Contemporary Clinical Trials Communications 15 (2019) 100360

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

journal homepage: www.elsevier.com/locate/conctc

Sample size calculation for small sample single-arm trials for time-to-event data: Logrank test with normal approximation or test statistic based on exact chi-square distribution?

Milind A. Phadnis

Department of Biostatistics, University of Kansas Medical Center, 3901 Rainbow Boulevard, Kansas City, 66160, KS, USA

ARTICLE INFO

Keywords:
Clinical trial
Exact test
Single-arm
Survival
Weibull

ABSTRACT

Background: Sample size calculations are critical to the planning of a clinical trial. For single-arm trials with time-to-event endpoint, standard software provides only limited options. The most popular option is the log-rank test. A second option assuming exponential distribution is available on some online websites. Both these approaches rely on asymptotic normality for the test statistic and perform well for moderate-to-large sample sizes.
Methods: As many new treatments in the field of oncology are cost-prohibitive and have slow accrual rates, researchers are often faced with the restriction of conducting single arm trials with potentially small-to-moderate sample sizes. As a practical solution, therefore, we consider the option of performing the sample size calculations using an exact parametric test with the test statistic following a chi-square distribution. Analytic results of sample size calculations from the two methods with Weibull distributed survival times are briefly compared using an example of a clinical trial on cholangiocarcinoma and are verified through simulations.
Results: Our simulations suggest that in the case of small sample phase II studies, there can be some practical benefits in using the exact test that could affect the feasibility, timeliness, financial support, and 'clinical novelty' factor in conducting a study. The exact test is a good option for designing small-to-moderate sample trials when accrual and follow-up time are adequate.
Conclusions: Based on our simulations for small sample studies, we conclude that a statistician should assess sensitivity of his calculations obtained through different methods before recommending a sample size to their collaborators.

1. Introduction

Two-arm randomized clinical trials are the gold standard in biomedical research as they allow performance assessment of a new experimental treatment relative to a standard control. However, there are situations where conducting a two-arm trial is not possible and a single-arm trial may be the preferred choice. For single-arm trials with a time-to-event endpoint, surprisingly few options for sample size calculation are available in literature or in standard software. The most popular option is the log-rank test [1] and its weighted versions. It has been used for sample size calculations by Finkelstein et al. [2], Kwak and Jung [3], Jung [4], Sun et al. [5] and more recently by Wu [6]. Likewise, sample size calculations for exponentially distributed survival times have been proposed by Lawless [7] (available as online calculators; see SWOG [8]). Both approaches rely on asymptotic normality of the test statistic and perform well for moderate-to-large sample sizes. As many new treatments in the field of oncology are cost-prohibitive and have slow accrual rates, researchers are often restricted to conducting single-arm trials with small-to-moderate sample sizes.

The sample size formula proposed by Wu [6] is based on the exact variance of the test statistic and hence is an improvement on the earlier versions of the logrank test. Wu [6] has mentioned in his concluding remarks that his one-sample logrank test is conservative when dealing with small samples and that the correctness of its use depends on the correct specification of the underlying distribution of the standard population. In this context, we bring to the reader's attention that a parametric method of calculating sample size for exponentially distributed times was first published by Epstein and Sobel [9]. This method uses a test statistic that follows a chi-square distribution. Later, Narula and Li [10] have shown how to extend the calculations to the case of gamma, Weibull, and Laplace distributions in the uncensored case. One important point to note is that an iterative search algorithm may be needed to calculate the sample size given the value the other fixed parameters using their approach and to avoid this Narula and Li

E-mail address: mphadnis@kumc.edu.

https://doi.org/10.1016/j.conctc.2019.100360
Received 1 February 2019; Received in revised form 29 March 2019; Accepted 5 April 2019
Available online 13 April 2019

Table 1
Number of events/sample size for exact vs Wu's method (administrative censoring adjustment by equation [4]) for different values of Weibull shape parameter β, accrual time a, and follow-up time f.

