Design and Analysis of Equivalence Clinical Trials Via the SAS® System

Pamela J. Atherton Skaff, Jeff A. Sloan,
Mayo Clinic, Rochester, MN 55905

ABSTRACT

An equivalence clinical trial typically is conducted to demonstrate that there is no clinically significant difference between a standard and an experimental treatment. The study is designed with the desired outcome being equivalence in efficacy, while immediate toxicity, long-term adverse effects or costs may demonstrate to be advantageous for the experimental treatment. For such experiments the usual hypothesis testing framework, which tests the null hypothesis of no difference in efficacy, is inappropriate. Instead, one tests for the presence a specified difference between the efficacy of the two treatments to be no more than some value delta, δ. While the theory underlying equivalence trials has been developed and applied in clinical environments (Blackwelder, 1997), dedicated statistical software is lacking. In this paper, we summarize the statistical methodology and present a series of SAS macros for equivalence trials. Topics covered include sample size determination in the design of equivalence studies as well as hypothesis testing and confidence intervals for the efficacy endpoint.

INTRODUCTION

HYPOTHESES

The standard comparison trial has a null statistical hypothesis which holds that two quantities are indifferentiable. The alternative hypothesis states the two quantities are different. The null hypothesis will be rejected in favor of the alternative if the data provide enough evidence that the alternative is true.

The goal of an equivalence trial is to prove that two quantities are equal, i.e. prove the null hypothesis from the standard comparison. But, the null hypothesis in the standard methodology for a comparison trial can never be accepted, rendering the standard hypothesis testing framework inapplicable. In order to test precisely for the equivalence of two treatments, the alternative hypothesis must reflect that the treatments are the same. This fact necessitates a role reversal in the defining of the hypotheses. The hypothesis testing framework for an equivalence trial requires the specification of no difference in the alternative and a difference in the null.

The equivalence testing framework contains an additional parameter delta, δ, to indicate the maximum clinical difference allowed for an experimental therapy to be considered equivalent with a standard therapy.

The equivalence hypotheses are typically one-sided as follows:

$$H_0: \ \theta \geq \delta$$
$$H_1: \ \theta < \delta$$

Where θ is the difference or ratio of the success measures for the experimental treatment and the standard treatment. We test that the success differential between the two treatments is smaller that the specified criteria (δ). The alternative hypothesis may be accepted with 100(1-α)% confidence and the two agents may be considered equal.

ERROR RESPECIFICATION

The type I error (α), in the standard comparison trial context is the probability that an experimental therapy is thought to be inferior to a standard therapy, when no difference actually exists, i.e. rejecting H_0 when it is true. A type II error (β), is the probability that the experimental therapy is concluded to be no better than the standard when in fact there is a difference, i.e. failing to reject H_0 when it is false.

For an equivalence trial, stating that an experimental treatment is no better than a standard when it is not, is failing to reject H_0 when it is false, which is equivalent to the type II error from the standard hypothesis. Thus, just as the null and alternative hypothesis have reversed roles, so do the definitions of the errors. The type I error of an equivalence trial is the type II error of the standard comparison trial. However, the significance level (α) of the equivalence trial is chosen in the usual manner.

STATISTICAL PROCEDURES

Statistical methodology that exists for the equivalence test includes performing the test of the statistical hypothesis, construction of a confidence interval and predicting sample size. In this section we briefly describe the motivation for each technique involved in cases of binomial (Blackwelder, 1982), continuous (Jennison & Turnbull, 1993), and survival (Com-Nougue et al, 1993) endpoints.

BINOMIAL ENDPOINTS (PHASE I/II TRIALS)

For binomial endpoints is defined as the positive difference of the probability of success for the standard group (Pe) and the probability of success for the

experimental group (Ps) of patients, i.e. $\theta = Ps - Pe$. The test statistic involved is the usual Z statistic derived from the normal approximation to the binomial distribution.

$$Z_0 = \frac{\theta - \delta}{SE} \text{ where}$$

$$SE = \left[\frac{Ps(1-Ps)}{Ns} + \frac{Pe(1-Pe)}{Ne} \right]^{1/2}$$

Ns refers to the standard sample size and Ne refers to the experimental sample size. Note that SE is a function of Ps and Pe rather than the pooled success rate in the standard hypothesis testing framework because here the null hypothesis assumes a difference of in the success rates. We will reject the null and conclude equivalence of success rates between the standard and experimental treatments if $Z_0 < Z(\alpha)$ or alternatively if the p-value of the result $P(Z < Z_0)$ is less than α.

