Published in final edited form as:

J Biopharm Stat. 2018 ; 28(3): 437–450. doi:10.1080/10543406.2017.1315820.

A simulation-based sample size calculation method for preclinical tumor xenograft experiments

Jianrong Wu¹, Shengping Yang²

¹Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, Tennessee, USA

²Department of Pathology, Texas Tech University Health Science Center, Lubbock, Texas, USA

Abstract

Pre-clinical tumor xenograft experiments usually require a small sample size that is rarely greater than 20, and data generated from such experiments very often do not have censored observations. Many statistical tests can be used for analyzing such data, but most of them were developed based on large sample approximation. We demonstrate that the type I error rates of these tests can substantially deviate from the designated rate, especially, when the data to be analyzed has a skewed distribution. Consequently, the sample size calculated based on these tests can be erroneous. We propose a modified signed log-likelihood ratio test (MSLRT) to meet the type I error rate requirement for analyzing pre-clinical tumor xenograft data. The MSLRT has a consistent and symmetric type I error rate that is very close to the designated rate, for a wide range of sample sizes. By simulation, we generated a series of sample size tables based on scenarios commonly expected in tumor xenograft experiments, and we expect that these tables can be used as guidelines for making decisions on the numbers of mice used in tumor xenograft experiments.

Keywords

Log-normal distribution; Pre-clinical tumor xenograft experiment; Modified signed log-likelihood ratio test; Sample size calculation

1 Introduction

In cancer drug development, demonstrating anticancer activity in preclinical tumor xenograft models is important. The National Cancer Institute has promoted a cancer drug screening program since the mid-1950s. The Pediatric Preclinical Testing Program (PPTP) recently conducted a pediatric cancer drug screening program for both *in vitro* and *in vivo* models (Houghton et al., 2007). In *in vivo* testing, human cancer cells from standard tumor lines are engrafted into mice to produce xenograft models. Tumor-bearing mice are randomized into control (*C*) and treatment (*T*) groups, and the maximum tolerated doses of the cytotoxic agents are administered. The volume of each tumor is measured at the initiation of the study and weekly throughout the study period. Mice are euthanized when the tumor volume reaches four times its initial volume (tumor quadrupling). An real tumor xenograft data from

a PPTP study is given in Table 1. Because control tumor grows often quickly and treated tumor grows relatively slow, assessment of treatment effect based on tumor volume is often very inefficient. Stuschke et al. (1990) demonstrated that survival (time to tumor quadrupling) analysis should be used to analyze such data. However, the survival data generated from tumor xenograft models are quite different from that of from clinical trial. Consequently, applying sample size calculation method developed for clinical studies to tumor xenograft experiments in general results in very large sample size requirement due to the difference between clinical and tumor xenograft studies (Borma et al., 2010). Specifically, pre-clinical tumor xenograft experiments are unique due to the followings:

- a. Unlike patients, which usually have heterogeneous genetic background, pre-clinical tumor xenograft subjects (mice), are genetically very similar due to inbreeding, and thus are expected to have homogeneous treatment responses;
- b. Unless there are substantial sub-clones, engrafted tumors in general have much less subtype tissues than tumors in patients, which is another reason that homogeneous responses are expected;
- c. Pre-clinical tumor xenograft studies are usually conducted in controlled environments during a manageable short period of time, and thus there are very few censored observations-in fact, many times, no censored observations.

Consequently, some considerations have to be made in pre-clinical tumor xenograft data analysis and sample size calculation.

- (1) The exponential or proportional hazards assumption, which is commonly used in survival trial sample size calculation (Schoenfeld, 1981), is very often violated in pre-clinical tumor xenograft experiments (mainly due to a and b), for example, the survival curves can completely separate between untreated and treated groups if the treatment is effective.
- (2) The number of subjects required in tumor xenograft experiments is in general small which renders most of the existing statistical tests, including the Wald, Score, likelihood ratio and log-rank test, etc., problematic (e.g., inaccurate type I error rate) since they were all developed based on large sample approximation.
- (3) In reality, the number of mice used in tumor xenograft experiments is almost never greater than 20, and thus it is feasible and straightforward to use simulation, rather than derive a formula, to obtain the required sample size.

In order to adopt a simulation-based strategy in sample size calculation, it is important that the test to be used have consistent type I error rate for a wide range of sample sizes. However, since sample size requirement for tumor xenograft experiments can be as small as 3 (Mruphy et al., 2016), most of tests developed based on large sample approximation would not be appropriate. In addition, although minimal/no censored observation in tumor xenograft experiments allows us to use a parametric test for data analysis, it is critical to make valid assumptions on the distribution of the survival time. To address this issue, we performed extensive evaluation on the distribution of tumor quadrupling time using data from various studies. Being in consistent to a and b mentioned above, our observations show that tumor quadrupling time of the control group in general approximately follows a normal

distribution, while the treated group usually has a distribution skewed to the right. We consulted several biologists, oncologists and pharmacologists, all of them acknowledged that they had the same observations. One possible interpretation is that, without treatment, all engrafted mice have similar responses to the dominate tumor cells; however, with treatment, more heterogeneous responses can be triggered due to subtle differences in drug absorption, metabolism, as well as potential tumor cell heterogeneity, which results in differences in drug resistance. This obvious difference between the distributions of the control and treated groups introduces increased difficulty in making statistical comparisons, and similarly, in sample size calculations. Although non-parametric test (for example, Wilcoxon test) seems to be a choice for such a situation, its statistical power can be 0 if the number of subjects in each group is less than 4. Neother (1987) developed a sample size calculation formula for the Wilcoxon test. However, the formula has serious limitations, including: 1) a plateau on power regardless of increased effect size; 2) obvious incorrect power for sample size < 4 . In this paper, we propose a MSLRT for tumor xenograft data analysis, and provide a series of sample size tables that can be used for determining the appropriate sample size for conducting tumor xenograft experiments. The rest of the paper is organized as the follows: In Section 2, we introduce the MSLRT statistics. In section 3, we compare the power and type I error rates of a number of existing test statistics with MSLRT, under a variety of scenarios. In section 4, we use an example to demonstrate the performance of the MSLRT. In Section 5, we generate a series of sample size tables for experiments to be used for testing commonly made hypotheses. In addition, we provide some brief discussions on issues associated with sample size calculation for tumor xenograft experiments in section 6.

