Phase II Trials for Heterogeneous Patient Populations with a Time-to-Event Endpoint

Author:

Sin-Ho Jung

Affiliation:

Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina, 27710, U.S.A.

Correspondence address:

E-mail: sinho.jung@duke.edu

SUMMARY

In this paper, we consider a single-arm phase II trial with a time-to-event end-point. We assume that the study population has multiple subpopulations with different prognosis, but the study treatment is expected to be similarly efficacious across the subpopulations. We review a stratified one-sample log-rank test and present its sample size calculation method under some practical design settings. Our sample size method requires specification of the prevalence of subpopulations. We observe that the power of the resulting sample size is not very sensitive to misspecification of the prevalence.

KEY WORDS: Censoring, prevalence, sample size formula, historical control, stratified 1-sample log-rank test

1 Introduction

Phase II trials are to sort out efficacious experimental therapies before proceeding to large scale phase III trials. The patient population of a phase II trial often consists of multiple subpopulations, called strata, with different prognosis. In this case, the final decision on the study treatment should adjust for the heterogeneity of the patient population.

If we randomize patients between a control arm and an experimental arm, then the distribution of patient characteristics defining the strata is expected to be similar between the two arms, so that a univariate analysis ignoring the heterogeneity of patient population is still valid, e.g. Jung (2013). In order to expedite the procedure, however, phase II cancer clinical trials are traditionally designed using a single-arm design treating patients with an experimental treatment only whose efficacy will be compared with a historical control. In a single-arm phase II trial, we hardly can expect the distribution of patient characteristics to be similar to that of a historical control.

Stratified analysis is a popular statistical method to handle the heterogeneity of a study population. One of the most common primary endpoints in phase II cancer clinical trials is tumor response which is a binary variable indicating the size of an index tumor has changed substantially during or following treatment (Simon 1989, Jung et al. 2004). When the clinical outcome is tumor response, London and Chang (2005) and Sposto and Gaynon (2009) propose stratified testing method for single-arm phase II trials. Jung, Chang and Kang (2012) investigate the impact of the standard unstratified testing on type I error and power control when the prevalence of strata are misspecified at the design stage.

Sometimes, tumor response is not appropriate as an endpoint. For examples, in studies of adjuvant chemotherapies, the tumor is completely resected before chemotherapy, so that tumor response is not a meaningful endpoint. Also, tumor response is not a good endpoint for cytotoxic therapies which are meant to prevent the growth of tumor rather than shrinking it. In these cases, a reasonable endpoint is a time to event, such as disease recurrence recurrence or death. Because of the loss to follow-up or termination of the study, event times may be censored. Following the standard terminology, we will use time-to-event, failure time, and survival time as synonymous in this paper.

The one-sample log-rank test (Woolson 1981; Berry 1983; Finkelstein et al. 2003) has been used for single-arm phase II trials to compare the survival distribution of an experimental therapy with that of a historical control. Kwak and Jung (2013) proposed optimal two-stage designs for single-arm phase II trials to be analyzed with the one-sample log-rank test.

In this paper, we review a stratified one-sample log-rank test for single-arm phase II trials with heterogeneous patient populations, and propose its sample size calculation method. The sample size calculation requires specification of the prevalence of strata at the design stage of a phase II trial. We investigate the impact of the erroneously specified prevalence on the statistical power of single-arm phase II trials. We demonstrate our methods with a real phase II cancer clinical trial.

2 Stratified One-Sample Log-Rank Test

Suppose that there are J strata with different survival distributions because of different risk levels. For strata $j(= 1, ..., J)$, let $\Lambda_{0j}(t)$ denote the cumulative hazard function of a selected historical control which are obtained from a previous study or by a retrospective record study. If, for the historical control, we assume an exponential distribution with hazard rate λ_{0j}, then we have $\Lambda_{0j}(t) = \lambda_{0j}t$.