Study specific parameters: median time under H_0 = 2.5 months; median time under H_A = 3.75 months; alpha = 0.05 (one-sided); target power = 0.80;

Shape parameter	a	Method	f=1	f = 2	f = 4	f = 6	f = 9	f = 12
β = 0.50	a = 0	Exact	148/492	148/373	148/290	148/254	148/225	148/209
		Wu	144/478	145/365	147/287	148/254	149/227	151/212
	a = 3	Exact	148/350	148/307	148/262	148/238	148/216	148/203
		Wu	146/344	147/303	148/261	149/239	150/219	151/207
	a = 6	Exact	148/300	148/274	148/245	148/227	148/209	148/198
		Wu	147/297	147/273	149/245	150/229	151/213	152/203
	a = 9	Exact	148/272	148/255	148/232	148/218	148/204	148/194
		Wu	148/271	149/255	149/234	150/221	151/208	152/199
	a = 12	Exact	148/254	148/241	148/223	148/211	148/199	148/191
		Wu	149/254	149/242	150/226	151/215	152/204	152/196
	a = 15	Exact	148/241	148/230	148/216	148/206	148/196	148/188
		Wu	149/242	150/233	150/219	151/210	152/200	153/194
β = 0.75	a = 0	Exact	66/292	66/188	66/128	66/106	66/90	66/82
		Wu	62/273	63/179	66/105	66/105	68/92	69/85
	a = 3	Exact	66/169	66/139	66/97	66/97	66/86	66/79
		Wu	64/163	64/135	67/97	67/97	68/88	69/83
	a = 6	Exact	66/134	66/118	66/91	66/91	66/82	66/78
		Wu	65/131	66/117	68/93	68/93	68/85	70/82
	a = 9	Exact	66/117	66/106	66/87	66/87	66/80	66/76
		Wu	66/116	67/107	68/89	68/89	70/84	70/80
	a = 12	Exact	66/107	66/99	66/84	66/84	66/78	66/75
		Wu	67/107	67/100	69/87	69/87	70/82	71/80
	a = 15	Exact	66/100	66/94	66/81	66/81	66/77	66/74
		Wu	68/101	68/96	69/85	69/85	70/81	71/79
β = 1	a = 0	Exact	37/220	37/120	37/71	37/56	37/46	37/42
		Wu	34/199	35/112	36/69	38/56	39/48	40/45
		Lawless	184	103	64	52	45	42
	a = 3	Exact	37/103	37/79	37/59	37/50	37/44	37/41
		Wu	35/97	36/76	37/59	39/52	40/47	41/45
		Lawless	90	71	55	48	44	42
	a = 6	Exact	37/75	37/64	37/53	37/47	37/42	37/40
		Wu	36/73	37/64	38/54	39/49	41/46	41/44
		Lawless	68	59	50	46	43	41
	a = 9	Exact	37/63	37/56	37/49	37/45	37/41	37/40
		Wu	37/63	38/58	39/51	40/48	41/45	42/44
		Lawless	59	53	47	44	42	41
	a = 12	Exact	37/56	37/52	37/46	37/43	37/41	37/39
		Wu	39/58	39/54	40/49	41/47	42/45	42/44
		Lawless	54	50	46	43	42	41
	a = 15	Exact	37/52	37/49	37/45	37/42	37/40	37/39
		Wu	40/55	40/52	40/48	41/46	41/44	42/44
		Lawless	51	48	45	43	41	41
β = 1.25	a = 0	Exact	24/193	24/89	24/46	24/34	24/28	24/26
		Wu	21/166	21/79	23/44	24/34	26/30	28/29
	a = 3	Exact	24/72	24/52	24/36	24/30	24/27	24/25
		Wu	22/66	23/49	24/36	26/32	27/29	28/29
	a = 6	Exact	24/48	24/40	24/32	24/28	24/26	24/25
		Wu	24/47	24/40	25/33	27/31	27/29	28/29
	a = 9	Exact	24/39	24/35	24/30	24/27	24/26	24/25
		Wu	25/40	25/36	26/32	27/30	28/29	28/29
	a = 12	Exact	24/35	24/32	24/29	24/27	24/25	24/25
		Wu	26/37	26/34	27/31	28/30	28/29	29/29
	a = 15	Exact	24/32	24/30	24/28	24/26	24/25	24/24
		Wu	26/35	27/33	27/31	28/30	28/29	29/29
β = 1.50	a = 0	Exact	16/176	16/68	16/30	16/22	16/18	16/17
		Wu	14/151	14/61	16/30	17/23	19/21	21/21
	a = 3	Exact	16/52	16/35	16/23	16/19	16/17	16/17
		Wu	15/48	16/34	17/25	19/22	20/21	21/21
	a = 6	Exact	16/32	16/26	16/21	16/18	16/17	16/17
		Wu	17/33	17/28	18/23	19/21	20/21	21/21
	a = 9	Exact	16/26	16/23	16/19	16/18	16/17	16/17
		Wu	18/28	18/25	19/22	20/21	21/21	21/21
	a = 12	Exact	16/23	16/21	16/18	16/17	16/17	16/17
		Wu	19/26	19/24	20/22	20/21	21/21	21/21
	a = 15	Exact	16/21	16/20	16/18	16/17	16/17	16/16
		Wu	19/25	19/23	20/22	20/21	21/21	21/21

