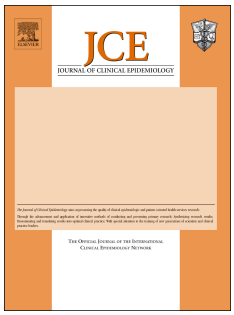
Accepted Manuscript



The Selection of Comparators for Randomized Controlled Trials of Health-Related Behavioral Interventions: Recommendations of an NIH Expert Panel

Kenneth E. Freedland, Abby C. King, Walter T. Ambrosius, Evan Mayo-Wilson, David C. Mohr, Susan M. Czajkowski, Lehana Thabane, Linda M. Collins, George W. Rebok, Sean P. Treweek, Thomas D. Cook, Jack D. Edinger, Catherine M. Stoney, Rebecca A. Campo, Deborah Young-Hyman, William T. Riley, the National Institutes of Health Office of Behavioral and Social Sciences Research Expert Panel on Comparator Selection in Behavioral and Social Science Clinical Trials

PII: S0895-4356(18)30539-0

DOI: https://doi.org/10.1016/j.jclinepi.2019.02.011

Reference: JCE 9832

To appear in: *Journal of Clinical Epidemiology*

Received Date: 13 June 2018

Revised Date: 20 December 2018

Accepted Date: 16 February 2019

Please cite this article as: Freedland KE, King AC, Ambrosius WT, Mayo-Wilson E, Mohr DC, Czajkowski SM, Thabane L, Collins LM, Rebok GW, Treweek SP, Cook TD, Edinger JD, Stoney CM, Campo RA, Young-Hyman D, Riley WT, the National Institutes of Health Office of Behavioral and Social Sciences Research Expert Panel on Comparator Selection in Behavioral and Social Science Clinical Trials, The Selection of Comparators for Randomized Controlled Trials of Health-Related Behavioral Interventions: Recommendations of an NIH Expert Panel, *Journal of Clinical Epidemiology* (2019), doi: https://doi.org/10.1016/j.jclinepi.2019.02.011.

Supplement 1: Full Report

The Selection of Comparators for Randomized Controlled Trials of Health-Related Behavioral Interventions: Recommendations of an NIH Expert Panel

Kenneth E. Freedland[1], Abby C. King[2], Walter T. Ambrosius[3], Evan Mayo-Wilson[4], David C. Mohr[5], Susan M. Czajkowski[6], Lehana Thabane[7], Linda M. Collins[8], George W. Rebok[4], Sean P. Treweek[9], Thomas D. Cook[10], Jack D. Edinger[11], Catherine M. Stoney[12], Rebecca A. Campo[13], Deborah Young-Hyman[13], and William T. Riley[13], for the National Institutes of Health Office of Behavioral and Social Sciences Research Expert Panel on Comparator Selection in Behavioral and Social Science Clinical Trials

1 Washington University School of Medicine, St. Louis, Missouri, United States of America, **2** Stanford University School of Medicine, Stanford, California, United States of America, **3** Wake Forest School of Medicine, Winston-Salem, North Carolina, United States of America, **4** Johns Hopkins University Bloomberg School of Public Health, Baltimore, Maryland, United States of America, **5** Northwestern University Feinberg School of Medicine, Chicago, Illinois, United States of America, **6** National Cancer Institute, Bethesda, Maryland, United States of America, **7** McMaster University Department of Health Research Methods, Evidence, and Impact, Hamilton, Ontario, Canada, **8** Pennsylvania State University College of Health and Human Development, University Park, Pennsylvania, United States of America, **9** University of Aberdeen Health Services Research Unit, Aberdeen, Scotland, **10** Northwestern University Institute for Policy Research, Evanston, Illinois United States of America, **11** National Jewish Health, Denver, Colorado, United States of America, **12** National Heart, Lung, and Blood Institute, Bethesda, Maryland, United States of America, **13** National Institutes of Health Office of Behavioral and Social Sciences Research, Bethesda, Maryland, United States of America.

Address correspondence to: Kenneth E. Freedland, PhD, Department of Psychiatry, Washington University School of Medicine, 4320 Forest Park Avenue, Suite 301, St. Louis, Missouri 63108, United States of America. Telephone: 314-286-1311; Fax: 314-286-1301; email: freedlak@wustl.edu .

Abstract

Objectives: To provide recommendations for the selection of comparators for randomized controlled trials of health-related behavioral interventions.

Study Design and Setting: The National Institutes of Health Office of Behavioral and Social Science Research (OBSSR) convened an expert panel to critically review the literature on control or comparison groups for behavioral trials and to develop strategies for improving comparator choices and for resolving controversies and disagreements about comparators.

Results: The panel developed a Pragmatic Model for Comparator Selection in Health-Related Behavioral Trials. The model indicates that the optimal comparator is the one that best serves the primary purpose of the trial, but that the optimal comparator's limitations and barriers to its use must also be taken into account.

Conclusion: We developed best practice recommendations for the selection of comparators for health-related behavioral trials. Use of the Pragmatic Model for Comparator Selection in Health-Related Behavioral Trials can improve the comparator selection process and help to resolve disagreements about comparator choices.

1. Introduction

Controversies and disagreements often surround the selection of comparators for randomized controlled trials (RCTs) of health-related behavioral interventions. This creates problems for investigators, reviewers, and funding agencies; impedes progress in intervention research; and diminishes the perceived quality of behavioral RCTs.

A scientific priority of the National Institutes of Health Office of Behavioral and Social Sciences Research (OBSSR) is to enhance and promote the research infrastructure, methods, and measures needed to support a more cumulative and integrative approach to behavioral and social sciences research [1]. To advance this scientific priority, OBSSR recently assembled a multidisciplinary expert panel on comparator selection in health-related behavioral RCTs. A steering committee (WTR, DYH, KEF, ACK) identified candidates for the panel who were statisticians, clinical trial investigators or methodologists, or NIH staff who had previously published relevant work on RCT methodology or who had responsibilities for research on health-related behavioral interventions. The final selections were made by the Director of OBSSR (WTR); three individuals declined the invitation because of other commitments.

Following preliminary discussions, the panel convened at the NIH campus in Bethesda Maryland on April 12 and 13, 2017. The goals of the meeting were to develop recommendations for researchers and reviewers and to produce a report to address key questions about comparator choices. The panel considered diverse areas of behavioral intervention research, including clinical treatment trials and community-based prevention trials. Disagreements were discussed, and votes were taken on the major issues, but the entire panel agreed on all major points. The main strength of this process is that it integrated the views of leading experts from diverse fields. Its main limitation is that public comments were not obtained. The recommendations reflect the

perspective of the expert panel convened by the NIH but does not represent official policy or guidance of the NIH.

This paper presents the consensus view and recommendations of the expert panel.[1] It focuses primarily on trials in which individuals are the units of randomization, but many of the principles also apply to cluster-randomized trials. Its goals are to clarify the reasons why comparators have been controversial in health-related behavioral intervention research, and to present the Pragmatic Model for Comparator Selection in Health-Related Behavioral Trials **(Figure S1)** to help resolve controversies about comparators and to refine decision-making strategies. The recommendations included in this report are intended primarily for researchers who are planning or proposing randomized trials of behavioral interventions and for peer reviewers of trial proposals and publications, rather than for meta-analysts.

[1] This document presents the full report of the expert panel. The published manuscript presents a summary of this report. Some of the table and figure numbers differ between these two versions.

