Version 1

Mediana Designer Manual

1 Introduction

Mediana Designer is a free Windows-based software tool that supports traditional and simulation-based power/sample size calculations in fixed-sample and group-sequential trials.

The current version of Mediana Designer comes with the following three module:

- TraditionalSampleSize module (Analytical calculations in fixed-sample trials): This module supports analytical evaluation of operating characteristics in clinical trials with a fixed-sample design.
- TraditionalSimulations module (Simulation-based calculations in fixed-sample trials): This module implements the simulation-based Clinical Scenario Evaluation approach in clinical trials with a fixed-sample design.
- GroupSequential module (Analytical and simulation-based calculations in group-sequential trials): This module implements analytical and simulation-based evaluation of operating characteristics in clinical trials that employ group-sequential designs with several decision points.

This manual provides a detailed summary of statistical methods implemented in each module.

Developer

Mediana Designer was developed by and is maintained by Mediana Inc. For more information on Mediana Designer, please visit the Biopharmaceutical Network site at

`http://biopharmnet.com/mediana-designer/`

The latest version of the manual can be downloaded from this web site.

Reviewers

Beta versions of Mediana Designer have been reviewed by the following biopharmaceutical statisticians (in alphabetical order):

Thomas Brechenmacher (IQVIA), Jian Chen (Tesaro), Qiqi Deng (Boehringer Ingelheim), Miguel Garcia (Boehringer Ingelheim), Wei Guo (Tesaro), Jim Love (Boehringer Ingelheim), Yi Liu (Boehringer Ingelheim), Kaushik Patra (Alexion), Gautier Paux (Sanofi), Dooti Roy (Boehringer Ingelheim), Kyle Wathen (Johnson & Johnson), Ron Yu (Gilead).

The development team would like to thank the reviewers for the valuable feedback.

Validation

Multiple commercially available and open-source software tools have been used to test the implementation of statistical methods presented in this manual. This includes:

- EAST software (Cytel, 2016).
- SAS software, e.g., PROC POWER (SAS, 2017).
- R packages, e.g., TrialSize (Zhang et al., 2013) and gsDesign (Anderson, 2016).

For a detailed summary of procedures that were carried out to test Mediana Designer, please visit

`http://biopharmnet.com/mediana-designer-validation/`

Notation and conventions

The following notation will be used throughout this manual:

- $\Phi(x)$ is the cumulative distribution function of the standard normal distribution.
- z_{1-x} is the upper $100x$th percentile of the standard normal distribution, i.e., $z_{1-x} = \Phi^{-1}(1-x)$.

In addition, a clinical trial will be designed to ensure power of $1-\beta$ (e.g., 90% power with $\beta = 0.1$) with a one-sided Type I error rate set to α (e.g., $\alpha = 0.025$).

2 TraditionalSampleSize module: Analytical calculations in fixed-sample trials

2.1 Introduction

This section focuses on sample size and power calculations in clinical trials that utilize a traditional design with a fixed number of patients (if the primary endpoint is continuous or binary) or a fixed number of events (if the primary endpoint is a time-to-event endpoint).

A two-arm clinical trial with a parallel design will be assumed. Let n_1 and n_2 denote the number of patients enrolled into the control arm and the treatment arm,

respectively. The total number of enrolled patients is denoted by $n = n_1 + n_2$. The randomization ratio (r) is defined as the ratio of the number of patients in the treatment arm to that in the control arm. For example, with $r = 2$, twice as many patients are assigned to the treatment arm compared to the control arm. In other words, $n_2 = rn_1$ and thus

$$n_1 = \frac{n}{1+r} \text{ and } n_2 = \frac{rn}{1+r}.$$

In addition, d will denote the target number of events in trials with a time-to-event endpoint.

Analytical frequentist approaches to sample size and power calculations are discussed in Sections 2.2 through 2.4 (these sections describe trials with continuous, binary and time-to-event endpoints). Within the frequentist framework, Mediana Designer supports the following calculations:

- Clinical trials with continuous or binary endpoints: Calculate power for a given sample size or calculate the required sample size for a given power level.
- Clinical trials with time-to-event endpoints: Calculate power for a given number of events or calculate the target number of events for a given power level. If the patient accrual and dropout parameters are specified, calculate power for a given sample size or calculate the required sample size for a given power level

A simulation-based Bayesian approach to evaluating assurance (probability of success) in clinical trials is presented in Sections 2.5. Mediana Designer supports the following assurance calculations:

- Clinical trials with continuous or binary endpoints: Calculate assurance for a given sample size.
- Clinical trials with time-to-event endpoints: Calculate assurance for a given number of events.

2.2 Frequentist calculations in trials with continuous endpoints

It is assumed that the continuous primary endpoint is normally distributed and the treatment effect is evaluated using the standard Z-test. Power and sample calculations based on this popular test are supported in several R packages (TrialSize and gsDesign), POWER procedure in SAS and EAST.

The true values of the mean effects in the control arm and treatment arm are denoted by μ_1 and μ_2, respectively, and similarly, the standard deviations in the control arm and treatment arm are denoted by σ_1 and σ_2. The mean treatment difference is given by $\delta = \mu_2 - \mu_1$.

