ELSEVIER

Urologic Oncology: Seminars and Original Investigations 37 (2019) 305−312

UROLOGIC
ONCOLOGY

Seminars article

An investigator's introduction to statistical considerations in clinical trials

Kathryn Winter, M.S.*, Stephanie L. Pugh, Ph.D.

NRG Oncology Statistics and Data Management Center, ACR, Philadelphia, PA

Received 19 May 2018; received in revised form 22 December 2018; accepted 25 December 2018

Abstract

The purpose of this paper is to provide an introduction for investigators to many of the statistical considerations for clinical trials that will aid in their collaborations with statisticians for clinical trial research endeavors and when reading the clinical trials literature. The purpose of this paper is not to turn a physician into a statistician, that takes formal training and education, as well as day in and day out immersion in the statistical design and analysis of clinical trials, hence statistician as a profession.

Successful clinical trials, not to be confused with only positive clinical trials, are ones that are well designed to answer the trial question, well conducted, and appropriately reported and published, regardless of the results. Physicians and statisticians each play integral roles in the realm of clinical trials and successful clinical trials are the result of collaborations between physicians and statisticians from the beginning of an idea through the manuscript publication. © 2018 Published by Elsevier Inc.

Keywords: Statistical considerations; Determining sample size; Analyses

1. Introduction

The goal of this paper is to present some fundamental statistical considerations for clinical trials. "The purpose of a clinical trial is to provide valid and convincing evidence about the effects of medical therapy" [1]. One of the most important statistical considerations is that an investigator that wants to do a clinical trial needs to work with a statistician. The statistician plays an integral role from the beginning of a trial idea/design, through the publication of the trial results. This paper will touch on a few of the many topics under the broad heading of statistics in clinical trials. There are many good courses and text books on this topic. One recommendation for oncology clinical trials is *Clinical Trials in Oncology* by [2].‘

A clinical trial is any research study that prospectively assigns patients to one or more interventions in order to evaluate the effects of those interventions on define health outcomes. The process starts when an investigator has an idea for evaluating a particular intervention. Where can an investigator go to turn that idea into a clinical trial? As part of the National Cancer Institute National Clinical Trials Network, there are 5 Network Groups funded to do some phase I, and primarily phase II and III cancer treatment clinical trials, 4 of which focus on adult cancers (Alliance, ECOG-ACRIN, NRG Oncology, and SWOG) and 1 on pediatric cancers (COG). Getting involved with one or more of these groups is a good way to do clinical trials, especially phase III and larger randomized phase II trials, which are challenging to do at single institutions. Each of those 5 Network Groups has a dedicated Statistics and Data Management Center, which provides statistical expertise throughout the clinical trials process, as well as the infrastructure for all statistical and data management needs for a trial. When an investigator brings forth an idea for a new trial, it gets discussed/vetted through applicable disease, modality, and other committees and then goes through the specific Group's review processes. The responsible Statistics and Data Management Center statistician is an important voice in all of those conversations.

Disclosures—none.

*Corresponding author.

E-mail address: winterk@nrgoncology.org (K. Winter).

https://doi.org/10.1016/j.urolonc.2018.12.025

1.2. Types of trials

Brief descriptions of Phase I, II, and III trials are as follows:

