Medical Biostatistics: Basic Concepts

2

Konstantinos I. Bougioukas
and Anna-Bettina Haidich

Introduction

When physicians begin to read the research literature in their chosen field, one of the first things they will discover is that knowledge of statistics is essential. This chapter provides an overview of essential statistical methods available to landmarks trials that investigate hypertension. It begins with an introduction to the types of variables and then demonstrates methods for summarizing, visualizing, and understanding data. The chapter continues with basic principles in the context of hypothesis testing and interpretation of effect sizes, confidence intervals and p-values. The authors also describe the process of selecting the appropriate statistical test in bivariable analysis (e.g., t-test, ANOVA, Kruskal-Wallis, chi-squared test) and outline basic regression models (multivariable analysis), with a special emphasis on survival analysis and Cox proportional hazards model. It also briefly covers topics such as intention-to-treat and per protocol analyses, interim analysis, subgroup and sensitivity analyses, sample size calculation and power of the study. The focus of the chapter is not on computational formulas, but on basic concepts and ideas with practical examples from published trials. At the end, the reader will have learned the essential principles and tools of biostatistics required for research in hypertension field.

K. I. Bougioukas · A.-B. Haidich (✉)
Department of Hygiene, Social-Preventive Medicine and Medical Statistics, Medical School, Aristotle University of Thessaloniki, Thessaloniki, Greece
e-mail: haidich@gapps.auth.gr

Population, Sampling, Study Design and Randomization

In epidemiology and biostatistics, the term *population* is used for any collection of units, which are often people, but may be, for example, institutions, events, etc. about which the researcher wish to investigate particular properties and draw some conclusions [1].

The *sample* is a finite part or subset of the accessible population that participates in the study [1]. The gold standard for ensuring generalizability (external validity) is the *probability sampling.* These methods use a random process to guarantee that each unit of the population has the same probability of being chosen in the sample [2, 3].

Clinical trials can be classified in several ways, depending on their design. From most to least common in the healthcare literature [4], the major categories of randomized trials are summarized in Table 2.1.

The random procedure that is used by a trial design is called *randomization method*. The most usual randomization methods are simple randomization, block randomization and stratified randomization [11, 12]. A combination of these methods can also be used in trial designs, and other special methods do exist.

An example from the literature with these basic concepts is presented below:

V. Papademetriou et al. (eds.), *Management of Hypertension*,
https://doi.org/10.1007/978-3-319-92946-0_2

Example: Treatment of Hypertension in Patients 80 Years of Age or Older-Hypertension in the Very Elderly Trial (HYVET) [13]

Population: Patients of 80 years of age or older with persistent hypertension (defined as a sustained systolic blood pressure of 160 mm Hg) living in Western and Eastern Europe, China, Australasia, and North Africa.

Sample (size): Entry of 4761 patients

Study design: 3845 patients; 1933 patients assigned to diuretic indapamide and 1912 patients to matching placebo alone (parallel groups).

Randomization method: Randomization was stratified according to age (80–89 years and 90 years or older) and sex; permuted blocks of 4 and 6 of any 10 patients were used to ensure roughly equal assignment to each of the two groups within large centers. (Stratified and blocked randomization).

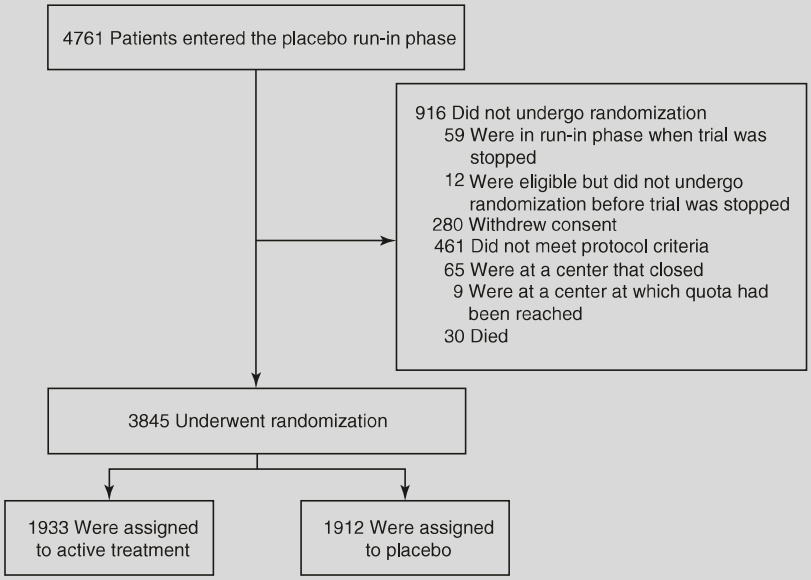


Table 2.1 Classification of randomized trials based on design

Study design (guidelines)	Defining characteristics	Example of trial
Parallel group (CONSORT 2010 [5])	Each participant is randomized to one of the intervention arms.	The Valsartan Antihypertensive Long-Term Use Evaluation (VALUE) Trial [6].
Crossover (under development[a])	Each participant is exposed to each intervention in a random sequence (all patients eventually get all treatments in varying order).	Spironolactone versus placebo, bisoprolol, and doxazosin to determine the optimal treatment for drug-resistant hypertension (PATHWAY-2): a randomized, double-blind, crossover trial [7].
Cluster (CONSORT extension 2012 [8])	Clusters of individuals (for example, clinics and schools) are randomly allocated to different study arms.	Patient education and provider decision support to control blood pressure in primary care: A cluster randomized trial [9].
Factorial (under development[a])	Participants are randomly assigned to individual interventions or a combination of interventions	Effect of blood pressure lowering and antihypertensive drug class on progression of hypertensive kidney disease: results from the AASK trial [10]. (3 × 2 factorial design)

[a]http://www.equator-network.org/library/reporting-guidelines-under-development/

Characteristics of the Subjects as Variables: Basic Types

In any particular trial, several characteristics (such as sex, age, ethnicity, smoking status, systolic blood pressure, or even an event like death from cardiovascular causes) are recorded for the participants of the study. In statistical terms these characteristics are called *variables* since they vary from subject to subject or from time to time [1]. Variables can take either categorical (qualitative data) or numerical (quantitative data) values (Fig. 2.1). The type of data is crucial for the decision regarding presentation (summary measures and graphs) and the techniques of data analysis that are employed [14, 15].

Categorical Variables

Variables are categorical when their data are placed into distinct groups with appropriate labels according to some qualitative characteristic or attribute, for instance place of birth, ethnic group, or type of drug [1, 16]. Categorical variables can be further divided into either nominal or ordinal variables. *Nominal* variables have two or more categories such as sex (male or female) or blood group (A, B, AB, or O) without natural ordering, while *ordinal* variables have an intrinsic order such as degree of pain (none, mild, moderate, or severe) [14, 15]. Note: A nominal variable like sex (male or female) or survival status (alive or dead) which can take only two possible categories is also called *binary* or *dichotomous* [14, 16].

Numerical Variables

Numerical variables take arithmetic values. They can be subdivided in discrete or continuous variables. *Discrete* variables can only take values of a countable set of numbers (which are usually the whole numbers 0, 1, 2, 3, etc.). Examples of discrete variables are the heart rate (beats/min), the number of visits to a GP in a year and the hospitalization days. In contrast, *continuous* variables have no limitation on the values that they can take. The baseline characteristics such as weight, height or blood pressure of the participants are examples of common continuous variables in a trial. However, the actual measurements are restricted by the accuracy of the method used for measuring the value [14, 15].

Moreover, categorization of numerical variables is also common in clinical research although there is a cost of loosing information [1]. For example, in the Systolic Blood Pressure Intervention Trial (SPRINT) [17] the continuous variable of systolic blood pressure was categorized into three categories ($\leq$132 mm Hg, >132 mm Hg to <145 mm Hg, $\geq$145 mm Hg) and was presented in the table of baseline characteristics.

Variables from a Different Point of View

Information on a particular variable is usually collected for one of two reasons. The first is when the variable is an *outcome* of interest. An outcome variable is a characteristic which is believed to be affected by the values taken by other

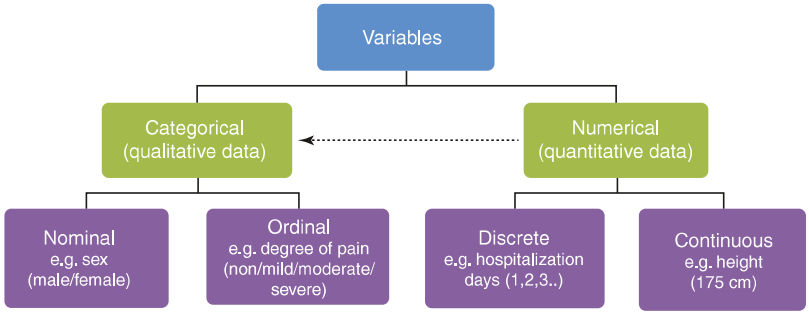


Fig. 2.1 Different types of variables. They can take either categorical (qualitative data) or numerical (quantitative data) values. The numerical variables can be converted into categorical variables. (Adapted from Aviva and Sabin [14])

variables [18]. It is also called a *response* or *dependent* variable. The outcomes measures can be defined as *primary* or *secondary endpoints* [19] based on the primary or secondary objectives of the trial, respectively. For example, in the main Hypertension in the Very Elderly Trial (HYVET) [13] that investigated the relative benefits and risks of antihypertensive treatment in patients 80 years of age or older, the primary endpoint was any stroke (fatal or nonfatal). Secondary endpoints included death from any cause, death from cardiovascular causes, death from cardiac causes, and death from stroke.

The second type of variable that the researcher would want to collect information on is an *explanatory* variable [18]. This is a factor that may influence the outcome or the association of the exposure and outcome (confounding factor). Such a variable partly explains the variability of the outcome. They are also called *independent* or *predictor* variables. In the HYVET study some of the explanatory variables were sex, age, baseline systolic blood pressure while seated, and previous cardiovascular disease.

Presenting Summaries of Variables

Summary Measures and Graphs for Categorical Variables

Categorical data are typically summarized by reporting the number (absolute frequencies) and the percentage (relative frequencies) of cases occurring into each category. The information from two categorical variables at once can be presented in a *two-way table* (*cross table* or *contingency table*), such as the one shown in Table 2.2 (one categorical variable is the race/ethnicity with five categories and the other is the treatment groups of the trial; data are from the Controlled Onset Verapamil Investigation of Cardiovascular End Points [CONVINCE] trial [20]) and this display of data is called a *frequency distribution* [21].

From this table we can see that in both treatment groups the majority of patients were white (COER verapamil group = 84.2%, Atenolol or Hydrochlorothiazide group = 84.5%). Another,

Table 2.2 Cross table of the race/ethnicity and the treatment group for the participants in CONVINCE trial

Race/ethnicity	Treatment No. (%) of participants	
	COER Verapamil (n = 8179)	Atenolol or Hydrochlorothiazide (n = 8297)
White	6864 (84.2)	6981 (84.5)
Black	559 (6.9)	563 (6.8)
Asian	99 (1.2)	100 (1.2)
Hispanic	592 (7.3)	579 (7.0)
Other	36 (0.4)	41 (0.5)

Data are available in CONVINCE trial [20]
Abbreviation: COER controlled-onset extended-release

important feature of this table is that the percentages within each treatment group (column percentages) add up to 100% (e.g., COER Verapamil group 84.2% + 6.9% + 1.2% + 7.3% + 0.4% = 100%). Moreover, this is a way to cross-check that the calculations have been performed correctly.

For categorical demographic variables, such as race/ethnicity, authors may well find that tables suffice for simple and concise recording of data. However, if the information in the table is sufficiently important, communicating it graphically may be a better choice [22]. For example, the reader of CONVINCE study [23] can immediately recognize from the presented side by side (or grouped) bar graph (Fig. 2.2) that for the primary end point, in each treatment group, more participants had primary events between 6 AM and noon than any other 6-h period.

Bar graphs are frequently used to present categorical variables. Another study that provides informative bar graphs is the Symplicity HTN-1 study [24]. This open-label study investigated the long term changes in blood pressure after renal denervation (RDN) in patients with treatment-resistant Hypertension. The proportion of patients with systolic blood pressure of 180 mm Hg or higher decreased over the duration of the study, from 30% at baseline to 5% at 36 months. The proportion who achieved target systolic blood pressure values of less than 140 mm Hg increased significantly at all time points (Fig. 2.3; stacked bar graph). At 1 month after RDN, 55 of 80 (69%) patients had reductions in systolic blood pressure of at least 10 mm Hg, which rose

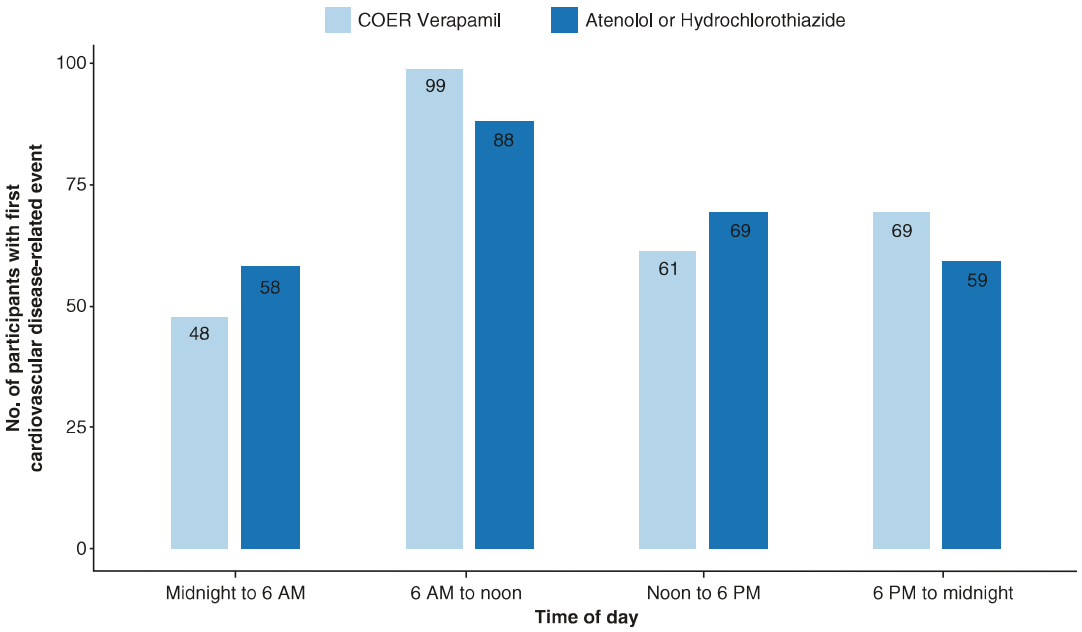


Fig. 2.2 Time of onset of first cardiovascular disease–related event was determined for 277 participants in the controlled onset extended-release (COER) verapamil group and 274 participants in the atenolol or hydrochlorothiazide group. (Grouped bar graph with frequencies; adapted from CONVINCE study [23])

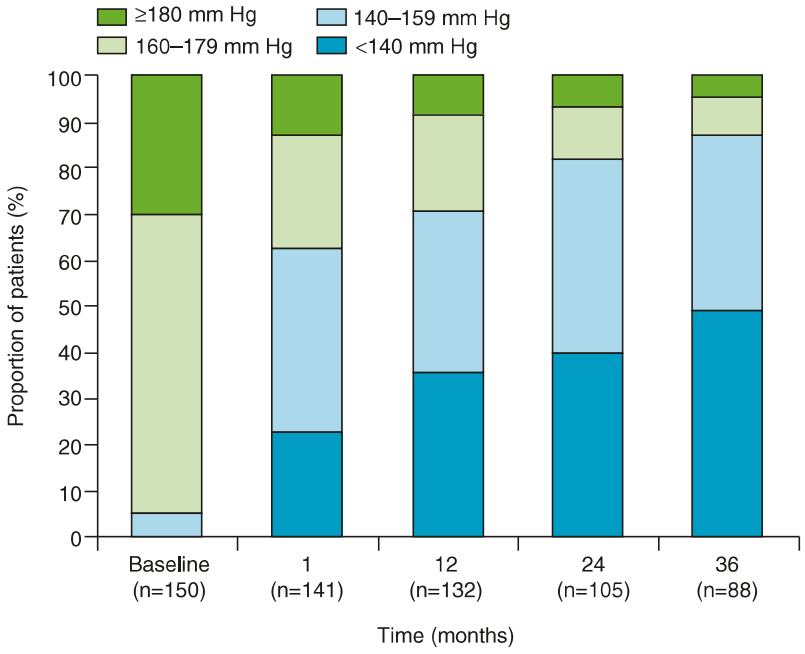


Fig. 2.3 Distribution of changes in systolic blood pressure for all treated patients (stacked bar graph). (Adapted from Symplicity HTN-1 study [24])

progressively to 82 of 88 (93%) at 36 months. Reductions of 20 mm Hg or more were seen in 68 of these 88 patients (77%) (Fig. 2.4; grouped bar graph with percentages).

