Published in final edited form as:
J Psychiatr Res. 2011 May ; 45(5): 626–629. doi:10.1016/j.jpsychires.2010.10.008.

The Role and Interpretation of Pilot Studies in Clinical Research

Andrew C. Leon, Ph.D.[1], **Lori L. Davis, M.D.**[2], and **Helena C. Kraemer, Ph.D.**[3]

[1] Weill Cornell Medical College, Department of Psychiatry, New York, NY

[2] University of Alabama School of Medicine, Birmingham, AL VA Medical Center, Tuscaloosa, AL

[3] Stanford University, Department of Psychiatry and Behavioral Sciences, Stanford, CA

Abstract

Pilot studies represent a fundamental phase of the research process. The purpose of conducting a pilot study is to examine the feasibility of an approach that is intended to be used in a larger scale study. The roles and limitations of pilot studies are described here using a clinical trial as an example. A pilot study can be used to evaluate the feasibility of recruitment, randomization, retention, assessment procedures, new methods, and implementation of the novel intervention.

A pilot study is not a hypothesis testing study. Safety, efficacy and effectiveness are not evaluated in a pilot. Contrary to tradition, a pilot study does not provide a meaningful effect size estimate for planning subsequent studies due to the imprecision inherent in data from small samples. Feasibility results do not necessarily generalize beyond the inclusion and exclusion criteria of the pilot design.

A pilot study is a requisite initial step in exploring a novel intervention or an innovative application of an intervention. Pilot results can inform feasibility and identify modifications needed in the design of a larger, ensuing hypothesis testing study. Investigators should be forthright in stating these objectives of a pilot study. Grant reviewers and other stakeholders should expect no more.

Keywords

Pilot study; feasibility study; proof of concept study

INTRODUCTION

Over the past several years, research funding agencies have requested applications for pilot studies that are typically limited to a shorter duration (one to three years) and a reduced budget using, for example, the NIH R34 funding mechanism (NIMH, 2010). Pilot studies play a key role in the development or refinement of new interventions, assessments, and other study procedures. Commonly, results from pilot studies are used to support more expensive and lengthier pivotal efficacy or effectiveness studies. Importantly, investigators, grant reviewers, and other stakeholders need to be aware of the essential elements,

Corresponding Author: Andrew C. Leon, Ph.D., Weill Cornell Medical College, Department of Psychiatry, Box 140, 525 East 68th Street, New York, NY 10065, telephone: (212)746-3872, fax (212)746-8754, acleon@med.cornell.edu.

appropriate role, and exceptional strengths and limitations in the interpretation of pilot studies.

A pilot study is, "A small-scale test of the methods and procedures to be used on a larger scale …" (Porta, 2008). The fundamental purpose of conducting a pilot study is to examine the feasibility of an approach that is intended to ultimately be used in a larger scale study. This applies to all types of research studies. Here we use the randomized controlled clinical trial (RCT) for illustration. Prior to initiating a full scale RCT an investigator may choose to conduct a pilot study in order to evaluate the feasibility of recruitment, randomization, retention, assessment procedures, new methods, and/or implementation of the novel intervention. A pilot study, however, is not used for hypothesis testing. Instead it serves as an earlier-phase developmental function that will enhance the probability of success in the larger subsequent RCTs that are anticipated.

For purpose of contrast, a hypothesis testing clinical trial is designed to compare randomized treatment groups in order to draw an inference about efficacy/effectiveness and safety in the patient *population*, based on *sample* results. The primary goal in designing such a study is to minimize the bias in the estimate of the treatment effect. (Leon et al., 2006; Leon & Davis, 2009). That is, the trial is designed to ask the question, "Is the treatment efficacious, and if so, what is the magnitude of the effect?". Features of RCTs that help us achieve this goal are randomized group assignment, double-blinded assessments, and control or comparison groups.