- □ Phase I trials are generally done to determine the maximum tolerated dose for a regimen of interest. One of the most important aspects of designing a phase I trial is determining the specific definition of what will count as a dose-limiting toxicity. While patients on phase I trials can be followed for outcomes such as survival, given the small sample sizes, statistical testing is not done on efficacy outcomes in these trials.
- □ Phase II trials use regimens that have phase I level safety data and further evaluate them for efficacy, as well as expanding the safety profile data, in order to identify regimens that are promising enough to consider for a definitive phase III trial. More recently, due to biases that can occur when comparing new regimens to historical controls, phase II trials have been conducted as randomized phase II trials. Randomized phase II trials are not definitive trials, as they generally have higher type I errors [3], but are done with the goal of determining if there is a sufficient signal in favor of the experimental regimen, as compared to the standard of care, to move to a definitive phase III trial, which is much more resource and cost intensive.
- □ Phase III trials are the definitive trials for evaluating an experimental regimen. These are generally large, multicenter, randomized trials with the goal of changing clinical practice. Two main phase III trial designs are:
  - ○ Superiority—to determine if treatment A is better than treatment B. These are the most common phase III designs.
  - ○ Noninferiority—to determine if treatment A is noninferior to treatment B for a given outcome, where treatment A has other benefits. These designs are more often used in disease areas where the efficacy outcomes are very high and such that new regimens of interest can give a little on an efficacy outcome while focusing on reducing side effects, treatment time, etc.
- □ For example, the NRG Oncology trial RTOG 0415, which is a phase III trial in favorable risk prostate cancer patients looking to replace conventionally fractionated 3D-CRT/IMRT treatment (41 fractions) with a shorter course of hypofractionated 3D-CRT/IMRT treatment (28 fractions). The trial is specifically designed to determine if hypofractionated 3D-CRT/IMRT is noninferior to conventionally fractionated 3D-CRT/IMRT, with respect to disease-free survival. The null hypothesis in noninferiority trials is that the experimental treatment is significantly worse than the standard treatment, by a specified amount, and the alternative is that it is not. The determination of the size of that “specified amount,” often referred to as the non-inferiority margin, is a critical part of designing a noninferiority trial [18].

2. Determining sample size

One important responsibility of the study statistician is to work with the principal investigator of a developing trial to determine the required sample size. Investigators have been overheard saying things like, “I have a clinical trial and can fund fifty patients. That should be enough, right?” or “I did a quick calculation and this trial is going to need 140 patients, right?” Determining a sample size isn’t done by guessing a number. Contrary to popular belief, statisticians do not have magic key boards that spit out sample sizes with the push of a single button (Fig. 1). The sample size for a clinical trial is a collaborative process that requires significant input from the principal investigator, but potential sample size calculations should be done by the statistician. The final decision of which sample size to use for a given study should be based on discussions between the principal investigator, statistician, and other applicable investigators, as described further in this paper. The main components that a statistician needs to determine potential sample sizes are: hypotheses/primary endpoint, control arm data, effect size of interest, and type I and II error rates.

2.1. Hypotheses

The *null hypothesis*, often denoted H_0, is the hypothesis assumed to be true (e.g., no difference between treatment arms). The *alternative hypothesis*, often denoted H_A, is what the investigator is hoping to confirm (e.g., the new treatment (experimental) is better than the standard treatment (control)). Statistical tests are then used to determine if the observed data are not consistent with H_0 and therefore supportive of H_A. It is not practical to include all patients with a particular disease, say intermediate prostate cancer, in a trial to evaluate a new treatment, so *hypothesis testing*

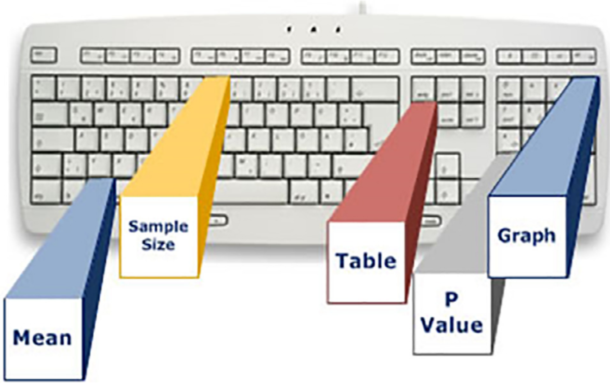

Fig. 1. Keyboards, like the imaginary one pictured here, to do statistical design and analysis at the push of a button don’t exist.

provides a systematic and formal way to help investigators come to a conclusion about a *population* by studying a *sample*. For example, NRG Oncology trial GU002 will accrue a sample of patients to make conclusions about adding chemotherapy to the standard radiation/androgen deprivation regimen in the high risk, postprostatectomy cancer patient population.

