THE USE OF WEIGHTED Z-TESTS IN MEDICAL RESEARCH

K. K. Gordon Lan and Yuhwen Soo
Sanoﬁ-aventis, Bridgewater, New Jersey, USA

Cynthia Siu
Data Power (DP), Inc., Ringoes, New Jersey, USA

Mey Wang
Center for Drug Evaluation, Taipei, Taiwan

Traditionally the un-weighted Z-tests, which follow the one-patient-one-vote principle, are standard for comparisons of treatment effects. We discuss two types of weighted Z-tests in this manuscript to incorporate data collected in two (or more) stages or in two (or more) regions. We use the type A weighted Z-test to exemplify the variance spending approach in the first part of this manuscript. This approach has been applied to sample size re-estimation. In the second part of the manuscript, we introduce the type B weighted Z-tests and apply them to the design of bridging studies. The weights in the type A weighted Z-tests are pre-determined, independent of the prior observed data, and controls alpha at the desired level. To the contrary, the weights in the type B weighted Z-tests may depend on the prior observed data; and the type I error rate for the bridging study is usually inflated to a level higher than that of a full-scale study. The choice of the weights provides a simple statistical framework for communication between the regulatory agency and the sponsor. The negotiation process may involve practical constrains and some characteristics of prior studies.

Key Words: Adapted design; Bridging study; Interim analysis; Sample size re-estimation; Type I error; Weighted Z-test.

1. INTRODUCTION

Recent advances in statistical methodologies that guide decision making in planning, data monitoring, and early termination have contributed to increasing use of flexible methods in designing clinical studies. In clinical trials, it is often desirable, and may be ethically necessary, to utilize ﬁndings from accumulating data to either recommend early termination of the study, modify assumptions underlying the original design (e.g., endpoints and treatment arms), or recalculate sample sizes. Loosely speaking, adaptive design generally refers to the use of accrued interim data or external information to modify an on-going study. A typical application of this idea was Wei and Durham’s (1978) adaptive allocation procedure. Under this
procedure, the treatment arm with the current best outcome is assigned to the next patient with higher probability than the other arms. In this paper, we describe a different approach to modify clinical trial designs based on the observed treatment effect.

In many clinical trials to develop new drugs, the main goal of the study is to test \( H_0: \) the new drug is as effective as placebo (or a standard treatment) versus \( H_a: \) the new drug is better than the placebo. This is a two-sample problem and the decision will be based on a test \( Z_1 = \frac{\sum_{i=1}^{N} X_i}{\sqrt{N}} \) which, in the large sample case, is (approximately) \( N(0,1) \), where \( \delta = 0 \) under \( H_0 \) and is positive under \( H_a \).

The sequentially evaluated values of \( Z_1 \) over time behave like the sequentially evaluated values of a one-sample test \( Z = \frac{\sum_{i=1}^{N} X_i}{\sqrt{N}} \). There is a heuristic way to understand why the sequential distributions of a two-sample test and a one-sample test are similar. Assume that there are \( N \) patients in each of the two treatment groups. Mathematically, we may pair the patients and consider the \( N \) differences \( X_i \) as one-sample observations. When the numbers of patients are different, the details can be found in Lan and Zucker (1993). Therefore, we will only consider the theory for the one-sample statistic, but the results apply to the two-sample comparisons, including the comparison of two means and the use of the log-rank test for survival analysis under proportional hazards assumption.

Suppose \( X_1, X_2, \ldots, X_N \) comprise a sample from a population with mean \( \mu \) and variance \( \sigma^2 = 1 \), where \( \mu = 0 \) under \( H_0 \) and \( \mu > 0 \) under \( H_a \). The final test statistic \( Z_1 \) is \( N(0,1) \) under \( H_0 \). Lan and Wittes (1988) considered the decomposition of the test \( Z = Z_1 \) into the sum of two \( B \)-values. To review this decomposition, let us consider an example when the total sample size \( N \) and an interim analysis are taken at \( n \leq N \), or at information time \( t = n/N \), \( Z = Z_1 \) can be expressed as

\[
Z = \frac{X_1 + X_2 + \cdots + X_N}{\sqrt{N}} = \frac{X_1 + X_2 + \cdots + X_n}{\sqrt{N}} + \frac{X_{n+1} + X_{n+2} + \cdots + X_N}{\sqrt{N}}
\]

\[
= \sqrt{\frac{n}{N}} \frac{X_1 + X_2 + \cdots + X_n}{\sqrt{n}} + \sqrt{\frac{N-n}{N}} \frac{X_{n+1} + X_{n+2} + \cdots + X_N}{\sqrt{N-n}}
\]

\[
= \sqrt{t}Z_{[1-n]} + \sqrt{1-t}Z_{[n+1-N]} \tag{1}
\]

where \( Z_{[1-n]} \) and \( Z_{[n+1-N]} \) are two independent standard normal random variables. The last expression corresponds to the \( B \)-value decomposition in Lan and Wittes (1988). In this paper, we present two types of weighted \( Z \)-tests.

