

## THE ROLE OF INTERNAL PILOT STUDIES IN INCREASING THE EFFICIENCY OF CLINICAL TRIALS

JANET WITTES AND ERICA BRITTAIN

*Biostatistics Research Branch, Division of Epidemiology and Clinical Applications, National Heart, Lung, and Blood Institute, Federal Building, Room 2A11, Bethesda, MD 20817, U.S.A.*

### SUMMARY

Investigators often design clinical trials without knowing precisely the values of such necessary parameters as the variances or the event rates in the control group. In order to determine reasonable values for such parameters, they may design a small pilot study external to the main trial. In this paper we propose designs, which we term internal pilot studies, that designate a portion of the main trial as a pilot phase. At the end of the internal pilot study, the investigators recompute preselected parameters and recalculate required sample size. The study then proceeds with the modifications dictated by the internal pilot. Final analyses of the results incorporate all data, disregarding the fact that part of the data came from a pilot phase. As one example of this type of design, we consider a study to compare two normally distributed means. By simulation, we show a numerical example for which the effect of the procedure on the  $\alpha$ -level is negligible, but the potential gain in power considerable. We urge considering a similar approach for a number of types of endpoints.

### INTRODUCTION

W. G. Cochran used to introduce his course on experimental design not only by teaching about such desirable characteristics of a study as randomization, balance, replication and blindedness, but also by warning of some actions one should not take. Listed among the undesirable actions were pilot studies. With a twinkle in his eyes, Cochran would announce that pilot studies, especially in clinical trials, would often lead to regret. A pilot that showed a statistically significant, or nearly significant, treatment effect would prevent one from embarking on a large scale study, for administrators would ask, 'Why should you perform a large study when a small one sufficed?' Yet a small study would not be persuasive enough to influence medical practice. Similarly, a pilot study that showed no effect would also prevent one from initiating a large study, because the administrator would challenge, 'Why should you embark on a major investigation when the pilot did not appear promising?' Yet the power of such a small pilot would be too low to exclude even important differences among treatments. He would cite the polio field trial as an example of a very large and influential study that had not included a pilot phase.<sup>1</sup> Cochran taught us that pilot studies were all too often mischief makers. The lesson at least one of his students took home was that if a study was worth doing, it was worth doing big. Those of us who had entered statistics from experimental science were at first shocked; a basic tenet of all experimental method we had previously learned was that because science proceeded incrementally, pilot studies were wise investments in the stepwise acquisition of new knowledge.

In this paper, we take Cochran's admonitions to heart as we outline some of the uses for pilot studies in clinical trials. We define two types of pilot studies. First is the usual kind, which we term an 'external' pilot – a study that is structurally distinct from the main study. The other type, which we term an 'internal' pilot, is an integral part of the main investigation. We argue that external pilot studies are appropriate for a variety of purposes related to the mechanics of a study, while internal pilots allow refinement of estimates of parameters used to design the study. In particular, an investigator may opt for an external pilot if he is truly unsure whether a process intended for the trial is feasible. On the other hand, an investigator who expects the process very probably to succeed may opt instead for an internal pilot. He can therefore retain the flexibility to change the design if necessary. We suggest that in many cases building an internal pilot study into a clinical trial will likely have at most a very small adverse effect on the significance level and may greatly improve the power of the study. As a specific example of the type of design we propose, we consider the effect of an internal pilot designed to recalculate the variance in a study with a normally distributed outcome variable. Although the essential ideas are relevant to binary outcome variables as well as to other variables with variances related to the means, the details in those cases are much more complicated.