2.2. Type I and II errors

There is always an underlying truth for any hypothesis test; however it is not known, which is why clinical trials are conducted. Given this, there are 2 possible errors that can result from a clinical trial, *Type I and Type II Errors*, as well as 2 correction decisions (Table 1). A *type I error* (α), is a false positive rejecting the H_0 when it is true, i.e., concluding a difference in the treatment arms, when the underlying truth is that they are the same. A *type II error* (β), is a false negative not rejecting the H_0 when the H_A is true, i.e., not concluding a difference in the treatment arms, when it does exist. Note that the 2 decisions based on the data in Table 1 are "Reject H_0" and "Fail to Reject H_0." Accepting the null hypothesis is not a correct statistical decision for a hypothesis test. The trial data either provides sufficient evidence that is not consistent with the H_0 (reject the H_0 and conclude the H_A), or it does not (do not reject the H_0 and do not conclude the H_A). Ideally, a clinical trial would keep both errors as small as possible; however, α and β are inversely proportional to each other, so, for a given sample size, as the type II error is decreased, the type I error increases. The general rules for prospectively designed clinical trials are: type II error $\leq$0.20 and type I error $\leq$0.05 [2] for phase III and 0.05 to 0.20 in phase II [3].

2.3. Statistical power

Statistical power for a clinical trial is related to the type II error. If the underlying truth in a given clinical trial is that the H_A is true, then the data either support H_A (correct decision) or do not support H_A (type II error). Statistical power is the probability of a making a correct decision, as depicted in Table 1, and calculated by 1—type II error. As mentioned above, prospectively designed clinical trials should have a type II error $\leq$0.20, which corresponds to a minimum of 80% statistical power. When reading the results of a prospectively designed clinical trial, it is uncommon not to have information on the statistical power. It is important to know the statistical power for retrospective analyses as well, especially when "negative" results are reported. "Negative" retrospective analyses reported without the statistical power available to detect the hypothesized difference, should be interpreted cautiously.

Table 1
Depiction of statistical power and type I and type II errors based on what is true vs. the data.

Decision based on data	Truth	
	H_0 true	H_0 false (H_A true)
Reject H_0	False positive: type I error (α)	Correct action (power)
Fail to reject H_0	Correct action	False negative: type II error (β)

2.4. One-sided and 2-sided hypothesis tests

Hypothesis tests can be either 1-sided or 2-sided. If an investigator is interested in testing 2 treatments A and B to conclude if there is a difference between them in either direction, then a *2-sided* hypothesis test is appropriate. When an investigator is testing an experimental treatment (A) and is interested only in concluding if it is better than the control arm treatment (B), then a *1-sided* hypothesis tests is appropriate. The statistically allowable conclusions, based on a 1-sided or 2-sided test, are shown in Table 2. For a 2-sided test, the α is equally divided between the 2 possible statistically significant differences, such that an α 0.05 test provides 0.025 α for A being better than B and 0.025 α for B being better than A. In a 1-sided test, all of the α goes to the one possible statistically significant differences, so in an α 0.05 test, from the example above, A being better than B has a 0.05 false positive rate. Therefore, a 1-sided test of the same α level as a 2-sided test has a higher false positive rate [4]. Given that, a 1-sided test for definitive phase III trials frequently use an α of 0.025, however it is not uncommon to see a 1-sided test with an alpha of 0.05, especially in trials of more rare tumors.

2.5. Endpoints

The primary endpoint is the main question that powers the study (i.e., provides the basis for sample size estimation). Secondary endpoints are important, sometimes hypothesis-driven, outcomes; while tertiary/exploratory endpoints are hypothesis-generating. For example, a randomized trial was conducted to determine if short-term androgen deprivation therapy (ADT) would improve survival in patients with localized prostate cancer [5]. The primary endpoint was overall survival and secondary endpoints included disease-specific mortality, biochemical failure, distant metastases, adverse events, and erectile dysfunction. The study found that the addition of short-term ADT to radiation improved survival. It also decreased the rates of disease-specific mortality, biochemical failure, and distant metastases. However, the patients receiving short-term ADT experienced hot flashes and higher rates of erectile dysfunction.

