# Glossary

**$2 \times 2$ contingency table**. A $2 \times 2$ contingency table is used to describe the association between a binary independent variable and a binary response variable. See [ADAPT] **gsdesign twoproportions**.

**acceptance region**. In classical hypothesis testing, an acceptance region is the complement of the rejection region and is defined as a set of values of a test statistic for which the null hypothesis cannot be rejected. Group sequential designs further differentiate between the acceptance region, where the null hypothesis is accepted and the trial is terminated early for futility, and the continuation region, where the trial is continued due to insufficient evidence to accept or reject the null hypothesis. Also see *rejection region* and *continuation region*.

**accrual period** or **recruitment period** or **accrual**. The accrual period (or recruitment period) is the period during which participants are being enrolled (recruited) into a study. Also see *follow-up period*.

**active control** or **active comparator**. In a clinical trial of an experimental treatment for a condition where there is an existing standard of care, there is often an ethical argument against giving study participants a placebo, so the control group is given the standard of care and the experimental treatment is compared with the active control. Also see *placebo control*.

**adaptive design**. As defined by the U.S. Food and Drug Administration (2019), an adaptive design is a "clinical trial design that allows for prospectively planned modifications to one or more aspects of the design based on accumulating data from subjects in the trial."

**administrative censoring**. Administrative censoring is the right-censoring that occurs when the study observation period ends. All participants complete the course of the study and are known to have experienced one of two outcomes at the end of the study: survival or failure. This type of censoring should not be confused with withdrawal or loss to follow-up. Also see *censored, uncensored, left-censored, and right-censored*.

**adverse event**. Adverse events are harmful side effects of a treatment and negative medical outcomes not associated with an underlying disease. Clinical trials must closely track the incidence and severity of adverse events to ensure that a treatment is safe as well as effective.

**allocation ratio**. The allocation ratio, $n_2/n_1$, is the number of study participants in the experimental (treatment) group divided by the number of participants in the control (reference) group.

**alpha**. Alpha, $\alpha$, denotes the significance level. Also see *familywise significance level*.

**alternative hypothesis**. In hypothesis testing, the alternative hypothesis represents the counterpoint to which the null hypothesis is compared. When the parameter being tested is a scalar, the alternative hypothesis can be either one-sided or two-sided. Also see *null hypothesis*.

**arm**. In the context of a clinical trial, groups of study participants given the same treatment are often called arms. In a classic two-arm randomized controlled trial, the experimental arm is given the experimental treatment and the control arm is given the control treatment. Also see *single-arm trial* and *two-arm trial*.

**attained power**. When calculating the required sample size for a specified significance level and power, the resulting sample size is often fractional and must be rounded up to a whole number. This causes the attained power to be slightly greater than the requested power. Also see *power*.

**attained sample-size ratio**. When specifying a sample-size ratio that results in noninteger sample sizes, `gsdesign` will round up the computed sample sizes to the nearest integers. The attained sample-size ratio is computed using the rounded sample sizes. Also see *sample-size ratio*.

**balanced design**. A balanced design represents an experiment in which the numbers of treated and untreated study participants are equal. For many types of two-sample hypothesis tests, the power of the test is maximized with balanced designs. Balanced designs may also be called equal-allocation designs.

**Bernoulli trial**. A Bernoulli trial is an experiment with only two possible outcomes, "success" or "failure", recorded as 0 and 1, respectively. In a clinical trial with a binary outcome, each participant's response is viewed as an independent Bernoulli trial with a fixed probability of success. See [ADAPT] **gsdesign oneproportion** and [ADAPT] **gsdesign twoproportions**.

**beta**. Beta, $\beta$, denotes the probability of committing a type II error, namely, failing to reject the null hypothesis even though it is false. Also see *type II error*.

**binding futility boundaries** or **binding futility bounds**. In a group sequential clinical trial with binding futility bounds, if the test statistic at an interim analysis crosses the futility boundary, the trial must be stopped for futility; otherwise, it risks overrunning the specified significance level. Group sequential designs with binding futility bounds require smaller efficacy critical values than equivalent group sequential designs with nonbinding futility boundaries. Also see *nonbinding futility boundaries*.

**binary outcome**. When the response of each participant in a clinical trial is either "success" or "failure", we say the trial has a binary outcome. Analysis of binary clinical trial data treats each response as a Bernoulli trial with a fixed probability of success. A test of proportions, such as a binomial test or Pearson's $\chi^2$ test, is conducted to determine if the data are compatible with the null hypothesis. See [ADAPT] **gsdesign oneproportion** and [ADAPT] **gsdesign twoproportions**. Also see *composite endpoint*.

**binomial test**. A binomial test is a test for which the exact sampling distribution of the test statistic is binomial. See [R] **bitest**. Also see [ADAPT] **gsdesign oneproportion**.

**biomarker**. A biomarker is a characteristic of the body that can be measured objectively and that serves as an indicator of healthy biological processes, disease status, or response to a therapeutic treatment.

Response biomarkers are frequently used as surrogate endpoints for clinical trials where the clinical outcome of interest is too difficult, time consuming, or expensive to measure. For example, instead of relying on autopsy to diagnose Alzheimer's disease, we can use medical imaging to measure brain glucose metabolism as a response biomarker.

Biomarkers can also serve as risk factors used to define a population of interest. For example, the APOE $\epsilon 4$ gene is a known risk factor for Alzheimer's disease, and it can be used as a biomarker to define the target population of a clinical trial.