a = accrual time in months, f = follow-up time in months.
'Exact' refers to the exact calculations done using the chi-square distribution.
Note: The calculations given by Lawless [6] were done using an online calculator by SWOG [7] and only show the total sample size and not the number of events.

[10] also mention five different closed-form solutions based on a normal approximation. Surprisingly, popular statistics software does not have options for such calculations though PASS [11] has incorporated the logrank calculations of Wu [6].

2. Methods

The Weibull distribution is a two-parameter distribution with its pdf given by:

$$f(t) = \frac{\beta}{\theta^{\beta}} t^{\beta-1} e^{-(t/\theta)^{\beta}} \qquad \theta, \beta > 0, \ t > 0 \tag{1}$$

here θ is a scale parameter and β is a shape parameter that determines the shape of the hazard function ($\beta > 1$ gives hazard that is increasing over time, and, $\beta < 1$ gives hazard that is decreasing over time with $\beta = 1$ representing the special case of exponential distribution with constant hazard).

With modern computational tools, we can write an efficient SAS program for an iterative approach using the formula given in Narula and Li [10] accounting for administrative censoring. That is, following Narula and Li [10], the problem of calculating sample size n (without censoring) in the Weibull case to test the hypothesis H_0: $\theta = \theta_0$ against the alternative H_A: $\theta = \theta_1 (< \theta_0)$ at level of significance α and probability of type II error γ reduces to solving for δ using

$$\delta = \chi^2_{1-\gamma}(v)/\chi^2_{\alpha}(v) \tag{2}$$

with $\delta = \theta_0/\theta_1$ and $v = 2n$. Our program then adjusts their method for administrative censoring accounting for study-specific accrual and follow-up times in the following way:

Assuming a uniform accrual, the censoring distribution function $G(t)$ is given by

$$G(t) = \begin{Bmatrix} 1 & \text{if } t \le f \\ \frac{a+f-t}{a} & \text{if } f \le t \le a+f \\ 0 & \text{otherwise} \end{Bmatrix} \tag{3}$$

where a and f are the accrual and follow-up time respectively. Then the probability that a subject experiences a failure during the study is given by

$$d = \int_0^{\infty} G(t).\, f_1(t) dt \tag{4}$$

where $f_1(t)$ is $f(t)$ with $\theta = \theta_1$. Dividing the number of events by d gives the sample size adjusted for administrative censoring. Alternatively, d can be calculated using Simpson's rule by

