Paper 4641-2020

Testing Hypotheses for Equivalence and Non-inferiority with Binary and Survival Outcomes

Joseph C. Gardiner, Department of Epidemiology and Biostatistics,
Michigan State University, East Lansing, MI 48824

ABSTRACT

The classic superiority test of comparison of two treatment groups seeks to show that they differ on a measure of their efficacy. However, when a new treatment has a similar therapeutic action to an existing standard treatment, there might be little difference in their efficacy. An equivalence study is designed to show that the difference in a measure of efficacy between the new and standard treatments is within a pre-specified margin of clinical indifference. In another context, the new treatment may offer lower cost and/or better patient compliance but might have a lower efficacy than the standard treatment. A non-inferiority study is designed to show that the new treatment is not less effective than the standard treatment to within a pre-specified margin of clinical indifference. In this paper, we discuss the formulation of tests of hypotheses for equivalence and non-inferiority studies in the context of a two-sample design for binary and survival outcomes. For a binary outcome, the comparison between the new and standard treatments can be assessed by the difference in the probability of response to treatment, the relative risk or odds ratio. For a survival outcome, the comparative assessment can be made using survival probabilities or where appropriate, the hazard ratio. SAS POWER and FREQ procedures offer options for performing the tests of hypotheses, and assessment of statistical power and sample size for conducting equivalence and non-inferiority studies.

INTRODUCTION

Consider the situation where two groups, called here treatment (*T*) and control (*C*) are compared on a response or outcome variable whose distribution has a parameter θ. For a continuous outcome θ is typically the mean response, whereas when the outcome variable is binary, θ is a probability.

Denoting by θ_T, θ_C the group-specific parameters, the objective of a superiority test is to assesses if θ_T, θ_C are different. Formally, we test $H_0 : \theta_T = \theta_C$ versus $H_1 : \theta_T \neq \theta_C$. If the alternative H_1 is tenable, then the two groups differ in their mean responses, and if 'bigger is better' then either treatment is better than control $(\theta_T > \theta_C)$ or control is better than treatment $(\theta_T < \theta_C)$. If the null H_0 cannot be rejected, then the groups do not differ in their mean responses. The classic two-sample t-test of H_0 versus H_1 is based on independent normally distributed samples from the treatment and control populations.

There are circumstances where we do not expect the groups to differ substantively. For example, when a new drug *T* has a similar formulation as an existing drug *C,* a small difference in θ_T, θ_C would be clinically inconsequential. An equivalence study is designed to show the difference in θ_T, θ_C is within a pre-specified margin Δ (>0) of clinical indifference. Formally, we test

$H_0 : \theta_T - \theta_C \le -\Delta$ *or* $\theta_T - \theta_C \ge \Delta$ versus $H_1 : -\Delta < \theta_T - \theta_C < \Delta$. If the alternative H_1 is tenable, we would claim that there is no substantive difference in mean responses of the two drugs because the difference is within the margin of indifference. Hence the drugs are declared equivalent. Instead of the symmetric indifference zone $(-\Delta, \Delta)$ we may specify it has $(-\Delta_L, \Delta_U)$ with lower and upper margins.

A third scenario arises when the new drug *T* is expected to be no worse than the existing standard drug *C*. For example, drug *T* could offer better patient compliance, lower cost and fewer side-effects but might have a lower efficacy than the standard drug *C*. Thinking of a positive difference $\theta_T - \theta_C > 0$ as being favorable to *T* relative to *C*, a non-inferiority study tests $H_0 : \theta_T - \theta_C \le -\Delta$ versus $H_1 : \theta_T - \theta_C > -\Delta$ where $\Delta > 0$ is a pre-specified non-inferiority margin. If the alternative H_1 is tenable, we would claim that *T* is not inferior than *C*.

The terms equivalence and non-inferiority apply to different contexts and these names are synchronous with their respective alternative hypothesis H_1. However, the term *equality test* is used to describe the test of $H_0 : \theta_T = \theta_C$ versus $H_1 : \theta_T \ne \theta_C$. This makes sense when referring to the null H_0. When reference is to the alternative hypothesis H_1 the term *inequality test* is apropos. SAS reserves the term *superiority* for describing a test complementary to non-inferiority. The *superiority test* is cast as $H_0 : \theta_T - \theta_C \le \Delta$ versus $H_1 : \theta_T - \theta_C > \Delta$ where $\Delta > 0$ is a pre-specified superiority margin. This is no different than non-inferiority testing with the margin $-\Delta$ replaced by Δ.

The focus of this paper is on equivalence and non-inferiority tests comparing two groups, treatment to control. Analysis is based on independent samples from each group. When the outcome is binary, θ_T, θ_C are relabeled as p_T, p_C for the respective probabilities of a favorable outcome. Tests can be formulated in terms of

(i) risk difference, $p_T - p_C$,

(ii) relative risk, p_T / p_C, or

(iii) odds ratio, $\{p_T / (1 - p_T)\} / \{p_C / (1 - p_C)\}$.

When the outcome is a time-to-event, we will use survival probabilities or where appropriate, the hazard ratio to formulate the tests.

This paper does not discuss an entirely analogous framework for equivalence and non-inferiority tests based on response means for continuous outcomes.

DATA LAYOUT -BINARY OUTCOME

From independent samples we have binomial counts in treatment (*T*) and Control (*C*): $Y_T \sim BIN(N_T, p_T)$ and $Y_C \sim BIN(N_C, p_C)$ with observed counts in the layout in Table 1. The row totals N_T, N_C are fixed by design.

TABLE 1: Data Layout

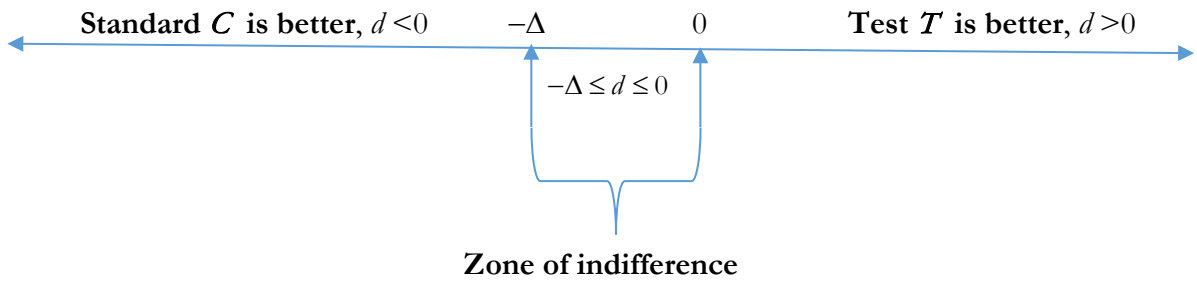
Group	Response (Column 1)	Non-response (Column 2)	Total
Treatment (Row 1)	n_{11}	n_{12}	$n_{1\cdot} = N_T$
Control (Row 2)	n_{21}	n_{22}	$n_{2\cdot} = N_C$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$n = N$

The favorable response is in column 1. The risk difference $d = p_T - p_C$ is row 1 minus row 2 estimated by sample proportions $\hat{p}_T = n_{11} / n_{1\cdot}, \hat{p}_C = n_{21} / n_{2\cdot}$. Appendix 1 provides an outline of the derivation of score tests for non-inferiority and equivalence. A less computationally heavy approach applies the Wald test. However, PROC FREQ carries out these tests.

Non-inferiority Tests based on the Risk Difference

The test is formulated as $H_0 : p_T - p_C \le -\Delta$ versus $H_1 : p_T - p_C > -\Delta$ with a pre-specified positive value for the margin Δ. Figure 1 shows where the test treatment T is better $d \equiv p_T - p_C > 0$; where the standard treatment C is better $d < 0$, and the indifference zone $-\Delta \le d \le 0$.

FIGURE 1: Regions of the Risk Difference



The test statistic is

$$(0) \qquad Z = \frac{\hat{p}_T - \hat{p}_C + \Delta}{\left[\frac{(\tilde{p} - \Delta)(\tilde{q} + \Delta)}{n_{1\cdot}} + \frac{\tilde{p}\tilde{q})}{n_{2\cdot}}\right]^{1/2}}$$

where $\tilde{p} \in (\Delta, 1)$ is the feasible solution to the cubic equation $Ap^3 + Bp^2 + Cp + D = 0$,

$$(2) \quad A = 1 + \theta, \ \ \theta = n_{1\cdot} / n_{2\cdot}, \ B = -\left[1 + \theta + \Delta(\theta + 2) + \theta\hat{p}_T + \hat{p}_C\right],$$
$$C = \left[\Delta^2 + \Delta\left(\theta + 1 + 2\hat{p}_C\right) + \theta\hat{p}_T + \hat{p}_C\right], \ D = -\Delta(1 + \Delta)\hat{p}_C.$$

The solution can be obtained from the POLYROOT function in PROC IML.