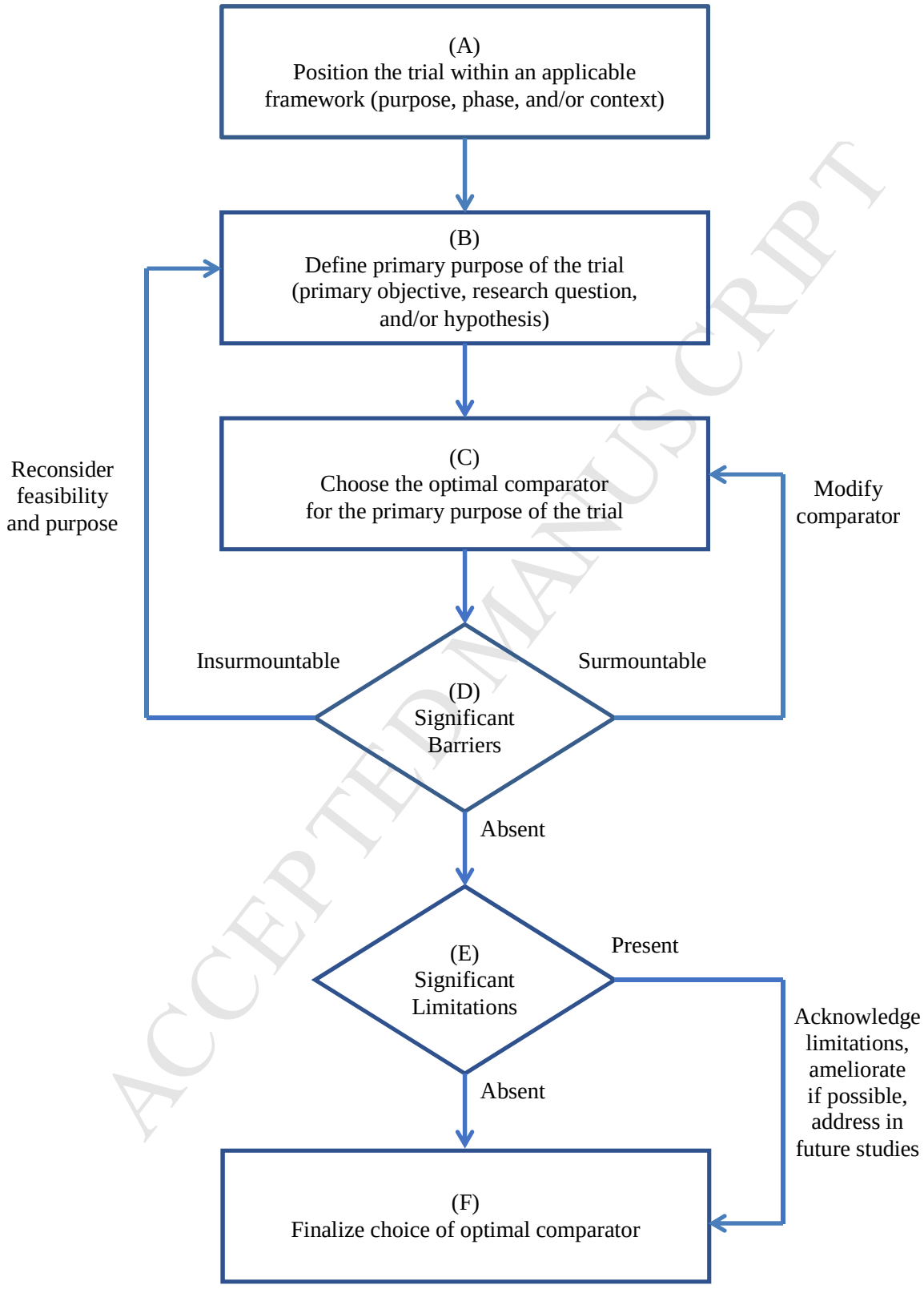


Figure S1. The Pragmatic Model for Comparator Selection in Health-Related Behavioral Trials.

2. Types and Characteristics of Comparators

The study arms to which interventions are compared are called "control" groups, arms, or conditions in some articles, and "comparison" groups, arms, or conditions in others. The distinctions between these terms are neither clear nor consistent. To minimize confusion, we use the generic term *comparator* instead of control or comparison, whenever possible throughout this report. Also, when we use the term "trial", we are referring to an RCT unless otherwise stated.

Table S1 lists comparators that are often used in health-related behavioral RCTs. A survey of behavioral and social science trials found that usual care, no treatment, and active interventions were the most frequently-used comparators in protocols published between 2012 and 2016 [2].

Comparators with the same name may take different forms in different trials. For example, a variety of conditions have been called "attention controls" or "attention-placebo control groups." One trial compared a relaxation intervention to a "health enhancement program" [3], while another compared a self-regulation intervention to "support and discussion" [4], yet both reports referred to the comparators as "attention-placebo" conditions. Some comparators, such as wait lists, have been used in trials of many kinds of behavioral interventions. Others have been reserved for more specialized applications. For example, pill placebo comparator arms have been added to trials in which behavioral interventions were compared to medications [5], but behavioral interventions are almost never compared to pill placebos in two-arm trials.

Table S1. Comparators that are often used in health-related behavioral trials.

Name(s)	Description
Comparators Often Used to Evaluate Whether an Intervention Works at All	
No treatment	No intervention is provided.
Wait list	The same intervention that is provided to the experimental group is subsequently provided to the comparator group, after the post-test evaluation has been completed.
Comparators Often Used to Determine How Well an Intervention Works Relative to a Clinically Relevant Alternative	
Usual care Routine (standard) care Treatment as usual	Treatments or services that are routinely provided in the settings from which trial participants are recruited. Often differs across individuals and settings, and in some trials may be enhanced or restricted for trial participants.
Standard of care	State-of-the-art, guideline-adherent treatments or services that are routinely provided or recommended in the settings from which participants are recruited. In some circumstances, the standard of care is provided in a uniform fashion across individuals; in others, it may be individually tailored or personalized.
Optimized care Standardized care	If a "standard of care" exists but it is not being routinely practiced in the setting(s) in which the RCT is conducted, usual care may be "optimized" or "standardized" by the investigators for the RCT to approximate the current standard of care.
Alternative intervention	Used when one intervention is compared to another intervention for the same problem, such as in comparative effectiveness trials.
Alternative modality	A condition that is identical to the experimental intervention except for the modality, channel, or source of delivery, e.g., an intervention that is delivered in person instead of by videoconference (i.e., a different delivery channel), or an intervention that is delivered by a human vs. a computer (i.e., a different intervention source).
Alternative content	A comparator that uses the same technology as the experimental intervention but that differs in content or in other details, e.g., a health behavior coaching smartphone app vs. a symptom monitoring smartphone app.

Name(s)	Description
Supported or unsupported intervention	A comparator that uses the same technology as the experimental intervention and that provides the same content, but that adds or subtracts additional components, such as a human-delivered component (e.g., counselor guidance).
Comparators Often Used to Investigate How or Why an Intervention Works	
Attention control Attention-placebo Nonspecific therapy	Umbrella terms for a variety of conditions that are usually designed to provide the same amount of contact with the intervention staff or program that will be given to the participants in the treatment arm of the trial. May be designed to control for common factors of therapy such as support, and/or other ingredients such as educational materials.
Placebo Sham	A condition that structurally resembles the experimental intervention but that lacks its putative active ingredient(s) and that is intended to alter expectancies or to stimulate a placebo response.
Component	A condition that is identical to the experimental intervention except lacking one or more of its putative active ingredients or elements.
Dosage	A condition that is identical to the experimental intervention except for one or more of its dosage parameters, e.g., frequency or duration of contacts.

As shown in **Table S2**, comparators can be characterized along several dimensions [6], including *acceptability, feasibility, formidability, relevance, resemblance, stringency, and uniformity*. The ethical *acceptability* of a comparator depends on the risks to which the participants will be exposed and whether the benefits of the research outweigh these risks. Key stakeholders, such as participants, clinicians, or community leaders, may judge the acceptability of comparators on other grounds, such as *justice*, *equity*, or *interference with customary practices*. For example, an underserved population might reject a proposed RCT in which some members of the community would receive free treatment while others would receive no treatment, because they would consider that to be unjust. As another example, physicians might be unwilling to participate in an RCT if they were concerned that random assignment of their patients to the comparator arm would interfere in some way with their clinical care.

Table S2. Key characteristics of comparators used in health-related behavioral trials.

Characteristic	Definition
Acceptability	Whether the comparator meets current standards for the ethical conduct of research and satisfies stakeholder requirements.
Feasibility	Whether the comparator can be successfully implemented, given the trial's resources and environment.
Formidability	How much pre-post change the comparator induces (or is expected to induce) in the outcome.
Relevance	How closely the comparator corresponds to "real world" interventions, services, or programs.
Resemblance	How similar the comparator is to the intervention to which it is being compared.
Stringency	How well the comparator controls for threats to internal validity and helps to minimize biases.
Uniformity	How homogeneous the comparator is across participants or sites.

The *feasibility* of a comparator pertains to the practicability of implementing it. This depends, to a large extent, on the resources available for the trial and the environment in which it will be conducted. For example, a small research grant might provide sufficient funds to compare an intervention to no treatment, but not enough to compare it to an alternative intervention.

Formidability refers to the magnitude of pre-post change that the comparator induces (or that it is expected to induce). The more (beneficial) change the comparator induces, the more difficult it is for an intervention to outperform it in a superiority trial or to approximate its performance in a noninferiority trial [7]. For example, a well-established, evidence-based intervention would be a more formidable comparator for a novel intervention than a wait list would be.

Relevance is the extent to which the comparator corresponds to real-world interventions, services, or programs. For example, an evidence-based intervention that is used routinely to treat a certain condition would be a clinically relevant comparator for novel interventions for that condition. In contrast, a "psychoeducation program" that is designed *de novo* as an attention-placebo comparator for an RCT but that is never provided as part of routine patient care or community-based services would have low clinical or public health relevance.

The *resemblance* of the comparator refers to the extent to which it resembles the intervention. Trials in which the comparator strongly resembles the intervention provide some of the only examples of double-blinding in health-related behavioral intervention research. For example, sham biofeedback equipment and procedures are so similar to genuine biofeedback, except for the verity of the feedback, that most trial participants cannot tell which group they are in. In contrast, the participants in an "aerobic exercise vs. cognitive therapy" trial could easily tell which group they are in because the interventions do not resemble one another.

Stringency refers to how well threats to internal validity and other biases are controlled. Stringent comparators are ones that help to minimize or eliminate these problems. A stringently controlled trial is one in which the outcomes can be attributed with reasonable certainty to the effects of the intervention rather than to biases, confounds, or other rival explanations.

Finally, *uniformity* refers to the homogeneity or heterogeneity of the comparator across trial participants or sites. For example, everyone in the comparator arm of a standard drug trial receives identical pill placebos; in contrast, the content of a nonstandardized "supportive counseling" condition may differ considerably across sites, counselors, and participants.

3. Sources of Controversy

The panel identified several reasons why comparators have been controversial in health-related behavioral intervention trials. First, comparator choices can have profound effects on the purpose, feasibility, fundability, results, and impact of RCTs. Disagreements about comparators are not simply methodological disputes; they have broader ramifications for intervention research [8]. In many cases, disagreements about comparator choices are proxies for disagreements about the primary purpose of the trial. Such disagreements are often resolved by replacing the planned comparator with a different one. Unfortunately, the result may be a trial that cannot answer the original research question, or one whose de facto primary purpose is not the one that the investigator had intended to pursue.