### PURPOSES FOR PILOT STUDIES

External pilot studies in clinical trials have several uses. First, a pilot study can teach about the mechanics of the trial, for example how to administer therapy, how to identify potential participants, which of a number of possible drugs has fewest side effects, or whether a data entry form is well designed. For instance, phase I studies in cancer are essentially external pilot studies that determine appropriate doses. As another example, the SHEP pilot study was a randomized trial designed in part to determine whether elderly people would be willing to participate in a clinical trial testing the efficacy of antihypertensive medication in treatment of isolated systolic hypertension and whether medication would lower systolic blood pressure.<sup>2</sup> The investigators designed the study as an external pilot because they felt that a great many procedural issues had to be studied before embarking on a major clinical trial. We believe that investigators should routinely consider pilot studies to examine the components of a clinical trial. When feasible, we recommend that such pilot studies be randomized trials. Eliciting informed consent to randomization in the pilot helps gain insight into the acceptability of the trial to prospective participants. Moreover, any surprises that arise in the course of the study can be better assessed if a control group is present. Finally, a randomized study, but not an unrandomized one, is a candidate for pooling with other related studies.<sup>3</sup>

This paper deals with internal pilot studies that estimate parameters necessary for the design of the trial. The designers of a clinical trial should have reliable prior estimates of three classes of parameters, related to (1) the administration of the study, (2) the process of the disease and (3) the effect of treatment. The boundaries separating these classes overlap somewhat, but the classes do provide a guideline for categorizing variables. Examples of administrative variables are (1) the number of patients that the participating clinic can expect to identify, (2) the willingness of patients and their physicians to join the trial and (3) the recruitment rate in the clinics. Because these parameters pertain to the state of the participants before randomization, as long as the outcome variable is not a time-to-event measure, changes in them during the course of the trial do not affect the  $\alpha$ -level of the test of treatment effect. For example, one can usually add new clinics if recruitment is low without affecting the operating characteristics of the significance tests.

Next are the parameters that relate to the process of the disease in the control group, for example: (1) the likely degree of compliance to the protocol, as measured by the number of missed

visits or the willingness to take placebo medicine; (2) the variance of the outcome variable, or, for binary outcomes, the event rate in the control group; (3) for both continuous and binary variables, the rate of progression of the disease in the control group during the time course of the study; and (4) the rates of competing risks. We consider these variables to be candidates for both internal and external pilots.

Finally, some variables important to design are related to treatment itself, for example the treatment effect one desires to detect. A change in this type of variable during the course of a study has the potential to affect the  $\alpha$ -level considerably. Therefore we do not regard such variables as candidates for internal pilot studies.

The sample size required to detect a given effect is sensitive to all the parameters mentioned above and to others as well; unfortunately, their values are extremely difficult to specify accurately before the trial begins. All of us involved in clinical trials can remember studies in which estimates have been wildly wrong—an observed variance five times that projected,<sup>4</sup> an observed recruitment rate one-third of the promised rate,<sup>5</sup> a mortality rate only one-eighth of the prior estimate.<sup>6</sup> Where do our projections come from? A dash of epidemiology added to some case histories, clinical impressions, statistical hunches, budgetary nudges, and unwarranted optimism. The reality of the trial is often very different from our expectations, for we select patients and they select us. In fact, selection itself is one important reason for the difference between what we estimate from the available data and what we see in our trials. Informed consent and our entry criteria select a specific subset of patients. As a result, our study group is much more homogeneous than the cohorts in epidemiology. By contrast, the process of selection, especially in multi-centre studies, tends to lead to a more heterogeneous group than in most reports of clinical case series.

In order to ensure that the parameters used for the design will be accurate, many people recommend performing a pilot study prior to the trial. This external pilot should select a group of patients large enough to achieve good estimates of the necessary parameters. The problems with such an approach hark back to Cochran's warning: a pilot that is too small cannot provide reliable enough estimates of the parameters of interest; a pilot that is too large is not a pilot. Suppose, for example, one performs a pilot study in order to estimate an event rate. Because the pilot is small, the estimated rate has a very wide confidence limit. The SHEP pilot study observed in the placebo group an annual rate of stroke of 0.016.<sup>2</sup> There were roughly 100 participants in the placebo group; the standard error of the event rate was 0.012. External pilot studies are often very unrepresentative of the population to which they refer. They are small; the participants tend to come from a very few clinics; the follow-up is short. Indeed, a pilot study that takes too long will delay the onset of the main trial, perhaps so long that the therapy in question may have entered practice to an extent that precludes a study.