An α-level test rejects H_0 if the observed $Z > z_{1-\alpha}$ where $z_{1-\alpha}$ is the $100(1-\alpha)$ percentile of the standard normal distribution. If z^* denotes the computed value of Z from the data, the p-value of the test is $P[Z > z^*]$. The non-inferiority test is carried out in PROC FREQ as demonstrated in the following example.

EXAMPLE 1

Cornely et al, (2012) report on a multicenter double-blind randomized trial comparing fidaxomicin as alternative to vancomycin for treatment of *Clostridium difficile* infection. The two drugs have similar efficacy and safety. Patients were 16 years or older, with acute toxin-positive infection. Clinical cure, a binary outcome was defined as resolution of diarrhea and no further need for treatment. The pre-specified margin is 10% to assess non-inferiority of fidaxomicin (**Group** =1) compared to vancomycin (**Group** =2). In the per-protocol analysis, 198 of 216 patients treated with fidaxomicin achieved clinical cure by end of the monitoring of the study, compared to 213 of 235 patients treated with vancomycin. The input data set is created by

```
data Mycin;
length group $5.;
input Group response count total@;
 do i=1 to 2;
if i=1 then output;
if i=2 then do; response=0; count=total-count;
output;
 end;
 end;
datalines;
FIDAX 1 198 216
VANCO 1 213 235
;
run;

proc format;
value response 1='yes' 0='no';
run;
```

The entire non-inferiority analysis is called by

```
proc freq data=mycin order=data;
tables group*response/riskdiff(noninf margin=.1 method=fm);
weight count;
format response response.;
run;
```

If is important to specify the MARGIN= and METHOD= options, because the defaults are the value Δ=0.2, and the Wald method, described later in Table 3. First, we must be assured that the 2×2 table has the test treatment (fidaxomicin) and favorable response (=1) in the north-west cell.

Frequency Row Percent	Table of group by response		
group	**response**		
	yes	**no**	**Total**
FIDAX	198 91.67	18 8.33	216
VANCO	213 90.64	22 9.36	235
Total	411	40	451

The RISKDIFF options request a non-inferiority test, margin Δ=.10 and the Farrington-Manning (1990) score test (`method=fm).` These options carry out the computations in equations (1), (2) and implements the test.

TABLE 2: Non-inferiority Analysis based on the Risk Difference

H0: P1 – P2 <= –Margin Ha: P1 – P2 > –Margin						
Margin = 0.1 Score (Farrington–Manning) Method						
Risk Difference	**ASE (F–M)**	**Z**	**Pr > Z**	**Noninferiority Limit**	**90% Confidence Limits**	
0.0103	0.0296	3.7207	<.0001	–0.1000	–0.0385	0.0590

The test is significant indicating that we may declare non-inferiority. A test-based two-sided 90% confidence interval (CI), (–0.0385, 0.0590) is calculated as $\left(0.013-z_{.0.95}\times 0.0296, 0.013+z_{.0.95}\times 0.0296\right)$. The CI lies within the non-inferiority limit (–0.10, 1).

The Farrington-Manning approach is the preferred choice for non-inferiority tests. Other options are available in PROC FREQ which affect the construction of the test statistic, notably how the variance $Var(\hat{p}_T-\hat{p}_C)$ is estimated. The default is the Wald test statistic. A continuity correction c may be requested that improves the large-sample approximation to the normal distribution of the test statistic All test-statistics are constructed as $Z=\dfrac{\hat{p}_T-\hat{p}_C+\Delta-c}{\sqrt{Var(\hat{p}_T-\hat{p}_C)}}$.

A summary of these options is shown in Table 3. They are placed in RISKDIFF(*options*). By default, each analysis produces a two-sided 90% confidence interval with lower limit $\hat{p}_T-\hat{p}_C-c-z_{.95}\times\sqrt{Var(\hat{p}_T-\hat{p}_C)}$ and upper limit $\hat{p}_T-\hat{p}_C+c+z_{.95}\times\sqrt{Var(\hat{p}_T-\hat{p}_C)}$.

TABLE 3: Methods for Non-inferiority Tests

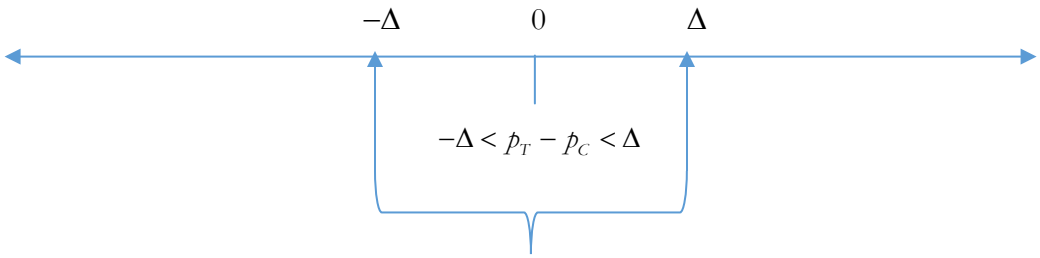
RISKDIFF (*options*)	$Var(\hat{p}_T - \hat{p}_C)$		**Correction**, c
noninf margin= method=wald var=sample	$\frac{\hat{p}_T\hat{q}_T}{n_{1\cdot}} + \frac{\hat{p}_C\hat{q}_C}{n_{2\cdot}}$		none
noninf margin= method=wald var=sample correct	$\frac{\hat{p}_T\hat{q}_T}{n_{1\cdot}} + \frac{\hat{p}_C\hat{q}_C}{n_{2\cdot}}$		$\frac{1}{2}\left(\frac{1}{n_{1\cdot}} + \frac{1}{n_{2\cdot}}\right)$
noninf margin= method=wald var=null	$\frac{(\bar{p}-\Delta)(\bar{q}+\Delta)}{n_{1\cdot}} + \frac{\bar{p}\bar{q}}{n_{2\cdot}}$	$\bar{p} = \frac{n_{11} + n_{21} + \Delta n_{1\cdot}}{n}$	none
noninf margin= method=wald var=null correct	$\frac{(\bar{p}-\Delta)(\bar{q}+\Delta)}{n_{1\cdot}} + \frac{\bar{p}\bar{q}}{n_{2\cdot}}$	$\bar{p} = \frac{n_{11} + n_{21} + \Delta n_{1\cdot}}{n}$	$\frac{1}{2}\left(\frac{1}{n_{1\cdot}} + \frac{1}{n_{2\cdot}}\right)$
noninf margin= method=HA correct	$\frac{\hat{p}_T\hat{q}_T}{n_{1\cdot}-1} + \frac{\hat{p}_C\hat{q}_C}{n_{2\cdot}-1}$		$\frac{1/2}{\min(n_{1\cdot}, n_{2\cdot})}$

HA=Hauck-Anderson (1986); $\hat{q}_T = 1 - \hat{p}_T$, $\hat{q}_C = 1 - \hat{p}_C$; $\bar{q} = 1 - \bar{p}$

Equivalence Tests based on Risk Difference

The test is formulated as $H_0 : |\, p_T - p_C\,| \geq \Delta$ versus the alternative $H_1 : |\, p_T - p_C\,| < \Delta$ with a pre-specified margin Δ. Rewrite as two one-sided tests (TOST): $H_{0a} : p_T - p_C \leq -\Delta$ versus $H_{1a} : p_T - p_C > -\Delta$ *and* $H_{0b} : p_T - p_C \geq \Delta$ versus $H_{1b} : p_T - p_C < \Delta$. Rejection of *both* H_{0a} and H_{0b} would mean that the risk difference lies in the zone of equivalence, that is, $-\Delta < p_T - p_C < \Delta$ as in Figure 2.