Second, the attributes discussed above are often in tension with one another. There are situations, for example, in which it might be desirable to have a usual care comparator that is both clinically relevant and very uniform. However, this combination may be unattainable because usual care tends to be clinically relevant but heterogeneous. In contrast, artificial

comparators that are designed *de novo* to balance nonspecific treatment effects or parameters between groups may be homogeneous but clinically irrelevant. When it is impossible to find or design comparators with an ideal combination of attributes, one is left to choose among imperfect alternatives. Consequently, disagreements about comparator choices often reflect differences of opinion about unavoidable tradeoffs among comparator attributes.

Third, many trials have compared health-related behavioral interventions to no-treatment or wait-list conditions. These trials have been criticized as merely showing that behavioral interventions are better than nothing [9]. They have also been contrasted with double-blind, placebo-controlled drug trials, and judged to be less stringent. This has led some researchers to conclude that health-related behavioral trials should always control for attention or placebo effects and that comparators should always be standardized (i.e., uniform). Others disagree with this solution and argue that it is likely to create more problems than it prevents.

This debate reflects a misguided tendency to equate the stringency and formidability of the comparator with the overall scientific rigor of an RCT. NIH defines scientific rigor as the strict application of the scientific method to ensure robust and unbiased experimental design, methodology, analysis, interpretation and reporting of results [10]. By extension, RCTs are rigorous to the extent that they produce trustworthy, informative, and replicable findings. This requires tighter control over explanatory variables in some trials than in others, and more formidable comparators in some trials than in others.

Many of the earliest psychotherapy and behavioral intervention studies were stringently-controlled laboratory experiments. They had the advantage of tight experimental control and high internal validity but were criticized for lacking generalizability to clinical or community-based interventions [11]. Most of these studies randomly assigned participants to conditions that

were identical except for one specific difference, to isolate the effect of the independent variable on the dependent variable. For example, a trial of desensitization of spider phobia compared massed vs. spaced exposure; there were no other differences between the conditions [12].

In contrast, contemporary health-related behavioral intervention research emphasizes generalizability and utility. This has led to a shift toward comparisons between conditions that differ in *multiple* ways rather than in single, discrete characteristics. RCTs that are conducted in clinical or community settings and that compare complex behavioral interventions to usual care exemplify these trends [13]. These are adaptive responses to growing societal needs for clinically useful research [14], but they clash with the traditional experimental paradigm of discrete and tightly-controlled independent variables.

4. Existing Approaches to Comparator Selection

4.1 Background

The panel reviewed the scientific literature on comparators as well as select research methodology textbooks [15-20] and other methodology training materials to identify existing guidance frameworks and informative perspectives. The search yielded a variety of recommendations based on *study purpose, research phase, research ethics, research context, empirical evidence, trial quality, and cumulative science*.

Table S3 lists a variety of translational research frameworks that help to clarify how the purposes, phases, and contexts of health-related intervention research affect comparator choices. An underlying premise of these frameworks is that intervention research progresses (often in a nonlinear or recursive fashion) from basic science to intervention development and refinement, to efficacy tests, to more pragmatic trials, and ultimately to implementation research. This occurs

along a dimension of *intervention maturity*. **Figure S2** shows how the frameworks map onto different regions of the intervention maturity dimension.

Table S3. Translational research frameworks.

Title	Brief Summary
International Conference on Harmonization **[21]**	Classifies drug studies by their objectives: human pharmacology, therapeutic exploratory, therapeutic confirmatory, therapeutic use. Developed primarily for pharmaceutical research, but much of it is applicable to health-related behavioral trials. Correlates the major objectives of trials with the traditional phases of drug development (Phases I, II, III, and IV). Adopted by the FDA **[22].**
Institute of Medicine's Operational Phases of Translational Research **[23]**	Arrays phases of health-related research along a continuum that ranges from basic science (T0), translation to humans (T1), translation to patients (T2), translation to practice (T3), and translation to community (T4).
Stakeholder-Informed Framework **[24]**	Argues that the risks to stakeholders of erroneous findings shift between early and late phases of research. The dominant early-phase threats are Type II errors that can stifle research on innovative interventions. The threats in later phases are Type I errors and clinical implementation of ineffective treatments.
Phase Framework for Behavioral Interventions **[25, 26]**	Suggests that comparators are often irrelevant in both the earliest and the latest phases of research. Stresses comparisons of different dosages of new interventions in Phase II and of new vs. established interventions in Phase III.
Obesity Related Behavioral Intervention Trials (ORBIT) model [49]	Provides an iterative framework for intervention development and testing. Starts with identification of a significant clinical question that may require multiple studies to address. Development and testing the intervention progresses from Phase I (design) through II (preliminary testing), III (efficacy), and IV (effectiveness). Allows returns to earlier phases for further development and testing when needed.
NIH Stage Model for Behavioral Intervention Development **[27]**	Spans Stage 0 (basic research), I (intervention generation/refinement), II (efficacy in research clinics), III (efficacy in community clinics or other relevant settings), IV (effectiveness), and V (implementation and dissemination). Encourages examination of mechanisms of behavior change at every stage of intervention development, as well as careful consideration of the intervention's implementation potential.
Multiphase Optimization Strategy (MOST) [51]	Emphasizes identification of the most efficient and effective combination of intervention components prior to testing of the intervention in an RCT. Components are initially evaluated in the screening phase, investigated in greater detail in the refining phase, and tested as a complete intervention in the confirming phase.

Title	Brief Summary
Purpose-Guided Trial Design (PGTD) [6]	Identifies four types of behavioral trials with different primary purposes. Links the dominant purpose of the comparator to the primary purpose of the trial.
Pragmatic Explanatory Continuum Indicator Summary (PRECIS-2) [52]	Describes a continuum from very explanatory to very pragmatic trials. Explanatory trials are conducted in ideal settings to give interventions the best chance of showing beneficial effects. Pragmatic trials are conducted in real world settings to inform decisions about whether to use interventions in practice.
Medical Research Council Framework for the Development and Evaluation of RCTs for Complex Interventions to Improve Health [53]	Includes four stages. 1) Development: complex intervention processes and outcomes are derived from existing evidence. 2) Feasibility and Piloting: recruitment, retention, and other procedures are tested and sample size requirements are determined. 3) Evaluation: effectiveness and cost-effectiveness are assessed and change processes are characterized. 4) Implementation: intervention is disseminated, surveillance and monitoring systems are established, and long-term follow-up studies are conducted.

#	Framework	Intervention Maturity					
		Low	⟷	**Intermediate**	⟷		**High**
1	International Conference on Harmonization (ICH) / Food and Drug Administration (FDA)	Human Pharmacology		Therapeutic Exploratory		Therapeutic Confirmatory	Therapeutic Use
2	Institute of Medicine (IOM) Phases of Translational Research	T0 Basic Science		T1 Translation to Humans		T2 Translation to Patients	T3 Translation to Practice	T4 Translation to Community
3	Stakeholder-Informed Framework		Phase I Development and Feasibility	Phase II Preliminary Testing		Phase III Large Multicenter Efficacy Trials	Phase IV Effectiveness Trials
4	Phase Framework for Behavioral Interventions		Phase I Feasibility	Phase II Dose-Response	Phase III Efficacy	Phase IV Effectiveness	Phases V-VII Implementation
5	Obesity-Related Behavioral Intervention Trials (ORBIT) Model		Phase I Design	Phase II Preliminary Testing		Phase III Efficacy	Phase IV Effectiveness
6	National Institutes of Health (NIH) Stage Model	Stage 0 Basic Research	Stage I Intervention Generation & Refinement	Stage II Efficacy in Research Clinics	Stage III Efficacy in Community Clinics	Stage IV Effectiveness	Stage V Implementation & Dissemination
7	Multiphase Optimization Strategy (MOST)		Intervention Component Screening Phase	Intervention Refinement Phase		Confirmatory Phase	
8	Purpose-Guided Trial Design (PGTD) framework		Intervention-Oriented Trials			Outcome-Oriented Trials	Utility-Oriented and Experimental Trials

#	Framework	Intervention Maturity					
		Low	↔	**Intermediate**		↔	**High**
9	Pragmatic Explanatory Continuum Indicator Summary (PRECIS-2)		Very Explanatory Trials	Rather Explanatory Trials	Rather Pragmatic Trials	Very Pragmatic Trials	
10	MRC Framework for the Development and Evaluation of RCTs for Complex Interventions	Development	Feasibility/ Piloting		Evaluation		Implementation

Fig S2. The intervention maturity dimension in translational research frameworks. The alignment of frameworks in this figure is a rough approximation; it is not intended to be interpreted as an exact map.

4.2 Study purpose

A consistent theme in the literature on RCT methodology is that *trials should be designed to serve the study's primary purpose (i.e., its main aim, objective, hypothesis, or research question).* This principle is implicit in many discussions of comparators, but it plays an especially prominent role in three methodological frameworks. First, the International Conference on Harmonisation [21, 28] and the U.S. Food and Drug Administration [22, 29] assert that the choice of the comparator for an RCT should be based on the objective of the trial. Their recommendations stress the importance of choosing comparators that are "adequate to the task," and they do not assume that that certain comparators are either inherently inadequate or inherently useful for every type of trial and every therapeutic research objective.

