PASI-2011

Cancer Survival Trial Design

John(Jianrong) Wu
St Jude Children's Research Hospital
August 19-20, 2011

Introduction of survival analysis and cancer clinical trial design

Part I: Introduction of cancer clinical trial design

Cancer clinical trial is a planned experiment for cancer patients based on limited sample of patients to make inferences about how treatment should be conducted in the general population of patients who will require treatment in the future.

Traditional classification of cancer clinical trials:

- Phase I trial is often a drug safety study through drug dose-escalation to determine (estimate) the maximum tolerated dose (MTD) based on observed dose limiting toxicities (DLT). Common designs of phase I trial: traditional 3+3 design or a model based dose escalation design (CRM, EWOC).

- Phase II trial is often a small-scale and single-arm (non-randomized) study to determine the safety and efficacy of the drug (or drug combination treatment) and to see weather the new treatment has sufficient antitumor activity to warrant a further large-scale randomized study.
- Phase III trial is often a large-scale randomized study to determine the efficacy of a new treatment (experimental arm) compared with the best current standard therapy (control arm) where patients' allocation is based on a randomization procedure.

There are many aspects of the design of a clinical trial that must be considered. These include a good idea for improving treatment, study objectives, specific hypotheses, treatment plan, patient's eligibility criteria, method of randomization and blinding, sample size calculation and statistical analysis plan (ASP), protocol development, protocol review (PRC) and approval (IRB, FDA),

database issue, protocol coordinate and trial monitoring (DSMB) for safety and ethic consideration.

The major elements in designing a clinical trial

- Define the study objectives: primary and secondary objectives
- Specify the eligibility, treatments, and endpoints
- Determining the magnitude of difference (effect size) to be detected
- Specify how treatment assignment will be accomplished (randomization)
- Examine the historical data to identify distribution assumptions used for sample size calculation

A good clinical trial minimizes variability of the evaluation and provides an unbiased evaluation of the treatment by avoiding confounding from other factors. Randomization insures that each patient have an equal chance of receiving any of the treatments under study, generate comparable treatment groups which are alike in all important aspects except for the treatment each group receives.

- Randomization and stratification

In general, a randomized trial is an essential tool for testing the efficacy of the treatment. A simple (complete) randomization (tossing a coin, tables of random numbers) is completely unpredictable however it is not quite sufficient by itself to guarantee comparable treatment arms unless the sample size is large. In small or moderate size studies, major imbalances in important patient characteristics can occur by chance. Patient characteristics incorporated into the randomization scheme to achieve balance are

called stratification factors. Stratification factors should be those know to be strongly associated with outcome (eg, tumor response, survival). In general, we suggest no more than three stratification factors used in cancer clinical trials. Stratified randomization is achieved by performing a separate randomization procedure within each strata. For example, in a study, age and gender are the two most important risk factors. To make sure the balance of age and gender between two treatment groups, the study is to be stratified on age (< 10 vs ≥ 10) and gender (Male vs Female) (two stratification factors), a simple (or blocked) randomization would be done within each of the four defined patients strata:

- age$<$ 10 and M
- age$<$ 10 and F
- age $\geq$ 10 and M
- age $\geq$ 10 and F

Stratified randomization not only ensures that the treatment numbers in each treatment groups are closely balanced within each stratum but also ensures that treatment groups are similar with respect to the important prognostic factors.

The basic benefits of randomization include

- Eliminates selection bias.
- Balances arms with respect to prognostic variables (known and unknown).
- Forms basis for statistical tests.

Some issues of randomization

- Is randomization feasible? ethical issue
- Agree to randomization? (both clinician and patient)

- Blinding

Human behavior is influenced by what we know or believe. In research there is a particular risk of expectation influencing findings and leading to biased results. Blinding is used to try to eliminate such bias. For example, medical staff caring for patients in a randomized trial should be blinded to treatment allocation to minimize possible bias in patient management and in assessing disease status. Only patient blinded trial is called single blinding. Neither the patient nor the clinician are aware of the treatment assignment is called double blinded trial. Sometime to avoid bias from data analysis, statisticians are blinded from the outcome too.