The following two scenarios that correspond to two different alternative hypotheses of a beneficial effect will be considered in this section (as well as other sections):

- Upper one-sided alternative: A positive value of the mean treatment difference corresponds to a beneficial treatment effect.
- Lower one-sided alternative: A negative value of the mean treatment difference indicates treatment benefit.

Superiority assessment

Consider first the superiority setting where the goal is to demonstrate that the treatment provides a statistically significant improvement over the control. The

hypothesis testing problem with an upper one-sided alternative is defined as

$$H_0 : \delta = 0 \text{ versus } H_1 : \delta > 0$$

and, with a lower one-sided alternative,

$$H_0 : \delta = 0 \text{ versus } H_1 : \delta < 0.$$

The test statistic for evaluating the strength of evidence in favor of the alternative hypothesis is given by

$$Z = \frac{\widehat{\mu}_2 - \widehat{\mu}_1}{\sqrt{\widehat{\sigma}_1^2/n_1 + \widehat{\sigma}_2^2/n_2}}.$$

When the upper one-sided alternative is assumed, a larger value of the test statistic leads to a decision to reject the null hypothesis of no effect, i.e., H_0 is rejected if $Z \geq z_\alpha$. With the upper one-sided alternative, the null hypothesis is rejected if $Z \leq -z_\alpha$.

If the average standard deviation is defined as

$$\sigma = \sqrt{\sigma_1^2 + \frac{\sigma_2^2}{r}},$$

power is easily computed as a function of δ and σ, i.e.,

$$\psi(\delta, \sigma) = \Phi\left(\sqrt{\frac{n}{1+r}}\frac{\delta}{\sigma} - z_\alpha\right) \text{ (upper one-sided alternative),}$$

$$\psi(\delta, \sigma) = \Phi\left(-\sqrt{\frac{n}{1+r}}\frac{\delta}{\sigma} - z_\alpha\right) \text{ (lower one-sided alternative).}$$

If power is set to $1 - \beta$, the required total number of patients is equal to

$$n = (1+r)\frac{(z_\alpha + z_\beta)^2\sigma^2}{\delta^2}.$$

Non-inferiority assessment

In a non-inferiority setting, the trial's objective is to demonstrate that the treatment is not substantially worse than, i.e., not inferior to, the control. The degree of non-inferiority is determined using a pre-set constant, known as the non-inferiority margin. The margin is denoted by γ.

The null hypothesis of inferiority and alternative hypothesis of non-inferiority are defined as

$$H_0 : \delta = \gamma \text{ versus } H_1 : \delta = 0,$$

where γ is negative under the upper one-sided alternative and positive under the lower one-sided alternative. The corresponding non-inferiority test statistic is given by

$$Z = \frac{\widehat{\mu}_2 - \widehat{\mu}_1 - \gamma}{\sqrt{\widehat{\sigma}_1/n_1 + \widehat{\sigma}_2/n_2}}.$$

It is easy to verify that, under the null hypothesis of inferiority, the test statistic follows the standard normal distribution. As above, the null hypothesis is rejected if $Z \geq z_\alpha$ with the upper one-sided alternative and if $Z \leq -z_\alpha$ with the lower one-sided alternative.

Power as a function of the total sample size n is given by

$$\psi(\delta, \sigma|\gamma) = \Phi\left(\sqrt{\frac{n}{1+r}}\frac{\delta-\gamma}{\sigma} - z_\alpha\right) \text{ (upper one-sided alternative)},$$

$$\psi(\delta, \sigma|\gamma) = \Phi\left(-\sqrt{\frac{n}{1+r}}\frac{\delta-\gamma}{\sigma} - z_\alpha\right) \text{ (lower one-sided alternative)}.$$

The total sample size in the trial as a function of β is given by

$$n = (1+r)\frac{(z_\alpha + z_\beta)^2\sigma^2}{(\delta-\gamma)^2}.$$

Example

The following numeric example illustrates the process of computing the sample size in a clinical trial with a normally distributed endpoint. Consider an antihypertension Phase III trial with two arms (experimental treatment versus active control). The primary analysis in the trial is based on the change in systolic blood pressure (measured in mmHg) and aims to demonstrate that the treatment is non-inferior to the active control. A larger reduction in the mean systolic blood pressure is desirable and thus a lower one-sided alternative will be considered in the hypothesis testing problem. Under the alternative hypothesis, $\mu_1 = \mu_2 = -9$ (i.e., $\delta = 0$) and, under the null hypothesis of inferiority, $\mu_1 = -9$ and $\mu_2 = -6$ (i.e., $\delta = 3$). The trial's design is balanced ($r = 1$) and the common standard deviation is $\sigma_1 = \sigma_2 = 10$, therefore

$$\sigma^2 = \sigma_1^2 + \frac{\sigma_2^2}{r} = 200.$$

The non-inferiority margin is set to $\gamma = 3$ (note that the margin is positive since the lower one-sided alternative is considered in this hypothesis testing problem). Using a one-sided $\alpha = 0.025$ and 90% power ($\beta = 0.1$), the total sample size in the trial is

$$n = (1+r)\frac{(z_\alpha + z_\beta)^2\sigma^2}{(\delta-\gamma)^2} = 467.$$

The resulting sample size matches the example presented in the EAST user manual (Cytel, 2016, Chapter 12).

