their design and testing that would warrant more in-depth coverage and are out of the scope of the current report. For example, we did not cover digital behavioural interventions, dissemination and implementation research and specific considerations for applications to paediatric or other digestive diseases. We recognise that many of our recommendations may seem overwhelming to the average gastroenterologist. A goal of our consensus statement is to increase knowledge of best practices that applies to gastroenterologists who are planning to develop or test BGBT, reviewing proposals for or results from BGBT studies or writing management guidelines for DGBI.

1 | Methods

After a proposal initiated by HBM and LK was approved by the Rome Foundation Board of Directors in December 2022, we assembled a multidisciplinary committee of international experts (gastropsychologists, nurse scientist and gastroenterologist). Committee expertise covered the various aspects of behavioural intervention design and testing, including experimental/mechanistic discovery (LG, IT and BL); intervention design and adaptation, early testing and refinement (KK, IT and HBM); consideration and evaluation of mediators and moderators of behaviour change (HBM and BL); and the conduct of large efficacy and effectiveness studies (BL, LK and MS). In a series of team video calls and email exchanges between February 2023 and May 2024, a modified Delphi method was used to create consensus on the goals, structure and framework of the report.

There are many models that exist to guide aspects of behavioural intervention development and testing, such as the National Institutes of Health (NIH) [19] and the Medical Research Council framework [20, 21] (see Czajkowski [22]). The committee unanimously voted to frame our report around the Obesity-Related Behavioural Intervention Trials (ORBIT) model [23, 24] for a few reasons: (1) it aligns with the United States Food and Drug Administration (FDA) clinical trial vernacular; (2) the ORBIT model is iterative and flexible, which suits a multidisciplinary field of professionals (i.e., flexible movement forwards and backwards between basic science, early development/refinement and efficacy/effectiveness testing); (3) the use of a single framework will allow for the future development of a 'quality checklist' for the development and testing of gastrointestinal (GI) behavioural interventions at various stages [23].

We used best practice guidelines for developing and testing non-pharmacological interventions for medical conditions, specifically heavily influenced by *Behavioural Clinical Trials for Chronic Diseases* [24], which may be viewed as the corollary to guidelines followed for pharmacotherapy trials (e.g., Friedman et al. [25]). We included a leading expert in behavioural clinical trial guidelines (LP) who is the Co-Director of the ORBIT Institute and faculty member and former director of multiple National Institute of Health-sponsored trainings through the Office of Behavioural and Social Sciences Research. In a series of virtual meetings, Co-Chairs HBM and LK met with LP to review figures, identify gaps and confirm that committee consensus was keeping within the extant behavioural intervention science best practices.

Once the team had agreed on key considerations for each treatment phase, we broke into groups of 2–3 working team members, based on our specific areas of experience or expertise, to write each ORBIT phase-based section. Each section was then

reviewed by a second small group, and differences in opinion across phases were resolved through virtual meetings. HBM and LK met regularly to ensure cohesion and consistency in recommendations throughout the document. The final report was reviewed and approved by the Rome Foundation Board of Directors before submission.

RECOMMENDATION #1: Investigators should follow a treatment development framework, such as the ORBIT model (see Figure 1). Reviewers should expect to see justification for a phase of testing within a model/framework in grant proposals and publications.

Below, we summarise the foci of each phase within the ORBIT model framework (Figure 1). We also provide a summary of our recommendations in Table 1.

1.1 | Discovery/Experimental Phases

RECOMMENDATION #2: A clearly defined clinically significant question should be generated through clinical observation, existing literature and experimental research. Experimental research is critical to form the scientific premise for the behavioural intervention components that could be used to answer the clinically significant question.

Discovery/experimental efforts are essential, but often overlooked. While the ORBIT model is not linear per se, discovery/experimental efforts typically occur at the earliest stage of intervention development, prior to behavioural intervention design and testing. However, discovery/experimental work, such as basic behavioural research, may continue to interact bidirectionally with other phases of the ORBIT model throughout behavioural intervention development (Figure 1) [23]. Below, we describe the key elements for discovery/experimental work.

1.1.1 | Identify a Clinically Significant Question

The first step to behavioural intervention development is the identification of a clinically significant question that could potentially be addressed using a BGBT. This can be informed by both clinical and basic science perspectives through literature review and clinical observation [24] and includes:

- Disease-related considerations (e.g., the brain–gut axis contributes to DGBI symptom maintenance, DGBI symptoms are associated with high healthcare costs/utilisation).
- Drawbacks in current treatment (e.g., access to behavioural intervention is limited, uncertain efficacy or effectiveness of an existing intervention for a subgroup, intervention adherence is poor).
- Promise of a new treatment (e.g., new treatment targets a novel hypothesised pathway, intervention shows promise in preliminary work, new intervention is less costly than the standard existing treatment).

A clinically significant question may be identified through clinical observation, bolstered by the existing literature, and experimental research. For further reading, table 3.4 in Behavioural Clinical Trials for Chronic Diseases [24] provides a comprehensive overview. While identification of a clinically significant question is the first step, it is the common thread that runs through the entire BGBT design and testing process. That is, the clinically significant question is relevant for all phases—for the immediate next steps of intervention, such as determining clinically meaningful milestones in Phase Ia,

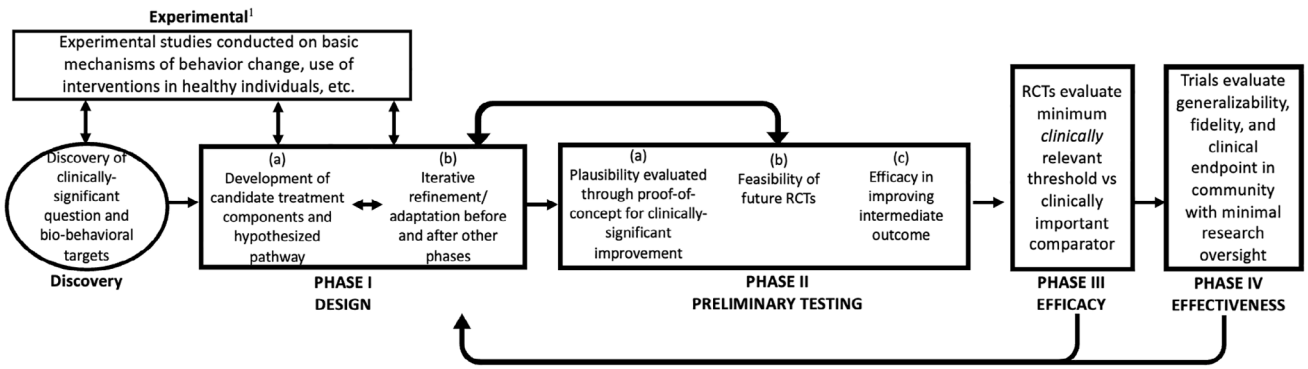

FIGURE 1 | The future of brain–gut behaviour therapy development and testing within the ORBIT Model. Modified from the ORBIT model [24]. Phases do not always occur sequentially. For example, refinement trials may be conducted prior to Phase II testing or may be conducted after a phase IIc efficacy trial to revise the treatment to be more efficient or parsimonious. See Figure S2 for an example progression from Phase Ia through Phase IV. BGBT, brain–gut behaviour therapy; ORBIT, Obesity-Related Behavioural Intervention Trials; RCT, randomised controlled trial. (1) The label 'Experimental' is not used by the ORBIT model, but we included it for clarity. (2) Phase IIc trials are the corollary to Phase II drug trials and aim to identify if the BGBT can achieve a clinically significant improvement in an intermediate (e.g., behavioural risk or maintenance factor that contributes to the primary outcome) or surrogate outcome, compared to the passage of time, other non-specific factors, or other available treatments. Phase III trials which aim to identify if the BGBT is better at achieving a clinically significant improvement in the primary clinical outcome than some established form of care. In some cases, a Phase IIc trial may be the end goal for efficacy testing, when there is a well-established threshold for improvement in the intermediate or surrogate outcome that is associated with a clinically significant improvement in the primary outcome. If a Phase III trial is needed, a Phase IIc trial is a milestone before moving to Phase III testing.

but also in future phases, such as evaluating if an intervention is efficacious in clinical practice (i.e., Phase III efficacy) [26, 27].

Example in Action: A clinician notices that a subgroup of patients with DGBI report notable fear of GI symptoms and looks at the literature to find a series of studies that support elevated levels of fear of symptoms (clinically significant problem). With no empirically-supported treatments that target fear of GI symptoms, the clinician then questions if using a BGBT intervention to reduce fear of symptoms could lead to improved clinical outcomes (clinically significant question). This question may then become the basis for a behavioural intervention and inform the Phase Ia design of the intervention.

1.1.2 | Identify a Basic Scientific Premise for Intervention

The other feature of the Discovery/Experimental Phase includes the discovery of a basic scientific premise for an intervention. Basic behavioural research can include a variety of research types, including experimental studies, self-report measures (e.g., surveys and questionnaires) and observational methods (e.g., ecological momentary assessment) [19, 28]. Basic behavioural research can help to identify the specific drivers of the behavioural outcome that we are interested in changing [22, 23]. Similar to how drug trials use pre-clinical basic science studies in animal models to understand a disease process, basic behavioural research provides insight into if and how behavioural mechanisms are linked to clinical outcomes and informs future treatment development,

TABLE 1 | Consensus statement on the design and conduct of behavioural clinical trials for disorders of gut–brain interaction.