On the other hand, let $\Lambda_j(t)$ denote the unknown cumulative hazard function of the experimental therapy for stratum j. We want to test

$$H_0 : \Lambda_j(t) \geq \Lambda_{0j}(t) \text{ for all } j = 1, ..., J$$

against

$$H_1 : \Lambda_j(t) < \Lambda_{0j}(t) \text{ for some } j = 1, ..., J.$$

Let n_j denote the number of patients from stratum j, and $n = \sum_{j=1}^{J} n_j$ the total sample size. For patient $i(= 1, ..., n_j)$ in stratum j, T_{ji} and C_{ji} denote the survival and censoring times, respectively, that are independent within each stratum. In a real clinical trial, we observe censored survival time $X_{ji} = \min(T_{ji}, C_{ji})$ and event indicator $\delta_{ji} = I(T_{ji} \leq C_{ji})$. We define event and at risk processes $N_{ji}(t) = \delta_{ji} I(X_{ji} \leq t)$ and $Y_{ji}(t) = I(X_{ji} \geq t)$,

respectively, for each patient in stratum j, and $N_j(t) = \sum_{i=1}^{n_j} N_{ji}(t)$ and $Y_j(t) = \sum_{i=1}^{n_j} Y_{ji}(t)$ for stratum j.

Under H_0 for large n, the stratified 1-sample log-rank test

$$W = n^{-1/2} \sum_{j=1}^{J} \int_0^\infty \{dN_j(t) - Y_j(t) d\Lambda_{0j}(t)\}$$

is approximately normal with mean 0 and its variance can be consistently estimated by

$$\hat{\sigma}^2 = n^{-1} \sum_{j=1}^{J} \int_0^\infty Y_j(t) d\Lambda_{0j}(t)$$

refer to, e.g., Fleming and Harrington (1991). So, we reject H_0 with one-sided α if $Z = W/\hat{\sigma} < -z_{1-\alpha}$, where $z_{1-\alpha}$ denotes the $100(1-\alpha)$ percentile of the standard normal distribution.

Note that, for stratum j, $O_j \equiv \int_0^\infty dN_j(t) = \sum_{i=1}^{n_j} \delta_j$ is the observed number of events. Let $S_{0j}(t) = \exp\{-\Lambda_{j0}(t)\}$ denote the survivor function of survival times in stratum j under H_0 and $G(t) = P(C_{ji} \geq t)$ denote the survivor function of the common censoring distribution. Since $n_j^{-1} Y_j(t)$ uniformly converge to $S_{0j}(t)G(t)$,

$$E_j \equiv \int_0^\infty Y_j(t) d\Lambda_{0j}(t) \approx -\int_0^\infty G(t) dS_{0j}(t) = P(T_{ji} < C_{ji} | H_0)$$

is the expected number of events under H_0. Hence, the standardized test statistic is expressed as

$$\frac{W}{\hat{\sigma}} = \sum_{j=1}^{J} \frac{O_j - E_j}{\sqrt{E_j}}.$$

3 Sample Size Calculation

Sample size calculation is one of the key components of a study design for clinical trials. To this end, we propose a method to calculate the required sample size of the stratified one-sample logrank test, $n = \sum_{j=1}^{J} n_j$, for a specified power under a specific alternative hypothesis $H_1 : \Lambda_j(t) = \Lambda_{1j}(t)$ for $j = 1, ..., J$.

Let $\gamma_j = n_j/n$ denote the expected prevalence of stratum j ($\gamma_j > 0$ and $\sum_{j=1}^{J} \gamma_j = 1$), $S_{1j}(t) = \exp\{-\Lambda_{1j}(t)\}$ denote the survivor function of T_{ji} under H_1. Under H_1, $n^{-1} Y_j(t)$

uniformly converges to $\gamma_j G(t)S_{1j}(t)$, so that $\hat{\sigma}^2$ converges to

$$\sigma_0^2 = \sum_{j=1}^{J} \gamma_j \int_0^\infty G(t)S_{1j}(t)d\Lambda_{0j}(t). \tag{1}$$

Under H_1, W is approximately normal with mean

$$\sqrt{n}\omega \equiv \sqrt{n}\sum_{j=1}^{J} \gamma_j \int_0^\infty G(t)S_{1j}(t)d\{\Lambda_{1j}(t) - \Lambda_{0j}(t)\}$$

and variance

$$\sigma_1^2 = \sum_{j=1}^{J} \gamma_j \int_0^\infty G(t)S_{1j}(t)d\Lambda_{1j}(t) = -\sum_{j=1}^{J} \gamma_j \int_0^\infty G(t)dS_{1j}(t). \tag{2}$$

Note that σ_1^2 equals the probability that a patient has an event during the study period when H_1 is true, and $\omega = \sigma_1^2 - \sigma_0^2$.