2 Test Statistics

In tumor xenograft experiments, the tumor quadrupling time of the control group in general has a normal distribution, and the treated group has a skewed distribution, the comparison is thus in favor of median other than mean because comparing median makes sense for both groups. Given the uniqueness of tumor xenograft data, we will derive four test statistics for testing the hypothesis of equal medians of two independent groups. Let x_i be the tumor quadrupling time of the i^{th} subject in the first group ($i = 1, \dots, n$) and y_j be the tumor quadrupling time of the j^{th} subject in the second group ($j = 1, \dots, m$), and we assume there is no censored observations. The corresponding medians are M_1 and M_2 , respectively. Assume that x_i and $\log y_j$ are independently and normally distributed with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 , respectively. More specifically,

$$x_i \sim N(\mu_1, \sigma_1^2), \text{ and } \log y_j \sim N(\mu_2, \sigma_2^2).$$

The null hypothesis of interest is

$$H_0: M_1 = M_2.$$

where $M_1 = \mu_1$ and $M_2 = e^{\mu_2}$ are the medians of two groups. Let $\psi = \log(M_1/M_2) = \log \mu_1 - \mu_2$ be the logarithm ratio of the medians. Testing the hypothesis H_0 is equivalent to testing the hypothesis of $H_0 : \psi = 0$. In the rest of the scenario, let $\psi = \log(M_1/M_2) = \log \mu - \mu_2$ be a parameter of interest, and let $\lambda = (\mu, \sigma_1^2, \sigma_2^2)'$ be a vector nuisance parameter, where $\mu = \mu_1$. Then the log-likelihood function of $\theta = (\psi, \lambda)' = (\psi, \mu, \sigma_1^2, \sigma_2^2)'$ based on sample (x, y) can be written as

$$\begin{aligned} \ell(\theta) = & -n \log \sigma_1 - m \log \sigma_2 + \frac{\mu}{\sigma_1^2} t_1 - \frac{1}{\sigma_1^2} t_3 + \frac{1}{\sigma_2^2} (\log \mu - \psi) t_2 \\ & - \frac{1}{2\sigma_2^2} t_4 - \frac{m}{2\sigma_2^2} (\log \mu - \psi)^2 - \frac{n\mu^2}{2\sigma_1^2}, \end{aligned} \quad (1)$$

where $t = (t_1, t_2, t_3, t_4)' = \left(\sum_{i=1}^n x_i, \sum_{j=1}^m \log y_j, \sum_{i=1}^n x_i^2, \sum_{j=1}^m (\log y_j)^2 \right)'$ is a minimum sufficient statistic. It can be shown that the maximum likelihood estimate $\hat{\theta} = (\hat{\psi}, \hat{\mu}, \hat{\sigma}_1^2, \hat{\sigma}_2^2)'$ is given by

$$\begin{aligned} \hat{\psi} &= \log \hat{\mu} - \bar{t}_2, \\ \hat{\mu} &= \bar{t}_1, \\ \hat{\sigma}_1^2 &= \bar{t}_3 - \bar{t}_1^2, \\ \hat{\sigma}_2^2 &= \bar{t}_4 - \bar{t}_2^2. \end{aligned}$$

Under the null hypothesis $H_0 : \psi = \psi_0$, the constrained maximum likelihood estimate $\hat{\theta}_{\psi_0} = (\psi_0, \hat{\mu}_{\psi_0}, \hat{\sigma}_{1\psi_0}^2, \hat{\sigma}_{2\psi_0}^2)'$ can be obtained numerically by using R function "nlmin" (R core team, 2013).

The expected Fisher information matrix is derived as

$$I(\theta) = \begin{pmatrix} \frac{m}{\sigma_2^2} & -\frac{m}{\mu\sigma_2^2} & 0 & 0 \\ -\frac{m}{\mu\sigma_2^2} & \frac{n}{\sigma_1^2} + \frac{m}{\mu^2\sigma_2^2} & 0 & 0 \\ 0 & 0 & \frac{2n}{\sigma_1^2} & 0 \\ 0 & 0 & 0 & \frac{2m}{\sigma_2^2} \end{pmatrix}. \quad (2)$$

To derive the test statistics for testing the hypothesis of equal medians, we will consider three large-sample statistics: Score statistic, Wald statistic, and signed log-likelihood ratio

statistic; and a small-sample test statistic, known as the MSLRT (Barndorff-Nielsen, 1986, 1991; Fraser, et al., 1999).

For the null hypothesis $H_0: \psi = \psi_0$ (often $\psi_0 = 0$), the Score statistic is defined as the standardized score function

$$S(\psi_0) = \ell_{\psi}(\hat{\theta}_{\psi_0}) \left\{ I^{\psi\psi}(\hat{\theta}_{\psi_0}) \right\}^{1/2}, \quad (3)$$

where $\ell_{\psi}(\theta) = \partial \ell(\theta) / \partial \psi$ is the score function, and $I^{\psi\psi}(\theta)$ is the inverse expected Fisher information matrix $\mathcal{I}(\theta)$ corresponding to the (ψ, ψ) partition. Based on the log-likelihood function (1), the Score statistic can be derived as

$$S(\psi_0) = \left\{ \frac{m}{\hat{\sigma}_{2\psi_0}^2} (\log \hat{\mu}_{\psi_0} - \psi) - \frac{t_2}{\hat{\sigma}_{2\psi_0}^2} \right\} \left\{ \frac{n}{\hat{\sigma}_{1\psi_0}^2} + \frac{m}{\hat{\mu}_{\psi_0}^2 \hat{\sigma}_{2\psi_0}^2} \right\}^{1/2} \left(\frac{nm}{\hat{\sigma}_{1\psi_0}^2 \hat{\sigma}_{2\psi_0}^2} \right)^{-1/2}. \quad (4)$$