In contrast to external pilot studies, an internal pilot can be large with essentially no increase in time or money. In an internal pilot, the protocol designates the first phase of the study as a 'pilot' phase; at the end of the pilot, the investigator estimates the parameters of interest and recalculates the sample size. The analysis of the data at the end of the trial treats all observations as if they had come from a single study. Clearly, most administrative parameters can be handled this way, and we suspect they are. For instance, if recruitment is too slow, investigators often extend their trials. (One must be a little careful, however, before assuming that extending recruitment has no effect on the statistical properties of a trial. Specifically, in time-to-event data, a change in the recruitment period alters the time at risk and therefore has some effect on the operating characteristics of the test of treatment effect.) Also, clearly, one must not adjust sample size on the basis of internal data that estimate parameters related to the treatment variables, for that would importantly bias the results. Our interest in this paper focuses on those parameters that relate to

the process of disease in the control group. We consider here one example of such an internal pilot: the case of a normally distributed outcome variable.

#### EXAMPLE: AN INTERNAL PILOT FOR ESTIMATING VARIANCE

Consider a clinical trial to compare two groups, treated and control, on the basis of a normally distributed outcome variable. Perusal of the literature before embarking on the study leads the investigators to project a variance of  $\tau^2$ ; unbeknownst to them, the true variance is  $\sigma^2$ . In our own experience, the projected variance  $\tau^2$  is likely to be considerably less than  $\sigma^2$  because the literature tends to report more homogeneous case series than will be entered in a large clinical trial, especially an efficient one with inclusive entry criteria.

Using a type I error  $\alpha$  and a power of  $1 - \beta$  for the likely treatment effect  $\delta$ , one calculates the sample size per group  $n_0$  as usual. Further, the investigator selects a proportion  $\pi$ ; the first  $\pi n_0$  patients and  $\pi n_0$  controls constitute the internal pilot. When the study has assessed the endpoints for all these patients, the investigator calculates the estimated variance  $s^2$  as the pooled variance from the two groups. If  $s^2 \leq \tau^2$ , the study continues as planned so that the total sample size remains  $2n_0$ . If, on the other hand,  $s^2 > \tau^2$ , one adjusts the sample size using the newly estimated variance. At the end of the trial, the data analyses incorporate all observations – those collected during the internal pilot phase as well as those collected subsequently – to assess the effect of treatment. Although the procedure inflates the true  $\alpha$ -level (see Lohr<sup>7</sup> for a proof in a more general setting), the degree of inflation will be small in most practical settings. Intuitively, if  $\tau^2 \ll \sigma^2$ , that is if one has grossly underestimated the variance, then at the end of the pilot phase  $s^2$  is likely to be much larger than  $\tau^2$ , the sample size will increase greatly, and the pilot phase will have become a small portion of the entire study. Thus the true  $\alpha$ -level will be very close to the nominal level. Conversely, if  $\tau^2 \gg \sigma^2$ , that is if one has grossly overestimated the variance, a change in sample size is highly unlikely at the end of the pilot, so that once more the study has preserved its  $\alpha$ -level. Ironically, the  $\alpha$ -level is adversely affected only when the prior guess  $\tau^2$  is not very far from the true variance  $\sigma^2$ . In that case, a change in sample size is highly likely, but if  $\pi$  is large (say 1/2) the increment in sample size will be low, the pilot will represent a large portion of the total data, and the  $\alpha$ -level will be inflated. Under most practical situations, even when  $\tau^2$  is close to  $\sigma^2$ , the inflation of the  $\alpha$ -level will be only very slight. We recommend either ignoring this bias in type I error, or setting the critical value at the beginning high enough to ensure that the type I error is less than  $\alpha$  for all possible values of  $\sigma^2$ .