A pie chart is another graph that can be used for the presentation of the categorical variables. The chart consists of a circle subdivided into sections, one for each category or group, so that the

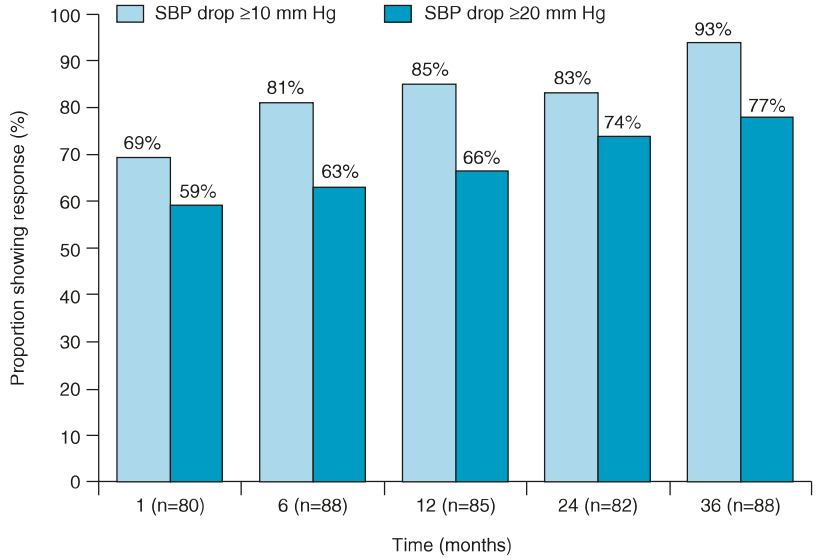


Fig. 2.4 Proportion of patients assessed to 36 months who showed treatment responses at different time points in the study (grouped bar graph with percentages). (Adapted from Symplicity HTN-1 study [24])

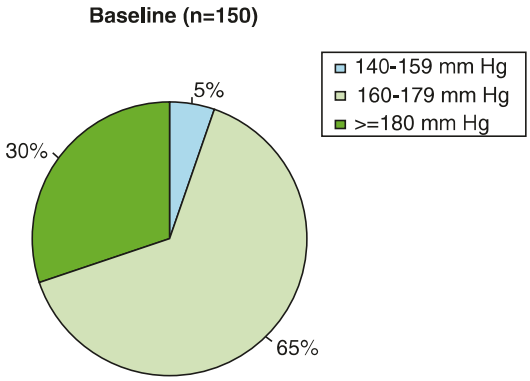


Fig. 2.5 Distribution of patients (n = 150) according to their blood pressure value at baseline

area of each section is proportional to the frequency in that category [14]. This type of chart is rarely used in scientific papers [25] because it requires too much space to present too little information, whereas there are better visualization alternatives such as bar charts. For example, the pie chart in Fig. 2.5 represents the distribution of patients (n = 150) according to their blood pressure value at baseline, that is the first stacked bar (baseline) of the Fig. 2.3. If the researcher had decided to present all the information of the graph, it would be needed multiple pie charts (five different pie graphs). However, multiple pie charts takes up a lot of the limited manuscript space and are difficult to analyze and interpret, especially when comparing adjacent pies [26, 27].

Summary Measures and Graphs for Numerical Variables

Two basic *summary measures* should be reported for a numerical variable. The first measure indicates where the center of the distribution of the values lies. This is an index of location (or central tendency) because it defines the center, or middle, of the sample data. The second measure describes the 'spread' of the observations, how widely the values are spread above and below the central value, and is called *variability* (or *dispersion*) of the distribution.

Measures of Location

There are three measures commonly used to describe the location or 'center' of the distribution of a numerical variable [28]:

1. *Arithmetic Mean*

The mean (or average), of a set of values is calculated by summing up all the values of observations and dividing by the total number of observations. The mathematical formula for n observations is

$$\overline{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_{n-1} + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where x_i is the ith observation of the sample. The capital Greek sigma Σ is a summation sign and is simply a short way of writing the quantity $x_1 + x_2 + x_3 + \ldots + x_{n-1} + x_n$.

The arithmetic mean is, in general, a natural measure of location and uses all the data values. However, it is influenced by extreme values, known as outliers or distorted by skewed data.

2. *The Median*

The median of a variable is the place that divides the data in half, once the data are ordered from smallest to largest. It is thought to be the "middle" value.

The sample median is:

- The $\left(\frac{n+1}{2}\right)$th largest observation if n is odd
- The average of the $\left(\frac{n}{2}\right)$th and $\left(\frac{n}{2}+1\right)$th largest observations if n is even

The median can be more appropriate for distributions that are skewed. When the distribution is symmetrical, the median equals the mean.

3. *The Mode*

The mode is the most frequently occurring value among all the observations in a sample. There may be more than one mode if two values are equally frequent. The major disadvantage is that it ignores most of the information.

Measures of Variability

Several different measures can be used to describe the variability of a sample [28]. Two different variables can have the same arithmetic mean but can be made up of very different values.

1. *Range*

Perhaps the simplest measure is the range. It is defined as the difference between the largest and the smallest observations in a sample. The range is markedly influenced by extreme values.

2. *Variance*

The average of the squares of the deviations from the sample mean. The resulting measure of spread, denoted by s^2, is:

$$s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n-1}$$

The units of the variance are the square of the units of the original observations. Variance is sensitive to outliers and it is inappropriate for skewed data.

3. *Standard deviation*

The standard deviation (SD) is defined as the square root of variance.

$$sd = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n-1}}$$

It is evaluated in the same units as the raw data. As the variance, it is sensitive to outliers and thus it is inappropriate for skewed data.

4. *Interquartile range (IQR)*

Range between percentiles (percentile is the value below which a given percentage of the data observations occur. e.g., the 25th percentile is the value below which 25% of the observations may lie) can also measure the variability of the sample. A common range used is the *interquartile range* (IQR), which is the range between the 25th and 75th percentile, thus the middle 50% (75%–25%) of the data is included between these two values. Again, data have to be ordered first from smallest to largest value. It is appropriate for skewed data.

Example
Suppose 11 baseline home systolic blood pressure records (mm Hg) in a trial with resistant hypertension:

134.9, 143.7, 151.0, 132.4, 150.2, 148.0, 148.7, 162.3, 131.5, 162.3, 137.8

Mean

$$\bar{x} = \frac{134.9+143.7+151.0+132.4+150.2+148.0+148.7+162.3+131.5+162.3+137.8}{11} =$$

$$\bar{x} = \frac{1602.8}{11} = 145.7\,\text{mm Hg}$$

Median
131.5, 132.4, 134.9, 137.8, 143.7, **148.0**, 148.7, 150.2, 151.0, 162.3, 162.3
Md = 148.0

Mode
134.9, 143.7, 151.0, 132.4, 150.2, 148.0, 148.7, **162.3**, 131.5, **162.3**, 137.8
Mo = 162.3 mm Hg

Range
Range = 162.3–131.5 = 30.8 mm Hg

Standard deviation (sd) and percentiles can be easily calculated with a statistical package such as R program:

Standard deviation (sd)
sd = 10.8 mm Hg

Interquartile range (IQR)
Percentiles

0%	**25%**	50%	**75%**	100%
131.5	**136.4**	148.0	**150.6**	162.3

Therefore
IQR = (150.6–136.4) = 14.2 mm Hg.

Histogram

The *histogram* is a graphical representation of the distribution of numerical data. They are typically used as tools for inspecting the distribution of numerical variables and get a "feel" for the data. A histogram gives information about [29]:

- How the data are distributed (Fig. 2.6): (a) left-skewed, (b) symmetric (e.g., normal distribution), (c) right-skewed.
- The amount of variability in the data
- Where the "center" of the data is (approximately) located

In an approximately normal distribution such as Fig. 2.6b, the mean (red line), the median (blue line) and the mode (green line) have very close values and the histogram is symmetric about the mean. Moreover, "nearly all" values (99.7%) of a normal distribution are within the interval ($\bar{x} - 3sd$, $\bar{x} + 3sd$).

Box Plot

A *box plot* chart is another graph that can be used for conveying location and variation information for continuous data, particularly for detecting changes between different groups of data before any formal analyses are performed. Figure 2.7 illustrates such a diagram that examines the yearly risk of recurrent lobar intracerebral hemorrhage (ICH) based on systolic blood pressure categories [30].

Box lower and upper margins indicate 25th (known as Q1; the value at which 25% of the data fall below) and 75th percentiles (known as Q3; the value at which 25% of the data fall

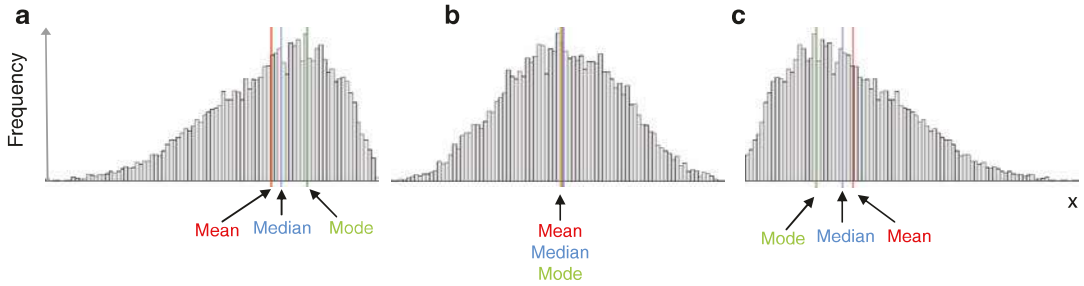


Fig. 2.6 Histograms of (**a**) a negative asymmetric distribution (left-skewed), (**b**) a normal (bell-shaped) distribution and (**c**) a positive asymmetric distribution (right-skewed)

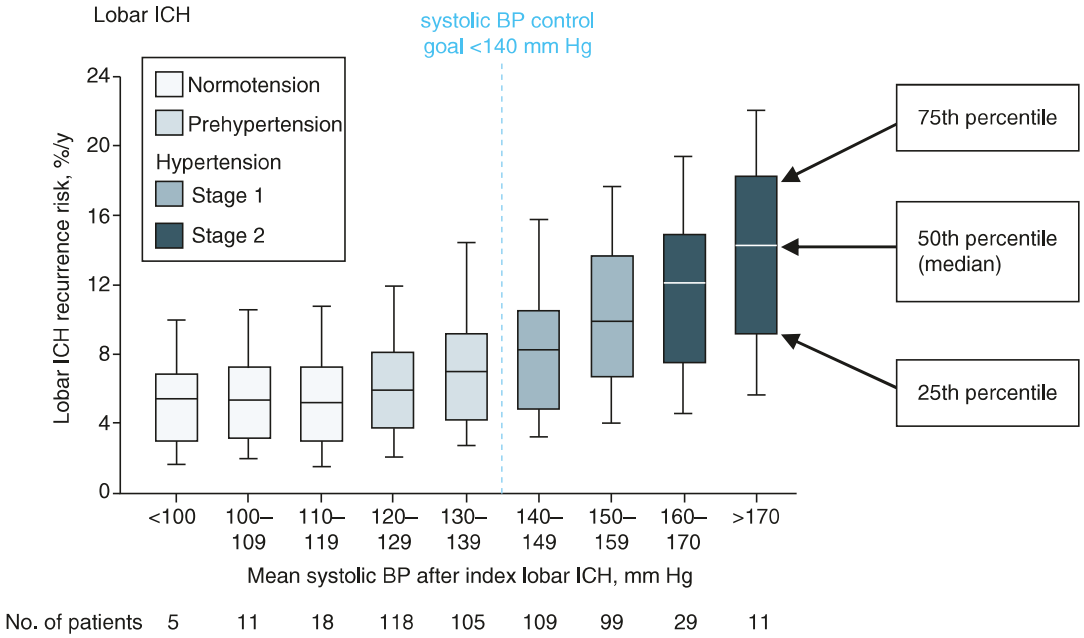


Fig. 2.7 Estimated yearly risk of recurrent lobar ICH (intracerebral hemorrhage) based on systolic BP (blood pressure) measurements during follow-up. (Adapted from Biffi et al. [30])

above) of yearly risk of recurrent lobar ICH, respectively; therefore, the boxes include the middle 50% of the observations. Horizontal lines in boxes indicate median values (50th percentile; Q2); error bars (or whiskers) indicate maximum and minimum estimated values in each distribution (Spear-style). Note: Tukey-style whiskers extend to a maximum of 1.5 × IQR beyond the box, while Altman-style whiskers can also be defined to span the 95% central range of the data [31].

A boxplot that is symmetric with the median line at approximately the center of the box and with symmetric whiskers suggests that the data may have come from a normal distribution.

In the statistical analysis methodology of a research paper, it should be indicated clearly how demographic data and clinical

outcomes will be summarized. The following format is recommended [32]:

- Mean (standard deviation [sd]) for continuous or discrete variables with symmetric distributions.
- Median (first quartile [Q1], third quartile [Q3] or minimum [min], maximum [max]) for those with skewed distributions.
- Number (percentage) for categorical variables.

Examples from the literature:

"The Mean (SD), or median (interquartile range) values are quoted for the biometric and biochemical variables" (UKPDS 1998 [33]).

"We expressed continuous variables as means and standard deviations and qualitative variables as percentages" (Duran-Cantolla J. et al. [34]).

Other Popular Charts and Graphs for Numerical Data Used in Trials

Other commonly used graphs for presenting important numerical data are time charts (or line graphs), scatter plots and dynamite plots.

Time Chart (or Line Graph)

Typically a time chart has some unit of time on the horizontal axis (year, day, month, and so on) and a numerical variable on the vertical axis (usually systolic or diastolic blood pressure in mm Hg for trials with hypertension patients). At each time period, the amount is shown as a dot, and the dots connect to form the time chart [29]. Moreover, error bars can be added to each dot (Fig. 2.8).

