Guidelines for Designing and Evaluating Feasibility Pilot Studies

Jeanne A. Teresi, EdD, PhD,† Xiaoying Yu, MD, PhD,‡ Anita L. Stewart, PhD,§ and Ron D. Hays, PhD‖*

Background: Pilot studies test the feasibility of methods and procedures to be used in larger-scale studies. Although numerous articles describe guidelines for the conduct of pilot studies, few have included specific feasibility indicators or strategies for evaluating multiple aspects of feasibility. In addition, using pilot studies to estimate effect sizes to plan sample sizes for subsequent randomized controlled trials has been challenged; however, there has been little consensus on alternative strategies.

Methods: In Section 1, specific indicators (recruitment, retention, intervention fidelity, acceptability, adherence, and engagement) are presented for feasibility assessment of data collection methods and intervention implementation. Section 1 also highlights the importance of examining feasibility when adapting an intervention tested in mainstream populations to a new more diverse group. In Section 2, statistical and design issues are presented, including sample sizes for pilot studies, estimates of minimally important differences, design effects, confidence intervals (CI) and nonparametric statistics. An in-depth treatment of the limits of effect size estimation as well as process variables is presented. Tables showing CI around parameters are provided. With small samples, effect size, completion and adherence rate estimates will have large CI.

Conclusion: This commentary offers examples of indicators for evaluating feasibility, and of the limits of effect size estimation in pilot studies. As demonstrated, most pilot studies should not be used to estimate effect sizes, provide power calculations for statistical tests or perform exploratory analyses of efficacy. It is hoped that these guidelines will be useful to those planning pilot/feasibility studies before a larger-scale study.

Key Words: guidelines, feasibility, pilot studies, statistical issues, confidence intervals, diversity

(*Med Care* 2022;60: 95–103)

From the *Columbia University Stroud Center at New York State Psychiatric Institute, New York; †Research Division, Hebrew Home at Riverdale, Riverdale, NY; ‡Department of Preventive Medicine and Population Health, University of Texas Medical Branch at Galveston, Galveston, TX; §Department of Social and Behavioral Sciences, University of California, San Francisco, Institute for Health & Aging, San Francisco, CA; and ‖Division of General Internal Medicine and Health Services Research, University of California, Los Angeles, Los Angeles, CA.

This article was a collaboration of the Analytic Cores from several National Institute on Aging Centers: Resource Centers for Minority Aging Research (UCSF, grant number 2P30AG015272-21, Karliner; UCLA, grant number P30-AG021684, Mangione; and University of Texas, grant number P30AG059301, Markides), an Alzheimer's Disease—RCMAR Center (Columbia University, grant number 1P30AG059303, Manly, Luchsinger) an Edward R. Roybal Translational Research Center (Cornell University, grant number 5P30AG022845, Reid, Pillemer, and Wethington), and the Measurement Methods and Analysis Core of a Claude D. Pepper Older Americans Independence Center (National Institute on Aging, 1P30AG028741, Siu). These funding agencies played no role in the writing of this manuscript. X.Y. is supported by a research career development award (K12HD052023: Building Interdisciplinary Research Careers in Women's Health Program-BIRCWH; Berenson, PI) from the National Institutes of Health/Office of the Director (OD)/National Institute of Allergy and Infectious Diseases (NIAID), and Eunice Kennedy Shriver National Institute of Child Health & Human Development (NICHD).

The authors declare no conflict of interest.

Correspondence to: Jeanne A. Teresi, EdD, PhD, Columbia University Stroud Center at New York State Psychiatric Institute, 1051 Riverside Drive, Box 42, Room 2714, New York, NY 10032-3702. E-mails: teresimeas@aol.com; jat61@cumc.columbia.edu.

Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website, www.lww-medicalcare.com.

ISSN: 0025-7079/22/6001-0095

Pilot studies are a necessary first step to assess the feasibility of methods and procedures to be used in a larger study. Some consider pilot studies to be a subset of feasibility studies,[1] while others regard feasibility studies as a subset of pilot studies. As a result, the terms have been used interchangeably.[2] Pilot studies have been used to estimate effect sizes to determine the sample size needed for a larger-scale randomized controlled trial (RCT) or observational study. However, this practice has been challenged because pilot study samples are usually small and unrepresentative, and estimates of parameters and their standard errors may be inaccurate, resulting in misleading power calculations.[3,4] Other questionable goals of pilot studies include assessing safety and tolerability of interventions and obtaining preliminary answers to key research questions.[5]

Because of these challenges, the focus of pilot studies has shifted to examining feasibility. The National Center for Complementary and Integrative Health (NCCIH) defines a pilot study as "a small-scale test of methods and procedures to assess the feasibility/acceptability of an approach to be used in a larger scale study."[6] Others note that pilot studies aim to "field-test logistical aspects of the future study and to incorporate these aspects into the study design."[5] Results can inform modifications, increasing the likelihood of success in the future study.[7]

Although pilot studies can still be used to inform sampling decisions for larger studies, the emphasis now is on confidence intervals (CI) rather than the point estimate of effect sizes. However, as illustrated below, CIs will be large for small sample sizes. Addressable questions are whether data collection protocols are feasible, intervention fidelity is maintained, and participant adherence and retention are achieved.

