Published in final edited form as:

Clin Trials. 2016 December ; 13(6): 632–640. doi:10.1177/1740774516654861.

Non-inferiority Tests for Anti-Infective Drugs using Control Group Quantiles

Michael P. Fay¹ and Dean A. Follmann¹

¹National Institute of Allergy and Infectious Diseases

Abstract

Background/Aims—In testing for non-inferiority of anti-infective drugs, the primary endpoint is often the difference in the proportion of failures between the test and control group at a landmark time. The landmark time is chosen to approximately correspond to the q th historic quantile of the control group, and the non-inferiority margin is selected to be reasonable for the target level q . For designing these studies, a troubling issue is that the landmark time must be pre-specified, but there is no guarantee that the proportion of control failures at the landmark time will be close to the target level q . If the landmark time is far from the target control quantile then the pre-specified non-inferiority margin may not longer be reasonable. Exact variable margin tests have been developed by Röhm and Kieser to address this problem, but these tests can have poor power if the observed control failure rate at the landmark time is far from its historic value.

Methods—We develop a new variable margin non-inferiority test where we continue sampling until a pre-specified proportion of failures, q , have occurred in the control group, where q is the target quantile level. The test does not require any assumptions on the failure time distributions, and hence no knowledge of the true q th control quantile for the study is needed.

Results—Our new test is exact and has power comparable to (or greater than) its competitors when the true control quantile from the study equals (or differs moderately from) its historic value. Our *nivm* R package performs the test and gives confidence intervals on the difference in failure rates at the true target control quantile. The tests can be applied to time to cure, or other numeric variables as well.

Conclusions—A substantial proportion of new anti-infective drugs being developed use non-inferiority tests in their development, and typically a pre-specified landmark time and its associated difference margin is set at the design stage to match a specific target control quantile. If through changing standard of care or selection of a different population, the target quantile for the control group changes from its historic value, then the appropriateness of the pre-specified margin at the landmark time may be questionable. Our proposed test avoids this problem by sampling until a pre-specified proportion of the controls have failed.

Keywords

Difference in proportions; Exact test; Non-inferiority margin; Variable margin

Reprints and permission: sagepub.co.uk/journalsPermissions.nav.

Corresponding author: Michael P. Fay, National Institute of Allergy and Infectious Diseases, Biostatistics Research Branch, 5601 Fishers Lane MSC 9820, Bethesda MD 20892, USA. mfay@niaid.nih.gov.

Introduction

New anti-infective drugs are often tested using non-inferiority clinical trials against previously developed drugs. In 2014 of the 10 drugs approved by the U.S. Food and Drug Administration for treatment of infectious diseases, half used a non-inferiority trial in their clinical development^{1,2}. Consider Isavuconazole (under development), used for treating invasive aspergillosis, a life-threatening infection seen in immunocompromised patients³. A non-inferiority study compared the test drug (Isavuconazole) to a control drug (Voriconazole) for all-cause mortality by the landmark time (day 42), with a pre-specified margin of 10%. Isavuconazole would be declared non-inferior if the difference between the test and control failure rates at day 42 was statistically significantly less than 10%. This 10% margin was based on historic failure rates at day 42 of 18.8% for the control treatment, and for this historic rate, the margin of 10% was reasonable to the regulatory agency. In the non-inferiority study, the control failure rates at day 42 matched the historical rates fairly well ($52/258 = 20.2\%$), and the test drug had similar rates ($48/258 = 18.6\%$). Since the upper limit of the 95% confidence interval on the percent difference was less than 10%, Isavuconazole was declared non-inferior.

Consider the potential problem of the control treatment failure rates at the landmark time being considerably less than the historic rates. If the observed failure rates for the control treatment were 10% in the non-inferiority study, then the 10% margin may be too large to reasonably declare non-inferiority. When planning a study, we cannot guarantee that the failure rates for control treatment will be close to the historic rates, so we would like to be able to adjust the margin in a statistically rigorous way according to the observed control failure rates. One approach is to use variable non-inferiority margins that change based on the control rate^{4,5}. For example, Röhm and Kieser⁵ continue to use a pre-specified landmark time, but through the variable non-inferiority margin make it difficult to declare non-inferiority if the control failure rate is less than the historic one. Although that approach protects against unreasonable non-inferiority declarations, this approach has low power if the control failure rate at the landmark time was poorly predicted. Our approach, the *control quantile* method, is different. We follow patients until a pre-specified *proportion* of the control treatment group fail, then develop tests and confidence intervals using data up until that time. The *control quantile* method retains proper type I error rates, protects against unreasonable non-inferiority findings, and importantly retains reasonable power if the control failure rates have changed since the historic rates were measured.