Figure 2.8 shows that the two therapeutic strategies quickly resulted in different systolic blood-pressure levels. After the first year of therapy, the average systolic blood pressure at the 4-month protocol visits that both groups attended was 119.3 mm Hg in the intensive-therapy group and 133.5 mm Hg in the standard-therapy group, resulting in an average between-group difference of 14.2 mm Hg.

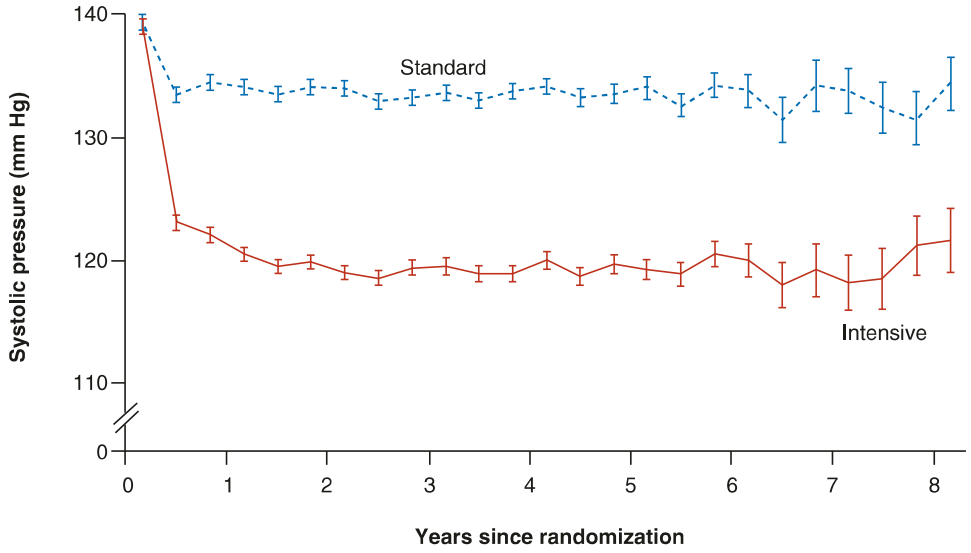
Scatter Plots

When two continuous variables are measured, the nature of the association between them can be explored graphically with a scatter plot [22]. Scatter plots are usually used prior to analyses to help assess the association of the variables to particular analytical procedures. Figure 2.9 is an example of a (grouped) scatter plot that investigates the association of office systolic blood pressure (SBP) and 24-h mean SBP in individual patients at baseline for sustained hypertension (red dots) and white-coat hypertension (blue dots) (European Lacidipine Study on Atherosclerosis [ELSA] trial that investigated white-coat hypertension [36]).

As shown in Fig. 2.9, in SH, 24-h mean SBP correlated with office SBP values (progressively lower values of one pressure were associated with progressively lower values of the other). In striking contrast, in WCH, 24-h mean SBP values were all in a narrow range, independently of the level of office SBP. In both SH and WCH patients, 24-h SBP was always lower than office SBP, the two values becoming progressively closer as SBP became less and the difference seems to disappear at about 130–120 mm Hg office SBP.

Dynamite Plots

Dynamite plots are bar plots where group means (usually mean changes from baseline blood pressure for studies examining hypertension) are represented by the tops of bars or columns and are very often presented in trials [37]. In this graph from ROX CONTROL HTN study [38] the reader can see that mean changes in office and 24 h ambulatory systolic blood pressure at 6 months were greater in the arteriovenous coupler group than in the control group (Fig. 2.10). Net mean differences were all in favour of the arteriovenous coupler group (office blood pressure −23.2 mm Hg systolic and −17.7 mm Hg diastolic, and ambulatory blood pressure −13.0 mm Hg systolic, and −13.4 mm Hg diastolic). However, the body of the bar has no logical interpretation and this



Mean no. of medications prescribed								
Intensive	3.2	3.4	3.4	3.5	3.5	3.5	3.4	3.4
Standard	1.9	2.1	2.1	2.2	2.2	2.3	2.3	2.3
No. of patients								
Intensive	2174	2071	1973	1792	1150	445	156	156
Standard	2208	2136	2077	1860	1241	504	203	201

Fig. 2.8 Mean systolic blood-pressure levels at each study visit with error bars. (Adapted from ACCORD [35])

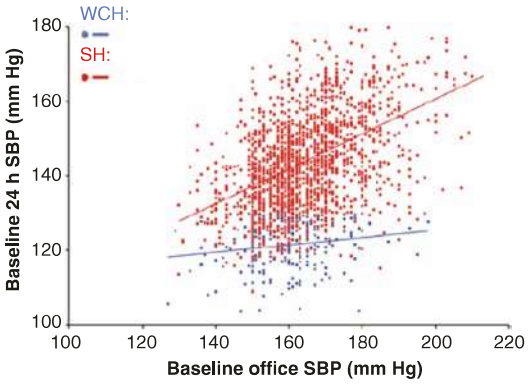


Fig. 2.9 Correlation between office systolic blood pressure (SBP) and 24-h mean SBP in individual patients at baseline. Data are shown separately for sustained hypertension (SH) and white-coat hypertension (WCH). (Adapted from Mancia et al. [36])

graph may not be appropriate for representing means [31]. Better alternative graphs are box plots or means plots (means presented by points with error bars).

Common Measures of Association

Measures of association such as risk ratio (RR) and odds ratio (OR) can be defined by constructing two-by-two contingency tables [39]. However, the most common measure that is reported in hypertension landmark trials is the hazard ratio (HR) that is derived from Cox models.

Risk Ratio

In a clinical trial to assess the impact of a new treatment on occurrence of an event, the risk ratio (RR) (or relative risk) could be calculated. For example in ALLHAT trial [40] for the heart failure outcome (Table 2.3) the risk ratio was reported comparing the Lisinopril treatment with Clorthalidone treatment (the hazard ratio could not be estimated because proportional hazards assumption was violated-see the section "Survival

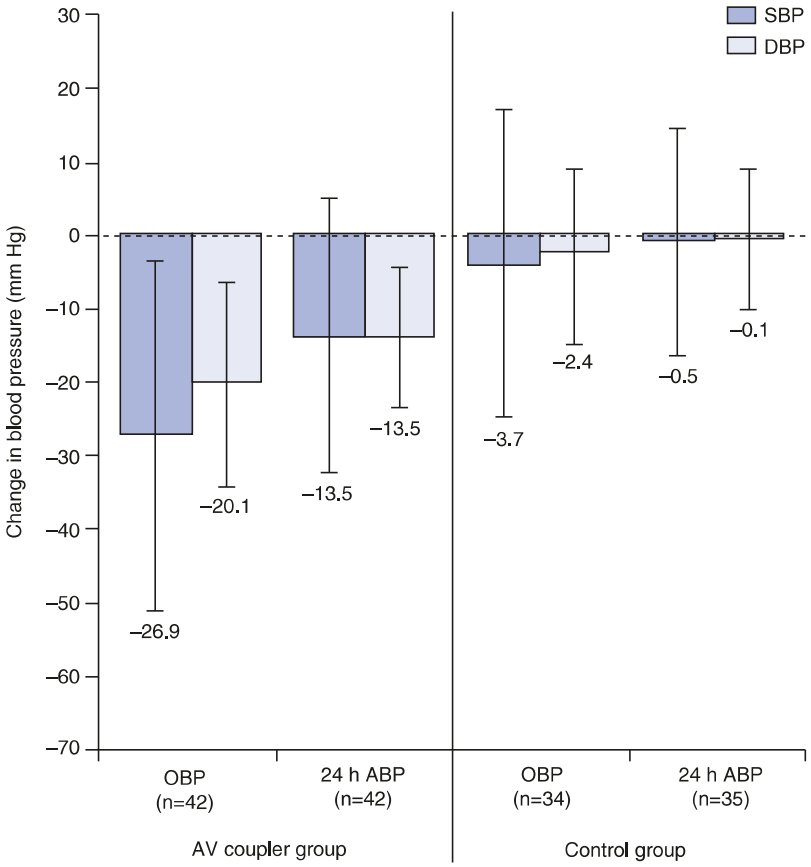


Fig. 2.10 Change from baseline in blood pressure at 6 months. Data are mean (SD). SBP systolic blood pressure, DBP diastolic blood pressure, OBP office blood pressure, ABP ambulatory blood pressure, AV arteriovenous. (Adapted from ROX CONTROL HTN study [38])

Table 2.3 Two-by-two contingency table for calculating risk ratio (data are available in ALLHAT trial [40])

		Outcome *Heart failure*		
		Yes	No	Total
Exposure	Treatment A: *Lisinopril*	a = 612	b = 6053	a + b = 6665
	Treatment B (control): *Chlorthalidone*	c = 870	d = 10491	c + d = 11361

Analysis and Cox Regression" p. 41). In this situation, the risk ratio is the ratio of the probability of occurrence of an event (risk) between two groups (e.g., treatment A vs treatment B; treatment B can be considered as control group):

$$Risk\ \ Ratio\left(RR\right)=\frac{risk\,in\,treatmentA}{risk\,in\,treatmentB\left(control\right)}$$

A risk ratio of 1 occurs when the risks are the same in the two groups and is equivalent to no association between the exposure to different treatments and the outcome. A risk ratio greater than 1 occurs when the risk of the outcome is higher among those exposed to the treatment A than among the treatment B. A risk ratio less than 1 occurs when the risk is lower among those exposed to treatment A, suggesting that the treatment A may be more protective than B when the outcome is negative e.g., heart failure. The further the risk ratio is from 1, the stronger the association between treatment and outcome. Note that a risk ratio is always a positive number (0, ∞).

The risk ratio in ALLHAT trial (Table 2.3) is calculated:

$$Risk\ \ in\ \ treatment\ \ A=\frac{a}{a+b}=\frac{612}{6665}=0.0918$$

$$Risk\ in\ treatment\ B\left(control\right)=\frac{c}{c+d}$$

$$=\frac{870}{11361}=0.0766$$

$$\textbf{Risk Ratio}: RR=\frac{\frac{a}{a+b}}{\frac{c}{c+d}}=\frac{\frac{612}{6665}}{\frac{870}{11361}}=\frac{0.0918}{0.0766}=1.19$$

The risk of heart failure was 1.19 times higher in Lisinopril than in Clorthalidone group. In other words, the Lisinopril group had a 19% (1.19–1 = 0.19) higher risk for heart failure compared to Clorthalidone group.

Odds Ratio

The odds ratio can also be calculated from the Table 2.3. The odds ratio (OR) is the ratio of the odds of the outcome event in the treatment group compared to the control group:

$$Odds\ \ in\ \ treatment\ \ A=\frac{a}{b}$$

$$Odds\ \ in\ \ treatment\ \ B\ \left(control\right)=\frac{c}{d}$$

$$Odds\ \ Ratio\left(OR\right)=\frac{odds\ in\ treatmentA}{odds\ in\ treatment\ B\left(control\right)}$$

OR = 1 Exposure does not affect odds of outcome (no association)
OR > 1 Exposure to treatment A associated with higher odds of outcome
OR < 1 Exposure to treatment A associated with lower odds of outcome

For ALLHAT trial example (Table 2.3) the odds ratio is calculated:

$$\textbf{Odds Ratio}: OR=\frac{\left(\frac{a}{b}\right)}{\left(\frac{c}{d}\right)}=\frac{a*d}{b*c}=\frac{612*10491}{6053*870}=1.2$$

The odds of heart failure were 1.2 times higher in Lisinopril group than in Clorthalidone group. In other words, the odds of heart failure were 20% (1.2–1 = 0.2) higher in Lisinopril group. In this example the incidence of the heart failure was low and the OR is similar to RR.

Hazard Ratio

The majority of clinical trials record the length of time from study entry to a disease endpoint for a treatment and a control group. In this occasion, the hazard ratio is the most appropriate measure to be calculated and reported. Hazard ratio is a measure of relative risk over time in circumstances where we are interested not only in the total number of events, but in their timing as well. It is an estimate of the ratio of the hazard rate in the treated versus the control group. The Cox proportional hazards model (see the section "Survival Analysis and Cox Regression" p. 41) is usually used to calculate the hazard ratio, as it cannot be calculated directly from the crude numbers.

The hazard ratio is interpreted in a similar manner to the risk or odds ratio therefore values above one indicate a raised hazard, values below one indicate a decreased hazard and values equal to one indicate that there is no increased or decreased hazard of the endpoint. For example, in the SPRINT trial the hazard of heart failure was 38% (HR = 0.62; 0.62–1 = −0.38) lower in the intensive treatment than in standard treatment.

The hazard ratio is sometimes used interchangeably to mean a relative risk; however, this interpretation is not correct. The hazard ratio incorporates the change over time, whereas the relative risk can only be computed at single time points, generally at the end of the study.

Parameters and Statistics

Parameters are the summary description of the characteristics of the population (Fig. 2.11). For example the mean baseline home systolic blood pressure, μ_{SBP}, and its standard deviation, σ_{SBP}, in the population of patients assigned in the PATHWAY-2 study [7] were unknown parameters of interest. These unknown values were estimated from the sample data (n = 335) using the statistics $\bar{x}$ (sample mean) and *sd* (sample standard deviation) that equal to 147.6 mm Hg and 13.2 mm Hg, respectively. Therefore, *statistics* are measures of

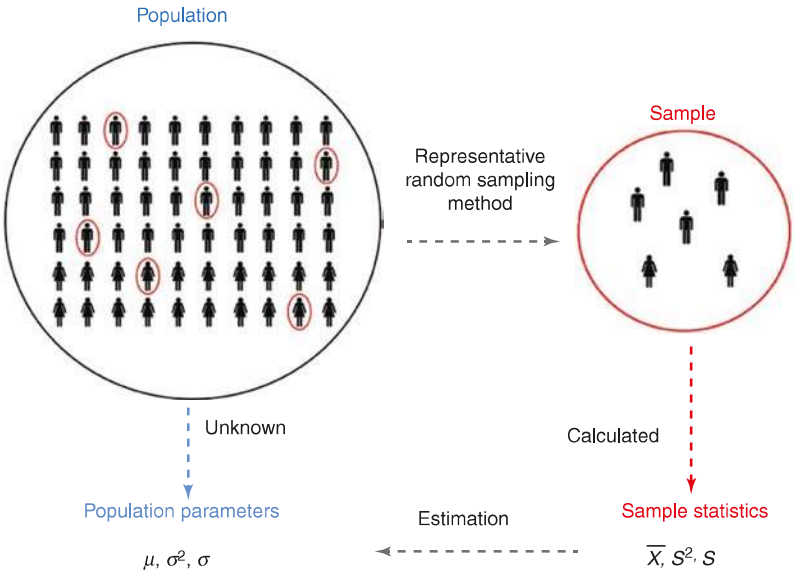


Fig. 2.11 Parameters are referred to the population while statistics are referred to the sample

numerical characteristics that describe the sample and can be considered as estimates of unknown population parameters. (Note: Greek letters refer to population attributes while their sample counterparts are Roman letters).

A **test statistic,** such as t statistic, is a quantity derived from the sample and is used in statistical hypothesis testing.

General Method for Hypothesis Testing (p-Value Approach)

In the statistical methods of the ACCORD BP trial [35] is reported that "the ACCORD BP trial was designed to have 94% power to detect a 20% reduction in the rate of the primary outcome for participants in the intensive therapy group, as compared with those in the standard-therapy group, assuming a two-sided alpha level of 0.05".