$$d = 1 - \frac{1}{6}\{S_1(f) + 4S_1(f + 0.5a) + S_1(f + a)\} \tag{5}$$

where $S_1(t)$ is the survival function of the Weibull with $\theta = \theta_1$

For the Weibull, this allows comparison with Wu [6] and for the special case of the exponential, this allows comparison with Lawless [7]. To do so, we consider a real-life example about designing a phase II clinical trial for treating patients suffering from chemotherapy refractory advanced metastatic biliary cholangiocarcinoma, a “rare” but aggressive neoplasm. Such patients have metastatic disease and undergo an initial treatment followed by a second-line treatment which has a progression-free survival (PFS) rate of 5–10% by 1 year. Oncologists are therefore working towards improving the PFS by using new combination therapies. Historically, published literature mentions a median PFS of 2.5 months with an IQR of around 2–5 months. Due to dismal survival rates, they consider an improvement in 25th, 50th and 75th percentile of PFS by a factor of 1.5 as clinically meaningful and holding promise for future large sample studies. The rarity of disease poses recruitment problems with typical accrual rates being approximately 12–15 patients/year. Based on financial and administrative limitations, researchers envision a study with an accrual time of 2 years and follow-up time of 3 years. Loss to follow-up is anticipated to be 15–20%. It should be noted that as the researchers hypothesize a consistent improvement in PFS for all quantiles of the survival curve - of the historical controls by a factor of 1.5, the Weibull distribution is a good choice for performing the sample size calculations as is evident from the definition of δ in (1). Following Wu [6], the shape parameter β for the Weibull is estimated from the historical controls as 1.25 (increasing hazard).

3. Results

For the study-specific design features in this example, Table 1 shows the comparisons between the two methods for various values of the shape parameter β ranging from 0.5 to 1.5. The conservative nature of the logrank test can be studied by observing how in Table 1 the sample sizes vary as a function of accrual and follow-up time (keeping other design parameters fixed) for the different values of the shape parameter. When $\beta = 0.5$, we see that Wu's logrank test gives smaller sample sizes compared to the exact method only when both a and f are small in magnitude. On the other hand, as either a or f increases, the exact test yields smaller sample sizes. This general pattern is even more accentuated as β increases from 0.5 to 1.5. In fact, for $\beta = 1.5$, only $a \le 3$ and $f \le 3$ allow the logrank test to have smaller sample sizes than the exact method. As in the cholangiocarcinoma example under consideration, researchers hypothesize improvement in median PFS by a factor of 1.5, small values of a and f are impractical as based on the accrual rate, very few patients can participate in this study.

For this example, where $\beta = 1.25$ is chosen, the exact method gives a sample size of 24 when $a \ge 15$ and $f \ge 12$ whereas the logrank test finds a lower bound at 29 no matter how big a and f are chosen. That is, even with the flexibility to follow patients for a hypothetically large amount of time and thereby observe all events, Wu's method does not go below a threshold value of 29 events. Through simulations (we used 10,000 simulations) using the Weibull distribution, it can be shown that with large follow-up times, 80% power is achieved only with 24 subjects and the exact method is analytically able to yield a sample size of 24. By adopting the popular ad-hoc method of inflating the sample size to accommodate drop-outs (conservatively assuming they provide no extra information), the adjusted sample size can be calculated as 24/0.8 = 30. That is, if the researcher's ‘optimistic’ estimate of accruing 15 patients/year is true, it appears likely that this study can be completed within the stipulated timeframe. A similar ad-hoc adjustment for Wu's method would require 37 patients to be enrolled, which is outside the practical timeframe of the study. However, by assuming that drop-outs occur randomly over the study period (assuming a uniform distribution), for an anticipated drop-out rate of 20%, our simulations gave a sample size of $n = 28$ with 80.6% power. Thus, a combination of analytical calculations using the exact method aided by further simulations can enable a statistician to design a small sample trial with adequate power. If additional information from similar such studies is available, a statistician can also incorporate other drop-out mechanisms (such as exponentially distributed drop-out times with a specific mean).

Similar comparisons can be performed for other values of β such as $\beta = 0.75$ (decreasing hazard) and $\beta = 1$ (constant hazard – exponential distribution). In the case of $\beta = 1$, it can be seen that the normal approximation proposed by Lawless [7] gives smaller n than the exact method for small-to-moderate values of follow-up time. However, with large values of follow-up time, this is no longer the case and the normal approximation cannot yield sample sizes below $n = 41$. Through simulations (we used 10,000 simulations) using the exponential distribution, it can be shown that with large follow-up times, 80% power is achieved only with 37 subjects and the exact method is analytically able to yield a sample size of 37 whereas the normal approximation method and the logrank test yield sample sizes of 40 and 43 respectively.

