# Feasibility and Pilot Studies

Kenneth E. Freedland, PhD
Professor of Psychiatry and Psychology
Washington University School of Medicine
St. Louis, Missouri, USA

*Seminar on*
The Nuts and Bolts of Behavioral Intervention Development
Susan Czajkowski, PhD, Chair
37th Annual Meeting & Scientific Sessions of the Society of Behavioral Medicine
Washington, DC
March 30, 2016

# Disclosure

- Current research funding from NHLBI.
- No significant financial interests to disclose.

# Controversy and Confusion

- Feasibility & pilot studies may seem like no-brainers: easy to design, conduct, analyze, and interpret.
- But many investigators and reviewers are confused about feasibility and pilot studies, and frustrated by frequent disagreements about them.
- Statisticians and other expert methodologists disagree on key points; they don't even agree on terminology.
- I'll give you my own perspective, with the proviso that it isn't based on a solid consensus .

# Feasibility Studies

- The main purpose of a feasibility study, in my opinion, is to determine whether it's possible to successfully conduct a larger study.
  - E.g., assume that you've developed a novel intervention for cancer survivors.
  - You want to test it in an RCT.
  - Reviewers will want to know (and so should you) whether the chances are good that you'll be able to successfully conduct this trial.

# Feasibility Studies

- Note that the question is *not* whether the intervention will turn out to be efficacious.
- It's about whether the trial will be successfully conducted, no matter what the results of the trial turn out to be.
- What sorts of questions might reviewers have about the chances of success?

# Feasibility Studies

- Feasibility questions (partial list):
  - Will you be able to recruit enough patients?
  - Are these patients willing to be randomized to your intervention vs. some other condition?
  - How many are likely to be nonadherent to the study protocol or to drop out?
  - Are the therapists able to follow the protocol?
  - Are the measures too burdensome?
  - Etc.

# Feasibility Studies

- Feasibility studies address these sorts of questions.
- But they aren't the only source of information that reviewers rely upon.
  - E.g., your biosketch and the preliminary studies of your RCT proposal will tell them how much experience you've had with trials like this one.
  - E.g., your Facilities & Other Resources page will tell them whether the setting is conducive.
  - E.g., your budget will constrain the size of the study.
- If you've been doing similar trials for years, you might not even need to conduct a feasibility study.

# Feasibility Studies

- What sort of statistical analysis plan should you include in a feasibility study?
- If the larger study is going to be *really* large – e.g., a large, multicenter trial with thousands of patients, to be conducted at quite a few sites around the country – then *inferential statistics* should probably be part of the plan.
  - E.g., if the average recruitment is X% in a small feasibility study run at 2 sites, what are the confidence intervals around X%?
  - This provides a range of plausible recruitment rates for the larger multicenter trial.

# Feasibility Studies

- But what if you're simply hoping to conduct a plain-old single-site study with a relatively modest sample size (e.g., $50 \leq n \leq 200$).
- What would inferential statistics (confidence intervals, p values, t-tests, etc.) tell you about the feasibility of the RCT you'd like to conduct?
  - In my opinion, they won't tell you much more about what to expect in the main trial than you'd learn from ordinary descriptive statistics like percentages.

# Feasibility Studies

- Small feasibility studies are about *you*.
  - How likely is it that *you'll* be able to conduct the trial in a rigorous & successful manner,
  - In *your* setting, with *your* procedures, patients, *your* intervention, measures, etc.?
- You're only trying to generalize to yourself and to your anticipated RCT, not to the world at large.
- Inferential statistics won't help you with that.*

**Caution: Some reviewers may disagree with this.*

# Feasibility Studies

- What should you do instead?
- Set criteria for success, based on the needs of the anticipated RCT, and collect data to see if you can meet these criteria.
  - E.g., you expect to need about n=200 patients for the anticipated RCT.
  - The RCT will have a one-year enrollment phase.
  - You're recruiting from a single clinic that sees 5,000 patients a year.
  - Your feasibility study enrollment rate has to exceed 4%.
  - If it doesn't, reviewers will doubt your ability to enroll 200 patients in a year for the RCT.