- Endpoints

In cancer clinical trial, tumor response (CR + PR) and survival are often the two major primary endpoints, where survival including overall survival, progression-free survival or event-free survival.

- Sample Size/Power calculation

The determination of the sample size of a survival trial depends on the following factors

- An estimate for the endpoint of interest in the control group (based on historical data, for example, 5 survival probability).
- A clinically meaningful minimal effect size
- Type I error α (false positive rate; reject null hypothesis when null is true) and
- type II error β (false negative rate; accept null hypothesis when null is not true). $1 - \beta$ is the power (reject null hypothesis when null is not true).
- Accrual rate and duration of follow up.

Part II: Introduction of survival data analysis

In most cancer clinical trials, the primary outcome is patient survival. There is an accrual period to accrue patients into study. Then patients are randomized to two (or more) groups for treatment. There is some additional follow-up time prior to analysis of the data. At the final analysis, some patients will have died, while some patients will remain alive. For those patients who remain alive, the total time of observation will vary, depending upon when in the accrual period they were registered to the trial. The actual survival time for these patients is unknown.

For more precise, survival time is defined as a time interval between the time origin (date of patient registered on study or date of randomization) and an interested event (disease progression, death) or date end of study (or date of final analysis).

Survival data from clinical trials are often subjected to the right censoring, in which the trial ends before the event of interest is observed in the study. For example, patient without disease progression or still alive at end of study or withdraw from study. Therefore, what can be observed is not the true survival time (T) but the minimum of true survival time and potential censoring time C, as well as an indication of whether the observed time was a true survival time (T) or a censored observation (C). That is the observed survival time $T^0 = min(T, C)$ and the censoring indicate $\Delta = I(T < C)$. Here we assume that T and C are independent which is often called independent censoring model.

Survival data are generally described in terms of two related probabilities: survival and hazard. The survival probability (function) $S(t) = P(T > t)$ is the probability that a patient survives from the time origin to a specified future time t. The hazard function $h(t)$ is defined as

$$h(t) = \lim_{\Delta t \to 0^+} P(t \leq T < t + \Delta t | T \geq t)/\Delta t$$

which specifies the instantaneous rate (risk or hazard) at which death occurs for patient who is surviving at time t. $h(t)\Delta t$ is the approximate probability that an individual dies in the interval $(t, t + \Delta t)$, conditional on that patient having survived to time t (Cox and Oakes, 1984).

The relationship between survival function and hazard is given as follow

$$S(t) = e^{-\int_0^t h(t)dt}$$

The simplest survival distribution is exponential survival function with a constant hazard $h(t) = \lambda$ and survival function

$$S(t) = e^{-\lambda t}, \quad t > 0$$

The survival probability can be estimated nonparametrically from observed survival times using Kaplan and Merier (KM) method. Suppose that k patients have events in the period of follow-up at distinct times $t_1 < t_2 < \cdots < t_k$ in a sample of size n with unknown survival function S. Suppose that d_j events at t_j and m_j are censored in the interval $[t_j, t_{j+1})$ at times $t_{j1}, ..., t_{jm_j}, j = 0, ..., k$, where $t_0 = 0$ and $t_{k+1} = \infty$. Let $n_j = (m_j + d_j) + \cdots + (m_k + d_k)$ be the number of individuals at risk at a time just prior to t_j. The probability of event at t_j is

$$P(T = t_j) = S(t_j^-) - S(t_j)$$

The contribution to the likelihood of a censored survival time at t_{jl}

is

$$P(T > t_{jl}) = S(t_{jl})$$

Therefore the likelihood of data is given

$$L = \prod_{j=0}^{k} \left\{ [S(t_j^-) - S(t_j)]^{d_j} \prod_{l=1}^{m_j} S(t_{jl}) \right\}$$

The maximum likelihood estimate (MLE) is defined as (K and Prentice, 2003)

$$\hat{S} = \text{argmax}_S L.$$

This MLE is given as the Kaplan-Meier estimator of the survival function of T with a form of

$$\hat{S}(t) = \prod_{t_j \leq t} (\frac{n_j - d_j}{n_j})$$

Following recursive formula is useful to calculate the probability of

being alive at time t_j

$$\hat{S}(t_j) = \hat{S}(t_{j-1})(1 - \frac{d_j}{n_j})$$

where $t_0 = 0$ and $\hat{S}(0) = 1$). The value of $\hat{S}(t)$ is constant between times of events, and therefore it is a step function that changes value only at the time of each event.