The Wald statistic is defined as the standardized maximum likelihood estimate

$$W(\psi_0) = (\hat{\psi} - \psi_0) \left\{ I^{\psi\psi}(\hat{\theta}_{\psi_0}) \right\}^{-1/2}, \quad (5)$$

which can be simplified as

$$W(\psi_0) = (\hat{\psi} - \psi_0) \left(\frac{n}{\hat{\sigma}_{1\psi_0}^2} + \frac{m}{\hat{\mu}_{\psi_0}^2 \hat{\sigma}_{2\psi_0}^2} \right)^{-1/2} \left(\frac{nm}{\hat{\sigma}_{1\psi_0}^2 \hat{\sigma}_{2\psi_0}^2} \right)^{1/2}.$$

The signed log-likelihood ratio statistic can be calculated from the log-likelihood function (1) as

$$R(\psi_0) = \text{sgn}(\hat{\psi} - \psi_0) \left\{ 2 \left[\ell(\hat{\theta}) - \ell(\hat{\theta}_{\psi_0}) \right] \right\}^{1/2}. \quad (6)$$

The distribution of all three statistics, $S(\psi_0)$, $W(\psi_0)$, and $R(\psi_0)$ is approximately distributed as standard normal when the sample size is large, but they all suffer from an asymmetric, liberal (or conservative) type I error rate when the sample size is small. However, the loss in accuracy of the large sample approximations could be recovered by using recently developed high-order asymptotic normality theory. Applications of such high-order accurate methods to statistical inference may be found in Wong and Wu (2000) and Wu et al. (2002, 2003, 2006). In this article, we consider a modified signed log-likelihood ratio statistic introduced by Barndorff-Nielsen (1986,1991) and Fraser et al. (1999). It is generally known as the R^* -formula, which has the form

$$R^*(\psi_0) = R(\psi_0) + R(\psi_0)^{-1} \log \left\{ \frac{U(\psi_0)}{R(\psi_0)} \right\}, \quad (7)$$

where $R(\psi_0)$ is the signed log-likelihood ratio statistic given in (6), and $U(\psi_0)$ is a new statistic.

In general, the statistic $U(\psi_0)$ can be hard to obtain. However, it has been shown (Fraser, et al., 1999) that if the log-likelihood function $l(\theta) = l(\theta, x, y)$ can be written as $l(\theta, t)$, where t is a minimum sufficient statistic with the same dimension as the parameter θ , then $U(\psi_0)$ could be simplified as

$$U(\psi_0) = \frac{\left| \ell_{;t}(\hat{\theta}) - \ell_{;t}(\hat{\theta}_{\psi_0}) \ell_{\lambda;t}(\hat{\theta}_{\psi_0}) \right|}{\left| \ell_{\theta;t}(\hat{\theta}) \right|} \left(\frac{\left| j_{\theta\theta}(\hat{\theta}) \right|}{\left| j_{\lambda\lambda}(\hat{\theta}_{\psi_0}) \right|} \right)^{1/2}, \quad (8)$$

where the sample space derivatives are defined as

$$\ell_{;t}(\theta) = \frac{\partial}{\partial t} \ell(\theta; t);$$

the mixed derivatives are defined as

$$\ell_{\theta;t}(\theta) = \frac{\partial}{\partial \theta'} \ell_{;t}(\theta);$$

$j_{\theta\theta}(\theta) = -\partial^2 \ell(\theta) / \partial \theta \partial \theta'$ is the observed information matrix; and $j_{\lambda\lambda}(\theta) = -\partial^2 \ell(\theta) / \partial \lambda \partial \lambda'$ is the observed nuisance information matrix.

Hence, the Score statistic $\mathcal{S}(\psi_0)$, Wald statistic $W(\psi_0)$, signed log-likelihood ratio statistic $R(\psi_0)$, and the MSLRT $R^*(\psi_0)$ can be obtained from (3) through (8).

3 Type I Error Rate and Power Assessment

In this section, we carry out a series of simulations to compare the performance of the t -test, Wilcoxon test, Score, Wald, signed log-likelihood ratio tests and MSLRT. Since the objective of this paper is to develop sample size calculation method for pre-clinical tumor xenograft experiments, we expect the range of the sample size for each group to be between 3 at the minimum and 20 at the maximum. Therefore, it is informative to assess type I error rates based on two sample sizes ($n = 3$ and $n = 15$, respectively) (Tekindal et al., 2016; Tekindal and Yazici, 2016). Furthermore, based on extensive examination of existing pre-clinical tumor xenograft data, we conclude that it is meaningful to use three different values, i.e., 2, 3 and 4, to simulate the standard deviation of the tumor quadrupling time for the control group. Also based on our observation that the treated group in general has larger standard deviation than the control group, we included four standard deviation (treated/control) ratios, i.e., 1, 2, 3 and 4, in our simulation.

The simulation results (Table 2) show that when samples size is small ($n = 3$ scenario), both the t -test and MSLRT (denoted as R^* in the table) test have well controlled type I error rate. As a comparison, both the Score and Wald tests are too conservative, while the Wilcoxon test has no power due to the rank-based nature of its test statistic. In addition, both the left-tail and right-tail error rates for the signed log-likelihood ratio statistic (denoted as R in the table) are inflated. Under the scenario of larger sample size ($n = 15$), many of the tests have improved type I error rate, including the t -test, Score, Wald, and signed log-likelihood ratio tests. However, the performance of the t -test becomes less desirable, i.e., while the type I errors are inflated for the left-tail test, the right-tail test is too conservative. Meanwhile, the Wilcoxon test is very conservative. As a comparison, the proposed MSLRT performed consistently well under all scenarios, with the type I error rate close to the nominal level of 0.05 for both left and right trial. Furthermore, the MSLRT test possesses not only nearly exact, but also symmetric two-tail error rates.