**Type A:** This is a direct generalization of Eq. (1). In general, \( Z_w = \sqrt{w}Z_I + \sqrt{1-w}Z_{II} \), where \( Z_I \) and \( Z_{II} \) are independent, but observation in each may come from two different populations or two stages of a study. Note that if both \( Z_I \) and \( Z_{II} \) are \( N(0,1) \), then \( Z_w \) is also \( N(0,1) \), even if the sample size in \( Z_{II} \) depends on \( Z_I \). The use of Type A weighted \( Z \)-test in sample size re-estimation will be reviewed in Section 3. A special case of \( w = 1/\sqrt{2} \) was proposed by Proschan et al. (2003). They consider restricted designs in which the sample size must be at least as large as originally
planned. The advantage of their method is that if the sample size is not increased, then the test reduces to the usual t-test. A general investigation of the two-stage adaptive designs was given by Liu et al. (2002). A recent manuscript by Hung et al. (2005) gives the motivation behind sample size re-estimation and reviews the Cui et al. (1999) design, comparing its performance to that of the traditional fixed or group sequential design.

Type B: The weighted Z-test is also expressed as $Z_w = \sqrt{w}Z_I + \sqrt{w}Z_{II}$, but the weight $w$ depends on the information contained in $Z_I$. This approach allows for the use of prior information, $Z_I$, for modifying the design parameters and/or partitioning of the future sample space for $Z_{II}$. Note that even if $Z_I$ and $Z_{II}$ are $N(0,1)$ unconditionally, $Z_w$ is not $N(0,1)$. The use of Type B weighted Z-test is also motivated by practical need. For instance, a current issue in drug development is how to bridge evidence of efficacy and safety from old to new patient populations, or to combine results in Phase IIB and III studies. This could be handled by increasing $\alpha$-level, which is determined and made acceptable to local regulatory authorities at the design stage, prior to study start. The weighted Z-test provides a useful framework for deliberations between sponsor and regulatory authorities regarding the best use of existing evidence, which can be translated into a smaller sample size rather than a full scale study, an increasing of the $\alpha$-level, or assigning the weight, $w$, to pre-existing evidence. The purpose here is to best utilize existing information to allow us to conduct a confirmatory study at a fraction of the sample size required in a full-scale trial. For bridging studies, since drug effect has been established in the foreign region, emphasis should be put upon the bridging region $\alpha$-level, not the unconditional type I error for the combined weighted Z-test. More details are given in Section 4.

2. NEED FOR ADAPTIVE DESIGN

In the past, applications of statistical methods to scientific research have proven helpful, however, in drug development many statistical methods in clinical trial design and data analysis need modifications. For example, in clinical trials, endpoints appropriate for treatment assessment are chosen based on clinical judgment. This assessment may sometimes involve multiple endpoints, but very often we do not know how to pool them into one index. Therefore, we gamble on one by picking a primary endpoint, and apply available statistical methods to the selected primary endpoint at the end of the study. Additional information gathered during the course of the study might trigger changing the primary endpoint. Statistical mechanism does not seem to be well established for this purpose. Even when adaptation was applied, information from other endpoints might not have been adequately taken into consideration for evaluation of the treatment.

In many practical situations, the design of a clinical trial is based on prior knowledge of the studied disease. The prior information might include the theory of a mechanism (e.g., medical, biological, genetic, and pharmacological) and the real data (pre-clinical or clinical). When the original idea (motivation of the study) is not appropriate, or the prior information is non-informative, it is natural to evaluate the data at interim stage and use the updated information to modify the design of the study. For example, the risk of linking the practice of medicine to what seems logical
comes from the assumption made in the 1970s and 1980s that reducing premature ventricular contractions would lower arrhythmic deaths. To our surprise, we learned from the Cardiac Arrhythmia Suppression Trial (CAST), that the drugs slightly effective in reducing premature ventricular beats actually increased patient mortality due to proarrhythmia (The Cardiac Arrhythmia Suppression Trial, 1989; Echt et al., 1991). In April 1989, the Data and Safety Monitoring Board (DSMB) for CAST recommended discontinuing encainide and flecainide, the two drugs that seemed to be associated with the excess mortality. Two years later, they recommended stopping the third drug, moricizine (The Cardiac Arrhythmia Suppression II Investigators, 1992). The early termination protected additional patients from being put at risk, and allowed early dissemination of results to physicians. The complexity of the study design and a mid-course protocol modification raise several data-monitoring issues not previously discussed. These include how to handle apparently dramatic yet unexpected results, the need for flexibility in modifying study design and goals, and the conflict between existing study data and both conventional wisdom and medical practice.

In medical research, modification of the design of an experiment based on accrued data has been in practice for hundreds, if not thousands, of years. In the past, we had a tendency to adopt statistical procedures available in the literature and apply them directly to the design of clinical trials. Because those procedures were not motivated by clinical trial practice, they may not be the best tools to handle certain situations.