Clearly defining endpoints in a clinical trial protocol and subsequent publications is very important, as not all endpoints with the same name are defined the same way across trials. Some are straightforward, for example it is universal that failure for overall survival (OS) is death due to any

Table 2
Possible conclusions based on 1- and 2-sided hypotheses.

Increase in median OS to	Possible conclusions (overall survival example)
2-sided H_0: OS treatment A = OS Treatment B H_A: OS treatment A ≠ OS Treatment B	■ OS is significantly different between treatments A and B, with treatment A having better OS than treatment B. ■ OS is significantly different between treatments A and B, with treatment B having better OS than treatment A. ■ OS is not significantly different between treatments A and B.
1-sided H_0: OS treatment A = OS Treatment B H_A: OS treatment A ≠ OS Treatment B	■ OS for the experimental treatment A is significantly better than OS for treatment B. ■ OS for the experimental treatment A is *not* significantly better than OS for treatment B.

cause. Some people use disease-free survival (DFS) and progression-free survival (PFS) interchangeably, but there is a difference. As a general rule, for DFS it is expected that the treatment will eradicate the disease, whereas for PFS the expectation is that the treatment will keep the disease from getting worse and it is a bonus if it actually gets better. While OS is considered a gold standard for phase III trials, in this day and age of limited resources, and unfortunately low adult clinical trial participation rates, it is even more important to think about the best clinical endpoint for a given patient population. For example, in early or even intermediate stage prostate cancer, where the patients' median age is commonly 70 years, OS may not be the right primary endpoint for finding clinically meaning treatment improvements, as the vast majority of these patients will die of something other than their treated cancer. The goal is to keep the patient from dying of the treated cancer, but there is no treatment that will keep a prostate cancer patient alive forever. The trial conducted by Jones et al required 1,979 eligible patients and the median follow-up at the time of analysis was 9.1 years. Lengthy follow-up and considerably large sample sizes are required in order to detect clinically meaningful differences in OS and those trials are increasingly rare. Prostate cancer specific death may be an ideal endpoint, however, given how few number of those events occur in early/intermediate prostate cancer patients, the required sample size is often much too high to be feasible. Selection of appropriate second endpoints is also important, as they can help provide a more comprehensive picture of the treatment being evaluated. In a recently reported NRG prostate trial [6], which did use OS as the primary endpoint, high dose radiation did not show an improvement in the primary endpoint of OS, but did show a benefit for secondary endpoints of biochemical failure, distant metastases, and rate of salvage therapy; although high dose RT also had more late grade 2+ toxicities.

Using a surrogate endpoint can reduce the sample size and length of trial, thereby decreasing the cost. Surrogate endpoints are essentially an intermediate endpoint; one that occurs during the course of a disease and is known to be a forerunner to the outcome of interest [7]. Royce et al evaluated 4 potential surrogates for all-cause mortality, PSA failure, PSA nadir >0.5 ng/ml, PSA doubling time <9 months, and interval to PSA failure <30 months [8]. The Prentice criteria, a set of conditions required to validate a potential surrogate endpoint, was used to assess each surrogate [9]. PSA nadir >0.5 ng/ml after radiation was selected as the optimal surrogate for all-cause mortality. The study team concluded that using this as an entry criteria would "enhance the likelihood that the study will be able to answer the question of whether survival is prolonged when novel treatment is added to standard of care as compared with standard of care and over a shorter time period." Many researchers are investigating surrogate endpoints that can be used in clinical trials, especially for prostate cancer.

2.6. *Effect size*

Once the primary endpoint is finalized, the effect size needs to be determined. The investigator provides the statistician information about the control arm, generally current standard of care, and how much they think the experimental treatment will improve the endpoint of interest. For example, it may be that the control arm has a survival distribution corresponding to a median OS of 23 months and the investigator thinks that the experimental treatment will increase the survival distribution corresponding to a median OS to 31 months (hazard ratio [HR] = 0.74) or the standard of care has a disease-free survival distribution corresponding to a 5-year DFS of 75% and the investigator thinks that the experimental treatment will increase the DFS distribution corresponding to a 5-year DFS of 83% (HR = 0.65). The difference to be detected and the required sample size are inversely proportional, meaning the smaller the difference, the larger the required sample size.