FIGURE 2: Zone of equivalence



Zone of equivalence

The first component H_{0a} versus H_{1a} is the previously discussed non-inferiority test. Apply the test statistic Z in equation (1) with equation (2). We reject H_{0a} if $Z > z_{1-\alpha}$. Note that it is α, not $\frac{1}{2}\alpha$. For the second component H_{0b} versus H_{1b} the test statistic is Z' defined by equations (3) and (4).

$$(3) \qquad Z' = \frac{\hat{p}_T - \hat{p}_C - \Delta}{\left[\frac{(\tilde{p}+\Delta)(\tilde{q}-\Delta)}{n_{1\cdot}} + \frac{\tilde{p}\tilde{q}}{n_{2\cdot}}\right]^{1/2}}$$

where $\tilde{p} \in (0, 1-\Delta)$ is the feasible solution to the cubic equation $A\tilde{p}^3 + B\tilde{p}^2 + C\tilde{p} + D = 0$

$$(4) \quad A = 1+\theta,\ \theta = n_{1\cdot}/n_{2\cdot},\ B = -\left[1+\theta-\Delta(\theta+2)+\theta\hat{p}_T+\hat{p}_C\right]$$
$$C = \left[\Delta^2 - \Delta\left(\theta+1+2\hat{p}_C\right)+\theta\hat{p}_T+\hat{p}_C\right],\ D = \Delta(1-\Delta)\hat{p}_C$$

Reject H_{0b} if observed $Z' < -z_{1-\alpha}$.

The overall p-value of the equivalence test is $P[Z > z, Z' < z']$ where z, z' are the observed values of the test statistics. But $P[Z > z, Z' < z'] \le \min\left(P[Z > z], P[Z' < z']\right)$. The overall p-value is taken as the larger of the two p-values $P[Z > z], P[Z' < z']$.

This Farrington-Manning score test is implemented in PROC FREQ with the METHOD=FM option. Because there are two standard errors computed in the denominators of Z and Z', a test-based 100(1−2α)% confidence interval for the risk difference (default α=0.05) applies the larger of these two standard errors. Evidence of equivalence is declared if the confidence interval lies wholly within the equivalence zone (−Δ, Δ).

Returning to Example 1, we will assess the equivalence of fidaxomicin and vancomycin. The margin is set at Δ=0.10. The entire equivalence analysis is called by

```
proc freq data=mycin order=data;
tables group*response/riskdiff(equivalence margin=.1 method=fm);
weight count;
format response response.;
run;
```

TABLE 4: Equivalence Analysis based on the Risk Difference

H0: P1 – P2 <= Lower Margin or >= Upper Margin					
Ha: Lower Margin < P1 – P2 < Upper Margin					
Lower Margin = –0.1 Upper Margin = 0.1 Score (Farrington–Manning) Method					
Risk Difference	**ASE (F–M)**	**Equivalence Limits**		**90% Confidence Limits**	
0.0103	0.0296	–0.1000	0.1000	–0.0385	0.0590

Two One-Sided Tests (TOST)			
Test	**Z**	**P-Value**	
Lower Margin	3.7207	**Pr > Z**	<.0001
Upper Margin	–3.1614	**Pr < Z**	0.0008
Overall			0.0008

The Lower Margin Test is $H_{0a}: p_T - p_C \le -\Delta$ versus $H_{1a}: p_T - p_C > -\Delta$. The Upper Margin Test is $H_{0b}: p_T - p_C \ge \Delta$ versus $H_{1b}: p_T - p_C < \Delta$. The ASE(F-M) refers to the larger of the two standard errors. A separate calculation shows a standard error SE=0.0284 (the denominator of Z'), whereas SE=0.0296 (the denominator of *Z*). The p-value of the test is the larger of the two p-values—here 0.0008. Because the 90% CI (–0.0385, 0.0590) lies wholly within the equivalence zone (–0.10, 0.10), we would declare equivalence.

General Comments

Other methods for equivalence testing use similar test statistics as in Table 3, *but* we must carry out two one-sided tests (TOST) of significance (Schuirmann, 1987). Also, note that for an α-level test the critical values are $\pm z_{1-\alpha}$, that is, α is not halved. Terms such as upper right-sided test and lower-left sided test, borrow from the direction of the alternative hypotheses.

An asymmetric equivalence zone, (Δ_L, Δ_U) is specified as MARGIN=(Δ_L, Δ_U) where $\Delta_L<\Delta_U$. It is good practice to state the margin because the default Δ_L= –0.2, Δ_U= +0.2 may not apply universally to all applications.

Non-inferiority Tests based on the Relative Risk

Using the relative risk $p_T / p_C = \rho$, the non-inferiority test is $H_0: \rho \le 1-\Delta$ versus $H_1: \rho > 1-\Delta$ where 0<Δ<1. PROC FREQ uses $\rho_0 = 1-\Delta$ as the MARGIN. By default, MARGIN=0.8 . For the favorable response, comparing test treatment to control (standard), we should set ρ_0 high. The Farrington-Manning score test statistic is

$$(5) \qquad Z = \frac{\hat{p}_T - \rho_0 \hat{p}_C}{\left[\frac{\tilde{p}_T(1-\tilde{p}_T)}{n_{1\cdot}} + \rho_0^2 \frac{\tilde{p}_C(1-\tilde{p}_C)}{n_{2\cdot}}\right]^{1/2}}$$

where $\tilde{p}_T = \rho_0 \tilde{p}_C$ and $\tilde{p}_C$ is the solution to the quadratic equation $Ap^2 + Bp + C = 0$. The feasible solution is $\tilde{p}_C = \frac{-B - \sqrt{B^2 - 4AC}}{2A}$, where

$$(6) \qquad A = \rho_0(1+\theta),\ \theta = n_{1\cdot}/n_{2\cdot},\ B = -\left[1 + \theta\hat{p}_T + \rho_0(\theta + \hat{p}_C)\right],\ C = \theta\hat{p}_T + \hat{p}_C .$$

We reject $H_0: \rho \le \rho_0$ if $Z > z_{1-\alpha}$, for an α-level test. The p-value is $P[Z > z^*]$ where z^* is the computed value of *Z* from the data.

A test-based 100(1–α)% confidence interval is generated as the region $\{\rho_0 : Q(\rho_0) < c\chi^2_{1-\alpha}\}$ with $Q(\rho_0) = Z^2$ as a function of ρ_0 and $c = n/(n-1)$ a bias reduction factor that can be optionally suppressed. The default is α=0.10.

EXAMPLE 2

For illustration we use the study described in Example 1, but data from the modified intent-to-treat (ITT) analysis comparing fidaxomicin as alternative to vancomycin for treatment of *Clostridium difficile* infection. Here 221/252 patients treated with fidaxomicin achieved clinical cure by the end of the monitoring of the study, compared to 223/257 patients treated with vancomycin. Create a data set (**`mycin_mITT`**) to generate the 2×2 table:

Frequency Row Pct	Table of group by response		
group	response		
	yes	no	Total
FIDAX	221 87.70	31 12.30	252
VANCO	223 86.77	34 13.23	257
Total	444	65	509

With the relative risk as metric, and a margin=0.9 (= ρ_0) the non-inferiority analysis is called by

```
proc freq data=mycin_mITT order=data;
tables group*response/relrisk(noninf margin=.9 method=fm);
weight count;
format response response.;
run;
```

TABLE 5: Noninferiority Analysis based on Relative Risk

H0: P1 / P2 <= Margin Ha: P1 / P2 > Margin						
Margin = 0.9 Score (Farrington-Manning) Method						
Relative Risk	ASE (F-M)	Z	Pr > Z	Noninferiority Limit	90% Confidence Limits	
1.0107	0.0298	3.2236	0.0006	0.9000	0.9550	1.0699

The estimated relative risk is 1.011. The score test is significant, p-value<.001; we declare non-inferiority of fidaxomicin. The test-based 90% CI (0.955, 1.07) is contained in (0.90, +∞). To suppress the bias reduction factor apply

```
tables group*response/relrisk(CL=score(correct=no)) alpha=.10;
```

An option to perform the likelihood ratio test is available (**`method=LR`**) as well as the standard Wald test (**`method=WALD`**). The latter test statistic is $Z_W = [\log(\hat{p}_T / \hat{p}_C) - \log(\rho_0)] / \sqrt{v_W}$ where $v_W = n_{11}^{-1} + n_{21}^{-1} - n_{1\cdot}^{-1} - n_{2\cdot}^{-1}$ estimates the variance of $\log(\hat{p}_T / \hat{p}_C)$ .The results from either method do not change our conclusions.

Equivalence Tests based on Relative Risk

We carry out two one-sided tests of significance (TOST) for the pair of hypotheses,

$H_{0a}: \rho \le \rho_0$ versus $H_{1a}: \rho > \rho_0$ *and* $H_{0b}: \rho \ge \rho_1$ versus $H_{1b}: \rho < \rho_1$. Rejection of *both* H_{0a} and H_{0b} would mean that the relative risk lies in the zone of equivalence, that is, $\rho_0 < \rho < \rho_1$. In practice pre-specified equivalence limits would satisfy $\rho_0 < 1 < \rho_1$. The defaults are $\rho_0 = 0.8$, $\rho_1 = 1.25$.