**bisection method**. This method finds a root $x$ of a function $f(x)$ such that $f(x) = 0$ by repeatedly subdividing an interval on which $f(x)$ is defined until the change in successive root estimates is within the requested tolerance and function $f(\cdot)$ evaluated at the current estimate is sufficiently close to 0.

**blinding**. Blinding refers to clinical trials where the identity of the treatment is hidden. In an open-label trial, participants are told which treatment they are receiving. In a single-blinded trial, participants do not know which treatment they receive, but the researchers administering the treatments and the data analysts are unblinded, meaning they know which treatment each participant receives. If the study design of a blinded trial calls for the experimental treatment to be compared with no intervention, then the control group is given a placebo so that they do not know they are members of the control group. In a double-blinded trial, both the participants and the researchers administering the treatments are blinded to the identity of the treatments, and in a triple-blinded trial, even the data analysts are blinded.

**boundary** or **bound**. See *stopping boundary*.

**boundary-calculation procedure** or **boundary-calculation method**. In the context of a group sequential design, the boundary-calculation procedure refers to the method used to create a stopping boundary. Boundary-calculation procedures fall into two broad categories: classical stopping bounds and error-spending bounds. Classical stopping bounds calculate boundary critical values directly, while error-spending bounds define an error-spending function that partitions type I or type II error between the planned looks. Also see *classical Wang–Tsiatis bounds*, *classical Pocock bounds*, *classical O'Brien–Fleming bounds*, *error-spending Pocock bounds*, *error-spending O'Brien–Fleming bounds*, *error-spending Kim–DeMets bounds*, and *error-spending Hwang–Shih–de Cani bounds*.

**censored**, **uncensored**, **left-censored**, and **right-censored**. An observation is censored when the exact time of failure is not known, and it is uncensored when the exact time of failure is known.

An observation is left-censored when the exact time of failure is not known; it is merely known that the failure occurred before $t_l$. Suppose that the event of interest is becoming employed. If a subject is already employed when first interviewed, his outcome is left-censored.

An observation is right-censored when the time of failure is not known; it is merely known that the failure occurred after $t_r$. If a patient survives until the end of a study, the patient's time of death is right-censored.

Also see *administrative censoring*.

**clinical trial**. A clinical trial is an experiment testing the effect of a treatment or procedure on human participants.

**clinically meaningful difference** or **clinically meaningful effect** or **clinically significant difference**. Clinically meaningful difference represents the magnitude of an effect of interest that is of clinical importance. What is meant by "clinically meaningful" may vary from study to study.

**clinical outcome**. The clinical outcome is an outcome that confers direct clinical benefit, such as overall survival. In practice, clinical outcomes are often expensive or time consuming to measure, so surrogate endpoints are frequently measured instead. Also see *endpoint* and *target parameter*.

**composite endpoint**. Sometimes, when designing a clinical trial, there are multiple endpoints of interest. One solution is to combine multiple endpoints into a single composite endpoint. For example, a clinical trial of a treatment for COVID-19 might use a composite endpoint such as "death or intubation", where each participant's response is an indicator of whether they died or were intubated. Also see *binary outcome*.

**continuation region**. In group sequential designs, a continuation region is defined as a set of values of a test statistic that provide insufficient evidence to accept or reject the null hypothesis. If the test statistic from an interim analysis of clinical trial data lies within the continuation region, the trial will continue as planned (as opposed to stopping early if the test statistic lies within the acceptance region or the rejection region). There is no continuation region at the final analysis, because at this stage, the null hypothesis must be either accepted or rejected. Also see *acceptance region* and *rejection region*.

**control arm**. See *control group*.

**control group**. A control group (or arm) comprises study participants who are randomly assigned to a group where they receive the control treatment, which is either no treatment or a standard treatment. In hypothesis testing, this is usually the reference group. Also see *experimental group*.

**control treatment**. In a clinical trial, the control treatment is the reference treatment against which an experimental treatment is judged. If there are no existing treatments that are comparable with the experimental treatment, then the control group will typically receive a placebo. When a standard of care exists, there is often an ethical argument against using a placebo; in this case, an active control is used, in which control-group participants receive the existing standard of care. Also see *experimental treatment*.

**critical value**. In classical hypothesis testing, a critical value is a boundary of the rejection region. In the context of a group sequential design, there are two types of critical values: efficacy critical values, which are boundaries of the rejection region, and futility critical values, which are boundaries of the acceptance region. Also see *efficacy critical values* and *futility critical values*.

**Data Monitoring Committee (DMC)** or **Data and Safety Monitoring Committee (DSMC)** or **Data and Safety Monitoring Board (DSMB)**. In the context of a clinical trial, a DMC is a panel of experts that is tasked with periodically reviewing data collected by the trial. The DMC will analyze data on safety concerns, such as adverse events suffered by study participants, and the DMC will advise the sponsor of the trial if the study is believed to pose unnecessary risk to participants. In adaptive clinical trials that allow stopping for efficacy or futility, the DMC will perform interim analyses of incomplete trial data to evaluate the effectiveness of the experimental treatment. Not all clinical trials require the use of a DMC.

**delta**. Delta, $\delta$, in the context of power and sample-size calculations, denotes the effect size. In the context of a Wang–Tsiatis efficacy or futility boundary, capital Greek letter Delta, $\Delta$, represents the parameter of the boundary calculation. See *Classical (Wang–Tsiatis) bounds* in *Methods and formulas* of [ADAPT] **gsbounds** for the formula. Also see *effect size*.