2.7. *Interim analyses*

As part of the trial design, it needs to be determined if any interim analyses will be included. If a treatment really is extremely effective (efficacy) or will never be positive (futile), it is better to know sooner rather than later; however it is also important, even more so on the futility side, not to erroneously conclude futility too early based on results with very little information on the primary endpoint. Phase III trials often include interim analyses for both efficacy and futility. There are several methods that can be used for both and various philosophies on how conservative or liberal a rule should be used. As interim efficacy analyses as taking a look at the primary hypothesis comparison and at any given look there is a chance for a false positive, each of these looks "spends" some of the overall study alpha (type I error). It is not a one-size fits all, but rather needs to be prospectively determined on a study by study basis as part of the statistical design of the trial, as the method and number of interim analyses impact the required sample size.

2.8. *Time-to-event endpoints*

For time-to-event endpoints, although a sample size is determined, what is really driving that sample size is the number of required primary endpoint events (e.g., number of deaths for OS). If 250 OS events are needed, which is driven by the effect size and type I and II errors, clearly the smallest sample size would be 250 patients, although the study would not be able to report out until all of the entered patients had died. How much higher than the required number of events does the sample size need to be? There is not a fixed answer for that. The accrual rate and how quickly events occur, corresponding to the required follow-up, both play a role in determining the overall sample size. For the same number of required events and being able to report the results in a reasonable timeframe, a trial that can accrue quickly, but has a low failure rate (e.g., early stage prostate cancer) will have a larger sample size than a quickly accruing trial with a high failure rate (e.g., metastatic prostate cancer).

2.9. *Stratification and blinding*

Statisticians are all about reducing bias in clinical trials, in order to minimize any factors that could confound trial's treatment effect interpretation. Two topics along these lines are the use of stratification factors and whether patients/physicians will be blinded to the treatment that the patient will receive. While randomization helps with balancing factors between treatment arms that could impact the trial's endpoints, stratification will ensure balance for factors known to be associated with the primary outcome. There is a common misconception that stratifying by say entry level prostate-specific antigen (PSA) of <10 vs. ≥10 will result in 50% PSA <10 and 50% PSA ≥10 on each treatment arm. Wrong! The actual distribution of the levels of a given stratification factor are not known until the trial has finished accruing. Stratification provides balance between the treatment arms with respect to the levels of the factor. For the PSA example, if the overall distribution of entry level PSA <10 and ≥10 is ∼25% and 75%, then stratifying by that factor provides that each treatment arm will have ∼25% of patients with entry level PSA <10 and ∼75% ≥10. There is not a hard and fast rule for how much stratification is too much, but over stratification can reduce precision. If a tumor marker is going to be used for stratification, there are logistical issues that need to be considered. Will the marker determination be done centrally or at each treating institution? If centrally, how long will it take, as it is important to get the marker assessment done timely.

Blinding in a clinical trial is when the patient and/or physician are not aware of what treatment the patient is receiving. Obviously there is the potential for biases when the patient and/or physician are aware of the treatment that the patient is receiving, which may vary depending on the specific endpoint being assessed. However, it is not always logistically feasible or even realistic to include blinding in all randomized clinical trials. Trials randomizing to with/without radiation or surgery would require "sham" RT or surgery for the without treatment arm. However, in a drug trial in which a patient may be taking a pill, it is more logistically feasible to have identical pills for both the active drug and placebo. Trials where the drug is delivered via IV pose additional logistical challenges. Unblinding rules need to be determined. As a general rule, unblinding prior to the patient failing the primary endpoint should only be done if the information is necessary to treat the patient for an emergent issue. Rules for unblinding after the patient fails the primary endpoint should be consider as well (e.g., for a PFS primary endpoint, will control arm patients that progress be crossed over to the experimental treatment and if so, what will the impact be on a secondary endpoint of OS?). The decision to blind or not is determined on a trial by trial basis, based on various resource, logistical, and sometimes regulatory requirement, factors.