The syntax below carries out the score test, with results shown in Table 6.

```
proc freq data=mycin_mITT order=data;
tables group*response/relrisk(equivalence margin=(.90, 1.10)
                              method=FM);
weight count;
format response response.;
run;
```

TABLE 6: Equivalence Analysis based on the Relative Risk

H0: P1 / P2 <= Lower Margin or >= Upper Margin				
Ha: Lower Margin < P1 / P2 < Upper Margin				
Lower Margin = 0.9 Upper Margin = 1.1 Score (Farrington-Manning) Method				
Relative Risk	**Equivalence Limits**		**90% Confidence Limits**	
1.0107	0.9000	1.1000	0.9550	1.0699

Two One-Sided Tests (TOST)				
Test	**ASE (F-M)**	**Z**	**P-Value**	
Lower Margin	0.0298	3.2236	**Pr > Z**	0.0006
Upper Margin	0.0323	–2.4024	**Pr < Z**	0.0081
Overall				0.0081

Much of the analysis is already shown in Table 5 as a part of the non-inferiority analysis. The overall p-value for the equivalence test is the larger of the two p-values for the lower margin test (right-hand side), and upper margin test (left-hand side).

Non-inferiority Tests based on the Odds Ratio

In terms of the odds ratio (OR), $\omega = \dfrac{p_T / q_T}{p_C / q_C}$, the non-inferiority test is $H_0: \omega \le \omega_0$ versus $H_1: \omega > \omega_0$ where $\omega_0 < 1$ is specified. Here $q_T = 1 - p_T$, $q_C = 1 - p_C$. The score test is derived in the usual way and leads to the test statistic

$$(7) \qquad Z = \frac{\left[\frac{(\hat{p}_T - p_T)}{p_T q_T} - \frac{(\hat{p}_C - p_C)}{p_C q_C}\right]}{\left[\frac{1}{n_1 . p_T q_T} + \frac{1}{n_2 . p_C q_C}\right]^{1/2}}.$$

To be operational, replace p_C by an estimator $\tilde{p}_C$ and take $\tilde{p}_T = \frac{\omega_0 \tilde{p}_C}{1+(\omega_0 - 1)\tilde{p}_C}$, where $\tilde{p}_C = \frac{-B + \sqrt{B^2 - 4AC}}{2A}$ is the feasible solution to the quadratic equation $Ap^2 + Bp + C = 0$, and

$$(8) \qquad A = n_2 . (\omega_0 - 1), B = n_1 . \omega_0 + n_2 . - (\omega_0 - 1)(n_{11} + n_{21}), C = -(n_{11} + n_{21}).$$

For an α-level test we reject $H_0 : \omega \le \omega_0$ if the observed $Z > z_{1-\alpha}$. The p-value of the test is $P[Z > z^*]$ where z^* is the observed value of the test statistic.

The formulation with the odds ratio is like the previous discussion using the relative risk. Equations (7) and (8) replace equations (5) and (6). However, PROC FREQ does not have an option for testing using the odds ratio unlike the RELRISK, but confidence intervals for the OR by the score method can be obtained from

```
tables group*response/oddsratio(CL=score);
```

This constructs a two-sided 100(1−α)% confidence interval for the odds ratio ω as the region $\{\omega_0 : Q(\omega_0) < c\chi^2_{1-\alpha}\}$ with $Q(\omega_0) = Z^2$ as a function of ω_0 and $c = n/(n-1)$ a bias reduction factor that can be optionally suppressed by `oddsratio(CL=score(correct=no));`

An equivalence test is formulated as TOST: $H_{0a} : \omega \le \omega_0$ versus $H_{1a} : \omega > \omega_0$ *and* $H_{0b} : \omega \ge \omega_1$ versus $H_{1b} : \omega < \omega_1$ where the margins satisfy $\omega_0 < 1 < \omega_1$.

EXAMPLE 3

Wellek (2010, Table 6.26 page 191) summarizes results from a randomized trial of a calcium channel blocker verapamil and a classic diuretic drug in patients with mild to moderate hypertension (Holzgreve et al, 1989). Outcomes were achieving a target diastolic blood pressure response (<90 mm Hg) at different time points in the course of the trial. The data used here comprise the response to any trial medication at 8 weeks, with patients stratified by previous treatment with anti-hypertensive drugs. In 225 patients who had previous use of anti-hypertensive drugs (Group POS), the favorable response was achieved in 108 patients, compared to 63 of 119 patients who had no previous use (Group NEG). Our objective is to assess equivalence of the two groups with response to treatment. From the cell counts in the VERDI data set the estimated OR is 0.8205.

Frequency Row Pct	Table of Group by response		
Group	**response**		
	yes	**no**	**Total**
POS	108 48.00	117 52.00	225
NEG	63 52.94	56 47.06	119
Total	171	173	344

Carrying out an equivalence test with margins $\omega_0 = 0.6667$ and $\omega_1 = 1.500$ is accomplished by performing the calculations for (7) and (8) in a data step, for ω_0 (upper test) and for ω_1 (lower test). The overall p-value (larger of the two p-values) is 0.18. We cannot reject the null hypothesis and declare equivalence. A 90% confidence interval for the OR with the bias reduction factor applied, is (0.565, 1.192). This interval is not contained within (ω_0, ω_1).

Conditional test based on the odds ratio

Return to the data layout in Table 1. The row totals are fixed by design: in theory we have binomial response counts in treatment (T) and control (C): $Y_T \sim BIN(N_T, p_T)$ and $Y_C \sim BIN(N_C, p_C)$.

From elementary probability the conditional distribution of Y_T given $Y = Y_T + Y_C = n_{\bullet 1}$ is

$$(9) \qquad P[Y_T = n_{11} \mid Y = n_{\bullet 1}] = \frac{\binom{N_T}{n_{11}}\binom{N_C}{n_{\bullet 1} - n_{11}}\omega^{n_{11}}}{\sum_u \binom{N_T}{u}\binom{N_C}{n_{\bullet 1} - u}\omega^u}$$

on the domain $\max(0, n_{\bullet 1} - N_C) \le n_{11} \le \min(N_T, n_{\bullet 1})$ where the summation in the denominator is over this domain. The discrete distribution is called Fisher's Univariate Noncentral Hypergeometric Distribution, or Conditional (Extended) Hypergeometric Distribution. It is the engine that drives calculations of Fisher's exact test of homogeneity $H_0 : \omega = 1$ in PROC FREQ. SAS function PDF("HYPER", $n_{11}, N, N_T, n_{\bullet 1}, \omega$) and the corresponding CDF and QUANTILE functions can be applied to carry out an equivalence test. As with all multi-argument SAS functions, the order of the arguments is critically important.

EXAMPLE 3 (CONTINUED)

Use the margins $\omega_0 = 0.6667$ and $\omega_1 = 1.500$ with observed n_{11}=108, $n_{\bullet 1}$=171. The lower test $H_0 : \omega \ge \omega_1$ *vs* $H_1 : \omega < \omega_1$ has p-value $P[Y_T \le n_{11} \mid Y = n_{\bullet 1}, \omega_1]$ <.0006 and the upper test $H_0 : \omega \le \omega_0$ *vs* $H_1 : \omega > \omega_0$ has p-value $P[Y_T \ge n_{11} \mid Y = n_{\bullet 1}, \omega_0]$ =0.2099.

Because the discrete conditional hypergeometric distribution is used, in general we cannot guarantee an exact α-level significance test. Wellek (2010) has an extensive discussion on constructing a randomized decision rule for the equivalence test based on the odds ratio, with the objective of having an exact α-level test. The rejection rule for the composite hypothesis, $H_0 : \omega \le \omega_0 \; or \; \omega \ge \omega_1$, rejects H_0 if $k_1 < Y_T < k_2$, and with probability γ_1 if $Y_T = k_1$, with probability γ_2 if $Y_T = k_2$. We accept H_0 if $Y_T < k_1 \; or \; Y_T > k_2$. Note the strict inequalities. The constants $k_1, k_2, \gamma_1, \gamma_2$ depend on (α, ω_0, ω_1 $N_T, N_C, n_{\cdot 1}$). We need to solve

$$\gamma_1 P[Y_T = k_1 \mid \omega_0] + \gamma_2 P[Y_T = k_2 \mid \omega_0] + \sum_{u=k_1+1}^{k_2-1} P[Y_T = u \mid \omega_0] = \alpha \text{ and}$$

$$\gamma_1 P[Y_T = k_1 \mid \omega_1] + \gamma_2 P[Y_T = k_2 \mid \omega_1] + \sum_{u=k_1+1}^{k_2-1} P[Y_T = u \mid \omega_1] = \alpha$$

with the probabilities in (9). The solution is $k_1 = 110, k_2 = 113$, γ_1 =.92315, γ_2 =.88134. Wellek provides SAS programs for calculations. In real-world applications, randomized rules are not widely used.