Note: If the largest observation is a censored survival time, say t^*, then $\hat{S}(t)$ is undefined for $t > t^*$. If the largest observed survival time t_K is an uncensored observation, then $\hat{S}(t)$ is zero for $t > t_K$.

The variance of $\hat{S}(t)$ is estimated using Greenwoods formula

$$\hat{var}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}$$

For large sample sizes $\hat{S}(t)$ has an approximately normal distribution. Therefore, a simple way to obtain the $100(1 - \alpha)\%$

confidence interval for the survival probability $\hat{S}(t)$ is

$$\hat{S}(t) \pm z_{1-\alpha/2}\sqrt{\hat{var}(\hat{S}(t))}$$

where z_α is the α quantile of the standard normal distribution.

A better interval can be obtained by using a log-log transformation

$$\hat{S}(t)^{exp\{\pm z_{1-\alpha/2}\sqrt{\hat{var}(log[-log(\hat{S}(t))])}\}}$$

where $\hat{var}(log[-log(\hat{S}(t))])$ can be derived by delta method.

The hazard function in the interval $t_j \leq t < t_{j+1}$ can be estimated by

$$\hat{h}(t) = \frac{d_j}{n_j \tau_j}, \quad t_j \leq t < t_{j+1}$$

where $\tau_j = t_{j+1} - t_j$.

A smoothed estimate of hazard function can be obtained using kernel smooth function.

An example

Table 1: Survival time (months) on an cancer trial

1	2	4*	6	6	7*	9	11	15*	16
17	18*	24	24*	25*	26	28	31*	32*	35*

Table 2: KM estimate of the cancer survival data

time interval	n_j	d_j	c_j	$(n_j - d_j)/n_j$	$\hat{S}(t)$
$[0, 1)$	20	0	0	(20-0)/20	1.00
$[1, 2)$	20	1	0	(20-1)/20	0.95
$[2, 6)$	19	1	1	(19-1)/19	0.90
$[6, 9)$	17	2	1	(17-2)/17	0.79
$[9, 11)$	14	1	0	(14-1)/14	0.74
$[11, 16)$	13	1	1	(13-1)/13	0.68
$[16, 17)$	11	1	0	(11-1)/11	0.62
$[17, 24)$	10	1	1	(10-1)/10	0.56
$[24, 26)$	8	1	2	(8-1)/8	0.49
$[26, 28)$	5	1	0	(5-1)/5	0.39
$[28, 35)$	4	1	3	(4-1)/4	0.29

The median survival time can be estimated from K-M survival function

$$\hat{m} = \min\{t_j, \hat{S}(t_j) < 0.5\}$$

If the survival function $\hat{S}(t) = 0.5$ at an interval $[t_j, t_{j+1})$, then the K-M median estimate is $\hat{m} = (t_j + t_{j+1})/2$.

It has been shown that the asymptotic variance of $\hat{m}$ can be estimated by (Reid, 1981)

$$\hat{var}(\hat{m}) = \frac{1}{\hat{f}^2(\hat{m})} \hat{var}\{\hat{S}(\hat{m})\},$$

where the $\hat{f}$ is an estimate of density function f and $\hat{var}\{\hat{S}(\hat{m})\}$ is given by Greenwood's formula at $t = \hat{m}$. To use this asymptotic variance formula, we have to estimate the density function f. A common type of density estimation is a window estimate using a kernel function. For example, Kosorok (1999) proposed an optimal

window estimate based on the kernel function

$$\hat{f}(\hat{m}) = \int n^{1/5} \hat{Q}^{-1} K\left(\frac{\hat{m} - x}{n^{-1/5}\hat{Q}}\right) d\hat{F}(x),$$

where $\hat{F} = 1 - \hat{S}$, $\hat{Q}$ is twice the estimated interquartile range of F, and the kernel $K(\cdot)$ is triangular function on $[-1, 1]$.