Description	This consensus statement is intended to provide guidance to those in the gastroenterology field who are planning to develop or test BGBT, reviewing proposals for or results from BGBT studies, or are writing management guidelines for DGBI. We focus on key recommendations applicable to BGBT. Readers should refer to other sources [24, 25] for detailed guidance on other necessary trial design elements
Methods	A multidisciplinary committee of international experts used a modified Delphi method, framing the consensus statement around the ORBIT model [23, 24], an iterative and flexible framework for the development and testing of behavioural treatments
Recommendations	
#1	Investigators should follow a treatment development framework, such as the ORBIT model (see Figure 1). Reviewers should expect to see justification for a phase of testing within a model/framework in grant proposals and publications.
#2	A clearly defined clinically significant question should be generated through clinical observation, existing literature and experimental research. Experimental research is critical to form the scientific premise for the behavioural intervention components that could be used to answer the clinically significant question.
#3	Candidate intervention characteristics, target population and outcomes should be iteratively refined/optimised. Refinement *before* efficacy testing can increase the likelihood of success in later confirmatory trials as well as facilitate future intervention affordability and scalability. During this process, decisions should be documented and stakeholders in some form should be included.
#4	Plausibility of clinically significant change should be demonstrated *before* moving to efficacy testing and typically does not require large samples. The term 'proof-of-concept' is appropriate, while 'preliminary efficacy' and 'pilot trial' should not be used. Clearly defined and pre-specified benchmarks or minimum clinically important difference thresholds in the outcome should be used to test plausibility.
#5	Once plausibility is demonstrated, threats to the feasibility of a future efficacy trial (not feasibility of the intervention) should be identified and addressed (and tested, as applicable). If a pilot feasibility trial is conducted, it should mirror the planned future efficacy trial and should test pre-specified feasibility benchmarks.
#6	Once plausibility is demonstrated and feasibility is addressed, efficacy trials prioritising internal validity are appropriate. Efficacy trials should be powered based on an ability to detect *both* a clinically and a statistically significant benefit, using an MCID in response between a treatment and control. Trials testing effects on intermediate or surrogate outcomes (e.g., ORBIT Phase IIc) are sufficient in some cases, otherwise serve as milestone before a confirmatory trial testing effects on the primary clinical outcome (e.g., ORBIT Phase III).
#7	Comparators in BGBT trials are not one-size-fits-all and should have a documented rationale based on the *purpose of the trial* and the *purpose of the comparator.*
#8	Given that it is not possible for BGBT trials to be double-masked, efforts to extend masking must be conducted, including masking staff and participants to hypotheses, utilising masked outcome assessors and investigators holding a scientific mindset of equipoise about the treatment arms' effects.
#9	Once efficacy of the intervention is confirmed, effectiveness/pragmatic studies prioritising external validity (real-world practice) are appropriate. If the intervention needs to be adapted for the population/environment, efficacy of that adapted intervention is needed *before* testing effectiveness.

Abbreviations: BGBT, brain–gut behaviour therapy; DGBI, disorders of brain–gut interaction; Equipoise, the belief that whether one trial arm is better than another is unknown; ORBIT, Obesity-Related Behavioural Intervention Trials.

particularly in Phase Ia design and refinement. Indeed, the drivers of behaviour initially identified by basic science can be translated into mechanisms of action (targets) and/or a model for how to change behaviour (components) in an intervention [23]. Thus, the next step would be to translate such findings into (1) candidate mechanisms of action that could then (in Phase Ia) be modified via behavioural intervention and (2) candidate components of that behavioural intervention that offer promise for modifying these mechanisms. Figure 2 provides an overview of this process, including the common methods used in DGBI research, the candidate biological and behavioural target mechanisms they are measuring and how these targets relate to a behavioural intervention.

Notably, behavioural research interacts dynamically across several aspects of the ORBIT model (Figure 1) [23, 24]. Within the Discovery/Experimental Phase, basic research findings can provide a scientific basis for a clinically significant question, while clinical experience may shed light on a problem that can be further understood through basic research [23]. A similar reciprocal relationship exists between basic research and Phase Ia. Indeed, findings from basic research can be used to make decisions about various aspects of the intervention, such as the treatment components to include and the dose of delivery [23, 24]. Meanwhile, outcomes from intervention trials inform basic research, such as providing information on how mechanistic targets perform in clinically meaningful settings (i.e., outside of the research lab).

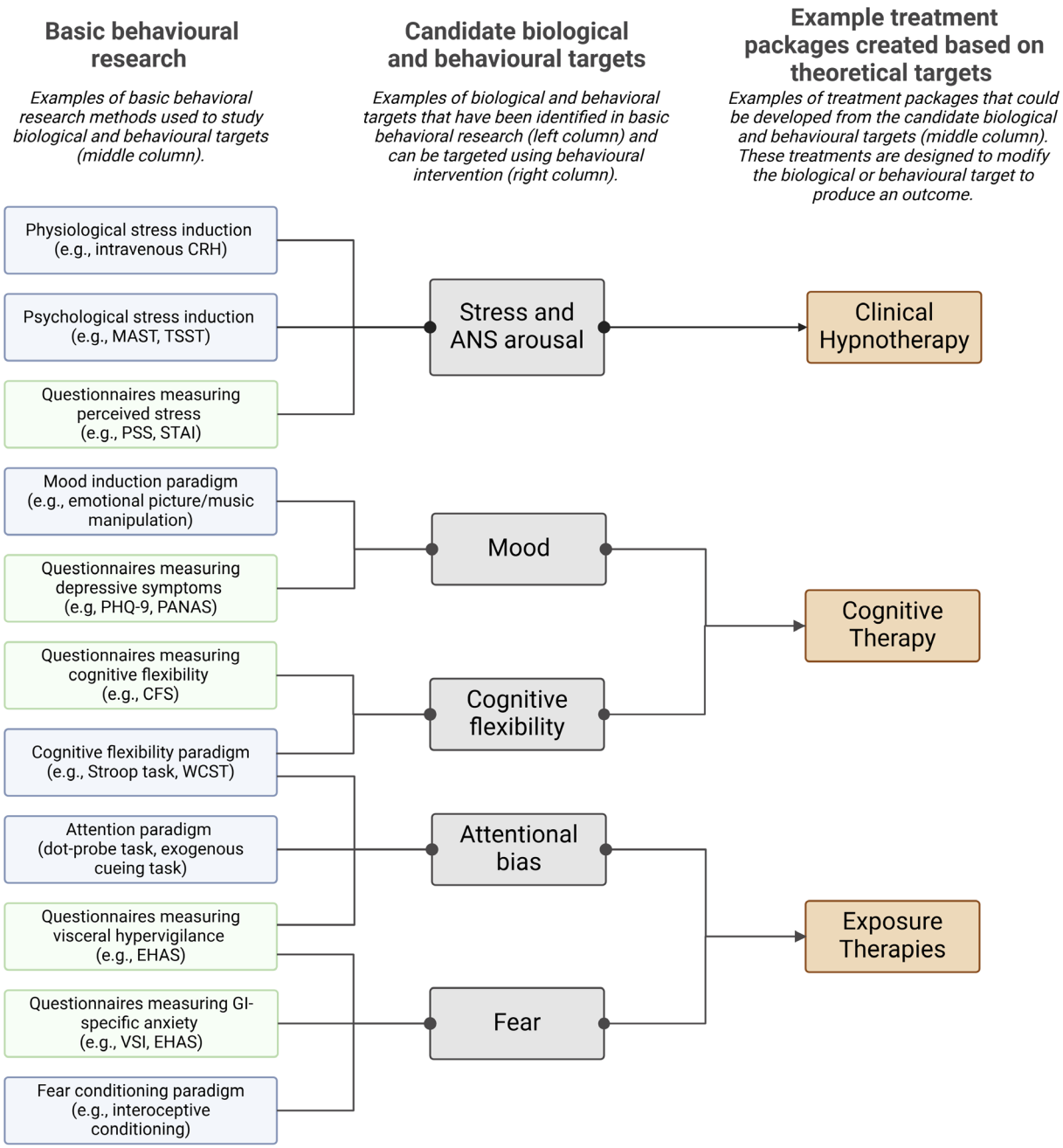

FIGURE 2 | Legend on next page.

FIGURE 2 | Overview of the relationship between basic behavioural research, the candidate biological and behavioural targets they are measuring and how these targets relate to behavioural intervention. This figure provides examples of basic behavioural research methods, candidate biological and behavioural targets and treatment packages that can be developed from these targets. Starting from the left, the figure demonstrates common methods used in disorders of gut–brain interaction (DGBI) basic behavioural research, including both experimental paradigms (blue rectangles) and survey/questionnaire data (green rectangles). Basic behavioural research provides evidence for specific candidate biological and behavioural targets (grey rectangles), including constructs such as stress, mood, cognitive flexibility, attentional bias and fear. For example, questionnaires assessing fear as well as experimental research using fear-conditioning paradigms have both demonstrated that fear is an important mechanism in DGBI. The biological and behavioural targets can then be translated into a behavioural intervention (orange rectangles), where it is modified to produce an outcome. Different behavioural interventions target different biological and behavioural mechanisms. For example, one way fear can be targeted is through exposure therapy, whereas cognitive flexibility may be targeted using a cognitive-based intervention; this could then serve as a rationale to create a combined treatment. However, it is also possible that exposure therapies can engage the target of cognitive flexibility, which would be advantageous for treatment parsimony. Of note, for simplicity, the figure provides one example of a treatment package per theoretical target. However, in reality, several different classes of behavioural intervention can be used to address one theoretical target. For example, both cognitive therapy and exposure therapy can be used to target mood. Further, some of treatment modalities can be broken down into several components (e.g., there are different types of exposure therapy—see Figure S1). CFS, Cognitive Flexibility Scale; CRH, corticotropin releasing hormone; EHAS, Oesophageal Hypervigilance and Anxiety Scale; MAST, Maastricht Acute Stress Task; PANAS, Positive and Negative Affect Schedule; PHQ-9, Patient Health Questionnaire-9 item; PSS, Perceived Stress Scale; STAI, State Trait Anxiety Inventory; TSST, Trier social stress test; VSI, Visceral Sensitivity Index; WCST, Wisconsin Card Sorting Task. Created in https://BioRender.com.