Second, Purpose-Guided Trial Design (PGTD) [7] is a heuristic framework for behavioral trial design. Its fundamental principle is that the design of a behavioral trial should fit its primary purpose. It posits that the primary purpose of the trial outweighs all other considerations. If the comparator that best fits the purpose of the trial cannot be used (e.g., due to ethical constraints), the PGTD framework argues that it should not be replaced by an alternative comparator that does not serve the primary purpose of the trial, because doing so would leave the primary research question unanswered. Unless there is a suitable alternative, the investigator should reconsider whether it is feasible under the circumstances to answer the primary research question.

Third, the Obesity-Related Behavioral Intervention Trials (ORBIT) model provides a systematic framework for developing and testing behavioral interventions for preventing and treating chronic diseases [30]. The model includes four phases (Phases I-IV). The primary purposes of intervention studies differ across these phases, but all of them are designed to answer a "significant clinical question" as their overriding, long-term purpose.

4.3 Research phase

Schwartz et al. [26] described four phases of behavioral intervention research. Phase I studies are for the development and testing of interventions. They are usually not RCTs, and they do not require the kind of comparators that are discussed in this report. Phase II studies compare different dosages of the same intervention. Phase III efficacy trials compare new interventions to existing ones. Phase IV studies use case-control designs to confirm effectiveness in "real-world" settings. Later phases include practice-based implementation studies (Phase V), widespread diffusion (Phase VI), and institutional- and policy-level interventions (Phase VII). Of these three, only Phase V studies typically require comparators [25]. This framework shows that different comparators are needed in different phases of behavioral intervention research.

Mohr et al. [24] noted that stakeholder interests shift across the phases of intervention research, and that this affects comparator choices. When a novel intervention is developed or an existing intervention is revamped or repurposed, stakeholder interests are best served in the early phases of research by giving the intervention a chance to demonstrate its potential value. The use of excessively formidable comparators in Phase I studies and in initial Phase II trials can decrease the chances of finding a signal, quash further work on promising interventions, and thereby deprive stakeholders of potential benefits. In contrast, later Phase II and Phase III efficacy trials are designed to inform decision-making by service providers and policy-makers. Type I errors, i.e., conclusions that interventions have benefits when in fact they do not, pose a greater risk to stakeholders in late-phase research. The use of comparators that are not very formidable in advanced Phase II and Phase III trials of behavioral interventions, including ones that have passed low-formidability tests in earlier studies, can increase the risk that ineffective

interventions will be adopted. Thus, comparators with different degrees of formidability may be needed at different phases of research on an intervention, because stakeholder interests shift.

This means that intervention developers should plan ahead during their early-phase work for the higher-formidability tests that will be conducted in later phases. An intervention may have to be refined and strengthened if early, low-formidability tests have shown promising results, to improve the chances that it will be able to withstand subsequent, higher-formidability trials.

The fractional factorial optimization trials that may be conducted as part of the multiphase optimization strategy (MOST) are based on a different logic of experimental control than an RCT, and they typically do not require traditional comparators [31, 32]. These trials examine the *components* of a complex intervention, rather than the intervention as a whole. Consequently, a comparator for the intervention as a whole, such as a placebo condition, may be unnecessary [31]. Similarly, in the Design phase (Phase I) of the Obesity-Related Behavioral Intervention Trials (ORBIT) framework for developing behavioral interventions [30], small-N or single-case designs examine the intervention's components, dosage parameters, targets, and modes of delivery without between-group comparators. In the Proof-of-Concept phase (Phase IIa), quasi-experimental, within-subject designs evaluate the intervention's ability to produce a clinically significant improvement on the behavioral target. Comparators are not essential in these studies. Thus, the phase of research affects not only choices among comparators, but also whether a comparator is needed at all.

4.3 Research ethics

The ethics of placebo- and sham-controlled trials have been vigorously debated [33-38]. Ethical issues also have been raised about other comparators used in behavioral RCTs [39].

Against this backdrop, a comparator choice that seems reasonable to an investigator may nevertheless be questioned by an institutional review board (IRB). Discussions of the potential risks and harms associated with various comparators, and recommendations for addressing ethical concerns about comparators, may be found in textbooks on the responsible conduct of research [40] and in journal articles on clinical research ethics [e.g., 34, 41, 42]. Key considerations include inequity of care, inadequate transparency (e.g., deception), risks from exposure to the comparator condition, and opportunity costs, especially if participants in the comparator arm are discouraged from seeking nonstudy care for the target of the intervention.

4.5 Research context

Contextual factors such as current standards of practice, the setting in which a trial is conducted, or the characteristics of the population, can affect the feasibility, acceptability, and stringency of comparators [43, 44]. For example, resistance to the use of certain comparators (e.g., no treatment) is common in community-based participatory research (CBPR) [45]. As another example, some comparators are infeasible in certain clinical research settings; e.g., it may be impossible to ensure that a “no treatment” group will indeed receive no treatment when interventions are available to participants through health care providers who are not involved in the trial [6]. Social context effects also play important roles in cluster randomized trials [46].

4.6 Empirical evidence

The growing empirical literature on comparators is challenging some long-held assumptions. For instance, it is often assumed that no-treatment, wait-list, and placebo comparators are interchangeable in terms of their formidability. However, recent network meta-analyses of

interventions for depression [47] and social anxiety disorder [48] found greater improvement in participants who were assigned to no-treatment or placebo than in those who were assigned to wait-list comparators. Thus, these conditions differ with respect to their formidability. This illustrates the value of empirical research to inform comparator choices.

4.7 Trial quality

Meta-analysts and clinical guideline panels often use instruments such as the Cochrane Risk of Bias tool [49] or the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) system [50] to rate their confidence in the evidence. If a trial's methodological rating could be improved by choosing one kind of comparator rather than another, it would be advantageous to take this into account. GRADE, for example, downgrades the evidence rating for "indirectness" if two interventions of interest have not been tested in head-to-head trials and their relative efficacy can only be judged in relation to other comparators [51]. For the most part, however, these rating scales do not reward certain comparator choices or penalize others. Thus, they provide few clues as to which comparators to choose for RCTs of behavioral interventions.

4.8 Cumulative science

High-quality meta-analyses of RCTs are among the best sources of evidence to guide clinical and public health policies and practices [52, 53]. Network meta-analyses can make indirect comparisons between conditions that have not been directly pitted against one another in RCTs. However, indirect meta-analytic comparisons are more vulnerable to selection biases than are meta-analyses of conditions that have been directly compared to one another in RCTs [54].

Direct, pairwise meta-analytic comparisons are possible only if enough RCTs of an intervention employ the same comparator. If comparator choices were dictated by this consideration, researchers would be obliged to employ whichever comparator had been used in previous trials of the same intervention. However, they have a higher responsibility to ensure that their own trial is optimally designed to answer the research question that it is intended to answer, even if that means using a different comparator than has been used in previous trials. Whether the trial will meet the inclusion criteria for future meta-analyses is a less important concern for clinical trialists.

Comparative effectiveness research (CER) and comparative efficacy trials have raised some additional questions about the cumulative science of intervention research. One way to compare two interventions is to conduct a randomized, head-to-head trial. Another way is to conduct separate trials in which each intervention is tested against the same kind of comparator (e.g., a placebo control condition), and then to aggregate the findings in a meta-analysis. Although network meta-analyses can be very informative, they are more vulnerable than are direct, head-to-head trials to biases due to differences in the samples, interventions, comparators, and other factors. All else being equal, direct comparisons receive higher GRADE "quality of evidence" scores than indirect comparisons [55].

However, there are also some problems with head-to-head comparisons. One is that relatively new interventions rarely yield dramatically better outcomes than well-established, evidence-based interventions. Consequently, trials that test whether a newer intervention is superior to an established intervention usually have to be powered to detect relatively small differences between the interventions, so they require large samples [56]. Another is that it is not uncommon for the sponsor or the investigator(s) to have a vested interest in one of the interventions but not

in the other, a situation that can lead to biased outcomes. This has been problematic in industry-sponsored pharmaceutical research [57, 58], but it can also affect behavioral intervention research when the investigators have an allegiance to a particular intervention [59].

5. General Principles of Comparator Selection

5.1 Optimal comparator for the research question

Based on its critical review of the literature on comparators, the panel unanimously agreed that *compatibility with the primary purpose of the trial is the single most important consideration in choosing a comparator*. The *optimal comparator* is the one that will provide the clearest answer to the primary research question or the strongest test of the trial's primary hypothesis. The rationale for the choice of the comparator should start with the primary purpose of the trial, and it should not be premised on less important considerations or on arbitrary rules.

However, there may be barriers to the use of the optimal comparator in some circumstances. Also, the comparator that best fits the trial's primary purpose may leave other questions unanswered or impose other limitations on the study. If the comparator's optimality for the primary purpose of the trial would be diminished by addressing these questions or limitations, it may be better to address them in subsequent trials instead. Thus, it is necessary to consider not only the trial's primary purpose but also barriers and limitations when choosing a comparator.

Investigators should clearly explain their choice of comparator, disclose any alternatives that were considered, explain why they were rejected, and acknowledge any salient limitations related to the comparator. Reviewers who are called upon to evaluate research, should, in turn, judge comparator choices first and foremost in relation to the trial's primary purpose, hypothesis, or research question, while recognizing the limitations such decisions often incur. They should

also be wary of requesting changes in comparators that would change the primary purpose of a proposed trial.