Hence, under H_1, we have

$$\frac{W}{\hat{\sigma}} \approx \frac{W - \sqrt{n}\omega}{\sigma_1} \times \frac{\sigma_1}{\sigma_0}$$

and $(W - \sqrt{n}\omega)/\sigma_1$ is approximately $N(0,1)$. Using this result, we can derive the power function for given n,

$$1-\beta = P(\frac{W}{\hat{\sigma}} < -z_{1-\alpha}|H_1) \approx P\Big(\frac{W - \sqrt{n}\omega}{\sigma_1} < -\frac{\sqrt{n}\omega}{\sigma_1} - \frac{\sigma_0 z_{1-\alpha}}{\sigma_1}\Big|H_1\Big) = \Phi\Big(-\frac{\sqrt{n}\omega}{\sigma_1} - \frac{\sigma_0 z_{1-\alpha}}{\sigma_1}\Big)$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. By solving this equation and replacing $\omega = \sigma_1^2 - \sigma_0^2$, we obtain the required sample size

$$n = \frac{(\sigma_0 z_{1-\alpha} + \sigma_1 z_{1-\beta})^2}{(\sigma_1^2 - \sigma_0^2)^2}. \tag{3}$$

We consider more practical situations that will simplify the formula (3) in the following subsections.

3.1 Proportional hazards model with a common hazard ratio across strata

Suppose that the survival distributions between experimental and historical control therapies have a proportional hazards model within each stratum. Furthermore, suppose that we

expect similar efficacy improvement across the strata, so that we assume a common hazard ratio across strata, i.e. $\Lambda_{0j}/\Lambda_{1j} = \Delta$ for $j = 1, ..., J$. Then, from (1) and (2), we have $\sigma_0^2 = \Delta\sigma_1^2$ and $\omega = (1-\Delta)\sigma_1^2$. Under this assumption, (3) is simplified to

$$n = \frac{(\sqrt{\Delta}z_{1-\alpha} + z_{1-\beta})^2}{\sigma_1^2(\Delta-1)^2}. \tag{4}$$

Since $\sigma_1^2 = P(T < C|H_1)$ is the probability that a patient experience an event during the study, the expected number of events at the analysis time, $D = n\sigma_1^2$, is expressed as

$$D = \frac{(\sqrt{\Delta}z_{1-\alpha} + z_{1-\beta})^2}{(\Delta-1)^2}. \tag{5}$$

3.2 Under uniform accrual and exponential survival models

Exponential distribution has been one of the most popular parametric models in survival analysis because it fits real survival data relatively well and the computation is easy. Suppose that in the statistical testing we assume exponential survival distributions for the historical control, i.e. $d\Lambda_{0j}(t) = \lambda_{0j}dt$. For the sample size calculation, we assume exponential survival distributions for both experimental and historical control therapies with hazard rates λ_{hj} under H_h for $h = 0, 1$. The survival and cumulative hazard functions are given as $S_{hj}(t) = \exp(-\lambda_{hj}t)$ and $\Lambda_{hj}(t) = \lambda_{hj}t$, respectively. Note that exponential distributions satisfy the proportional hazards assumption.

Furthermore, we assume that patients are accrued with a constant rate during period a and followed for an additional period of b after completion of accrual. Then, the censoring distribution is $U(b, a+b)$ with survivor function $G(t) = P(C \geq t) = 1$ if $t \leq b$; $= (a+b)/a - t/a$ if $b \leq t \leq a+b$; $= 0$ otherwise. Under these assumptions, we have

$$\sigma_1^2 = 1 - \sum_{j=1}^{J} \gamma_j \frac{e^{-b\lambda_{1j}}}{a\lambda_{1j}}(1 - e^{-a\lambda_{1j}}). \tag{6}$$

Similarly, we can show that

$$\sigma_0^2 = \sum_{j=1}^{J} \gamma_j \frac{\lambda_{0j}}{\lambda_{1j}}\{1 - \frac{e^{-b\lambda_{1j}}}{a\lambda_{1j}}(1 - e^{-a\lambda_{1j}})\}. \tag{7}$$

Note that σ_0^2 is equal to σ_1^2 when λ_{0j} is replaced by λ_{1j}. By plugging (6) and (7) in (4), we calculate a sample size under uniform accrual and exponential survival model.

If we further assume a common hazard ratio $\Delta = \lambda_{0j}/\lambda_{1j}$ as in the previous subsection, we have $\sigma_0^2 = \Delta\sigma_1^2$ and $\omega = \sigma_1^2 - \sigma_0^2 = (1-\Delta)\sigma_1^2$. Hence, (4) is expressed as

$$n = \frac{(\sqrt{\Delta}z_{1-\alpha} + z_{1-\beta})^2}{\sigma_1^2(\Delta-1)^2} \tag{8}$$

which has the identical form of (4) but with a simpler expression (6) for σ^2.