2.10. *Determining a sample size is a process*

Even given all of the information above, the statistician does not just compute a single number and say "here it is." Potential sample sizes are calculated under different scenarios determined by varying parameters such as the effect size, error rates, and accrual rate and length of follow-up for time-to-event endpoints, among others, and are then discussed with the trial's principal investigator. Ideally, a trial should be designed to test the smallest clinically meaningful difference with acceptable error rates; however due to resource limitations, that required sample size is not always feasible. Then it becomes a balancing act between a feasible sample size and a realistic effect size. It does not make any sense, nor is it really ethical, to do a trial that is looking for such a large difference that no one believes would ever occur just to get to a sample that can meet funding constraints. Table 3 shows potential sample sizes when looking

Table 3
Sample sizes and hazard ratios for selected increases in median OS.

Increase in median OS to	Hazard ratio	Required evaluable sample size
37	0.62	256
34.5	0.67	321
33	0.70	381
31	0.74	513

for an increase in OS, assuming an exponential distribution where the control arm median OS is 23 months, a 2-sided alpha of 0.05 and 85% power. In the discussions between the investigator and statistician, perhaps 513 patients really is not feasible and an OS increase corresponding to a median OS of 37 months is overly optimistic. That would eliminate the top and bottom row options and then further discussions would determine if whether to go with a sample size of 321 or 381, or even to discuss further options.

3. Analyses

3.1. *What is a* P *value?*

The answer is not "a number that needs to be <0.05 so the trial can be published in a good journal." A *P* value is simply the probability of obtaining a result from the data (which comes from a sample) that is equal to or more extreme than the observed result, *given that the H_0 is true*. Generally a small *P* value indicates the observed results are not likely to have occurred if the H_0 is true, so either the H_0 is not true, or a type I error has occurred. In hypothesis testing a significance level, also known as type I error and denoted α, is set a priori. The *P* value from the data is compared such that if the data *P* value $\leq\alpha$, the H_0 is rejected. A *P* value only indicates how likely the result would be if the H_0 is true, it gives absolutely no information about the magnitude of the observed difference or about the number of patients from which the data came. Consider the example in Table 4, which shows the results of 4 example trials that all had overall survival distributions corresponding to 40% and 42% 2-year overall survival for the control and experimental arms, respectively.

Any difference, however small, can be shown to be "statistically significant" with enough patients, but statistically significant does not always equal clinically meaningful. This is why statisticians design clinical trials with the appropriate number of patients to answer a clinically meaningful hypothesis of interest. This also illustrates why the *P* value should never be reported alone. In a time-to-event scenario comparing 2 treatments, it should always be accompanied by the effect size (HR), estimates at meaningful time points (e.g., 2 and/or 5-year OS) with relevant confidence intervals, and often graphs showing the full distributions.

Table 4
Impact of sample size on p-value for the same treatment difference.

Trial	*P* value	#Patients in trial
A	0.77	200
B	0.39	2,000
C	0.16	5,000
D	0.004	20,000

3.2. *Interpreting* P *values*

While the official statistical conclusion of a hypothesis test is based on the *P* value, please remember, that the *P* value alone does tell the whole story for the analysis. That message understood, for a result to be statistically significant, the *P* value from the test based on the collected data must be ≤the apriori-specified α, with the key word there being apriori. It is completely inappropriate and downright wrong to bump up the apriori α because the resulting *P* value was a little higher than the original α and equally as wrong to collect the data, do the hypothesis test and then choose an α. Consider a trial designed with a hypothesis test for an OS endpoint with an overall α of 0.05. If the results of the trial produce a *P* value of 0.036, then statistical significance has been met. If the results produce a *P* value of 0.062, then statistical significance has not been met. There is no such thing as "almost statistically significant." Much like a woman either is or is not pregnant, hypothesis test results either are or are not statistically significant. However, remember the important fact that the *P* value alone does not tell the whole story. From the nonstatistically significant example above, a *P* value of 0.062 or 0.29 or 0.75 would all result in statistical significance not being met; but, the additional important information (HR, confidence intervals, etc.) would be very different. Just as statistically significant does not always equal clinically meaningful; nor does *not* statistically significant equal clinically meaningless.