The control quantile method may be applied to time to cure as well as other numeric endpoints. For example, in the second Application Section, we consider a trial with the endpoint of percent reduction in skin lesions at day 3. For ease of exposition, we primarily refer to the endpoint of analysis as time to failure.

Review of Non-inferiority Tests with Fixed or Variable Margins

Fixed Margin Tests at Landmark Time

Let $F_1(t)$ and $F_2(t)$ be the cumulative distributions for the control and test treatment failure times, respectively. Let t_0 be a pre-specified landmark time. A simple non-inferiority hypotheses on the failure times at t_0 is

$$\begin{aligned} H_0: F_2(t_0) - F_1(t_0) &\geq \delta \\ H_1: F_2(t_0) - F_1(t_0) &< \delta, \end{aligned} \quad (1)$$

where $\delta > 0$ is the non-inferiority margin.

Let $x_1 \equiv x_1(t_0)$ and $x_2 \equiv x_2(t_0)$ be the number of failures by t_0 in the control group and the treatment group, respectively. The associated random variables have $X_a(t_0) \sim \text{Binomial}(n_a, F_a(t_0))$, where n_a is the number of subjects in group a , and $a = 1, 2$. We can use tests or confidence intervals on the difference in proportions for these hypotheses^{6,7}; however, these methods on the difference in proportions may not be testing the most appropriate hypotheses, as we discuss in the next section.

Motivation and Choice of a Variable Margin

For the fixed margin test, ideally the landmark time t_0 is approximately $\tau \equiv F_1^{-1}(q)$, where q is some target control failure rate. For the Isavuconazole study, $t_0 = 42$ days, $\delta = 0.10$ and the historical control rate was 18.8% so a reasonable q is $q = 0.20$.

In practice, F_1 for the upcoming study is unknown so we cannot pre-specify both t_0 and q . If F_1^{hist} is the historical control distribution, then t_0 is chosen such that $\hat{F}_1^{\text{hist}}(t_0) \approx q$. However, even if $q \approx \hat{F}_1^{\text{hist}}(t_0) \approx F_1^{\text{hist}}(t_0)$, through either selection of different patients, change of care over time, or evolution of infecting pathogens, we could have $F_1(t_0)$ very different from $F_1^{\text{hist}}(t_0)$. Under these new conditions, the margin δ may no longer be sensible. As an extreme case, if after the non-inferiority study was completed, we estimate that $\hat{F}_1(t_0) = 0.01$, then the margin $\delta = 0.10$ would not make sense, since if $\hat{F}_1(t_0) = F_1(t_0)$ then only 1% of the patients on control treatment are expected to die by t_0 , but a test treatment with up to 10 times higher failure rates at t_0 would be defined as non-inferior. The current practice is to negotiate with the regulatory agency a landmark time, t_0 , with $q \approx \hat{F}_1^{\text{hist}}(t_0)$, and hope that the control failure rates at t_0 are close to q in the non-inferiority study. This is not an ideal situation, since there is uncertainty about what to do if the control rates from the non-inferiority study are very much less than the target (or even how to define “very much less”).

When t_0 is not approximately $\tau = F_1^{-1}(q)$, another approach is to use variable non-inferiority margins for the difference in proportions that change based on the control rate^{4,5}. These methods redefine the hypotheses as

$$\begin{aligned} H_0: & F_2(t) - F_1(t) \geq d\{F_1(t)\} \text{ for all } t \\ H_1: & F_2(t) - F_1(t) < d\{F_1(t)\}, \text{ for some } t \end{aligned} \quad (2)$$

where $d\{\cdot\}$ is the variable margin function for the difference. These hypotheses are typically calibrated so that $d\{\cdot\}$ is smooth with $d\{F_1(\tau)\} = \delta$. We rewrite Hypotheses (2) as,

$$\begin{aligned} H_0: & F_2(t) \geq g\{F_1(t)\} \text{ for all } t \\ H_1: & F_2(t) < g\{F_1(t)\}, \text{ for some } t \end{aligned} \quad (3)$$

where $g\{F_1(t)\} = d\{F_1(t)\} + F_1(t)$. The Hypotheses (3) do not require a pre-specified t_0 close to $\tau = F_1^{-1}(q)$ like Hypotheses (1). Further Hypotheses (3) have more flexibility than comparing hazard ratios under proportional hazards (see⁴ Chapter 13). The null hypothesis in (3) holds for all t and hence represents the entire distributions, F_1 and F_2 . If we reject the null, then for *some* t , we conclude that $F_2(t) - F_1(t) < d\{F_1(t)\}$. If $t_0 \approx \tau$, then the test will be powerful when $F_2(\tau) - F_1(\tau)$ is small. The main issue with these hypotheses is defining $g\{\cdot\}$ so that for every t we get a sensible non-inferiority margin, though typically our focus is on the region around t_0 or τ .