**directional test**. See *one-sided test*.

**dropout**. Dropout is the withdrawal of participants before the end of a study and leads to incomplete or missing data. Also see *withdrawal*.

**effect size**. The effect size is the size of the clinically meaningful difference between the treatments being compared, typically expressed as a quantity that is independent of the unit of measure. For example, in a one-sample mean test, the effect size is a standardized difference between the mean and its reference value. In other cases, the effect size may be measured as an odds ratio or a risk ratio. Also see *delta*.

**efficacy** or **clinical efficacy**. Efficacy, the capacity to produce a desired result, is an important concept in clinical trials. In the context of a clinical trial, efficacy is quantified by measuring one or more endpoints. The efficacy of an experimental treatment is most commonly established by demonstrating that the experimental treatment compares favorably against a control treatment, but in the case of single-arm clinical trials, the endpoint from the group receiving the experimental treatment is compared against a prespecified reference value. In a clinical trial designed to demonstrate efficacy, the null hypothesis is that the experimental treatment lacks efficacy, and efficacy is established by rejecting $H_0$. Also see *efficacy boundaries*, *efficacy stopping*, and *futility*.

**efficacy boundaries** or **efficacy bounds**. In the context of group sequential designs for clinical trials, efficacy bounds are boundaries of the rejection region. If a test statistic is equal to or more extreme than the efficacy critical value, the test statistic is within the rejection region and the null hypothesis is rejected, allowing the trial to be terminated for treatment efficacy. Also see *futility boundaries*, *efficacy*, and *efficacy critical values*.

**efficacy critical values**. Efficacy critical values define efficacy boundaries in a group sequential design. At each look, a hypothesis test is conducted. If the test statistic is a $z$ statistic, it is compared directly with the efficacy critical value; if not, the significance level approach is used to compare the significance level of the test statistic with the significance level of the efficacy critical value. Also see *efficacy boundaries* and *futility critical values*.

**efficacy stopping**. In the context of group sequential designs for clinical trials, efficacy stopping refers to the early termination of a clinical trial due to treatment efficacy. This occurs when the test statistic calculated at an interim analysis lies within the rejection region, so the null hypothesis is rejected. Also see *efficacy* and *futility stopping*.

**endpoint**. The endpoint of a clinical trial is the target parameter that is used for hypothesis testing. Often, the clinical outcome of interest is difficult, time consuming, or expensive to measure, so a surrogate endpoint is measured instead. If there are multiple endpoints of interest, it is common to combine them into a single composite endpoint or to designate a primary endpoint that is used for sample-size determination. Also see *surrogate endpoint* and *composite endpoint*.

**equal-allocation design**. See *balanced design*.

**error-spending approach** or **error-spending function**. Instead of calculating boundary critical values directly, the error-spending approach to group sequential designs defines an error-spending function that partitions the alpha (for efficacy bounds) or beta (for futility bounds) into per-look probabilities of committing a type I or type II error. The critical value at each look is calculated based on the error spent, and the critical value at a look does not depend on critical values of future looks.

**error-spending O'Brien–Fleming-style bound**. In a group sequential clinical trial, one technique for calculating efficacy or futility boundaries is to use an error-spending O'Brien–Fleming-style bound. Boundary critical values from an error-spending O'Brien–Fleming-style bound are very similar to those of classical O'Brien–Fleming bounds, but they are obtained using an error-spending function. Also see *O'Brien–Fleming bounds* and *error-spending approach*.

**error-spending Pocock-style bound**. In a group sequential clinical trial, one technique for calculating efficacy or futility boundaries is to use an error-spending Pocock-style bound. Boundary critical values from an error-spending Pocock-style bound are very similar to those of classical Pocock bounds, but they are obtained using an error-spending function. Also see *Pocock bounds* and *error-spending approach*.

**ESS**. See *expected sample size*.

**exact test**. An exact test is one for which the probability of observing the data under the null hypothesis is calculated directly, often by enumeration. Exact tests do not rely on any asymptotic approximations and are therefore widely used with small datasets. See [ADAPT] **gsdesign oneproportion** and [ADAPT] **gsdesign twoproportions**.

**expected sample size (ESS)** or **average sample number**. In the context of a group sequential design, the ESS is the average sample size that would be required if the trial were to be repeated many times with the same design and with a given effect size. The ESSs under the null and alternative hypotheses are denoted as $\text{ESS}_0$ and $\text{ESS}_1$, respectively. Also see *maximum sample size*.

**experimental arm**. See *experimental group*.

**experimental group**. An experimental group (or arm) is a group of participants that receives a treatment or procedure of interest defined in a controlled experiment. In hypothesis testing, this is usually a comparison group. Also see *control group*.

**experimental study**. In an experimental study, as opposed to an observational study, the assignment of participants to treatments is controlled by investigators. For example, a study that compares a new treatment with a standard treatment by assigning each treatment to a group of participants is an experimental study. Also see *observational study*.

**experimental treatment**. In a clinical trial, an experimental treatment is a new treatment, such as a drug, medical device, or medical procedure, that is being tested. Typically, the experimental treatment is compared with a control treatment, but in the case of single-arm clinical trials, the endpoint from the group receiving the experimental treatment is compared with a prespecified reference value. Also see *control treatment*.

**failure function**. When analyzing time-to-event data, the failure function is the probability of experiencing a failure event at or before time $t$. If we denote the time of failure as $T$, we can define the failure function as the cumulative distribution function of $T$, where $F(t) = \Pr(T \leq t)$. The probability density function of $T$ is the derivative of the failure function with respect to time, written as $f(t) = \partial F(t)/\partial t$. Also see *hazard function* and *survivor function*.