Though the exact method yields smaller sample sizes for many situations, it is necessary to assess whether or not the empirical type I

Table 2
Evaluation of empirical type I error and empirical power using the exact method for the cholangiocarcinoma study with H_0: $t_{med} \leq 2.5$ months and effect size equal to improvement in median time by a factor of 1.5 - using 10,000 simulations (nominal type I error 5%, target power 80%).

Shape parameter β	Accrual time a in months	Follow-up time f in months	Total sample size n	Average # events observed under H_0	Empirical type I error	Average # events observed under H_A	Empirical power
$\beta = 0.50$	$a = 0$	$f = 1$	492	174.80	0.0453	148.17	0.8245
		$f = 3$	321	170.73	0.0452	148.25	0.8259
		$f = 6$	254	167.21	0.0457	148.33	0.8247
		$f = 12$	209	163.18	0.0472	148.50	0.8193
	$a = 3$	$f = 1$	350	171.64	0.0454	148.39	0.8235
		$f = 3$	280	168.57	0.0434	148.07	0.8169
		$f = 6$	238	166.01	0.0477	148.43	0.8186
		$f = 12$	203	162.31	0.0446	148.41	0.8169
	$a = 6$	$f = 1$	300	169.53	0.0448	148.44	0.8218
		$f = 3$	257	166.99	0.0466	148.11	0.8214
		$f = 6$	227	165.15	0.0456	148.59	0.8139
		$f = 12$	198	161.44	0.0485	148.22	0.8191
	$a = 12$	$f = 1$	254	166.59	0.0458	148.53	0.8141
		$f = 3$	231	164.95	0.0461	148.37	0.8179
		$f = 6$	211	162.78	0.0441	148.01	0.8096
		$f = 12$	191	160.49	0.0472	148.38	0.8099
$\beta = 0.75$	$a = 0$	$f = 1$	291	85.72	0.0452	66.13	0.8425
		$f = 3$	149	81.74	0.0454	66.12	0.8366
		$f = 6$	106	78.21	0.0512	66.55	0.8425
		$f = 12$	82	73.34	0.0494	66.36	0.8227
	$a = 3$	$f = 1$	169	82.41	0.0513	66.17	0.8432
		$f = 3$	122	79.86	0.0467	66.43	0.8403
		$f = 6$	97	76.82	0.0471	66.62	0.8296
		$f = 12$	79	72.17	0.0472	66.01	0.8210
	$a = 6$	$f = 1$	134	80.26	0.0448	66.49	0.8418
		$f = 3$	108	78.10	0.0484	66.49	0.8394
		$f = 6$	91	75.54	0.0459	66.57	0.8235
		$f = 12$	78	72.39	0.0499	66.84	0.8230
	$a = 12$	$f = 1$	106	77.05	0.0490	66.56	0.8347
		$f = 3$	93	75.06	0.0460	66.13	0.8234
		$f = 6$	84	73.83	0.0485	66.70	0.8284
		$f = 12$	75	71.06	0.0505	66.58	0.8236
$\beta = 1$	$a = 0$	$f = 1$	220	53.41	0.0475	37.22	0.8509
		$f = 3$	87	49.18	0.0418	37.08	0.8405
		$f = 6$	56	45.40	0.0439	37.55	0.8343
		$f = 12$	42	40.51	0.0535	37.46	0.8419
	$a = 3$	$f = 1$	103	50.09	0.0471	39.33	0.8413
		$f = 3$	67	47.24	0.0478	37.52	0.8436
		$f = 6$	50	43.60	0.0485	37.36	0.8359
		$f = 12$	41	40.02	0.0446	37.60	0.8236
	$a = 6$	$f = 1$	75	47.35	0.0490	37.40	0.8377

(*continued on next page*)

Table 2 (*continued*)