For example, the primary purpose of a randomized comparative effectiveness trial is to compare two interventions to one another, not to evaluate the effects that either one might have relative to no intervention [60]. Consequently, a no-treatment comparator arm would be superfluous in this type of trial. The absence of a no-treatment arm might leave some potentially interesting questions unanswered, but it would not prevent the trial from achieving its main aims.

5.2 Barriers

If there is an insurmountable barrier to the use of the comparator that best fits the purpose of the trial, the investigator should consider whether it can be modified to overcome the barrier without sacrificing its goodness-of-fit. If that is not possible (e.g., if it would be unethical to randomize participants to a no-treatment arm that would deprive them of essential care), the investigator should determine whether a different comparator could overcome the barrier while still fitting the primary purpose. If one cannot be identified, then it may not be feasible to conduct the trial in a way that would answer the primary research question or test the primary hypothesis. Faced with this dilemma, the investigator should reconsider the purpose and design of the study, and decide whether a different question, hypothesis, or design should be pursued instead.

It may be possible to conduct an informative test of the primary hypothesis or to answer the primary research question if minor modifications of the optimal comparator would suffice, but not if a major modification or a replacement comparator is required. In this circumstance, it might be better to consider pursuing other objectives instead. Also, when ethics board members,

grant reviewers, and others who are charged with evaluating trial designs ask for substantial modification or replacement of an optimal comparator, they should consider whether this will prevent the investigator from testing the primary hypothesis or answering the primary research question. Of course, it is to the investigator's advantage to try to anticipate and address any ethical issues or other concerns that might be raised about the comparator, *before* submitting a proposal to conduct an RCT of a health-related behavioral intervention.

5.3 Limitations

A comparator may be the best choice for testing a trial's primary hypothesis or answering its primary research question yet be less than ideal in other respects. It may leave some secondary or exploratory questions unanswered or create opportunities for certain biases to affect the trial. Unlike ethical or resource constraints or vigorous stakeholder objections, such limitations do not create absolute barriers to the use of the comparator of choice. They may, however, affect the validity or utility of some of the conclusions that may be drawn from the results.

Thus, the justification for the choice of a comparator should address any important and foreseeable limitations and clarify any methodological compromises that may have to be made in order to answer the primary research question or test the primary hypothesis. It should also consider whether the comparator can be modified to minimize its limitations without affecting its compatibility with the primary purpose of the trial, and whether an alternative comparator might be equally compatible but with fewer or less severe limitations. If significant limitations remain after modifications and alternatives have been explored, their implications should be discussed.

As a hypothetical example, assume that there is an established intervention (A), and a newer one (B). The best way to determine whether B is superior to A would be a direct, head-to-head

comparison in an RCT. However, A is a 6-session intervention and B is a 10-session intervention. Thus, the active ingredients of the interventions will not be the only difference between the arms of the trial; the amount of contact will also differ.

The choice of A as the comparator for B might be criticized because the type of intervention would be "confounded" with the amount of contact. This would not be a valid criticism, because the amount of contact is an integral feature of each intervention, not a confounder. Nevertheless, if B turns out to be superior to A, critics could argue that A might have produced better results if it were four sessions longer. The fact that this question was left unanswered could be considered a limitation of the trial. When the trial was being designed, it might have been possible to artificially lengthen A (or artificially shorten B), to equalize the amount of contact. This strategy was employed, for example, in a trial comparing psychodynamic therapy to CBT for social anxiety disorder [61]. Unfortunately, this approach leaves the primary research question unanswered; the results are uninformative as to whether unaltered B, which reflects how B is delivered in real-world circumstances, is superior to unaltered A. If the dosages of these interventions had been evaluated and optimized *before* they were compared to each other in a head-to-head RCT, it would be easier to justify the differential-dosage design. Regardless, having to justify this design would be a small price to pay for preserving the real-world relevance of both arms of the RCT.

6. The Pragmatic Model for Comparator Selection

The Pragmatic Model for Comparator Selection in Health-Related Behavioral Trials (**Figure S1**) is a new approach developed by the expert panel. Its core principle is that *the optimal comparator is the one that best serves the primary purpose of the trial*. If a comparator is chosen

on any other basis, it may not fit the trial's primary purpose. This can happen, for example, if a barrier to the use of the optimal comparator is encountered and as a result, it is replaced with one that is feasible to use but that does not serve the primary purpose of the trial. It can also happen if the optimal comparator's limitations are given greater weight than its compatibility with the trial's primary purpose, and as a result, a suboptimal comparator is chosen instead.

These problems can be prevented by following the model's decision-making algorithm, in which the comparator that best serves the primary purpose of the trial is identified *before* barriers and limitations are addressed. It may be possible to overcome some barriers through minor modifications of the optimal comparator. Others may be insurmountable, and this may compel the investigator to revisit the feasibility or purpose of the trial. The limitations of the comparator are considered only after it has been determined that there are no insurmountable barriers to its use and any surmountable barriers have been resolved. The limitations are acknowledged, ameliorated if possible, and/or addressed in future studies if it would be both feasible and informative to do so. The output of the algorithm is the comparator that best fits the primary purpose of the trial, despite its limitations.

The Pragmatic Model for Comparator Selection encourages researchers to choose *clinically relevant* rather than artificial comparators that are unlikely to ever be used in practice, unless the primary purpose of the study requires an artificial comparator to isolate a mechanism of change or a specific component of the intervention. It acknowledges that tradeoffs among comparator attributes can create unavoidable methodological limitations, and asserts that tolerable limitations should not be allowed to stand in the way of informative research. On the other hand, resource constraints and unacceptable risks to participants or to other stakeholders can pose

legitimate barriers to the use of otherwise optimal comparators. The algorithm includes a path for researchers who encounter such barriers and who must therefore seek alternatives.

7. Selected Applications of the Model

7.1 Introduction

The Pragmatic Model for Comparator Selection provides a general strategy for the selection of optimal comparators. This section discusses the application of the model to some specific challenges in behavioral RCT design. It follows the same sequence as the algorithm.

7.2 Positioning and justifying the study within an applicable research framework

The translational research frameworks listed in **Table S3** and **Figure S2** suggest that advances in intervention research typically emerge from phased research *programs,* including ones that span multiple research teams, rather than from isolated studies. By positioning and justifying their trial within an applicable research framework and by reviewing relevant studies and meta-analyses, investigators can show what led up to their trial and what is likely to follow. This helps to clarify the purpose of the trial and explicate the reasons why a certain comparator should be used and why others should not.

For example, a proposal to conduct an uncontrolled trial might be rejected on the grounds that such studies are inherently uninformative. They are indeed uninformative with respect to the efficacy or effectiveness of mature interventions (e.g., Stages II-IV in the NIH Stage Model or Phases II-IV in the ORBIT Model), but they can be very useful in Stage I (intervention development and refinement) or Phases I and II (intervention design and preliminary testing). Positioning and justifying an early-phase study within the MOST framework, the NIH Stage

model, the ORBIT model, or other applicable framework can help the investigator define its primary purpose and explain why other research questions or hypotheses (such as ones that would be more appropriate for a later phase of research) will be deferred to later trials.

As another example, pragmatic trials are typically conducted on relatively mature interventions. Because they are intended to test them under real-world conditions, usual care or treatments that are routinely provided in clinical or community practice settings are typically the most appropriate comparators in these trials [62]. The heterogeneity of usual care makes it unsuitable for RCTs in which stringent and uniform control over explanatory variables is critical, but it is consistent with the primary purpose of pragmatic trials that test the effectiveness of interventions against the heterogeneous care that participants would ordinarily receive in real-world practice. The PRECIS-2 tool [62] can be used to position a proposed pragmatic trial along the explanatory-pragmatic continuum, and thereby help to build a case for choosing an appropriately realistic comparator and for defending its methodological limitations.

7.3 Defining the primary purpose of the trial

This step in the algorithm moves from the generic objectives of the translational research frameworks discussed above to a specific aim, research question, or hypothesis. For example, a research team that has been developing and refining a novel intervention might decide, based on the ORBIT model, that their next study should be a Phase III efficacy trial. Their next task is to define the specific aims of this trial.

The Population, Intervention, Comparator, Outcome, and Timing (PICOT) format is helpful for framing clear research questions [63, 64]. It bridges the gap between the generic questions that emerge from translational research models and the specific ones that investigators have

about interventions and outcomes. PICOT makes close connections between research questions and comparators, particularly in RCTs of relatively mature interventions. It also acknowledges that RCTs estimate the *relative* efficacy or effectiveness of interventions, and that they do so in comparison to specific comparators and within particular contexts. The PICOT format thereby helps to counteract the erroneous assumption that the purpose of an RCT is to determine the absolute, comparator-independent and context-free efficacy or effectiveness of an intervention.

7.4 Choosing the optimal comparator

Once the trial's aims have been specified and justified, the next step is to choose the optimal comparator, if a comparator is needed. In some cases, several different comparators may be potential candidates. In other cases, the primary purpose or hypothesis may be tightly coupled with a particular comparator. For example, at least 15 RCTs have tested the hypothesized contribution of the eye movement component of Eye Movement Desensitization and Reprocessing (EMDR) to therapeutic outcomes [65]. All of these trials have compared EMDR with versus without eye movements; no other comparator would have fit their primary purpose.