Although many in the scientific community have accepted the new focus on feasibility for pilot studies, there has not been universal adoption. Numerous articles describe guidelines for conducting feasibility pilot studies,[8–10] both randomized and nonrandomized.[2,11] A useful next step is to augment general guidelines with specific feasibility indicators and describe strategies for evaluating multiple aspects of feasibility in one pilot study. In addition, studies of health disparities face special feasibility issues. Interventions that were tested initially in mainstream populations may require adaptation for use in ethnically or sociodemographically diverse groups and measures may not be appropriate for those with lower education or limited English proficiency.

Building on a framework developed by the NCCIH,[6] Figure 1 presents an overview of questions to address. Section 1 of this commentary provides guidelines for assessments, data collection, and intervention implementation. Section 2 addresses statistical and design issues related to conducting pilot studies.

These guidelines were generated to assist investigators from several National Institutes of Health Centers that fund pilot studies. Presenters at Work in Progress meetings have expressed the need for help in framing pilot studies consistent with current views about their use and limitations. A goal of this commentary is to provide guidance to early and mid-stage investigators conducting pilot studies.

SECTION 1: ASSESSING FEASIBILITY IN PILOT STUDIES

Assessments and Data Collection

Can Participants Comply With Data Collection Protocols?

Data can be obtained via questionnaires, performance tests (eg, cardiopulmonary fitness, cognitive functioning), lab tests (eg, imaging), and biospecimens (eg, saliva, blood). Data may vary in complexity (eg, repeated saliva samples over 3 d, maintaining a food diary), and intrusiveness (eg, collecting mental health data or assessing cognition). The logistics can be challenging, for example, conducting assessments at a clinic or university or scheduling imaging scans. With the COVID pandemic, an important issue is the feasibility of conducting assessments remotely, for example, using telehealth software.

A detailed protocol is needed to test data collection feasibility, assure assessment completion, and track compliance. Measures may require administration via tablet or laptop in the community, with secure links for uploading and storing data; links and data collection software require testing during pilot studies. For biospecimens, the protocol should include details on storing and transferring samples (eg, some may require refrigeration).

Feasibility indicators can include completion rates and times for specific components, perceived burden, inconvenience, and reasons for noncompletion,[9] all of which may inform assessment protocol modification. Assessments can be scheduled in community settings for convenience, and briefer measures may be used to reduce respondent burden. Instructions to interviewers and participants can be tested in the pilot study. For example, to facilitate compliance with a complex biospecimen collection protocol, a video together with in-person support and instruction were provided to Spanish-speaking Latinas.[12]

Are needed Data Available From Administrative Records and How Are Variables Defined and Scored?

Studies in clinical settings may use medical record data or administrative sources to assess medical conditions, and health care provider data may be used to determine eligibility. Feasibility issues include obtaining permission, demonstrating access, and capability to merge data across sources. Also important is how demographic or clinical characteristics are measured, and their accuracy and completeness. Race and ethnicity data are often obtained through the medical record, possibly as a stratification variable, but may be assessed in ways that make it of questionable validity (eg, by observation). When an important clinical measure is not available in the medical records, one can explore the feasibility of self-report measures, which may be reliable and valid in relation to objective measures, for example, weight and height[13] or CD4 counts.[14]

Conceptual and Psychometric Adequacy of Measures

Are the Measures Acceptable, Appropriate, and Relevant to the Target Population?

Measures developed primarily in mainstream populations may not be culturally appropriate for some race and ethnic groups. There can be group differences in the interpretations of the meaning of questions, or in relevance of the concept measured. In a physical activity intervention study, the activities assessed excluded those typically performed by bariatric surgery patients, thus missing important changes in activity level.[15] In a feasibility patient safety survey, respondents evaluated the usefulness, level of understanding, and whether the survey missed important issues.[16]

Qualitative methods such as cognitive interviews and focus groups are key to determining conceptual adequacy and equivalence,[17] and to ensure that the targeted sample members understand the questions.[18,19] For example, the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) methodology uses the Delphi method.[20]

Is There Evidence of Reliability and Validity (Including Responsiveness to Change) of Measures in the Target Population?

Do measures developed in mainstream populations meet standard psychometric criteria when applied to the target population? This includes potentially testing the equivalence of administering measures via paper/pencil and electronically.

Section 1: Assessing Feasibility in Pilot Studies

Assessments and data collection

Data collection protocols and data availability

Can participants comply with assessments and data collection protocols?
Are needed data available from administrative records and how are variables defined and scored?

Conceptual and psychometric adequacy of measures

Are the measures acceptable, appropriate, and relevant to the target population?
Is there evidence of reliability, validity (including responsiveness to change) of measures in the target population?

Intervention implementation

Can interventionists be recruited, trained, and retained?
Can interventionists deliver the intervention as intended (per protocol)?
Are the treatment conditions (intervention and control) acceptable to participants and interventionists?
Will participants adhere to and engage in the intervention components?

Section 2: Statistical and Design Issues in Planning Pilot Studies

Sample sizes for pilot feasibility studies

What sample size is needed for a pilot study to address feasibility questions?

Group differences and effect sizes

Can the pilot study be used to estimate group differences and generate effect sizes?
What types of statistical analyses should be proposed for pilot studies?

Specifying the minimally important difference (MID)

Are there available estimates of minimally important differences?

Variance estimates

Are there estimates from earlier studies of the variances of outcomes?

Confidence intervals

How large will confidence intervals be for process outcomes?

Problems with use of non-parametric statistics

Are non-parametric statistics a rescue method for small pilot studies?

Evaluation of randomization algorithms and specification of design features and MIDs

Randomization algorithm

Is the randomization algorithm working correctly?

Dose and separation

Is there separation between groups in terms of dose delivered?