		$f = 3$	57	44.93	0.0455	37.27	0.8299
	$a = 6$	$f = 6$	47	42.70	0.0489	37.66	0.8359
		$f = 12$	40	39.32	0.0482	37.41	0.8250
		$f = 1$	56	43.77	0.0453	37.38	0.8257
		$f = 3$	49	42.85	0.0446	37.73	0.8260
	$a = 12$	$f = 6$	43	40.68	0.0481	37.35	0.8277
		$f = 12$	39	38.60	0.0487	37.33	0.8201
		$f = 1$	193	38.32	0.0513	24.06	0.8843
		$f = 3$	59	34.33	0.0494	24.15	0.8664
	$a = 0$	$f = 6$	34	29.73	0.0448	24.25	0.8545
		$f = 12$	26	25.80	0.0494	24.67	0.8415
		$f = 1$	72	34.96	0.0480	24.25	0.8596
		$f = 3$	42	31.61	0.0512	24.17	0.8595
	$a = 3$	$f = 6$	30	27.93	0.0505	24.08	0.8491
$\beta = 1.25$		$f = 12$	25	24.91	0.0490	24.17	0.8218
		$f = 1$	48	31.73	0.0467	24.13	0.8516
		$f = 3$	35	29.48	0.0445	24.13	0.8419
	$a = 6$	$f = 6$	28	26.81	0.0524	24.09	0.8372
		$f = 12$	25	24.94	0.0490	24.43	0.8208
		$f = 1$	35	28.62	0.0506	24.58	0.8438
		$f = 3$	30	27.49	0.0494	24.49	0.8400
	$a = 12$	$f = 6$	27	26.40	0.0547	24.81	0.8409
		$f = 12$	25	24.97	0.0489	24.68	0.8228
		$f = 1$	176	28.41	0.0506	16.05	0.8803
		$f = 3$	41	24.58	0.0487	16.09	0.8731
	$a = 0$	$f = 6$	22	20.33	0.0483	16.60	0.8537
		$f = 12$	17	16.99	0.0449	16.67	0.8224
		$f = 1$	52	25.23	0.0495	16.28	0.8719
		$f = 3$	28	22.30	0.0451	16.52	0.8609
	$a = 3$	$f = 6$	19	18.39	0.0456	16.21	0.8401
$\beta = 1.5$		$f = 12$	17	17.00	0.0439	16.84	0.8209
		$f = 1$	32	21.88	0.0464	16.21	0.8560
		$f = 3$	23	20.28	0.0453	16.59	0.8442
	$a = 6$	$f = 6$	18	17.69	0.0507	16.30	0.8288
		$f = 12$	17	17.00	0.0446	16.90	0.8216
		$f = 1$	23	19.28	0.0470	16.61	0.8358
		$f = 3$	19	17.86	0.0495	16.11	0.8282
	$a = 12$	$f = 6$	17	16.85	0.0442	16.14	0.8181
		$f = 12$	17	17.00	0.0446	16.95	0.8222

error rate and empirical power are close to their nominal values. To do this, a simulation study (with 10,000 simulations) was conducted with the study-specific design parameters of the cholangiocarcinoma study. For varying values of a and f for β ranging from 0.5 to 1.5, time-to-event data was simulated with the sample sizes calculated by the exact method and the results are displayed in Table 2. From this table it can be see that for almost all scenarios the empirical type I error rates were close to the nominal 5% alpha level. Likewise, empirical power always slightly exceeded the target 80% power. Except for seemingly impractical values such as $a = 0$ (all subjects are available for recruitment at the start) and $f = 1$ (a very short follow-up time for the study under consideration), there was not much difference in the magnitude by which empirical power exceeded the target power. As the results displayed in Table 2 are in the context of a specific example with a fixed effect size, a similar evaluation of empirical type I error and power was conducted for the hypothetical example discussed in Wu [6] where in

Table 3
Comparing empirical type I error (α) and empirical power with Wu's method for the example in Wu (2015 – accrual time = 3, follow-up time = 1, nominal type I error 5%, nominal power 90%)) using the exact adjustment for administrative censoring given by equation (4) – number of simulations = 10,000.