As another example, the comparator in a noninferiority trial must be an active intervention that represents the current standard of care for the problem or disorder of interest [66]. When there is more than one well-established, evidence-based intervention, the investigator can select the one that is best suited to the study's goals and resources. However, if only one intervention is generally considered to be the standard of care, it is the only comparator that should be used to evaluate the noninferiority of a novel intervention.

Some research funding opportunities also stipulate comparators that are either acceptable or unacceptable. For example, the methodology standards for Patient-Centered Outcomes Research

Institute (PCORI) proposals state that, "generally, usual care or nonuse comparator groups should be avoided unless these represent legitimate and coherent clinical options" [67].

Many research questions that can be asked about health-related behavioral interventions pertain to *whether* the intervention works at all, *how well* it works relative to clinically relevant alternatives, or *how or why* it works. A comparator that is optimal for addressing one of these questions is unlikely to be optimal for addressing any of the others.

Whether the intervention works at all: Very low *formidability* comparators such as wait lists and no-treatment conditions are useful in early-phase trials in which the primary research question focuses on whether the intervention has any effect at all. These comparators control for changes in the outcome variable(s) that may occur without intervention (e.g., due to spontaneous recovery or regression to the mean). Consequently, between-group differences reveal whether the intervention has any effects that cannot be explained by the natural history of the target problem. This question cannot be answered if the comparator includes an intervention of some sort, even if it is a relatively weak intervention.

How well the intervention works relative to a clinically relevant alternative: *Relevance* is a key attribute of comparators in trials that are designed to determine whether an intervention produces better outcomes than a clinically relevant alternative produces. The reason is that this research question is only worth asking if the comparator is a genuine alternative. The most relevant alternatives tend to be services, interventions, or programs that are already established and available to the population of interest. Relevant comparators include a) ones that reflect existing clinical or public health practices or services (e.g., usual care or standard of care), b) alternative interventions (e.g., a well-established, evidence-based intervention as a comparator for a newer intervention), and c) clinically-relevant variations on the experimental intervention

(e.g., the same intervention except delivered via an alternative modality, such as when a face-to-face intervention is compared to the same intervention delivered via remote telehealth technology).

If multiple relevant alternatives exist, refinement of the hypothesis or research question can help to narrow the range of options. Also, some of the possible alternatives may be more informative or may have greater translational value than others.

For example, family physicians have a clinical practice guideline for promoting smoking cessation [68], but their adherence to it is inconsistent. An RCT comparing family medical practice with versus without an adjunctive behavioral intervention could be designed either with a usual care or with a standard-of-care comparator. Usual care would reflect the inconsistent approach to smoking cessation counseling that is typically found in family practice settings and would thus be a clinically relevant comparator. A standard-of-care comparator, in contrast, would require all participating physicians to adhere to the clinical practice guidelines. It would be more uniform than usual care, and probably more formidable, but it would also be less relevant to typical family medical practice.

How or why the intervention works: "How" or "why" questions often concern the extent to which the effects of complex interventions may be explained by certain components or by underlying mechanisms that these components purportedly target, rather than by other components or mechanisms. The most common variant concerns whether the components that are purportedly the unique, target-specific, or otherwise distinctive "active" ingredients of an intervention have effects on outcomes of interest that cannot be explained by nonspecific features such as attention.

Resemblance is one of the most important attributes of comparators that are used to test this kind of hypothesis, but it is difficult if not impossible to construct high-resemblance comparators for most complex behavioral interventions. Consequently, investigators often settle for ones that resemble the experimental intervention in some respects (e.g., the number of planned contacts) but not in others (e.g., the content of the sessions). For example, a multicenter RCT was conducted to determine whether the effects of maintenance CBT for recurrent depression "…might at least partially be explained by nonspecific factors"[69]. Manualized psychoeducation was used to represent these nonspecific factors. However, there were multiple differences in the content, format, and delivery of CBT and psychoeducation. Thus, the comparator did not literally consist of the nonspecific ingredients of CBT minus the special, distinctive, or target-specific ingredients of CBT. As is usually the case in this kind of trial, the "nonspecific" comparator was actually a *different intervention*, not a condition consisting entirely of the nonspecific components of the experimental intervention.

There were two other differences between the groups, one planned and the other unplanned. Both interventions included 16 sessions over 8 months, but the psychoeducation sessions were 30 minutes shorter than the CBT sessions. Only 73% of the participants completed at least 12 sessions of psychoeducation, whereas 87% completed at least 12 sessions of CBT. This reflects another common drawback of attention-placebo and other common-factors comparators in behavioral trials: they tend to be less credible and engaging than the interventions to which they are compared. Consequently, they are vulnerable to differential adherence and attrition.

Comparators such as psychoeducation that are used to address "why" questions are usually low in relevance. They are often artificial conditions that are constructed for RCTs rather than ones that are based on real-world, evidence-based practices. On the other hand, they tend to be

more formidable than wait lists, no-treatment groups, or usual care comparators in which many participants may receive no, minimal, or inadequate treatment. This helps to explain why they have also been used to address questions about "how well" interventions work. Unfortunately, their low relevance makes these comparators poor choices for such applications. In general, clinical trials have greater translational value when they compare health-related behavioral interventions to practice-relevant alternatives instead of to irrelevant ones.

Because of the pervasive problem of low resemblance, typical artificial comparators may also be poor choices for trials that ask "how" or "why" questions about the active ingredients or underlying mechanisms of complex interventions. If a suitable high-resemblance comparator does not exist and cannot be created, it may not be feasible to parse the specific and nonspecific ingredients of the intervention. This may appear to leave the investigator at a dead end, but it can be taken instead as an opportunity to pursue one of several different research strategies. First, it may be possible to turn the intervention itself into a comparator and to work towards developing a refined or novel intervention that can outperform it. This would pivot the research program from questioning the merits of an existing intervention to making progress towards the development of a more effective one. By repeating this approach over several cycles, a research program could produce a series of advances in treatment efficacy. In the long run, it will not matter whether the original intervention offered nothing more than attention, placebo, or common-factors effects, because better, more effective interventions will have supplanted it.

Second, interventions that are well grounded in the science of behavior change [27] and carefully developed within frameworks such as MOST, ORBIT, MRC, or the NIH Stage Model, have usually been thoroughly scrutinized and optimized by the time they are tested in efficacy

trials. Basing interventions on well-specified models and compelling empirical evidence can help to prevent questions about their specific components from being raised in late-phase RCTs.

Third, one of the reasons why the ostensibly special, active, or target-specific ingredients of complex behavioral interventions are often questioned is the suspicion that it might be possible to achieve similar results with simpler, less burdensome, or less expensive interventions. Nonsignificant results from superiority trials with attention-placebo or common-factors comparators should not be interpreted to mean that the interventions are equivalent. A better approach would be to design a noninferiority trial comparing the intervention to another *bona fide* intervention, i.e., one that is simpler, less expensive, and/or more convenient, and that is also suitable for real-world utilization [66]. For example, noninferiority trials have compared telephone-delivered vs. in-person genetic counseling [70, 71].

Another approach is to evaluate the cost-effectiveness of the intervention, compared to other alternatives [72, 73]. The MOST framework can also incorporate constraints on interventions such as cost and dosage limits [74]. These approaches can identify interventions whose costs are justifiable in relation to the health outcomes they produce. They are particularly relevant to technology-based interventions that are often developed not to be superior to in-person interventions but to deliver noninferior results at less cost and with greater reach and scalability.

Mismatches and multiple comparators: Mismatches between comparators and study purposes often occur when investigators attempt to answer more than one of three these types of questions in the same trial, or when reviewers ask them to do so. Answering one question well is better than answering multiple questions poorly.

When investigators want to answer more than one of these kinds questions in a single trial, or when reviewers ask them to do so, a common solution is to compare the intervention to two

different comparators (i.e., in a three-arm trial). This may seem like an efficient research strategy but it can cause more problems than it solves. For example, a proposal to evaluate whether a novel intervention is superior to usual care may be criticized for "failing to control for attention." However, a three-arm trial could be problematic in several ways. First, it is hard to justify asking whether an intervention works better than a clinically relevant alternative such as usual care at the same time that one is asking whether it provides nothing more than attention or placebo effects. Second, it is more expensive and difficult to conduct a three-arm than a two-arm trial, and it exposes more participants to experimentation. Third, three-arm trials can produce many different patterns of between-group differences and/or lack of differences; some of these patterns can be difficult to interpret and are potentially misleading [75-77]. Thus, the decision to add a third arm (i.e., a second comparator arm) to an RCT is one that should be carefully considered.

7.5 Addressing significant barriers

When a barrier to the use of a particular comparator is encountered, the investigator's first task is to determine whether it is surmountable. If a relatively minor modification can make the comparator acceptable or feasible while preserving the trial's ability to answer the primary research question or test the primary hypothesis, then the modification should be made. If that is not possible, it may be necessary to choose a different comparator, if there are others that are compatible with the trial's primary purpose. For example, if a no-treatment comparator would be unacceptable for equity or ethical reasons but a wait-list would be acceptable, the investigator should use a wait-list comparator as long as it is compatible with the primary research question.

The investigator would also have to consider the feasibility of this solution. If the goal were to estimate a short-term effect, then a wait-list comparator might be feasible; if the goal were

instead to estimate a long-term effect, then a wait-list might not be feasible. Resource constraints can also create barriers to the use of certain kinds of comparators, particularly for early-phase trials with small budgets. The investigator might have sufficient funds to provide the intervention to the experimental group but not enough to provide it to the comparison group. Consequently, a no-treatment comparator would be financially feasible, but a wait-list comparator would be infeasible. Thus, the investigator would be caught between the unacceptability of "no treatment" and the unaffordability of a wait-list comparator. In this circumstance, the impracticality of answering the original research question could become part of the rationale for pursuing a different yet still timely and important question.