# Feasibility Studies

- You don't need to know the confidence interval around that 4% number, in the sense that you're not trying to generalize from it.
- BUT CIs may help you answer the question of how big should the feasibility study be.
  - Smaller CIs make reviewers more confident.
  - 4% of 10 patients: huge confidence interval.
  - 4% of 20 patients: smaller confidence interval.
  - 4% of 200 patients: feasibility study is way to big.

# Feasibility Studies

- A common and serious omission in feasibility proposals:
  - Failure to describe the anticipated RCT.
- "Feasibility of what?" is not a question you should leave unanswered for the reviewers.
  - Provide the rationale and enough detail about the anticipated RCT so that they know what you're hoping to do.
  - They'll judge your feasibility proposal both on its own merits and on their reaction to whether your trial would be worth conducting, assuming that it turns out to be feasible.

# Feasibility Studies

- RCTs are often full of surprises, and few of the surprises are loads of fun.
- You're probably going to encounter some unexpected problems in an RCT, even if your feasibility study results were great.
- So, don't let good feasibility data lull you into complacency; be prepared to deal with whatever challenges come up in the RCT.

# Pilot Studies

- The literature on pilot study methodology is a bewildering maze of contradictions.
- Reasons:
  - Disagreements among experts about methods.
  - Differences between drug & behavioral research.
  - Different ideas in different places (e.g., UK, US).
  - Different kinds of pilot data may be needed for small vs. large RCTs.

# Pilot Studies

- Some experts assert that a pilot study should be a miniature version of the main trial.
  - Same procedures but smaller sample.
- If the main trial is a large one with clinical end points (e.g., primary outcome = stroke), a pilot may focus on intermediate or surrogate end points (e.g., reduction in blood pressure) because they're quicker to obtain and/or they can be studied in a smaller sample.
- I'm going to focus on pilot studies for smaller (e.g., single-site) trials, and on the "miniature trial" concept.