This manuscript will focus on pilot studies, those used to shape some, but not all aspects of the design and implementation of hypothesis testing clinical trials. It is the feasibility results, not the efficacy or safety results, that inform subsequent trials. The objective of this manuscript is to elaborate on each of these points: efficacy, safety, and feasibility. We discuss both the design of a pilot study and the interpretation and application of pilot study results. What is discussed here applies to pilot studies, feasibility studies and proof of concept studies, terms that have been used somewhat interchangeably in the literature and henceforth are referred to here as "pilot studies".

WHAT A PILOT STUDY CAN DO: ASSESS FEASIBILITY

Pilot study results can guide in the design and implementation of larger scale efficacy studies. There are several aspects of RCT feasibility that are informed by conducting a pilot study. A pilot study can be used to evaluate the feasibility of recruitment, randomization, retention, assessment procedures, and implementation of the novel intervention and each of these can be quantified (Table 1). Study components that are deemed infeasible or unsatisfactory should be modified in the subsequent trial or removed altogether.

Rationale for a Control or Comparison Group in a Pilot

The inclusion of a control or comparator group in a full scale trial accounts for the passage of time, the increased attention received in the study, the expectation of a therapeutic intervention, and the psychological consequences of legitimized sick role.(Klerman, 1986) Nevertheless, an investigator might wonder what purpose a control group serves in a pilot if no inferential comparisons are to be conducted. Although not essential for many aspects of the study, inclusion of a control group allows for a more realistic examination of recruitment, randomization, implementation of interventions, blinded assessment procedures, and retention in blinded interventions. Each aspect of feasibility could be quite different from an uncontrolled study when intervention assignment is randomized and blinded, particularly if placebo is a distinct possibility. In an open label pilot that has no control group, participants are recruited to a known, albeit experimental, intervention with

no risk of receiving placebo. Assessments are conducted in an unblinded fashion. The implementation of only one intervention can be evaluated. Retention information is based on those receiving unblinded treatment. With these issues in mind, a pilot study can better address its goals if a control group is part of the design. A control group would also be particularly illuminating for psychotherapy or psychosocial interventions, whereby the control group's aspects and procedures are also tested for feasibility, consistency, and acceptability.

Good Clinical Practices

A pilot study provides opportunity to develop consistent practices to enhance data integrity and the protection of human subjects. These good clinical practices include the refinement of source documentation, informed consent procedures, data collection tools, regulatory reporting procedures, and monitoring/oversight procedures, especially when multiple sites and investigators are engaged in the study. A pilot study can be critical in research staff training and provide experiences that strengthen and confirm competencies and skills required for the investigation to be conducted with accuracy and precision.

SAMPLE SIZE DETERMINATION IN DESIGNING A PILOT STUDY

A pilot study is not a hypothesis testing study. Therefore, no inferential statistical tests should be proposed in a pilot study protocol. With no inferential statistical tests, a pilot study will not provide *p*-values. Power analyses are used to determine the sample size that is needed to provide adequate statistical power (typically 80% or 90%) to detect a clinically meaningful difference with the specified inferential statistical test. However, power analyses should not be presented in an application for a pilot study that does not propose inferential tests. A pilot sample size is instead based on the pragmatics of recruitment and the necessities for examining feasibility.

Pilot Data for a Pilot Study

Pilot studies are exploratory ventures. Pilot studies generate pilot data, their design need not be guided with the support of prior pilot data. It is quite reasonable and expected that a pilot study is proposed with no pilot or other preliminary data supporting the proposal and that its proposed sample size is based on pragmatics such as patient flow and budgetary constraints. This does not preclude the need for a theoretical rationale for the intervention or the methodology being proposed for a pilot study.

Are Pilot Data Included in the Larger Trial?

Pilot study data generally should not be combined with data from the subsequent larger scale study. This is because it is quite likely that the methods will be modified after the pilot, even if minimally. Such changes in protocol risk adding additional, perhaps unknown, source of variation. However, if a well-specified adaptive design were explicated prior to the start of a pilot study, and the risk of elevated type I error appropriately controlled, it is conceivable that data from before and after protocol changes could, in fact, be pooled. This is a rare exception.