In another trial (UKPDS 38) [33] is referred that "in the text relative risks are quoted as risk reductions and significance tests were two sided. For aggregate end points 95% confidence intervals are quoted, whereas for single end points 99% confidence intervals are quoted to allow for potential type 1 errors."

The reader of the articles comes across with concepts such as power, type of errors, significance tests, confidence intervals, two-sided tests and alpha level of 0.05. The theory behind these elements is based on hypothesis testing. The basic steps of this theory is outlined below [14, 41]:

Basic Steps of Hypothesis Testing

Step-by-step hypothesis testing is following with more details:

Steps in hypothesis testing

1. From the research question, determine the appropriate null hypothesis, H_0, and the alternative, H_a.
2. Set the level of significance, α (usually 0.05).
3. Identify the appropriate test statistic and calculate the observed test statistic from the data.
4. Using the known distribution of the test statistic, calculate the p-value.

   Compare the *p*-value to significant level α. If *p*-value < α, reject the null hypothesis. If *p*-value ≥ α, do not reject the null hypothesis.
5. Interpret the results.

Step 1: State the Null and Alternative Hypotheses

The two types of hypotheses are the null and alternative. The null hypothesis (H_0) is a statement that indicates that no difference exists between conditions, groups, or variables while the alternative hypothesis (H_a) indicates a difference or association. The alternative hypothesis may be one-tailed or two-tailed, depending on the context of the research. One-tailed, hypothesis indicates a statistically significant change in a particular direction. For example, a treatment that is expected to show an improvement would be one-tailed. A two-tailed, hypothesis indicates a statistically significant change, but in no particular direction. For example, a researcher may compare two new conditions with no assumed difference between them. However, it is not known which condition would show the largest result.

Step 2: Set the Level of Significance Associated with the Null Hypothesis

When the researcher performs a particular statistical test, there is always a chance that the result is due to chance instead of any real difference. This means that the findings will lead to reject the null hypothesis when it is actually true. In this situation, a *type I error* is occurred. Therefore, statistical tests assume some level of uncertainty that it is called level of significance or alpha (the Greek letter α). The researcher chooses, before the data are collected, the level of significance (usually $\alpha = 0.05$) associated with the null hypothesis. In situations in which the clinical implications of incorrectly rejecting the null hypothesis are severe, it may require stronger evidence (more strict criteria) before rejecting the null hypothesis e.g., $\alpha = 0.01$ or $\alpha = 0.001$.

Step 3: Choose and Calculate the Appropriate Test Statistic Specific to H_0

The researcher should choose a particular type of test statistic based on characteristics of the data. For example, some tests are appropriate for comparing independent groups, while other tests are appropriate for dependent groups. Normal distribution of the data also plays an important role in choosing an appropriate test statistic.

Step 4: Compare the p-Value to Significant Level α. Reject or Not Reject the Null Hypothesis

The test statistic follows a known theoretical probability distribution. The value of the test statistic obtained from the sample is related to a known distribution to obtain the p-value, the area in both (or occasionally one) tails of the probability distribution. **The p-value is the probability of obtaining the observed results, or something more extreme, if the null hypothesis is true.** Most statistical packages provide the two-tailed p-value automatically. The smaller the p-value, the greater the evidence against the null hypothesis.

Conventionally, if p-value < 0.05, there is sufficient evidence to reject the null hypothesis, as there is only a small chance of the results occurring if the null hypothesis was true. In this occasion, the results are significant at the 5% level.

In contrast, if *p*-value ≥ 0.05, there is insufficient evidence to reject the null hypothesis and the results are not significant at the 5% level. This does not mean that the null hypothesis is true; but simply that there is not enough evidence to reject it.

Quoting a result only as significant at a certain cut-off level (e.g., stating only that $p < 0.05$) can be misleading. For example, H_0 would be rejected for $p = 0.049$ but not for $p = 0.051$. Therefore, it is recommended quoting the exact p-value obtained from the computer output. P-values less than 0.001 usually are reported as $p < 0.001$.

Step 5: Interpret the Results

Communicating results in a meaningful and comprehensible manner makes the research useful to others.

We present an example (e.g., chi-squared test) with the steps of hypothesis testing in Table 2.4.

In decision making, there are four possible scenarios that can happen regarding the truth in the population versus the results in the study sample, which are shown in Table 2.5:

Table 2.4 Practical procedures for hypothesis testing

Steps	Procedure	Example from ACCORD trial [35]
Step 1	State the null and alternative hypotheses.	H_0: There is no association between serious adverse events attributed to antihypertensive treatment and treatment strategy (adverse events and treatment strategy are independent). Ha (two-sided): There is an association between serious adverse events and treatment strategy (adverse events and treatment strategy are dependent).
Step 2	Set the level of significance associated with the null hypothesis.	Two-sided alpha level of $\alpha = 0.05$.
Step 3	Choose and calculate the appropriate test statistic specific to H_0.	Chi-squared test ($\chi^2 = 20.4$).
Step 4	Compare the p-value to the level of significance. Reject or not reject the null hypothesis.	The p-value < 0.001 is smaller than 0.05. Reject the null hypothesis.
Step 5	Interpret the results.	Adverse events and treatment strategy are dependent. As compared with the standard-therapy group, the intensive-therapy group had significantly higher rates of serious adverse events attributed to antihypertensive treatment.

Table 2.5 Type I and II errors in hypothesis testing

		In population the null hypothesis is	
		True (there is no difference)	False (there is difference)
Decision based on the sample	Not reject the null hypothesis	Correct decision: 1−α	Type II error: β (False negative)
	Reject the null hypothesis	Type I error: α (False positive)	Correct decision: 1−β (power of the study)

Errors in Hypothesis Testing

Types of error in hypothesis testing

Type I error: the null hypothesis is rejected while it is true; it is also called a false positive result. It is concluded that there is an effect when, in reality, there is none. The maximum chance (probability) of making a Type I error is denoted by α (alpha). This is the significance level of the test.

Type II error: we do not reject the null hypothesis when it is false, and conclude that there is no effect when one really exists; it is also called a false negative result. The chance of making a Type II error is denoted by β (beta); its compliment, $(1 - \beta)$, is the power of the test. The ***power***, therefore, is the probability of rejecting the null hypothesis when it is false.

Hypothesis Testing and Confidence Interval (CI)

In medical statistics, a confidence interval (CI) is a type of interval estimate that shows the precision of an effect of interest. For example, a 95% confidence interval (95% CI) of the effect of interest (e.g., the difference in means) indicates that if the experiment was repeated many times under the same conditions (same sample sizes, same sampling method) on the same population and the 95% CIs of the effect (difference in means) were calculated, then 95% of these CIs would be expected to capture the true effect (that is the true parameter value) [1, 42]. In Fig. 2.12 the confidence intervals of the 100 randomly generated samples (sample size = 60) are presented. Each vertical bar is a confidence interval, centered on a sample mean (green point). The intervals all have the same length, but are centered on different sample means as a result of random sampling. The five red confidence intervals do not cover the true population mean (the horizontal red line $\mu = 2.25$). This is what we would expect using a 95% confidence level–approximately 95% of the intervals covering the population mean.

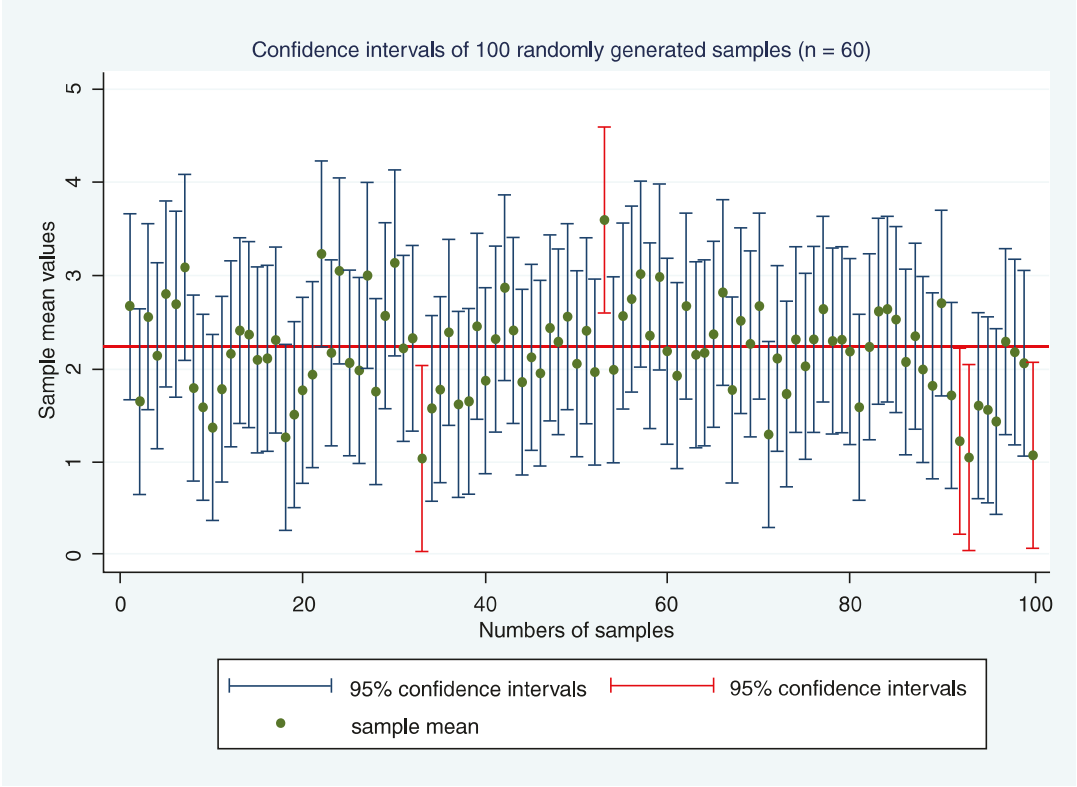


Fig. 2.12 The confidence intervals of the 100 randomly generated samples (sample size = 60). Ninety-five of them covering the population mean (the horizontal red line μ = 2.25). (Source: http://sites.nicholas.duke.edu/statsreview/ci/)

Confidence intervals and hypothesis tests are closely linked. The primary aim of a hypothesis test is to make a decision and provide an exact p-value. A confidence interval quantifies the effect of interest (e.g., the difference in means), and enables us to assess the clinical implications of the results. However, because it provides a range of plausible values for the true effect, it can also be used to make a decision although an exact p-value is not provided. If the confidence interval does not contain the null hypothesis value (e.g., the value zero when the effect of interest is the mean difference, or the value 1 for ratios, such as the risk or odds ratio), the p-value is less than the alpha level and the results are statistically significant ($p < \alpha$). If the confidence interval contains the null hypothesis value, the p-value is equal or larger than the alpha level and the results are not statistically significant ($p \geq \alpha$).

For example in ALLHAT [40], no significant difference (HR = 0.98; 95% CI: 0.90–1.07; p = 0.65) was observed between amlodipine and chlorthalidone for the primary outcome (fatal coronary heart disease or nonfatal myocardial infarction) as the confidence interval included the value 1 and $p = 0.65 > 0.05$.

Basic Statistical Tests

Principles for Choosing Statistical Test in Bivariable Analysis

Basic statistical tests can be categorized as parametric or non-parametric. A *parametric statistical test* makes assumptions about the parameters of the population distribution from which one's data are drawn. Examples of such tests are t-test and analysis of variance (ANOVA) test. *Non-parametric tests* (also called distribution-free tests) are the ones that make no such assumptions. Examples of non-parametric tests include the Mann-Whitney test and Kruskal-Wallis test [43]. If the sample size is large enough, the

parametric tests usually have more statistical power than nonparametric tests counterparts.

We present the principles in guiding the choice of basic statistical tests in bivariable analysis with one dependent (numerical or categorical) variable and one categorical independent variable with levels-groups [44]. First of all, the researcher should answer what type of measurement is the dependent variable (numerical or categorical) and then how many groups are included in the independent variable. The next step is to inspect whether the groups of measurements are related (dependent groups) or not (independent groups). Measurements taken from the same participants at different time points (e.g., before-after studies, cross-over trials) must be analyzed using tests for paired data. The groups of measurements collected in these study designs are dependent.

In case of a numerical dependent variable (Table 2.6 and Fig. 2.13), the researcher further has to decide whether a parametric or a nonparametric test is more appropriate to be used. For example, if a continuous variable is approximately normally distributed in two independent groups (or the sample size per group is large enough), a two sample t-test can be conducted, else the corresponding non-parametric test, Mann-Whitney test, can be used (for highly skewed data or small sample sizes). There are various methods to check for normal distribution, e.g., plotting histograms or using one of the many "normality tests" (such as the Shapiro-Wilk test).

For examining the association between two categorical variables the tests presented in Table 2.7 are used. The flow chart for choosing each test is shown in Fig. 2.14.

Multiple Comparisons Problem

For comparison of three or more groups the researcher usually applies classical statistical tests such as analysis of variance (ANOVA) or Kruskal-Wallis test. For example in ANOVA the null hypothesis is that all means are equal while the alternative hypothesis is that there is at least

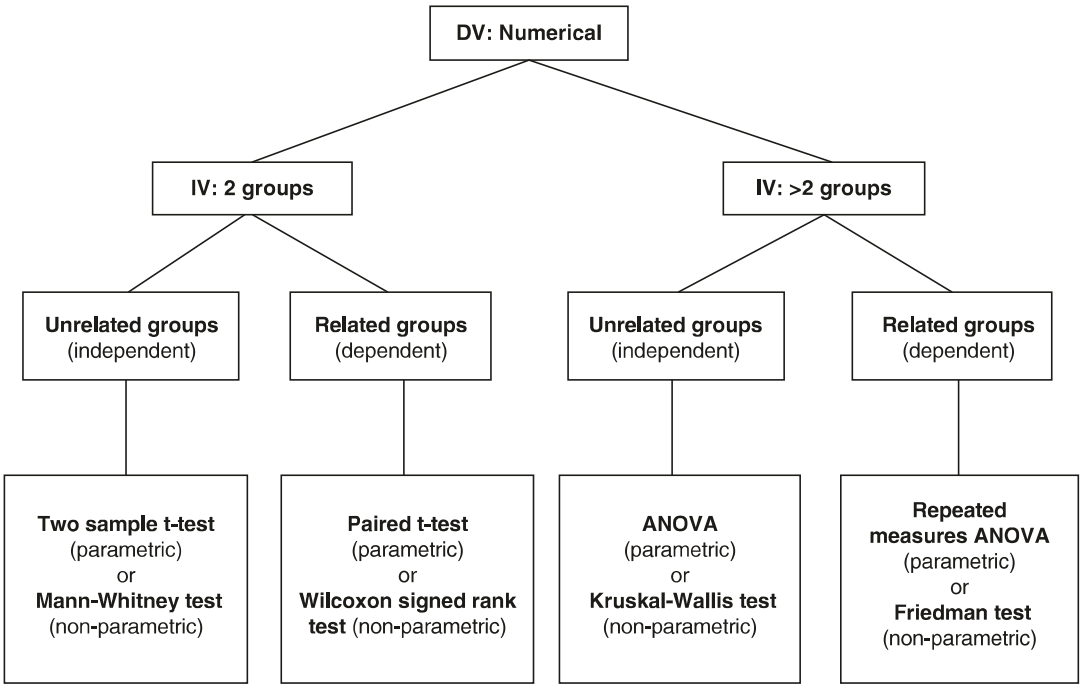


Fig. 2.13 Flow chart for choosing statistical when the dependent variable is numerical. DV Dependent variable, IV Independent variable

Table 2.6 Basic statistical tests in bivariable analysis with a numerical dependent variable

	Groups	Number of groups	Parametric statistical test	Description	Corresponding non-parametric statistical test	Description
Independent variable: categorical	Unrelated groups (independent)	2 groups	Two-sample t-test	Two sample t-test is used to compare the means of two independent samples. Each sample should follow an approximately normal distribution and the variances should be approximately equal. (The test can also be applied when each group has large number of observations). Example: Comparison of baseline systolic blood pressure in the intensive treatment group with the standard treatment group.	Mann-Whitney test	The Mann-Whitney is used to test the null hypothesis that two independent samples come from identical population distributions when the dependent variable is either ordinal or continuous.
		>2 groups	One-way analysis of variances (ANOVA)	One-way ANOVA is used to compare the means of several independent samples. Each sample has to follow an approximately normal distribution and the variances should be equal (homogeneity). (It can also be applied when each group has large number of observations). Example: Comparison of baseline systolic blood pressure in four race or ethnic groups (black, white, hispanic, and other) of patients at high risk for cardiovascular events.	Kruskal-Wallis test	The Kruskal-Wallis test is a rank-based non-parametric test that can be used to decide whether more than two population distributions are identical without assuming them to follow the normal distribution.
	Related groups (dependent)	2 groups (paired data)	Paired t-test	Paired t-test is used to assess whether the mean of the differences between two related measurements is significantly different from zero. Differences should follow approximately the normal distribution. (The test can also be applied when the number of patients is large). Example: Comparison of changes from baseline of 24-h mean systolic blood pressure at 12 and 24 months treatment in sustained hypertension patients.	Wilcoxon signed rank test	The Wilcoxon signed-rank test makes use of the sign and the magnitude of the rank of the differences between pairs of measurements.
		>2 groups (repeated measurements)	Repeated measurements ANOVA	Repeated measures analysis of variance (ANOVA) can be used when several measurements of the same dependent variable are taken at different time points. Assumptions: Normality of the residuals[a] by time point. Sphericity: variances of all possible difference scores are equal. (The test can also be applied for large number of measurements). Example: Comparison of changes from baseline of 24-h mean systolic blood pressure at 12, 24, 36 and 48 months treatment in sustained hypertension patients.	Friedman test	The Friedman test is used when the data arise from more than two related samples. It is based on ranks.