Identifying a clinically significant clinical question and a basic scientific premise for intervention (addressing the clinically significant question) are crucial to the early stages of intervention development, but can also be referred back to and applied throughout the design and refinement process. Currently, basic behavioural research in psychogastroenterology exists (Figure 2), which can and should inform clinically significant research questions and intervention design.

> **Example in Action**: The clinician who develops the clinically significant question looks to the basic science literature and finds that experimental fear-conditioning paradigms in *healthy* humans demonstrate that learned fear of GI sensations is associated with reduced visceral pain thresholds (i.e., increased visceral pain perception).[26] They then conduct studies (or look to other literature) that show:
>
> a. Individuals with DGBI have heightened visceral perception compared to controls. Based on these findings, they hypothesise that fear is a mechanism in visceral perception, serving as the foundation for an intervention for which reducing fear (via behavioural intervention) could potentially result in a clinical improvement for DGBI.
> b. Higher fear of GI symptoms (e.g., via specific fear items on the Visceral Sensitivity Index,[27] a measure of GI-specific anxiety) is associated with clinical outcomes of worse symptom severity and quality of life.

1.2 | Phase I: Design

1.2.1 | Overview of Phase I

The overall goal of Phase I is to assemble (Phase Ia) and refine (Phase Ib) the BGBT, including tinkering with the BGBT in an attempt to maximise strength and efficiency. In contrast to Phase I drug development model studies that may be conducted with healthy volunteers or patients, Phase I behavioural work is conducted with patients. However, a key similarity between behavioural and drug development in Phase I is the tinkering with the intervention to refine and optimise it [24]. Below, we describe key recommendations for Phase I.

RECOMMENDATION #3: Candidate intervention characteristics, target population and outcomes should be iteratively refined/optimised. Refinement *before* efficacy testing can increase the likelihood of success in later confirmatory trials as well as facilitate future intervention affordability and scalability. During this process, decisions should be documented and stakeholders in some form should be included.

1.2.2 | Selecting the Target Population (Ia)

Having an identified clinically significant question is a requirement for initiating a Phase Ia study, in which a population with DGBI symptoms that could be modified by behavioural intervention is selected, supported by a basic scientific premise. Ceiling effects can occur if the target population treated with the intervention does not have or has low levels of the targeted behavioural mechanism. To reduce the risk of the treatment not showing its intended effects, it can be helpful to identify inclusion criteria that support the selection of a population elevated on target mechanisms. The target population may be refined over time through the development process.

Notably, there is increasing recognition of the importance of involving stakeholders from the target population early in the development process and throughout later testing (e.g., individuals with lived experience)—there are many ways this can be implemented, from stakeholder focus groups or advisory committees, or including patients as investigators [29].

> **Example in Action**: Using the basic behavioural research findings that higher fear of GI symptoms is associated with worse clinical outcomes, a cut-off is used to define the target population. If this cut-off was not used, there is a high risk of seeing a later ceiling effect if a large majority of those treated cannot improve much on the behavioural target.

1.2.3 | Identify a Hypothesised Pathway (Ia)

After identifying the target population, a hypothesised pathway which comprehensively explains how the intervention may answer the clinically significant questions should be identified. The hypothesised pathway should be supported by the identified basic scientific premise, using established theoretical models and empirical evidence. While stated a priori (e.g., in a registered protocol) for any study that tests the pathway (or portions of it), it is not static—the pathway will be refined, and hopefully simplified, as treatment development continues.

> **Example in Action**: Based on the basic scientific premise that reducing fear of GI symptoms could lead to a clinically significant improvement in the DGBI, the investigator drafts an initial hypothesised pathway alongside the assembly of the candidate intervention with exposure therapies. **Supplemental Figure 1, Panel A** offers a simplified example. Some treatment components may target multiple behavioural targets to produce clinically significant change; for example, fear of GI symptoms is only one facet of GI-specific anxiety—other facets may also be targeted by exposure therapies, such as reducing avoidance behaviours around symptoms.

1.2.4 | Select Appropriate Outcome Measures (Ia)

Selection of outcome measures that align with the hypothesised pathway occurs in Phase Ia and should be clinically meaningful. Patient stakeholders from the target population may aid in determining outcomes. One challenge for BGBT in selecting outcome measures is that DGBI have no known biomarkers, thus subjective outcomes are often used. Subjective outcome measures may often include self-report surveys, daily diaries or clinician-ratings—all of these measurements have the same risk of bias as the participant who is reporting on the outcome is not masked to treatment allocation. Thus, subjective outcomes raise risk of bias from inflated estimates of improvement in a trial that inherently cannot be double-masked; thus, BGBT trials necessitate careful implementation of strategies to reduce risk of bias. Priority should be placed upon objective outcomes that could apply to a BGBT depending on the clinically significant question (e.g., weight change, rate of emergency department visits).

Below, we delineate categories for types of outcomes, to facilitate a shared nomenclature for outcome variables relevant for BGBT development and testing [24]. The type of outcome investigated in a study depends on the purpose of the study (and corresponding study phase). The 2022 Consolidated Standards of Reporting Trials (CONSORT) guidelines also provide a comprehensive review of different types of outcomes [30].

1.2.4.1 | Primary Outcome. A definitive, clinically relevant outcome used to test efficacy (i.e., Phase III, some Phase Ib) or effectiveness (i.e., in Phase IV) of the intervention.

A model by which the intervention is hypothesised to affect this outcome should have been described by previous research (i.e., the hypothesised pathway as above). The measure used to assess a primary outcome should have a distinct definition of a responder (e.g., minimum clinically important difference; MCID). A *single* primary outcome is used in definitive clinical trials (Phase III and Phase IV). Notably, primary outcomes are typically objective event rates with a long-term follow-up (e.g., incidence of coronary artery disease, death rate and healthcare costs), as these are the outcomes that drive changes in clinical care and are of interest to third-party payers in the United States. Other outcomes below typically are tested in earlier phase studies, including Phase IIc efficacy studies, where the duration of follow-up tends to be shorter.

1.2.4.2 | Surrogate Outcome. Associated with the primary outcome—this outcome can be used as a substitute, when assessment of the primary outcome (e.g., mortality) is not possible.

1.2.4.3 | Secondary Outcomes. Any other hypothesised outcome of the intervention.

An *intermediate outcome* is a type of secondary outcome specifically hypothesised to mediate the change on the primary outcome, affected by the intervention. For Phase III trials, one intermediate outcome is recommended, reserving exploration of mediators and moderators for earlier phase work. In fact, showing that the intervention can plausibly produce clinically significant change in target mediators (i.e., intermediate outcome) earlier on justifies progression to a Phase IIc efficacy trial. By identifying candidate moderators early on, the intervention can be refined so it optimally can work for identified subgroups—either by targeting the intervention to a specific subgroup or modifying it to have equal effects across subgroups (e.g., to work equally well for men and women). Other secondary outcomes could be outcomes that the intervention targets to mediate change in the intermediate outcome (the intermediate outcome then mediates change in the primary outcome).

1.2.4.4 | Exploratory Outcome. Used to explore new associations with the intervention or other outcomes in Phase I and Phase IIa studies. In confirmatory efficacy (Phase IIc and Phase III) and effectiveness (Phase IV) studies, exploratory outcomes should be kept to a minimum to keep the attention on the intervention and avoid responder burden.

1.2.5 | Assemble the Candidate Intervention (Ia)

An intervention may be based on a validated model/framework or adapted from an existing intervention to the needs of a specific population. In the case of the latter, the adaptation takes place in Phase Ib refinement studies. A patient-centred approach should be considered during the assembly of the intervention. Focus group interviews, patient advisory groups or other similar methods can be used to put together an intervention. Patients may provide important input about aspects such as unmet needs, preferred intervention structure or type of delivery. Throughout the process of assembling the intervention, *documentation of decisions* is important to keep track of the intervention development/adaptation process. Assembly of the candidate intervention might be returned to when refinement (see Phase Ib) is necessary—for example, an intervention with established efficacy may

Example in Action: Based on the example hypothesised pathway in **Supplemental Figure 1**, the below outcome measures are identified.

- **Primary outcome:** Self-report scale of DGBI symptom severity with an established MCID target for clinically significant improvement
- **Intermediate outcome:** One self-report scale of fear of symptoms with a known relationship with the primary outcome (and ideally an MCID target).
- **Exploratory outcome(s):** In some Phase I and II studies, the investigator chooses the following for other candidate mechanisms of action (i.e., intermediate outcomes): self-report scales of avoidance around symptoms and fear of eating, and a visceral pain perception task. They also choose a self-report scale of symptom-related quality of life to explore change in another clinical outcome.

Note these are examples; symptom severity could be an intermediate outcome in other hypothesised pathways with the primary outcome being a threshold change in healthcare utilisation. We also do not have specific recommendations on which types of outcomes (and MCIDs) should be used, as this is out of this paper's scope. However, current FDA guidance recommends use of a threshold change in symptom severity as the clinically significant target for irritable bowel syndrome.[39] A recent review of DGBI symptom severity-specific patient-reported outcome measures also summarises available MCIDs.[58]

need to be modified (and subsequently tested for efficacy) before testing effectiveness in a real-world setting (see Phase IV). Some considerations during the candidate intervention assembly are:

- Treatment integrity—fidelity to the theoretical model and procedures specified in a treatment protocol [31].
- Timing of delivery (e.g., how long after diagnosis and timing of medication initiation).
- Acceptability by recipients and stakeholders, which can be affected by duration of sessions, structure of the intervention and skill of the therapist, among other factors [32]. Focus group interviews may be conducted in Phase Ia with the target population and stakeholders (e.g., clinicians) to inform acceptability when writing the manual [33].
- Short- and long-term feasibility, including how feasible the delivery method will be in various populations (e.g., to reach populations with limited geographic access to healthcare, group vs. individual delivery).