Let \( \Delta \) be the treatment effect of a new drug. For the comparisons of two means, \( \Delta = (\mu_T - \mu_C)/(\sqrt{2}\sigma) \), and for the comparisons of two survival distributions using the log-rank test, \( \Delta = 1/\sqrt{2} \log(\text{hazard ratio}) \). For a 20% hazard reduction, \( \Delta = 0.1578 \). In our experience, we almost always dealt with a |\( \Delta | \leq 3 \). Problems arise in the sample size calculation of clinical trials using either the frequentist approach or the Bayesian approach:

1. The frequentist approach requires that we specify a test procedure and a partition of the sample space \( \Omega \) into two sets: \( C \), the critical region, and \( A \), the acceptance region. Furthermore, we assume that the observed treatment difference can be summarized by a test statistics, \( Z \), the distribution of which usually can be approximated by \( N(\eta\Delta, 1) \), where \( \Delta \) is the treatment effect and \( \eta \) depends on the amount of information in the sample. Under the null hypothesis \( H_0: \Delta = 0 \). Type I error occurs if the null hypothesis \( H_0 \) were true, but it is rejected. The hypothesis testing method, thus, yields a decision to reject \( H_0 \) or not, depending on whether \( Z \) falls within the critical region \( C \). For one-sided type I error rate \( \alpha = 0.025 \), \( C = \{Z \geq 1.96\} \). For a given treatment effect \( \Delta > 0 \) and desired power = 100(1 - \( \beta \))%, sample size per group \( N \) is determined by solving the equation \( E(Z|\Delta) = \sqrt{N}\Delta = z_\alpha + z_\beta \), where \( z_\alpha \) (\( z_\beta \)) is the upper \( \alpha \)-quantile (\( \beta \)-quantile) for the standard normal distribution.

Ideally, we try to find an accurate estimate of treatment effect \( \Delta \) at time 0, which is almost implausible. In practice, if \( \Delta \) can be estimated accurately in advance, then there is no need to start the clinical trial; and if we truly believe that \( \Delta > 0 \), it is not ethical to randomize patients into the placebo group. It has been recommended that we may use a “minimum meaningful difference \( \Delta_0 \)” to design a trial, but that approach has some problems as well, since the choice of \( \Delta_0 \) may change during the
course of the study. From the clinical perspective, sometimes it may not be possible
to pre-specify such a $\Delta_0$. In public health, even a small value of positive $\Delta_0$ would
be important. Dr. Robert O’Neill of the FDA once made a comment, “I haven’t
seen a (beneficial) effect that I didn’t like.” From the financial perspective, $\Delta_0$ is the
minimum value of benefit for the product to get a fair market share. However, if
the sponsor observes much less toxic effect during interim analysis than expected,
the value of $\Delta_0$ may change.

Even in the ideal situation when $\Delta_0$ can be pre-specified, the design of the
study will still have problems. Note that if we use the traditional way of designing
a study with one-sided $z = 0.025$ and power of $1 - \beta = 85\%$, for a given treatment
effect $\Delta_0$, we choose a sample size for each treatment group $N$ so that the test
statistic $Z$ has an expected value $E(Z|\Delta_0) = z_\alpha + z_\beta = 1.96 + 1.04 = 3$. For a fixed
design of the study, if the final observed treatment effect $\hat{\Delta} = \Delta_0$, then the observed
$Z = 3$, as expected. If $\hat{\Delta} = 0.8\Delta_0$, then $Z = 0.8 \times 3 = 2.4$, which is still greater than
the critical value 1.96. However, since we specified $\Delta = \Delta_0$ as the minimum clinically
meaningful difference and the observed $\hat{\Delta} < \Delta_0$, we cannot declare this a positive
study. In general, if the minimum clinically meaningful difference $\Delta_0$ is specified
to design a study, then the real “power” of the test is $P(Z \geq 1.96$ and $\hat{\Delta} \geq \Delta_0) \leq
P(\hat{\Delta} \geq \Delta_0) \approx 0.5$ when $\hat{\Delta}$ is symmetric with respect to mean $\Delta_0$. Note that in the
large sample case, the maximum likelihood estimate $\hat{\Delta}$ is $\approx N(\Delta, \xi^2)$. The classical
power of $P(Z \geq 1.96)$ is for the test of $\Delta > 0$, not for $\Delta \geq \Delta_0$. We do not have a
good solution for this problem unless the real treatment effect $\Delta$ is bigger than $\Delta_0$,
and the difference $\Delta - \Delta_0$ is moderate. If $\Delta$ cannot be evaluated satisfactorily at the
beginning, how do we choose the sample size of the study? We will discuss this in
the next section, where after an interim unblinded data analysis, sample size could
be re-estimated.

2. In contrast, the **Bayesian approach** considers the treatment effect $\Delta$ being
random that has distribution known as the prior distribution. This prior distribution
is combined with data to form the posterior distribution. The new drug is concluded
to be beneficial if, under the posterior distribution, $P(\Delta > 0)$ is “high” (e.g., 95%).
The sample size is determined such that a specified desired power is reached. To
simplify discussion, we consider normal prior and normal responses. Specifically, let
$X_1, X_2, \ldots, X_{100}$ be i.i.d. $N(\Delta, 1)$, then the sample mean

$$
\bar{X} = \frac{X_1 + X_2 + \cdots + X_{100}}{100} \sim N(\Delta, 1/100)
$$

Assume the prior distribution of $\Delta$ be $N(0, \tau^2)$, then the posterior distribution of $\Delta$
given $\bar{X}$ is

$$
\Delta \mid \bar{X} \sim N\left(\frac{\tau^2}{\tau^2 + 1/100} \bar{X}, \frac{\tau^2/100}{\tau^2 + 1/100}\right)
$$

(2)

The posterior belief of the treatment effect is not a single number, rather it is
a random variable. The most likely posterior point estimate of $\Delta$ is $\frac{\tau^2}{\tau^2 + 1/100} \bar{X}$, which
is less than the sample mean $\bar{X}$. We find that this is hard to convince the clinicians,
especially when $\tau^2$ is smaller. On the other hand, if we let $\tau^2 \to \infty$, then $E(\Delta \mid \bar{X}) \to \bar{X}$. 