[a]residuals: the difference between the observed value and the estimated value of the quantity of interest

Table 2.7 Basic statistical tests in bivariable analysis with a categorical dependent variable

	Groups	Number of groups	Statistical test	Description
Independent variable: categorical	Unrelated groups (independent)	≥2 groups	Chi-squared test or Fisher's exact test (if there are expected frequencies less than 5)	Chi-squared test is used to determine whether there is a significant association between two categorical variables. It works well when the expected frequencies are large, otherwise Fisher's exact test can be applied. Example: Investigate the association between serious adverse events (yes/no) and treatment (Placebo, Spironolactone, Doxazosin or Bisoprolol).
	Related groups (dependent)	2 groups	McNemar's test or Exact binomial test (for small samples)	McNemar's test, for 2 × 2 tables, is used to assess whether there is a significant change in proportions over time for paired data or whether there is a significant difference in proportions between matched cases and controls. Alternatively, Exact binomial test can be used for small samples. Example: Investigate if the proportion of patients with systolic blood pressure > 130 mm Hg (yes/no) differs between the baseline and completion of the treatment (paired groups).
		>2 groups	Cochran's Q test	Cochran's *Q* test is used to determine if there are differences on a dichotomous dependent variable between three or more related groups. It is an extension to the McNemar's test. Example: Investigate if the proportion of patients with systolic blood pressure > 140 mm Hg (yes/no) differs between baseline, 6, 12 and 18 months treatment (four dependent groups of measurements).

one mean which is different from others. Rejecting the null hypothesis does not indicate which groups differ from the other groups. Consequently, researcher needs to examine patterns of differences among groups. However, this requires multiple comparisons of group means that lead to inflation of the Type I errors (the probability of falsely rejecting H_0) [45, 46]. For example, if there are g = 4 groups and pairwise comparisons are conducted with individual t tests at the significant level of 0.05 (5%) (individual error rate, IER), there are k = g(g−1)/2 = 12/2 = 6 possible pairwise comparisons (1 with 2, 1 with 3, 1 with 4, 2 with 3, 2 with 4 and 3 with 4). In this situation, the probability of at least one Type I error for the family of 6 tests (familywise error rate, FWER) is approximately $FWER = 1 - (1 - IER)^k = 1 - (1 - 0.05)^6 = 1 - (0.95)^6 = 0.265$ (26.5%).

Post-hoc Adjustment (Bonferroni Correction)

It is possible to use some form of post-hoc adjustments to take account of the number of tests performed. Many methods exist to manage the multiplicity problem [46, 47]. A commonly used approach is the Bonferroni correction which adjusts the statistical significance threshold by the number of tests [45]. For example, for a

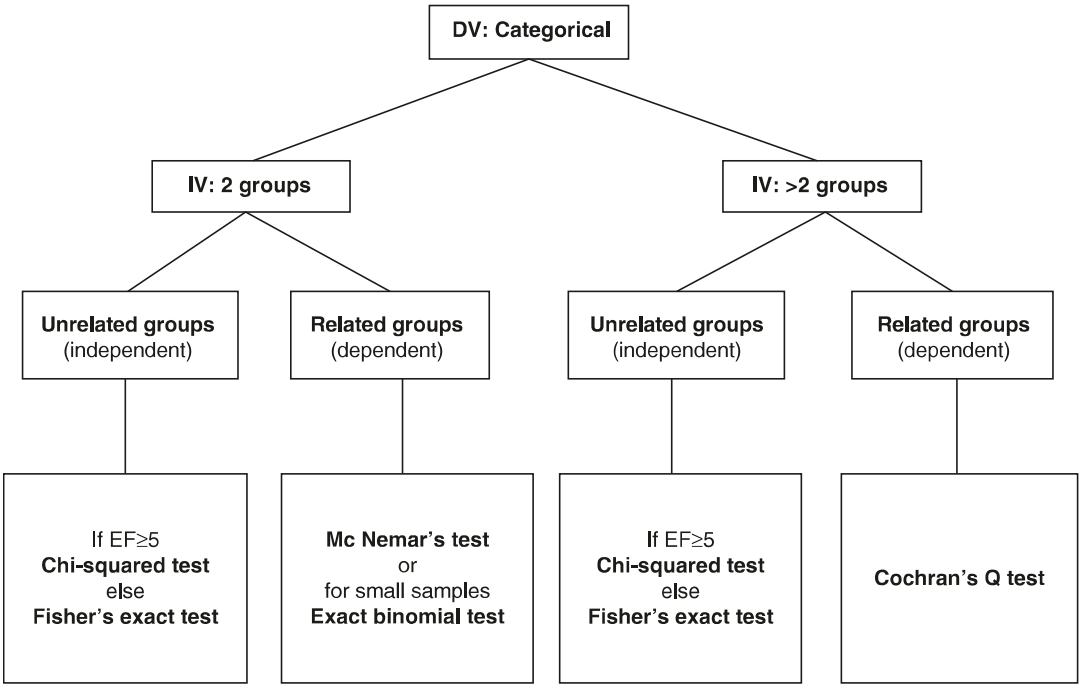


Fig. 2.14 Flow chart for choosing statistical when the dependent variable is categorical. DV Dependent variable, IV independent variable, EF expected frequencies

FWER fixed at 5%, the IER in a group of six tests is set at FWER/k = 0.05/6 = 0.008; therefore, an individual t-test must have a *p*-value less than 0.008 to be considered statistically significant. Even though the Bonferroni test controls the FWER, in many situations it may be too conservative and not have enough power to detect significant differences [46].

(Note: Equivalently, the researcher may multiply each individual p-value by the number of tests carried out in order to compute p_{adj}-values; any decisions about significance are then based on these adjusted p-values [14]. For example, if the researcher conducts six comparisons with t tests while keeping FWER = 0.05, each p-value of the t tests should be multiplied by 6 and then be compared with 0.05).

Basic Regression Models (Multivariable Analysis)

Clinical outcomes come in a variety of different types. Some are continuous, such as systolic blood pressure, and can be analyzed with linear regression. If the observed outcome is dichotomous/binary such as if a patient dies from a specific disease or not, logistic regression can be applied. However, if the information on the time to death is the observed outcome of interest, data are analyzed using statistical methods for survival analysis [48]. This statistical analysis uses the generic term "survival", although this method can be applied for any *time-to-event* outcome other than mortality. For example the outcome measured could be the time that the patient remains free of certain complications (*event-free survival-EFS)* or the time that the disease does not get worse (*progression free survival-PFS)*.

Regression analysis is used for explaining or modeling the association between a single variable *Y*, let's call this *response* variable (or just outcome), and one or *more explanatory* variables, $X1,\ldots,Xp$. When the number of parameters (p) is $p = 1$, it is called simple regression but when $p > 1$ it is called multiple regression. The most common models that are used in medical research are the linear, logistic and Cox regression models [28]. The characteristics of these models are outlined in Table 2.8. The survival analysis and

Table 2.8 Common regression models that are used in medical research

Type of outcome variable	Regression model	Measure of exposure effect	Basic assumptions	Description
Numerical	**Linear** $\mathbf{y = b_o + b_1x_1 + b_2x_2 + \ldots + b_px_p}$ The number of subjects per variable (SPV) should be at least 10.	Mean difference	1. Independence of the residuals 2. Linearity: associations between explanatory variables and response variable are linear 3. Homoscedacity 4. No multicollinearity 5. Normality of the residuals	Association between a numerical response variable and one or more (numerical or categorical) explanatory variables. Example: We would like to see if the treatment strategy adjusted for other covariates (age, sex, weight and smoking status) has an effect on 24-h mean systolic blood pressure (SBP) in sustained hypertension patients.
Dichotomous	**Logistic** $\mathbf{logit(P) = b_o + b_1x_1 + b_2x_2 + \ldots + b_px_p}$ 10 outcome events per variable (EPV) with the least common event determining the maximum number of independent variables. For example, in a study of 170 participants, assume that 100 patients achieved the goal of home systolic blood pressure of <135 mm Hg and 70 patients did not achieve it after antihypertensive treatment. Therefore at most, seven independent variables can be included (since 70 is the smallest outcome).	Odds ratio (OR)	1. Linearity of independent continuous variables with the log odds of the model 2. No multicollinearity	Association between a dichotomous response variable and one or more (numerical or categorical) explanatory variables. Example: We would like to see if the treatment strategy adjusted for other prespecified baseline covariates (sex, age, height, weight, smoking history, and the baseline value of home systolic blood pressure) has an effect on achieving a home systolic blood pressure of <135 mm Hg.
Time to a binary event	**Cox (proportional hazards)** $\mathbf{ln(h(t)) = ln(ho(t)) + b_1x_1 + b_2x_2 + \ldots + b_px_p}$ 10 outcome events per variable (EPV)	Hazard ratio (HR)	1. Linearity of independent continuous variables with the ln (hazard) 2. No multicollinearity 3. The hazard ratios are assumed to be constant over time (proportionality assumption).	Association between time to a binary event (death, failure, relapse) and one or more (numerical or categorical) explanatory variables. Example: We would like to see if the antihypertensive treatment strategy adjusted for other covariates (body size, age, sex, diabetes, and previous history of cardiovascular events, stroke, or chronic kidney disease) has an effect on time to cardiovascular death for hypertension high risk patients.

Residuals: the difference between the observed value and the estimated value of the quantity of interest interest; logit(P): ln(P) = ln(P/1−P) where P is the probability for the event to occur and ln is the natural logarithm

Homoscedacity: constant variance of the residuals

Multicollinearity: can occur in the regression model if two or more explanatory variables are significantly related to each other

Cox regression is presented more analytically in the following sections because these statistical approaches are used in the majority of the landmark trials that examine hypertension.

Survival Analysis and Cox Regression

In analyzing survival or time-to-event data, there are several important quantities of interest to define.

Event Definition

A clearly defined event is crucial for the presentation of survival data. Examples of potential events are: (1) death from any cause, (2) disease progression, (3) diagnosis with a specific disease, or (4) recovery (e.g., return to work) [49].

Start Time

Another crucial component of any survival analysis is the start time or time zero. This is the time point that is most important in relation to the time at which the event under study occurs. In a clinical trial this is typically the time of randomization and ensures comparability of the treatment arms. Because the time variable used in plots and analyses is the time since time zero, it also defines all the times at which surviving subjects are assumed to be comparable [50].

Censoring

The distinguishing feature of survival data is that at the end of the follow-up period, the event will probably not have occurred for all participants in the study (Fig. 2.15) (only participants 1, 5, and 7 experienced the event). This can be because the participant is lost to follow-up (e.g., has moved away) (participant 6) or is withdrawn (participant 3), or because the end of the study observation period is reached without the subject having an event (participants 2 and 4). For these participants, survival time is said to be right-censored. Although these may seem to be cases of missing data as the time-to-event is not actually observed, these subjects are highly valuable as the observation that they went a certain amount of time without experiencing an event is itself informative. One of the most important properties of survival methods is their ability to handle such censored observations which are ignored by methods such as a Mann-Whitney test (non-parametric test because time-to-event data are usually skewed)

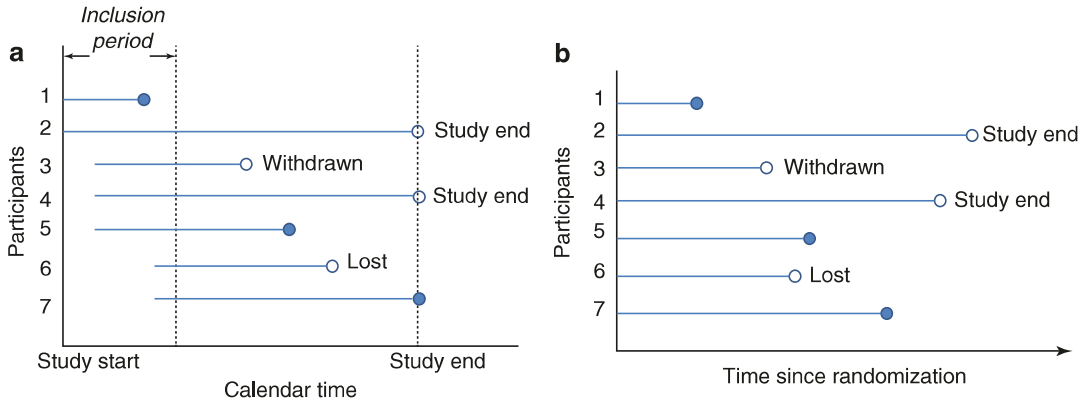


Fig. 2.15 Diagram (**a**) shows participants profiles in calendar time. The participants may enter the study at different times during the inclusion period. Diagram (**b**) ignores the different starting times and convert calendar time into survival time. Solid blue circles indicate participants who had the event while white circles indicate those who had censored data

for comparing survival times of two independent groups.

Survival analysis takes into account censored data and, therefore, utilizes the information available from a clinical trial more fully [18, 49, 51].