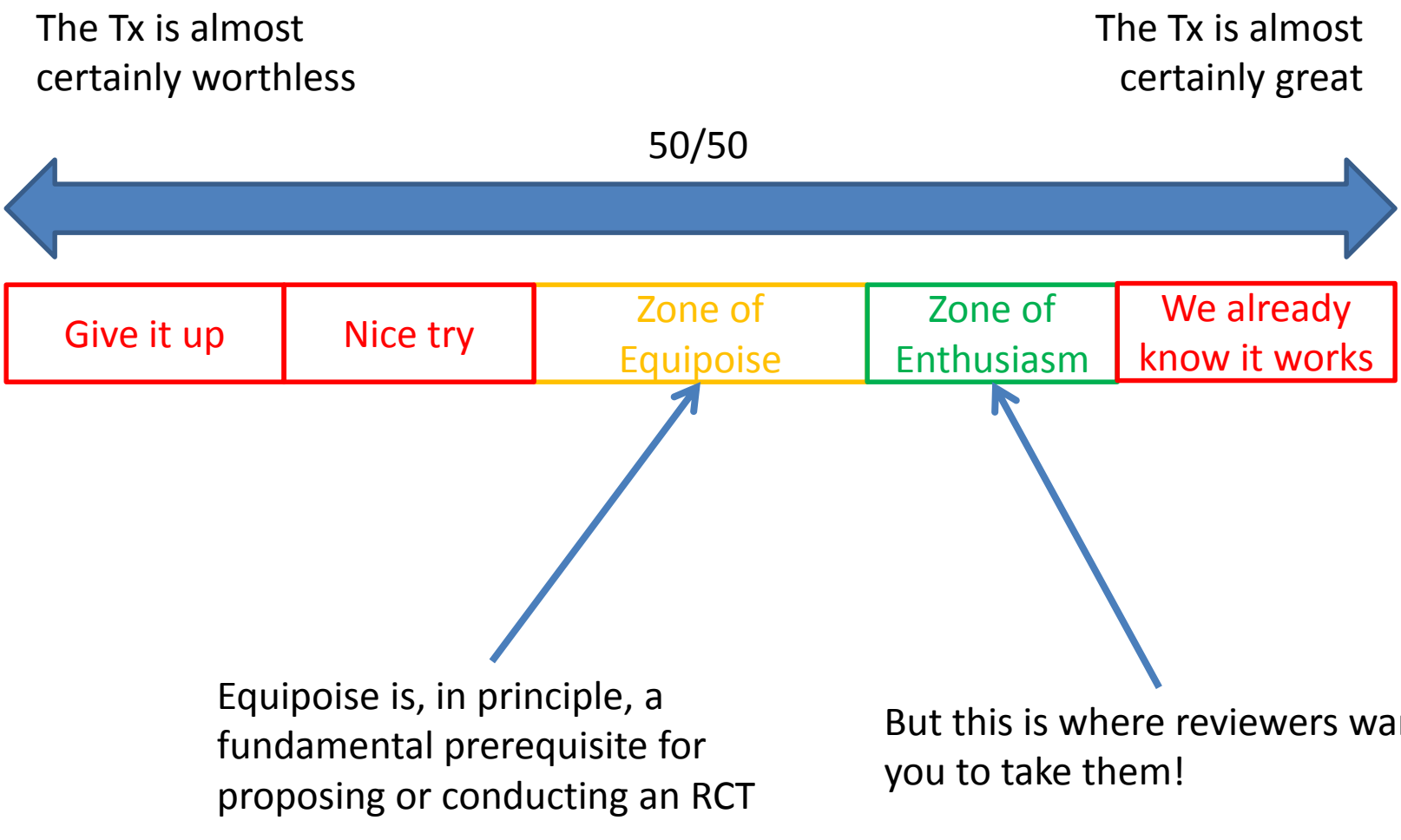
# Pilot Studies

- Two very common (but very questionable) reasons for conducting pilot studies:
  - To convince reviewers that the larger trial will turn out to show evidence of efficacy.
  - To obtain an effect size estimate for use in the power analysis that will determine the target sample size for the larger trial.

# Raising Efficacy Expectations

- Reviewers don't want NIH or other funding agencies throw away money on interventions that seem unlikely to be efficacious.
- So they want preliminary evidence of efficacy.
- Two problems with this desire:
  - It is at odds with the principle of equipoise.
  - Like Goldilocks, they want the preliminary evidence to be "just right" – not too bad, not too good.

# Raising Efficacy Expectations



# How to Play the Expectancy Game

- Provide preliminary evidence of efficacy if you must (and if you can), but don't get carried away and produce evidence that's too "definitive".
- Use other strategies to build a strong case that the trial is needed and worth conducting. E.g.,
  - The trial is needed to resolve a clinically important question
    - Informed if possible by practice-based research
  - There's a strong rationale for the intervention
    - Ingredients that make sense & that have mechanistic plausibility
  - Efficacy has been established in other populations; trial is needed to evaluate generalizability to a different population.

# Pilot Data and Effect Size Estimates

- Before 2006, most of us blithely assumed that we should just plug our pilot study effect size into the power analysis for our larger trial.
- A landmark paper by Kraemer et al. blew us completely out of the water.
  - Kraemer HC et al. Caution regarding the use of pilot studies to guide power calculations for study proposals. *Arch Gen Psychiatry* 2006;63(5):484-89.

# Kraemer et al. (2006)

"Clinical researchers often propose (or review committees demand) pilot studies to determine whether a study is worth performing and to guide power calculations.

**"The most likely outcomes are that (1) studies worth performing are aborted and (2) studies that are not aborted are underpowered.**

"The argument herein is not to discourage clinical researchers from performing pilot studies (or review committees from requiring them) but simply to caution against their use for the objective of guiding power calculations."

# Kraemer et al. (2006)

- The reason is that effect size estimates based on small pilot studies are *much too imprecise.*
- This fact has been gradually sinking in among clinical trialists and causing other statisticians to devise workarounds (more on this later).
- Kraemer et al.'s recommended *alternative* is less well known and less well understood.
- It raises some profound questions of its own.

# Kraemer et al. (2006)

- In behavioral research, we tend to conduct tiny pilot studies to pave the way for single-site trials that are themselves pretty small by the standards of medical trials.
- In medical research, it's not uncommon for Phase 2 studies (that are larger than many of our R01-level RCTs) to produce very good results, only to be followed by large Phase III multicenter trials with bad results.
- We should probably be more cautious about drawing firm conclusions from tiny pilot studies and small RCTs.