**familywise error rate** or **familywise type I error**. When multiple hypothesis tests are conducted, the familywise error rate is the probability of committing a type I error during at least one test. Also see *type I error* and *familywise significance level*.

**familywise significance level**. When multiple hypothesis tests are conducted, the familywise significance level is an upper bound to the familywise error rate. Also see *significance level* and *familywise error rate*.

**finite population correction**. When sampling is performed without replacement from a finite population, a finite population correction is applied to the standard error of the estimator to reduce sampling variance.

**Fisher–Irwin exact test**. See *Fisher's exact test*.

**Fisher's exact test**. Fisher's exact test is an exact small-sample test of independence between rows and columns in a $2 \times 2$ contingency table. Conditional on the marginal totals, the test statistic has a hypergeometric distribution under the null hypothesis. See [ADAPT] **gsdesign twoproportions** and [R] **tabulate twoway**.

**Fisher information** or **information**. When estimating parameters from data, the Fisher information for those parameters is a matrix that quantifies the precision with which the parameters can be estimated from the data. Technically, the Fisher information is the expected value of the negative Hessian matrix of the log likelihood. In the context of clinical trials, it is common to conduct a hypothesis test of a single parameter, in which case the Fisher information is a scalar. In this case, a larger value of the Fisher information indicates that more is known about the parameter (typically due to a larger sample size). Also see *information ratio*.

**fixed-sample design (FSD)** or **fixed study design** or **fixed-sample study design** or **fixed design**. An FSD is an experimental design where the sample size is fixed. See [ADAPT] **GSD intro** for a comparison of FSDs versus group sequential designs.

**follow-up period** or **follow-up**. The (minimum) follow-up period is the period after the last participant entered the study until the end of the study. During the follow-up period, existing participants are under observation and no new participants enter the study. If $T$ is the total duration of a study and $r$ is the accrual period of the study, then follow-up period $f$ is equal to $T - r$. Also see *accrual period*.

**fractional sample size**. Sample-size calculations that compute sample size as a continuous quantity will often produce noninteger sample sizes. In practice, a fractional sample size must be rounded up to a whole number of participants. This rounding can cause the attained power to exceed the requested power. Also see *sample size*.

**futility**. Futility, defined as a lack of the ability to produce a desired result, has particular importance in the context of a clinical trial designed to demonstrate treatment efficacy. In this case, futility refers to the inability of the clinical trial to reject the null hypothesis and demonstrate efficacy. Clinical trials allowing for futility stopping may be terminated early for futility if the result of an interim analysis supports accepting the null hypothesis. Also see *futility boundaries*, *futility stopping*, and *efficacy*.

**futility boundaries** or **futility bounds**. In the context of group sequential designs for clinical trials, futility bounds are boundaries of the acceptance region. If a test statistic is less extreme than the futility critical value, the test statistic is within the acceptance region and the null hypothesis can be accepted, allowing the trial to be terminated for treatment futility.

There are two types of futility boundaries, binding futility boundaries and nonbinding futility boundaries. If the test statistic at an interim analysis crosses a binding futility boundary, the trial must be stopped for futility; otherwise, it risks overrunning the specified significance level. If a nonbinding futility boundary is used, the familywise type I error is controlled even if the trial continues after crossing the futility boundary. Also see *efficacy boundaries*, *futility*, and *futility critical values*.

**futility critical values**. Futility critical values define futility boundaries in a group sequential design. At each look, a hypothesis test is conducted. If the test statistic is a $z$ statistic, it is compared directly with the futility critical value; if not, the significance level approach is used to compare the significance level of the test statistic to the significance level of the futility critical value. Also see *futility boundaries* and *efficacy critical values*.

**futility stopping**. In the context of group sequential designs for clinical trials, futility stopping refers to the early termination of a clinical trial due to treatment futility, often described as "abandoning a lost cause". This occurs when the test statistic calculated at an interim analysis lies within the acceptance region and the null hypothesis is accepted. Also see *futility* and *efficacy stopping*.

**group sequential clinical trial** or **group sequential trial**. A group sequential clinical trial is a clinical trial that uses a group sequential design. Also see *group sequential design (GSD)*.

**group sequential design (GSD)**. A GSD is an experimental design where the sample size is not fixed in advance, and preplanned interim analyses of the partial dataset are conducted (typically during the accrual period) to allow early stopping for efficacy or futility. GSDs are frequently used in clinical trials.

**GSD**. See *group sequential design (GSD)*.

**hazard function**. When analyzing time-to-event data, the hazard function at time $t$ is the instantaneous rate of failure at time $t$, conditional on survival until time $t$. The hazard function is written as $h(t) = f(t)/S(t)$, where $f(t)$ is the derivative of the failure function with respect to time, written as $f(t) = \partial F(t)/\partial t$, and $S(t)$ is the survivor function. Also see *failure function* and *survivor function*.

**hazard ratio** and **log hazard-ratio**. The hazard ratio is the ratio of the hazard functions of two different populations. If the hazard functions are proportional, then $h_2(t) = \Delta h_1(t)$ for all $t$ or, equivalently, $S_2(t) = \{S_1(t)\}^{\Delta}$. Here $h_1(t)$ and $h_2(t)$ are the hazard functions for the control group and the experimental group, respectively; $\Delta$ is the hazard ratio; and $S_1(t)$ and $S_2(t)$ are the survivor functions of the control and the experimental groups, respectively.