Survival Function

One of the most important quantities is the *survival function*, denoted by S(t), which provides the probability of surviving beyond a specific point in time (denoted t) [52]. As t gets larger, the probability of an event increases and therefore $S(t)$ decreases. Plotting a graph of probability against time produces a *survival curve*, which is a useful component in the analysis of such data (Fig. 2.16). Since $S(t)$ is a probability, it is always between zero and one for all values of t, $0 \leq S(t) \leq 1$. When $t = 0$, $S(0) = 1$, indicating that all patients are event-free at the start of study and theoretically, if the study period increased without limit, everyone will experience the event, so the survivor curve must eventually fall to zero. In practice, when using actual data, we usually obtain graphs that are step functions. Cumulative survival drops with every experienced event, whereas it remains unchanged with every censored observation (indicated by the red plus signs). Moreover, because the study period is never infinite in length, it is possible that not everyone studied gets the event. Thus, $S(t)$ may not go all the way down to zero at the end of the study [49].

Hazard Function

Hazard is defined as the immediate risk of event occurrence. The hazard function *h(t)* gives the instantaneous potential per unit time for the event to occur, given that the individual has survived up to time t. Note that, in contrast to the survivor function, which focuses on not failing, the hazard function focuses on failing, that is, on the event occurring. It is always non-negative and has no upper bound [49].

Regardless of which function S(t) or h(t) one prefers, there is a clearly defined relationship between the two. In fact, if one knows the form of S(t), one can derive the corresponding h(t), and vice versa.

Kaplan-Meier Approach and Log-Rank Test

In comparing the survival distributions of two or more groups (for example, new therapy vs standard of care), Kaplan-Meier estimation and the log-rank test are the basic statistical methods of analyses.

Kaplan-Meier approach is usually presented in placing curves for different treatment groups on the same graph that allows the reader to graph-

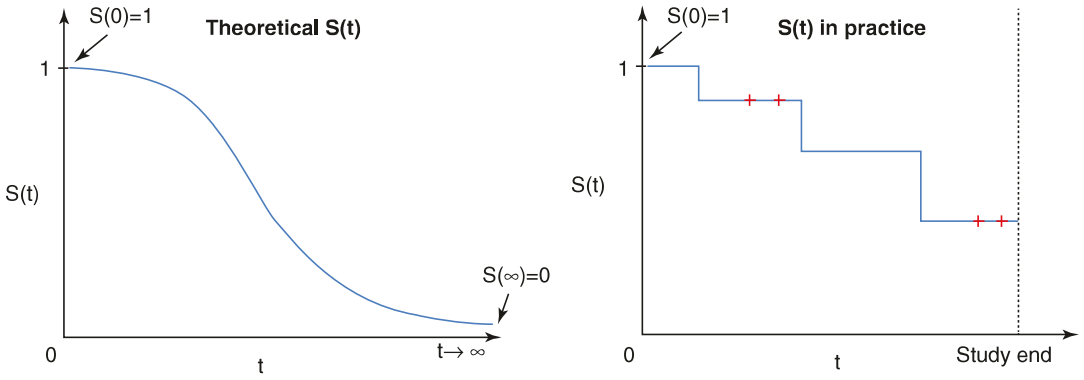


Fig. 2.16 The theoretical survival curve (on the left) and the step chart for the S(t) that is produced in practice (on the right)

ically review any treatment differences. This can be done in one of two ways [53]:

Kaplan-Meier Survival Plots

A Kaplan-Meier survival (K-M) plot presents an estimate (Kaplan-Meier estimator) of the probability of surviving beyond each time (vertical axis) versus time (horizontal axis). The curves in a K-M plot decrease with time from 1 (or 100%), displaying the proportion of patients that have survived (or remain event-free) (Fig. 2.17).

Samples of survival times are frequently highly skewed, therefore, the median is generally a better measure of central location than the mean. This value (the point at which half the patients have experienced the event) can be estimated for each curve by proceeding horizontally from the 0.5 point on the Y-axis until the survivor curve is reached and then proceeding vertically downward until the X-axis is crossed at the median survival time. For example in Fig. 2.17 the median overall survival (OS) time for metastatic colorectal cancer patients treated with bevacizumab-containing therapy was estimated to be 19.9 months in patients with hypertension (HTN) and 12.3 months in normotensive patients [54]. (Note: each patient's HTN status was determined 3 months after date of initiation of bevacizumab-containing therapy and from that time point was calculated the OS time).

Figure 2.18 depicts event free Kaplan-Meier curves for the participants in the SPRINT trial in a paper that conducted subgroup analysis [55]. It shows that the proportion of event free acute decompensated heart failure (ADHF) for the treatment group consistently lies above that for the placebo group; this difference indicates that the intensive treatment has better prognosis than the standard treatment at all time points of follow-up. Notice, however, that the two survival functions are somewhat closer together in the six months of follow-up, but thereafter they are quite spread apart. This widening gap suggests that the treatment is more effective later during follow-up than at the start. However, the estimated median event free time was not reached either for the

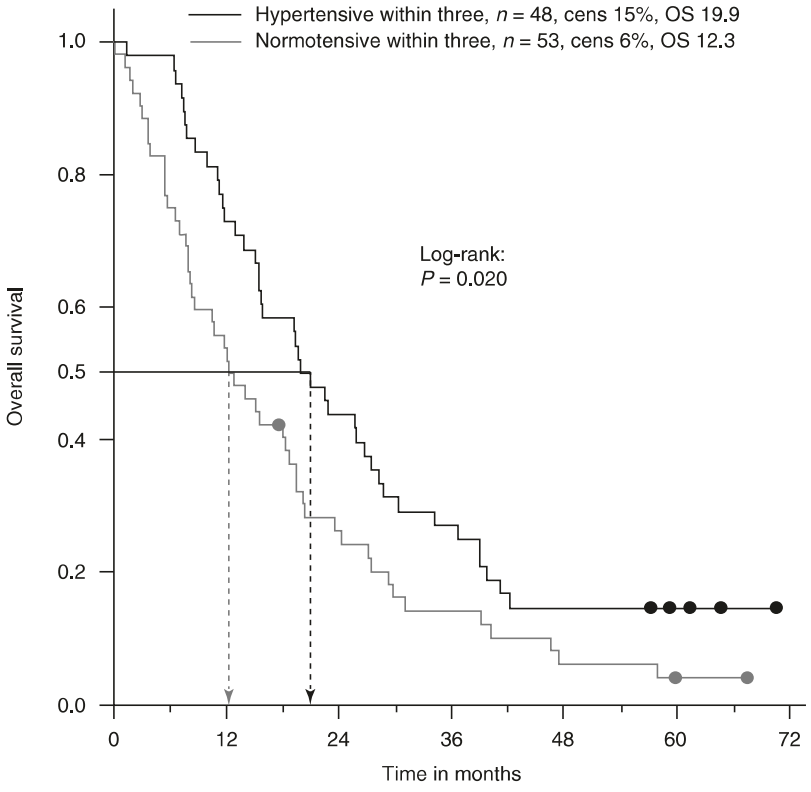


Fig. 2.17 Kaplan Meier survival curves of overall survival (OS) for metastatic colorectal cancer patients with hypertension and no hypertension treated with bevacizumab-containing therapy. The median overall survival time is calculated (on-study date: after 3 months of the initiation of bevacizumab-containing therapy) for each group. (Adapted from Österlund et al. [54])

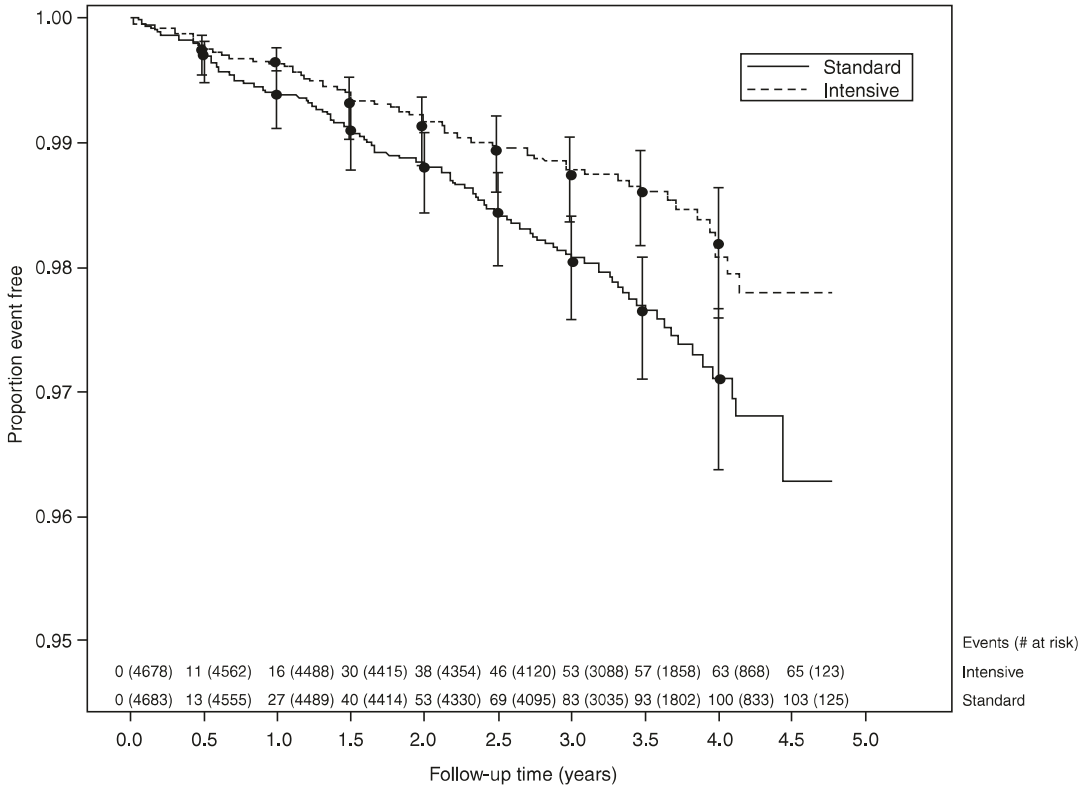


Fig. 2.18 Kaplan–Meier curves for the SPRINT (Systolic Blood Pressure Reduction Intervention Trial [55]) acute decompensated heart failure outcome by treatment group. Vertical bars indicate 95% confidence intervals. Number at risk and number of events is shown every 6 months

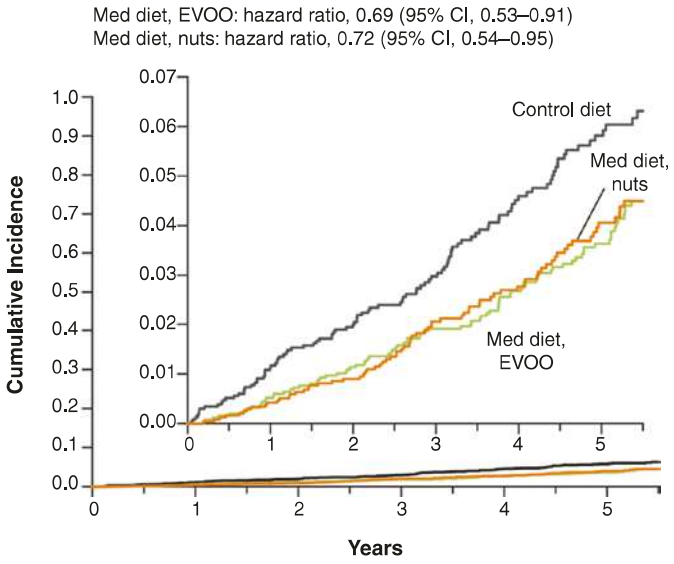
intensive treatment or for the standard treatment. (Note: for this reason, studies sometimes report the estimated time point at which a lower percentile (e.g., 25th) of the study population has the event [50]).

The 95% confidence intervals of the event free survival curves are shown with vertical bars in Fig. 2.18. In practice, there are usually patients who are lost to follow-up or event free at the end of follow-up, and confidence intervals are often wide at the tail of the curves due to lower number of patients, making meaningful interpretations difficult [56].

Cumulative Incidence Plots

Cumulative incidence plots are an alternative to Kaplan-Meier survival curve plots and are used very often in clinical trials. This plot shows the cumulative probability that the event of interest has occurred over the course of the observation period. The increase in event rates is shown starting from 0 (or 0%) subjects at time zero with an increasing curve over time [18, 53]. An informative example is shown in Fig. 2.19 obtained from the PREDIMED trial [57] that compares three different groups (Mediterranean Diet with extra-virgin olive oil (EVOO), Mediterranean Diet with Nuts and Control Diet). Actually, it presents the same plot in two different vertical scales. The y-axis of the nested plot, in the right side of the graph, is limited to the maximum estimated incidence (without using the full range 0–1 for the y-axis) in order to provide more detail [50, 53].

If two survival curves (or cumulative incidence plots) cross at any point, such as Med diet,



No. at Risk

Control diet	2450	2268	2020	1583	1268	946
Med diet, EVOO	2543	2486	2320	1987	1687	1310
Med diet, nuts	2454	2343	2093	1657	1389	1031

Fig. 2.19 The plot shows the incidence of the primary end point (a composite of acute myocardial infarction, stroke, and death from cardiovascular causes) for the three groups of the study (Mediterranean Diet with extra-virgin olive oil (EVOO), Mediterranean Diet with Nuts and Control Diet) [57]. EVOO extra-virgin olive oil, Med Mediterranean

nuts and Med diet, EVOO (see the nested scaled graph in the right side of the Fig. 2.19) this might suggest that the hazard ratio between the two groups has reversed and the proportionality assumption (an assumption that is required in Cox analysis) has been violated.

Log-Rank Test

While a Kaplan-Meier plot elegantly represents differences between various groups' survival curves over time, it gives little indication of their statistical significance. The most common method of comparing independent groups of survival times is the log-rank test. This test, however, does not account for confounding variables, such as differences in patient demographics (e.g., age, sex) between groups [58].

Cox Regression Model

If an investigator is interested in quantifying or investigating the effects of known covariates (e.g., age, or race) or predictor variables (e.g., blood pressure), regression models are utilized. A rule of thumb is that Cox models should have a minimum of 10 outcome events per predictor variable.

Compared to the Kaplan–Meier method where only categorical variables can be used to predict the event, with the Cox regression analysis a combination of categorical and/or continuous variables can be used to predict survival. In addition, Cox regression models can also manage censored data.

Among the available survival regression models, the Cox proportional hazards model developed by Sir David Cox is the most commonly used and it is given by the following equation:

$$\ln\left(h\left(t\right)\right) = \ln\left(h_0\left(t\right)\right) + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

or

$$h\left(t\right) = h_0\left(t\right) \cdot e^{\left(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p\right)}$$

where h(t) is the hazard at time t, $h_0(t)$ is an arbitrary baseline hazard (in which we are not interested),

$x_1, \ldots, x_p$ are explanatory variables in the model and $\beta_1, \ldots, \beta_p$ are the corresponding coefficients. The exponential values of the coefficients, $e^{\beta i}$, are the estimated **hazard ratios (HR)**. The hazard ratio is assumed to be constant over time in this model (i.e. the hazards for the groups to be compared are assumed to be *proportional*). It is important to check this assumption either by using graphical methods (e.g., Schoenfeld residuals plot) or specific statistical techniques (e.g., interaction between time and treatment) [48, 49, 59]. For example in ALLHAT trial [40] it is reported that "The Cox proportional hazards regression model assumption was examined by using log-log plots and testing a treatment×time (time-dependent) interaction term."