Example in Action: Three different types of exposure therapy are identified (treatment components) delivered in 1:1 treatment sessions (dose) that may reduce symptom severity (primary outcome) via decreased fear of symptoms (intermediate outcome) (**Supplemental Figure 1, Panel A**).

1.2.6 | Iterative Adjustment of Hypotheses and Outcomes (Ia/Ib)

In our opinion, more time spent in and documentation of (e.g., in publications) Phase I work for BGBT is needed. Phase I work is essential, but like Discovery/Experimental work, it is often overlooked (and underfunded). Phase I requires embracing the high chance of failure, but at a minimal cost—rather than compiling a treatment (Phase Ia) and then jumping to test it in 50 patients, Phase I encourages tinkering with the intervention before testing efficacy. Once plausibility is achieved in Phase IIa, the researcher is ready to move towards efficacy testing.

Example in Action: Through Phase Ia exploratory work, the clinician tinkers with delivering the assembled intervention multi-component exposure therapy by experimenting with content and selected outcome measures in their clinical practice to solidify the hypothesised pathway (see **Supplemental Figure 1, Panel A)**. Then, in a case series of 5 consecutive patients (who represent the target population) in their clinic, they demonstrate plausibility of clinically significant improvement by 100% of patients achieving the target clinical endpoints in both the primary outcome (symptom severity) and the intermediate outcome (fear of GI symptoms) (Phase IIa proof-of-concept). However, after conducting a series of refinement studies (Phase Ib below), they revisit their hypothesised pathway to adjust hypotheses with a reduced number of treatment components and altered delivery mode (see **Supplemental Figure 1, Panel B**). **Supplemental Figure 2** offers an example progression of studies.

1.2.7 | Refinement Studies (Ib)

Refinement studies aim to improve promising interventions' efficiency while preserving efficacy, which can facilitate intervention affordability and scalability. For example, studying moderators in Phase Ib optimises the intervention for specific subgroups, such that the most potent intervention is then later studied in a confirmatory trial [24]. Notably, the use of Phase Ib is not typically linear (see Figure 1). Some common foci of refinement studies and example methods are:

- Parsimony (e.g., number of components, dose and delivery mode): single-case designs, Bayesian estimation, factorial and fractional factorial (e.g., Multiphase Optimization Strategy) designs [34].
- Timing (e.g., when an intervention component is most effectively delivered, when assessments are ideally timed): 'Just-in-time' adaptive interventions and time series designs.
- Heterogeneity in response (i.e., how to identify moderators and optimise an intervention for certain subgroups): Sequential multiple assignment randomised trial (SMART) and adaptive interventions (e.g., micro-randomised trial).
- Acceptability (e.g., how to ensure patients adhere to the intervention): qualitative studies, focus groups with

stakeholders and randomised comparative effectiveness designs.

Example in Action: Qualitative interviews with clinician and patient stakeholders indicate that a group-based delivery format (rather than 1:1) would be more acceptable. The investigators return to Phase Ia to revise the treatment material to be conducive to group delivery. To identify how to optimize and refine the treatment in group format, the investigators test out different treatment components and different doses in a Phase Ib MOST design trial. They find that interoceptive exposure alone and the addition of live coaching via text message outside of sessions is the best combination for facilitating between-session exposures and achieving clinically significant improvement in both fear of symptoms and symptom severity. They may not necessarily need to conduct a separate Phase IIa proof-of-concept, as the MOST design trial findings support plausibility of the refined treatment.

1.3 | Phase II and Phase III—Preliminary Testing and Efficacy

RECOMMENDATION #4: Plausibility of clinically significant change should be demonstrated *before* moving to efficacy testing and typically does not require large samples. The term 'proof-of-concept' is appropriate, while 'preliminary efficacy' and 'pilot trial' should not be used. Clearly defined and pre-specified benchmarks or MCID thresholds in the outcome should be used to test plausibility.

RECOMMENDATION #5: Once plausibility is demonstrated, threats to the feasibility of a future efficacy trial (not feasibility of the intervention) should be identified and addressed (and tested, as applicable). If a pilot feasibility trial is conducted, it should mirror the planned future efficacy trial and should test pre-specified feasibility benchmarks.

RECOMMENDATION #6: Once plausibility is demonstrated and feasibility is addressed, efficacy trials prioritising internal validity are appropriate. Efficacy trials should be powered based on an ability to detect *both* a clinically and a statistically significant benefit, using an MCID in response between a treatment and control. Trials testing effects on intermediate or surrogate outcomes (e.g., ORBIT Phase IIc) are sufficient in some cases and otherwise serve as a milestone before a confirmatory trial testing effects on the primary clinical outcome (e.g., ORBIT Phase III).

1.3.1 | Overview

Investigators may iteratively refine a BGBT by toggling between early Phase II studies and Phase I refinement—with the goal of moving to a definitive trial (Phase II and/or III efficacy testing) with the most potent, efficient BGBT. Figure S2 provides an example (iterative) progression of BGBT testing between Phase 1b and Phase IV. Before arriving at a confirmatory Phase IIc or III trial, substantially more time will be spent tinkering with the treatment package, specified outcomes and trial procedures and then testing the efficacy of the treatment in improving its behavioural risk factor targets in Phase I and early Phase II. Historically (among all behavioural treatments, not only BGBT), a common pitfall has been to compile a treatment manual based on theoretical benefit and to quickly move to efficacy testing. This approach has been associated with many behavioural treatments failing because they have been underdeveloped. One reason for this is the common use of underpowered, miniature randomised controlled trials (RCTs) proposed as tests of 'preliminary efficacy', which are now no longer recommended [24, 35]. First, the use of 'preliminary efficacy' trials have resulted in many treatments never proceeding to true efficacy testing in confirmatory trials. Second, these trials provide false reassurance for power estimates in a later confirmatory trial—the standard errors of effect size estimates generated from these trials will under- or overestimate power [28]. Thus, once a treatment is compiled in Phase I, investigators should not jump to efficacy testing. Below, we provide an overview of each Phase II and III and then summarise key recommendations.

1.3.1.1 | Phase II. The overall goal of Phase II is to conduct initial evaluations of the completed intervention's clinical impact and prepare for efficacy testing, akin to the drug development model. However, in contrast to the drug development model, Phase II may be the end goal for efficacy testing for some behavioural treatments [24]. Phase II specifically has three sub-phases:

1.3.1.2 | Phase IIa. Proof-of-concept studies identify if the treatment can *plausibly* produce *clinically* significant change, with a specific benchmark target for the intermediate or the primary/surrogate outcome. Note that a proof-of-concept study does *not* necessitate a comparator condition, but some studies may have a randomised design (see Mathew et al. [36] as an example). Typically, Phase IIa studies test the treatment in a small sample usually with time series, quasi-experimental or uncontrolled designs. If plausibility of producing clinically significant change is achieved, then it is appropriate to proceed to efficacy testing (Phase IIc efficacy and/or Phase III efficacy). If a Phase IIa study is *not* conducted before proceeding to testing efficacy, it is possible that the BGBT is not yet potent enough to produce clinically significant change—conducting an RCT instead would be waste of time and money, so Phase IIa study failure facilitates ease in going back to Phase I studies.

1.3.1.3 | Phase IIb. Feasibility studies that aim to determine if the anticipated RCT can be conducted as planned—by identifying, testing and addressing threats to feasibility of the planned RCT. Data from feasibility studies can be used to support applications for funding the future RCT. They are *not* about the intervention itself—for example, adherence to between-session homework with text messaging would be explored in earlier phase testing. Feasibility studies may include data mining study of electronic health records, stakeholder surveys/focus groups, experimentation with different trial procedures and non-randomised (open-label) trials. In select cases, a randomised pilot trial is recommended—for example, a feasibility pilot RCT should be strongly considered to test feasibility for a planned Phase III multi-site efficacy trial.

A pilot feasibility RCT is a smaller version of the planned RCT. See the 2010 CONSORT guidelines for more guidance [37]. However, within feasibility RCTs, other phase work could be done—see Hooker et al. [38] for an example pilot feasibility RCT protocol to test feasibility of recruitment and retention, while also refining fidelity of training interventionists.

1.3.1.4 | Phase IIc. The primary question posed by a Phase IIc efficacy trial is if a behavioural treatment is causally related to clinically significant improvement in an intermediate outcome (e.g., key behavioural or biologic mechanism contributing to primary outcome) or a surrogate outcome (i.e., when the primary outcome is not measurable) better than chance, the passage of time or compared to non-specific attention/other behavioural or medical treatments available [24]. When the primary outcome (for testing in a Phase III trial) is not as readily measurable (e.g., long-term impacts on healthcare costs), a Phase IIc efficacy trial would be a milestone towards Phase III testing by testing a readily measurable intermediate outcome that has a known association with the primary outcome. The idea is that any efficacy shown from a Phase IIc trial would hopefully mean that the BGBT would be efficacious to produce clinically significant improvement in the primary outcome (e.g., symptom severity) in a Phase III trial. However, in some cases, a Phase IIc trial may be the end goal for efficacy testing, when there is a well-established threshold for improvement in the intermediate or surrogate outcome that is associated with a clinically significant improvement in the primary outcome. Whether or not a Phase III trial is needed is ultimately up to the investigator based on their knowledge of the field. Phase III efficacy data may be needed to justify change in practice to convince developers of clinical guidelines, third party payors and institutional decision makers.

1.3.2 | Phase III

Highly similar to Phase III drug development, the primary question posed by a Phase III efficacy trial is if a behavioural treatment is causally related to clinically significant improvement in the primary outcome, offering a level of evidence to support (or not) incorporation of the BGBT into clinical practice. Accordingly, efficacy should be established based not only on statistical significance but also on *clinical* significance, which is aligned with the FDA's guidance [39], and a previous Rome Foundation Working Team Report on patient reported outcomes [40]. To establish clinical significance, best practice recommendations for behavioural trials suggest evaluation of the MCID [41–43] in the primary outcome (see the MCID section for more information).