The design we suggest is a variant of Stein's classic two-stage procedure for estimating a normal mean.<sup>8</sup> Stein's method differs from ours in two important respects. First, he permits stopping at  $\pi n_0 + 1$ ; we require that the sample size per group be at least  $n_0$  even if the observed variance at the end of the pilot phase is much lower than projected. Second, Stein's  $t$ -test excludes the data after the pilot phase in the estimation of the variance; our  $t$ -test includes all the data, from the pilot phase as well as the later phase, in estimates of the means and the variance. The advantage of Stein's approach is that his  $t$ -test is of exact level  $\alpha$ ; our type I error is slightly inflated because we treat the data as if the two phases were independent when in fact they are not. We believe that our approach is preferable to Stein's in clinical trials for several practical reasons. Cutting the size of a study in the middle of the trial can lead to problems in interpretation later. If the results of the study turn out to be inconclusive, the investigators are likely to question whether their decision to cut the trial back was warranted. If, on the other hand, strong trends in the treatment effect emerge early in the study, the trial may well be terminated before the planned end anyhow. Also, we are uncomfortable using only the initial phase of the study for estimating the variance, because

Table I. Effect of an internal pilot study on  $\alpha$ -level, power, and expected sample size  $E(N)$  as a function of true variance  $\sigma^2$ 

Fixed design		$E(N)$	Internal pilot	
$\sigma^2$	Power		$\alpha$	Power
1.0	0.995	86.0	0.050	0.996
1.5	0.96	86.5	0.050	0.97
2.0	0.90	93.2	0.050	0.93
3.0	0.74	128.4	0.051	0.89
4.0	0.61	170.0	0.052	0.90

Study design:  $\delta = 1$ ,  $\tau^2 = 2$ , initial total sample size 86; pilot phase 43.  
 Number of replications in simulations: power 5000;  $\alpha$ -level 40,000.

we believe that sacrifices too much information and hence power. Because our simulations suggest that the bias in  $\alpha$ -level is usually trivial, we opt for the higher power at the expense of slight bias in type I error.

Several questions remain for this design. We are trying to determine the ratio  $\sigma^2/\tau^2$  that corresponds to the maximum true  $\alpha$  and also that maximum value of  $\alpha$ . Further, we are interested in learning how to select  $\pi$ . Moshman has discussed a similar problem in connection with Stein's two-stage procedure.<sup>9</sup>

### Numerical example

Consider an experiment in which the true treatment difference  $\delta$  is 1, the projected variance  $\tau^2$  is 2, and  $\pi = 1/2$ . A two-tailed test at  $\alpha = 0.05$  and power of 0.9 requires a sample size of 43 patients per group. Table I compares the simulated power of the fixed sample study to that of the study with an internal pilot as a function of the true variance  $\sigma^2$ . The table also presents the exact expected sample size  $E(N)$  and the simulated  $\alpha$ -levels arising from the use of an internal pilot. Because our sample sizes were small and we wished to calculate  $\alpha$  precisely, we used  $t$ - and non-central  $t$ -distributions rather than their normal approximations to compute sample sizes at the end of each simulated internal pilot study. In our simulations, we used 40,000 replications for  $\alpha$  and 5000 for the power. The table demonstrates the potential attractiveness of the internal pilot. Although theoretically the  $\alpha$ -level is greater than the nominal level, in our simulations, when  $\tau^2 = \sigma^2$ , the design incorporating an internal pilot led to only a very small increase in average sample size with no detectable effect on either  $\alpha$  or power. (In fact, our simulations yielded  $\alpha = 0.050$  with an SE of 0.0011.) Thus, the loss was quite small when our prior estimate of variance was essentially correct. When the true variance was less than projected, the expected sample size was barely greater than the fixed value, the  $\alpha$  was only very slightly inflated, and the power was, of course, higher than the design required. This last fact has nothing to do with the internal pilot, but is simply a consequence of the smaller variance.

The internal pilot was very beneficial when  $\sigma^2 > \tau^2$ , which, we believe, is typical of clinical trials. In that case, the fixed design study is in fact underpowered while the design using an internal pilot has power essentially 0.9 and an  $\alpha$ -level very close to 0.05. Of course, the benefit of the internal pilot can only be realized if the investigators are actually able to increase the sample size in accordance with the variance observed at the end of the pilot phase. Note that although the simulated  $\alpha$ -levels appear to increase with increasing  $\sigma^2$ , none of the  $\alpha$ s differs significantly from any other.