The Cox regression model was used in the subanalysis of the ACCOMPLISH randomised controlled trial [60] as well. The Fig. 2.20 shows the hazard ratios for different endpoints comparing two treatments groups (benazepril and amlodipine vs benazepril and hydrochlorothiazide) within each of the three BMI categories. In the obese group, the primary or secondary endpoints did not differ between treatment arms ($p > 0.05$). However, in the overweight category, the hazard of primary endpoint was 24% (HR = 0.76; 0.76–1 = −0.24) lower in patients assigned to benazepril and amlodipine than to benazepril and hydrochlorothiazide, adjusted for all the other covariates (age, sex, diabetes, and previous history of cardiovascular events, stroke, or chronic kidney disease) in the model. The 95% confidence interval (0.59–0.94) does not include one, which indicates that the result is significant. As expected, the corresponding p-value = 0.037 is less than 0.05.

Differences between treatment arms were greatest in the normal weight category. Hazard rates for both the composite primary endpoint and total myocardial infarction were lower in patients assigned to benazepril and amlodipine than to benazepril and hydrochlorothiazide ($p < 0.05$).

Intention to Treat Analysis and Per Protocol Analysis

Intention-to-treat (ITT) analysis is a comparison of the treatment groups that includes all subjects as originally allocated after randomization, regardless of whether they completed the trial or even received the treatment after randomization [1, 61]. It serves to protect from biases in RCTs

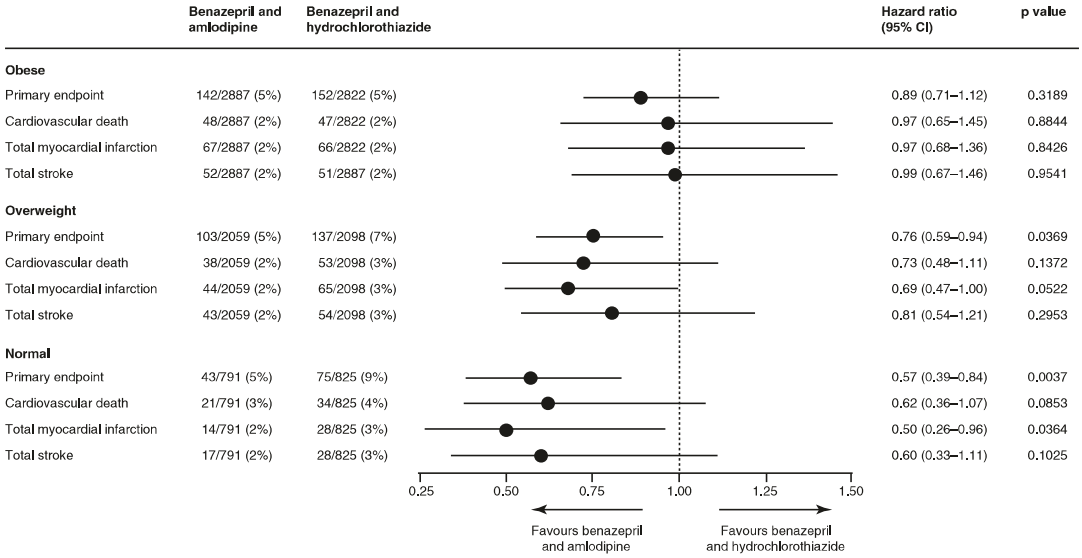


Fig. 2.20 Comparison of hazard rates within obese, overweight, and normal weight categories**.** Hazard ratio was calculated by Cox regression and adjusted for age, sex, diabetes, and previous history of cardiovascular events, stroke, or chronic kidney disease. BMI: body-mass index. (Adapted from Weber et al. [60])

associated with noncompliance and missing outcomes [62–64]. This method is recommended in superiority trials and assumes that if the subjects are randomized adequately then noncompliant subjects will be balanced across all the treatment groups [18].

Per-protocol analysis is a comparison of treatment groups that includes only those participants who completed the treatment originally allocated (fulfill the protocol in the terms of the eligibility, interventions, and outcome assessment) [64]. By focusing only on the fully compliant subjects, one can determine the maximal efficacy of a treatment [18]. If done alone, this analysis leads to bias.

In noninferiority trials, both intention to treat and per-protocol analyses are recommended; both approaches should support noninferiority [64].

Interim Analysis

Many trials recruit participants over a long period of time. Interim analyses of randomized controlled trials involve early looks at the data, usually by an independent data monitoring committee to protect the welfare of subjects [18, 65]. In practice, this can be done by stopping enrollment/treatment as soon as a drug is determined to be harmful (e.g., a large number of serious adverse events in one of the treatment groups), highly beneficial (e.g., large effect size suggests superiority of one treatment over the other and clinical equipoise no longer exists) or have negligible chance of demonstrating efficacy if fully enrolled, given results to date (that is, stopping for futility) [5, 66]. For example ACCOMPLISH [67] and SPRINT [17, 55] trials were stopped early at the recommendation of the data and safety monitoring board because the observed difference between the treatment groups exceeded the boundary of the pre-specified stopping rule.

However, performing multiple statistical examinations of accumulating data without appropriate correction can lead to erroneous results and interpretations. If the accumulated data from a trial are examined at five interim analyses that use a p-value of 0.05, the overall false positive rate is nearer to 19% than to the nominal 5% [5]. Adjustment for multiple analyses in interim analysis can be conducted using group sequential methods. The approaches described by Pocock [68, 69], O'Brien & Fleming [70] and Lan and DeMets [71] are popular implementations of group sequential testing for clinical trials. For example, ACCORD trial [35] quoted that "during the trial, an independent data and safety monitoring committee appointed by the NHLBI monitored the primary outcome (11 times) and total rate of death (7 times) with the use of O'Brien–Fleming boundaries determined by the Lan–DeMets approach. For these two outcomes, P values were adjusted to account for the number, timing, and results of interim analyses".

Subgroup and Sensitivity Analyses

Subgroup Analysis

Patients recruited in a major trial (such as SPRINT trial [17]) are not a homogeneous bunch: demographics (e.g., age, gender, race), their medical history (e.g., previous chronic kidney disease or previous cardiovascular disease), and other baseline characteristics (e.g., systolic blood pressure) may vary. Hence, it is reasonable to undertake subgroup analyses to inspect whether the overall result of the trial appears to apply to all eligible patients, or whether there is evidence that real treatment effects depend on certain baseline features [18, 72].

Subgroup analyses for the SPRINT trial [17] are shown in Fig. 2.21. This kind of figure, called a forest plot [72], is the usual way of documenting the estimated treatment effect within each subgroup (an HR in this case) together with its 95% CI. It shows that the effects of the intervention on the rate of the primary outcome was consistent across the six pre-specified subgroups, all being in the direction of superiority for intensive treatment compared with standard treatment. For reference, the results for all patients, with their inevitably tighter CIs, are shown at the top of Fig. 2.21.

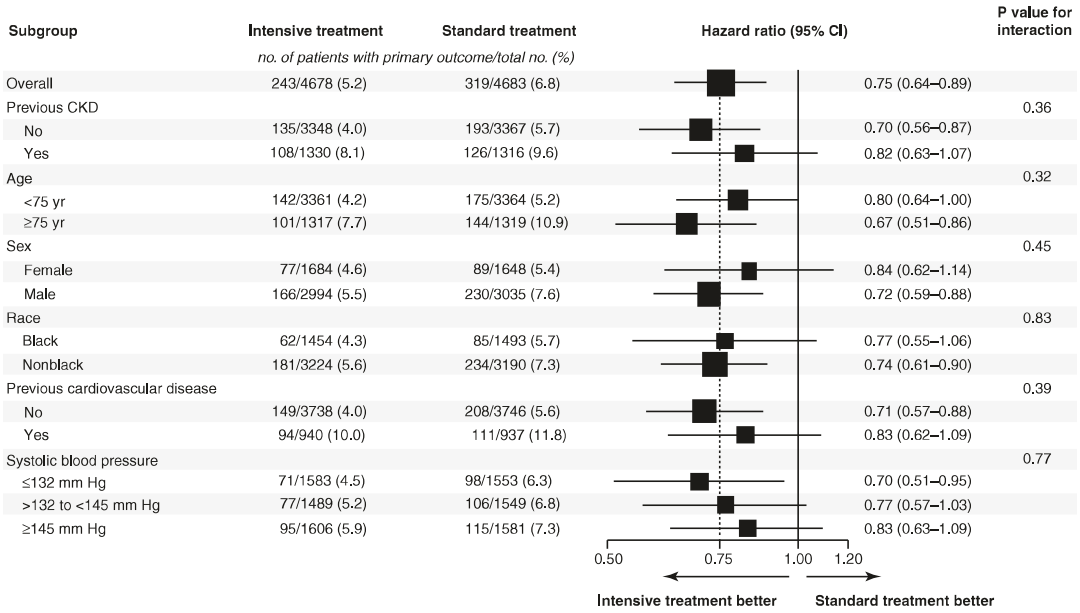


Fig. 2.21 Forest plot of primary outcome according to subgroups. The dashed vertical line represents the hazard ratio for the overall study population. The box sizes are proportional to the precision of the estimates (with larger boxes indicating a greater degree of precision [larger studies]). CKD Chronic kidney disease. (Adapted from SPRINT trial [17])

Scanning across subgroups, one can see that estimated HRs vary by chance and CIs are wider for smaller subgroups which have tinier squared bolb. Some CIs overlap the line of unity, indicating that the subgroup p-value does not reach 5% significance level; this will inevitably happen, especially in smaller subgroups, and is not helpful in interpreting subgroup findings [72]. For example for the subgroup analysis of sex (female/male) in the SPRINT trial, the effect in females was not statistical significant (95%CI: 0.62 to 1.14, $p > 0.05$) but in males (95%CI: 0.59 to 0.88, $p < 0.05$) it was. Comparing p-values for separate analyses of the treatment effect in each group can be misleading [73].

Instead, a statistical test of interaction should accompany each subgroup display as shown in Fig. 2.21. This interaction test examines the extent to which the observed difference in HRs across subgroups may be attributed to chance variation. However, the greater the number of statistical test for interaction performed, the greater the probability of a false-positive finding caused by chance alone (the overall Type I error rate for all subgroup analyses is inflated), so the p-values for interaction test should be adjusted [74]. The SPRINT trial reported that "Interactions between treatment effect and prespecified subgroups were assessed with a likelihood-ratio test for the interaction with the use of Hommel-adjusted p-values" (Hommel adjustment method is a modification of Bonferroni correction [75]). In conclusion, for the SPRINT trial there were no significant interactions between treatment and subgroup with respect to the primary outcome.

Sensitivity Analysis

Sensitivity analysis is an approach to inspect the impact, effect or influence of key assumptions or variations—such as different methods of analysis, different cut-offs or definitions of

outcomes, protocol deviations, different missing data management and manipulation of outliers—on the final interpretations and overall conclusions of a study. In other words, it is a technique to assess the robustness of the findings based on primary analyses of data in clinical trials [76].

For example for the longitudinal analysis of systolic blood pressure in ACCORD trial [35], a sensitivity analysis was presented that compared Maximum Likelihood Repeated Measures Analysis (ML) under the assumption that the missing data were missing at random (MAR) with analysis of observed data under the assumption that the missing values were missing completed at random (MCAR).

Sample Size Calculation and Power

The sample size of a randomized controlled trial (RCT) is the number of subjects needed to detect a clinically relevant treatment effect. Usually, the number of participants in a trial is restricted because of scientific and ethical reasons. However, if the sample size is too small, one may not be able to detect an important existing effect (low power), whereas samples that are unduly large may waste time, research resources, money, and raise ethical considerations. It is therefore important to plan carefully and optimize the sample size of a clinical trial [5, 77].

Elements of the sample size calculation are [5, 18, 77, 78]:

1. The estimated outcomes in each group (which implies the clinically important target difference between the intervention groups, e.g., minimum expected difference of means, $\mu_1 - \mu_2$, or proportions, $p_1 - p_2$)
2. The significance level alpha and whether a one-tailed or two-tailed statistical analysis is planned, usually 5%
3. The desired statistical power $(1 - \beta)$, usually 80% or 90%
4. The estimated measurement variability (e.g., standard deviation) for continuous outcomes
5. The study design (parallel or crossover, etc.)
6. The expected dropout rate of subjects during the study
7. Adjustments for interim or/and subgroup analyses

The power of the study is a measure of how likely it is that the hypothesis test will produce a statistically significant result, for a population effect of a given magnitude, if an effect truly exists. For example, a study power of 80% means that if the study were to be repeated many times, a statistically significant result would be obtained 8 times out of 10, if there truly is an effect of the specified size.

The power needed is usually decided before the start of the study for calculating the sample size. However, it is also possible to work backwards, to estimate the power of a study given a fixed sample size. For example, there may be circumstances where the number of participants who are available or affordable (due to cost or time constraints) is limited. The power of a clinical trial is increased when the level of alpha, the expected difference, or the sample size are increased [79].

Authors should indicate how the sample size was estimated. They should identify the primary endpoint on which the calculation was based, all the elements used in the computation, and report the resulting target sample size [5]. This information is usually provided in statistical methods of the research paper. For example, in PATHWAY-2 trial [7] is reported that "the sample size was estimated to be 294 patients, based on detecting a difference of 3 mm Hg (SD 12) in home systolic blood pressure between each of the experimental drugs and the placebo treatment, with 90% power using a single sample *t* test at the 0.003 significance level (this was chosen in order that the 0.01 level could be adjusted for three planned comparisons)".

Nowadays, the sample size calculation can be conducted with specialized tools such as G*Power or software environment R (Fig. 2.22).

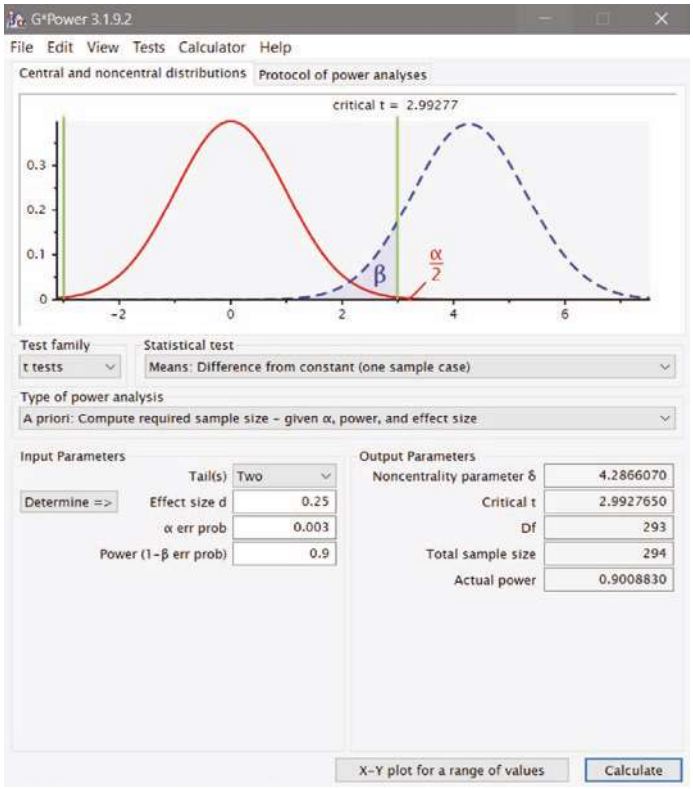


Fig. 2.22 The sample size calculation can be conducted with specialized tools such as G*Power or R programming language

R programming language

```
library(pwr)

pwr.t.test(d=0.25,n=NULL,power=0.9, sig.level=0.003,type="one.sample",alternative="two.sided")


     One-sample t test power calculation

              n = 293.3132
              d = 0.25
      sig.level = 0.003
          power = 0.9
    alternative = two.sided
```

where $d = \frac{|\mu_1 - \mu_2|}{\sigma}$ is the standardized difference or effect size.