2.3 Frequentist calculations in trials with binary endpoints

Two-arm trials with a binary primary endpoint are considered in this section. The Z-test with an unpooled variance estimate is supported in this setting. Power and sample calculations for this popular test are presented in multiple papers and books, including Chow, Shao and Wang (2008) and Julious (2010). This approach is implemented in SAS, EAST and multiple R packages (TrialSize and gsDesign). For example, this test is implemented in the TwoSampleProportion.NIS function of the TrialSize package.

Hypothesis testing problems

Let π_1 and π_2 denote the true values of the proportions of interest, e.g., response rates, in the control arm and treatment arm, respectively. The true treatment difference is equal to $\delta = \pi_2 - \pi_1$.

As in Section 2.2, the following two scenarios will be considered:

- Upper one-sided alternative: A positive value of the treatment difference corresponds to a beneficial treatment effect.
- Lower one-sided alternative: A negative value of the treatment difference corresponds to treatment benefit.

In a superiority setting, the corresponding null hypotheses of no effect and alternative hypotheses of a beneficial effect are defined as in Section 2.2, i.e.,

$$H_0 : \delta = 0 \text{ versus } H_1 : \delta > 0 \text{ (upper one-sided alternative)},$$

$$H_0 : \delta = 0 \text{ versus } H_1 : \delta < 0 \text{ (lower one-sided alternative)}.$$

Similarly, if the trial is conducted to pursue a non-inferiority objective with a prespecified non-inferiority margin γ, the null hypothesis of inferiority and alternative hypothesis of non-inferiority are again set up as in Section 2.2, i.e.,

$$H_0 : \delta = \gamma \text{ versus } H_1 : \delta = 0.$$

The margin is negative if the upper one-sided alternative is considered and is positive otherwise.

Superiority and non-inferiority assessments

The test statistic for evaluating the significance of the treatment effect in a superiority setting is given by

$$Z = \frac{\widehat{\pi}_2 - \widehat{\pi}_1}{\sqrt{\widehat{\pi}_1(1-\widehat{\pi}_1)/n_1 + \widehat{\pi}_2(1-\widehat{\pi}_2)/n_2}}.$$

The null hypothesis of no effect is rejected if $Z \geq z_\alpha$ provided a higher value of the proportion indicates a beneficial treatment effect (upper one-sided alternative) and if $Z \leq -z_\alpha$ otherwise (lower one-sided alternative).

The standard deviation corresponding to this test is naturally defined as follows

$$\sigma = \sqrt{\pi_1(1-\pi_1) + \frac{\pi_2(1-\pi_2)}{r}}.$$

Using this definition of σ, the power function of the test is given by

$$\psi(\pi_1, \pi_2) = \Phi\left(\sqrt{\frac{n}{1+r}}\frac{\delta}{\sigma} - z_\alpha\right) \text{ (upper one-sided alternative)},$$

$$\psi(\pi_1, \pi_2) = \Phi\left(-\sqrt{\frac{n}{1+r}}\frac{\delta}{\sigma} - z_\alpha\right) \text{ (lower one-sided alternative)}.$$

The total sample size of the trial with a superiority objective is

$$n = (1+r)\frac{(z_\alpha + z_\beta)^2\sigma^2}{\delta^2}.$$

If the trial's goal is to show that the treatment is non-inferior to the control, the test statistic is easily modified as follows

$$Z = \frac{\widehat{\pi}_2 - \widehat{\pi}_1 - \gamma}{\sqrt{\widehat{\pi}_1(1-\widehat{\pi}_1)/n_1 + \widehat{\pi}_2(1-\widehat{\pi}_2)/n_2}}.$$

The power function is defined as follows

$$\psi(\pi_1, \pi_2|\gamma) = \Phi\left(\sqrt{\frac{n}{1+r}}\frac{\delta-\gamma}{\sigma} - z_\alpha\right) \text{ (upper one-sided alternative)},$$

$$\psi(\pi_1, \pi_2|\gamma) = \Phi\left(-\sqrt{\frac{n}{1+r}}\frac{\delta-\gamma}{\sigma} - z_\alpha\right) \text{ (lower one-sided alternative)}.$$

The total sample size in the trial is given by

$$n = (1+r)\frac{(z_\alpha + z_\beta)^2\sigma^2}{(\delta-\gamma)^2}.$$

Example

As an example, consider a two-arm Phase III trial in patients with HIV. The primary endpoint in the trial is binary (24-week disease-free rate) and a higher rate corresponds to a beneficial effect. Suppose that the trial is designed to perform a non-inferiority assessment for a novel treatment compared to an active control. The disease-free rate in the control arm is assumed to be 80% ($\pi_1 = 0.8$). Under the null hypothesis of inferiority, the disease-free rate in the treatment arm is 75%, i.e., $\pi_2 = 0.75$ or $\delta = -0.05$. An upper one-sided alternative is considered in the trial and states that the disease-free rate equals 80% in both trial arms ($\pi_2 = 0.8$ or $\delta = 0$). The non-inferiority margin is set to $\gamma = -0.05$. Assuming a balanced design ($r = 1$) with a one-sided $\alpha = 0.025$ and 90% power ($\beta = 0.1$), the required total number of patients in the trial is

$$n = (1+r)\frac{(z_\alpha + z_\beta)^2\sigma^2}{(\delta-\gamma)^2} = 2690,$$

where $\sigma = 0.5657$. This sample size calculation matches that presented in the EAST user manual (Cytel, 2016, Chapter 24).