THE USE OF WEIGHTED Z-TESTS

629
Does it make more sense to let $\tau^2 \to \infty$? When $\tau^2 \to \infty$, say $\tau^2 = 1,000,000$, then unconditionally $P(\Delta > 3) = P(N(0, 1) > 0.003) \approx 0.5$ and $P(\Delta < -3) \approx 0.5$. That is, the prior belief of the treatment effect is arbitrarily vague (or non-informative) and on the average is 0. On the other hand, the new drug under development is either “extremely beneficial” or “extremely harmful”, which is very informative. Note that in our experience, $|\Delta| \leq 3$. If the investigator’s prior belief justifies $\tau^2 \to \infty$, then the patients ought to be informed of this belief before randomization.

In some cases, ideas motivated from a frequentist approach can be translated into a Bayesian argument, and vice versa. For a frequentist procedure, one can construct an equivalent Bayesian procedure that, having implicitly chosen a prior distribution, yields the same rejection region. Such a partition of sample space will depend on the choice of prior distribution in Bayesian paradigm. Using the previous example, without prior information, in a hypothesis testing setting we conclude that the new drug is effective if the test statistic $Z = \sqrt{100} \ X \geq 1.96$ for a one-sided $\alpha = 0.025$. If $\tau^2 = 0.0238296$, then the Bayesian decision rule $P(\Delta > 0) \bar{S}(Z) \geq 95\%$ is equivalent to the frequentist one, with the same partition of the sample space. In addition, since we are more familiar with the frequentist approach, we will discuss in the following sections only the hypothesis testing approach.

## 3. ADAPTIVE DESIGN FOR SAMPLE SIZE RE-ESTIMATION

The determination of sample size with adequate power is an integral part of clinical trial design. Incorrect assumptions about the effect size $\Delta$ may lead to lack of power to detect treatment effect and, hence, increase the likelihood of making false negative conclusions (type II error). Sample size re-estimation methods can be applied to determine a revised sample size estimate based on the interim findings. However, changing the sample size based on unblinded interim results may inflate the type I error rate. Several approaches have been investigated to control the type I error rate, namely, the use of interim variance estimate (Wittes and Brittain, 1990), the use of weighted $Z$-tests, and the conditional power approach (consider stopping early for futility), among others. When observations are i.i.d. the idea of applying the weighted $Z$-test for sample size re-estimation has been given by Fisher (1998), Shen and Fisher (1999), and Cui et al. (1999).

As indicated in Section 1, mathematically it is easier to consider the one-sample problem that $X_1, X_2, \ldots, X_n, \ldots, X_N$ are i.i.d. $N(\Delta, 1)$, and test the hypothesis $H_0 : \Delta = 0$ vs. $H_1 : \Delta > 0$ with a one-sided type I error rate $\alpha$. Note that for two-sample comparison with equal allocation of patients into two group, $\Delta = (\mu_T - \mu_C)/\sqrt{2\sigma}$ and $N$ is the sample size for each group. The test statistic $Z = \sum_{i=1}^{N} X_i / \sqrt{N}$, as given in Eq. (1), can be expressed as

$$Z = \sqrt{t} Z_{[1\to n]} + \sqrt{1-t} Z_{[n+1\to N]}$$

where $t = n/N$, and

$$Z_{[1\to n]} = \frac{X_1 + X_2 + \cdots + X_n}{\sqrt{n}} \quad \text{and} \quad Z_{[n+1\to N]} = \frac{X_{n+1} + X_{n+2} + \cdots + X_N}{\sqrt{N-n}}$$
are two independent standard normal random variables, under $H_0$. This $Z$-statistic is “unweighted” in the sense that each patient contributes the same weight to the evaluation of test.

For example, with prior information obtained from sample \{X^*_1, X^*_2, \ldots, X^*_n\} ~ i.i.d. $N(\Delta, 1)$ and $\sum X^*_i = 22.475$ (or $Z^* = 2.2475$), let \{X_1, X_2, \ldots, X_{100}\} ~ i.i.d. $N(\Delta, 1)$ be the new observations, then the critical region using the frequentist approach is

$$Z = \sum_{i=1}^{100} X_i / \sqrt{100} \geq 0.5244$$

(derived from $(22.475 + \sum_{i=1}^{100} X_i) / \sqrt{200} \geq 1.96$ for a one-sided $\alpha = 0.025$), which corresponds to a new type I error rate of 0.30. While, by bringing in the prior information, the Bayesian decision rule $P(\Delta > 0 | [Z^*, X_1, X_2, \ldots, X_{100}]) \geq 95\%$ is equivalent to

$$Z = \sum_{i=1}^{100} X_i / \sqrt{100} \geq 0.3113$$

when $\tau^2 = 0.0238296$ (derived from $(22.475 + \sum_{i=1}^{100} X_i) / \sqrt{200} \geq \sqrt{\frac{200\tau^2 + 1}{200\tau^2}} \times 1.645$). That is, when the prior information is informative, the frequentisists use a larger type I error rate for testing, while the Bayesians use a “favorable” prior. The frequentist and the Bayesian approaches differ in the partition of the sample space.