References

1. Everitt BS. Medical statistics from A to Z. Cambridge, UK: Cambridge University Press; 2006.
2. Peters TJ, Eachus JI. Achieving equal probability of selection under various random sampling strategies. Paediatr Perinat Epidemiol England. 1995;9:219–24.
3. Martinez-Mesa J, Gonzalez-Chica DA, Duquia RP, Bonamigo RR, Bastos JL. Sampling: how to select participants in my research study? An Bras Dermatol Brazil. 2016;91:326–30.
4. Hopewell S, Dutton S, Yu L-M, Chan A-W, Altman DG. The quality of reports of randomised trials in 2000 and 2006: comparative study of articles indexed in PubMed. BMJ [Internet]. 2010;340. Available from: http://www.bmj.com/content/340/bmj.c723.abstract.
5. Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. Int J Surg England. 2012;10:28–55.
6. Julius S, Weber MA, Kjeldsen SE, McInnes GT, Zanchetti A, Brunner HR, et al. The Valsartan

Antihypertensive Long-Term Use Evaluation (VALUE) trial. Hypertension [Internet]. 2006;48:385–391. Available from: http://hyper.ahajournals.org/content/48/3/385.abstract.
7. Williams B, MacDonald TM, Morant S, Webb DJ, Sever P, McInnes G, et al. Spironolactone versus placebo, bisoprolol, and doxazosin to determine the optimal treatment for drug-resistant hypertension (PATHWAY-2): a randomised, double-blind, crossover trial. Lancet [Internet]. Elsevier; 2017;386:2059–68. Available from: https://doi.org/10.1016/S0140-6736(15)00257-3.
8. Campbell MK, Piaggio G, Elbourne DR, Altman DG. Consort 2010 statement: extension to cluster randomised trials. BMJ. England. 2012;345:e5661.
9. Bosworth HB, Olsen MK, Dudley T, Orr M, Goldstein MK, Datta SK, et al. Patient education and provider decision support to control blood pressure in primary care: a cluster randomized trial. Am Heart J United States. 2009;157:450–6.
10. Wright JTJ, Bakris G, Greene T, Agodoa LY, Appel LJ, Charleston J, et al. Effect of blood pressure lowering and antihypertensive drug class on progression of hypertensive kidney disease: results from the AASK trial. JAMA. United States. 2002;288:2421–31.
11. Vickers AJ. How to randomize. J Soc Integr Oncol Canada. 2006;4:194–8.
12. Suresh KP. An overview of randomization techniques: an unbiased assessment of outcome in clinical research. J Hum Reprod Sci [Internet]. India: Medknow Publications Pvt Ltd; 2011;4:8–11. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3136079/.
13. Beckett NS, Peters R, Fletcher AE, Staessen JA, Liu L, Dumitrascu D, et al. Treatment of hypertension in patients 80 years of age or older. N Engl J Med [Internet]. Massachusetts Medical Society; 2008;358:1887–98. Available from: https://doi.org/10.1056/NEJMoa0801369.
14. Aviva P, Caroline S. Medical statistics at a glance. 3rd ed. Oxford, UK: Wiley Blackwell; 2009.
15. Thomas E. An introduction to medical statistics for health care professionals: describing and presenting data. Musculoskeletal Care England. 2004;2:218–28.
16. Kirkwood BR, Sterne JAC. Essential medical statistics. 2nd ed. Hoboken: Wiley-Blackwell; 2003.
17. SPRINT Research Group. A randomized trial of intensive versus standard blood-pressure control. N Engl J Med [Internet]. Massachusetts Medical Society; 2015;373:2103–16. Available from: https://doi.org/10.1056/NEJMoa1511939.
18. Wang D, Bakhai A. Clinical trials: a practical guide to design, analysis, and reporting. London: Remedica; 2006.
19. Paul S. Clinical endpoint. Encycl Biopharm Stat. 3rd ed. [Internet]. CRC Press; 2012. p. 273–5. Available from: https://doi.org/10.1201/b14674-43.
20. Black HR, Elliott WJ, Grandits G, Grambsch P, Lucente T, White WB, Neaton JD, Grimm Jr. RH, Hansson L, Lacourcière Y, Muller J. Principal results of the controlled onset verapamil investigation of cardiovascular end points (CONVINCE) trial. JAMA [Internet]. 2003;289:2073–82. Available from: https://doi.org/10.1001/jama.289.16.2073.
21. Duquia RP, Bastos JL, Bonamigo RR, González-Chica DA, Martínez-Mesa J. Presenting data in tables and charts. An Bras Dermatol [Internet]. Sociedade Brasileira de Dermatologia; 2014;89:280–5. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4008059/.
22. Rice K, Lumley T. Graphics and statistics for cardiology: comparing categorical and continuous variables. Heart. England. 2016;102:349–55.
23. Black HR, Elliott WJ, Grandits G, Grambsch P, Lucente T, White WB, et al. Principal results of the controlled onset verapamil investigation of cardiovascular end points (CONVINCE) trial. JAMA. United States. 2003;289:2073–82.
24. Krum H, Schlaich MP, Sobotka PA, Böhm M, Mahfoud F, Rocha-Singh K, et al. Percutaneous renal denervation in patients with treatment-resistant hypertension: final 3-year report of the Symplicity HTN-1 study. Lancet [Internet]. Elsevier; 2017;383:622–9. Available from: https://doi.org/10.1016/S0140-6736(13)62192-3.
25. Annesley TM. Bars and pies make better desserts than figures. Clin Chem United States. 2010;56:1394–400.
26. Hink JK, Eustace JK, Wogalter MS. Do grables enable the extraction of quantitative information better than pure graphs or tables? Int J Ind Ergon [Internet]. 1998;22:439–47. Available from: http://www.sciencedirect.com/science/article/pii/S0169814197000176.
27. Kozak M, Hartley J, Wnuk A, Tartanus M. Multiple pie charts: unreadable, inefficient, and overused. J Sch Publ [Internet]. University of Toronto Press; 2015;46:282–9. Available from: https://doi.org/10.3138/jsp.46.3.05.
28. Rosner B. Fundamentals of biostatistics. Boston: Cengage Learning; 2010.
29. Rumsey DJ. Statistics for dummies. Hoboken: Wiley; 2011.
30. Biffi A, CD A, TK B, et al. Association between blood pressure control and risk of recurrent intracerebral hemorrhage. JAMA [Internet]. 2015;314:904–12. Available from: https://doi.org/10.1001/jama.2015.10082.
31. Nuzzo RL. The box plots alternative for visualizing quantitative data. PM&R [Internet]. 2016;8:268–72. Available from: http://www.sciencedirect.com/science/article/pii/S1934148216000678.
32. Thabane L, Akhtar-Danesh N. Guidelines for reporting descriptive statistics in health research. Nurse Res England. 2008;15:72–81.
33. Tight blood pressure control and risk of macrovascular and microvascular complications in type 2 diabetes: UKPDS 38. BMJ [Internet]. 1998;317:703–713. Available from: http://www.bmj.com/content/317/7160/703.abstract.
34. Duran-Cantolla J, Aizpuru F, Montserrat JM, Ballester E, Teran-Santos J, Aguirregomoscorta JI, et al. Continuous positive airway pressure as treatment for systemic hypertension in people with obstructive sleep apnoea: randomised controlled trial. BMJ. England. 2010;341:c5991.

35. Effects of intensive blood-pressure control in type 2 diabetes mellitus. N Engl J Med [Internet]. Massachusetts Medical Society; 2010;362:1575–85. Available from: https://doi.org/10.1056/NEJMoa1001286.
36. Mancia G, Facchetti R, Parati G, Zanchetti A. Effect of long-term antihypertensive treatment on white-coat hypertension. Hypertens (Dallas, Tex 1979) United States. 2014;64:1388–98.
37. Logan M. Biostatistical design and analysis using R: a practical guide. Chichester: Wiley; 2011.
38. Lobo MD, Sobotka PA, Stanton A, Cockcroft JR, Sulke N, Dolan E, et al. Central arteriovenous anastomosis for the treatment of patients with uncontrolled hypertension (the ROX CONTROL HTN study): a randomised controlled trial. Lancet (London, England) England. 2015;385:1634–41.
39. Kirkwood BR, Sterne JAC. Essential medical statistics. Hoboken: Wiley; 2010.
40. Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic: The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). JAMA United States. 2002;288:2981–97.
41. Chernick MR, Friis RH. Introductory biostatistics for the health sciences: modern applications including bootstrap. Hoboken: Wiley; 2003.
42. Dunn OJ, Clark VA. Basic statistics: a primer for the biomedical sciences. Hoboken: Wiley; 2009.
43. Sedgwick P. A comparison of parametric and non-parametric statistical tests. BMJ. England. 2015;350:h2053.
44. du Prel J-B, Rohrig B, Hommel G, Blettner M. Choosing statistical tests: part 12 of a series on evaluation of scientific publications. Dtsch Arztebl Int Germany. 2010;107:343–8.
45. Cao J, Zhang S. Multiple comparison procedures. JAMA. United States. 2014;312:543–4.
46. Kim H-Y. Statistical notes for clinical researchers: post-hoc multiple comparisons. Restor Dent Endod Korea (South). 2015;40:172–6.
47. Write PS. Adjusted P-values for simultaneous inference. Biometrics. 1992;43:1005–13.
48. George B, Seals S, Aban I. Survival analysis and regression models. J Nucl Cardiol [Internet]. 2014;21:686–94. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4111957/.
49. Kleinbaum D, Klein M. Survival analysis: a self-learning text (Statistics for Biology and Health) [Internet]. Springer; 2005. Available from: citeulike-article-id:3504416.
50. May S, McKnight B. Graphics and statistics for cardiology: survival analysis. Heart England. 2017;103:335–40.
51. Altman DG. Practical statistics for medical research. 1st ed. London/New York: Chapman and Hall; 1991.
52. Selvin S. Survival analysis for epidemiologic and medical research [Internet]. Pract Guid Biostat Epidemiol. Cambridge: Cambridge University Press; 2008. Available from: https://www.cambridge.org/core/books/survival-analysis-for-epidemiologic-and-medical-research/021027404E37FCD99D5A9176D9EAB051.
53. Jager KJ, van Dijk PC, Zoccali C, Dekker FW. The analysis of survival data: the Kaplan-Meier method. Kidney Int United States. 2008;74:560–5.
54. Österlund P, Soveri L-M, Isoniemi H, Poussa T, Alanko T, Bono P. Hypertension and overall survival in metastatic colorectal cancer patients treated with bevacizumab-containing chemotherapy. Br J Cancer [Internet]. Nature Publishing Group; 2011;104:599–604. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3049598/.
55. Upadhya B, Rocco M, Lewis CE, Oparil S, Lovato LC, Cushman WC, et al. Effect of intensive blood pressure treatment on heart failure events in the systolic blood pressure reduction intervention trial. Circ Heart Fail United States. 2017;10:e003613.
56. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part I: basic concepts and first analyses. Br J Cancer England. 2003;89:232–8.
57. Estruch R, Ros E, Salas-Salvado J, Covas M-I, Corella D, Aros F, et al. Primary Prevention of Cardiovascular Disease with a Mediterranean Diet Supplemented with Extra-Virgin Olive Oil or Nuts. N Engl J Med [Internet]. 2018;378(25):e34. Available from: https://doi.org/10.1056/NEJMoa1800389.
58. Tolles J, Lewis RJ. Time-to-event analysis. JAMA. United States. 2016;315:1046–7.
59. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival analysis Part III: multivariate data analysis – choosing a model and assessing its adequacy and fit. Br J Cancer [Internet]. Nature Publishing Group; 2003;89:605–11. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2376927/.
60. Weber MA, Jamerson K, Bakris GL, Weir MR, Zappe D, Zhang Y, et al. Effects of body size and hypertension treatments on cardiovascular event rates: subanalysis of the ACCOMPLISH randomised controlled trial. Lancet [Internet]. Elsevier; 2017;381:537–45. Available from: https://doi.org/10.1016/S0140-6736(12)61343-9.
61. Newell DJ. Intention-to-treat analysis: implications for quantitative and qualitative research. Int J Epidemiol England. 1992;21:837–41.
62. Moher D, Hopewell S, Schulz KF. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ [Internet]. 2010;340. Available from: https://doi.org/10.1136/bmj.c869.
63. Lewis JA, Machin D. Intention to treat--who should use ITT? Br J Cancer England. 1993;68:647–50.
64. Shah PB. Intention-to-treat and per-protocol analysis. C Can Med Assoc J [Internet]. Canadian Medical Association; 2011;183:696. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3071397/.
65. Friedman LM, Furberg CD, DeMets DL. Fundamentals of clinical trials. New York: Springer; 2010.
66. Spieth PM, Kubasch AS, Penzlin AI, Illigens BM-W, Barlinn K, Siepmann T. Randomized controlled tri-

als – a matter of design. Neuropsychiatr Dis Treat New Zealand. 2016;12:1341–9.
67. Jamerson K, Weber MA, Bakris GL, Dahlöf B, Pitt B, Shi V, et al. Benazepril plus amlodipine or hydrochlorothiazide for hypertension in high-risk patients. N Engl J Med [Internet]. Massachusetts Medical Society; 2008;359:2417–28. Available from: https://doi.org/10.1056/NEJMoa0806182.
68. POCOCK SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika [Internet]. 1977;64:191–9. Available from: https://doi.org/10.1093/biomet/64.2.191.
69. Pocock SJ. Interim analyses for randomized clinical trials: the group sequential approach. Biometrics [Internet]. [Wiley, International Biometric Society]; 1982;38:153–62. Available from: http://www.jstor.org/stable/2530298.
70. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics [Internet]. [Wiley, International Biometric Society]; 1979;35:549–56. Available from: http://www.jstor.org/stable/2530245.
71. Gordon Lan KK, Demets DL. Discrete sequential boundaries for clinical trials. Biometrika [Internet]. 1983;70:659–63. Available from: https://doi.org/10.1093/biomet/70.3.659.
72. Pocock SJ, McMurray JJV, Collier TJ. Statistical controversies in reporting of clinical trials: part 2 of a 4-part series on statistics for clinical trials. J Am Coll Cardiol United States. 2015;66:2648–62.
73. Matthews JN, Altman DG. Statistics notes. Interaction 2: compare effect sizes not P values. BMJ England. 1996;313:808.
74. Lagakos SW. The challenge of subgroup analyses--reporting without distorting. N Engl J Med United States. 2006;354:1667–9.
75. Hommel G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika [Internet]. 1988;75:383–6. Available from: https://doi.org/10.1093/biomet/75.2.383.
76. Thabane L, Mbuagbaw L, Zhang S, Samaan Z, Marcucci M, Ye C, et al. A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. BMC Med Res Methodol England. 2013;13:92.
77. Noordzij M, Tripepi G, Dekker FW, Zoccali C, Tanck MW, Jager KJ. Sample size calculations: basic principles and common pitfalls. Nephrol Dial Transplant England. 2010;25:1388–93.
78. Eng J. Sample size estimation: how many individuals should be studied? Radiology United States. 2003;227:309–13.
79. Krzywinski M, Altman N. Points of significance: Power and sample size. Nat Meth [Internet]. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2013;10:1139–40. Available from: https://doi.org/10.1038/nmeth.2738.