Design effect estimates

Are there estimates of the design effects?

FIGURE 1. Framework of feasibility questions for pilot studies.

Inter-rater reliability should be established for interviewer-administered measures, and a certification form developed and tested in the pilot study.

For example, the Patient-Reported Outcome Measurement Information System (PROMIS) measures were developed with qualitative methods in an attempt to ensure conceptual equivalence across groups.[21] However, later work examined the psychometric properties in new applications and translations,[22] and physical function items were found to perform differently across language groups.[19,23] Translation or different cultural understanding of phrases or words could result in lack of measurement equivalence.

Quantitative methods include obtaining preliminary estimates of reliability (eg, test-retest, internal consistency, inter-rater), score distributions (range of values), floor or ceiling effects, skewness, and the patterns and extent of missing data, all of which are relevant for power calculations. Optimal qualitative methods to examine group differences in concepts, and quantitative methods for assessing psychometric properties and measurement equivalence were

described in a special issue of *Medical Care*[24] and later summarized.[25] Although pilot studies will not have sufficient sample sizes to test measurement equivalence, investigators can review literature describing performance in diverse groups. Identifying measures with evidence of conceptual and psychometric adequacy in the target population increases the likelihood that only minimal feasibility testing will be necessary. Feasibility testing can focus on multiple primary outcome measures to determine if one or more are not acceptable or understood as intended.

Intervention Implementation

Four aspects of the feasibility of implementing interventions are given in Figure 1. For interventionists, the questions are whether they can be recruited, trained, and retained, and whether they can deliver the intervention as intended. For participants, the main issue is whether they will adhere to and engage in the program components. The acceptability of treatment conditions pertains to both participants and interventionists. Testing feasibility is particularly important when evidence-based interventions found effective in a mainstream population are adapted or translated for a more diverse population.[26]

Specific steps for each question are summarized in Table 1, including feasibility assessment strategies and examples. A combination of quantitative and qualitative methods (mixed methods) is required for assessing implementation feasibility. Quantitative data can be obtained from structured surveys. Qualitative data are generated from open-ended interviews of interventionists or participants regarding adherence and acceptability, for example, reasons for not attending sessions or difficulty implementing program elements.

Recruiting and training interventionists, and assessing whether the intervention is delivered as intended are often overlooked, particularly when interventionists are recruited from community settings (eg, promotores, community health workers). Intervention delivery as intended (implementation or treatment fidelity) is determined by observation and structured ratings of delivery. In a feasibility study, investigators can focus on modifiable factors affecting treatment fidelity with the goal of modifying the intervention immediately if needed, thus improving the chances of resolution.

Acceptability by both intervention and control groups (how suitable, satisfying, and attractive)[33] is critical for diverse populations, to assure that treatment conditions are sensitive to cultural issues and relevant. Acceptability, reported by participants and interventionists, can be determined before implementation through formative research and debriefing postintervention interviews.

Although participant adherence to the intervention and retention are standard components of reporting (CONSORT), in a feasibility study, more detailed data are collected. Adherence can be tracked to each component, including assessment of reasons for non-adherence. If tracked in real time, results can highlight components that require modification. Interventionists can report whether participants can carry out the intervention activities[33] or have difficulty with some components, and participants can report whether components are too complicated or not useful. Adherence also includes engagement in the intervention (treatment receipt).[30] Engagement differs from adherence in that it is more focused on completion of all activities and/or practicing skills and understanding the material along the way.

SECTION 2: STATISTICAL AND DESIGN ISSUES IN PLANNING PILOT STUDIES

Sample Sizes for Pilot Feasibility Studies

What Sample Size is Needed for a Pilot Study to Address Feasibility Issues?

NCCIH notes that sample size should be based on "practical considerations including participant flow, budgetary constraints, and the number of participants needed to reasonably evaluate feasibility goals." For qualitative work, to reach saturation, sample sizes may be 30 or less. For quantitative studies, a sample of 30 per group (intervention and control) may be adequate to establish feasibility.[36]

Many rules of thumb exist regarding sample sizes for pilot studies,[37–41] resulting in a confusing array of recommendations. Using reasonable scenarios regarding statistics that may be generated from pilot studies examining process and outcome variables, relatively large samples are required. If estimates of parameters such as proportion within treatment groups adhering to a regimen or correlations among variables are to be estimated, CIs may be very large with sample sizes <70–100 per group. If the goal is to examine the CI around feasibility process outcomes such as acceptance rates, adherence rates, proportion of eligible participants who are consented or who agree to be randomized, then sample sizes of at least 70 may be needed, depending on the point estimate and CI width (Appendix Table 1, Supplemental Digital Content 1, http://links.lww.com/MLR/C359).

Group Differences and Effect Sizes

Can the Pilot Study be Used to Estimate Group Differences and Generate Effect Sizes?

Because the focus is on feasibility, results of statistical tests are generally not informative for powering the main trial outcomes. In addition, feasibility process outcomes may be poorly estimated.

Pilot study investigators often include a section on power and statistical analyses in grant proposals. Usually, the sections are not well-developed or justified. Often design features and measure reliability, 2 features affecting power are not considered. Most studies will require relatively large sample sizes to make inferential statements even for simple designs; complex designs and mediation and moderation require even larger samples. Thus, most pilot studies are limited in terms of estimation and inference. Some investigators have written acceptable analyses plans to be used in a future, larger study, and propose to test algorithms, software and produce results in an exploratory manner. This may be acceptable if the intent is to test the analytic procedures. If a statistical plan is provided for a future larger study, it should be clearly indicated as such. Some investigators provide exploratory analyses, which is not advised because the results will not be trustworthy.