## DISCUSSION

We have presented a design for an internal pilot study in a clinical trial that employs a normally distributed measure as an outcome variable. Our purpose was to demonstrate the advantages of this type of design. At only a very small sacrifice in  $\alpha$ , one may gain considerable flexibility and may increase importantly the chance of carrying out a study of reasonable power. We believe other situations are candidates for similar internal pilot studies. One might use an internal pilot in studying a binomial endpoint in a study with an acute response as the endpoint; the natural estimate to recalculate sample size would be the success rate in the control group at the end of the pilot phase. An application more difficult to implement would be to studies with an endpoint that measures survival. In comparing survival times with exponential survival and the logrank test, one might recalculate the exponential parameter at the end of the pilot phase and recompute sample size. In practice, however, such uses of internal pilots may not be feasible if the planned follow-up period extends much longer than the end of the recruitment period. Other possible uses for internal pilots include checking the assumptions regarding compliance in the control group and adjusting the sample size if non-compliance is greater than expected. Any such modification will affect the  $\alpha$ -level somewhat, but we suspect the effect will be negligible in many cases. Because the potential gain is considerable, we hope designers of clinical trials will consider carefully the use of internal pilots whenever they face considerable uncertainty in their prior estimates of important design parameters. Before using such an internal pilot, we recommend checking its likely effect on the type I error.

Although in this paper we have focused almost exclusively on internal pilots designed to recalculate sample size, the idea of using an internal pilot has much broader implications. One might design an internal pilot when one is unsure whether participants in the trial will comply with the protocol. In a study of the effects of reducing intake of alcohol, for instance, one could design an internal pilot to test whether the proposed intervention actually is effective in reducing intake. If the intervention has failed, one can terminate the study. If, on the other hand, the intervention has been effective in reducing intake, one may use the participants from the internal pilot in the main analysis of the data. Sometimes the design of a study is dependent upon the distribution of baseline parameters. Again, one can designate a portion of the sample as an internal pilot, measure the baseline parameters, and alter the design accordingly. For instance, in a study of the effects of cholesterol lowering on coronary artery graft occlusion following bypass surgery, investigators cannot know the distribution of the number of grafts before embarking on the study. Yet the optimal test statistic to compare treatments depends in part upon that distribution.<sup>10</sup> One can design an internal pilot that specifies how many angiograms to look at in order to predict the distribution of grafts. Then one can choose the test statistic on the basis of the distribution observed in the pilot.

The spirit behind internal pilots is simple: one uses the patients in the pilot to alter the main study, but one does not discard those data from those patients. As long as the  $\alpha$ -level is controlled, we believe these designs offer flexibility and power.

## REFERENCES

1. Francis, T., Korn, R. F., Voight, R. B., Boisen, M., Hemphill, F. M., Napier, J. A. and Tolchinsky, E. 'An evaluation of the 1954 poliomyelitis vaccine trials summary report', *American Journal of Public Health*, **45**, (part 2) 1-63 (suppl.), (1955).
2. Hulley, S. B., Furberg, C. D., Gurland B., McDonald R., Perry, H. M., Schnaper H. W., Schoenberger, J. A., Smith W. M. and Vogt, T. M. 'Systolic hypertension in the elderly program (SHEP): antihypertensive efficacy of chlorthalidone', *American Journal of Cardiology*, **56**, 913-920 (1985).