Table 3 presents the simulation results on power for small sample size $n = 3, 5, 8$ and 10. Because the score, Wald and signed log-likelihood ratio statistic did not preserve the correct type I error rate, thus, it is not meaningful to discuss the empirical power. The Wilcoxon test is a non-parametric test which is designed for comparing the medians. However, the Wilcoxon test is very conservative and consistently has lower power than that of the MSLRT as shown in the power simulation. Thus, the MSLRT is recommended for small sample tumor xenograft studies. In addition, we have also performed simulations for other parameter configurations; the results are consistent with those given in Tables 1 and 2 (data not shown).

4 Example

In this section, we demonstrate the performance of the MSLRT using data generated from a tumor xenograft experiment that had been conducted to demonstrate the antitumor activity of Rapamycin against an EW5 solid tumor model (Kolb, et al., 2012) (see Table 1). One of the goals of the study was to compare the tumor quadrupling time of the control and treated mice. In *in vivo* solid tumor xenograft experiments, the volume of each tumor is measured on a weekly schedule for logistical reasons. Therefore, the exact time of quadrupling of a tumor is not measured. The following interpolation formula can be used to calculate the tumor quadrupling time,

$$t_e = t_1 + (t_2 - t_1) \log(V_e/V_{t_1}) / \log(V_{t_2}/V_{t_1}),$$

where t_e is the interpolated quadrupling time, t_1 and t_2 are the lower and upper observation times bracketing the quadrupling tumor volume $V_e = 4V_0$, where V_0 is the initial tumor volume (Wu, 2009). The tumor quadrupling times calculated by using the interpolation formula for the EW5 tumor line are given in Table 4. The Q-Q plots show that tumor quadrupling time of the control group approximately follows a normal distribution, and the log transformed tumor quadrupling time of the treated group also approximately follows a normal distribution. The Shapiro-Wilk test for normality on the tumor quadrupling time gives a p-value of 0.162 (0.019) for the control (treated) group, and the test on the treated

group for the log transformed tumor quadrupling time has a p-value of 0.146. These findings confirm our assumption that the tumor quadrupling time of the control group has a normal distribution, while that of the treated group approximately has a log-normal distribution. The median tumor quadrupling times of the control and treated groups were 7.07 and 17.78 days, respectively. The corresponding standard deviations were 1.49 and 6.34, respectively. The two-sided p-values for testing equality of medians are presented in Table 5. Note that due to the relative large sample size ($n = 10$ for each group) in this study, the p-values for all the tests are quite similar, and we present this example mainly to demonstrate the commonly seen distributions of the tumor quadrupling times for both the treated and control groups.

5 Simulation-based Sample Size Calculation

Since the proposed MSLRT has well controlled type I error rate over a wide range of sample sizes, and the sample size required for tumor xenograft experiments is almost never greater than 20, it is straightforward to calculate sample size using simulations. Note that, the type I error rates for all other tests do not behave consistently well for small (e.g., $n = 3$) and large (e.g., $n = 15$) sample sizes, therefore, sample size calculated based on either simulation or a formula could be erroneous, and thus we are not interested in simulating sample size requirements for those tests. For the MSLRT, our simulation started with $n = 3$ for each group (with 10,000 simulations performed), and if the simulated statistical power is smaller than the pre-specified power (for example, 80% or 90%), then we performed another simulations using $n^* = n + 1$. We continued doing this until the simulated power exceeds the pre-specified power, then we used n^* as the required sample size. The sample size obtained based on the simulations show that under most of the simulated scenarios, three mice per group is necessary for the MSLRT to achieve 80% power, at the 0.05 significance level to detect a difference between the treated and control groups. The largest sample size from the simulation is 16, when the median tumor quadrupling times of the control and treated groups were 30 and 40 days (a small difference), respectively, and the standard deviations of both the control and treated groups were large. These results (Table 4) agree well with the sample size commonly used in pre-clinical tumor xenograft experiments. The R code for the simulation based sample size calculation is available upon request.

6 Discussions

Very different from cancer clinical study, the mice of tumor xenograft experiments are genetically homogeneous, and the tumor tested often comes from one patient. As a result, the tumor quadrupling times of the control group in general approximately follows a normal distribution with very small variation, and the tumor quadrupling times of the treated group has a skewed distribution with a larger variation due to subtle differences in drug response and the potential existence of tumor minor clones. In this paper, we introduced a novel test, named MSLRT, for comparing tumor quadrupling times with different distributions, and generated a series of sample size tables that could facilitate making sample size decisions for tumor xenograft experiments. Besides log-rank test, a number of parametric/non-parametric tests can be used in tumor xenograft data analysis due to minimal/none censored observations. Furthermore, since tumor xenograft experiments almost always require less than 20 subjects per treatment, due to genetic homogeneity, sample size calculation can be

straightforwardly performed via simulations. To perform sample size simulation, it is critical that the test to be used has consistently well-controlled type I error rate. We compared 6 commonly used statistical tests under the scenarios where the tumor quadrupling times of the two groups to be compared have different distributions, i.e., a normal distribution and a log-normal distribution, and found that the MSLRT outperformed all other tests. Assumptions have to be made in virtually all sample size calculations. The assumptions we have made for the MSLRT were that the tumor quadrupling time of the control group approximately follows a normal distribution, and that of the treated group approximately follows a log-normal distribution, which can be well justified based on tumor and host biology. We examined these assumptions using the existing data from a large number of tumor xenograft experiments, and the assumptions seem to be reasonable. We generated a series of sample size tables based on scenarios commonly expected in tumor xenograft experiments, and we anticipate that these tables provide valuable guidelines for researchers planning their tumor xenograft experiments.