A 100(1-α)% confidence interval may be a substitute approach to making the decision of accepting or rejecting the alternative hypothesis and is calculated as follows:

$$CI = (-\infty, \theta + Z_{1-\alpha} \cdot SE) \qquad (1)$$

Example: Assume we are testing a standard product, S, against an experimental product, E. We find that S is successful in 85 out of 100 patients and E is successful in 78 out of 99 patients. Are the two treatments equally effective to within a difference of $\delta = 0.10$? We calculate

$$Z_0 = (0.85 - 0.79 - 0.10) / 0.054 = -0.741$$

and calculate the p-value to be $P(Z < -0.74) = 0.229$.

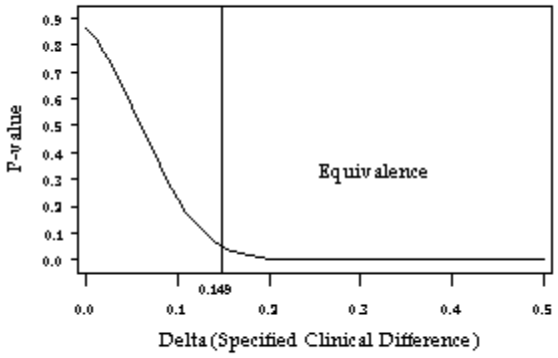
At the 5% level of significance we would not reject the null hypothesis and conclude that the difference in success rates between the two drugs could indeed be as much as $\delta = 10\%$. We can rule out that the drugs are more than 20% different in terms of success rate because if we were to test H_0: $\theta \geq \delta = 0.20$ we would get $Z_0 = -2.59$ with a corresponding p-value of $P(Z < -2.59) = 0.005$. Hence, the drugs would be deemed equivalent for $\delta = 0.20$.

The confidence interval upper limit for the population proportion difference is 0.149. This limit of the confidence interval is not less than our chosen $\delta = 0.10$ so we cannot conclude that the difference between the products is less than 10%.

This approach is different than the conventional confidence interval approach where we reject the null hypothesis and say two treatments are different if the confidence interval lies completely above the test statistic, i.e. we reject for large observable differences. In the equivalence test we reject for small observable differences. The CI tells us what values of delta are consistent with the data. This confidence interval may be interpreted as saying that the smallest value we can choose for delta and reject the null hypothesis of non-equivalence is $\delta = 0.149$. This value for delta may not be feasible clinically.

The power curve depicts the p-value of alpha and the corresponding delta for our example. The vertical reference line depicts the smallest value of delta with which equivalence will be detected. Values larger than delta = 0.149 will have a p-value smaller than alpha = 0.05, thus the null hypothesis will be rejected and equivalence concluded. Any value of delta less that 0.149 favors the non-rejection of the null hypothesis. The power of our test is 1-α, as explained in the previous section on error specification. The smallest value of delta for which equivalence will be concluded is also the upper limit of the 95% confidence interval.

Figure 1 Power Curve for Ps = 0.85 and Pe = 0.79



The Two Products are Equivalent for a Clinical Difference > 0.149

CONTINUOUS ENDPOINTS (PHASE II/III TRIALS)

The continuous endpoint case defines θ to be the difference in the population means between the standard and experimental groups for an appropriate variable of analysis, commonly referred to as the population mean difference, i.e. $\theta = \mu_s - \mu_e$. The standard and experimental treatments should at least have sufficiently similar distributions (Sheiner, 1992). In the continuous case, δ is sometimes defined as a percentage of the mean for the standard therapy group.

The formulations for the continuous data follows from statistical theory for testing the difference between two population means.

$$Z_0 = \frac{\theta - \delta}{SE} \text{ where}$$

$$SE = \left[\frac{Vs}{Ns} + \frac{Ve}{Ne} \right]^{1/2}$$

Vs and Ns refer to the variance and sample size of the standard group, and Ve and Ne refer to the same values in

the experimental group. The CI formula (1) applies with revised θ and SE terms.

A continuous case where a straight percentage difference is sought is not defined since the resulting testing procedures are identical to the ones defined for the binomial endpoints.