β	Method	$\delta = 1.2$			$\delta = 1.4$			$\delta = 1.6$			$\delta = 1.8$			$\delta = 2$		
		n	α	Power	n	α	Power	n	α	Power	n	α	Power	n	α	Power
$\beta = 0.1$	Exact	551	0.0458	0.9195	180	0.0503	0.9294	104	0.0495	0.9432	73	0.0472	0.9485	58	0.0492	0.9640
	Wu	534	0.048	0.9030	169	0.044	0.907	93	0.044	0.909	63	0.041	0.911	47	0.041	0.909
$\beta = 0.25$	Exact	504	0.0459	0.9191	164	0.0468	0.9297	94	0.0507	0.9430	66	0.0506	0.9480	52	0.0510	0.9600
	Wu	492	0.047	0.904	156	0.046	0.907	85	0.045	0.908	58	0.042	0.913	44	0.041	0.915
$\beta = 0.5$	Exact	438	0.0450	0.9177	141	0.0434	0.9258	81	0.0520	0.9382	56	0.0486	0.9489	44	0.0529	0.9763
	Wu	432	0.047	0.905	137	0.046	0.909	75	0.042	0.913	50	0.041	0.912	38	0.041	0.915
$\beta = 1$	Exact	351	0.0476	0.9096	110	0.0460	0.9158	62	0.0478	0.9324	42	0.0511	0.9434	33	0.0460	0.9474
	Wu	356	0.047	0.907	112	0.044	0.912	61	0.042	0.916	41	0.040	0.921	31	0.040	0.925
$\beta = 2$	Exact	289	0.0470	0.9092	87	0.0485	0.9145	47	0.0455	0.9186	31	0.0467	0.9285	23	0.0475	0.9325
	Wu	306	0.046	0.910	97	0.042	0.922	53	0.040	0.929	36	0.038	0.938	27	0.038	0.942
$\beta = 5$	Exact	267	0.0495	0.9024	79	0.0475	0.9036	42	0.0487	0.9176	27	0.0464	0.9121	20	0.0473	0.9241
	Wu	288	0.046	0.912	91	0.042	0.925	50	0.039	0.935	34	0.040	0.943	25	0.036	0.943

Note: In Wu's notation, $\delta = (m_1/m_0)^{\beta}$ where $m_0 = 1$ and m_1 are the median times under the null and alternate hypotheses respectively

$a = 3$ and $f = 1$ was kept fixed, but the effect size was varied from small to large, for β = 0.1, 0.25, 0.5, 1, 2, and 5. The results of these simulations are shown in Table 3 with target power at 90% and median time under the null hypothesis fixed at 1. Here too, for all scenarios, empirical type I error was close to the nominal level, and likewise, empirical power always slightly exceeded the 90% target level. For $\beta \leq 1$, the exact method yielded somewhat higher empirical power compared to Wu's method, while the converse was true for $\beta > 1$. Though at first glance Table 3 suggests that the exact method yields smaller sample sizes than the logrank test (with both methods having comparable empirical type I error and power) only for $\beta \geq 1$, it should be noted that the combination of $a = 3$ and $f = 1$ represent values that are quite small compared to the magnitude of the hypothesized improvement in median lifetime under the alternate hypothesis. For example, the combination of $\beta = 0.25$ and $\delta = 1.6$ indicates that the median under the alternate hypothesis is 6.55 times the median under the null (which is fixed at 1). Likewise, the combination of $\beta = 0.10$ and $\delta = 2$ indicates that the median under the alternate hypothesis is 1024 times the median under the null (which is fixed at 1). Thus, in such scenarios, it would be impractical to choose small values of a and f. As the values of a and f increase, compared to the logrank test, the exact method gives smaller sample sizes for $\beta \geq 0.5$ and same sample sizes for $\beta = 0.25$ (see results in Table 4). Though the logrank test gives smaller sample sizes for $\beta = 0.1$, such a small value of the shape parameter would require a very strong justification in a real life clinical trial.

For all results discussed in this section we used equation (4) to adjust for administrative censoring. For all scenarios shown in Table 3, we got very similar sample sizes (mostly same, sometimes differing by 1, rarely differing by 2) whether we used equation (4) or Simpson's rule mentioned in equation (5) except in the case when $\beta = 5$. In this case, adjustment by Simpson's rule yielded sample sizes of 284, 84, 44, 27 and 20 for δ= 1.2, 1.4, 1.6, 1.8 and 2 respectively. We thus recommend adjusting for administrative censoring using equation (4).