2.4 Frequentist calculations in trials with time-to-event endpoints

A clinical trial with an event-driven design will be considered in this section and it will be assumed that the time to the primary event follows an exponential distribution. Let λ_1 and λ_2 denote the hazard rates in the control and treatment arms, respectively. The hazard ratio is denoted by δ, i.e., $\delta = \lambda_2/\lambda_1$.

As in Sections 2.2 and 2.3, two scenarios based on lower and upper one-sided alternatives are considered to define the hypothesis testing problem. Under the upper one-sided alternative, treatment benefit is associated with a lower hazard ratio or, equivalently, with a longer time to the primary event in the treatment arm compared to the control arm. With the lower one-sided alternative, a beneficial effect is associated with a higher hazard ratio. This immediately implies that, in a trial pursuing a superiority objective, the null and alternative hypotheses are defined as follows:

$$H_0 : \delta = 1 \text{ versus } H_1 : \delta < 1 \text{ (upper one-sided alternative)},$$
$$H_0 : \delta = 1 \text{ versus } H_1 : \delta > 1 \text{ (lower one-sided alternative)}.$$

Within a non-inferiority framework, let γ denote the prospectively defined non-inferiority margin on the hazard ratio scale. Using this margin, the null and alternative hypotheses are set up as follows:

$$H_0 : \delta = \gamma \text{ versus } H_1 : \delta = 1.$$

Here γ is greater than 1 under the upper one-sided alternative and is less than 1 otherwise.

An important feature of clinical trials with time-to-event endpoints is that power calculations can be carried out to pursue two different two goals:

- Compute the target number of events in the trial. This calculation does not take into account patient accrual or patient dropout (in a sense, every patient is followed up until this patient experiences the event of interest).
- Compute the number of enrolled patients in the trial. To perform this calculation, assumptions about the patient accrual and dropout processes need be made.

These two goals will be discussed below. The calculation of the number of events is based on the Schoenfeld formula (Schoenfeld, 1981) and, to address the second goal, the approach developed in Lachin and Foulkes (1986) is applied. The same methodology is utilized in EAST, SAS as well as R packages (gsDesign and TrialSize).

Calculation of the number of events

Beginning with the problem of computing the target number of events in a trial where the primary endpoint is a time-to-event endpoint, consider a superiority setting. The treatment effect will be evaluated using the standard log-rank test. Let d denote the total number of events in the two trial arms. Assuming that there are no ties, let n_{1k} be the number of patients in the control arm who are at risk just before the kth event and, similarly, let n_{2k} be the number of patients in the treatment arm who are at risk just before the kth event, $k = 1, \ldots, d$. Finally, $I_k = 1$ if the kth event occurs in the control arm and 0 otherwise. The test statistic is given by

$$Z = \sum_{k=1}^{d} \left(I_k - \frac{n_{1k}}{n_{1k} + n_{2k}} \right) \Big/ \sqrt{\sum_{k=1}^{d} \frac{n_{1k} n_{2k}}{(n_{1k} + n_{2k})^2}}.$$

Assuming the upper one-sided alternative, a large value of this test statistic is inconsistent with the null hypothesis of no treatment effect and the null hypothesis is rejected if $Z \geq z_\alpha$. If the lower one-sided alternative is considered, the null hypothesis of no effect is rejected if $Z \leq -z_\alpha$.

If there is no censoring (and thus every patient ultimately experiences the event of interest) and the true hazard ratio δ is close to 1, the power function of the log-rank test can be approximated as follows:

$$\psi(\delta) = \Phi\left(-\frac{\sqrt{dr}}{1+r} \log\delta - z_\alpha \right) \text{ (upper one-sided alternative)},$$

$$\psi(\delta) = \Phi\left(\frac{\sqrt{dr}}{1+r} \log\delta - z_\alpha \right) \text{ (lower one-sided alternative)},$$

where $\log\delta$ is the natural logarithm of the hazard ratio.