Example: Assume we are testing a standard drug, S, with $\bar{x}_s$=2.70, Vs=0.36 and Ns=100 versus an experimental drug, E, with $\bar{x}_e$=2.61, Ve=0.40 and Ne=100. If δ is defined as 10 percent of the mean of the standard group then δ=0.27. If we set α=0.05 to test H_0: $\theta \geq \delta = 0.27$ we have

$$Z_0 = (2.70 - 2.61 - 0.27) / 0.087 = -2.069$$

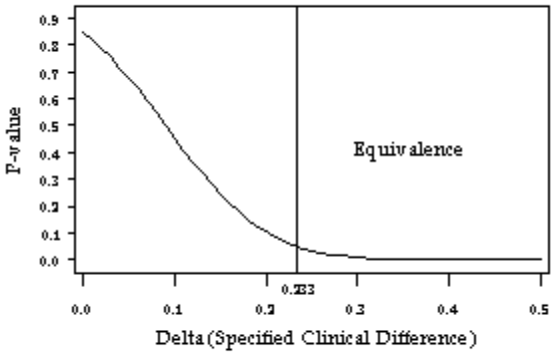
and calculate the p-value to be P(Z < -2.07) < 0.019.

At the 5% significance level we reject the null hypothesis in favor of the alternative and conclude that the difference between means of the two drugs is less than 10% of the standard mean.

The confidence interval upper limit for the difference of population means is 0.233. This upper limit of the CI around the actual difference is less than the specified clinical significance of 0.27 thus we are 95% confident that the true difference is less than delta therefore we conclude that we reject H_0 and say the two drugs are equivalent.

The power curve below relates the p-value to delta for the continuous data example. Values of delta larger than 0.233 will have a p-value smaller than α= 0.05. If we wish to have α= 0.10, then we could allow for a clinical difference of approximately 0.20, i.e. 7.5% of the standard mean, and still conclude equivalence.

Figure 2 Power Curve for μs = 2.70 and μe = 2.61



The Two Products are Equivalent for a Clinical Difference > 0.233

SURVIVAL ENDPOINTS (PHASE III TRIALS)

The case of a survival endpoint defines θ to be the relative risk which is the ratio of the hazard rates, i.e. $\theta = \frac{h_e(t)}{h_s(t)}$, at a particular time, t. The test statistic for allows for censored data, where a difference in survival probabilities is actually a case of binomial endpoints. If the hazard rates of the two treatments are identical the ratio will be equal to one, but if the experimental product has a higher hazard rate, the ratio will exceed one. Thus, δ is now defined as the highest clinically acceptable hazard rate to determine equivalence, i.e. δ = 2 or 3. The statistical procedure is derived from the usual logrank test based upon the observed number of deaths at each time point and the number expected in the new treatment, where the true relative risk is assumed to be equal to δ. Let us define omega, Ω, as

$$\Omega = \sum O_{E_i} - \sum E_{E_i}(\delta),\ i=1,...,k \text{ then}$$

$$Z_0 = \frac{\Omega}{SE}, \text{ where}$$

$$SE = \sqrt{\sum \frac{n_{E_i} n_{S_i} \delta}{\left(n_{E_i} \cdot \delta + n_{S_i}\right)^2}},\ i=1,...,k \text{ and}$$

$$E_E = \sum \left(\frac{n_{E_i} \cdot \delta}{n_{E_i} \cdot \delta + n_{S_i}} \right),\ i=1,...,k.$$

The statistics are summed over the observed distinct times of death, i, and O_{E_i} is number of patients who died at time t_i in the experimental group and E_{E_i} is the number of deaths at time t_i expected in the experimental group. The time is chosen in small enough increments so the value of O_{E_i} is either 1 or 0. Again, we will reject the null hypothesis in favor of equivalence if $Z_0 < Z(\alpha)$ or alternatively, the p-value of the result $P(Z < Z_0)$ is less than α.

A 100(1-α) confidence interval for the difference of survival probabilities follows from (1) with θ = Ω and the revised SE term. The CI for the relative risk function follows from the normal approximation of the hypergeometric distribution and the inverse of the relationship of δ to the risk ratio. It is calculated as

$$CI = \left[-\infty, \exp\left(\frac{4(O_E - E_E^*)}{k} + \frac{2 \cdot Z_{1-\alpha}}{\sqrt{k}} \right) \right]$$

where E_E^* is the estimate established by previous studies and k is the number of distinct times of death.