EXCEEDING THE LIMITATIONS OF A PILOT STUDY: WHAT PILOT STUDIES CANNOT DO

Although a pilot study will undoubtedly incorporate relevant outcome measures and can serve a vital role in treatment development, it is not, and should not, be considered a preliminary test of the intervention hypothesis. There are two fundamental reasons that

hypothesis testing is not used in a pilot study: the limited state of knowledge about the methods or intervention in the patient population to be studied and the smaller proposed sample size.

Tolerability and Preliminary Safety

Only in an extreme, unfortunate case, where a death occurs or repeated serious adverse events surface, do pilot studies inform the safety of testing an intervention due to the small sample size. However, pilot studies provide an opportunity to implement and examine the feasibility of the adverse event reporting system. Nevertheless, if some safety concerns are detected in a pilot study group-specific rates (with 95% confidence intervals) should be reported for adverse events, treatment emergent adverse events and serious adverse events. When event rates are reported and no adverse event is observed for a particular category, the *rule of three* should be applied to estimate the upper bound of the 95% CI, where the upper bound is approximately 3/n. (Jovanovic & Levy, 1997; Jovanovic et al., 1997) For example, consider a study with *N=15* receiving medication and no suicidal ideation was reported for that group. Although the observed rate of suicidal ideation is 0%, the upper bound of the 95% confidence interval is 3/15 or 20%. This imprecision, seen in the wide confidence interval, underscores the limited value of safety data from a pilot study.

Pilot Study Effect Sizes and Sample Size Determination

There has been a venerable tradition of using pilot studies to estimate between group effect sizes that, in turn, are used to inform the design of subsequent larger scale hypothesis testing studies. Despite it widespread use, it has been argued that the tradition is ill-founded. (Friedman, Furberg and DeMets, 1998; Kraemer et al., 2006) Pilot study results should not be used for sample size determination due to the inherent imprecision in between treatment group effect size estimates from studies with small samples. Furthermore, pilot results that are presented to grant review committees tend to be selective, overly optimistic and, at times, misrepresentational.

The adverse consequences of using a pilot study effect size for sample size estimation correspond with the two errors of inferential testing: false positive results (Type I error) and false negative results (Type II error). If a pilot study effect size is unduly large (i.e., a false positive result), subsequent trials will be designed with an inadequate number of participants to provide the statistical power needed to detect clinically meaningful effects and that would lead to negative trials. If a pilot study effect size is unduly small (i.e., a false negative result), subsequent development of the intervention could very well be terminated – even if the intervention eventually would have proven to be effective. Unfortunately, a false negative result could preclude the opportunity to further examine its latent efficacy.

An essential challenge of therapeutic development is that the true population effect size is unknown at the time a pilot study is designed. It is this gap in knowledge that motivates much research. An enthusiastic investigator may well believe that a series of cases provides evidence of efficacy, but such data are observational and uncontrolled; realistically, they are seldom replicated in RCTs -- as seen years ago with reserpine (Kinross-Wright, 1955; Campden-Main, Wegielski, 1955; Goller 1960). A case series estimate tends to be steeped in optimism, particularly if an estimate of such magnitude is seldom, if ever, seen in full scale trials for psychiatric disorders. Nevertheless, it is not unusual for research grant applications and pilot study publications to convey such optimism, particularly when based on pilot data.

It is possible, but highly unlikely, that the between group effect size *(d)* from a pilot study sample will provide a reasonable estimate of the population effect size (Δ), but that cannot be known based on the pilot data. (It is the population effect size, not the sample effect size,

that an RCT is designed to detect with sufficient power.) This estimation problem has to do with the precision of *d* and its relation to sample size. Estimates become more precise with larger sample sizes. Therefore, estimates of effect sizes should not be a specific aim of a pilot proposal. This applies to effects sizes for any type of outcome, be it a severity rating, a response status, or survival status. The reasoning for this is as follows.