TABLE 1. Methods for Examining Feasibility of Implementing Interventions

Step	Definition/Indicators	Examples of Feasibility Data Sources: Quantitative	Examples of Feasibility Data Sources: Qualitative
Can interventionists be recruited, trained, and retained?			
Recruit interventionists	Job description, sources of interventionists, qualifications, difficulties recruiting[27]	Administrative data: # recruited	Administrative data: difficulties finding interventionists
Train interventionists	Develop standardized training protocol, training manual (+ accompanying participant program manual); conduct training sessions[27]	Administrative data: # starting and completing training; adherence to training sessions Observer ratings: training observation checklist Structured ratings by trainees: training quality	Semistructured (open-ended) interviews: usefulness of training, suggestions for improvement of training, satisfaction with training
Assess interventionist competence	Assess outcomes of training, posttraining knowledge, core skills[27,28]	Knowledge test after training; observer ratings of interventionist's skill acquisition; certification	
Retain interventionists	Interventionists stay to the end of the intervention[29]	Administrative data: # retained through intervention	Administrative data: reasons for loss of interventionists
Can interventionists deliver the intervention as intended (per protocol)?			
Treatment fidelity (NIH Behavior Change Consortium)	Intervention delivered with fidelity to protocol; protocol defined by program manual, study protocol, and training manual; difficulties delivering intervention; consistency of delivery across interventionists[27,30]	Structured fidelity ratings or intervention delivery checklists by observers (direct observation or audio/video taping of sessions)	Reasons for nonfidelity
Feedback to interventionists, and ongoing training during intervention	System for giving feedback to interventionists when fidelity ratings are low; re-train or provide technical assistance as needed; possibly modify intervention; ongoing training and recertification[31,32]	Administrative data: amount of feedback and provision of additional interventionist training	Description of intervention components requiring feedback or additional training
Feedback from interventionists	Design method to identify program components that are hard to implement, challenges delivering intervention, components needing more time to deliver[27,29]	Feedback surveys completed by participants and interventionists after each session	Debriefing or focus groups of interventionists during or after pilot study; open-ended queries about difficulties
Are the treatment conditions (intervention and control) acceptable to participants and interventionists?			
Acceptability—before intervention	Formative research before designing intervention to determine needs of target population, how intervention could address needs[33]	Structured interviews/surveys	Focus groups, semi-structured interviews
Acceptability—after intervention	Postintervention debriefing about program overall and specific components: usefulness, ease of use, burden, comprehensibility, most/least helpful components, met expectations, suggestions for improvement, intent to continue behaviors[15,29,30,33–35]	Structured ratings	Focus groups, semistructured interviews with open-ended questions
Will participants adhere to and engage in the intervention components?			
Adherence/receipt of intervention	Track participant attendance/adherence and reasons for nonadherence; identify minimum adherence rate[15,29,34]	Tracking data: % completing intervention, # sessions completed	Tracking data: reasons for nonadherence (related or not to intervention)
Engagement in intervention	Track participant completion of each component, level of engagement; have interventionists rate skills mastery; open-ended reports of difficulties engaging[27,33–35]	% completing homework, practicing exercises, mastering skills; understanding material at each session	Open-ended queries about ability to engage in each component
Retention	Track dropout/retention[35]	Number of dropouts, number completing final assessment	Reasons for dropout (related or not to intervention)

What Types of Statistical Analyses Should be Proposed for Pilot Studies?

Descriptive statistics may be examined. For example, the mean and SD for continuous measures, and the frequency and percentage for categorical measures can be calculated overall and by subgroups. In large pilot trials, CIs may be provided to reflect the uncertainty of the main feasibility outcome by groups.

It may be possible to ascertain the minimally important difference (MID), to power a future trial.[42] For larger pilot trials, preparatory to large multisite studies, the variance of the primary outcome measure might be useful to determine

the standardized effect size. The MID does not account for the variance estimate required to calculate effect size.

Specifying the MID

Are There Available Estimates of MID?

While it is recognized that a MID cannot be generated using pilot data, such a specification based on earlier research may be important in planning a larger study. Methods for determining MIDs and treatment response have been reviewed.[43,44] The MID is "the average change in the domain of interest on the target measure among the subgroup of people deemed to change a minimal (but important) amount according to an 'anchor'."[45] Estimating the MID is a special case of examining responsiveness to change: the ability of a measure to reflect underlying change, for example, in health (clinical status), intervening health events, interventions of known or expected efficacy, and retrospective reports of change by patients or providers. In estimating the MID, the best anchors (retrospective measure of change and clinical parameters) are ones that identify those who have changed but not too much. Clinical input may be useful to identify the subset of people who have experienced minimal change.[6,46]

Variance Estimates

Are There Estimates of the Variances of Outcomes in Study Arms/Subgroups?

Variance estimates have an important impact on future power calculations. One could use the observed variance to form a range of estimates around that value in sensitivity analyses, and check if variances are similar to those of other studies using the same measures. The CI around that estimate should be calculated, rather than just the point estimate. However, values derived from small pilot studies may change with larger sample sizes and may be inaccurate. Thus, this estimation will only apply to large pilot studies.

CI

How Large Will CI be for Process Outcomes?