3.3. *Censoring*

For time-to-event end points for a given patient either (1) the event has occurred and the time of that occurrence is known or (2) the event has not occurred up to the time of the last available information. The latter scenario is called a censored observation. For censored patients, it is unknown if the event would have occurred, given additional observation time. One reason for having censored observations is due to not following patients forever on a study. Another reason is due to patients being lost-to-follow-up on a study. Take overall survival, for example. When a patient dies, the event of interest has occurred and the date of death is recorded. Patients who have not died are considered to be censored at their date of last follow-up. Consider the following 3 patients, one that died at 9 months and two that are censored. The patient that died at 9 months has complete information for overall survival. The first censored patient entered the study 2 years ago and is known to be alive at 2 years. While that patient is censored, that is as complete information as can be gotten, until their next follow-up is due. The second censored patient also entered the study 2 years ago, but the last follow-up information available is at 6 months. That patient is also censored, but there is information that should be had that is missing. It could be that the patient is alive at 2 years or the patient could have died sometime after 6 months.

The most common approach for analyzing survival data is the Kaplan Meier method [10]. All patients contribute to the reporting of overall survival up to the point that they are an event or are censored. When specific time points are reported from a Kaplan-Meier curve, it is important to know how much censoring there is prior to that time point. If there is a lot of censoring (“a lot” is relative to the study sample size) prior to a given time point, the overall survival rate may change based on how many of those censored patients end up having an event by that time point, based on additional follow-up information. Sometimes survival curves will indicate censored patients with tick marks on the curve at the time the patient is censored.

3.4. Intent-to-treat vs. as treated analyses

For randomized trials, intent-to-treat analysis means that the patients are analyzed based on the arm to which they were randomized, regardless of whether they received that treatment or not, which preserves balance of known and unknown factors that may affect outcomes, as afforded by randomization. An as-treated analysis means that patients are analyzed according to the treatment they received. Intent-to-treat is the standard analysis for randomized clinical trials, as it protects against potential biases, and should be the primary analysis used. However, it may be applicable to do additional as-treated analyses in some situations [11].

3.5. Missing data

A common, yet unfortunate, issue in analysis of longitudinal data is when one or more of the sequences of measurements are incomplete. Missing data is often found in quality of life data, such as the patient completed Expanded Prostate Cancer Index Composite questionnaire, and laboratory values, such as PSA. The main problem that arises with missing data is that the distribution of the observed data may not be the same as the distribution of the complete data. How this is handled in the analysis depends on the type of missingness: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).

Examples of MCAR data: A patient may miss a clinical visit due to a funeral or a QOL questionnaire may not have been completed because the printer broke at the clinic [12]. MCAR occurs when missing values are randomly distributed across all observations. No adjustments to the analysis are needed. In MAR data, missingness depends on the observed data and, when given the observed data, but it does not depend on the unobserved data [12,13]. For example, a patient without a car may take the bus to his clinic visit. He may miss a visit if he misses his bus by even a minute. Here, the missing data depends on a covariate, say socioeconomic class. Accounting for this covariate in analyses can make the missing data ignorable. . MNAR data is considered nonignorable since missingness depends on the unobserved data and may also depend on the observed data [12,13]. There are several available methods to handle MNAR data. A common type of missing data seen in clinical trials is drop out, where a patient is observed from baseline up until a certain point in time with no more measurements made. Analysis of MNAR data must take the non-ignorable missing into account. Further details on analysis of missing can be found in [14−17].