Hypothetical Example

Precision is embodied in the confidence interval (CI) around *d*. By definition, there is a 95% probability that Δ falls within the range of the 95% CI. Consider some examples, initially a hypothetical example. For simplicity, assume that two groups (e.g., active and placebo) of equal size ($n_i = n_j$, where the total sample size is $N=2n_i$) will be compared on a normally distributed outcome measure for which the groups have a common variance. The between group effect size, Cohen's *d*, is estimated as: $d=\frac{\overline{X}_1 - \overline{X}_2}{s}$. With equal sample sizes, the 95% CI for *d* is approximately: $d +/ - (4/\sqrt{N})$. (Note that the 4 in the numerator of the final term is derived from $2*t_{(N-2,\ \alpha/2)}$.) For example, if the sample effect size is *d=.50* (i.e., the two groups differ by one-half standard deviation unit) and there are 18 participants per group, the 95% CI is 0.50+/− 0.67: $-0.17 \leq \Delta \leq 1.17$ (i.e., $.50+/ - 4/\sqrt{36}$). This intervals denotes that the true effect of active relative to placebo (Δ) is somewhere between is *slightly detrimental* (−0.17) to *tremendously beneficial* (1.17). The corresponding estimates of sample size/group range from as many as 576 to as few as 12. Hence, with imprecision comes a vast disparity in sample size estimates and, if sample size determination for a subsequent study is based on an imprecise estimate, there is an enormous risk of underpowered or overpowered design. In other words, the efficacy data from a pilot study of this size are uninformative. Many pilot studies have far fewer than 18 participants/group and therefore even greater imprecision. We learn little if anything about the efficacy of an intervention with data from a small sample; yet, as discussed earlier, a great deal can be learned from a pilot study.

An Alternative to Using Pilot Data for Sample Size Determination

An alternative approach is to base sample size estimates for a full scale RCT on what is deemed a *clinically meaningful effect*. For example, the investigator must use clinical experience to describe a *clinically meaningful difference* on the primary outcome, in the case of MDD trials, the HAMD. How many HAMD units represent a meaningful between treatment group difference? Assume that the pre-post difference on the HAMD total has an *sd=6.0*. Then *d=.20* represents 1.2 units of HAMD change, *d=.40* represents 2.4 units of HAMD change, and *d=.50* represents 3.0 units of HAMD change. The respective sample sizes needed per group for 80% power with two-tailed t-test (alpha=.05) are: 393, 100, and 64. (N/group ≈ $16/d^2$; Lehr, 1992) The clinical interpretation of HAMD change of 1.2 to 3.0 would drive the choice among possible sample sizes in planning a study. Ideally, a hypothesis testing study should be designed to detect the smallest difference that is generally agreed to be *clinically meaningful*.

DISCUSSION

The primary role of a pilot study is to examine the feasibility of a research endeavor. For instance, feasibility of recruitment, randomization, intervention implementation, blinded assessment procedures, and retention can all be examined. Investigators should be forthright in stating these objectives of a pilot study and bravely accept the limitations of a pilot study. Grant reviewers should expect no more.

The choice of the intervention for a pilot study should be based on theory, mechanism of action, a case series, or animal studies that justify a rationale for therapeutic effect. Well-conceived and implemented pilot studies will reduce the risk of several problems that are commonly faced in clinical trials. These include the inability to recruit the proposed sample size and a corresponding reduction in statistical power, excessive attrition due to intolerable procedures or interventions, and the need to modify a protocol midway through a trial. As a consequence, pilot studies can reduce the proportion of failed trials and allow research funds to be spent on projects for which feasibility has been demonstrated and quantified.