Although we advise against calculating effect sizes for efficacy outcomes, and caution about calculating feasibility outcomes involving proportions, information on CIs is included below because there are specialized pilot studies that are designed to be large enough to accurately estimate these indices. In addition, it is instructive to show how wide the CI could be if used to examine group differences in feasibility indices or outcomes. CIs are presented for feasibility process outcomes such as recruitment, adherence and retention rates, and for correlations of the outcomes before and after an intervention. In general, point estimates will not be accurate. There are several rules of thumb.[37,47] Leon et al[7] provide examples of how wide CIs will be with small samples.

Examples of CI Estimation for Process Outcomes

The 95% Clopper Pearson Exact CI for one proportion and Wald Method with Continuity Correction CI for differences in 2 proportions were calculated under various scenarios. Setting the α level at 0.05, the limits for the 95% CI for 1 proportion are given by Leemis and Trivedi,[48] and the Wald Method CI for the difference in two proportions by Fleiss et al.[49]

$$P_L = \left(1 + \frac{n - n_1 + 1}{n_1 \; F(\alpha/2,\; 2n_1,\; 2(n - n_1 + 1))}\right)^{-1}$$

$$P_U = \left(1 + \frac{n - n_1}{(n_1 + 1) \; F(1 - \alpha/2,\; 2(n_1 + 1),\; 2(n - n_1))}\right)^{-1}$$

where n is the total sample size, n_1 is the number of events. $F(\alpha/2, b, c)$ is the $(\alpha/2)$th percentile of the F distribution with b and c degrees of freedom.

As shown in Appendix Table 1 (Supplemental Digital Content 1, http://links.lww.com/MLR/C359), the 95% CI for a single proportion of 0.1 with a total sample size of 30 is (0.021, 0.265) with width of 0.244. The width is narrower with increased sample size, but it is relatively large (0.185) even with sample sizes of 50.

For the difference between 2 proportions (0.2 vs. 0.1), when the sample size per group is 10, the 95% CI is (−0.310, 0.510) and the width is 0.820 (Appendix Table 2, Supplemental Digital Content 1, http://links.lww.com/MLR/C359). When the group sample size is 30, the width is 0.425. Even with 50 per group, the CI width is relatively large (0.317).

The tables and figures provide other examples. As shown in Table 2, Appendix Figure 1 (Supplemental Digital Content 1, http://links.lww.com/MLR/C359) and Appendix Table 1 (Supplemental Digital Content 1, http://links.lww.com/MLR/C359), the minimum width for a CI for a single proportion is large for sample sizes <70. Table 3, Appendix Figure 2 (Supplemental Digital Content 1, http://links.lww.com/MLR/C359) and Appendix Table 2 (Supplemental Digital Content 1, http://links.lww.com/MLR/C359) show that if one wished to estimate the difference in retention rates with accuracy, a sample size of at least 50 per group would be required.

Correlations of the Outcomes Before and After the Intervention

Table 4 shows the formulas and minimum and maximum length for the 95% CI for the Pearson correlation coefficient from 0.100 to 0.900. As shown in Appendix Table 3

TABLE 2. Minimum and Maximum Length for 95% Clopper Pearson Exact Confidence Intervals for a Single Proportion

Sample Size (n)	Minimum	Maximum
10	0.44	0.63
20	0.30	0.46
30	0.24	0.37
40	0.21	0.32
50	0.18	0.29
60	0.17	0.26
70	0.15	0.24
80	0.14	0.23
90	0.13	0.21

The minimum and maximum values for the confidence interval (CI) width were computed for proportions ranging from 0.1 to 0.9 by 0.1. The maximum width of a CI for a single proportion can be as large as 0.37 for a sample size of 30. For a given sample size, the 95% CI is widest for a proportion of 0.5 and narrowest when proportions are further away from 0.5. For example, when the proportion is 0.5, the maximum is 0.37 for n of 30; the minimum length is 0.24 when the proportion is 0.10 or 0.90.

TABLE 3. Minimum and Maximum Length for 95% Confidence Intervals for a Difference in 2 Proportions

Sample Size/Group (n)	Minimum	Maximum
10	0.82	1.07
20	0.54	0.71
30	0.42	0.57
40	0.36	0.48
50	0.32	0.43
60	0.29	0.39
70	0.26	0.36
80	0.24	0.33
90	0.23	0.31

The Wald method with continuity correction was used to calculate 95% confidence intervals (CI) for the difference (d) in 2 proportions (p2−p1 = d, set p1 = 0.1, 0.2, 0.3, 0.4, d = 0.1, 0.2, 0.3, then p2 = 0.2, 0.3, 0.4, 0.5, 0.6, 0.7 based on the value of d). The proportions are selected based on clinically relevant estimates and their differences. Setting p1 = 0.9, 0.8, 0.7, 0.6, given the same d = p1−p2 and corresponding p2 = 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, will yield the same estimates of the width of CI (differing only in the label of the events). The maximum width of a CI for a difference in 2 proportions can be as large as 0.57 for a group sample size of 30.

For example, given n = 30, the maximum width occurs when p2 = 0.55 and p1 = 0.45 and the minimum width occurs when p2 = 0.2 and p1 = 0.1.

Note also that in this example, p1 and p2 were restricted to the less extreme values indicated above. If p1 and p2 are not limited, and any 2 proportions are selected, the maximum values occur when p1 and p2 are close to 0.5 and within the range of proportions we considered; thus the value is still very close to the numbers in the table. If we consider more extreme proportions close to 0 and 1 then the Wald method of calculating confidence intervals for their difference can underestimate the width of the interval. For example, for n = 30, the maximum occurs when p1 and p2 are very close to 0.5; for p1 = 0.5001 and p2 = 0.4999, the width is 0.5727. The minimum occurs when p1 and p2 are very close to 0 or 1; for p1 = 0.0001 and p2 = 0.0002 (or p1 = 0.9999 and p2 = 0.9998), the width is 0.0791.