7.6 Addressing significant limitations

It is often impossible to design comparators that have a perfect combination of attributes. Consequently, comparator-related limitations are unavoidable in many trials. When this occurs, the investigator may have to make some tradeoffs or compromises among comparator attributes.

Comparator choices can give rise to two different classes of limitations, structural and conditional. *Structural* limitations are built into the trial and can usually be recognized at the design or proposal stage. For example, a differential dosage of contact or attention between groups is a certainty when the experimental arm receives a behavioral intervention and the comparison group receives no treatment. In contrast, *conditional* limitations are not preordained, and they may or may not occur. For example, unanticipated differential attrition might occur in an RCT because the participants happen prefer one condition over the other. Conditional limitations are often associated with post-randomization biases or confounds, such as differential expectancies, differential attrition, or co-intervention biases.

Structural limitations can be taken into account in the comparator selection process, but conditional limitations can only be taken into account if they are foreseeable. For example, prevention of differential attrition is one of the reasons why an investigator might choose an attention-placebo comparator, but differential attrition might occur anyway. This is not to say that conditional limitations should be ignored when trials are being designed, proposed, or reviewed. Foreseeable limitations should be carefully elucidated, especially ones that could seriously jeopardize the validity of the trial's primary results.

It may also be helpful to consider whether an apparent limitation "is a feature, not a bug." Differential attention, for example, could be problematic in a dismantling trial [65], but it is inseparable from the primary research question in eHealth and mHealth trials in which an app is augmented with counselor support in one group but not in the other [78]. It is also an essential feature of the trial if the research question is whether the outcome of an intervention is superior to that of existing practices, services, or treatments that provide less (or more) attention [6].

The comparator's limitations should be judged in relation to whether they would leave the primary research question unanswered. Less serious limitations rarely constitute sufficient grounds to reject a comparator that is well suited to the primary research question or hypothesis, particularly if they cannot be eliminated and there are no better alternatives. In some cases, concerns about the limitations of a comparator are misplaced proxies for concerns about the primary research question or hypothesis. The Pragmatic Model for Comparator Selection suggests that in this circumstance, it is more productive to question the primary purpose of the trial than to criticize the comparator.

Also, many secondary questions can be addressed *before* an intervention is tested in a randomized controlled trial. However, trials are often conducted at times when a variety of

secondary questions remain to be addressed. If the limitations of a comparator leave important but secondary questions about an intervention unanswered, the model suggests that preventing the primary question from being answered is not the best solution. The unanswered questions of one trial may become the primary questions of future trials. This kind of programmatic research can, over time, overcome many of the limitations of individual trials.

7.7 Finalizing the choice of the comparator

The Pragmatic Model for Comparator Selection yields a comparator that addresses the primary research question, has no insurmountable barriers to its implementation, and no limitations that outweigh its compatibility with the primary research question or hypothesis. Investigators should document their decision process, so that readers understand why the comparator was selected and why other potential comparators were rejected.

8. Summary and Conclusions

Comparators are the lightning rods of health-related behavioral intervention research; they attract thunderbolts of controversy while diverting us from scrutinizing the purposes and goals of our trials. The Pragmatic Model for Comparator Selection in Health-Related Behavioral Trials provides a way to resolve many of the disagreements and controversies that surround the comparators that are used in behavioral intervention trials. It gives greater weight to the compatibility of the comparator with the primary research question or hypothesis than it does to the limitations of the comparator. It stresses the importance of carefully defining the primary research question or hypothesis, and of positioning and justifying every RCT within an applicable translational research framework. It also recognizes that there may be barriers to the

use of otherwise optimal comparators, and it provides a pathway a pathway to follow when such barriers are encountered. The developers of this model hope that its adoption will help investigators, reviewers, oversight boards, and other stakeholders to address comparator-related disagreements and controversies that could impede progress in health-related behavioral intervention research if left unresolved.

References

[1] Riley WT. Behavioral and social sciences at the National Institutes of Health: Methods, measures, and data infrastructures as a scientific priority. Health Psychol. 2017;36:5-7.

[2] Wang M, Sun G, Chang Y, Jin Y, Leenus A, Maaz M, et al. A systematic survey of control groups in behavioral and social science trials. Research on Social Work Practice. 2017:1-8.

[3] Howren MB, Kellerman QD, Hillis SL, Cvengros J, Lawton W, Christensen AJ. Effect of a Behavioral Self-Regulation Intervention on Patient Adherence to Fluid-Intake Restrictions in Hemodialysis: a Randomized Controlled Trial. Ann Behav Med. 2016;50:167-76.

[4] Vranceanu AM, Riklin E, Merker VL, Macklin EA, Park ER, Plotkin SR. Mind-body therapy via videoconferencing in patients with neurofibromatosis: An RCT. Neurology. 2016;87:806-14.

[5] DeRubeis RJ, Hollon SD, Amsterdam JD, Shelton RC, Young PR, Salomon RM, et al. Cognitive therapy vs medications in the treatment of moderate to severe depression. Arch Gen Psychiatry. 2005;62:409-16.

[6] Freedland KE, Mohr DC, Davidson KW, Schwartz JE. Usual and unusual care: existing practice control groups in randomized controlled trials of behavioral interventions. Psychosom Med. 2011;73:323-35.

[7] Freedland KE. Design and conduct of randomized behavioral clinical trials: Purpose-guided trial design. International Journal of Behavioral Medicine. 2014;21:S1.

[8] Mohr DC, Ho J, Hart TL, Baron KG, Berendsen M, Beckner V, et al. Control condition design and implementation features in controlled trials: a meta-analysis of trials evaluating psychotherapy for depression. Transl Behav Med. 2014;4:407-23.

[9] Rifkin A. Randomized controlled trials and psychotherapy research. Am J Psychiatry. 2007;164:7-8.

[10] National Institutes of Health. Rigor and reproducibility. Bethesda, Maryland: U.S. Department of Health & Human Services; 2016.

[11] Kazdin AE. Evaluating the generality of findings in analogue therapy research. J Consult Clin Psychol. 1978;46:673-86.

[12] Rowe MK, Craske MG. Effects of an expanding-spaced vs massed exposure schedule on fear reduction and return of fear. Behaviour research and therapy. 1998;36:701-17.

[13] Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M, et al. Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ (Clinical research ed). 2008;337:a1655.

[14] Ioannidis JP. Why Most Clinical Research Is Not Useful. PLoS Med. 2016;13:e1002049.

[15] Areán PA, Kraemer HC. High-quality psychotherapy research: from conception to piloting to national trials. Oxford ; New York: Oxford University Press; 2013.

[16] Friedman LM, Furberg C, DeMets DL. Fundamentals of clinical trials. 4th ed. New York: Springer; 2010.

[17] Gitlin LN, Czaja SJ. Behavioral intervention research: designing, evaluating, and implementing. New York: Springer Publishing Company; 2016.

[18] Kazdin AE. Methodological issues & strategies in clinical research. 4th ed. Washington, DC: American Psychological Association; 2016.

[19] Kazdin AE. Research design in clinical psychology. Fifth ed. Boston: Pearson; 2017.

[20] Shadish WR, Cook TD, Campbell DT. Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin; 2001.

[21] International Conference on Harmonisation. ICH harmonized tripartite guideline: general considerations for clinical trials. International Conference On Harmonisation Of Technical Requirements For Registration Of Pharmaceuticals For Human Use; 1997.

[22] Food and Drug Administration. International Conference on Harmonisation: Guidance on general considerations for clinical trials. Federal Register. 1997;62:66113-9.

[23] Leshner AI, Terry S, Schultz AM, Liverman CT. The CTSA program at NIH : opportunities for advancing clinical and translational research. Washington, D.C.: National Academies Press; 2013.

[24] Mohr DC, Spring B, Freedland KE, Beckner V, Arean P, Hollon SD, et al. The selection and design of control conditions for randomized controlled trials of psychological interventions. Psychother Psychosom. 2009;78:275-84.

[25] Rebok GW. Selecting control groups: to what should we compare behavioral interventions? In: Gitlin LN, Czaja SJ, editors. Behavioral Intervention Research: Designing, Evaluating, and Implementing. New York: Springer Publishing Company; 2016. p. 139-59.

[26] Schwartz CE, Chesney MA, Irvine MJ, Keefe FJ. The control group dilemma in clinical research: applications for psychosocial and behavioral medicine trials. Psychosom Med. 1997;59:362-71.

[27] Onken LS, Carroll KM, Shoham V, Cuthbert BN, Riddle M. Reenvisioning clinical science: Unifying the discipline to improve the public health. Clin Psychol Sci. 2014;2:22-34.

[28] International Conference on Harmonisation. Choice of control group and related issues in clinical trials. International Conference on Harmonization; 2000.

[29] Food and Drug Administration. Choice of control group and related issues in clinical trials. Rockville, MD: Food and Drug Administration; 2001.

[30] Czajkowski SM, Powell LH, Adler N, Naar-King S, Reynolds KD, Hunter CM, et al. From ideas to efficacy: The ORBIT model for developing behavioral treatments for chronic diseases. Health Psychol. 2015;34:971-82.