# Effect Sizes for Power Analyses

- If you can't get a usable effect size (ES) estimate out of your pilot study, where *can* you get one from?
- Kraemer et al. recommend that you use the **threshold of clinical significance (TCS).**
- This is the smallest effect size that might affect clinical decision making.

# Effect Sizes for Power Analyses

- The TCS is not a number that you can derive from a small pilot study.
  - Not because of imprecision.
  - Because pilot studies are inherently uninformative about TCS.
- TCS values should be derived from other sources, such as
  - research literature on the target disorder
  - relationship of the target outcome to downstream outcomes (e.g., how much of a decrease in anxiety is needed to prevent heart attacks?
  - efficacy of treatments that are already available for the target disorder
  - input from clinicians and patients (e.g., practice-based research)

# Threshold of Clinical Significance

- How to determine TCS values is a complicated question and an underdeveloped topic.
- A core problem is how to relate clinical significance at the individual patient level to clinical significance at the between-group level.
- Kraemer et al. (and many others) recommend the Number Needed to Treat (NNT) as a good way to think about this.

# Threshold of Clinical Significance

- NNT = the number of patients who would have to be exposed to the intervention to achieve one more successful outcome than would occur if everyone were given the alternative (comparison condition) instead.
  - Justification for providing relatively complex, risky, expensive, or otherwise burdensome treatments depends on *small* NNTs.
  - Easy, cheap, low-risk treatments can be worthwhile even with a very large NNT.

# Threshold of Clinical Significance

## NNT examples from systematic reviews

| Problem or Disorder | Treatment | Comparator | Outcome | NNT |
|---|---|---|---|---|
| Migraine | Sumatriptan | Placebo | Relief within 2 hours | 2.6 |
| Dog Bite | Antibiotics | Placebo | Prevention of Infection | 16 |
| Myocardial Infarction | Low-dose Aspirin | No treatment | Prevention of vascular death | 40 |
| Myocardial Infarction | Rapid Thrombolytic Therapy | Delayed Thrombolytic Therapy | Prevention of vascular death | 100 |

# Threshold of Clinical Significance

## NNT examples from two of my own trials*

| Problem or Disorder | Treatment | Comparator | Outcome | NNT |
|---|---|---|---|---|
| Post-CABG Depression | CBT | Usual Care | Remission on HAM-D | 2.6 |
| Depression in Heart Failure | CBT | Usual Care | Remission on HAM-D | 3.3 |

*These are relatively favorable outcomes compared to what other trials have shown, so the "true" NNT for these treatments may be higher.

# Threshold of Clinical Significance

- Calculating NNT is easy, but you have to base it on a clinically meaningful outcome at the *individual patient level. E.g.,*
  - Remission on HAM-D in my trials
  - Generally accepted cutpoint for “relief” on a widely used headache measure.
  - Percentage adherence that previous studies and/or expert consensus panels say is necessary for a drug to be effective.

# Threshold of Clinical Significance

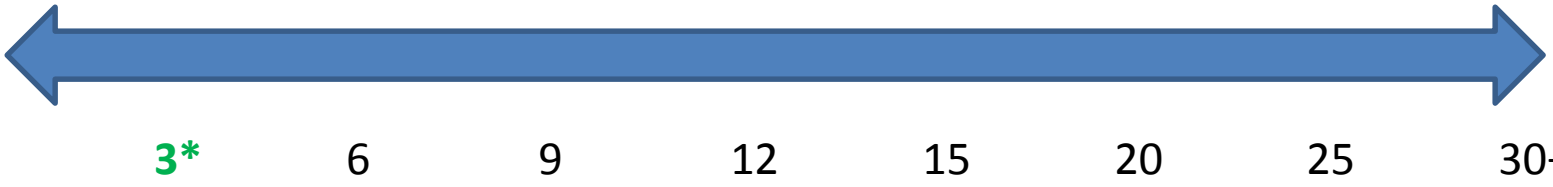
- Statisticians say that you can't or shouldn't plug an NNT into a power analysis.
- So, you'll still need a between-group effect size estimate that you *can* use.
- But you have to take two different effect sizes into account; this is widely misunderstood.