A 90% exact confidence interval for the OR is obtained from

```
proc freq data=VERDI order=data;
tables GROUP*response/oddsratio(CL=exact)alpha=.1;
exact OR;
*tables GROUP*response/oddsratio(CL=score) alpha=.1;
*tables GROUP*response/oddsratio(CL=score(correct=no)) alpha=.1;
*tables GROUP*response/oddsratio(CL=WALD) alpha=.1;
weight count;
format response affirm.;
run;
```

The EXACT statement is required for **CL=exact.** Because the discrete conditional hypergeometric distribution is applied to obtain the exact CL, the confidence level is at least 90% , i.e., the interval is conservative. Three other types of confidence intervals in Table 7, including the traditional WALD are based on asymptotic distribution of the odds ratio estimator.

TABLE 7: Confidence Limits for the Odds Ratio

Odds Ratio = 0.8205		
Type	**90% Confidence Limits**	
Exact	0.5511	1.2217
Score	0.5649	1.1918
Score (correct=no)	0.5652	1.1911
Wald	0.5648	1.1919

DISCUSSION

Our focus has been on non-inferiority and equivalence tests for a binary endpoint based on two independent samples. We framed the tests based on the risk difference, relative risk and odds ratio (Chowdhury et al, 2019a) applying the Farrington-Manning (1990) score test and options available in PROC FREQ. If the large sample normal approximation to the test statistic is not tenable, we could use exact procedures based on the binomial distribution (Wellek, 2010), or at least make a continuity adjustment to the normal approximation.

Methods for assessing power and sample size are addressed in PROC POWER for the risk difference and relative risk but not directly for the odds ratio. Castelloe and Watts (2015) and SAS Usage Notes describe the procedures with applications. They also provide details of non-inferiority and equivalence tests for a continuous endpoint where the mean difference or mean ratio is the effect measure for comparison between two groups. For analyses, PROC TTEST is the natural choice.

Several authors have addressed tests for equivalence and non-inferiority in matched-pair designs (Nam, 1997, Tang et al, 2007, Tsong et al, 2013) and crossover designs (Li et al, 2016), framing the procedures by score and likelihood ratio tests. Sidik (2003) addresses exact unconditional tests for the matched-pairs design.

Equivalence and non-inferiority studies comparing a new drug (T) to a control drug (C) require specification of the indifference margin Δ. An appropriate value is based on clinical considerations and statistical evidence drawn from historical studies of C in placebo-controlled studies that provide information of the effects θ_C, θ_P in C and placebo P. Regulatory agencies such the FDA and EMA may include public health needs in arriving at a judicious value of Δ. Because the control condition is an active standard, Δ should be smaller than the difference in response seen between active control and placebo, for example $\Delta \leq ½(\theta_C - \theta_P)$. The synthesis-margin approach attempts to incorporate this evidence in the current non-inferiority test or corresponding equivalence test, by using the estimates of θ_C, θ_P and their variances from studies comparing C to P. See Chang (2011, Chapter 3) and Liao (2015) for context and examples on how to use some fraction of the evidence from the comparison in specifying Δ for the current study of T and C.

Finally, a recent paper discusses a three-arm study for non-inferiority using the risk ratio and odds ratio (Chowdhury et al, 2019b). Equivalence testing procedures for continuous endpoints from independent K-samples, both parametric and nonparametric, as well as from dependent multivariate samples are discussed by Wellek (2010) Chapters 7 and 8 together with SAS programs for carrying out some daunting calculations. For the statistical analyst the realm of software programs for non-inferiority and equivalence testing is somewhat limited. Enhancements to the procedures FREQ and POWER are forthcoming that would extend their capabilities to handle more complex study designs.

Non-inferiority and Equivalence Tests for a Survival Endpoint

We adopt a separate notation for this section. In the context of a time-to-event, subject to right censoring, a generic datum on an individual is (T,δ) where δ is the indicator of whether T is the actual time to event $(\delta=1)$ or the censoring time $(\delta=0)$. Survival data comprise independent samples from two groups. Let S_1, S_2 denote the survival functions in group 1 (test) and group 2 (standard), respectively, with corresponding hazard functions h_1, h_2. With the single indicator $x=1$ for group 1, $x=0$ for group 2, the proportional hazards (PH) assumption is $h_1(t)=h_2(t)\exp(\beta x)$ for all t. This leads to $S_1(t)=\left(S_2(t)\right)^{\theta}$, where $\theta=\exp(\beta)$ is the hazard ratio (HR). Testing the hypothesis of equality, $H_0: S_1(t)=S_2(t),$ for all t, is entirely equivalent to $H_0: \beta=0$.

Without reference to the PH assumption, PROC LIFETEST performs the log-rank test and its variants (Klein & Moeschberger, 2003). PROC PHREG is dedicated to analysis of the PH model $h(t\,|\,\mathbf{x})=h_0(t)\exp(\mathbf{x}\boldsymbol{\beta})$ where $h_0(t)$ denotes a baseline hazard corresponding to covariates $\mathbf{x}=\mathbf{0}$.

Wellek (2010) citing Freitag (2005) and Freitag et al, (2006) develops an equivalence test and a non-inferiority test based on the distance metric: $\|S_1-S_2\|=\sup\{t>0:|S_1(t)-S_2(t)|\}$ under the PH assumption and continuity of the survival functions. The assumptions provide a neat translation from survival to the HR. Martinez et al, (2017) adopt a similar approach for the proportional odds survival model.

Let $\Delta>0$ denote a specified equivalence margin. The equivalence test is formulated as testing $H_0: \|S_1-S_2\|\geq\Delta$ versus $H_1: \|S_1-S_2\|<\Delta$. Under H_0, for *at least one* $t>0$, $|S_1(t)-S_2(t)|\geq\Delta$ whereas under H_1 we have $|S_1(t)-S_2(t)|<\Delta$ for *all* $t>0$. Hence the alternative says that the two survival functions are "equivalent" with respect to the maximal difference Δ.

Next, consider the non-inferiority test. Suppose survival could be slightly worse in group 1 than in group 2. With a margin $\Delta>0$ the non-inferiority test is formulated as testing $H_0: S_1(t)-S_2(t)\leq-\Delta$, for *at least one* $t>0$ versus $H_1: S_1(t)-S_2(t)>-\Delta$ *for all* $t>0$. Hence the alternative says that survival in group 1 is non-inferior to group 2. If $t=t_0$ is fixed, the analogy to the non-inferiority test of two proportions is not accidental (da Silva, Logan and Klein, 2008).

Because higher hazards are indicative of worse survival, under PH, the non-inferiority hypothesis in terms of the HR is stated as $H_0: \theta\geq\theta_0$ versus $H_1: \theta<\theta_0$ where $\theta_0>1$ is pre-specified.

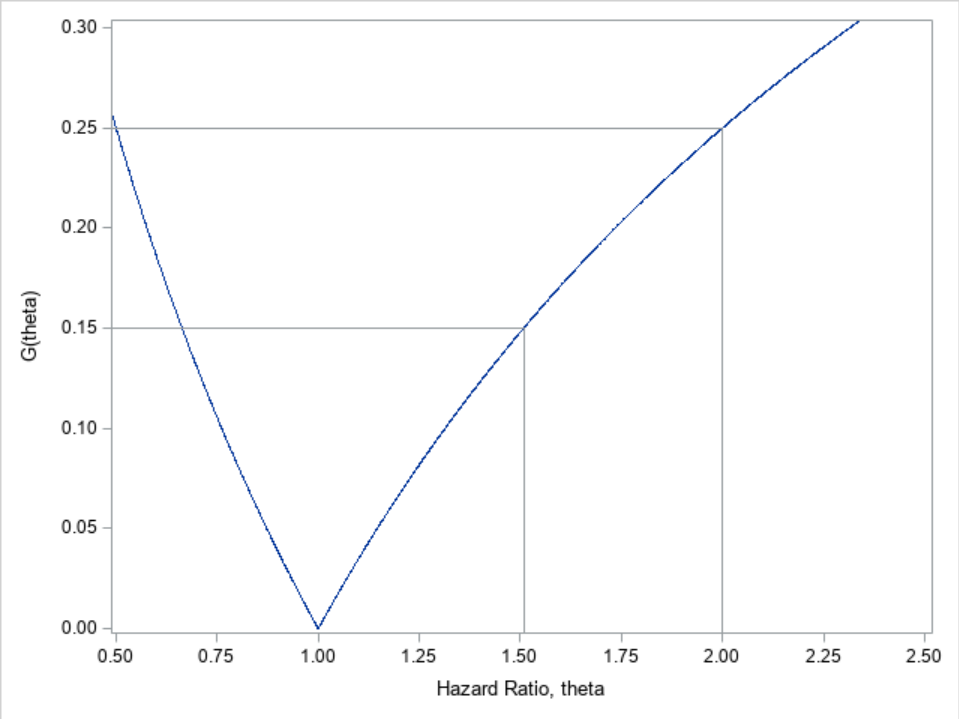
Continuity and the PH assumption ensure $\|S_1-S_2\|=\sup\{0<u<1:|u-u^{\theta}|\}$. Let $g(u)=u-u^{\theta}$, $u\in[0,1]$. The sign of $g(u)$ depends on θ; if $\theta<1$, $g(u)<0$; if $\theta>1$, $g(u)>0$. From calculus, $G(\theta)\equiv\sup\{|g(u)|:u\in[0,1]\}=\left|\theta^{(1-\theta)^{-1}}-\theta^{\theta(1-\theta)^{-1}}\right|$, as a function of $\theta\in(0,\infty)$. $G(\theta)$ is invariant under transformation $\theta\rightarrow\theta^{-1}$, strictly decreasing in (0, 1) and strictly increasing in (1,∞). See Figure 3.