$$K(x) = \begin{cases} x + 1 & -1 \leq x \leq 0 \\ 1 - x & 0 < x \leq 1 \\ 0 & |x| > 1. \end{cases}$$

In a cancer survival trial, it is often interested in comparison of survival functions between two treatment groups. Assume survival function S_i, density f_i and median m_i for two treatment groups $i = 1, 2$ and let $\hat{S}_i(t)$ be the Kaplan-Meier survival function estimate and $\hat{m}_i$ be the median estimate. We are interested in difference in medians (treatment effect is quantified by difference in medians)

$\tau = m_2 - m_1$. An estimate of τ is $\hat{\tau} = \hat{m}_2 - \hat{m}_1$, and a $100(1-\alpha)\%$ large sample asymptotic confidence interval of τ can be obtained by

$$\hat{\tau} \pm z_{1-\alpha/2}\hat{se}(\hat{\tau}),$$

where

$$\hat{se}(\hat{\tau}) = \{\hat{var}(\hat{m}_1) + \hat{var}(\hat{m}_2)\}^{1/2}.$$

Here the variances $\hat{var}(\hat{m}_1)$ and $\hat{var}(\hat{m}_2)$ can be estimated by a kernel density estimation or bootstrap method (discussion on this topic can be found from Wu, 2012).

Significant treatment effect can also be detected by the Log-rank test which is a non-parametric procedure to test hypothesis of equality of two or more survivor functions. For example, the hypothesis of interest in a trial is

$$H_0 : S_1(t) = S_2(t)$$

$$H_a : S_1(t) \neq S_2(t)$$

Assume that the unique, ordered failure times for two groups are denoted by $t_1 < t_2 < \cdots < t_k$. Let d_{1j} be the number of failures and n_{1j} the numbers at risk in group 1 at time t_j. Let d_{2j} and n_{2j} be the corresponding numbers of group 2. Then $d_j = d_{1j} + d_{2j}$ represents the number of failures in both groups at time t_j and $n_j = n_{1j} + n_{2j}$ is the numbers at risk in both groups at time t_j. The situation is summarized in following table.

Table 3: Number of failures at the j^{th} failure time in each of two groups

Group	Number of failures at t_j	Number of survivors beyond t_j	Number at risk just before t_j
I	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
II	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}

The conditional distribution for d_{1j}, d_{2j} give d_j is a hypergeometric distribution

$$\binom{n_{1j}}{d_{1j}}\binom{n_{2j}}{d_{2j}}\binom{n_j}{d_j}^{-1}$$

with mean and variance of d_{1j} are

$$e_{1j} = n_{1j}d_j/n_j$$

and

$$v_{1j} = \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}$$

The log-rank test is based on the difference between the observed d_{1j} and the expected e_{1j} of the form

$$d_L = \sum_{j=1}^{k} (d_{1j} - e_{1j}).$$

The variance can be estimated as

$$\hat{var}(d_L) = \sum_{j=1}^{k} \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}$$

The log-rank test is then given by the statistics

$$S = \frac{d_L^2}{\hat{var}(d_L)}$$

Which is known to be χ^2 distribution with 1 degree of freedom.

The log-rank test is most powerful when the hazard functions in two treatment groups are proportional to each other, but it can be less efficient then other tests (such as Wilcoxon test or weighted log-rank test) when the proportionality assumption is violated.

Phase III Cancer Survival Trial Design

Phase III cancer trials compare two (or more) treatments for a particular kind of cancer. Typically an experimental treatment is compared to a standard treatment. The primary objective is to see if the new treatment produces better survival.

It is widely acknowledged that the most appropriate way to compare treatments is through a randomized clinical trial. This randomization guarantees that there is no systematic selection bias in treatment allocation.