There are many elements of rigour to consider for Phase II and III studies, particularly depending on the study's purpose. Below, we describe key considerations for efficacy testing.

1.3.3 | Pre-Specified Benchmarks for Success

Outcomes should be clearly specified a priori. The benchmark for the outcome, if used, will depend on the type of trial.

1.3.3.1 | Feasibility Benchmarks. If a trial is conducted to evaluate feasibility (Phase IIb), assessments of feasibility (and benchmarks as relevant) should be pre-specified. Not all feasibility parameters need to have benchmarks, so investigators may only pre-specify the benchmarks most relevant to the success of their next step trial. For more information on feasibility, see the 2010 CONSORT guidelines [37] and affiliated online resource [44].

Example in Action: The investigators planned a Phase IIc multi-site efficacy trial based on plausibility of the refined treatment in improving fear of symptoms (intermediate outcome). Because the planned RCT is multi-site, they conduct the below feasibility studies:

a. Semi-structured feedback interview groups with 5 patients from each planned site. They present the planned RCT procedures to the patients and adjust it as needed based on their feedback (e.g., decrease number of assessments, increase remuneration).

b. Pilot feasibility RCT (a smaller version of the planned efficacy trial) from which they descriptively report findings to inform multi-site implementation.
   - Estimate the rate of recruitment, reporting on the number screened, number eligible, and number enrolled per month over one-year at each site. They set a benchmark of ≥X patients per site per month.
   - Ability to retain participants through X-month follow-up (no benchmark).
   - Ability to conduct multi-site fidelity training and assessments with interventionists (no benchmark).

The investigators use the feasibility data to support their grant application for the efficacy trial and plan to report the feasibility study findings within the methods of their efficacy trial paper.

1.3.3.2 | The MCID. Clinically significant improvement can be established with an MCID target in the outcome—this can apply to all outcomes, not just the primary outcome. A binary MCID target would be a response rate, the proportion of participants achieving the MCID outcome [40] (e.g., ≥X% reporting 'adequate relief' at post-treatment, Figure 3). For a continuous MCID, the target may be achievement of a certain level of change (e.g., X point change, ≥X% change or achievement of <X score). The target for continuous MCIDs can be identified in earlier pre-Phase I and Phase I work or based on psychometric studies of the continuous outcomes or clinical recommendations. The MCID target will depend on the clinically significant question—for example, if the end goal is to demonstrate efficacy of the BGBT for a subgroup of patients who have a certain characteristic associated with a high non-response rate to existing treatments, the MCID response rate may be based on existing treatments' response rates in the target population.

A review of MCID and targets that have been identified for DGBI (mostly IBS) is out of the scope of this review, but various methods have been used to identify MCID for DGBI. Anchor- and distribution-based methods are optimal (especially compared

to consensus approaches) [45, 46]. Below, we provide considerations for the MCID depending on the type of study.

1.3.3.3 | MCIDs for Proof-of-Concept. In Phase IIa, plausibility of the treatment in producing clinically significant change is determined by a pre-specified response rate for achievement of the MCID. Typically, the response rate is reported as the proportion of participants who achieved the binary or continuous threshold in the outcome. Appropriate benchmarks for *what* proportion of individuals achieving the MCID is needed to demonstrate plausibility in a Phase IIa trial are subjective, but may be identified based on the success rates of established treatments.

> **Example in Action**: The proportion of participants achieving ≥X% improvement in symptom severity (4/5 participants in **Figure 3, Panel A**) would demonstrate *plausibility* of the intervention producing clinically important change.

1.3.3.4 | MCIDs in Confirmatory Trials. Confirmatory efficacy trials (Phase IIc and Phase III) should test an intervention that has demonstrated plausibility of producing the chosen MCID and achieved a certain response rate in Phase IIa.

> **Example in Action**: In a confirmatory trial, efficacy of the intervention in producing clinically significant improvement would be shown by the treatment arm achieving the MCID target but not the comparator arm(s) (**Figure 3, Panel B**). The investigators may choose to have a Phase IIc trial be the end goal for efficacy testing because of a well-known association between the intermediate outcome MCID and the primary outcome MCID (**Supplemental Figure 2**).

We want to particularly highlight an issue with the use of MCIDs for efficacy testing. The use of a response rate suggests that the statistical test should evaluate if a larger proportion of patients in the treatment group achieved the MCID vs. the comparator. However, in BGBT trials to date, typically the response rate is relegated to a secondary outcome, and the primary outcome is the difference in the continuous outcome between arms—this may be because when we dichotomise an underlying continuous improvement into improved/not-improved and perform statistical tests on this binary outcome, we lose a considerable amount of statistical power [47], compared to using parametric statistics (such as linear mixed models) on the continuous outcome. We want to underscore that investigators *can* use an MCID target as the primary outcome and still have the main hypothesis testing be based on comparison of mean differences in the outcome between groups using parametric statistics. There may be an added benefit of smaller sample size requirements than comparing binary response rates. See Supporting Information for a further description of how to power a study based on an MCID that ties statistical significance with clinical significance.

1.3.4 | Strong Internal Validity—Selection of a Clinically Meaningful Comparator

RECOMMENDATION #7: Comparators in BGBT trials are not one-size-fits-all and should have a documented rationale based on the *purpose of the trial* and the *purpose of the comparator.*

Phases II and III trials are implemented in a controlled fashion with a population selected to minimise threats to internal validity, thereby maximising a causal inference between a treatment and the clinically significant improvement. Selection of a clinically meaningful comparator here is an important consideration for internal validity in BGBT trials. Not all Phase II studies will have a

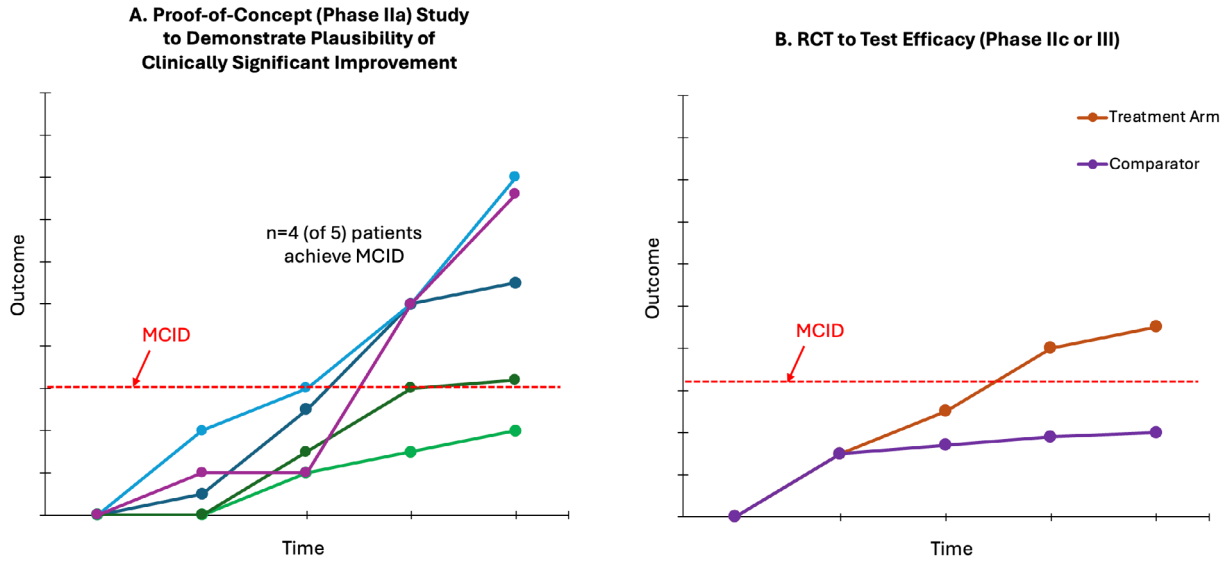

FIGURE 3 | Use of an MCID in pre-efficacy testing (Panel A) and efficacy testing (Panel B). Panel A displays a proof-of-concept study with lines for each participant; four of five participants surpass the pre-specified response rate for the minimum clinically important difference (MCID) in the outcome. The proof-of-concept data then supports efficacy testing (Panel B). We do not provide specific numbers for the outcome on the *y*-axis, as the measurement values for targets varies—for example, a target may be continuous (e.g., MCID could be % change, average change, or average absolute score) or binary (e.g., the proportion of participants achieving the MCID). RCT, randomised controlled trial.

comparator condition. A Phase IIb study that is specifically a pilot feasibility RCT or a Phase IIc efficacy trial will have comparators, and ideally the same comparators that will be tested in a planned Phase III efficacy trial. Notably, there is no (and there does not need to be) one-size-fits-all comparator for BGBT. We recommend investigators use the NIH OBSSR-supported Pragmatic Model for Comparator Selection [48] and the Purpose-Guided Trial Design Framework [49] to inform their comparator selection. These tools take a pragmatic approach based on purpose:

- *Purpose of the trial*: The choice of a comparator would likely differ between an intervention-oriented trial (e.g., Phase Ib refinement study), an outcome-oriented trial (e.g., Phase IIc efficacy trial) and a utility-oriented trial (e.g., Phase IV effectiveness study).
- Purpose of the comparator: Some comparators serve a control function (i.e., to rule out other explanations for the effect of the BGBT), while others serve a comparative function (i.e., to identify whether there is a difference between the conditions, e.g., a new treatment vs. the standard of care in a comparative effectiveness trial). In some studies, the comparator may serve both functions, possibly with differing degrees.