Acknowledgments

The author acknowledges two anonymous reviewers for their valuable comments that improved an earlier version of the paper. The first author's work was supported in part by the National Cancer Institute support grant P30CA021765 and ALSAC.

Appendix:: R code for the sample size calculation

```
library("rootSolve")
lik=function(theta)
{ psi=theta[1]
  mu=theta[2]
  sig1=theta[3]
  sig2=theta[4]
  lik0=-n*log(sig1)-m*log(sig2)+(mu/sig1^2)*t1-t3/(2*sig1^2)+
    (-psi+log(mu))*(t2/sig2^2)-t4/(2*sig2^2)-
    n*mû2/(2*sig1^2)-m*(-psi+log(mu))^2/(2*sig2^2)

  return(lik0)}

nlik=function(theta)
{ -lik(theta)}

llik=function(lam)
{ mu=lam[1]
  sig1=lam[2]
  sig2=lam[3]
  llik0=-n*log(sig1)-m*log(sig2)+(mu/sig1^2)*t1-t3/(2*sig1^2)+
    (-psi+log(mu))*(t2/sig2^2)-t4/(2*sig2^2)-
    n*mû2/(2*sig1^2)-m*(-psi+log(mu))^2/(2*sig2^2)
```

```

    return(llik0)}

nllik=function(lam)
{ -llik(lam) }

jj=function(theta)
{ psi=theta[1]

  mu=theta[2]
  sig1=theta[3]
  sig2=theta[4]
  j11=m/sig2^2
  j12=-m/(mu*sig2^2)
  j13=0
  j14=-2*t2/sig2^3+2*m*(-psi+log(mu))/sig2^3
  j22=t2/(m02*sig2^2)+n/sig1^2-m*(-psi+log(mu))/(m02*sig2^2)+
    m/(m02*sig2^2)
  j23=2*t1/sig1^3-2*n*mu/sig1^3
  j24=2*t2/(mu*sig2^3)-2*m*(-psi+log(mu))/(mu*sig2^3)
  j33=-n/sig1^2+3*(t3-2*mu*t1+n*m02)/sig1^4
  j34=0
  j44=-m/sig2^2-6*(-psi+log(mu))*(t2/sig2^4)+
    3*t4/sig2^4+3*m*(-psi+log(mu))^2/sig2^4
  jj0=matrix(c(j11,j12,j13,j14,j12,j22,j23,j24,j13,j23,j33,j34,
    j14,j24,j34,j44), 4,4)
  return(jj0)}

II=function(theta)
{ psi=theta[1]
  mu=theta[2]
  sig1=theta[3]
  sig2=theta[4]
  I0=matrix(0,4,4)
  I0[1,1]=m/sig2^2
  I0[1,2]=I0[2,1]=-m/(mu*sig2^2)
  I0[2,2]=n/sig1^2+m/(m02*sig2^2)
  I0[3,3]=2*n/sig1^2 I0[4,4]=2*m/sig2^2
  return(I0)}

I11=function(theta)
{ psi=theta[1]
  mu=theta[2]
  sig1=theta[3]
  sig2=theta[4]

```

```
I0=(n/sig1^2+m/(m02*sig2^2))/((n*m)/(sig1*sig2)^2)
return(I0)}
```

```
Score=function(theta)
{ psi=theta[1]
  mu=theta[2]
  sig1=theta[3]
  sig2=theta[4]
  score0=-t2/sig2^2+m*(-psi+log(mu))/sig2^2
  return(score0)}
```

```
lt=function(theta)
{ psi=theta[1]
  mu=theta[2]
  sig1=theta[3]
  sig2=theta[4]
  lt1=mu/sig1^2
  lt2=(-psi+log(mu))/sig2^2
  lt3=-1/(2*sig1^2)
  lt4=-1/(2*sig2^2)
  lt0=c(lt1,lt2,lt3,lt4)
  return(lt0)}
```

```
llamt=function(theta)
{ psi=theta[1]
  mu=theta[2]
  sig1=theta[3]
  sig2=theta[4]
  llamt0=matrix(0, 4, 3)
  llamt0[1,1]=1/sig1^2
  llamt0[1,2]=-2*mu/sig1^3
  llamt0[2,1]=1/(mu*sig2^2)
  llamt0[2,3]=-2*(-psi+log(mu))/sig2^3
  llamt0[3,2]=1/sig1^3
  llamt0[4,3]=1/sig2^3
  return(llamt0)}
```

```
rr=function(theta1, theta0)
{ rr0=sign(theta1[1]-theta0[1])*(2*(lik(theta1)-lik(theta0)))^(0.5)
  return(rr0)}
```

```
uu=function(theta1, theta0)
{ sig1=theta1[3]
  sig2=theta1[4]
```

```

    lt1=lt(theta1)
    lt0=lt(theta0)
    llamt0=llamt(theta0)
    mat0=cbind(lt1-lt0, llamt0)
    uu0=det(mat0)
    uu1=1/(sig1*sig2)^5
    uu2=det(jj(theta1))
    uu3=det(jj(theta0)[-1,-1])
    u0=(uu0/uu1)*(uu2/uu3)^(1/2)
    return(u0)}

# set median survival serie of the treated group
md2c=c(20,30,40)
# set median survival of the untreated group
mdl=10
# set the standard deviation serie of the untreated group
siglc=c(2,3,4)
# set the treated/untreated standard deviation ratios
ratioc=c(2,3,4)
# set the statistical power serie
powerc=c(0.8, 0.9)
# set the type I error rate serie
alfac=c(0.05, 0.1)

for(ite1 in 1:length(powerc))
{ power=powerc[ite1]
  for (ite2 in 1:length(alfac))
  {alfa=alfac[ite2]
    for (ite3 in 1:length(ratioc))
    {ratio=ratioc[ite3]
      for (ite4 in 1:length(siglc))
      {sig1=siglc[ite4]
        for (ite5 in 1:length(md2c))
        {md2=md2c[ite5]
          mu2=log(md2)
          mul=mdl
          fun5<-function(x){sqrt((exp(x^2)-1)*exp(2*mu2+x^2))-sig1*ratio}
          sig2<-uniroot(fun5, c(0, 2))$root
          psi=0

B=10000
for (nx in 3:20)
{ # initiate sample size, starting from 3 n=m=nx
  for (b in 1:B)