Detailed values are provided in Appendix Table 2 (Supplemental Digital Content 1, http://links.lww.com/MLR/C359).

(Supplemental Digital Content 1, http://links.lww.com/MLR/C359), the 95% CI for a correlation coefficient of 0.500 with a total sample size of 30 is (0.170, 0.729), the width is 0.559. When the sample size is 50, the width is 0.426. What is obvious from Table 4, Appendix Table 3 (Supplemental Digital Content 1, http://links.lww.com/MLR/C359) and Figure 3 (Supplemental Digital Content 1, http://links.lww.com/MLR/C359) is that with sample sizes below 100, one cannot estimate a correlation coefficient with accuracy except for conditions with a high correlation of 0.900 and sample size over 50.

TABLE 4. Minimum and Maximum Length for 95% Confidence Interval (CI) for Pearson Correlation Coefficient (0.1–0.9 by 0.1)

Sample Size (n)	Minimum	Maximum
10	0.35	1.25
20	0.20	0.88
30	0.15	0.71
40	0.13	0.62
50	0.11	0.55
60	0.10	0.50
70	0.09	0.47
80	0.09	0.44
90	0.08	0.41

The 95% CI for the correlation coefficient was obtained by using the Fisher Z transformation.[3] First, compute a 95% CI for the parameter $\frac{1}{2}\ln\frac{1+\rho}{1-\rho}$ using the formula $\frac{1}{2}\ln\frac{1+r}{1-r} \pm \frac{1.96}{\sqrt{n-3}}$, where r is the sample correlation coefficient and n is the sample size.

Denote the limits for the 95% CI for this interval as $(L_z,\ U_z)$. Then the limits of the 95% CI for the original scale $(L_\rho,\ U_\rho)$ can be calculated by using the conversion formulas below:

$$L_\rho = \frac{e^{2L_z}-1}{e^{2L_z}+1} \text{ and } U_\rho = \frac{e^{2U_z}-1}{e^{2U_z}+1}.$$

Problems With Use of Nonparametric Statistics

Are Nonparametric Statistics a Rescue Method for Small Pilot Studies?

Some investigators believe incorrectly that they may use nonparametric tests to get around the problem of poor estimation using parametric tests. Parametric tests rely on distributional assumptions; for example, the normality assumption is assumed for a 2-sample t test comparing the means between 2 independent groups when the population variance is unknown. If the normality assumption is violated, a nonparametric test such as the Wilcoxon rank-sum test is often used; one important assumption is equality of population variances. Pilot studies are typically conducted with small sample sizes, and tests of normality are not reliable due to either lack of power to detect non-normality or small sample-induced non-normality. Nonequal variances may be observed, and the 2-sample t test with Satterthwaite's approximation of the degrees of freedom is robust, except for severe deviation from normality. Although the nonparametric test has higher power if the true underlying distribution is far from normal given that other assumptions are met, it typically has lower statistical power than the parametric test if the underlying distribution is truly or close to normal. Unless there is strong evidence of the violation of normality based on the given data (with a reasonable sample size) and/or established knowledge of the underlying distribution, the parametric test is generally preferred. Non-parametric tests are not free of assumptions and not a rescue method, nor a substitution for parametric tests with small sample sizes.

Evaluation of Randomization Algorithms and Specification of Design Features and MIDs

The preceding presentation provided caveats regarding generating effect sizes, calculating power, estimating CI, and use of nonparametric statistics. Below is a discussion of statistical or design factors that may be examined in pilot studies.

Randomization Algorithm

Is the Randomization Algorithm Working Correctly?

One can check procedures and protocol for randomization and whether the correct group assignment was made after randomization. Small sample sizes can result in imbalance between arms or within subgroups that cannot be detected with pilot data or early on in studies. Therefore, examination of balance between groups does not inform about randomization procedure performance.

Dose and Separation

Is There Separation between Groups in Terms of Dose Delivered?

Does the dose need adjustment? Is there a difference between groups in program delivery? For example, in a study of behavioral interventions of diet and exercise changes to reduce blood pressure, did the usual care group members also change their diets or increase exercise, thus reducing the potential effects of the study? Group separation on intervention variables

may be examined in studies that have an indicator of whether the intervention is affecting the targeted index, for example, determining if blood levels of a drug are actually different between usual care and intervention groups.

Design Effect Estimates

Are There Estimates of the Design Effects?

The cluster size and intracluster correlation coefficient can affect power. These may be difficult to estimate with small pilot studies; however, one can usually get some idea about the cluster size from other information, which can be used in planning a larger study. For example, in a study of a pain intervention, patients will be clustered within physicians/practices. Investigators can determine in advance about how many patients are cared for within a practice that may be sampled.

DISCUSSION

A goal of this commentary was to provide guidelines for testing multiple components of a pilot study,[2] a likely strategy for early and mid-stage investigators conducting studies as part of a training grant or center. Guidelines on recruitment feasibility are also available,[50] including issues faced when studying disparities populations.