[31] Collins LM, Baker TB, Mermelstein RJ, Piper ME, Jorenby DE, Smith SS, et al. The multiphase optimization strategy for engineering effective tobacco use interventions. Ann Behav Med. 2011;41:208-26.

[32] Collins LM, MacKinnon DP, Reeve BB. Some methodological considerations in theory-based health behavior research. Health Psychol. 2013;32:586-91.

[33] Niemansburg SL, van Delden JJ, Dhert WJ, Bredenoord AL. Reconsidering the ethics of sham interventions in an era of emerging technologies. Surgery. 2015;157:801-10.

[34] Millum J, Grady C. The ethics of placebo-controlled trials: methodological justifications. Contemp Clin Trials. 2013;36:510-4.

[35] Miller FG, Wendler D. The ethics of sham invasive intervention trials. Clinical trials (London, England). 2009;6:401-2.

[36] Horng S, Miller FG. Ethical framework for the use of sham procedures in clinical trials. Critical care medicine. 2003;31:S126-30.

[37] Street LL, Luoma JB. Control groups in psychosocial intervention research: ethical and methodological issues. Ethics & behavior. 2002;12:1-30.

[38] Charney DS, Nemeroff CB, Lewis L, Laden SK, Gorman JM, Laska EM, et al. National Depressive and Manic-Depressive Association consensus statement on the use of placebo in clinical trials of mood disorders. Arch Gen Psychiatry. 2002;59:262-70.

[39] Kim SY, Miller FG. Varieties of standard-of-care treatment randomized trials: ethical implications. JAMA. 2015;313:895-6.

[40] Shamoo AE, Resnik DB. Responsible conduct of research. Third edition. ed. Oxford ; New York: Oxford University Press; 2015.

[41] Hey SP, Weijer C. What questions can a placebo answer? Monash bioethics review. 2016;34:23-36.

[42] Brim RL, Miller FG. The potential benefit of the placebo effect in sham-controlled trials: implications for risk-benefit assessments and informed consent. Journal of medical ethics. 2013;39:703-7.

[43] Bishop FL, Fenge-Davies AL, Kirby S, Geraghty AW. Context effects and behaviour change techniques in randomised trials: a systematic review using the example of trials to

increase adherence to physical activity in musculoskeletal pain. Psychology & health. 2015;30:104-21.

[44] Estellat C, Tubach F, Seror R, Alfaiate T, Hajage D, De Rycke Y, et al. Control treatments in biologics trials of rheumatoid arthritis were often not deemed acceptable in the context of care. Journal of clinical epidemiology. 2016;69:235-44.

[45] De las Nueces D, Hacker K, DiGirolamo A, Hicks LS. A systematic review of community-based participatory research to enhance clinical trials in racial and ethnic minority groups. Health Serv Res. 2012;47:1363-86.

[46] Fuller D, Potvin L. Context by treatment interactions as the primary object of study in cluster randomized controlled trials of population health interventions. International journal of public health. 2012;57:633-6.

[47] Furukawa TA, Noma H, Caldwell DM, Honyashiki M, Shinohara K, Imai H, et al. Waiting list may be a nocebo condition in psychotherapy trials: a contribution from network meta-analysis. Acta psychiatrica Scandinavica. 2014;130:181-92.

[48] Mayo-Wilson E, Dias S, Mavranezouli I, Kew K, Clark DM, Ades AE, et al. Psychological and pharmacological interventions for social anxiety disorder in adults: a systematic review and network meta-analysis. The Lancet Psychiatry. 2014;1:368-76.

[49] Higgins JP, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ (Clinical research ed). 2011;343:d5928.

[50] Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. Journal of clinical epidemiology. 2011;64:383-94.

[51] Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence--indirectness. Journal of clinical epidemiology. 2011;64:1303-10.

[52] Bero LA, Jadad AR. How consumers and policymakers can use systematic reviews for decision making. Ann Intern Med. 1997;127:37-42.

[53] Mulrow CD, Cook DJ, Davidoff F. Systematic reviews: critical links in the great chain of evidence. Ann Intern Med. 1997;126:389-91.

[54] Glenny AM, Altman DG, Song F, Sakarovitch C, Deeks JJ, D'Amico R, et al. Indirect comparisons of competing interventions. Health technology assessment (Winchester, England). 2005;9:1-134, iii-iv.

[55] Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence--indirectness. Journal of clinical epidemiology. 2011;64:1303-10.

[56] Vieta E, Cruz N. Head to head comparisons as an alternative to placebo-controlled trials. Eur Neuropsychopharmacol. 2012;22:800-3.

[57] Flacco ME, Manzoli L, Boccia S, Capasso L, Aleksovska K, Rosso A, et al. Head-to-head randomized trials are mostly industry sponsored and almost always favor the industry sponsor. Journal of clinical epidemiology. 2015;68:811-20.

[58] Gartlehner G, Morgan L, Thieda P, Fleg A. The effect of study sponsorship on a systematically evaluated body of evidence of head-to-head trials was modest: secondary analysis of a systematic review. Journal of clinical epidemiology. 2010;63:117-25.

[59] Cuijpers P, Driessen E, Hollon SD, van Oppen P, Barth J, Andersson G. The efficacy of non-directive supportive therapy for adult depression: a meta-analysis. Clin Psychol Rev. 2012;32:280-91.

[60] Jakicic JM, Sox H, Blair SN, Bensink M, Johnson WG, King AC, et al. Comparative Effectiveness Research: A Roadmap for Physical Activity and Lifestyle. Med Sci Sports Exerc. 2015;47:1747-54.

[61] Leichsenring F, Salzer S, Beutel ME, Herpertz S, Hiller W, Hoyer J, et al. Psychodynamic therapy and cognitive-behavioral therapy in social anxiety disorder: a multicenter randomized controlled trial. Am J Psychiatry. 2013;170:759-67.

[62] Loudon K, Treweek S, Sullivan F, Donnan P, Thorpe KE, Zwarenstein M. The PRECIS-2 tool: designing trials that are fit for purpose. BMJ (Clinical research ed). 2015;350:h2147.

[63] Thabane L, Thomas T, Ye C, Paul J. Posing the research question: not so simple. Can J Anaesth. 2009;56:71-9.

[64] Haynes RB, Sackett DL, Guyatt GH, Tugwell P. Clinical epidemiology : how to do clinical practice research. LWW medical book collection. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2006. p. xv, 496 p.

[65] Lee CW, Cuijpers P. A meta-analysis of the contribution of eye movements in processing emotional memories. Journal of behavior therapy and experimental psychiatry. 2013;44:231-9.

[66] Mauri L, D'Agostino RB, Sr. Challenges in the design and interpretation of noninferiority trials. N Engl J Med. 2017;377:1357-67.

[67] Patient-Centered Outcomes Research Institute. PCORI Methodology Standards. 2017.

[68] Larzelere MM, Williams DE. Promoting smoking cessation. Am Fam Physician. 2012;85:591-8.

[69] Stangier U, Hilling C, Heidenreich T, Risch AK, Barocka A, Schlosser R, et al. Maintenance cognitive-behavioral therapy and manualized psychoeducation in the treatment of recurrent depression: a multicenter prospective randomized controlled trial. Am J Psychiatry. 2013;170:624-32.

[70] Kinney AY, Butler KM, Schwartz MD, Mandelblatt JS, Boucher KM, Pappas LM, et al. Expanding access to BRCA1/2 genetic counseling with telephone delivery: a cluster randomized trial. Journal of the National Cancer Institute. 2014;106.

[71] Schwartz MD, Valdimarsdottir HB, Peshkin BN, Mandelblatt J, Nusbaum R, Huang AT, et al. Randomized noninferiority trial of telephone versus in-person genetic counseling for hereditary breast and ovarian cancer. J Clin Oncol. 2014;32:618-26.

[72] Muennig P, Bounthavong M. Cost-effectiveness analyses in health : a practical approach. Third edition. ed. San Francisco, CA: Jossey-Bass, a Wiley Brand; 2016.

[73] Neumann PJ, Sanders GD, Russell LB, Siegel JE, Ganiats TG. Cost effectiveness in health and medicine. Second edition. ed. Oxford ; New York: Oxford University Press; 2017.

[74] Bernstein SL, Dziura J, Weiss J, Miller T, Vickerman KA, Grau LE, et al. Tobacco dependence treatment in the emergency department: A randomized trial using the Multiphase Optimization Strategy. Contemp Clin Trials. 2017.

[75] Freedland KE. Demanding attention: reconsidering the role of attention control groups in behavioral intervention research. Psychosom Med. 2013;75:100-2.

[76] McDermott MM, Reed G, Greenland P, Mazor KM, Pagoto S, Ockene JK, et al. Activating peripheral arterial disease patients to reduce cholesterol: a randomized trial. Am J Med. 2011;124:557-65.

[77] Pagoto SL, McDermott MM, Reed G, Greenland P, Mazor KM, Ockene JK, et al. Can attention control conditions have detrimental effects on behavioral medicine randomized trials? Psychosom Med. 2013;75:137-43.

[78] Romero-Sanchiz P, Nogueira-Arjona R, Garcia-Ruiz A, Luciano JV, Garcia Campayo J, Gili M, et al. Economic evaluation of a guided and unguided internet-based CBT intervention for major depression: Results from a multi-center, three-armed randomized controlled trial conducted in primary care. PloS one. 2017;12:e0172741.

Figure 1. NIH Model for Comparator Selection in Health-Related Behavioral Trials