The function $g\{\cdot\}$ partitions the space of all values of the points $\{F_1(t), F_2(t)\}$ into a null region (gray in Figure 1) or an alternative region (white in Figure 1). For a specific landmark time t_0 , if you can show that the point $\{F_1(t_0), F_2(t_0)\}$ is significantly in the alternative region, then you have rejected the null hypothesis. This does not require choosing t_0 to equal $F_1^{-1}(q)$.

Consider some choices for $g\{\cdot\}$. Suppose all we cared about is Hypotheses (1) except with t_0 replaced with $\tau = F_1^{-1}(q)$:

$$\begin{aligned} H_0: & F_2(\tau) \geq F_1(\tau) + \delta \\ H_1: & F_2(\tau) < F_1(\tau) + \delta. \end{aligned} \quad (4)$$

Since τ is unknown and there are no restrictions on F_1 and F_2 , in order to rigorously test Hypotheses (4) we choose $g\{\cdot\}$ as a least-favorable-like function. Figure 1a plots this $g\{\cdot\}$ for the hypothesis with $q = 0.20$ and $\delta = 0.10$, showing that Hypotheses (4) are not a practical set of hypotheses; the point t with $F_1(t) = 0.80$ and $F_2(t) = 0.40$ is in the null region despite this point signifying that the test drug has substantially lower failure rates than the control drug after 80% of the controls have failed.

Figure 1b depicts a constant difference margin of 10%. This is not sensible, since the point with $F_1(t) = 0.01$ and $F_2(t) = 0.10$ would be in the alternative region, showing that a test drug with 10 times the failure rate of the control drug is non-inferior. Figure 1c depicts a constant odds ratio margin of $\Psi = 1.71$ (so that at $F_1(t) = 0.20$ the value of $F_2(t) = 0.30$ would be on the boundary). This margin avoids the problem with the constant difference boundary

because as $F_1(t) \rightarrow 0$ the boundary approaches the point $(F_1(0) = 0, F_2(0) = 0)$. Finally, Figure 1d gives a margin with constant odds ratio of $\Psi = 1.71$ for $t \leq \tau$ (with $F_1(\tau) = 0.20$), and a constant difference of 0.10 for $t > \tau$, as proposed by Röhmel and Kieser⁵.

Testing on Variable Margins with t_0 Fixed

An exact test for Hypotheses (3) with fixed t_0 and $x_a = x_a(t_0)$ for $a = 1, 2$ is given as follows. Define $W(x_1, x_2)$ such that smaller values are less likely under the null hypothesis, and $W(\cdot, \cdot)$ meets Barnard's convexity requirement^{5,8}:

$$\begin{aligned} W(x_1+1, x_2) &\leq W(x_1, x_2), \text{ for } x_1 \in \{0, 1, \dots, n_1 - 1\}, x_2 \in \{0, 1, \dots, n_2\}, \\ &\text{and} \\ W(x_1, x_2+1) &\geq W(x_1, x_2) \text{ for } x_1 \in \{0, 1, \dots, n_1\}, x_2 \in \{0, 1, \dots, n_2 - 1\}. \end{aligned} \quad (5)$$

Then a valid (i.e., exact) one-sided p-value can be calculated by maximizing over the margin the probability of equal or more extreme responses (see⁵ equation 3). Specifically, a one-sided exact p-value for testing the margin g with fixed t_0 is

$$p_{ft}(\mathbf{x}, g, t_0) = \max_{F_2(t_0) = g\{F_1(t_0)\}} \left\{ \sum_{i,j: W(i,j) \leq W(x_1, x_2)} f_{bin}(i; n_1, F_1(t_0)) f_{bin}(j; n_2, F_2(t_0)) \right\} \quad (6)$$

where f_{bin} is a binomial probability mass function, and the maximum is over pairs $\{F_1(t_0), F_2(t_0)\}$ such that $F_2(t_0) = g\{F_1(t_0)\}$ (for example, the thick solid black lines in Figure 1).

Röhmel and Kieser⁵ discuss four tests for variable margins based on the observed proportions at a landmark time. Two of these tests are exact (T_3 and T_4 in their notation). The first of these (T_3) is motivated by a margin similar to Figure 1d, and in our notation $W(\cdot, \cdot)$ is the maximum of two p-values: one from an exact conditional test for the odds ratio and one from an asymptotic test on the difference. The second exact test (T_4) is more powerful and is a generalization of a test of Barnard. That test orders the possible values of the points (x_1, x_2) by defining $W(n_1, 0)$ as the lowest value, and sequentially defining the next higher W value as at the point that meets Barnard's convexity requirement and increases by the smallest amount the probability of rejection on the margin with ties going to fewer events in the treatment group. The rule for breaking of ties is somewhat arbitrary, but having fewer ties will increase the power of the test without sacrificing the exactness property. This Barnard-Röhmel-Kieser test is quite computationally intensive.