The log hazard-ratio is the natural logarithm of the hazard ratio. If a log-rank test is used to compare the survivor functions of the two populations, under the proportional-hazards assumption the null hypothesis is $H_0 : \Delta = 1$ or, equivalently, $H_0 : \ln(\Delta) = 0$. See [ADAPT] **gsdesign logrank**.

Also see *hazard function* and *time-to-event data*.

**Hwang–Shih–de Cani bound** or **error-spending Hwang–Shih–de Cani bound** or **Hwang–Shih–de Cani design**. In a group sequential clinical trial, one technique for calculating efficacy or futility boundaries is to use an error-spending Hwang–Shih–de Cani design. Hwang–Shih–de Cani bounds are defined by an error-spending function indexed by parameter $\gamma$, and smaller values of $\gamma$ yield bounds that are more conservative at early looks. Also see *error-spending approach*.

**hypothesis**. A hypothesis is a statement about a population parameter of interest.

**hypothesis testing** or **hypothesis test**. This method of inference evaluates the validity of a hypothesis based on a sample from the population. See *Hypothesis testing* in *Remarks and examples* of [PSS-2] **Intro (power)**.

**information fraction**. In a group sequential clinical trial, the information fraction is the proportion of the maximum information that has been collected at the time of a scheduled look at the clinical trial data. In most cases, the information fraction is the proportion of the maximum sample size that has been collected. For time-to-event data, the information fraction is the proportion of the total number of failure events that have been observed, not the total number of study participants.

**information ratio**. In the context of a group sequential clinical trial, the information ratio is the ratio of the maximum information of the group sequential trial to the Fisher information of an equivalent fixed study design. In most cases, this is the ratio of the maximum sample size of the group sequential trial to the sample size of the fixed design, but for trials with time-to-event endpoints, the information ratio corresponds to the ratio of the maximum number of failure events observed in the group sequential trial to the number of failures observed in a fixed-design trial.

**interim analysis** or **interim look**. In the context of an adaptive clinical trial, an interim look is an analysis of trial data that is conducted while the trial is still under way and before the maximum sample size has been reached.

**Kim–DeMets bound** or **error-spending Kim–DeMets bound** or **Kim–DeMets design**. In a group sequential clinical trial, one technique for calculating efficacy or futility boundaries is to use an error-spending Kim–DeMets design. Kim–DeMets bounds are defined by an error-spending function indexed by parameter $\rho$, and larger values of $\rho$ yield bounds that are more conservative at early looks. Also see *error-spending approach*.

**likelihood-ratio test**. The likelihood-ratio test is one of the three classical testing procedures used to compare the fit of two models, one of which, the constrained model, is nested within the full (unconstrained) model. Under the null hypothesis, the constrained model fits the data as well as the full model. The likelihood-ratio test requires one to determine the maximal value of the log-likelihood function for both the constrained and the full models. See [ADAPT] **gsdesign twoproportions** and [R] **lrtest**.

**look**. In the context of a group sequential clinical trial, a look is an analysis of the clinical trial data that has been collected up to that point. Looks conducted while the trial is still collecting data are called interim looks, and the final look is performed when data from the maximum sample size have been collected. Also see *interim analysis*.

**loss to follow-up**. Participants are lost to follow-up if they do not complete the course of the study for reasons unrelated to the event of interest. For example, loss to follow-up occurs if participants move to a different area or decide to no longer participate in a study. Loss to follow-up should not be confused with administrative censoring. If participants are lost to follow-up, the information about the outcome those participants would have experienced at the end of the study, had they completed the study, is unavailable. Also see *withdrawal*, *administrative censoring*, and *follow-up period*.

**lower one-sided test** or **lower one-tailed test**. A lower one-sided test is a one-sided test of a scalar parameter in which the alternative hypothesis is lower one-sided, meaning that the alternative hypothesis states that the parameter is less than the value conjectured under the null hypothesis. Also see *One-sided test versus two-sided test* in *Remarks and examples* of [PSS-2] **Intro (power)**.

**maximum information**. In a group sequential clinical trial, the maximum information is the Fisher information of the parameter estimated during the hypothesis test, calculated at the maximum sample size. Also see *information fraction*.

**maximum sample size**. In a clinical trial following an adaptive design, the sample size of the trial is often not fixed in advance. However, in many adaptive designs, such as group sequential designs, the maximum possible sample size can be calculated before the study begins. Also see *expected sample size* and *sample size*.

**nominal alpha** or **nominal significance level**. This is a desired or requested significance level. Also see *familywise significance level*.

**nonbinding futility boundaries** or **nonbinding futility bounds**. In a group sequential clinical trial with nonbinding futility bounds, if the test statistic at an interim analysis crosses the futility boundary, the trial may be stopped for futility or continued without risk of overrunning the specified significance level. Group sequential designs with nonbinding futility bounds use the same efficacy critical values as equivalent group sequential designs without futility stopping. Also see *binding futility boundaries*.

**noninferiority trial**. A noninferiority trial is a clinical trial where the goal is to determine whether the experimental treatment is unacceptably inferior to the control (or comparator) treatment, which is almost always an active control. If the experimental treatment has some advantageous characteristics (for example, it produces fewer side effects than the control, is less expensive, or is easier to administer), practitioners might prefer the experimental treatment even if it is not more efficacious than the control.