Given the margin Δ, we can get the values θ satisfying $G(\theta)=\Delta$. With $\theta=(1+\varepsilon)$, $\varepsilon>0$, solve the equation $(1+\varepsilon)^{-\varepsilon^{-1}}-(1+\varepsilon)^{-(1+\varepsilon)\varepsilon^{-1}}=\Delta$, using PROC FCMP. For $\theta<1$, apply $\theta=(1+\varepsilon)^{-1}$.

```
proc fcmp;
 function equiv(t);
 Gt=(1+t)**(-1/t)-(1+t)**(-(1+t)/t);
return(Gt);
endsub;
do del=.05 to .30 by .05;
t=solve("equiv", {.}, del,.);
HR=1+t;
log_HR=log(HR);
put del= @10 t= @20 HR= @30 log_HR=;
end;
format del F5.2 t F6.4 HR F6.4 log_HR F6.4;
run;
```

```
del= 0.05 t=0.1457  HR=1.1457 log_HR=0.1360
del= 0.10 t=0.3135  HR=1.3135 log_HR=0.2727
del= 0.15 t=0.5077  HR=1.5077 log_HR=0.4106
del= 0.20 t=0.7341  HR=1.7341 log_HR=0.5505
del= 0.25 t=1.0000  HR=2.0000 log_HR=0.6931
del= 0.30 t=1.3149  HR=2.3149 log_HR=0.8394
```

FIGURE 3: Plot of $\theta \to G(\theta)$ with droplines at (1.51, 0.15) and (2.00, 0.25)



With these preliminary calculations, the equivalence test expressed in terms of β=log (HR) is

$H_0 : |\beta| \geq \log(1+\varepsilon)$ versus $H_1 : |\beta| < \log(1+\varepsilon)$.

The non-inferiority test (group 1 non-inferior to group 2) is expressed as $H_0 : \beta \geq \log(1+\varepsilon)$ versus $H_1 : \beta < \log(1+\varepsilon)$ because favorable survival in group 1 relative to group 2 means β <0, and thus smaller β is desirable.

Survival data for hypothesis testing

From the sample of survival data $\{(T_k, \delta_k, x_k) : 1 \leq k \leq N\}$ on individuals, let $t_1 < t_2 < \ldots < t_D$ denote the distinct *event* times in both groups; at time t_i in group j, d_{ij} = # events , n_{ij} = # at risk, i.e., the number of individuals who have not had the event or been censored prior to t_i. Total number of events $d_{i+} = d_{i1} + d_{i2}$, and total number at risk $n_{i+} = n_{i1} + n_{i2}$.

The score function $U_N(\beta)$ and information function $I_N(\beta)$ from the partial log likelihood are

$$U_N(\beta) = \sum_{i=1}^{D} \left\{ d_{i1} - d_{i+} \frac{n_{i1} \exp(\beta)}{\left(n_{i1} \exp(\beta) + n_{i2}\right)} \right\} \ , \ I_N(\beta) = \sum_{i=1}^{D} \left\{ \frac{d_{i+} n_{i1} n_{i2} \exp(\beta)}{\left(n_{i1} \exp(\beta) + n_{i2}\right)^2} \right\} = \sum_{i=1}^{D} d_{i+} f_i(\beta)(1 - f_i(\beta))$$

where $f_i(\beta) = n_{i1} \exp(\beta) / \{n_{i1} \exp(\beta) + n_{i2}\}$. The maximum (partial) likelihood estimator $\hat{\beta}$ is the solution to $U_N(\hat{\beta}) = 0$. If β is the true parameter, from large sample theory (Andersen et al, 1993), $\sqrt{N}(\hat{\beta} - \beta) \rightarrow Normal(0, \{\sigma^2(\beta)\}^{-1})$ and both $N^{-1} I_N(\beta)$ and $N^{-1} I_N(\hat{\beta})$ converge to $\sigma^2(\beta)$ in probability. These results are the basis for testing the equality hypothesis $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$.

When applied to a data set, PROC PHREG gives us the following:

$\hat{\beta}$, its standard error estimated by $\{I(\hat{\beta})\}^{-1/2}$, the score statistic $\chi_S^2 = \{U_N(0)\}^2 / I_N(0)$ and Wald statistic $\chi_W^2 = I_N(\hat{\beta})\hat{\beta}^2$. There is no easy way of generating the function $I_N(\beta)$ other than by a direct computation from the output statistics (for example, from PROC LIFETEST).

Consider the equivalence test (Wellek, 2010): Set the margin $\beta_0 \equiv \log(1+\varepsilon)$. An asymptotically valid α-level test rejects $H_0 : |\beta| \geq \beta_0$ if and only if $I_N^{1/2}(\hat{\beta}) |\hat{\beta}| > C_\alpha(I_N^{1/2}(\hat{\beta})\beta_0)$ where $C_\alpha(\psi)$ is the square-root of the 100α percentile of the χ^2 distribution with 1 degree of freedom and non-centrality ψ^2. The approximation $C_\alpha(\psi) \approx \psi - z_{1-\alpha}$, is excellent for even moderate ψ. Then re-express the test as

TOST: (*a*) Reject $H_{0a} : \beta \leq -\beta_0$ in favor of $H_{1a} : \beta > -\beta_0$ if $I_N^{1/2}(\hat{\beta})(\hat{\beta} + \beta_0) > z_{1-\alpha}$

(*b*) Reject $H_{0b} : \beta \geq \beta_0$ in favor of $H_{1b} : \beta < \beta_0$ if $I_N^{1/2}(\hat{\beta})(\hat{\beta} - \beta_0) < -z_{1-\alpha}$

The corresponding non-inferiority test is $H_0 : \beta \geq \beta_0$ versus $H_1 : \beta < \beta_0$ recognizing that lower hazards (log hazards) are desirable. In practice, (*b*) is the common formulation. When the upper confidence limit $\hat{\beta} + I_N^{-1/2}(\hat{\beta}) z_{1-\alpha}$ is below β_0 we may declare non-inferiority.

The ingredients to carry out the tests are obtained from PROC PHREG. The next example using summary results is offered purely for illustration.

EXAMPLE 4

Scagliotti et al, (2008) report on time to event outcomes in a randomized study of cisplatin plus pemetrexed (CP) against cisplatin plus gemcitabine (CG) in patients with advanced stage non-small-cell-lung cancer. CG is the widely used effective treatment (Referent); CP has similar efficacy, safety and a more convenient dose schedule. The objective was to demonstrate that CP is non-inferior to CG. We use the results summarized in the article for two endpoints: overall survival (OS) and progression-free survival (PFS) in over 1,670 patients.

The PH model, $h_1(t \mid x) = h_2(t)\exp(\beta x)$ with a binary covariate x (x=1 for CP, x=0 for CG) . From the published results on the HR and 95% CI, we extracted the estimate $\hat{\beta}$, the information $I_N(\hat{\beta})$ and constructed a 90% CI for β. The reported adjusted HR for covariates is used as a proxy. The calculations are summarized below.