Example: Assume a standard cancer treatment regimen has Ns=84, O_S=11, and E_S=5.47 while an experimental

regimen has Ne=82, O_E=9, and E_E=14.53. If δ is defined as a relative risk proportion of 2.73, can the two treatments be considered equivalent? If we set α = 0.05 to test H_0: $\theta \geq 2.73$ we calculate

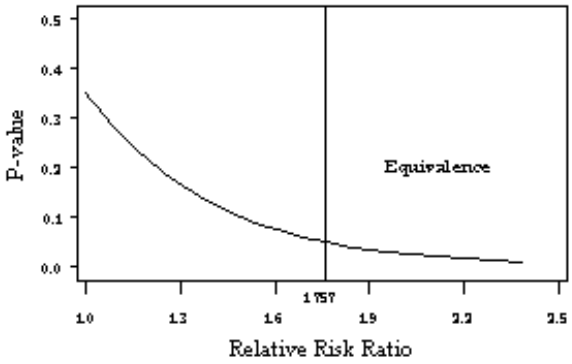
Z_0= (9 - 14.53) / 1.996 = -2.77

and calculate the p-value to be P(Z < -2.77) = 0.003.

At the 5% significance level we reject the null hypothesis and conclude that the relative risk ratio is small enough to conclude that the two regimens are equivalent. The upper limit of the confidence interval is 1.757 which confirms the decision.

The power curve depicts values of alpha and corresponding values for the relative risk ratio. A relative risk higher than 1.757 indicates equivalence.

Figure 3 Power Curve For Survival Endpoints



The Two Products are Equivalent for a Relative Risk > 1.757

SAMPLE SIZE

There is a general misconception that it requires a larger sample to prove equivalence than difference (Blackwelder, 1982). The veracity of this assumption is situation-specific and dependent upon the degree of difference, δ, one is willing to allow so that the standard and experimental treatments may still be considered equivalent. Formulas for the determination of the sample size for the various data types follow with n being the sample size per group.

Binomial Endpoints:

$$n = \frac{(Z_{1-\alpha} + Z_{1-\beta})^2 (Ps(1-Ps) + Pe(1-Pe))}{(Ps - Pe - \delta)^2}$$

where Ps and Pe are probabilities of success for the experimental and standard treatments, respectively.

Continuous Endpoints:

$$n = \frac{(Z_{1-\alpha} + Z_{1-\beta})^2 (Vs + Ve)}{(\mu_s - \mu_e - \delta)^2}$$

where Vs and Ve are the variances of the standard and experimental groups, respectively.

Survival Endpoints:

$$m = \frac{4(Z_{1-\alpha} + Z_{1-\beta})^2}{(\log(\delta))^2}$$

where m is the total number of events.

MACROS

A set of three SAS® macros has been developed for the purpose of calculating the statistical output. The user may analyze a data set, analyze a submitted set of values for output variables, or calculate a sample size for a particular type of data set: binomial, continuous, or survival. The three macros allow for the inclusion of several parameters which have default values, which are used if the parameters are omitted in the macro call.

ANALYSIS OF A DATA SET

The macro call parameters for the analysis of a data set are the type of data set, the data set name, the significance level, the clinical significance, the name of the success variable, the grouping variable, the values for the grouping variable. Defaults have been set as follows:

```
%DSANAL (       datatype=binomial,
                dataset=_last_,
                alpha=0.05, delta=0.10,
                variable=success, group=arm,
                value_e='a', value_s='b');
```

Before invoking the macro, a libname statement must be submitted to direct the macro to the proper data set or a data set must be created with which to use the macro (the default is the last data set created).

Assuming the data set has a variable `success' having a value of 1 for success and 0 for failure, and comparison groups of arm=a and arm=b, we may invoke the macro to analyze our binomial example by submitting the following statements using defaults for all but one of the parameters:

```
libname data `~path';
%dsanal(dataset=data.name);
```

Example output of the macro call follows:

TEST OF EQUIVALENCE

Experimental Sample Size 99

Standard Sample Size	100
Experimental P(success)	0.79
Standard P(success)	0.85
Std. Error for Success Rate Difference	0.054
Test Statistic	-0.74
Z(0.05)	-1.645
p-value	0.23
95% Confidence Interval Upper Limit	0.149

Result: Fail to reject the null hypothesis.
There is evidence to suggest that
the two treatments are non-equivalent.

Summary:

We test the hypotheses:
Ho: Ps-Pe >= 0.10
The two treatment success rates differ
by at least 10 %.
vs. H1: Ps-Pe < 0.10
The two treatment success rates differ
by no more than 10 %.