When designing a noninferiority trial, researchers define a noninferiority margin, denoted as $\delta$, to quantify an acceptable reduction in efficacy. The null hypothesis in a noninferiority trial is that the effect of the control treatment beats the effect of the experimental treatment by a margin of $\delta$ or more; the one-sided alternative hypothesis is that the effect of the control treatment does not beat the effect of the experimental treatment by a margin of at least $\delta$. For example, if the endpoint is a population mean and an upper one-sided test is desired, $\delta$ will be $< 0$ and the null and alternative hypotheses can be written as $H_0 : \mu_e - \mu_c \leq \delta$ and $H_a : \mu_e - \mu_c > \delta$, where $\mu_e$ is the mean response of the experimental group and $\mu_c$ is the mean response of the control group. Also see the related *substantial superiority trial*.

**null hypothesis**. In hypothesis testing, the null hypothesis typically represents the conjecture that one is attempting to disprove. Often, the null hypothesis is that a treatment has no effect or that a statistic is equal across populations.

**O'Brien–Fleming bounds** or **classical O'Brien–Fleming bounds** or **O'Brien–Fleming design**. In a group sequential clinical trial, one technique for calculating efficacy or futility boundaries is to use an O'Brien–Fleming design. O'Brien–Fleming efficacy bounds are characterized by being extremely conservative at early looks. O'Brien–Fleming bounds are a special case of classical Wang–Tsiatis bounds with parameter $\Delta = 0$. Also see *Wang–Tsiatis bounds*.

**observational study**. In an observational study, as opposed to an experimental study, the assignment of participants to treatments happens naturally and is thus beyond the control of investigators. Investigators can only observe participants and measure their characteristics. For example, a study that evaluates the effect of exposure of children to household pesticides is an observational study. Also see *experimental study*.

**observed level of significance**. See *p-value*.

**odds** and **odds ratio**. The odds in favor of an event are $\text{Odds} = p/(1-p)$, where $p$ is the probability of the event. Thus, if $p = 0.2$, the odds are 0.25, and if $p = 0.8$, the odds are 4.

The log of the odds is $\ln(\text{Odds}) = \text{logit}(p) = \ln\{p/(1-p)\}$, and logistic regression models, for instance, fit $\ln(\text{Odds})$ as a linear function of the covariates.

The odds ratio is a ratio of two odds: $\text{Odds}_2/\text{Odds}_1$. The individual odds that appear in the ratio are usually for an experimental group and a control group or for two different demographic groups.

**one-sample test**. A one-sample test compares a parameter of interest from one sample to a reference value. For example, a one-sample mean test compares a mean of the sample against a reference value.

**one-sided test** or **one-tailed test**. A one-sided test is a hypothesis test of a scalar parameter in which the alternative hypothesis is one-sided, meaning that the alternative hypothesis states that the parameter is either less than or greater than the value conjectured under the null hypothesis, but not both. Also see *One-sided test versus two-sided test* in *Remarks and examples* of [PSS-2] **Intro (power)**.

**overall significance level**. See *familywise significance level*.

**Pearson's $\chi^2$ test**. In the context of a clinical trial, Pearson's $\chi^2$ test is commonly used to test whether the observed event counts in a contingency table are consistent with the null hypothesis. See [ADAPT] **gsdesign twoproportions**. Also see *$2 \times 2$ contingency tables*.

**placebo** or **sham treatment**. In a clinical trial, a placebo is an inactive treatment, such as a sugar pill, that is designed to look like the experimental treatment. In studies of medical procedures, the term *sham treatment* is often used. Typically, study participants receiving a placebo are blinded, meaning that they are not told whether they are receiving the placebo or the experimental treatment. Also see *standard of care*.

**placebo control**. In a clinical trial, a placebo control is a control group that receives a placebo instead of an active control.

**Pocock bounds** or **classical Pocock bounds** or **Pocock design**. In a group sequential clinical trial, one technique for calculating efficacy or futility boundaries is to use a Pocock design. Pocock efficacy bounds are characterized by using the same critical value at all looks. Pocock bounds are a special case of classical Wang–Tsiatis bounds with parameter $\Delta = 0.5$. Also see *Wang–Tsiatis bounds*.

**population parameter**. See *target parameter*.

**power**. The power of a test is the probability of correctly rejecting the null hypothesis when it is false. It is often denoted as $1 - \beta$ in the statistical literature, where $\beta$ is the type II error probability. Commonly used values for power are 80% and 90%. See [PSS-2] **Intro (power)** for more details about power.

**power and sample-size (PSS) analysis**. Power and sample-size analysis investigates the optimal allocation of study resources to increase the likelihood of the successful achievement of a study objective. The focus of power and sample-size analysis is on studies that use hypothesis testing for inference. Power and sample-size analysis provides an estimate of the sample size required to achieve the desired power of a test in a future study. See [PSS-2] **Intro (power)**.

**probability of a type I error**. This is the probability of committing a type I error and incorrectly rejecting the null hypothesis. Also see *type I error* and *significance level*.

**probability of a type II error**. This is the probability of committing a type II error and incorrectly accepting the null hypothesis. Common values for the probability of a type II error are 0.1 and 0.2 or, equivalently, 10% and 20%. Also see *type II error*, *beta*, and *power*.

**PSS analysis**. See *power and sample-size (PSS) analysis*.

**p-value**. The $p$-value is the probability of obtaining a test statistic as extreme as or more extreme than the one observed in a sample assuming the null hypothesis is true.

**randomized controlled trial (RCT)**. In this experimental study, treatments are randomly assigned to two or more groups of participants, one of which is a control group.