3.3 When Accrual Rate is Given in stead of accrual period

In the previous subsections, we have assumed (i) uniform accrual during accrual period a, (ii) exponential survival model, and (iii) constant hazard ratios $\Delta = \lambda_{01}/\lambda_{11} = \cdots = \lambda_{0J}/\lambda_{1J}$. In this subsection, we assume that (i') accrual rate r is given instead of accrual period a in addition to the other assumptions. Note that (i') is more reasonable assumption than (i) because we can estimate the accrual rate based on the number of patients visiting the study institution recently, while accrual period depends on accrual rate and required sample size which is unknown.

From (6), $\sigma_1^2 = \sigma_1^2(a)$ is a function of unknown variable a. By replacing n with $a \times r$ in the left side of (8), we obtain an equation on a,

$$a \times r \times \sigma_1^2(a) = \Big(\frac{\sqrt{\Delta}z_{1-\alpha} + z_{1-\beta}}{\Delta - 1}\Big)^2$$

or, simply

$$a \times r \times \sigma_1^2(a) = D \tag{9}$$

from $n\sigma_1^2 = D$. In order to use (9), we should calculate D by (5) first. We solve one of these equations using a numerical method, such as the bisection method with respect to a. Let a^* denote the solution to the equations. Then, the required sample size and number of events are obtained as $n = a^* \times r$ and $D = a^* \times r \times \sigma_1^2(a^*)$, respectively.

Example: We consider a single-arm phase II clinical trial for pancreatic cancer patients to investigate weekly nab-paclitaxel plus gemcitabine, compared with gemcitabine only as the historical control. The study population consists of two subpopulations ($J = 2$), one with metastatic disease ($j = 1$) and the other with locally advanced disease ($j = 2$). The primary endpoint of the study is progression-free survival (PFS). We will not be interested in the

experimental therapy if its median PFS is $\theta_{01} = 4$ months or shorter for the metastatic disease group and $\theta_{02} = 6$ months or shorter for the locally advanced disease group. And, we will be highly interested in the experimental therapy if its median PFS is $\theta_{11} = 6$ months or longer for the metastatic disease group and $\theta_{12} = 9$ months or longer for the locally advanced disease group. We assume an exponential PFS for the historical control therapy in the statistical testing and for both historical control and experimental therapies in this sample size calculation. So, the annual hazard rates corresponding to these medians are $\lambda_{01} = 2.079$, $\lambda_{02} = \lambda_{11} = 1.386$, and $\lambda_{12} = 0.924$. The hazard ratio is commonly $\Delta = 1.5$ for the both disease groups. We expect an annual accrual of 60 patients from metastatic disease group and 30 patients from locally advanced disease group, i.e. $r = 90$ per year and $(\gamma_1, \gamma_2) = (2/3, 1/3)$. We plan an additional follow-up period of $b = 1$ year. Then, for 90% power, the stratified one-sample log-rank test with 1-sided $\alpha = 0.05$ requires $D = 45$ events (progressions or deaths) from (5) at the final analysis and $n = 57$ patients from (9).

We have generated 10,000 simulation data sets of size $n = 58$ under the design settings of the null and alternative hypotheses, and observed an empirical type I error rate of 0.041 (to be compared with $\alpha = 0.05$) and a power of 0.864 (to be compared with $1 - \beta = 0.9$).

3.4 Impact of Misspecification of Prevalence

In the sample size calculation of a phase II trial, an accurate specification of the prevalence of each strata may be critical to maintain an appropriate statistical power while we may control the type I error accurately using a stratified test statistic regardless of the observed prevalence. The sample size of a standard phase II trial is usually so small that the prevalence specified at the design stage can be quite different from that observed when the study is conducted.

We investigate the impact of misspecification of prevalence of strata at a sample size calculation. We assume the design setting of Example 1 except the prevalence. Let γ_{1j} denote the prevalence specified for sample size calculation and γ_{2j} the true one or the one observed from the study. We calculate the sample size n assuming prevalence of $(\gamma_{11}, \gamma_{12})$ and calculate the power of this sample size when the true prevalence is $(\gamma_{21}, \gamma_{22})$.

Table 1 reports the power of the sample sizes calculated for specified prevalence γ_{11} when

the true prevalence is γ_{21} for stratum 1.