Despite the convention, pilot studies do not provide useful information regarding the population effect size because the estimates are quite crude owing to the small sample sizes. Basing a decision to proceed or terminate evaluation of a particular intervention on pilot data is perilous because there is a very good chance that the decision will be derived from false positive or false negative results. In lieu of using pilot results, sample size determination should be derived based on that required for sufficient statistical power to detect a clinically meaningful treatment effect. The definition of clinically meaningful is not entirely empirically-based, but instead requires input from clinicians who treat the patient population of interest and perhaps from patients with the disorder.

By the very nature of pilot studies, there are critical limitations to their role and interpretation. For example, a pilot study is not a hypothesis testing study and therefore safety and efficacy are not evaluated. Further, a study can only examine feasibility of the patient type included in the study. The feasibility results do not necessarily generalize beyond the inclusion and exclusion criteria of the pilot.

In summary, pilot studies are a necessary first step in exploring novel interventions and novel applications of interventions – whether in a new patient population or with a novel delivery system (e.g., transdermal patch). Pilot results inform feasibility, which in turn, is instructive in that it points to modifications needed in the planning and design of a larger efficacy trial.

References

Campden-Main BC, Wegielski Z. The control of deviant behavior in chronically disturbed psychotic patients by the oral administration of reserpine. Ann N Y Acad Sci. Apr 15; 1955 61(1):117–122. [PubMed: 14377279]

Friedman, LM.; Furberg, CD.; DeMets, DL. Fundamentals of Clinical Trials. 3. New York: Springer; 1998.

Goller ES. A controlled trial of reserpine in chronic schizophrenia. J Ment Sci Oct. 1960; 106:1408–1412.

Jovanovic BD, Levy PS. A look at the rule of three. The American Statistician. 1997; 57:137–139.

Jovanovic BD, Zalenski RJ. Safety evaluation and confidence intervals when the number of observed events is small or zero. Annals of Emergency Medicine. 1997; 30:301–6. [PubMed: 9287891]

Kinross-Wright V. Chlorpromazine and reserpine in the treatment of psychoses. Ann N Y Acad Sci. Apr 15; 1955 61(1):174–182. [PubMed: 14377285]

Klerman GL. Scientific and ethical considerations in the use of placebo controls in clinical trials in psychopharmacology. Psychopharmacology Bulletin. 1986; 22:25–29. [PubMed: 3726071]

Kraemer HC, Mintz J, Noda A, Tinklenberg J, Yesavge JA. Caution regarding the use of pilot studies to guide power calculations for study proposals. Archives of General Psychiatry. 2006; 63:484–489. [PubMed: 16651505]

Lehr R. Sixteen s-squared over d-squared: A relation for crude sample size estimates. Statistics in Medicine. 1992; 11:1099–1102. [PubMed: 1496197]

Leon AC, Davis LL. Enhancing Clinical Trial Design of Interventions for Posttraumatic Stress Disorder. Journal of Traumatic Stress. 2009; 22:603–611. [PubMed: 19902462]

Leon AC, Mallinckrodt CH, Chuang-Stein C, Archibald DG, Archer GE, Chartier K. Attrition in randomized controlled clinical trials: Methodological issues in psychopharmacology. Biological Psychiatry. 2006; 59:1001–1005. [PubMed: 16503329]

National Institute of Mental Health. Pilot Intervention and Services Research Grants (R34). [Accessed October 4, 2010]. http://grants.nih.gov/grants/guide/pa-files/PAR-09-173.html

Porta, M. A Dictionary of Epidemiology. 5. Oxford: Oxford University Press; 2008.

Table 1

Aspects of Feasibility that Can be Examined with a Pilot Study

Study Component	Feasibility Quantification
Screening	Number screened per month
Recruitment	Number enrolled per month
Randomization	Proportion of screen eligible who enroll
Retention	Treatment-specific retention rates
Treatment adherence	Rates of adherence to protocol for each intervention
Treatment fidelity	Fidelity rates per unit monitored
Assessment process	Proportion of planned ratings that are completed; duration of assessment visit