**recruitment period** or **recruitment**. See *accrual period*.

**rejection region**. In hypothesis testing, a rejection region is a set of values of a test statistic for which the null hypothesis can be rejected. In the context of a group sequential design, a trial can be terminated early for efficacy if the test statistic falls within the rejection region during an interim analysis. Also see *acceptance region* and *continuation region*.

**relative risk**. See *risk ratio*.

**risk difference**. A risk difference is defined as the probability of an event occurring when a risk factor is increased by one unit minus the probability of the event occurring without the increase in the risk factor.

When the risk factor is binary, the risk difference is the probability of the outcome when the risk factor is present minus the probability when the risk factor is not present.

When one compares two populations, a risk difference is defined as a difference between the probabilities of an event in the two groups. It is typically a difference between the probability in the comparison group or experimental group and the probability in the reference group or control group.

**risk factor**. A risk factor is a variable that is associated with an increased or decreased probability of an outcome.

**risk ratio** or **relative risk**. A risk ratio, also called a relative risk, measures the increase in the likelihood of an event occurring when a risk factor is increased by one unit. It is the ratio of the probability of the event when the risk factor is increased by one unit over the probability without that increase.

When the risk factor is binary, the risk ratio is the ratio of the probability of the event when the risk factor occurs over the probability when the risk factor does not occur.

When one compares two populations, a risk ratio is defined as a ratio of the probabilities of an event in the two groups. It is typically a ratio of the probability in the comparison group or experimental group to the probability in the reference group or control group.

**sample size**. This is the number of participants in a sample. In a clinical trial with time-to-event data, the effective sample size is the number of events observed. In this case, sample-size calculations will determine the number of events that must be observed to achieve the specified power. If administrative censoring, loss to follow-up, or withdrawal are expected, the total required sample size can be estimated and will be larger than the number of events observed. Also see *expected sample size*, *fractional sample size*, and *maximum sample size*.

**sample-size determination**. This pertains to the computation of a sample size given power and effect size and any other study parameters.

**sample-size ratio**. The ratio of the experimental-group sample size relative to the control-group sample size, $n_2/n_1$.

**Satterthwaite's t test**. Satterthwaite's $t$ test is a modification of the two-sample $t$ test to account for unequal variances in the two populations. See [ADAPT] **gsdesign twomeans** for an example and see *Methods and formulas* of [PSS-2] **power twomeans** for formulas.

**score test**. A score test, also known as a Lagrange multiplier test, is one of the three classical testing procedures used to compare the fit of two models, one of which, the constrained model, is nested within the full (unconstrained) model. The null hypothesis is that the constrained model fits the data as well as the full model. The score test only requires one to fit the constrained model. See [ADAPT] **gsdesign oneproportion** and [R] **prtest**.

**sensitivity analysis**. Sensitivity analysis investigates the effect of varying study parameters on power, sample size, and other components of a study. The true values of study parameters are usually unknown, and analyses of power and sample size use best guesses for these values. It is therefore important to evaluate the sensitivity of the computed power or sample size in response to changes in study parameters.

**significance level**. In hypothesis testing, the significance level $\alpha$ is an upper bound for the probability of a type I error. Also see *alpha*, *probability of a type I error*, and *familywise significance level*.

**significance level approach**. The efficacy and futility critical values from a group sequential design are intended to be compared with $z$ statistics. If the test statistic used does not follow a standard normal distribution under the null hypothesis, the significance level approach is used to compare the significance level of the test statistic against the significance level of the efficacy critical value. This is done by comparing the $p$-value of the test statistic against the $p$-value corresponding to the efficacy or futility critical value. See *Significance level approach* in *Methods and formulas* of [ADAPT] **gsbounds** for details.

**single-arm trial**. A single-arm clinical trial is a trial where all study participants receive the experimental treatment. Because there is no control group, the endpoint is compared with a prespecified reference value. Also see *two-arm trial*.

**size of test**. See *significance level*.

**standard of care**. The standard of care is the medically accepted first-line treatment for a disease or condition. In a clinical trial of a treatment for a condition where there is a recognized standard of care, it is common to compare the experimental treatment to an active control consisting of participants who receive the standard of care. Also see *active control* and *placebo*.

**stopping boundary**. A stopping boundary is a set of critical values that define an efficacy or futility boundary. Also see *stopping rule*, *efficacy boundaries*, and *futility boundaries*.

**stopping rule**. In the context of a group sequential clinical trial, a stopping rule refers to an efficacy or futility boundary that allows the trial to be terminated before data from the maximum sample size have been collected. This occurs when the test statistic at an interim analysis crosses the efficacy or futility boundary, leading to the rejection or acceptance of the null hypothesis. Also see *efficacy stopping* and *futility stopping*.

**study participant**. Human subjects who volunteer to join a clinical trial are known as study participants.

**substantial superiority trial** or **superiority trial**. A substantial superiority trial is a clinical trial where the goal is to determine whether the experimental treatment is substantially superior to the control treatment. This is done by defining a clinically relevant superiority margin, denoted as $\delta$, before the trial begins. The one-sided alternative hypothesis is that the effect of the experimental treatment beats the effect of the control treatment by a margin greater than $\delta$; the null hypothesis is that it does not. For example, if the endpoint is a population mean and an upper one-sided test is desired, $\delta$ will be $> 0$ and the null and alternative hypotheses can be written as $H_0 : \mu_e - \mu_c \leq \delta$ and $H_a : \mu_e - \mu_c > \delta$, where $\mu_e$ is the mean response of the experimental group and $\mu_c$ is the mean response of the control (or comparator) group. Also see related concept *noninferiority trial*.