For these exact variable margin tests, any point $(F_1(t_0), F_2(t_0))$ that is in the null space will have probability of rejection less than or equal to the significance level, α . So the test is valid without any restrictions on F_1 and F_2 and regardless of how close $t = t_0$ is to τ . The problem is that if t_0 is chosen too small then the power can be very small under an alternative where $F_1 = F_2$. In the next section, we introduce the "control quantile test" that avoids this power problem, while retaining both the validity and flexibility of variable margins of the tests of Röhmel and Kieser.

Control Quantile Test

Continuous Failure Times

Let $T_{a(1)} < \dots < T_{a(n_a)}$ be the continuous failure times in group a . Our proposed test procedure is to reject if $x_2(T_{1(i)})$ is low, where i is specified using a pre-determined proportion in the control group, i/n_1 . For fixed $T_{1(i)}$ then

$$X_2(T_{1(i)}) \sim \text{Binomial} \{n_2, F_2(T_{1(i)})\}.$$

By the continuity assumption and the probability integral transformation, $F_1(T_{1(i)})$ is the i th order statistic from a sample of n_1 uniform random variables, and $F_1(T_{1(i)}) \sim \text{Beta}(i, n_1 - i + 1)$ (see⁹ p. 54 and p. 230). In Figure 2, we plot the densities of $F_1(T_{1(i)})$ when $i = (0.20n_1)$ for $n_1 = 100, 400$, and 1600 . Let $F_1(T_{1(i)}) = U$. Then on the boundary between the

hypotheses in (3), the binomial parameter for $X_2(T_{1(i)})$ is $g\{U\} \equiv F_2\{F_1^{-1}(U)\}$, and unconditionally its probability mass function is

$$f_{X_2}(x_2; g, i) = \int_0^1 f_{\text{bin}}(x_2; n_2, g\{u\}) f_{\text{beta}}(u; i, n_1 - i + 1) du,$$

where f_{beta} is the beta probability density function. Importantly, the distribution of X_2 does not depend on F_1 or F_2 , but only on their relationship on the boundary through $g\{\cdot\}$. We calculate f_{X_2} using numeric integration. The (one-sided) p-value of the control quantile test for hypotheses (3) is

$$p_{cq}(x_2; g, i) = \sum_{j=0}^{x_2} f_{X_2}(j; g, i). \quad (7)$$

For fixed i , the power to reject when $F_2\{F_1^{-1}(u)\} = h(u)$ is

$$\sum_{j=0}^{n_2} f_{X_2}(j; h, i) I\{p_{cq}(j; g, i) \leq \alpha\},$$

where $I(A) = 1$ if A is true and 0 otherwise.

Any i will give an exact test, and the choice of i will determine which part of the alternative space will drive the power to reject. Since our primary interest is at alternatives at the target quantile, τ [where $F_1(\tau) = q$], we choose i so that $i = \lceil n_1 q \rceil$, the largest integer less than or equal to $n_1 q$. By choosing this value, as the sample size gets larger, the test focuses more and more tightly around the difference in distributions at τ (see Figure 2). Alternatively, we could choose the smallest i that maximizes the power when $F_1 = F_2$, but that strategy is not recommended because it may emphasize different places in the alternative space which may require much longer follow-up leading to longer trials.

If the time to event is a positive event like cure, then we would test the alternative that $F_2(t) > g\{F_1(t)\}$ for some t , and the p-value is calculated analogously,

$$p_{cq}^*(x_2; g, i) = \sum_{j=x_2}^{n_2} f_{x_2}(j; g, i).$$

Allowing Tied Failure Times

Now consider the case where tied failure times are allowed. Although this section is notationally cumbersome, the basic idea is to assume that ties are caused by grouping of the failure times into observable bins. We consider two adjustments, a conservative one and a more approximate one.