Estimation issues for group differences in outcome measures as well as process indicators, for example, completion or adherence rates, were discussed, and it was demonstrated that both will have large CIs with small sample sizes. If a goal of a pilot study is to estimate group differences, this objective should be stated clearly, and the requisite sample sizes specified, often as large as 70–100 per group. A typical pilot study with 30 respondents per group is too small to provide reasonable power or precision. It has thus been argued that only counts, means, and percentages of feasibility outcomes should be calculated and later compared with albeit subjective thresholds that are specified a priori, such as achieving a retention rate of at least 80%.

It has been suggested that indicators of feasibility should be stated in terms of "clear quantitative benchmarks" or progression criteria by which successful feasibility is judged. For example, NCCIH guidelines suggest adherence benchmarks such as "at least 70% of participants in each arm will attend at least 8 of 12 scheduled group sessions."[6] For testing the feasibility of methods to reach diverse populations these data may be used to modify the methods rather than as strict criteria for progression to a full-scale study. For example, some research has shown that a trial can be effective with fewer sessions as long as key sessions are attended.

CONCLUSIONS

Several indicators that can be examined in pilot feasibility studies include recruitment, retention, intervention fidelity, acceptability, adherence, and engagement. Additional indicators include randomization algorithms, capability to merge data, reliability of measures, inter-rater reliability of assessors, design features such as cluster sizes, and specification of an MID if one exists. As demonstrated in this commentary, most pilot studies should not be used to estimate effect sizes, provide power calculations for statistical tests or perform exploratory analyses of efficacy. It is hoped that these guidelines may be useful to those planning pilot/feasibility studies preparatory to a larger-scale study.

REFERENCES

1. Eldridge SM, Lancaster GA, Campbell MJ, et al. Defining feasibility and pilot studies in preparation for randomised controlled trials: development of a conceptual framework. *PLoS One*. 2016;11:e0150205.
2. Lancaster GA, Thabane L. Guidelines for reporting non-randomised pilot and feasibility studies. *Pilot Feasibility Stud*. 2019;5:114.
3. Kleinbaum DG, Kupper LL, Nizam A, et al. *Applied Regression Analysis and Other Multivariable Techniques*, 5th ed. Boston, MA: PWS-Kent; 2014.
4. Kraemer HC, Mintz J, Noda A, et al. Caution regarding the use of pilot studies to guide power calculations for study proposals. *Arch Gen Psychiatry*. 2006;63:484–489.
5. Kistin C, Silverstein M. Pilot studies: a critical but potentially misused component of interventional research. *JAMA*. 2015;314:1561–1562.
6. National Center for Complementary and Integrative Health. NCCIH Research Blog [Internet] 2017. 2020. Available at: https://www.nccih.nih.gov/grants/pilot-studies-common-uses-and-misuses. Accessed November 11, 2021.
7. Leon AC, Davis LL, Kraemer HC. The role and interpretation of pilot studies in clinical research. *J Psychiatr Res*. 2011;45:626–629.
8. Moore CG, Carter RE, Nietert PJ, et al. Recommendations for planning pilot studies in clinical and translational research. *Clin Transl Sci*. 2011;4:332–337.
9. Eldridge SM, Chan CL, Campbell MJ, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. *Pilot Feasibility Stud*. 2016;2:64.
10. Lancaster GA, Dodd S, Williamson PR. Design and analysis of pilot studies: recommendations for good practice. *J Eval Clin Pract*. 2004;10:307–312.
11. Thabane L, Lancaster G. A guide to the reporting of protocols of pilot and feasibility trials. *Pilot Feasibility Stud*. 2019;5:37.
12. Samayoa C, Santoyo-Olsson J, Escalera C, et al. Participant-centered strategies for overcoming barriers to biospecimen collection among Spanish-speaking Latina breast cancer survivors. *Cancer Epidemiol Biomarkers Prev*. 2020;29:606–615.
13. Stewart AL. The reliability and validity of self-reported weight and height. *J Chronic Dis*. 1982;35:295–309.
14. Cunningham WE, Rana HM, Shapiro MF, et al. Reliability and validity of self-report CD4 counts-in persons hospitalized with HIV disease. *J Clin Epidemiol*. 1997;50:829–835.
15. Voils CI, Adler R, Strawbridge E, et al. Early-phase study of a telephone-based intervention to reduce weight regain among bariatric surgery patients. *Health Psychol*. 2020;39:391–402.
16. Scott J, Heavey E, Waring J, et al. Implementing a survey for patients to provide safety experience feedback following a care transition: a feasibility study. *BMC Health Serv Res*. 2019;19:613.
17. Stewart AL, Nápoles-Springer A. Health-related quality-of-life assessments in diverse population groups in the United States. *Med Care*. 2000;38(suppl):II102–II124.
18. Paz SH, Jones L, Calderon JL, et al. Readability and comprehension of the Geriatric Depression Scale and PROMIS® physical function items in older African Americans and Latinos. *Patient*. 2017;10:117–131.
19. Paz SH, Spritzer KL, Morales LS, et al. Evaluation of the Patient-Reported Outcomes Information System (PROMIS®) Spanish-language physical functioning items. *Qual Life Res*. 2013;22:1819–1830.
20. Mokkink LB, Terwee CB, Knol DL, et al. Protocol of the COSMIN study: COnsensus-based Standards for the selection of health Measurement INstruments. *BMC Med Res Methodol*. 2006;6:2.
21. Cella D, Yount S, Rothrock N, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS®): progress of an NIH Roadmap cooperative group during its first two years. *Med Care*. 2007;45(suppl 1):S3–S11.
22. Reeve B, Teresi JA. Overview to the two-part series: measurement equivalence of the Patient-reported Outcomes Measurement Information System (PROMIS®) short forms. *Psychol Test Assess Model*. 2016;58:31–35.
23. Jones RN, Tommet D, Ramirez M, et al. Differential item functioning in Patient-reported Outcomes Measurement Information System (PROMIS®)