```

```

{ x=rnorm(n,mu1, sig1)
  y=rnorm(m,mu2, sig2)
  t1=sum(x)
  t2=sum(y)
  t3=sum(x^2)
  t4=sum(y^2)
  t1.bar=t1/n
  t2.bar=t2/m
  mu.h=t1.bar
  sig1.h=sqrt((t3-n*t1.bar^2)/n)
  sig2.h=sqrt((t4-m*t2.bar^2)/m)
  psi.h=log(mu.h)-t2.bar
  theta.h=c(psi.h, mu.h, sig1.h, sig2.h)
  beta.h=theta.h[-1]
  beta0=beta.h
  beta.ch=nlminb(beta0, nllik)$par
  beta.ch=nlminb(beta0, nllik)$par
  theta.ch=c(psi, beta.ch)
  S0=Score(theta.ch)*(I11(theta.ch))^(1/2)
  Wa0=psi.h*(I11(theta.ch))^(1/2)
  rr0=rr(theta.h, theta.ch)
  uu0=uu(theta.h, theta.ch)
  mr0=rr0+log(abs(uu0/rr0))/rr0
  ans=round(c(b, 0, 0, 0, 0, 0, mr0, 0), digits=5)
  if (b==1) {anns=ans} else
  {anns=cbind(anns,ans)}
}
dat=t(anns)
mr0=dat[,7]
df=dat[,8]
N=dim(dat)[[1]]
mrl=length(mr0[mr0<qnorm(alfa)])/N
out=paste("power=",power,"alfa=",alfa,"md1=",md1,"md2=",md2,"sig1=",
          sig1,"ratio=",ratio,"n=",nx,"mrl=",mrl)
if (nx==20) {break}
if (nx<20 & mrl>power) {break}
}
print (out)}}}}

```

References

- Barndorff-Nielsen OE. (1986). Inference on full and partial parameters, based on the standardized signed log likelihood ratio. *Biometrika*, 73, 307–322.
- Barndorff-Nielsen OE. (1991). Modified signed log-likelihood ratio. *Biometrika*, 78, 557–563.
- Borma GF, Bloemb BR, Munnekeb M, Teerenstra S. (2010). A simple method for calculating power based on a prior trial. *Journal of Clinical Epidemiology*, 63:992–997. [PubMed: 20573484]

- Fraser DAS, Reid N, Wu J. (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika*, 86, 249–264.
- Kolb EA, Gorlick R, Maris JM, Keir ST, Morton CL, Wu J, Wozniak AW, Smith MA, Houghton PJ. (2012). Combination Testing (Stage 2) of the Anti-IGF-1 Receptor Antibody IMC-A12 With Rapamycin by the Pediatric Preclinical Testing Program. *Pediatr Blood Cancer*, 58:729–735. [PubMed: 21630428]
- Mantel N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 50:16370.
- Murphy B, Yin H, Maris JM, Kolb EA, Gorlick R, Reynolds CP, Kang MH, Keir ST, Kurmasheva RT, Dvorchik I, Wu J, Billups CA, Boateng N, Smith MA, Houghton PJ. (2016). Evaluation of Alternative In Vivo Drug Screening Methodology: single mouse analysis, *Cancer Research*, Accept.
- Noether GE. (1987). Sample Size Determination for Some Common Non-parametric Tests. *Journal of the American Statistical Association*, 82:645–647.
- R Core Team. (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
- Schoenfeld DA. (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika*, 68:316319.
- Tekindal MA, Gullu O, Yazici AC, Yavuz Y. (2016). The Cochran-Armitage Test To Estimate The Sample Size For Trend Of Proportions For Biological Data. *Turkish Journal of Field Crops*, 21:286–297.
- Tekindal MA, Yazici AC. (2016). Williams Test Required Sample Size For Determining The Minimum Effective Dose. (2016). *Turkiye Klinikleri Journal Biostatistics*, 8:53–81
- Wong ACM. and Wu J, (2000). Practical small sample asymptotics for distributions used in life-data analysis. *Technometrics*, 42, 149–156.
- Wu J, Houghton PJ. (2009). Assessing Cytotoxic Treatment Effects in Pre-clinical Tumor Xenograft Models. *Journal of Biopharmaceutical Statistics*, 19:755–762. [PubMed: 20183441]
- Wu J, Jiang G, Wei W, (2006). Confidence intervals of effect size in randomized comparative parallel-group studies. *Statistics in Medicine*, 25, 639–651. [PubMed: 16158411]
- Wu J, Jiang G, Wong ACM, Sun X, (2002). Likelihood analysis for the ratio of means of two independent log-normal distributions. *Biometrics*, 58, 463–469. [PubMed: 12071423]
- Wu J, Wong ACM, Jiang G, (2003). Likelihood-based confidence intervals for log-normal mean. *Statistics in Medicine*, 22, 1849–1860. [PubMed: 12754720]

Table 1:Tumor volumes (cm^3) measured in EW5 tumor xenograft model

Group	Days	Mouse									
		M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
Control	0	0.44	0.23	0.42	0.26	0.19	0.56	0.36	0.25	0.20	0.43
	7	2.01	0.37	1.73	0.72	0.55	2.70	1.73	0.98	0.68	1.75
	14	.	2.03	.	2.83	1.92	.	.	2.92	1.55	.
Rapamycin	0	0.43	0.6	0.42	0.29	0.21	0.54	0.26	0.31	0.18	0.24
	7	0.58	0.89	0.69	0.19	0.1	0.87	0.24	0.41	0.14	0.24
	14	1.03	2.11	0.93	0.19	0.16	1.72	0.32	0.86	0.17	0.45
	21	2.14	3.7	2.6	0.48	0.28	2.76	0.81	1.9	0.26	1.03
	28	.	.	.	0.7	0.4	.	1.03	.	0.35	.
	35	.	.	.	1.67	0.9	.	1.97	.	0.89	.