Therefore, Phase III trials are often conducted as

- Multiple arms (compare among treatment groups)
- Randomized (eliminate potential bias)
- Confirmatory (with power $\geq 90\%$)

- Multi-center study (large study)

Designs for Phase III trial

- Endpoint: Survival
- Hypothesis: Survival rate in treatment arm exceeds standard arm.
- Assign patients to treatment arms by randomization
- Often have a interim analysis plan
- Trial monitored by external DSMB

In this talk, I will focus on sample size calculation of phase III cancer survival trial. As in standard design, the power depends on

- The Type I error (significance level α)
- The difference of interest or effect size (hazard ratio).

The power of survival study depends on the number of events

(deaths), not the total number of patients.

In practice, designing a survival study involves deciding how many patients or individuals to enter, as well as how long they should be followed to insure enough number of events.

Suppose that in a study, two groups (standard vs new treatment) of patients to compare. Assuming a proportional hazards model for the survival times,

$$h_2(t) = \gamma h_1(t)$$

where $h_2(t)$ and $h_1(t)$ are the hazard functions for patients on new treatment and standard groups, respectively, γ is the unknown hazard ratio, which is equivalent to

$$S_2(t) = [S_1(t)]^\gamma$$

where $S_2(t)$ and $S_1(t)$ are the corresponding survival functions.

Define $\theta = log(\gamma)$ to be the log-hazard ratio. If $\theta = 0$, there is no

treatment difference. A negative value of θ indicate that survival is longer under new treatment. Testing $\theta = 0$ is equivalent to test two survival distribution same. Therefore two-sample log-rank test can be used.

Assume that the unique, ordered failure times for pooling the two groups are denoted by $t_1 < t_2 < \cdots < t_k$. Let d_{1j} be the number of failures and n_{1j} the numbers at risk in group 1 at time t_j. Let d_{2j} and n_{2j} be the corresponding numbers of group 2. Then $d_j = d_{1j} + d_{2j}$ and $n_j = n_{1j} + n_{2j}$ represent the number of failures and the number of at risk in both groups at time t_j.

The log-rank test is based on the difference between the observed d_{1j} and the expected e_{1j} of the form

$$U = \sum_{j=1}^{k} (d_{1j} - e_{1j}) = O - E.$$

where $e_{1j} = n_{1j} d_j / n_j$ is expected number of failures for group 1 at

t_j, O and E are the observed and expected number of failures.

The variance can be estimated of the log-rank statistics is

$$V = \sum_{j=1}^{k} \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}$$

Sellke and Siegmund (1983) showed that U, has an approximate normal distribution with mean θV and variance V.

Consider a two-sided hypothesis

$$H_0 : \theta = 0 \quad H_a : \theta \neq 0$$

To calculate the power, we have to calculate two probabilities

$$\alpha = P(|U| > c|H_0)$$
$$1 - \beta = P(|U| > c|H_a)$$

where $c > 0$ is a constant to be determined. Under null $\theta = 0$, then

$$\alpha = P(|U| > c|H_0) = 2P(U > c|\theta = 0) = 2\{1 - \Phi(\frac{c}{\sqrt{V}})\}$$

Therefore

$$c = z_{1-\alpha/2}\sqrt{V}$$

Under alternative $H_a : \theta = \theta_R < 0$, since $U \sim N(\theta_R V, V)$, then

$$1 - \beta = P(|U| > c|H_a) \approx P(U < -c|\theta = \theta_R) = \Phi(\frac{-c - \theta_R V}{\sqrt{V}})$$

so we get $-c - \theta_R V = z_{1-\beta}\sqrt{V}$. If we substitute for $c = z_{1-\alpha/2}\sqrt{V}$, and we have

$$V = \frac{(z_{1-\alpha} + z_{1-\beta})^2}{\theta_R^2}$$

recall

$$V = \sum_{j=1}^{k} \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}$$

When the number of events (deaths) is few relative to the number of at risk, we have

$$V \approx \sum_{j=1}^{k} \frac{n_{1j} n_{2j} d_j}{n_j^2}$$

Moreover, if θ is small, and recruitment to each group proceeds at a similar rate, then $n_{1j} \approx n_{2j}$, for $j = 1, \cdots, k$, then V is given by

$$V = \sum_{j=1}^{k} \frac{n_{1j} n_{2j} d_j}{(n_{1j} + n_{2j})^2} \approx \frac{d}{4}$$

where $d = \sum_{j=1}^{k} d_j$ is the total number of events of two groups in the study. Finally, the required number of events d for the study is given by

$$d = \frac{4(z_{1-\alpha} + z_{1-\beta})^2}{\theta_R^2}$$

Note: at later time points the approximation $n_{1j} \approx n_{2j}$ may be

poor, in general, $n_{1j}n_{2j}/n_j^2 < 1/4$, so that the required number of deaths will tend to be underestimated by using this formula (Collett, 2003).