# Power Analysis

- The TCS is the estimated effect size on which you should base your sample size calculation.
- The *other* effect size is your estimate of the effect you are *likely* to find (or that you could plausibly find), based on what's already known.
- Unless this projected ES is at least as big as the TCS, reviewers may lack confidence that you'll be able to detect a TCS-sized effect in your trial.

# Power Analysis

Between-Group Difference on the BDI-II* for Trials of CBT for Post-MI Depression




95% CI around an ES for a study that probably won't be funded because the likely effect might be smaller than the TCS

95% CI for a trial that might get funded because the outcome seems likely to be at least as good as the TCS.

*Total score range, 0 to 63; mild dep = 14-19, mod. = 20-28, severe = 29-63

*3 is considered by some researchers to be a reasonable TCS for this measure.

# Power Analysis and Pilot Studies

- This brings us back to pilot studies.
- Pilot studies are often used (despite Kraemer et al.'s well-justified warning) to estimate likely or plausible effects for larger trials.
- It's often difficult to find a better source of data upon which to base a plausibility estimate.
- What a dilemma!

# Power Analysis and Pilot Studies

- Some (e.g., Powell) recommend using data from small, uncontrolled proof-of-concept studies that show better results than historical controls or clinical experience.
  - E.g., typical adherence to a certain med is around 25% in a particular low SES adolescent population.
  - Physicians say that >70% is medically necessary.
  - A small (n=8) proof of concept study shows that your new intervention can achieve >70% adherence in some cases.
  - Reviewers now know that it's at least conceivable that your proposed RCT could demonstrate a clinically meaningful ES.

# Power Analysis and Pilot Studies

- Defining the TCS is a more complicated challenge than it may seem.
  - E.g., what if the best we've been able to do, with various treatments over a number years, is a 3-point difference on the BDI-II?
  - Would a 3-point difference in one more trial of one more intervention be clinically meaningful, simply because 3 points was good enough in the past?
  - Would a 4-point difference be *more* meaningful?
  - Patients won't notice a mere advantage of 1 point, but it's at least better than previous interventions have shown.

# Power Analysis and Pilot Studies

- Most progress in behavioral medicine is *incremental*, not *dramatic.*
  - Huge breakthroughs are very rare.
  - On the other hand, we collectively spend too much time aiming for outcomes that are no better than what we were able to achieve years ago.
  - We should design our studies with the goal of making incremental progress.

# Power Analysis and Pilot Studies

- Statisticians have been working on ways to tweak pilot study data to yield useable estimates of plausible effect sizes.
- These efforts also help to answer to the question of what size the pilot study should be to yield useable estimates.
- This is still a strategy for making reviewers feel confident.
- It doesn’t guarantee that the trial will turn out as predicted.
- It might support your argument for conducting an RCT.

# Statistical Magic

**Here are some examples of statistical strategies for estimating effect sizes from pilot data. Consult a statistician if at all possible for using any of these approaches.**

Lee EC et al. The statistical interpretation of pilot trials: should significance thresholds be reconsidered? *BMC Medical Research Methodology* 2014;14:41

Sim J and 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.

Viechtbauer W et al. A simple formula for the calculation of sample size in pilot studies. *J Clin Epidemiol* 2015;68:1375-1379.

# Summary

- If you're feeling confused about feasibility and pilot studies, welcome to the club.
- Feasibility studies are easier to understand and less controversial than pilot studies.
- Pilot studies raise some difficult issues but they can be very helpful nevertheless.
- Consult with a statistician, if possible, even when designing a feasibility or pilot study.

# Some Helpful References

Kraemer HC et al. Caution regarding the use of pilot studies to guide power calculations for study proposals. *Arch Gen Psychiatry* 2006;63(5):484-489.

Leon AC, Davis LL, Kraemer HC. The role and interpretation of pilot studies in clinical research. *J Psychiatr Res* 2011;45(5):626-629.

Moore CG et al. Recommendations for planning pilot studies in clinical and translational research. *Clin Trans Sci* 2011;4:332-337

Thabane L et al. A tutorial on pilot studies: the what, why, and how.
*BMC Research Methodology* 2010;10:1