Comparison, CP to CG	**Adjusted HR, 95% CI***	$\hat{\beta}$ **=Log(HR), 90% CI**	$I_N(\hat{\beta})$
OS, all patients	0.94, (0.84, 1.05)	–0.0619, (–0.1548, 0.0310)	313.669
PFS, all patients	1.04, (0.94, 1.15)	0.0392, (–0.0452, 0.1236)	380.021
OS, Squamous cell carcinoma	1.23, (1.00, 1.51)	0.2070, (0.0349, 0.3791)	91.324
PFS, Squamous cell carcinoma	1.36, (1.12, 1.65)	0.3075, (0.1453, 0.4697)	102.819

* Scagliotti et al, (2008), Fig 2.

The study assumed that CG would produce at least a 15% reduction in the risk of death over CP. It translates to a hazard ratio of CG vs CP=0.85. Invert to get the HR of CP vs CG, 1/0.85 =1.176. Hence, the non-inferiority hypothesis in terms of log(HR) is: $H_0 : \beta \geq \beta_0$ *vs* $H_1 : \beta < \beta_0$ where $\beta_0 = \log(1.176) = 0.16212$. From a calculation via PROC FCMP the maximal difference in the event-free survival functions of CP and CG is about 6% (i.e., Δ=0.06). With all patients analyzed, for both OS and PFS, we can declare non-inferiority of CP because the upper limit of the 90% CI is below β_0. The 5% level tests are significant (p<.009), so that H_0 can be rejected in favor of H_1.

Consider the endpoints in the subgroup of patients with squamous cell carcinoma. However, change the maximal difference in the two survival functions to 15%; Δ=0.15 corresponds to $\beta_0 = 0.4106$. For OS, we can declare non-inferiority (p<.025) but not for PFS where significantly *better* survival is demonstrated for CG over CP.

Calculations for non-inferiority and equivalence tests are carried out in a data step together with the separate TOST (*a*) and (*b*).

```
data PFS;
z975=QUANTILE("NORMAL",.975);
z95=QUANTILE("NORMAL",.95);
/*OS*/  HR=.94; LCL=.84; UCL=1.05;
/*PFS HR=1.04; LCL=.94; UCL=1.15;*/
b=log(HR);
STDERR=(log(UCL)-log(HR))/z975;
STDERR_=log(LCL/HR)/z975;
I=STDERR**(-2);
l_LCL90=b-z95*STDERR; l_UCL90=b+z95*STDERR;

b0=.1621;
Z_L=sqrt(I)*(b-b0);
Z_U=sqrt(I)*(b+b0);
pv_L=CDF("NORMAL", Z_L);
pv_U=SDF("NORMAL", Z_U);

psi=sqrt(I)*b0;
C=quantile("CHISQ",.05, 1, psi*psi);
C=sqrt(C);
z_crit=sqrt(I)*abs(b);
pv_chi=CDF("CHISQ", z_crit**2, 1, psi*psi);
run;
```

Remarks

(1) The tests for non-inferiority and equivalence described here are asymptotically valid α-level tests, that is, the size of the test does not exceed the nominal level α. The approximation $C_\alpha(\psi) \approx \psi - z_{1-\alpha}$ is excellent even for small ψ. From simulation studies Martinez et al, (2017) conclude that the actual type I error could be well below α. This is likely to happen when β_0 is large, corresponding to a large maximal difference Δ between the survival functions. The PH assumption is important when connecting β_0 to Δ. Practical applications are unlikely to specify Δ more than 0.15, perhaps much smaller. Martinez et al, (2017) analyze the POSM (proportional odds survival model) as an alternative to the PH model, and discuss robustness to misspecification of the true model.

(2) One could argue that in TOST (*a*), $I_N(-\beta_0)$ should replace $I_N(\hat{\beta})$ and in TOST (*b*), $I_N(\beta_0)$ should replace $I_N(\hat{\beta})$. In general, the limit $\sigma^2(\beta)$ of $N^{-1}I_N(\beta)$ depends on the design characteristics of the study such as, the rate of accrual of subjects, the distribution of entry time, survival and censoring distributions from entry and the total follow-up period of the study. Under assumptions on these design features, a formula can be derived for $\sigma^2(\beta)$, but numerical methods are needed to evaluate integrals. PROC SEQDESIGN documents some formulae on the cumulative expected number of events.

(3) Use of the null variance is seen in the non-inferiority test of a single binomial probability p against a known standard. The hypothesis test is $H_0 : p - p_0 \le -\Delta$ *vs* $H_1 : p - p_0 > -\Delta$ where p_0 is known and $\Delta > 0$ specified . Higher values of p are favorable. The rejection rule based on $Z(p) = N^{1/2}\left(\hat{p} - (p_0 - \Delta)\right) / \sqrt{pq}$, uses approximate normality to fix the size of the test:

Type I error $= P[Z(p) > k \mid H_0] = \Phi\left(-k + N^{1/2}(p - (p_0 - \Delta)) / \sqrt{pq}\right)$. The largest type I error (under the null) is the size of test $= \alpha = \Phi(-k)$ which gives $k = z_{1-\alpha}$. Hence,

Reject H_0 if $Z(p_0 - \Delta) = \dfrac{N^{1/2}\left(\hat{p} - (p_0 - \Delta)\right)}{\sqrt{(p_0 - \Delta)(1 - (p_0 - \Delta))}} > z_{1-\alpha}$. This default option VAREST=NULL is applied in PROC POWER. However, the option VAREST=SAMPLE, would use $\sqrt{\hat{p}(1-\hat{p})}$ in the denomination of the above statistic. In PROC FREQ the option is called VAR= and the default is VAR=SAMPLE.

Analogous comments apply to the equivalence test: $H_0 : | p - p_0 | \ge \Delta$ *vs* $H_1 : | p - p_0 | < \Delta$.

(4) The information function $I_N(\beta)$ could be constructed from the output from PROC LIFETEST. In general, for all β , $I_N(\beta) \le ¼D$ where D is the total number of observed events. The function is skewed bell-shaped but need not achieve its maximum at $\beta = \hat{\beta}$. If $n_{i1} \approx n_{i2}$ at all event times, then $I_N(\beta) \approx e^{\beta}(1 + e^{\beta})^{-2} D$, so that the upper bound is at $\beta = 0$. In assessing sample size for non-inferiority trials, Curtis and Crisp (2008) comment on which variance to use in construction of the test statistic based on the log hazard ratio of two exponential distributions of the time to event. They argue for applying the variance of the log hazard ratio under the alternative, rather than under the null, because the variance is larger under the alternative. This results in a conservative sample size. For exponentially distributed events times, the inverse of the variance of the log hazard ratio $\hat{\beta}_E$ is estimated by $I_N(\hat{\beta}_E) = w_1 w_2 D$, where w_1, w_2 are proportions of D in groups 1 and 2 ($w_1 + w_2 = 1$).

In conclusion, there are many features to be considered in formulating tests of non-inferiority and equivalence for comparing two survival functions. At the design stage, assessment of sample size and power is important. Two options in PROC POWER (**coxreg**, **twosamplesurvival**) offer some support for this effort.

REFERENCES

Andersen PK, Borgan O, Gill RD, Keiding N. *Statistical Models Based on Counting Processes.* New York: Springer-Verlag; 1993.

Castelloe J, Watts D. Equivalence and Noninferiority Testing Using SAS/STAT Software, Paper SAS1191-2015. SAS Global Forum; 2015; Dallas, TX.

Chang M. *Modern Issues and Methods in Biostatistics.* New York: Springer Science; 2011.

Chowdhury S, Tiwari RC, Ghosh S. Approaches for testing noninferiority in two-arm trials for risk ratio and odds ratio. *Journal of Biopharmaceutical Statistics.* 2019a;29(3):425-445.

Chowdhury S, Tiwari RC, Ghosh S. Non-inferiority testing for risk ratio, odds ratio and number needed to treat in three-arm trial. *Computational Statistics & Data Analysis.* 2019b;132:70-83.

Cornely OA, Crook DW, Esposito R, et al. Fidaxomicin versus vancomycin for infection with Clostridium difficile in Europe, Canada, and the USA: a double-blind, non-inferiority, randomised controlled trial. *Lancet Infectious Diseases.* 2012;12(4):281-289.

Crisp A, Curtis P. Sample size estimation for non-inferiority trials of time-to-event data. *Pharmaceutical Statistics.* 2008;7(4):236-244.

da Silva GT, Logan BR, Klein JP. Methods for Equivalence and Noninferiority Testing. *Biology of Blood and Marrow Transplantation.* 2009;15(1):120-127.

Farrington CP, Manning G. Test statistics and sample-size formulas for comparative binomial trials with null hypothesis of nonzero risk difference or non-unity relative risk. *Statistics in Medicine.* 1990;9(12):1447-1454.