As a result, the total number of events in the trial is equal to

$$d = \frac{(1+r)^2}{r} \frac{(z_\alpha + z_\beta)^2}{(\log\delta)^2}.$$

Switching to a clinical trial designed to demonstrate that the treatment is non-inferior to the control, the log-rank test needs to be modified to support a non-inferiority assessment as follows:

$$Z = \sum_{k=1}^{d} \left(I_k - \frac{n_{1k}}{n_{1k} + \gamma n_{2k}} \right) \Big/ \sqrt{\sum_{k=1}^{d} \frac{\gamma n_{1k} n_{2k}}{(n_{1k} + \gamma n_{2k})^2}}.$$

This implies that the power function of the non-inferiority test is given by

$$\psi(\delta|\gamma) = \Phi\left(-\frac{\sqrt{dr}}{1+r}\log\frac{\delta}{\gamma} - z_\alpha\right) \text{ (upper one-sided alternative)},$$

$$\psi(\delta|\gamma) = \Phi\left(\frac{\sqrt{dr}}{1+r}\log\frac{\delta}{\gamma} - z_\alpha\right) \text{ (lower one-sided alternative)},$$

and thus the total number of events needs to be set to

$$d = \frac{(1+r)^2}{r}\frac{(z_\alpha + z_\beta)^2}{(\log(\delta/\gamma))^2}.$$

Calculation of the number of patients

The approach presented above focuses on finding the target number of primary events and the required number of patients is not explicitly defined. To compute the number of patients to be enrolled into the trial, assumptions on the patient accrual and patient dropout processes need to be made. Suppose that the length of the accrual period is T_R and the total duration of the trial, i.e., the length of time from the enrollment of the first patient to the discontinuation of the last patient, is T_S.

Patients can be enrolled into the trial in a uniform fashion or, alternatively, a more general distribution can be introduced to describe the patient accrual. It is common to assume that the accrual is governed by a truncated exponential distribution with the following cumulative distribution function:

$$F(x|\tau) = \frac{1-\exp(-\tau x)}{1-\exp(-\tau T_R)},\ 0 \le x \le T_R.$$

Here τ is the parameter that defines the distribution's shape. With a positive value of τ, patients are initially enrolled at a high rate but the patient accrual slows down towards the end of the trial. On the other hand, if $\tau < 0$, the accrual rate increases over time. Lastly, if $\tau = 0$, this distribution simplifies to a uniform distribution, i.e., $F(x) = x/T_R$.

Secondly, there are two sources of censoring in the trial:

- Administrative censoring, i.e., a patient reaches the end of the trial without experiencing the primary event.
- Censoring due to dropout, i.e., a patient is lost to follow up before experiencing the primary event.

Assuming an exponential dropout, let η denote the common hazard rate of the dropout distribution in the control and treatment arms.

To define the formula for computing the total number of enrolled patients, let

$$\sigma_0 = \frac{1+r}{\sqrt{r\varphi(\lambda)}},$$

$$\sigma_1 = \sqrt{\frac{1+r}{\varphi(\lambda_1)} + \frac{1+r}{r\varphi(\lambda_2)}},$$

where

$$\varphi(x) = \frac{x}{x+\eta} + \frac{x\tau\exp(-(x+\eta)T_S)(1-\exp(-(x+\eta-\tau)T_R))}{(x+\eta)(x+\eta-\tau)(1-\exp(-\tau T_R))}$$

and λ is the average hazard rate, i.e.,

$$\lambda = \frac{\lambda_1 + r\lambda_2}{1+r}.$$

Now, assuming a superiority setting, the power function is given by

$$\psi(\delta) = \Phi\left(-\frac{\sqrt{n}}{\sigma_1}\log\delta - z_\alpha\frac{\sigma_0}{\sigma_1}\right) \text{ (upper one-sided alternative)},$$

$$\psi(\delta) = \Phi\left(\frac{\sqrt{n}}{\sigma_1}\log\delta - z_\alpha\frac{\sigma_0}{\sigma_1}\right) \text{ (lower one-sided alternative)}$$

and therefore the total number of enrolled patients is equal to

$$n = \frac{(z_\alpha\sigma_0 + z_\beta\sigma_1)^2}{(\log\delta)^2}.$$

With a non-inferiority setting, the power function is defined as follows

$$\psi(\delta|\gamma) = \Phi\left(-\frac{\sqrt{n}}{\sigma_1}\log\frac{\delta}{\gamma} - z_\alpha\frac{\sigma_0}{\sigma_1}\right) \text{ (upper one-sided alternative)},$$

$$\psi(\delta|\gamma) = \Phi\left(\frac{\sqrt{n}}{\sigma_1}\log\frac{\delta}{\gamma} - z_\alpha\frac{\sigma_0}{\sigma_1}\right) \text{ (lower one-sided alternative)}$$

and the total number of enrolled patients is equal to

$$n = \frac{(z_\alpha\sigma_0 + z_\beta\sigma_1)^2}{(\log(\delta/\gamma))^2}.$$

Example

To illustrate the methods for computing the target number of events and sample size in trials with time-to-event endpoints, consider a Phase III trial in patients with metastatic colorectal cancer. A two-arm design (experimental treatment plus best supportive care versus best supportive care) is employed in the trial and patients will be randomized in a 2:1 ratio to the treatment or control ($r = 2$). The primary objective of this trial is to demonstrate that the experimental treatment is superior to the control in terms of overall survival. It is assumed that median survival times in the control and treatment arms are 6 and 9 months, respectively. The hazard rates corresponding to these median survival times are

$$\lambda_1 = \frac{\log 2}{6} = 0.116,\ \lambda_2 = \frac{\log 2}{9} = 0.077,$$

and the hazard ratio is $\delta = \lambda_2/\lambda_1 = 0.667$. Assuming 90% power and a one-sided $\alpha = 0.025$, the target number of events in the trial is

$$d = \frac{(1+r)^2}{r}\frac{(z_\alpha + z_\beta)^2}{(\log\delta)^2} = 288.$$

Furthermore, the following assumptions will be made to find the required number of patients in the trial:

- The length of the patient accrual period is $T_R = 12$ months and the total length of the trial is $T_S = 24$ months. The patient accrual is governed by a truncated exponential distribution with the median accrual time of 9 months, which means that 50% of the patients are expected to be enrolled by the 9-month milestone. The corresponding parameter of the truncated exponential distribution is $\tau = -0.203$.