In general, the one-sample test statistic $Z$, which is a $N(0, 1)$ random variable under $H_0$, can be expressed as

$$Z_w = \sqrt{w} Z_I + \sqrt{1 - w} Z_{II}$$

(3)

where $w$ is a constant weight ($0 \leq w \leq 1$), and $Z_I$ and $Z_{II}$ are two independent standard normal random variables, under $H_0$. The distribution of $Z_w$ is unchanged even when we replace $Z_{II}$ by another $Z^{*}_{II}$, independent of $Z_I$, or change the weight $w$ to another fixed constant. A specific example is when an interim analysis produces a $Z_{[1\rightarrow n]}$ which may trigger sample size modification. Suppose a small treatment effect is observed during interim analysis. If the trend continues, there is only a small chance of concluding a positive study at the end of the trial. Increasing the sample size will give us a better chance for a statistically significant finding. If the sample size is increased from $N$ to $N^{*} (> N)$, based on findings in $Z_{[1\rightarrow n]}$ evaluated at time $t = n/N$, the weighted $Z$-test, $Z_w$, is defined as

$$Z_w = \sqrt{w} Z_{[1\rightarrow n]} + \sqrt{1 - w} Z_{[n+1\rightarrow N^{*}]}$$

(4)

which is also distributed as $N(0, 1)$ under the null hypothesis $H_0 : \Delta = 0$. If $N^* > N$, then the weighted $Z$-test assigns less weights to patients enrolled after the decision to increase the sample size than to those enrolled before the decision. The weighted $Z$-test is flexible in that the two $Z$ components are independent of each other and the weight $w$ is not dependent on the findings in $Z_{[1\rightarrow n]}$ evaluated at time $t$, although the increased sample size $N^*$ may depend on $Z_{[1\rightarrow n]}$. $N^*$ may be determined by conditional power consideration, and the computational formula is given in Wang...
et al. (2002). In this setting, $H_0$ is rejected if the weighted $Z$-statistic $Z_w$ lies within the original critical region and the type I error rate is preserved at the nominal level $z$.

Increasing the sample size based on unblinded interim result may inflate the type I error rate when using the unweighted $Z$-statistic approach. The weighted $Z$-test, on the other hand, controls the type I error rate and is extremely flexible, although it violates the one-patient-one-vote principle (an equal weight for each patient), which is preferred by some researchers. Lan and Trost (1997) and Chen et al. (2004) suggested consideration of both extending the trial and stopping the trial early due to futility. If the type I error inflation due to sample size increase is smaller than the type I error reduction from allowing early stopping due to futility, then the overall type I error rate can be preserved. More details of this approach are given in Section 6. The weighted $Z$-statistic is more advantageous over the unweighted $Z$-statistic with futility stopping criteria in that a consistent estimator of the treatment effect and a confidence interval can be constructed to make inference consistent with use of the weighted $Z$-statistic (Hung et al., 2002; Lawrence and Hung, 2003).

4. BRIDGING EVIDENCE TO NEW POPULATIONS

ICH E5 Guideline (International Conference on Harmonization, 1998) describes the use of a bridging study to allow extrapolation of foreign clinical data to a new region. Previous works in this area are based on the consistency approach (Shih, 2001), the equivalence and non-inferiority method (Liu, 2003), and the population bioequivalence approach (Chow et al., 2002). Alternatively, Ware et al. (2002) proposed the use of mixed effect model and Empirical Bayes estimator to “borrow strength” in other countries to estimate country-specific treatment effect. In this section, we propose a different approach, which allows the local bridging study to be analyzed separately from the original study by using the adjusted type I error determined from the weighted-$Z$ and the discount factor methods.

Consider the problem of planning a bridging study when efficacy data $Z_I = Z_{\text{Foreign}}$ is already available from the approval package in foreign countries. The principle of borrowing evidence $Z_{\text{Foreign}}$ from prior data to determine the acceptable weight $w$ for the new bridging data $Z_{II} = Z_{\text{Bridging}}$, can be applied to modifying the “traditional” $z$ (e.g., 0.025 one-sided) to a higher value ($z_{Bridge}$) for bridging studies. Let $Z_{\text{Bridging}}$ be the test statistic for the bridging study, which is $N(0, 1)$ under the null hypothesis $H_0$: $\Delta = 0$. The weighted $Z$-test with constant weight $0 \leq w \leq 1$, is expressed as

$$Z_w = \sqrt{w} Z_{\text{Foreign}} + \sqrt{1-w} Z_{\text{Bridging}} \quad (5)$$

In this setting, the weighted $Z$-test combines two $Z$-values from two different studies conducted in different ethnic populations, with constant weight $w$ determined by prior knowledge and clinical judgment regarding the impact of extrinsic and intrinsic factors on the inter-region variability. For $Z_w > z_z$, conditional on the observed value of the foreign study, $z_{\text{Foreign}}$,

$$Z_{\text{Bridging}} = (Z_w - \sqrt{w} z_{\text{Foreign}}) / \sqrt{1-w}$$

$$> (z_z - \sqrt{w} z_{\text{Foreign}}) / \sqrt{1-w} \quad (6)$$
which implies

\[ \alpha_{\text{Bridging}} = 1 - \Phi((z_{\alpha} - \sqrt{w} z_{\text{Foreign}}) / \sqrt{1 - w}) \]  

(7)

For example, for \( w = 0.5 \), when the significance level \( \alpha = 0.025 \) (i.e., \( z_{\alpha} = 1.96 \)) and the observed \( z_{\text{Foreign}} = 2.2475 \), the critical region of the bridging study will be \( z_{\text{Bridging}} \geq 0.5244 \), which associates with a significance level of \( \alpha_{\text{Bridging}} = 0.30 \). That is, by incorporating prior information from the foreign study, the bridging study may be conducted with a smaller sample size and tested at a higher significance level as compared with those required by a full-scale study. We would like to point out that the bridging \( \alpha \)-level of 0.30 in this example may not be acceptable to the local regulatory agency which, for example, may require a maximum local \( \alpha \)-level of 0.10.