physical functioning short forms. *Psychol Test Assess Model*. 2016;58: 371–402.
24. Teresi JA, Stewart AL, Morales LS, et al. Measurement in a multi-ethnic society: overview to the special issue. *Med Care*. 2006;44(suppl 3): S3–S4.
25. Teresi JA, Stewart AL, Stahl SM. Fifteen years of progress in measurement and methods at the Resource Centers for Minority Aging Research. *J Aging Health*. 2012;24:985–991.
26. Nápoles AM, Stewart AL. Transcreation: an implementation science framework for community-engaged behavioral interventions to reduce health disparities. *BMC Health Serv Res*. 2018;18:710.
27. Shafayat A, Csipke E, Bradshaw L, et al. Promoting Independence in Dementia (PRIDE): protocol for a feasibility randomised controlled trial. *Trials*. 2019;20:709.
28. Sternberg RM, Nápoles AM, Gregorich S, et al. Mentes Positivas en Accion: feasibility study of a promotor-delivered cognitive behavioral stress management program for low-income Spanish-speaking Latinas. *Health Equity*. 2019;3:155–161.
29. Griffin T, Sun Y, Sidhu M, et al. Healthy Dads, Healthy Kids UK, a weight management programme for fathers: feasibility RCT. *BMJ Open*. 2019;9:e033534.
30. Bellg AJ, Borrelli B, Resnick B, et al. Enhancing treatment fidelity in health behavior change studies: best practices and recommendations from the NIH Behavior Change Consortium. *Health Psychol*. 2004;23: 443–451.
31. Santoyo-Olsson J, Stewart AL, Samayoa C, et al. Translating a stress management intervention for rural Latina breast cancer survivors: the Nuevo Amanecer-II. *PLoS One*. 2019;14:e0224068.
32. Nápoles AM, Santoyo-Olsson J, Ortiz C, et al. Randomized controlled trial of Nuevo Amanecer: a peer-delivered stress management intervention for Spanish-speaking Latinas with breast cancer. *Clin Trials*. 2014;11:230–238.
33. Bowen DJ, Kreuter M, Spring B, et al. How we design feasibility studies. *Am J Prev Med*. 2009;36:452–457.
34. Nápoles AM, Santoyo-Olsson J, Stewart AL, et al. Evaluating the implementation of a translational peer-delivered stress management program for Spanish-speaking Latina breast cancer survivors. *J Cancer Educ*. 2018;33:875–884.
35. Nápoles AM, Santoyo-Olsson J, Chacon L, et al. Feasibility of a mobile phone app and telephone coaching survivorship care planning program among Spanish-speaking breast cancer survivors. *JMIR Cancer*. 2019;5: e13543.
36. Whitehead AL, Julious SA, Cooper CL, et al. Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable. *Stat Methods Med Res*. 2016;25:1057–1073.
37. Browne RH. On the use of a pilot sample for sample size determination. *Stat Med*. 1995;14:1933–1940.
38. Julious SA. Sample size of 12 per group rule of thumb for a pilot study. *Pharmaceutical Statistics*. 2005;4:287–291.
39. Cocks K, Torgerson DJ. Sample size calculations for pilot randomized trials: a confidence interval approach. *J Clin Epidemiol*. 2013;66:197–201.
40. Sim J, Lewis M. The size of a pilot study for a clinical trial should be calculated in relation to considerations of precision and efficiency. *J Clin Epidemiol*. 2012;65:301–308.
41. Stallard N. Optimal sample sizes for phase II clinical trials and pilot studies. *Stat Med*. 2012;31:1031–1042.
42. Keefe RS, Kraemer HC, Epstein RS, et al. Defining a clinically meaningful effect for the design and interpretation of randomized controlled trials. *Innov Clin Neurosci*. 2013;10(suppl A):4S–19S.
43. Hays RD, Spritzer KL, Reise SP. Using item response theory to identify responders to treatment: examples with the Patient-Reported Outcomes Measurement Information System (PROMIS®) physical function scale and emotional distress composite. *Psychometrika*. 2021;86:781–792.
44. Revicki D, Hays RD, Cella D, et al. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol*. 2008;61:102–109.
45. McLeod LD, Coon CD, Martin SA, et al. Interpreting patient-reported outcome results: US FDA guidance and emerging methods. *Expert Rev Pharmacoecon Outcomes Res*. 2011;11:163–169.
46. Hays RD, Farivar SS, Liu H. Approaches and recommendations for estimating minimally important differences for health-related quality of life measures. *COPD*. 2005;2:63–67.
47. Lehr R. Sixteen S-squared over D-squared: a relation for crude sample size estimates. *Stat Med*. 1992;11:1099–1102.
48. Leemis LM, Trivedi KS. A comparison of approximate interval estimators for the Bernoulli Parameter. *The American Statistician*. 1996;50:63–68.
49. Fleiss JL, Levin B, Paik MC. *Stastical Methods for Rates and Proportions*, 3rd ed. Hoboken, NJ: John Wiley & Sons Inc; 2003.
50. Stewart AL, Nápoles AM, Piawah S, et al. Guidelines for evaluating the feasibility of recruitment in pilot studies of diverse populations: an overlooked but important component. *Ethn Dis*. 2020;30(suppl 2): 745–754.