.: indicates a missing value due to mouse sacrificed after its tumor quadrupled.

Table 2:

Evaluation of one-tail empirical type I error rates of six methods for testing medians with nominal level $\alpha = 0.05$ for each tail.

Test Statistics								
<i>n</i>	SD	Ratio SDs	<i>t</i>	Wiicoxon	Score	Wald	<i>R</i>	<i>R</i> [*]
3	2	1	0.0382 [*]	0	0.0265	0.0379	0.0959	0.0404
3	2	1	0.0424 [†]	0	0.0222	0.0207	0.0951	0.0422
3	2	2	0.0418	0	0.0310	0.0394	0.1020	0.0472
3	2	2	0.0469	0	0.0275	0.0262	0.0983	0.0457
3	2	3	0.0423	0	0.0312	0.0376	0.1087	0.0489
3	2	3	0.0535	0	0.0272	0.0261	0.1066	0.0493
3	3	1	0.0373	0	0.0271	0.0445	0.0957	0.0424
3	3	1	0.0402	0	0.0230	0.0221	0.0878	0.0378
3	3	2	0.0388	0	0.0312	0.0408	0.1023	0.0475
3	3	2	0.0486	0	0.0263	0.0247	0.0958	0.0458
3	3	3	0.0393	0	0.0332	0.0394	0.1074	0.0499
3	3	3	0.0534	0	0.0244	0.0242	0.1004	0.0467
3	4	1	0.0396	0	0.0297	0.0523	0.0989	0.0439
3	4	1	0.0394	0	0.0191	0.0177	0.0901	0.0390
3	4	2	0.0413	0	0.0351	0.0501	0.1053	0.0495
3	4	2	0.0458	0	0.0233	0.0221	0.0908	0.0414
3	4	3	0.0410	0	0.0360	0.0445	0.1084	0.0545
3	4	3	0.0540	0	0.0266	0.0259	0.0988	0.0458
15	2	1	0.0631	0.0271	0.0548	0.0559	0.0595	0.0522
15	2	1	0.0391	0.0187	0.0520	0.0505	0.0553	0.0495
15	2	2	0.0785	0.0303	0.0535	0.0538	0.0591	0.0511
15	2	2	0.0267	0.0186	0.0459	0.0455	0.0524	0.0443
15	2	3	0.0981	0.0352	0.0548	0.0548	0.0631	0.0537
15	2	3	0.0235	0.0268	0.0496	0.0494	0.0548	0.0486
15	3	1	0.0707	0.0273	0.0571	0.0590	0.0607	0.0532
15	3	1	0.0345	0.0163	0.0508	0.0486	0.0561	0.0486
15	3	2	0.0966	0.0357	0.0597	0.0606	0.0652	0.0566
15	3	2	0.0255	0.0211	0.0534	0.0526	0.0586	0.0517
15	3	3	0.1146	0.0373	0.0531	0.0533	0.0603	0.0508
15	3	3	0.0202	0.0236	0.0485	0.0485	0.0554	0.0485
15	4	1	0.0723	0.0260	0.0501	0.0532	0.0546	0.0469
15	4	1	0.0329	0.0182	0.0486	0.0470	0.0532	0.0478
15	4	2	0.1108	0.0353	0.0561	0.0571	0.0611	0.0524
15	4	2	0.0226	0.0190	0.0534	0.0518	0.0585	0.0522
15	4	3	0.1399	0.0373	0.0553	0.0556	0.0621	0.0521

Test Statistics								
<i>n</i>	SD	Ratio SDs	<i>t</i>	Wiicoxon	Score	Wald	<i>R</i>	<i>R</i> [*]
15	4	3	0.0161	0.0216	0.0473	0.0469	0.0549	0.0465

SD: control (untreated) group standard deviation; Ratio SDs: ratio of standard deviations of treatment group vs. untreated group

^{*}, left trial empirical type I error

[†], right trial empirical type I error.

Table 3:

Evaluation empirical power of six methods for testing medians with sample sizes $n = 3, 5, 8$ and 10 and a one-sided type I error of $\alpha = 0.05$.