Freitag G. Methods for assessing noninferiority with censored data. *Biometrical Journal.* 2005;47(1):88-98.

Freitag G, Lange S, Munk A. Non-parametric assessment of non-inferiority with censored data. *Statistics in Medicine.* 2006;25(7):1201-1217.

Hauck WW, Anderson S. A comparison of large-sample confidence-interval methods for the difference of 2 binomial probabilities. *American Statistician.* 1986;40(4):318-322.

Holzgreve H, Distler A, Michaelis J, Philipp T, Wellek S. Verapamil versus hydrochlorothiazide in the treatment of hypertension - results of long-term double-blind comparative trial. *British Medical Journal.* 1989;299(6704):881-886.

Klein JP, Moeschberger ML. *Survival Analysis: Techniques for Censored and Truncated Data, 2nd Edition.* New York: Springer-Verlag; 2003.

Liao J. A constrained non-inferiority approach for assessing clinical efficacy to establish biosimilarity. *International Journal of Clinical Biostatistics and Biometrics.* 2015;1(2):2-7.

Li X, Li H, Jin M, Goldberg JD. Likelihood ratio and score tests to test the non-inferiority (or equivalence) of the odds ratio in a crossover study with binary outcomes. *Statistics in Medicine.* 2016;35(20):3471-3481.

Martinez EE, Sinha D, Wang W, Lipsitz SR, Chappell RJ. Tests for equivalence of two survival functions: Alternative to the tests under proportional hazards. *Statistical Methods in Medical Research.* 2017;26(1):75-87.

Nam J. Establishing equivalence of two treatments and sample size requirements in matched-pairs design. *Biometrics.* 1997;53(4):1422-1430.

Scagliotti GV, Parikh P, von Pawel J, et al. Phase III study comparing cisplatin plus gemcitabine with cisplatin plus pemetrexed in chemotherapy-naive patients with advanced-stage non-small-cell lung cancer. *Journal of Clinical Oncology.* 2008;26(21):3543-3551.

Schuirmann DJ. A comparison of the 2 one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics.* 1987;15(6):657-680.

Sidik K. Exact unconditional tests for testing non-inferiority in matched-pairs design. *Statistics in Medicine.* 2003;22(2):265-278.

Tang N-S, Tang M-L, Wang S-F. Sample size determination for matched-pair equivalence trials using rate ratio. *Biostatistics.* 2007;8(3):625-631.

Tsong Y, Yuan M, Dong X, Wu Y-t, Shen M. Comparing the response rates for superiority, noninferiority and equivalence testing with multiple-to-one matched binary data. *Journal of Biopharmaceutical Statistics.* 2013;23(1):98-109.

Wellek S. *Testing Statistical Hypotheses of Equivalence and Non-inferiority. Second Edition.* Boca Raton, FL: Chapman & Hall/CRC; 2010.

CONTACT INFORMATION

We welcome your comments and questions. Please contact

Joseph C. Gardiner
Department of Epidemiology and Biostatistics
College of Human Medicine
Michigan State University
East Lansing, MI 48824
gardine3@msu.edu

APPENDIX

The score tests for testing non-inferiority or equivalence are obtained from a common origin. Consider the two independent samples $n_{11} \sim BIN(n_{1\cdot}, p_T), n_{21} \sim BIN(n_{2\cdot}, p_C)$ shown in Table 1. The log likelihood $\log L(\boldsymbol{\theta}) = n_{11} \log\left(\frac{p_T}{1-p_T}\right) + n_{1\cdot} \log(1-p_T) + n_{21} \log\left(\frac{p_C}{1-p_C}\right) + n_{2\cdot} \log(1-p_C)$ where constants are dropped, and $\boldsymbol{\theta} = (p_T, p_C)'$. The score function is the derivative

$$\mathbf{U}(\boldsymbol{\theta}) = \frac{\partial(\log L(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} = \left[n_{1\cdot}\left(\frac{\hat{p}_T - p_T}{p_T(1-p_T)}\right), \quad n_{2\cdot}\left(\frac{\hat{p}_C - p_C}{p_C(1-p_C)}\right)\right]'.$$

Next, form the Hessian $\mathbf{H}(\boldsymbol{\theta}) = -\frac{\partial^2(\log L(\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = -diag\left(\frac{\partial U(p_T,p_C)}{\partial p_T}, \frac{\partial U(p_T,p_C)}{\partial p_C}\right)$.

Write $n_{1\cdot} = Nw_1$, $n_{2\cdot} = Nw_2$, with $w_1 + w_2 = 1$ and obtain the limits in probability as $N\to\infty$:

$$-N^{-1}\frac{\partial U(p_T,p_C)}{\partial p_T} \to w_1\{p_T(1-p_T)\}^{-1}, -N^{-1}\frac{\partial U(p_T,p_C)}{\partial p_C} \to w_2\{p_C(1-p_C)\}^{-1}.$$

Let $\mathbf{A}(\boldsymbol{\theta}) = diag\left[w_1\{p_T(1-p_T)\}^{-1}, w_2\{p_C(1-p_C)\}^{-1}\right]$.

Score tests are constructed for testing a null hypothesis, $H_0 : \mathbf{c}(\boldsymbol{\theta}_0) = \mathbf{0}$. For example, if the relative risk is specified as ρ_0, then $\mathbf{c}(\boldsymbol{\theta}_0) = \left(p_T - \rho_0 p_C\right)$, a scalar and $\boldsymbol{\theta}_0 = \{(p_T, p_C) : p_T = \rho_0 p_C\}$. Continuing with this case, $\{\mathbf{A}(\boldsymbol{\theta}_0)\}^{-1} N^{-½}\mathbf{U}(\boldsymbol{\theta}_0) = \left[w_1^{-½}\sqrt{n_{1\cdot}}(\hat{p}_T - \rho_0 p_C), \quad w_2^{-½}\sqrt{n_{2\cdot}}(\hat{p}_C - p_C)\right]'$ has an asymptotic bivariate normal distribution, zero means, zero correlation, and variances $\left(w_1^{-1}\rho_0 p_C(1-\rho_0 p_C),\ w_2^{-1}p_C(1-p_C)\right)$. For testing $H_0 : \mathbf{c}(\boldsymbol{\theta}_0) = \mathbf{0}$, form the test statistic

$\left[\left.\frac{\partial \mathbf{c}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}_0}\right]' \{\mathbf{A}(\boldsymbol{\theta}_0)\}^{-1} N^{-½} U(\boldsymbol{\theta}_0) = \sqrt{N}(\hat{p}_T - \rho_0\hat{p}_C)$. The asymptotic distribution is normal, mean zero and variance $\left(\frac{\rho_0 p_C(1-\rho_0 p_C)}{w_1} + \frac{\rho_0^2\, p_C(1-p_C)}{w_2}\right)$. The standardized statistic is

$$Z = \frac{\sqrt{N}(\hat{p}_T - \rho_0\hat{p}_C)}{\left[\left(\frac{\rho_0 p_C(1-\rho_0 p_C)}{w_1} + \frac{\rho_0^2\, p_C(1-p_C)}{w_2}\right)\right]^{½}} = \frac{\sqrt{Nw_1w_2}(\hat{p}_T - \rho_0\hat{p}_C)}{\left[w_2\rho_0 p_C(1-\rho_0 p_C) + \rho_0^2\, w_1 p_C(1-p_C)\right]^{½}}.$$

This is the same statistic applied in PROC POWER, but the group subscripts (T, C) are reversed.

To operationalize, replace p_C by its MLE $\tilde{p}(=\tilde{p}_C), \tilde{p}_T = \rho_0\tilde{p}_C$, under the null hypothesis, giving us equation (5): $Z = \frac{\hat{p}_T - \rho_0\hat{p}_C}{\left[\frac{\tilde{p}_T\tilde{q}_T}{n_{1\cdot}} + \rho_0^2\frac{\tilde{p}_C\tilde{q}_C}{n_{2\cdot}}\right]^{½}}$. To obtain the MLE of p_C under the constraint, $p_T = \rho_0 p_C$, the log likelihood has a single parameter p_C, that leads to the estimating equation: $(n_{11} - n_{1\cdot}\rho_0 p_C)(1-p_C) + (n_{21} - n_{2\cdot}p_C)(1-\rho_0 p_C) = 0$. The solution is in equation (6).

The score tests based on the risk difference and odds ratio are derived in the same manner by applying the appropriate constraint $\mathbf{c}(\boldsymbol{\theta}_0) = \mathbf{0}$. For a survival endpoint, and testing based on the log hazard ratio, the score function is obtained from the partial likelihood.