- The annual dropout rate is 5%, which means that the hazard rate of the exponential dropout distribution is $\nu = -\log 0.95/12 = 0.0043$.

The resulting number of enrolled patients is

$$n = \frac{(z_\alpha \sigma_0 + z_\beta \sigma_1)^2}{(\log \delta)^2} = 388.$$

2.5 Bayesian calculations

This section provides a short summary of simulation-based Bayesian calculations aimed at evaluating the probability of success, also known as assurance, in two-arm clinical trials with continuous, binary and time-to-event endpoints. In general, assurance calculations rely on averaging frequentist characteristics such as power with respect to prior distributions of the endpoint parameters, e.g., prior distributions of the true response rates in the control and treatment arms in a trial with a binary endpoint. The prior distributions are derived from historical data and a simple approach to deriving posterior distributions and carrying out assurance calculations in trials with continuous, binary and time-to-event endpoints is presented below. For more information on the use of assurance in clinical trials and calculation of posterior distributions, see O'Hagan, Stevens and Campbell (2005), Wang (2015) and Gelman et al. (2013).

It will be assumed throughout this section that posterior distributions of interest are found using information from a two-arm historical trial with the same experimental treatment and control as in the current trial. The indices corresponding to the control and treatment arms are $i = 1$ and $i = 2$, e.g., the number of patients in the control and treatment arms are denoted by k_1 and k_2, respectively. The historical trial may be either hypothetical, in which case this trial serves purely as a device for assessing the robustness of power calculations, or based on a real clinical trial. In the latter case it is important to remember that the assumed primary endpoint parameters, e.g., mean and standard deviation, do not need to be equal to the actual parameters observed in the real trial. The actual parameter values may be replaced by assumed values that will be utilized for power calculation in the current trial.

To compute assurance, let θ be the vector of endpoint parameters in the control and treatment arms, e.g., $\theta = (\pi_1, \pi_2)$ in a clinical trial with a binary primary endpoint, where π_1 and π_2 are the true values of the proportions in the control and treatment arms, respectively. If the probability of a statistically significant treatment effect in the current trial is denoted by $P(\theta)$, assurance is defined as

$$\int P(\theta) f(\theta) d\theta,$$

where $f(\theta)$ is the probability density function of the prior distribution of θ. This prior distribution is equal to the posterior distribution of θ given the data from the historical trial. This posterior distribution is often derived from a non-informative prior for θ. Within a simulation-based framework, the integral is approximated by

$$\frac{1}{s} \sum_{i=1}^{s} P(\theta_i),$$

where $\theta_1, \ldots, \theta_s$ are sampled from the prior distribution of θ and s is the number of simulation runs.

Bayesian calculations in trials with continuous endpoints

Assurance calculations are planned to be run a clinical trial with a continuous primary endpoint in addition to traditional frequentist calculations. The endpoint follows a normal distribution with the parameters (μ_1, σ_1^2) in the control arm and (μ_2, σ_2^2) in the treatment arm.

The joint distribution of the endpoint parameters in each trial arm is defined using the following two-step algorithm:

- The variance σ_i^2 follows a scaled inverse chi-square distribution with the degrees of freedom ν_i and scale parameter τ_i^2, $i = 1, 2$.
- The mean μ_i, conditional on the variance σ_i^2, follows a normal distribution with the mean μ_i and variance σ_i^2/κ_i, where κ_i is a pre-set scaling parameter, $i = 1, 2$.

A non-informative prior distribution can be defined by setting κ_i to 0, ν_i to -1 and τ_i^2 to 0, $i = 1, 2$.

Consider the ith trial arm, $i = 1, 2$, and let m_i and s_i denote the assumed mean and standard deviation that will be utilized in the power calculation in the current trial. These endpoint parameters are treated as if they were observed in the historical trial, i.e., these values are used to find the posterior distribution of the true means and standard deviations (as explained above, the assumed values may not be equal to the actual means and standard deviations in the historical trial). The posterior distribution of the endpoint parameters in the ith trial arm of the historical trial is then derived using the two-step approach described above. In particular, it can be shown that the posterior variance σ_i^2 follows a scaled inverse chi-squared distribution with the degrees of freedom parameter denoted by $k_i - 1$ and scale parameter s_i^2. Furthermore, the posterior mean μ_i, conditional on the posterior variance σ_i^2, is normally distributed with the mean m_i and variance σ_i^2/k_i.