We will reject Ho for large negative values of the test statistic and will conclude that the two treatments are not significantly different, therefore equivalent.

ANALYSIS OF A GIVEN SET OF VALUES

The macro call parameters for the analysis of a given set of values are the significance level, the type II error, the clinical significance, the sample sizes, the probabilities of success, the means, the variances, and the Kaplan-Meier survival probabilities for follow-up time, accrual plus follow-up time, and half of the accrual plus follow-up time. The defaults have been set as follows:

```
%VARANAL(   datatype=binomial,
            alpha=0.05,
            beta=0.10, delta=0.10,
            ne=100, ns=100,
            pe=0.75, ps=0.80,
            meane=0, means=0,
            ve=0, vs=0,
            survf_e=0, survf_s=0,
            survaf_e=0, survaf_s=0,
            surv5afe=0, surv5afs=0);
```

This macro will run independently of other coding and doesn't depend upon a formerly defined SAS data set. To invoke to macro to analyze our continuous endpoints example, we would submit the following code using default values for the sample size, alpha, and delta (note: the other variables are not needed in our analysis, thus are ignored):

```
%VARANAL(   datatype=continuous,
            meane=2.61,
            means=2.70,
            ve=0.40,
            vs=0.36);
```

The output of the macro call:

TEST OF EQUIVALENCE

Experimental Sample Size	100
Standard Sample Size	100
Experimental Mean	2.61
Standard Mean	2.70
Std. Error for Mean Difference	0.087
Test Statistic	-2.069
Z(0.05)	-1.645
p-value	0.019
95% Confidence Interval Upper Limit	-0.233

Result: Reject the null hypothesis.
There is evidence to suggest that
the two treatments are equivalent.

Summary:

We test the hypotheses:
Ho: Means-Meane >= 0.27
The two treatment means differ by at least
10 % of the standard mean.
vs. H1: Means-Meane < 0.27
The two treatment means differ by no more than
10 % of the standard mean.

We will reject Ho for large negative values of the test statistic and will conclude that the two treatments are not significantly different, therefore equivalent.

SAMPLE SIZE DETERMINATION

The macro call parameters for the sample size determination are the data type, the significance level, the clinical significance, the probabilities of success, the means, the probabilities of failure, the variances, and the standard error. Defaults have been set as follows:

```
%SAMPSIZE(  datatype=binomial,
            alpha=0.05,
            delta=0.10, omega=0,
            pe=0.75, ps=0.80,
            meane=0, means=0,
            pe_fail=0, ps_fail=0,
            ve=0, vs=0,
            se=0);
```

This macro will also run independently of other coding and doesn't depend upon a formerly defined SAS data set. To determine sample size for survival data in a test having significance level α and power(1-β) = 0.90, the following code is submitted:

```
%SAMPSIZE(datatype=survival,
            delta=2.73);
```

The output of the macro call:

SAMPLE SIZE FOR THE EQUIVALENCE TEST

Specifications:
Alpha = 0.05
Beta = 0.10
Delta = 2.73

Result: Required total number of events = 34
Required patients per group = 141

Summary:
To rule out the proportion of no more than delta = 2.73 between hazard rates, will require 141 patients per group, resulting in a total of 242 patients.

SUMMARY

In this paper we have provided SAS macros for convenient statistical analysis of binomial, continuous, and censored survival data in equivalence tests. Our purpose was to provide a comprehensive framework and supporting SAS code for researchers to carry out the basic analysis required for an equivalence trial.

The SAS code used for the macros is available from the first author via e-mail at skaff.pamela@mayo.edu.

REFERENCES

Blackwelder, WC "`Proving the null hypothesis' in clinical trials." Controlled Clinical Trials, 1982; 3: 345-353.

Blackwelder, WC `Equivalence trials.' In press: Encyclopedia of Biostatistics, 1997.

Jennison, C and Turnbull, BW `Sequential equivalence testing and repeated confidence intervals, with applications to normal and binary responses.' Biometrics, 1993; 49(1): 31-43.

Com-Nougue, C, Rodary, C and Patte, C `How to establish equivalence when data are censored: a randomized trial of treatments for B non-hodgkin lymphoma.' Statistics in Medicine, 1989; 8: 593-598.

Sheiner, LB `Bioequivalence revisited.' Statistics in Medicine, 1992; 11: 1777-1788.

ACKNOWLEDGEMENTS

We would like to thank William Blackwelder, National Institute of Health, for his support and suggestions in the development of this paper.