3. Yusuf, S., Simon, R. and Ellenberg, S (eds.) 'Proceedings of the Workshop on Methodological Issues in Overviews of Randomized Clinical Trials', *Statistics in Medicine*, **6**, 217–409 (1987).
4. Parrillo, J. E., Cunnion, R. E., Epstein, S. E., Parker, M. M., Suffredini, A. F., Brenner, M., Schaer, G. L., Palmeri, S. T., Cannon, R. O., Alling, D., Wittes, J. T., Ferrans, V. J., Rodriguez, E. R. and Fauci, A. S. 'A prospective, randomized, controlled trial of prednisone for dilated cardiomyopathy', *New England Journal of Medicine*, **321**, 1061–1068 (1989).
5. Probstfield, J. L., Wittes, J. T. and Hunninghake D. B. 'Recruitment in NHLBI population based studies and randomized clinical trials: data analysis and survey results', *Controlled Clinical Trials*, **8**, 141S–149S (1987).
6. The Steering Committee of the Physicians' Health Study. 'Preliminary report: findings from the aspirin component of the ongoing Physicians' Health Study', *New England Journal of Medicine*, **318**, 262–264 (1988).
7. Lohr, S. 'Accurate multivariate estimation using double and triple sampling', Technical Report, University of Minnesota School of Statistics, 1988.
8. Stein, C. 'A two-sample test for a linear hypothesis whose power is independent of the variance', *Annals of Mathematical Statistics*, **16**, 43–258 (1945).
9. Moshman, J. 'A method for selecting the size of the initial sample in Stein's two sample procedure', *Annals of Mathematical Statistics*, **29**, 1271–1275 (1958).
10. Zucker, D. and Wittes, J. 'Testing the effect of treatment in experiments with correlated binomial outcomes', manuscript in preparation.

## DISCUSSION

**Dr. Deykin:** What you are doing is to allow the ability to extend a clinical trial earlier than is done conventionally. You will, however, be making your recommendation for extension on the basis of a small sample. How would you try to convince a funding agency that your data are convincing enough to justify adding more patients? How far into the trial would you expect to see convincing data?

**Dr. Brittain:** That depends on the nature of the endpoint and the aim of the internal pilot. Moreover, the aim will differ depending on the stage of the study. Obviously, the farther the study has progressed the more we know. If, however, we are talking about a large study in which our pilot is designed to estimate variance, we would not need many observations to achieve a stable estimate of the variance, and hence a reasonable calculation of sample size.

**Dr. Meier:** If we have a little data and observe that the variance is about five times what we had thought it was going to be, the funding agency may tell us to quit. They are entitled to have that information and it's right to act upon it. My understanding of your talk, however, leaves me to draw an even sharper distinction between external and internal pilots. When you must simply reset uncertain parameters, then an internal pilot is probably the right approach. Why, after all, should you throw away data that you collect in the same manner with the same validity as data that you are going to get later? If, on the other hand, the question is whether some clinics are capable of getting quality ECGs, for example, we want an external pilot. We expect the pilot data will be junk and we don't want to contaminate the main study with it. So I see a fairly sharp distinction about the kinds of problems that you are discussing. Some problems will point you to internal and some to external pilots. Sometimes, in fact, you may perform a little external pilot to begin with, and then proceed with an internal one.

**Dr. Brittain:** I agree that in the case you just mentioned, an external pilot is clearly appropriate, but an internal pilot might address other feasibility issues.

**Dr. Lang:** I am interested in the asymmetry of your design. You stated that the data from your internal pilot should not be used to decrease your sample size. For example, I could imagine that

a funding agency might be willing to increase sample size if your internal pilot showed your size was too small. Similarly, why wouldn't the funding agency also want the right to reduce the sample size if indeed you had overestimated size?

**Dr. Brittain:** We were concerned decreasing sample size because of the adverse effect that could have on the  $\alpha$ -level.

**Dr. Wittes:** Further, if you allowed for a decrease in sample size, you might find it runs the risk of logical inconsistency. What would you do if, half-way through the trial, your new calculated sample size were less than you had already collected?

**Dr. Yusuf:** A practical answer to Janet Lang's interesting point is that if you had truly overestimated your sample size, you are likely to find a very convincing result early. In that case, you will stop your trial and save your money.

**Dr. Gordon:** Let me suggest another situation in which an internal pilot is more appropriate than an external one. Suppose you are interested in testing recruitment or compliance for a long study. If you perform a short external pilot study, you might get a misleading answer because people might be more willing to join a short study than a long study/or be willing to put up with a bad treatment in a short study than a long study. An internal pilot can test recruitment and compliance to the actual protocol.