**surrogate endpoint**. When the clinical outcome of interest is too difficult, time consuming, or expensive to measure, clinical trials often use a surrogate endpoint as their target parameter. A surrogate endpoint is an endpoint that is known to be associated with the clinical outcome of interest but is easier to measure. Many clinical trials use biomarkers as surrogate endpoints. Also see *endpoint*.

**survivor function**. When analyzing time-to-event data, the survivor function is defined as the probability of surviving beyond time $t$. If we denote the time of failure as $T$, we can define the survivor function as $S(t) = \Pr(T > t) = 1 - F(t)$, where $F(t)$ is the failure function. Also see *hazard function* and *failure function*.

**t test**. A $t$ test is a test for which the sampling distribution of the test statistic is a Student's $t$ distribution.

A one-sample $t$ test is used to test whether the mean of a population is equal to a specified value when the variance must also be estimated. The test statistic follows Student's $t$ distribution with $N - 1$ degrees of freedom, where $N$ is the sample size.

A two-sample $t$ test is used to test whether the means of two populations are equal when the variances of the populations must also be estimated. When the two populations' variances are unequal, a modification to the standard two-sample $t$ test is used; see *Satterthwaite's t test*.

**target parameter**. In power and sample-size analysis, the target parameter is the parameter of interest or the parameter in the study about which hypothesis tests are conducted. Also see *endpoint*.

**test statistic**. In hypothesis testing, a test statistic is a function of the sample that does not depend on any unknown parameters.

**time-to-event data** or **survival data**. Time-to-event data, also known as survival data, are collected from clinical trials where the endpoint is the amount of time elapsed before a participant experiences a failure event. See [ADAPT] **gsdesign logrank**.

**two-arm trial**. A two-arm clinical trial is a trial where participants are assigned to one of two treatment groups. Typically, one group is an experimental group and the other is a control group. Also see *single-arm trial*.

**two-sample test**. A two-sample test is used to test whether the parameters of interest of the two independent populations are equal, for example, a two-sample test of means, proportions, or hazard ratios. See [ADAPT] **gsdesign twomeans**, [ADAPT] **gsdesign twoproportions**, and [ADAPT] **gsdesign logrank**.

**two-sided test** or **two-tailed test**. A two-sided test is a hypothesis test of a parameter in which the alternative hypothesis is the complement of the null hypothesis. In the context of a test of a scalar parameter, the alternative hypothesis states that the parameter is less than or greater than the value conjectured under the null hypothesis.

**type I error**. The type I error of a test is the error of rejecting the null hypothesis when it is true. Also see *probability of a type I error* and *familywise type I error*.

**type II error**. The type II error of a test is the error of not rejecting the null hypothesis when it is false. Also see *probability of a type II error*.

**unbalanced design** or **unequal-allocation design**. An unbalanced design indicates an experiment in which the numbers of treated and untreated participants differ. Also see [PSS-4] **Unbalanced designs**.

**upper one-sided test** or **upper one-tailed test**. An upper one-sided test is a one-sided test of a scalar parameter in which the alternative hypothesis is upper one-sided, meaning that the alternative hypothesis states that the parameter is greater than the value conjectured under the null hypothesis. Also see *One-sided test versus two-sided test* in *Remarks and examples* of [PSS-2] **Intro (power)**.

**Wald test**. A Wald test is one of the three classical testing procedures used to compare the fit of two models, one of which, the constrained model, is nested within the full (unconstrained) model. Under the null hypothesis, the constrained model fits the data as well as the full model. The Wald test requires one to fit the full model but does not require one to fit the constrained model. Also see [ADAPT] **gsdesign oneproportion** and [R] **test**.

**Wang–Tsiatis bounds** or **classical Wang–Tsiatis bounds** or **Wang–Tsiatis design**. In a group sequential clinical trial, one technique for calculating efficacy or futility boundaries is to use a Wang–Tsiatis design. Wang–Tsiatis bounds are indexed by parameter $\Delta$, and smaller values of $\Delta$ yield bounds that are more conservative at early looks. Classical Pocock bounds and classical O'Brien–Fleming bounds are both special cases of the Wang–Tsiatis family of bounds. Also see *Pocock bounds* and *O'Brien–Fleming bounds*.

**withdrawal**. Withdrawal is the process under which participants withdraw from a study for reasons unrelated to the event of interest. For example, withdrawal occurs if participants move to a different area or decide to no longer participate in a study. Withdrawal should not be confused with administrative censoring. If participants withdraw from the study, the information about the outcome those participants would have experienced at the end of the study, had they completed the study, is unavailable. Also see *loss to follow-up* and *administrative censoring*.

***z* statistic**. A $z$ statistic is a test statistic that follows the standard normal distribution under the null hypothesis. Also see *$z$ test*.

***z* test**. A $z$ test is a test for which a potentially asymptotic sampling distribution of the test statistic is a normal distribution. For example, a one-sample $z$ test of means is used to test whether the mean of a population is equal to a specified value when the variance is assumed to be known. The distribution of its test statistic is normal. See [ADAPT] **gsdesign onemean** and [ADAPT] **gsdesign twomeans**.

# Reference

U.S. Food and Drug Administration. 2019. *Adaptive Designs for Clinical Trials of Drugs and Biologics: Guidance for Industry.* Docket No. FDA-2018-D-3124. Silver Spring, MD. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics-guidance-industry.