Example in Action: The investigators aimed to identify if their exposure therapy intervention augments existing GI practice. They considered three options below, deciding on option C.

a. Usual care (i.e., observation of what participants receive in routine care) vs. their treatment (usual care + exposure therapy). A limitation of this design is that the usual care alone group would have less study contact, which is not insurmountable but would be acknowledged as a limitation for efficacy testing; this would not be a limitation for a Phase IV effectiveness study.
b. First-line medication vs. their treatment (medication + exposure therapy). They identified for their DGBI of interest, this option had a high risk of recruitment difficulties because the first-line medication is already readily available and prescribed in general practice—for example, prospective participants could already be on the medication, have previously tried it, or not be willing to taper off other medications.
c. Enhanced usual care (usual care + education sessions) vs. their treatment (usual care + exposure therapy). This design potentially gets around degree of study contact but poses new threats related to feasibility of delivery (how to assign therapists, potential benefit of education).

1.3.5 | Basic Elements Necessary for Clinical Trial Designs

RECOMMENDATION #8: Given that it is not possible for BGBT trials to be double-masked, efforts to extend masking must be conducted, including masking staff and participants to hypotheses, utilising masked outcome assessors and investigators holding a scientific mindset of equipoise about the treatment arms' effects.

Phase II and III studies that involve a clinical trial are implemented in a controlled fashion with a population selected to minimise threats to internal validity, thereby maximising a causal inference between a treatment and the clinically important health outcome. In this section, we highlight *some* key elements that are of particular relevance to BGBT trials that should be considered in any trial—readers should refer to other sources [24, 25] for detailed guidance on other necessary trial design elements (e.g., adverse event reporting and intent-to-treat analysis). Investigators should consider describing what they decided to do (or not do) for each of these elements and state why they made the decision. For example, if an investigator has an unmasked study staff member to complete post-randomisation evaluations, a reason for this decision should be stated and any limitations of this decision noted.

1.3.5.1 | Protection of Random Assignment. Follow the CONSORT guidelines [50] to ensure concealed, truly random allocation sequence ideally with external sequence generation. Intent-to-treat analysis preserves random allocation is the primary analysis that pairs well with any per-protocol analyses as secondary. Sub-group analyses or covariate adjustment should be pre-specified. Potential confounders should be identified with stratification considered and justified. Evaluation of treatment adherence should be considered, as treatment arms may differ in their demands on participants.

1.3.5.2 | Preferences, Equipoise and Masking. It is not possible for a BGBT randomised trial to be double-masked. While not all drug trials are 100% double-masked, behavioural trials inherently have lower quality ratings due to inability to be double-masked, whereas in double-masked drug trials where treatments look identical, treatment arms in behavioural trials look different. Accordingly, BGBT trials are at risk for expectancy biases from participants, research staff and investigators as well as other stakeholders (e.g., the participant's non-study gastroenterologist) that can confound results particularly in randomised trials. For example, allocation is clearly visible in an online self-guided treatment compared to face-to-face treatment which may raise risk for treatment contamination if a participant in the online arm seeks out face-to-face treatment on their own. Similarly, naming treatment arm 'cognitive-behavioural therapy' versus 'education' seemingly suggests a preference for the former, raising risk for differential adherence and/or drop-out. However, there are ways to reduce these risks, including risks of placebo or nocebo response.

First, staff and participants should be *masked to hypothesis* (not necessarily aims), with investigators holding a scientific mindset of *equipoise* about the treatment arms' effects—for randomised trials, this may include careful selection of comparators (e.g., active comparators) and neutral naming of all trial arms. Second, investigators can push to extend the mask as much as possible by identifying *which components* of the trial (e.g., hypotheses, assignment and intervention fidelity) and *who* (e.g., post-randomisation assessors and co-investigators) can be masked. Third, a trial should include a detailed plan for discussion and assessment of participant's engagement in other interventions—the prevention of co-interventions could even occur by educating potential participants before randomisation about risks of co-interventions' negative impact on a trial's outcomes [24]. Finally, investigators can measure treatment credibility and expectancy post-randomisation [51]. For

evaluating risk of bias after trial completion, see the Cochrane Collaboration Risk of Bias-2 scale [52].

1.3.5.3 | Patient-Centred Eligibility. Patient-centred eligibility can facilitate recruitment and retention by identifying a representative target population who is motivated to engage in the intervention and requirements of the trial. Retention strategies may start at the beginning of potential participant engagement (e.g., an orientation session during recruitment) [53]. Selecting a population that is likely to adhere to both the intervention and the assessments is also key, which may be assessed with a run-in period. For randomised designs, using patient-centred eligibility may facilitate the protection of random assignment (see Powell et al. [24] for more information).

1.4 | Phase IV Effectiveness

RECOMMENDATION #9: Once efficacy of the intervention is confirmed, effectiveness/pragmatic studies prioritising external validity (real-world practice) are appropriate. If the intervention needs to be adapted for the population/environment, efficacy of that adapted intervention is needed *before* testing effectiveness.

1.4.1 | Overview of Phase IV

Effectiveness studies (sometimes referred to as pragmatic studies) have a primary goal of studying a treatment in real-world practice. The primary questions posed by a Phase IV trial include: 'does this intervention work under 'real-world' conditions?' and 'how does efficacy data translate into clinical practice?' In essence, effectiveness studies 'loosen the strictness' of traditional efficacy studies while maintaining scientific rigour. Distinguished from confirmatory efficacy trials (i.e., Phase IIc and Phase III), Phase IV trials may reduce internal validity to maximise external validity, highly similar to Phase IV drug development.

Phase IV effectiveness studies can range in study design. Within the context of pharmaceutical trials, Phase IV studies are often referred to as post-marketing studies. The type of designs used for a Phase IV study will be guided by the research question. Some examples of questions and designs are as follows:

- How does treatment X (which already has demonstrated efficacy) compare to a clinically relevant comparator such as an educational control in a gastroenterology clinic on achieving an MCID? (comparative effectiveness study)
- Can improvements in outcomes be maintained overtime for patients receiving treatment A in a clinic setting? (natural follow-up study)
- How does group-based treatment compare to individual treatment in terms of cost in dollars and quality-adjusted life-years? (cost-effectiveness study)

Due to the diverse group of study designs which address aspects of effectiveness along with the general lack of effectiveness studies, there have yet to be clear quality metrics developed for evaluating effectiveness studies. Frameworks, such as the stages in the Pragmatic Explanatory Continuum Indicator Summary-2 [54], have been developed to assist authors in distinguishing if their study contains more efficacy or effectiveness-like components [54, 55]. Although not specifically covered in this Working Team Report, a body of literature is focused on integrating components of implementation science (e.g., methods to facilitate uptake of evidence-based treatments into practice) earlier into the research trajectory through the use of hybrid designs that incorporate both effectiveness and implementation components [56, 57].

Several elements of rigour for Phase IV studies have been highlighted in previous sections including considerations for outcomes, and basic elements that are necessary for study designs. Below, we describe two elements unique to Phase IV studies.

1.4.2 | Test an Intervention That Already Has Proven Efficacy

Efficacy (i.e., from a confirmatory trial) is a prerequisite for effectiveness. To understand how an intervention works under real-world conditions, it is first necessary to establish efficacy in a controlled setting to avoid threats to external validity that could interfere with interpretation (e.g., factors related to delivery of treatment in the target population). Therefore, if the investigators decide to change the delivery method of the intervention or change patient populations, then the investigator should return to Phase I and progress through Phase III efficacy (or through Phase IIc efficacy) prior to embarking on a Phase IV trial. This is an essential point to understand. Within the DGBI literature, effective interventions are commonly modified to a new format and then progress to a Phase IV design without confirming efficacy of the new intervention format. A good rule of thumb is: if the intervention is being adapted, then you need to obtain efficacy data on the new intervention [54].

> **Example in Action**: The investigators want to take the exact same intervention tested in the confirmatory efficacy trial and expand it to other academic medical centers, so they proceed to Phase IV (see **Supplemental Figure 2**). However, if the investigators were changing the setting (e.g., to community health clinics), they would likely need to return to Phase I to identify how and what to adapt in the intervention.

1.4.3 | Maximise External Validity

One of the main goals of Phase IV is to maximise external validity—whether the findings of the study can be generalised to other situations. External validity can be considered in the selection of the patient population as well as intervention and comparator components.

1.4.3.1 | Population. Effectiveness studies should occur in 'real world' settings with an emphasis on generalisability and delivery to a diverse patient population. As such, inclusion and exclusion criteria are often 'loose' to enable recruitment of a diverse patient population. This contrasts with

efficacy testing's stricter criteria to maximise internal validity. In Phase IV, the patient population should be representative of patients seen in clinical practice. Another good rule of thumb: if you are expanding inclusion/exclusion criteria from the efficacy trial (i.e., Phases IIc or III), then you can proceed with a Phase IV trial. However, if you are *rewriting* efficacy trial inclusion/exclusion criteria (such as changing the patient population from those with irritable bowel syndrome to those with inflammatory bowel disease), then you should return to an earlier phase.

1.4.3.2 | Intervention and Comparator. There is often flexibility in intervention delivery and implementation as patients and providers experience different barriers outside of the clinical efficacy setting. Thus, any modifications to the intervention should be noted. For studies that have a comparator, it should mirror a clinically relevant option. However, having a comparator group or randomisation is not a key component of Phase IV effectiveness studies and studies may not have them (see examples of research questions and designs listed above). Careful consideration of the proposal and study design of the Phase IV trial is important, as the design will inform whether the study will have randomisation (as well as impact considerations for a control group). Using similar outcomes with efficacy trials will enable comparisons and identify implementation barriers that may influence intervention effectiveness.

> **Example in Action**: In the Phase IV study, the investigators loosen the inclusion/exclusion criteria to not require some of the stricter testing needed in the confirmatory efficacy trial—no longer requiring a negative upper endoscopy and allowing medication changes. They also expand the total timeframe for the treatment delivery.