3.6. Timing of analyses

On the day a trial completes accrual, investigators, understandable excited about that milestone, sometimes say “Great! How soon can we report?” Protocol interim and final analyses of the primary endpoint and of secondary endpoints are done at the times specified in the protocol, which are generally event or time driven. If a trial is designed to accrue for 3 years and requires 2 years of follow-up to reach the required number of primary endpoint events, and it accrued as planned, the answer to the above question is in about 2 years. Interim analyses are reported to the trial’s data monitoring committee and unless the trial results are released early by the data monitoring committee, the trial’s principal investigator does not see the results of interim analyses. When it is time to report out the trial’s results, the statistician provides an analysis report to the investigator, as based on the prospective statistical analysis plan. For example: information on accrual, patient and tumor characteristics, treatment delivery, adverse events, and the protocol endpoint specific analyses, along with the statistical interpretations of the results. Which endpoints are included is determined by the protocol-specified timing of the trial’s endpoints. For example, if the primary endpoint is complete response at 3 months post RT completion and there are secondary endpoints for overall survival and local failure after patients have been followed for 3 years, the initial analysis report would include the CR endpoint, but not the OS and local failure endpoints. After the analysis report is sent to the investigator, there is continued collaboration with the statistician in order to present and publish the results. It is an important responsibility of both the principal investigator and the statistician that the results of clinical trials are published, regardless of what those results are.

4. Summary

In summary, this paper has covered some of the fundamental statistical considerations for a clinical trial. Hopefully it provides a basis for future interactions with statisticians and an understanding of the integral role of the statistician as a coinvestigator from the beginning of a clinical trial idea, through study design, data collection, analyses, and the interpretation and publication of the results. Conversely, statisticians learn a lot about the disease/topic of the trial from the principal investigator, which aids in designing and analyzing trials. Neither the statistician nor the principal investigator is in it or can truly do it alone, it needs to be a true collaboration. Communication is key.

References

[1] Piantadosi S. Principals of clinical trial design. Semin Oncol 1988;15:423–33.

[2] Green S, Benedetti J, Cowley C. Clinical trials in oncology. 3rd ed. Chapman & Hall/CRC; 2012.

[3] Rubenstein LV, Korn EL, Freidlin B, et al. Design issues of randomized phase II trials and a proposal for phase II screening trials. J Clin Oncol 2005;23:7199–206.

[4] Ellenberg S. Biostatistics in clinical trials: Part 2 Determining sample sizes for clinical trials. Oncology 1989;3:39–46.

[5] Jones CU, Hunt D, McGowan DG, Amin MB, Chetner MP, Bruner DW, et al. Radiotherapy and short-term androgen deprivation for localized prostate cancer. N Engl J Med 2011;365:107–18.

[6] Michalski JM, Moughan J, Purdy J, Bosch W, Bruner DW, Bahary JP, et al. Effect of standard vs dose-escalated radiation therapy for patients with intermediate-risk prostate cancer: the NRG oncology RTOG 0126 randomized clinical trial. *JAMA Oncol* 2018:epub ahead of print.

[7] Ellenberg SS. Surrogate endpoints. *Br J Cancer* 1993;68:547-459.

[8] Royce TJ, Chen MH, Wu J, Loffredo M, Renshaw AA, Kantoff RW, et al. Surrogate end points for all-cause mortality in men with localized unfavorable-risk prostate cancer treated with radiation therapy vs radiation therapy plus androgen deprivation therapy: a secondary analysis of a randomized clinical trial. JAMA Oncol 2017;3:652–8.

[9] Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. Stat Med 1989;8:431–40.

[10] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Stat Assoc 1958;54:457–81.

[11] Ellenberg J. Intent-to-Treat Analysis versus As-Treated Analysis. Drug Information Journal 1996;30:535–44.

[12] Rubin DB. Inference and missing data. Biometrika 1976;63:581–92.

[13] Ibrahim JG, Molenberghs G. Missing data methods in longitudinal studies: a review. Test (Madr) 2009;18:1–43.

[14] Gibbons RD, Hedeker D, DuToit S. Advances in analysis of longitudinal data. Annu Rev Clin Psychol 2010;6:79–107.

[15] Fairclough D. Design and analysis of quality of life studies in clinical trials. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC; 2010.

[16] Schafer JL. Multiple imputation: a primer. Stat Meth Med Res 1999;8:3–15.

[17] Rizopoulos D. Joint models for longitudinal and time-to-event data with applications in R. 1st ed. Boca Raton FL: CRC Press; 2012.

[18] Lee WR, Dignam JJ, Amin MG, Bruner DW, Low D, Swanson GP, et al. Randomized phase III noninferiority study comparing two radiotherapy fractionation schedules in patients with low-risk prostate cancer. J Clin Oncol 2016;34:2325–32.