- Required number of patients

Typically patients are recruited over an accrual period a, and followed a follow-up period f. The total study duration is $a + f$. Assume entry times of a patient is uniformly distributed on interval $[0, a]$. Then the death of a patient in a group is

$$\begin{aligned} P(\text{death}) &= \frac{1}{a}\int_0^a P(\text{death}|\text{entry at t})dt \\ &= 1 - \frac{1}{a}\int_f^{a+f} S(u)du \end{aligned}$$

According to Simpson's rule

$$P(\text{death}) = 1 - \frac{1}{6}\{S(f) + 4S(0.5a + f) + S(a + f)\}$$

Therefore, the probability of death in the two treatment groups is

$$P(\text{death}) = 1 - \frac{1}{6}\{\bar{S}(f) + 4\bar{S}(0.5a + f) + \bar{S}(a + f)\}$$

where $\bar{S}(t) = \{S_1(t) + S_2(t)\}/2$.

Once the probability of an event (death) of a patient in the study has been evaluated, th required number of patients (two groups) for the study is given by

$$n = \frac{d}{P(\text{death})}$$

- Example

Clinical trial to asses new treatment for cancer patients. Under standard treatment, 5-year survival probability is 40%. Expect new treatment to increase 5-year survival to 60%. Assuming proportional hazards

$$S_2(t) = [S_1(t)]^\gamma$$

Therefore, the hazard ratio is

$$\gamma = \frac{log(0.60)}{log(0.40)} = 0.56$$

and the log hazard ratio $\theta_R = log(0.56) = -0.58$. The number of death required to have a 90% detecting a hazard ratio 0.56 (20% of 5-year survival difference) to be significant at the 5% level is given by

$$d = \frac{4(1.96 + 1.28)^2}{0.58^2} = 125$$

Allowing for possible underestimate, round it to 130 deaths in total.

Suppose that patients are to be recruited to the study over an $a = 18$ months period and a subsequent follow-up period of $f = 24$ months. Then the probability death is given by

$$P(\text{death}) = 1 - \frac{1}{6}\{\bar{S}(24) + 4\bar{S}(33) + \bar{S}(42)\}$$

Assume the estimate the survival probabilities of control group

(from historical data) are given by $S_1(24) = 0.70, S_1(33) = 0.57$ and $S_1(42) = 0.45$. Then we have
$\bar{S}(24) = (0.70+0.70^{0.56})/2 = 0.76, \bar{S}(33) = (0.57+0.57^{0.56})/2 = 0.65$ and $\bar{S}(42) = (0.45 + 0.45^{0.56})/2 = 0.54$. So the probability of death is $1 - (0.76 + 4 \times 0.65 + 0.54)/6 = 0.35$. Therefore, the required number of patients is

$$n = \frac{130}{0.35} = 372$$

and so 372 patients will need to be recruited to the study over the accrual period of 18 months. This demands a accrual rate of about 21 patients per month.

References

Collett D. *Modeling survival data in medical research*, 2nd ed. Chapman & Hall: London, 2003.

Cox DR and Oakes D. *Analysis of survival data.* Chapman & Hall: London, 1984.

Kalbfleisch JD and Prentice RL. *The statistical analysis of failure time data*, 2nd ed. John Wiley & Sons INC., 2002

Kosorok M. Two-sample quantile tests under general conditions. *Biometrika* 1999; **86**:909-921.

Reid N. Estimating the median survival time. *Biometrika* 1981; **68**:601-608.

Wu J. Confidence intervals for the difference of median failure times applied to censored tumor growth delay data. *Statistics in Biopharmaceutical Research* 2011; **3**:488-496.