The posterior distributions of the endpoint parameters derived from the historical trial will be used as the prior distributions when evaluating assurance in the current trial. An important feature of this approach to performing Bayesian calculations is that it depends only on the number of patients in each arm of the historical trial. If the sample size in the historical trial is large, the marginal posterior distributions will be tightly clustered around the assumed values of the endpoint parameters, i.e., around m_1 and s_1 in the control arm and around m_2 and s_2 in the treatment arm, and assurance will be reasonably close to frequentist power.

Bayesian calculations in trials with binary endpoints

Consider a two-arm historical trial with a binary endpoint and let π_1 and π_2 denote the true values of the proportions in the control and treatment arms, respectively. Conjugate distributions will be assumed for these proportions, i.e., π_i will be assumed to follow a beta distribution with the shape parameters given by α_i and β_i, $i = 1, 2$.

The assumed values of the proportions to be used in the power calculations in the current trial are denoted by p_1 and p_2. Treating these values as if they were observed in the historical trial and assuming non-informative priors for the true proportions, i.e.,

$$\alpha_i = 1,\ \beta_i = 1,\ i = 1, 2,$$

it is easy to derive the parameters of the posterior distributions of π_1 and π_2 in the historical trial. The posterior distribution of the true proportion π_i is also a beta

distribution with the shape parameters

$$\alpha_i = 1 + p_i k_i,\ \beta_i = 1 + (1 - p_i)k_i,\ i = 1, 2.$$

The resulting posterior distributions of the true proportions will be used as the prior distributions for computing assurance in the current trial.

Bayesian calculations in trials with time-to-event endpoints

Considering a two-arm historical trial with the same time-to-event endpoint as in the current trial, assume that the time to the event of interest is exponentially distributed and let λ_1 and λ_2 denote the true hazard rates in the control and treatment arms, respectively. Using a conjugate distribution approach, it will be assumed that λ_i follows a gamma distribution with the shape parameter α_i and rate parameter β_i, $i = 1, 2$.

An improper non-informative prior with $\alpha_i = 0$ and $\beta_i = 0$ will be assumed for the true hazard rate λ_i, $i = 1, 2$, in the historical trial. Let l_1 and l_2 denote the assumed hazard rates in the control arm and treatment arm of the current trial, respectively. If the observed hazard rates in the historical trial are set to these assumed values, the posterior distribution of λ_i in the historical trial is a gamma distribution with the following shape and rate parameters:

$$\alpha_i = k_i,\ \beta_i = k_i / l_i,\ i = 1, 2.$$

As above, the gamma distributions with these parameters will be used as the prior distributions for the hazard rates when performing assurance calculations in the current trial.

Example

Consider a development program for the treatment of rheumatoid arthritis. The primary endpoint in the Phase II and III trials included in this program is binary (ACR20 definition of improvement) and a higher response rate indicates a beneficial treatment effect. Suppose that the sample size in a Phase III trial will be computed assuming that the true control response rate is $\pi_1 = 0.3$ and the true treatment response rate is $\pi_2 = 0.5$. These response rates are based on the results observed in a recently conducted Phase II trial, which will serve as the historical trial. Using a one-sided $\alpha = 0.025$ and the Z-test for proportions, it is easy to check that the total number of patients needs to be set to 242 to ensure 90% power.

To support assurance calculations for this Phase III trial, it will be assumed that the true response rates in the two trial arms follow beta distributions. Non-informative priors will be considered in the historical trial to compute the posterior distributions for the response rates that will be ultimately utilized in the assurance calculation. The non-informative beta priors are defined using the following set of shape parameters:

$$\alpha_1 = 1,\ \beta_1 = 1 \text{ (control arm)},$$
$$\alpha_2 = 1,\ \beta_2 = 1 \text{ (treatment arm)}.$$

Assuming a balanced design in the Phase II trial, suppose that the sample size per arm is $m_1 = m_2 = 50$ patients. To compute the posterior distributions of the true response rates, the observed response rates in the Phase II trial are assumed to be

equal to $p_1 = 0.3$ and $p_2 = 0.5$. The posterior distributions are beta distributions with the following parameters:

$$\alpha_1 = 16,\ \beta_1 = 36 \text{ (control arm)},$$
$$\alpha_2 = 26,\ \beta_2 = 26 \text{ (treatment arm)}.$$

These posterior distributions are plotted in Figure 1. It can be seen from this figure that the posterior distributions of the true response rates are centered around the assumed values ($p_1 = 0.3$ and $p_2 = 0.5$) and demonstrate a fairly high amount of variability due to the fact that the estimated response rates are obtained using a relatively small historical trial.

The posterior distributions now serve as the prior distributions for π_1 and π_2 in the Phase III trial. A simulation-based algorithm can now be applied to generate samples from the prior distributions, compute response rates in the control and treatment arms and ultimately evaluate the significance of the treatment effect in each simulation run. By averaging over a large number of simulation runs, e.g., over 10,000 runs, the assurance is estimated to be 73.7%. This value is lower than 90%, which is the target for frequentist power, since the assurance calculation accounts for the uncertainty around the assumed response rates, i.e., $p_1 = 0.3$ and $p_2 = 0.5$. If the response rates came from a larger Phase II trial, assurance would be closer to frequentist power. For example, if the sample size per arm in the Phase II trial is $m_1 = m_2 = 100$ patients, assurance increases to 79.2%.