Assume that the failure times occur in continuous time, but we can only observe whether a failure time has occurred or not at a finite amount of assessment times. Let the assessment times be $0 = a_0 < a_1 < \dots < a_K < \infty$. We observe $x_1(a_k)$ and $x_2(a_k)$ for $k = 1, \dots, K$. For example, if we only observe the failure times to the nearest day, then the assessment times could be defined as midnight at each day. Let $m_{1k} = x_1(a_k) - x_1(a_{k-1})$ and $m_{2k} = x_2(a_k) - x_2(a_{k-1})$ be the number of failures in each group that occur in $(a_{k-1}, a_k]$. Let $\mathcal{A}$ be defined so that $T_{1(j)} \in (a_{\ell-1}, a_\ell]$, with $i = [n_1 q]$, so that $x_2(a_{\ell-1}) \leq x_2(T_{1(j)}) \leq x_2(a_\ell)$. Since for hypotheses (3) higher values of $x_2(T_{1(j)})$ make it harder to reject, the conservative adjustment is to use $x_{2c} = x_2(a_\ell)$ in place of $x_2(T_{1(j)})$ in equation 7. For testing an alternative in the opposite direction (i.e., $F_2(t) > g\{F_1(t)\}$), then lower values of $x_2(T_{1(j)})$ make it harder to reject, and the conservative adjustment is to use $x_{2c} = x_2(a_{\ell-1})$.

Now we develop an approximate method for when the conservativeness needed to ensure strict validity is not required. Let $T_{1(j)}$ be the (unobserved) j th largest of the $m_{1\ell}$ failure times for the control group, so that $j = i - x_1(a_{\ell-1})$. Since we cannot observe the failures within $(a_{\ell-1}, a_\ell]$, we make the convenience assumption that the failures for both groups are equally distributed within that interval. Imagine a pot of $m_\ell = m_{1\ell} + m_{2\ell}$ failures from both groups, and imagine drawing failures (from both groups combined) at random until we get the j th failure from the control group out of the $m_{1\ell}$. Then Y , the number of draws it takes to get that j th control failure, is distributed inverse hypergeometric with mean $j(m_\ell + 1)/(m_{1\ell} + 1)$ (see ¹⁰ p. 255-256). Then the number of treatment failure draws, $Y - j$, represents $x_2(T_{1(j)}) - x_2(a_{\ell-1})$. So to estimate the unobservable $x_2(T_{1(j)})$, for this approximate method we use x_{2a} , where

$$x_{2a} = \text{round} \left\{ x_2(a_{\ell-1}) + E(Y) - j \right\} = \text{round} \left\{ x_2(a_{\ell-1}) + \frac{\{i - x_1(a_{\ell-1})\} m_{2\ell}}{m_{1\ell} + 1} \right\}, \quad (8)$$

where we round up if the non-integer part is 0.5 for testing hypotheses 3, and round down if the inequality in the alternative is reversed. If $m_{1\ell} = 1$ then $x_{2a} = \text{round} \left\{ x_2(a_{\ell-1}) + \frac{m_{2\ell}}{2} \right\}$.

Confidence Intervals

The effect of interest is

$$\Delta(q) = F_2(\tau) - F_1(\tau),$$

where $F_1(\tau) = q$ and q is known and fixed, and $(q) \in [-q, 1 - q]$. We can obtain an upper confidence limit for (q) by inverting a series of hypothesis tests associated with a class of g functions in equation 3 (see⁹, Section 9.2.1 for general discussion on inverting hypothesis tests to get confidence sets). For example, one class is the constant difference functions given by $g(F_1(t)) = F_1(t) + \Delta$, with Figure 1b representing one element in the series, $g_{0.10}$. Then the upper $100(1 - \alpha/2)\%$ one-sided limit is the value Δ_U such that the control quantile test gives $\alpha/2 = P_{cq}(x, g_{\Delta_U})$ (or $\Delta_U = 1 - q$ if no such solution exists). The lower $100(1 - \alpha/2)\%$ one-sided limit, Δ_L , is analogous and solves $\alpha/2 = p_{cq}^*(x, g_{\Delta_L})$ (or $\Delta_L = -q$ if no such solution exists). Those two limits can be combined to give a $100(1 - \alpha)\%$ confidence interval (Δ_L, Δ_U) .

Now consider the constant odds ratio function (e.g., Figure 1c). Let $\psi(q, \Delta)$ be the odds ratio when $F_2(\tau) = F_1(\tau) + \Delta = q + \Delta$, i.e.,

$$\psi(q, \Delta) = \frac{F_2(\tau) \{1 - F_1(\tau)\}}{F_1(\tau) \{1 - F_2(\tau)\}} = \frac{(q + \Delta)(1 - q)}{q(1 - q - \Delta)}.$$

Solving for $F_2(\tau)$ we get a series of functions that depend on Δ (and a fixed q), and replacing τ with arbitrary t we get the series of g functions

$$g_{\Delta}(F_1(t)) = \frac{F_1(t)\psi(q, \Delta)}{F_1(t)\psi(q, \Delta) + 1 - F_1(t)}. \quad (9)$$

The confidence intervals is calculated exactly as for the constant difference functions except using equation 9 for $g(F_1(t))$.