Alternatively, given power of \( 1 - \beta \), the sample size for each group in a full-scale study is determined by

\[ N_{\text{Full}} = \frac{1}{\Delta^2} (z_{\alpha} + z_{\beta})^2 \]  

(8)

If the sample size per group for a bridging study, \( N_{\text{Bridging}} \), is determined based on a discount factor \( D \) \((0 < D < 1)\), such that

\[ N_{\text{Bridging}} = D \times N_{\text{Full}} \]  

(9)

then

\[ N_{\text{Bridging}} = D \times N_{\text{Full}} = \frac{1}{\Delta^2} (z_{\alpha_{\text{Bridging}}} + z_{\beta})^2 \]  

(10)

These imply

\[ \alpha_{\text{Bridging}} = \Phi(z_{\beta} - \sqrt{D(z_{\alpha_{\text{Bridging}}} + z_{\beta})}) \]  

(11)

For example, when the desired power is 85\%, that is, \( \beta = 0.15 \), and the significance level \( \alpha_{\text{Full}} = 0.025 \), with the discount factor of 0.8 (i.e., 20\% discount), the bridging study will need to be tested at the significance level of \( \alpha_{\text{Bridging}} = 0.0502 \approx 0.05 \). For the discount factor of 0.6 (i.e., 40\% discount), the significance level would be \( \alpha_{\text{Bridging}} = 0.0996 \approx 0.1 \).

Sometimes the local regulatory authorities may request conducting a bridging study to detect a fraction of the treatment effect observed in a foreign study, while agreeing with a smaller sample size. A different discount factor with a similar idea applies here as well. Assume \( N_{\text{Foreign}} = 5000 \) for each treatment group and the observed treatment effect \( \Delta = 0.06 \) with an extremely impressive \( Z \)-value of 4.24. Suppose the anticipated treatment effect in the local region is 75\% of \( \Delta \), that is, \( \tilde{\Delta} = 0.06 \times 0.75 = 0.045 \), and a discount factor of \( D = 0.8 \) is used for determining the sample size of each group in the bridging study, that is, \( N_{\text{Bridging}} = 5000 \times 0.8 = 4000 \). Applying Eq. (10), we have \( 4000 = \frac{1}{0.045^2} (z_{\alpha_{\text{Bridging}}} + z_{\beta})^2 \) or \( z_{\alpha_{\text{Bridging}}} + z_{\beta} = 2.84605 \). When \( \alpha_{\text{Bridging}} \) is chosen to be 0.1, power = 94.1\%. For another choice of \( \alpha_{\text{Bridging}} = 0.05 \), power = 88.5\%. A two-way table of \( \alpha_{\text{Bridging}} \) and power can be created and an \( \alpha \)-level may be negotiated between the sponsor and the local regulatory agency.
5. CHOICE OF WEIGHT \( w \) AND DISCOUNT FACTOR \( D \)

The weighted Z-test can be used to weigh two Z-values from two stages of a trial, and can also be used to weigh two Z-values from two independent studies or in the case of combining Phase IIB and III studies. The weight, \( w \), may be determined based on the observed \( Z_I \) and other findings obtained from previous studies. For the case of bridging studies, the values of \( Z_I \) may be derived from evidence established in foreign regions, while \( Z_{II} \) is the Z-test evaluated for the new bridging study in the local region. The amount of bridging information required depends on the established evidence and the given weight, \( w \), as well as the degree of certainty on the variability of treatment effects among regions.

Two approaches have been described for determining the value of increased type I error rate: the weighted Z-statistics and discount factor method. They correspond to choosing weight, \( w \), and discount factor, \( D \), respectively. Values of \( w \) and \( D \) should reflect our degree of consensus on the expected amount of variability in treatment effects among regions concerned.

This should guide our choice of default priors for weight, \( w \), and discount factor, \( D \). The goal is to facilitate the acceptance of foreign data to minimize duplication of clinical studies and expedite the availability of new therapies to patients for their benefit. The type of required bridging study depends on experience with a particular drug class, intrinsic factors, and extrinsic factors. Intrinsic factors identified in ICH E5 which might have a bearing on dosage, dose regimen, efficacy, and toxicity profiles across different populations, include the genetic, physiological, and pathological properties specific to particular ethnic groups or populations. Likewise, extrinsic ethnic factors pertain to prevailing medical practice, treatment guidelines, and cultural characteristics that vary from region to region. ICH E5 consensus prescribes that randomized clinical trial (RCT) would be necessary when there are:

- Doubts on the appropriate dose or dosage choices for patients from the targeted local population due to relevant intrinsic factors
- Predominant dissimilar extrinsic factors between the local and referenced regions
- New treatment or NCE belongs to unfamiliar drug class
- Insufficient, existing clinical experience with drugs of similar class

Table 1 describes the relationship between the choices of \( w \) and \( D \) and bridging strategy.