1.5 | Recommendations for Evaluating Rigour in Proposing and Reporting on the Development and Testing of BGBT

BGBT have inherent differences from pharmacotherapies that require different evaluation criteria in both the earlier developmental phases, efficacy and effectiveness testing. In this Rome Foundation Working Team Report, we provide the first consensus statement for BGBT development and testing, heavily informed by best practice recommendations for behavioural clinical trials for chronic disease [24] and organised by a selected model (the ORBIT model [23]). By following the framework described, BGBT development and testing will align with drug development, but offer a specific framework on which to evaluate behavioural interventions. We also believe that this working team report offers a shared vernacular for how the field describes BGBT development. While *not* a change to the current gold-standard evaluation models for efficacy trials (e.g., GRADE [17]), we provide considerations for elements in behavioural intervention development that are needed for evaluating rigour within earlier development and testing phases (i.e., experimental, discovery, Phases I and II). While some of these recommendations are aspirational and may not be reflected in current and near-future studies due to funding and other constraints, we still encourage investigators to spend *more* time tinkering with a treatment package, iteratively refining it to make it most efficient and potent enough to plausibly produce clinically significant change before proceeding to efficacy testing. We also suggest that further work be conducted to define MCIDs for clinically significant targets, which meta-analyses focused on evaluating the effects of BGBT (e.g., efficacy, effectiveness, moderators and mediators) should also utilise. Finally, we also advocate the position that before testing BGBT in real-world settings (i.e., Phase IV and implementation science), the specific BGBT (including the components and delivery format) should already have established efficacy.

Our proposed consensus statement does not suggest one formula for the development and testing of BGBT, as the progression should be iterative and specific to the clinically significant question and scientific premise at hand. We recommend that investigators who are proposing and/or reporting on work related to the development and testing of BGBT use this statement as a starting point and to seek further guidance from our referenced sources including Powell et al. [24].

Author Contributions

Helen Burton-Murray: conceptualization, visualization, writing – original draft, methodology, writing – review and editing, project administration. **Livia Guadagnoli:** conceptualization, visualization, writing – original draft, writing – review and editing. **Kendra Kamp:** conceptualization, visualization, writing – original draft, writing – review and editing. **Inês A. Trindade:** conceptualization, visualization, writing – original draft, writing – review and editing. **Lynda H. Powell:** methodology, writing – review and editing. **Magnus Simrén:** conceptualization, writing – review and editing. **Brjánn Ljótsson:** conceptualization, writing – review and editing, writing – original draft, supervision. **Laurie Keefer:** writing – original draft, writing – review and editing, supervision, conceptualization, methodology.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

References

1. L. Keefer, S. K. Ballou, D. A. Drossman, G. Ringstrom, S. Elsenbruch, and B. Ljótsson, "A Rome Working Team Report on Brain-Gut Behavior Therapies for Disorders of Gut-Brain Interaction," *Gastroenterology* 162, no. 1 (2022): 300–315, https://doi.org/10.1053/j.gastro.2021.09.015.

2. V. C. Goodoory, M. Khasawneh, E. R. Thakur, et al., "Effect of Brain-Gut Behavioral Treatments on Abdominal Pain in Irritable Bowel Syndrome: Systematic Review and Network Meta-Analysis," *Gastroenterology* 167 (2024): 934–943.e5, https://doi.org/10.1053/j.gastro.2024.05.010.

3. C. J. Black, E. R. Thakur, L. A. Houghton, E. M. Quigley, P. Moayyedi, and A. C. Ford, "Efficacy of Psychological Therapies for Irritable Bowel Syndrome: Systematic Review and Network Meta-Analysis," *Gut* 69 (2020): 1441–1451.

4. A. C. Ford, B. E. Lacy, L. A. Harris, E. M. M. Quigley, and P. Moayyedi, "Effect of Antidepressants and Psychological Therapies in Irritable

Bowel Syndrome: An Updated Systematic Review and Meta-Analysis," *American Journal of Gastroenterology* 114, no. 1 (2019): 21–39, https://doi.org/10.1038/s41395-018-0222-5.

5. E. Axelsson, D. Kern, E. Hedman-Lagerlöf, et al., "Psychological Treatments for Irritable Bowel Syndrome: A Comprehensive Systematic Review and Meta-Analysis," *Cognitive Behaviour Therapy* 52, no. 6 (2023): 565–584, https://doi.org/10.1080/16506073.2023.2225745.

6. B. E. Lacy, M. Pimentel, D. M. Brenner, et al., "ACG Clinical Guideline: Management of Irritable Bowel Syndrome," *American Journal of Gastroenterology* 116, no. 1 (2021): 17–44, https://doi.org/10.14309/ajg.0000000000001036.

7. L. Chang, S. Sultan, A. Lembo, G. N. Verne, W. Smalley, and J. J. Heidelbaugh, "AGA Clinical Practice Guideline on the Pharmacological Management of Irritable Bowel Syndrome With Constipation," *Gastroenterology* 163, no. 1 (2022): 118–136, https://doi.org/10.1053/j.gastro.2022.04.016.

8. E. Savarino, F. Zingone, B. Barberio, et al., "Functional Bowel Disorders With Diarrhoea: Clinical Guidelines of the United European Gastroenterology and European Society for Neurogastroenterology and Motility," *United European Gastroenterology Journal* 10, no. 6 (2022): 556–584, https://doi.org/10.1002/ueg2.12259.

9. K. A. Gwee, S. Gonlachanvit, U. C. Ghoshal, et al., "Second Asian Consensus on Irritable Bowel Syndrome," *Journal of Neurogastroenterology and Motility* 25, no. 3 (2019): 343–362, https://doi.org/10.5056/jnm19041.

10. D. H. Vasant, P. A. Paine, C. J. Black, et al., "British Society of Gastroenterology Guidelines on the Management of Irritable Bowel Syndrome," *Gut* 70, no. 7 (2021): 1214–1240, https://doi.org/10.1136/gutjnl-2021-324598.

11. J. H. Oh, J. G. Kwon, H. K. Jung, et al., "Clinical Practice Guidelines for Functional Dyspepsia in Korea," *Journal of Neurogastroenterology and Motility* 26, no. 1 (2020): 29–50, https://doi.org/10.5056/jnm19209.

12. C. J. Black, P. A. Paine, A. Agrawal, et al., "British Society of Gastroenterology Guidelines on the Management of Functional Dyspepsia," *Gut* 71, no. 9 (2022): 1697–1723, https://doi.org/10.1136/gutjnl-2022-327737.

13. H. Miwa, A. Nagahara, A. Asakawa, et al., "Evidence-Based Clinical Practice Guidelines for Functional Dyspepsia 2021," *Journal of Gastroenterology* 57, no. 2 (2022): 47–61, https://doi.org/10.1007/s00535-021-01843-7.

14. G. Barbara, C. Cremon, M. Bellini, et al., "Italian Guidelines for the Management of Irritable Bowel Syndrome: Joint Consensus From the Italian Societies of: Gastroenterology and Endoscopy (SIGE), Neurogastroenterology and Motility (SINGEM), hospital Gastroenterologists and Endoscopists (AIGO), digestive Endoscopy (SIED), general Medicine (SIMG), gastroenterology, Hepatology and Pediatric Nutrition (SIGENP) and Pediatrics (SIP)," *Digestive and Liver Disease* 55, no. 2 (2023): 187–207, https://doi.org/10.1016/j.dld.2022.11.015.

15. L. Wauters, R. Dickman, V. Drug, et al., "United European Gastroenterology (UEG) and European Society for Neurogastroenterology and Motility (ESNM) Consensus on Functional Dyspepsia," *Neurogastroenterology and Motility* 33, no. 9 (2021): e14238, https://doi.org/10.1111/nmo.14238.

16. P. Alonso-Coello, H. J. Schünemann, J. Moberg, et al., "GRADE Evidence to Decision (EtD) Frameworks: A Systematic and Transparent Approach to Making Well Informed Healthcare Choices. 1: Introduction," *BMJ* 353 (2016): i2016, https://doi.org/10.1136/bmj.i2016.

17. "GRADE Handbook," accessed May 11, 2025, https://gdt.gradepro.org/app/handbook/handbook.html.

18. H. Burton Murray and B. Ljótsson, "Future of Brain-Gut Behavior Therapies: Mediators and Moderators," *Gastroenterology Clinics of North America* 51, no. 4 (2022): 723–739.

19. L. Nielsen, M. Riddle, J. W. King, et al., "The NIH Science of Behavior Change Program: Transforming the Science Through a Focus on Mechanisms of Change," *Behaviour Research and Therapy* 101 (2018): 3–11, https://doi.org/10.1016/j.brat.2017.07.002.

20. K. Skivington, L. Matthews, S. A. Simpson, et al., "A New Framework for Developing and Evaluating Complex Interventions: Update of Medical Research Council Guidance," *BMJ* 374 (2021): n2061, https://doi.org/10.1136/bmj.n2061.

21. L. S. Onken, K. M. Carroll, V. Shoham, B. N. Cuthbert, and M. Riddle, "Reenvisioning Clinical Science: Unifying the Discipline to Improve the Public Health," *Clinical Psychological Science* 2, no. 1 (2014): 22–34, https://doi.org/10.1177/2167702613497932.

22. S. M. Czajkowski and C. M. Hunter, "From Ideas to Interventions: A Review and Comparison of Frameworks Used in Early Phase Behavioral Translation Research," *Health Psychology* 40, no. 12 (2021): 829–844, https://doi.org/10.1037/hea0001095.

23. S. M. Czajkowski, L. H. Powell, N. Adler, et al., "From Ideas to Efficacy: The ORBIT Model for Developing Behavioral Treatments for Chronic Diseases," *Health Psychology* 34, no. 10 (2015): 971–982, https://doi.org/10.1037/hea0000161.