Now consider the class of functions like in Figure 1d, some of which are plotted in Figure 3. We call one of these margins, a **Quantile Difference Margin with an Odds Ratio Taper** (QuaDMORT). When $\Delta = 0$ then

$$g_{\Delta}(F_1(t)) = \begin{cases} \min \left\{ F_1(t) + \Delta, \frac{F_1(t)\psi(q, \Delta)}{\psi(q, \Delta)F_1(t) + 1 - F_1(t)} \right\} & \text{if } t < \tau \\ F_1(t) + \Delta & \text{if } t \geq \tau, \end{cases} \quad (10)$$

and when $\Delta < 0$, we use equation 10, except we replace min with max. As with the other cases, we invert a series of hypothesis tests to get the confidence interval. For each of the margins in Figure 3, we can calculate a one-sided p-value for the associated test. All of the margins that fail to reject at the $\alpha/2$ level represent a one-sided $100(1 - \alpha/2)$ confidence set. Then we repeat this process for the other one-sided test. The two-sided $100(1 - \alpha)$ confidence set is the union of those sets. We represent the set as the interval at τ for $\Delta = F_2(\tau) - F_1(\tau)$.

Comparison of Control Quantile Test to Barnard-Röhmel-Kieser Test

The control quantile tests have power comparable to that of the Barnard-Röhmel-Kieser test using $t_0 = \tau$, the most powerful of the exact fixed landmark tests of Röhmel and Kieser⁵. Since both the control quantile test and the Barnard-Röhmel-Kieser test are exact, the salient comparison is power. Figure 4 shows the power of the Barnard-Röhmel-Kieser test depends on the value of $F_1(t_0)$, while the power of the control quantile test has a constant value of 86.8%, since the follow-up does not stop at t_0 but stops at the random variable $T_{1(j)}$. If $t_0 = \tau$ then the power of the Barnard-Röhmel-Kieser test is 85.7%, slightly smaller than the control quantile test. If t_0 is chosen at $F_1(t_0) = 0.25$ then the power of the Barnard-Röhmel-Kieser test is 89.8%, and slightly larger than the control quantile test. Importantly, if t_0 is much smaller than τ then the Barnard-Röhmel-Kieser test power can be greatly diminished (see Figure 4). Thus, by changing the study design, the control quantile test protects against poorly powered studies while achieving nearly the maximum power of its competitor.

Application 1: Simulated Data Modeled After Isavuconazole Study

Consider the Isavuconazole study¹ mentioned in the introduction. An appropriate control quantile test uses the QuaDMORT with $q = 0.20$ and $\delta = 0.10$ and testing hypotheses 3 at the 0.025 significance level. For designing a study with these parameters, we need a sample size of $n_1 = n_2 = 436$ to have at least 90% power under the alternative that $F_1 = F_2$. This sample size is calculated in R by first installing and loading our nivm R package, when using the call: `powerNicqTest(power=.90,q=.2,delta0=.1)`. In contrast, the sample size for an exact non-inferiority test on the difference in proportions (using the score test as the test statistic with StatXact 9, Cytel) at t_0 with margin 0.10 and $F_1(t_0) = F_2(t_0) = 0.20$ and one-sided significance level of 2.5% is $n_1 = n_2 = 340$ for 90% power, or $n_1 = n_2 = 257$ for 80.43% power. So, the QuaDMORT method needs a larger sample size because it protects against unreasonable rejections when $F_1(t_0) \ll 0.20$.

Because the raw data are not publicly available, we simulate data based on a simple model. Kaplan-Meier curves are available and show very similar distributions for both groups (Figure 20 of³), with about 20% (30%) of the failures occurring by day 42 (84). We use the same exponential cure model for both groups, where 60% of the subjects are cured, and the other 40% have exponential failure distributions with mean=60.59, giving $F_1(42) = F_2(42) = 0.20$ and $F_1(84) = F_2(84) = 0.30$. We simulate data using the sample sizes from the study, $n_1 = n_2 = 258$, assuming administrative censoring at day 100 using the nivm R package (see `demo("create.IsaSimData")`). That simulated data has $x_2(T_{1(j)}) = 57$ using $i = 52$, and the control quantile p-value is $p = 0.032$ (which is > 0.025) with 95% confidence intervals on $F_2(\tau) - F_1(\tau)$ of $(-0.051, 0.105)$. This simulated study just barely fails to demonstrate non-inferiority at the margin of 0.10.