For the case of dealing with samples from two subpopulations, assume \( \Delta_1 > 0 \) and \( \Delta_2 > 0 \) under \( H_n \), but \( \Delta_1 \) is suspected to be larger than \( \Delta_2 \), where \( X_1, X_2, \ldots, X_n \sim \text{i.i.d.} \ N(\Delta_1, 1) \) and \( X_{n+1}, X_{n+2}, \ldots, X_N \sim \text{i.i.d.} \ N(\Delta_2, 1) \), where \( n \) and \( N - n \) are the sample size per group for each of the two subpopulations.

<table>
<thead>
<tr>
<th>( w )</th>
<th>Bridging Strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>No bridging is required</td>
</tr>
<tr>
<td>0 &lt; ( w &lt; 1 ) or 0 &lt; ( D &lt; 1 )</td>
<td>Determine sample size and increase ( \alpha ) for bridging study</td>
</tr>
<tr>
<td>( w = 1 )</td>
<td>Replicate the original study conducted in foreign region</td>
</tr>
</tbody>
</table>
The weighted Z-statistic

\[ Z_w = \sqrt{w}Z_{[1\rightarrow n]} + \sqrt{1-w}Z_{[n+1\rightarrow N]} \]

has mean value

\[ E(Z_w) = \sqrt{w}\sqrt{n}\Delta_1 + \sqrt{1-w}\sqrt{N-n}\Delta_2 \]

The weight, \( w \), may be determined follows.

- For given \( n \) and \( N \), choose \( w \) such that \( E(Z_w) \) is maximized, that is,

\[ w = \frac{n\Delta^2_2}{n\Delta^2_2 + (N-n)\Delta^2_2} \tag{12} \]

- Combining two subpopulations (e.g., high/low risk, male/female, adult/pediatric, etc.) makes medical sense. In order to have good representation, each subpopulation should have sufficient number of patients included, for example, for each treatment group \( n \geq n_0 \).
- The choices of \( w \) cannot be solely a statistical decision. It requires medical and biological judgments, and regulatory authority’s agreement and acceptance as well. From local regulatory point of view, the maximum tolerable type I error rate may not be greater than 0.1. In that case, the proposed approach does provide a basis for discussion and negotiation between sponsors and regulatory agencies.

### 5.1. Clinical Assessment of Clinical Data Package (CDP)

Once the sponsor and a particular regulatory authority have agreed that a bridging study (or studies) based on RCT is necessary, both parties should then proceed and negotiate on how much bridging information (\( N_{\text{Bridging}} \) for each treatment group based on \( w \) and \( D \)) to collect in the new region. These should be based on the nature, quality, and strength of evidence from the body of knowledge obtained in the Clinical Data Package (CDP). There are several strategies to choose the weighting and discount factor:

1. Assign \( t(=n/N) \) to \( w \), the weight assigned to Z-value determined from evidence derived from prior foreign studies. Fisher (1998) called it the variance spending function.
2. The value assigned to weight, \( w \), can be based on subjective judgment or the degree of certainty one has on the expected amount of differences in treatment effect between the foreign and local regions. Based on our prior knowledge, if weight, \( w \), is proportional to the variability of treatment effects within the studies and between the two regions, then \( w \) is similar to the weight used in obtaining the posterior mean of combined effect in Bayesian approach.
3. Following ICH E5, the weight, \( w \), should be based on the body of knowledge obtained from clinical assessment of CDP in existing dossiers, of drugs in the same class, and both intrinsic and extrinsic factors. In this complex situation, it may be unrealistic to assume that a quantitative function for \( w \) or \( D \) based on
theoretical work can adequately simulate reality. A qualitative approach based on clinical assessment and regulatory review of CDP may be adopted to assign values to $w$ and $D$. The theoretical work should instead provide guidance on which subjective judgment could be made about $w$ and $D$.

5.2. Impact on the Design and Analysis of Local Bridging Studies

ICH E5 proposes the use of bridging studies to allow extrapolation of clinical data from one ethnic population to another. However, it does not provide specific statistical framework for the design and analysis of bridging trials. In ICH E5, a bridging study is defined as a supplemental study performed in a new region to provide pharmaco-dynamic or clinical data on efficacy, safety, dosage, and dose regimen that allow extrapolation of foreign clinical data to the new region.

In this paper, we describe a new paradigm for the design and analyses of bridging studies to facilitate cross approval in non-ICH countries, or within ICH regions (the United States, European Union, and Japan). This is based on the premise that substantial evidence of a new medication’s efficacy and safety has already been established when sponsors won marketing approval in the US or European Union. By borrowing evidence from these established data, this provides for the concept of using a bridging study, with sample size at a fraction of that required in the original, full-scale study, to demonstrate similar efficacy and safety profiles in other ethnic populations.

Key challenges in designing local bridging studies include variations in endpoints, medical practice, and target population in local as compared to foreign regions where original studies were conducted. Subjective endpoints (e.g., rating scales used in CNS trials) may be influenced by cultural differences. The use of a placebo arm may not be acceptable to local medical practice. Variations in standard treatment guidelines can also have an impact on the choice of appropriate control in comparative studies. Often, the target population may not be identical to those in the original study due to variations in inclusion and exclusion criteria.