24. L. Powell, K. E. Freedland, and P. Kauffman, *Behavioral Clinical Trials for Chronic Diseases*. (Switzerland: Springer, 2021).

25. L. Friedman, C. Furberg, and D. DeMets, *Fundamentals of Clinical Trials*, 4th ed. Switzerland: Springer, 2010).

26. J. Zaman, N. Weltens, H. G. Ly, et al., "Influence of Interoceptive Fear Learning on Visceral Perception," *Psychosomatic Medicine* 78, no. 2 (2016): 248–258, https://doi.org/10.1097/psy.0000000000000257.

27. J. S. Labus, R. Bolus, L. Chang, et al., "The Visceral Sensitivity Index: Development and Validation of a Gastrointestinal Symptom-Specific Anxiety Scale," *Alimentary Pharmacology & Therapeutics* 20, no. 1 (2004): 89–97, https://doi.org/10.1111/j.1365-2036.2004.02007.x.

28. L. O. Walker and S. M. Czajkowski, "Designing Interventions to Improve the Health of Women Using the ORBIT Model," *Journal of Obstetric, Gynecologic, and Neonatal Nursing* 48, no. 2 (2019): 197–204, https://doi.org/10.1016/j.jogn.2019.02.002.

29. K. L. Carman, P. Dardess, M. Maurer, et al., "Patient and Family Engagement: A Framework for Understanding the Elements and Developing Interventions and Policies," *Health Affairs (Millwood)* 32, no. 2 (2013): 223–231, https://doi.org/10.1377/hlthaff.2012.1133.

30. N. J. Butcher, A. Monsour, E. J. Mew, et al., "Guidelines for Reporting Outcomes in Trial Reports: The CONSORT-Outcomes 2022 Extension," *Journal of the American Medical Association* 328, no. 22 (2022): 2252–2264, https://doi.org/10.1001/jama.2022.21022.

31. A. M. Nezu and C. M. Nezu, *Evidence-Based Outcome Research: A Practical Guide to Conducting Randomized Controlled Trials for Psychosocial Interventions* (New York, NY: Oxford University Press, 2007).

32. M. Sekhon, M. Cartwright, and J. J. Francis, "Acceptability of Healthcare Interventions: An Overview of Reviews and Development of a Theoretical Framework," *BMC Health Services Research* 17, no. 1 (2017): 88, https://doi.org/10.1186/s12913-017-2031-8.

33. B. Lobe, D. Morgan, and K. A. Hoffman, "Qualitative Data Collection in an Era of Social Distancing," *International Journal of Qualitative Methods* 19 (2020).

34. L. M. Collins and K. C. Kugler, *Optimization of Behavioral, Biobehavioral, and Biomedical Interventions*, vol. 10 (Cham: Springer International Publishing, 2018), 978.

35. H. C. Kraemer, J. Mintz, A. Noda, J. Tinklenberg, and J. A. Yesavage, "Caution Regarding the Use of Pilot Studies to Guide Power Calculations for Study Proposals," *Archives of General Psychiatry* 63, no. 5 (2006): 484–489, https://doi.org/10.1001/archpsyc.63.5.484.

36. A. R. Mathew, E. F. Avery, C. Cox, et al., "Development of a Targeted Behavioral Treatment for Smoking Cessation Among Individuals With Chronic Obstructive Pulmonary Disease," *Journal of Behavioral Medicine* 46, no. 6 (2023): 1010–1022, https://doi.org/10.1007/s10865-023-00411-z.

37. S. M. Eldridge, C. L. Chan, M. J. Campbell, et al., "CONSORT 2010 Statement: Extension to Randomised Pilot and Feasibility Trials," *BMJ* 355 (2016): i5239, https://doi.org/10.1136/bmj.i5239.

38. J. E. Hooker, J. R. Brewer, K. A. McDermott, et al., "Improving Multimodal Physical Function in Adults With Heterogeneous Chronic Pain; Protocol for a Multisite Feasibility RCT," *Contemporary Clinical Trials* 138 (2024): 107462, https://doi.org/10.1016/j.cct.2024.107462.

39. L. B. Burke, D. L. Kennedy, P. H. Miskala, E. J. Papadopoulos, and A. M. Trentacosti, "The Use of Patient-Reported Outcome Measures in the Evaluation of Medical Products for Regulatory Approval," *Clinical Pharmacology and Therapeutics* 84, no. 2 (2008): 281–283, https://doi.org/10.1038/clpt.2008.128.

40. B. Spiegel, R. Bolus, L. A. Harris, et al., "Measuring Irritable Bowel Syndrome Patient-Reported Outcomes With an Abdominal Pain Numeric Rating Scale," *Alimentary Pharmacology & Therapeutics* 30, no. 11–12 (2009): 1159–1170, https://doi.org/10.1111/j.1365-2036.2009.04144.x.

41. R. Jaeschke, J. Singer, and G. H. Guyatt, "Measurement of Health Status. Ascertaining the Minimal Clinically Important Difference," *Controlled Clinical Trials* 10, no. 4 (1989): 407–415, https://doi.org/10.1016/0197-2456(89)90005-6.

42. S. S. Leopold and R. Porcher, "Editorial: The Minimum Clinically Important Difference-The Least We Can Do," *Clinical Orthopaedics and Related Research* 475, no. 4 (2017): 929–932, https://doi.org/10.1007/s11999-017-5253-5.

43. D. Ren, T. Wu, C. Wan, et al., "Exploration of the Methods of Establishing the Minimum Clinical Important Difference Based on Anchor and Its Application in the Quality of Life Measurement Scale QLICP-ES (V2.0) for Esophageal Cancer," *Health and Quality of Life Outcomes* 19, no. 1 (2021): 173, https://doi.org/10.1186/s12955-021-01808-7.

44. C. Chan, "Pilot and Feasibility Studies," 2024, https://pilotandfeasibilitystudies.qmul.ac.uk/.

45. A. Wright, J. Hannon, E. J. Hegedus, and A. E. Kavchak, "Clinimetrics Corner: A Closer Look at the Minimal Clinically Important Difference (MCID)," *Journal of Manual & Manipulative Therapy* 20, no. 3 (2012): 160–166, https://doi.org/10.1179/2042618612y.0000000001.

46. D. Revicki, R. D. Hays, D. Cella, and J. Sloan, "Recommended Methods for Determining Responsiveness and Minimally Important Differences for Patient-Reported Outcomes," *Journal of Clinical Epidemiology* 61, no. 2 (2008): 102–109, https://doi.org/10.1016/j.jclinepi.2007.03.012.

47. V. Fedorov, F. Mannino, and R. Zhang, "Consequences of Dichotomization," *Pharmaceutical Statistics* 8, no. 1 (2009): 50–61, https://doi.org/10.1002/pst.331.

48. K. E. Freedland, A. C. King, W. T. Ambrosius, et al., "The Selection of Comparators for Randomized Controlled Trials of Health-Related Behavioral Interventions: Recommendations of an NIH Expert Panel," *Journal of Clinical Epidemiology* 110 (2019): 74–81, https://doi.org/10.1016/j.jclinepi.2019.02.011.

49. K. E. Freedland, "Purpose-Guided Trial Design in Health-Related Behavioral Intervention Research," *Health Psychology* 39, no. 6 (2020): 539–548, https://doi.org/10.1037/hea0000867.

50. CONSORT checklist, "CONSORT Group," accessed May 11, 2024, https://www.consort-spirit.org/.

51. G. J. Devilly and T. D. Borkovec, "Psychometric Properties of the Credibility/Expectancy Questionnaire," *Journal of Behavior Therapy and Experimental Psychiatry* 31, no. 2 (2000): 73–86.

52. J. A. C. Sterne, J. Savović, M. J. Page, et al., "RoB 2: A Revised Tool for Assessing Risk of Bias in Randomised Trials," *BMJ* 366 (2019): l4898, https://doi.org/10.1136/bmj.l4898.

53. M. Mayhew, M. C. Leo, W. M. Vollmer, L. L. DeBar, and M. Kiernan, "Interactive Group-Based Orientation Sessions: A Method to Improve Adherence and Retention in Pragmatic Clinical Trials," *Contemporary Clinical Trials Communications* 17 (2020): 100527, https://doi.org/10.1016/j.conctc.2020.100527.

54. K. Loudon, S. Treweek, F. Sullivan, P. Donnan, K. E. Thorpe, and M. Zwarenstein, "The PRECIS-2 Tool: Designing Trials That Are Fit for Purpose," *British Medical Journal* 350 (2015): h2147, https://doi.org/10.1136/bmj.h2147.

55. L. S. Wieland, B. M. Berman, D. G. Altman, et al., "Rating of Included Trials on the Efficacy-Effectiveness Spectrum: Development of a New Tool for Systematic Reviews," *Journal of Clinical Epidemiology* 84 (2017): 95–104, https://doi.org/10.1016/j.jclinepi.2017.01.010.

56. G. M. Curran, M. Bauer, B. Mittman, J. M. Pyne, and C. Stetler, "Effectiveness-Implementation Hybrid Designs: Combining Elements of Clinical Effectiveness and Implementation Research to Enhance Public Health Impact," *Medical Care* 50, no. 3 (2012): 217–226, https://doi.org/10.1097/MLR.0b013e3182408812.

57. S. J. Landes, S. A. McBain, and G. M. Curran, "An Introduction to Effectiveness-Implementation Hybrid Designs," *Psychiatry Research* 280 (2019): 112513, https://doi.org/10.1016/j.psychres.2019.112513.

58. J. Tack, F. Carbone, L. Chang, and B. E. Lacy, "Patient-Reported Outcomes in Disorders of Gut-Brain Interaction," *Gastroenterology* 166, no. 4 (2024): 572–587.e1. https://doi.org/10.1053/j.gastro.2023.11.307.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.