Application 2: Percent Reduction in Lesion at Day 3 for Dalbavancin

Now consider two studies, DISCOVER 1 and 2, each comparing dalbavancin to control drugs (vancomycin and linezolid) for skin infections. A primary efficacy response is percent reduction in skin lesion size at 48-72 hours post-study drug initiation. For simplicity we

consider only subjects that have percent reduction measured at 48-72 hours (over 97% of subjects had that efficacy response measured). Thus, for this illustration the sample sizes of DISCOVER 1 are: Dalbavancin ($n=281$) and Vancomycin/Linezolid ($n=277$), and those for DISCOVER 2 are: Dalbavancin ($n=363$) and Vancomycin/Linezolid ($n=358$). We plot the proportion of the patients that had percent reduction less than or equal to different values in Figure 5. Those who had an increase are counted as no reduction in the graph.

There is a slight non-significant difference in the control curves between the two studies. For example, if we defined a success as having percent reduction in lesion size of greater than or equal to 20% as recommended by the FDA¹¹, then DISCOVER 1 has $22/281 = 7.8\%$ failed in the control group, while DISCOVER 2 has $38/363 = 10.5\%$ failed in the control group. The FDA Guidance¹² recommends a 10% margin for non-inferiority for this endpoint, but it is based in part on data from 1940 or earlier (see also Talbot, et al¹³ whose Figure 1 comes from a study published in 1940). This points to the difficulty of defining non-inferiority margins, since the 10% margin was picked without the benefit of knowing that the control group failure rates from the studies would be so low.

For the control quantile test, we define a failed response as percent reduction less than or equal to the percent reduction of the lowest 20% percentile of the control group so that $F_1(\tau) = 0.20$, and the margin is set to 10% at τ . We use the QuaDMORT variable margin function. For DISCOVER 1, we have $n_1 = 277$ so $i = 56$, and $X_2(T_{1(56)}) = 68$ out of $n_2 = 281$. The difference in proportion of failures is $68/281 - 56/277 = 0.2420 - 0.2022 = 0.0398$. So dalbavancin looks slightly worse. The control quantile test p-value is 0.081 with 95% confidence interval on $F_2(\tau) - F_1(\tau)$ equal to $(-0.029, 0.125)$, not quite showing non-inferiority at the 10% margin. For DISCOVER 2, the difference in failure proportions is $76/363 - 72/358 = 0.2094 - 0.2011 = 0.0082$. The control quantile test p-value is 0.0054 with 95% confidence interval on $F_2(\tau) - F_1(\tau)$ equal to $(-0.051, 0.078)$, showing non-inferiority at the 10% margin.

Discussion

We have developed the control quantile non-inferiority test that uses a variable margin motivated by Röhm and Kieser⁵. This test is exact and prevents potential problems with pre-specifying the landmark time based on historical control rates. Specifically, those pre-specified times may differ from the associated historical times due to secular shifts in the distribution due to changing supportive care or changing pathogens. The R package *nivm* to calculate the control quantile test and associated confidence intervals is available on <http://cran.r-project.org/>.

Our method avoids pre-specifying a landmark time where the difference in failure rates are compared, instead we pre-specify the target control proportion of failures that determines the time when we compare the failure rates. Just as the previous method requires extensive discussions with regulators and care with choosing the landmark time, our method requires similar discussions and care in choosing the pre-specified target proportion of control failures. That decision will depend on the application, and it is difficult to give general guidance across applications; however, we can say that the target proportion should be

chosen so that there is a very high probability that the target proportion of failures will be reached within the time frame planned for the study. If the study is stopped before the pre-specified proportion is reached, then only an *ad hoc* analysis may be done, and the resulting statistical properties will be unclear.

If there is right censoring before the pre-specified proportion of controls have failed, then the test cannot be applied as described. If the censoring is independent of the failure time and the group membership, and if there is only a small amount of censoring, then removing the censored individuals should not overly bias the comparison between the groups. A more rigorous method for handling this type of censoring requires further work.

Acknowledgments

We thank Michael Zelasky of Durata Therapeutics for providing the data from the Dalbavancin studies.