If the specified prevalence is identical to the observed one (the diagonal of Table 1), then we have the power of the nominal $1 - \beta = 0.9$. The lower diagonal cells of Table 1 denote the power of sample size calculated by overestimating the prevalence of the high risk group (stratum 1). In this case, the sample sizes are underpowered compared to the targeted $1 - \beta = 0.9$ since the trial will observe less events than expected at the study design. For example, if we design a trial assuming a prevalence of $\gamma_{11} = 0.7$, but observe $\gamma_{21} = 0.4$ from the trial, then we will have a power of 0.885. Overall, however, we observe that the impact of misspecified prevalence on statistical power is moderate over the wide range of specified and true prevalence. If the prevalence of the high risk group is overestimated (the upper diagonal of Table 1), then we have enough power. Anyhow, at the design stage of a trial, it will be safe to check the power of the calculated sample size for a wide range of prevalence, and to plan a sample size recalculation before completing accrual if necessary.

4 Discussions

In this paper, we have considered design of phase II clinical trials when the patient population consists of multiple subpopulations, called strata, with different prognosis. We assume that the study therapy is expected to be similarly beneficial for all strata. If the study therapy is expected to be efficacious only for a subset of strata (e.g. different disease type), then the eligibility criteria should be appropriately defined to exclude the strata that would not be expected to have the benefit of the study therapy.

We have proposed to account for the heterogeneity of patient population using a stratified testing method for single-arm phase II clinical trials with a time-to-event outcome, such as progression-free survival. We also present a sample size calculation method for the stratified test to be used when designing such trials.

When designing a trial to be analyzed using a stratified test, it is required to specify the prevalence of strata. Jung, Chang and Kang (2012) have shown that unstratified testing can severely distort the type I error and power when the prevalence of strata is different from the real one. In this paper, however, we observe that the power is not much influenced by

misspecification of the prevalence as far as the type I error rate is accurately controlled by using a stratified analysis.

REFERENCES

Berry G. The analysis of mortality by the subject-years methods. *Biometrics* 1983; 39:173-184.

Finkelstein DM, Muzikansky A, Schoenfeld DA. Comparing survival of a sample to that of a standard population. *Journal of the National Cancer Institute* 2003; 95:1434-1439.

Fleming TR, Harrington DP. (1991). *Counting Processes and Survival Analysis.* New York: Wiley.

Jung SH. (2013). *Randomized phase II cancer clinical trials.* New York: Chapman & Hall.

Jung SH, Chang M, Kang S. Phase II cancer clinical trials with heterogeneous patient populations. *Journal of Biopharmaceutical Statistics* 2012; 22:312-328.

Jung SH, Lee TY, Kim KM, George S. Admissible two-stage designs for phase II cancer clinical trials. *Statistics in Medicine* 2004; 23:561-569.

Kwak MJ, Jung SH. Phase II clinical trials with time-to-event endpoints: Optimal two-stage designs with one-sample log-rank test. *Statistics in Medicine* 2013; 33:2004-2016.

London WB, Chang MN. One- and two-stage designs for stratified phase II clinical trials. *Statistics in Medicine* 2005; 24:2597-2611.

Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* 1989; 10:1-10.

Sposto R, Gaynon PS. An adjustment for for patient heterogeneity in the design of two-stage phase II trials. *Statistics in Medicine* 2009; 28:2566-2579.

Woolson RF. Rank-tests and a one-sample log-rank test for comparing observed survival-data to a standard population. Biometrics 1981; 37:687-696.

Table 1: Power of the sample size calculated by specifying a prevalence of γ_{11} for stratum 1 when the true prevalence is γ_{21}

Specified	True Prevalence, γ_{21}								
γ_{11}	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	0.900	0.905	0.910	0.915	0.919	0.924	0.928	0.932	0.935
0.2	0.895	0.900	0.905	0.910	0.915	0.919	0.923	0.928	0.931
0.3	0.889	0.895	0.900	0.905	0.910	0.915	0.919	0.923	0.927
0.4	0.883	0.889	0.895	0.900	0.905	0.910	0.914	0.919	0.923
0.5	0.878	0.884	0.890	0.895	0.900	0.905	0.910	0.914	0.919
0.6	0.872	0.878	0.884	0.890	0.895	0.900	0.905	0.910	0.914
0.7	0.867	0.873	0.879	0.885	0.890	0.895	0.900	0.905	0.910
0.8	0.861	0.867	0.873	0.879	0.885	0.890	0.895	0.900	0.905
0.9	0.855	0.861	0.867	0.874	0.879	0.885	0.890	0.895	0.900