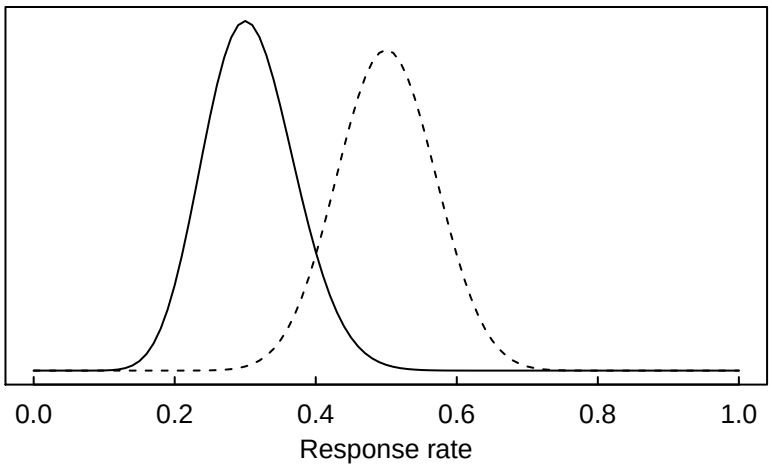

Figure 1 Posterior distributions of the true response rates in the Phase II trial (solid curve, control arm; dashed curve, treatment arm).

3 TraditionalSimulations module: Simulation-based calculations in fixed-sample trials

3.1 Introduction

This section introduces the simulation-based Clinical Scenario Evaluation approach in clinical trials with a fixed-sample design. For more information on the Clinical Scenario Evaluation framework, see Benda et al. (2010), Friede et al. (2010) and Dmitrienko and Pulkstenis (2017).

4 GroupSequential module: Analytical and simulation-based calculations in group-sequential trials

4.1 Introduction

This section discusses analytical and simulation-based evaluation evaluations of operating characteristics in a group-sequential trial with a continuous, binary or time-to-event primary endpoint. An analytical approach is employed to compute standard operating characteristics of a group-sequential trial without patient accrual or dropout modeling. The analytical evaluation of the trial's operating characteristics is complemented by a simulation-based approach that enables the trial's sponsor to model the patient accrual or dropout processes and perform sensitivity assessments.

References

[1] Anderson, K. (2016). gsDesign: Group sequential design. R package. https://CRAN.R-project.org/package=gsDesign.

[2] Benda, N., Branson, M., Maurer, W., Friede, T. (2010). Aspects of modernizing drug development using clinical scenario planning and evaluation. *Drug Information Journal*. 44, 299-315.

[3] Chow, S.C., Shao, J., Wang, H. (2008). *Sample Size Calculations in Clinical Research*. Second Edition. Chapman and Hall/CRC Press, New York.

[4] Cytel (2016). *EAST 6.4 User Manual*.

[5] Diegert, C., Diegert, K.V. (1981). Note on inversion of Casagrande-Pike-Smith approximate sample-size formula for Fisher-Irwin test on 2×2 tables. *Biometrics*. 37, 595.

[6] Dmitrienko, A., Pulkstenis, E. (editors). (2017). *Clinical Trial Optimization Using R*. Chapman and Hall/CRC Press, New York.

[7] Fleiss, J.L., Tytun, A., Ury, H.K. (1980). A simple approximation for calculating sample sizes for comparing independent proportions. *Biometrics*. 36, 343-346.

[8] Friede, T., Nicholas, R., Stallard, N., Todd, S., Parsons, N. R., Valdes-Marquez, E., Chataway, J. (2010). Refinement of the clinical scenario evaluation framework for assessment of competing development strategies with an application to multiple sclerosis. *Drug Information Journal*. 44, 713-718.

[9] Gelman, A., Carlin, J., Stern, S., Dunson, D., Vehtari, A., Rubin, D. (2013). *Bayesian Data Analysis* (Third Edition). Chapman and Hall/CRC Press, New York.

[10] Julious, S.A. (2010). *Sample Sizes for Clinical Trials*. Chapman and Hall, New York.

[11] Lachin, J.M., Foulkes, M. (1986). Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. *Biometrics*. 42, 507-519.

[12] O'Hagan, A., Stevens, J.W., Campbell, M.J. (2005). Assurance in clinical trial design. *Pharmaceutical Statistics*. 4, 187-201.

[13] SAS (2017). *SAS/STAT 14.3 User's Guide*. The POWER Procedure.

[14] Schoenfeld, D.A. (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika*. 68, 316-319.

[15] Wang, M.D. (2015).Applications of probability of study success in clinical drug development. Selected Papers from 2013 ICSA/ISBS Joint Statistical Meetings. 4, 185-196.

[16] Zhang, E., Wu, V.Q., Chow, S.C., Zhang, H.G. (2013). TrialSize: R functions for sample size calculations. https://CRAN.R-project.org/package=TrialSize.