				Test Statistics					
means	n	SD	Ratio SDs	t	Wilcoxon	Score	Wald	R	R^*
10, 20	3	2	1	0.9987	0	0.9816	0.9985	1	0.9988
10, 20	3	2	2	0.9165	0	0.8913	0.9088	0.9958	0.9663
10, 20	3	2	3	0.7043	0	0.6881	0.7061	0.9610	0.8338
10, 20	3	3	1	0.9401	0	0.8481	0.9721	0.9929	0.9463
10, 20	3	3	2	0.6762	0	0.6670	0.7382	0.9282	0.7787
10, 20	3	3	3	0.4641	0	0.4815	0.5342	0.8190	0.6057
10, 20	3	4	1	0.7884	0	0.6934	0.8905	0.9430	0.8103
10, 20	3	4	2	0.4921	0	0.5079	0.6189	0.8025	0.6007
10, 20	3	4	3	0.3225	0	0.3690	0.4394	0.6872	0.4628
10, 20	5	2	1	1	1	1	1	1	1
10, 20	5	2	2	0.9996	0.9951	0.9999	0.9999	1	0.9999
10, 20	5	2	3	0.9861	0.9412	0.9920	0.9924	0.9972	0.9916
10, 20	5	3	1	0.9996	0.9903	0.9995	0.9998	0.9999	0.9995
10, 20	5	3	2	0.9745	0.8919	0.9804	0.9829	0.9920	0.9767
10, 20	5	3	3	0.8591	0.7173	0.8993	0.9052	0.9428	0.8903
10, 20	5	4	1	0.9867	0.9117	0.9863	0.9922	0.9933	0.9838
10, 20	5	4	2	0.8724	0.6956	0.8962	0.9116	0.9334	0.8804
10, 20	5	4	3	0.6913	0.5165	0.7538	0.7703	0.8299	0.7272
10, 20	8	2	1	1	1	1	1	1	1
10, 20	8	2	2	1	1	1	1	1	1
10, 20	8	2	3	0.9999	0.9994	0.9999	0.9999	0.9999	0.9999
10, 20	8	3	1	1	1	1	1	1	1
10, 20	8	3	2	0.9995	0.9963	0.9994	0.9994	0.9996	0.9994
10, 20	8	3	3	0.9878	0.9607	0.9864	0.9872	0.9910	0.9840
10, 20	8	4	1	0.9996	0.9984	0.9995	0.9996	0.9996	0.9993
10, 20	8	4	2	0.9883	0.9574	0.9839	0.9857	0.9878	0.9817
10, 20	8	4	3	0.9334	0.8593	0.9215	0.9263	0.9388	0.9104
10, 20	10	2	1	1	1	1	1	1	1
10, 20	10	2	2	1	1	1	1	1	1
10, 20	10	2	3	1	0.9998	1	1	1	1
10, 20	10	3	1	1	1	1	1	1	1
10, 20	10	3	2	0.9999	0.9994	0.9999	0.9999	0.9999	0.9999
10, 20	10	3	3	0.9984	0.9861	0.9971	0.9971	0.9978	0.9964
10, 20	10	4	1	1	0.9998	1	1	1	0.9999
10, 20	10	4	2	0.9982	0.9862	0.9967	0.9970	0.9976	0.9961

Test Statistics									
means	<i>n</i>	SD	Ratio SDs	<i>t</i>	Wilcoxon	Score	Wald	<i>R</i>	<i>R</i> *
10, 20	10	4	3	0.9804	0.9234	0.9675	0.9693	0.9725	0.9628

SD: control (untreated) group standard deviation; Ratio SDs: ratio of standard deviations of treatment group vs. untreated group

Table 4:

The tumor quadrupling times of the control and treated mice for the data presented in Table 1.

Tumor quadrupling time										
Group	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
Control	6.42	10.67	6.84	8.98	8.68	6.16	6.19	7.22	8.54	6.93
Treated	12.65	28.60	16.15	18.32	19.57	33.19	17.24	16.00	15.74	18.56

Table 5:

Two-sided p-value for testing the hypothesis of equal medians.

Test Statistics						
	<i>t</i>	Wilcoxon	Score	Wald	<i>R</i>	<i>R</i> *
p-value	< 0.001	< 0.001	0.002	0.002	< 0.001	< 0.001

Table 6:

Sample size calculations for various type I errors and powers.

			Median of Control(day)									
			10			20			30			
Median of Treatment(day)			20	30	40	30	40	50	40	50	60	
(α, β)	SD_C	SD_T										
(0.05, 0.2)	2	4	3	3	3	3	3	3	3	3	3	3
		6	3	3	3	4	3	3	4	3	3	
		8	4	3	3	5	3	3	5	3	3	
	3	6	4	3	3	4	3	3	4	3	3	
		9	5	3	3	5	3	3	6	3	3	
		12	6	3	3	7	3	3	9	4	3	
	4	8	5	3	3	5	3	3	6	3	3	
		12	6	3	3	8	3	3	9	4	3	
		16	8	3	3	11	4	3	13	4	3	
	2	4	3	3	3	3	3	3	3	3	3	
		6	3	3	3	3	3	3	3	3	3	
		8	3	3	3	4	3	3	4	3	3	
(0.1, 0.2)	3	6	3	3	3	3	3	3	3	3	3	
		9	3	3	3	4	3	3	4	3	3	
		12	4	3	3	5	3	3	6	3	3	
	4	8	3	3	3	4	3	3	4	3	3	
		12	5	3	3	6	3	3	6	3	3	
		16	6	3	3	8	4	3	9	3	3	
(0.05, 0.1)	2	4	3	3	3	3	3	3	3	3	3	
		6	4	3	3	4	3	3	5	3	3	
		8	5	3	3	6	3	3	6	3	3	
	3	6	4	3	3	5	3	3	5	3	3	
		9	6	3	3	7	3	3	7	3	3	
		12	7	3	3	10	4	3	11	4	3	
	4	8	6	3	3	6	3	3	7	3	3	
		12	8	3	3	10	4	3	12	4	3	
		16	10	4	3	14	5	3	16	5	3	
(0.1, 0.1)	2	4	3	3	3	3	3	3	3	3	3	
		6	3	3	3	3	3	3	4	3	3	
		8	4	3	3	4	3	3	5	3	3	
	3	6	3	3	3	4	3	3	4	3	3	
		9	4	3	3	5	3	3	6	3	3	
		12	6	3	3	7	3	3	8	3	3	

Median of Control(day)										
			10			20			30	
Median of Treatment(day)	20	30	40	30	40	50	40	50	60	
(α, β)	SD_C	SD_T								
	4	8	5	3	3	5	3	3	5	3
		12	6	3	3	8	3	3	9	3

SD_C : control (untreated) group standard deviation; SD_T : treatment group standard deviation; α : type I error; β : type II error.