References

1. US Food and Drug Administration. [27-Jan-2015] Novel new drugs 2014 summary. 2014. www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/DrugInnovation/UCM430299.pdf
2. US Food and Drug Administration. [27-Jan-2015] 2015. www.accessdata.fda.gov/scripts/cder/drugsatfda/
3. US Food and Drug Administration. [27-Jan-2015] Advisory committee briefing document: Isavuconazonium, invasive aspergillosis and invasive mucormycosis, january 22,2015. 2015. www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/Anti-InfectiveDrugsAdvisoryCommittee/UCM430748.pdf
4. Rothmann, MD.; Wiens, BL.; Chan, IS. Design and analysis of non-inferiority trials. CRC Press; 2012.
5. Röhmel J, Kieser M. Investigations on non-inferioritythe food and drug administration draft guidance on treatments for nosocomial pneumonia as a case for exact tests for binomial proportions. Statistics in Medicine. 2013; 32(14):2335–2348. [PubMed: 22991269]
6. Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. Statistics in medicine. 1998; 17(8):873–890. [PubMed: 9595617]
7. Fagerland MW, Lydersen S, Laake P. Recommended confidence intervals for two independent binomial proportions. Statistical methods in medical research. 2011 0962280211415469.
8. Barnard G. Significance tests for 2× 2 tables. Biometrika. 1947:123–138. [PubMed: 20287826]
9. Casella, G.; Berger, RL. Statistical Inference. second edition. Duxbury Press; 2002.
10. Johnson, NL.; Kemp, AW.; Kotz, S. Univariate Discrete Distributions. 3. New York: John Wiley & Sons; 2005.
11. Durata Therapeutics. [02-Sep-2015] Dalbavancin for injection for treatment of acute bacterial skin and skin structure infections nda 021-883 briefing document presented to the fda anti-infective drugs advisory committee. 2014. <http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/Anti-InfectiveDrugsAdvisoryCommittee/UCM390793.pdf>
12. US Food and Drug Administration. [02-Sep-2015] Guidance for industry, acute bacterial skin and skin structure infections: Developing drugs for treatment. 2013. <http://www.fda.gov/downloads/Drugs/Guidances/ucm071185.pdf>
13. Talbot GH, Powers JH, Fleming TR, et al. Progress on developing endpoints for registrational clinical trials of community-acquired bacterial pneumonia and acute bacterial skin and skin structure infections: update from the biomarkers consortium of the foundation for the national institutes of health. Clinical infectious diseases. 2012; 55:1114–1121. [PubMed: 22744885]

Figure 1.

Different choices for the variable margin function, $g\{F_1(t)\}$. Gray is the null region, and white is the alternative region. All 4 panels have $F_1(\tau) = 0.20$ and $g\{F_1(\tau)\} = 0.10$. In terms of the alternative hypotheses, the panels represent:

- a. $H_1 : F_2(\tau) < F_1(\tau) + 0.10$;
- b. $H_1 : F_2(t) < F_1(t) + 0.10$, for all t ;
- c. $H_1 : OR(F_2(t); F_1(t)) < 1.71$, for all t , where $OR(A, B) = A(1 - B)/(B(1 - A))$; and
- d. $H_1 : OR(F_2(t); F_1(t)) < 1.71$ for $t \leq \tau$ and $F_2(t) < F_1(t) + 0.10$ for $t > \tau$.

Figure 2.

Probability density functions (scaled to fit nicely on the graph) for $F_1(T_{K(1)})$, the true (unknown) control distribution evaluated at the i th failure in the control group, with $i = 0.20n_1$ and $n_1 = 50, 100, 400$, and with failures observed in continuous time. The density lines (thin black) are overlaid onto the variable margin (thick black) of Figure 1d: margin with $F_1(\tau) = 0.20$ and $g\{F_1(\tau)\} = 0.10$ and $H_1: OR(F_2(t); F_1(t)) < 1.71$ for $t < \tau$ and $F_2(t) < F_1(t) + 0.10$ for $t > \tau$. Gray dotted line is $F_1 = F_2$.

Figure 3.

Quantile difference margins with odds ratio tapers (QuaDMORTs) calculated with $q = 0:20$ for different values of δ . The margins are solid for $F_1(t) < 0.20$ and dotted for $F_1(t) \geq 0.20$. The gray margin is the line of equality ($\delta = 0$). The margins have (from the bottom up): $\delta = -0.20, -0.10, 0, 0.10, 0.20, \dots, 0.70$. In constructing a confidence limit, we successively test each curve in the family until we just reject (see text for details).

Figure 4.

$F_1(t_0)$ by Power under $F_1 = F_2$ with one-sided significance level of 0.025 for $q = 0.20$ and $\delta_0 = 0.10$. The power of the Barnard-Röhmel-Kieser test (solid black) depends on $F_1(t_0)$, while the power of the control quantile test is constant (dashed black line). Gray reference line at $q = 0.20$.

Figure 5.

Proportion of patients at 48 to 72 hours with percent reduction from baseline of skin infection area greater than or equal to the value on horizontal axis. Patients with an increase are graphed as 0 percent reduction. Vertical line represents pre-defined success of 20% reduction or greater. Two squares on that line represent proportions that are compared with the traditional method. Horizontal line represents 80% success in the control group, and the two squares at the associated percent reduction represent the proportions at the estimated target comparison value (i.e., at $\hat{\tau} = \hat{F}_1^{-1}(0.80)$) for the control quantile test.