To address these design issues, the proposed, weighted $Z$-test provides a flexible framework that allows $Z$-tests from two different studies with possibly different designs to be combined in a simple manner. Once both sponsor and local regulatory authority have agreed on the level of increased $\alpha$, and the value of weighting or the discount factor, we show how sample size for the bridging study can easily be determined. The sample size to be adopted reflects the agreement between sponsor and local regulatory authority, based on the nature, quality, and strength of evidence borrowed from the body of knowledge, obtained in the approved clinical data package in the referenced country or region.

6. DISCUSSION

One of us served in NHLBI-NIH and was involved in animal experiments. Typically, the researcher requested a certain number of rats, say 100, for an experiment. The Animal Care and Use Committee, however, agreed to give the researcher only 50 rats to start with, and would give the researcher another 50 rats if the results from the first 50 rats was “promising”. If the results were not promising, then the experiment should be stopped. This is the way the committee dealt with the
THE USE OF WEIGHTED Z-TESTS

situation for financial and ethical reasons, when prior data did not exist. What if the results are somewhat promising? What measure can the researcher take to determine whether the results are “promising”? It is natural to calculate the probability of concluding a positive study at the end of a trial given the interim results, which is known as the conditional power (Lan and Wittes, 1988). Conditional power is a useful tool for adaptive design and sequential analyses that allows modification of sample size, or early stopping due to futile or beneficial outcomes.

Increasing the sample size based on unblinded interim result may inflate the type I error rate when using the unweighted $Z$-statistic approach. Lan and Trost (1997) and Chen et al. (2004) suggested consideration of both extending the trial, continuing as planned, and stopping the trial early due to futility during interim analysis. If the type I error inflation due to sample size increase is smaller than the type I error reduction from allowing early stopping for futility, then the overall type I error rate can be preserved. Extension of this approach to create one single interim analysis rule for both early termination and sample size re-estimation can be found in Siu and Lan (2001). It is worth noting that the futility boundary needs to be specified in advance in this unweighted $Z$ approach for sample size re-estimation.

The weighted $Z$-test introduces a convenient way to re-estimate sample sizes in clinical trials. It is flexible, in that the two $Z$ components are independent of each other and the weights are not dependent on the findings in the interim analysis evaluated at the information time $t$, although the increased sample size for each treatment group $N^*$ may do. $H_0$ is rejected if the weight $Z$-statistic $Z_w$ lies within the original critical region and the type I error rate is exactly at the nominal level $\alpha$. The weighted $Z$-statistic approach, unlike the unweighted $Z$-statistic approach, is flexible in that the decision to increase the sample size and the magnitude of sample size increment are not mandated by pre-specified rules. In either approach, if the calculated $N^*$ is too large, then modify $N^*$ or lower the desired conditional power and recalculate the sample size. Because the desired conditional power does not need to be fixed in advance, increasing sample size will not reveal the interim observed $Z_{[1-n]}$ value, therefore, the concerns of data integrity and potential bias can be addressed satisfactorily.

For a clinical trial sponsored by a pharmaceutical company, it is beneficial for the sponsor to have access to interim analysis results as soon as possible so that appropriate actions can be taken, such as modifying the design of the study for a more efficient evaluation, as discussed in this paper. The interim review of data is usually conducted by an independent DSMB not directly involved with the conduct of the trial. Unfortunately, this current practice does not allow the sponsor to have easy access to the interim results primarily because of the FDA’s concerns in (1) the control of the type I error rate, and (2) the bias the interim analyses may introduce. As pointed out by Lan (2003), the former concern can be handled by employing appropriate statistical procedures, but the latter has no simple solution, as no statistical procedures could possibly correct the bias caused by the practice of clinicians of the study when they are aware of or suspect the treatment effect. Ironically, the extremely valuable first hand information obtained at the DSMB meetings would be beneficial to the DSMB members in their own research (either in another DSMB or in their own laboratories or clinical centers). As the sponsor of a trial, a company deserves at least an equal right to access the information, although it may take an innovative approach to reach this goal. To ensure that the
sponsor has access to all the crucial interim analysis information and to address the concerns of the FDA, Lan (2003) recommended that selected, trained company employees be allowed to attend DSMB closed sessions, and important information of development should be communicated to some designated upper managers of the sponsor. All these should be governed by the FDA agreed standard operating procedures to ensure that the sensitive information obtained during DSMB meetings does not leak.

By promoting the use of adaptive designs we do not intend to encourage careless planning for a study and then using an adaptive design to rescue it. Suppose careful planning would yield a sample size of 300 for a certain study. If it starts with poor planning with \( N = 200 \) and then extends it to \( N^* = 300 \) using a weighted \( Z \)-test, then this design is not as efficient as the traditional design with \( N = 300 \) using a standard \( Z \)-test. However, we would like to point out that the loss of efficiency is due to poor planning, not to the use of an adaptive design.

ACKNOWLEDGMENTS

We learned the problem of a very flat prior from Dr. Michael Proschan of the National Heart, Lung, and Blood Institute, NIH. We would like to thank Dr. James Hung, who made some insightful comments to our earlier draft.

REFERENCES