4. Conclusion

For large phase II or III trials, it would not make much difference if any of these three methods were used for sample size calculations. However, in the case of small sample phase II studies, there could be practical differences that would affect the feasibility, timeliness, financial support, and 'clinical novelty' factor (the challenge faced by clinicians to be the first to conduct a clinical trial using a novel idea) of the study. Additionally, both Wu's method and Lawless' method do not yield sample sizes below a certain threshold no matter how long the accrual and follow-up times are. On the other hand, the exact method does not suffer from this shortcoming. As for the concluding remark by Wu [6] about correctly specifying the underlying distribution, we fully agree with this assertion. In this context, we provide two insights. The first is the possibility to extend the framework presented in Narula and Li [10] to using a generalized gamma distribution (a family of distributions with exponential, Weibull, lognormal, gamma, inverse-Weibull, and inverse-gamma, as special cases) to model the historical controls in the case where definitive well-established large sample studies have been conducted on such historical controls and then use the estimates of these parameters to design the current single-arm study. Just like the gamma, Weibull and Laplace distributions, the sample size calculation could be done using the framework presented in Narula and Li [10] as it would allow calculation of nk (in place of n in equation (1)) where $k > 0$ is the extra shape parameter estimated from historical controls. Then dividing by k would yield the sample size required for current single-arm study. The second insight is to conduct simulations during the design phase of a study to assess how, in the case of Weibull, partial information available from historical data impacts the sample size calculation for the current study. Often, historical control information is available in published literature in the form of a Kaplan Meier curve, or as point estimates of median and/or interquartile range. It would therefore be interesting to assess the sensitivity of using this partial information from prior studies on the sample size calculations of the current study. Additionally, the role of random censoring due to drop-outs/loss to follow-up needs to be addressed in a comprehensive manner.

In some disease-specific areas, single-arm trials are inevitable. For example, conducting a randomized controlled trial (RCT) may prove impractical when recruiting a small target population (rare disease). Ethical or practical considerations may dictate that researchers conduct a single-arm trial with all enrolled patients receiving the experimental treatment. Another example of a single-arm trial is the "window-of-opportunity" trial in which patients diagnosed with a disease are

Table 4
Sample size comparison of exact vs Wu's method for the example in Wu (2015) – with a = 18, f = 18 (all other parameters same as Table 3).

β	Method	Total Sample Size n				
		$\delta = 1.2$	$\delta = 1.4$	$\delta = 1.6$	$\delta = 1.8$	$\delta = 2$
$\beta = 0.1$	Exact	467	151	87	61	48
	Wu	457	145	79	53	40
$\beta = 0.25$	Exact	352	112	63	43	34
	Wu	355	112	61	41	31
$\beta = 0.5$	Exact	272	82	44	30	22
	Wu	288	90	49	33	24
$\beta = 1$	Exact	258	76	40	26	19
	Wu	279	88	48	32	24
$\beta = 2$	Exact	257	75	39	25	18
	Wu	279	88	48	32	24
$\beta = 5$	Exact	257	75	39	25	18
	Wu	279	88	48	32	24

Note: In Wu's notation, $\delta = (m_1/m_0)^{\beta}$ where $m_0 = 1$ and m_1 are the median times under the null and alternate hypotheses respectively

awaiting subsequent surgery and can be enrolled only in this waiting period. The number of such patients may be small requiring a single-arm study. Likewise, there are situations where the standard drug has been so well studied and documented that researchers may consider published results about its performance as reliable historical data. In this case, they may prefer a single-arm trial. Further, in the case of some rare diseases, it may be ethical to treat only those subjects with a new experimental drug who have stabilized on the standard-of-care treatment, thereby warranting a single-arm study.

Overall, we feel that statisticians should be aware of the issues we have discussed in planning a single-arm trial for time-to-event data. Our calculations and simulation study suggests that the exact method is a good option for designing small-to-moderate sample trials when accrual and follow-up time is adequate. Thus, in cancer trials where accrual rates are low, it may be necessary to have longer accrual times and in such trials, the exact method may be preferred as it yields smaller sample sizes with sufficient power while maintaining the type I error rate. The statistician should strive to use the most appropriate method considering various practical considerations in consultation with the researchers. Especially, in the case of small sample studies, they should assess sensitivity of their calculations obtained through different methods.

Acknowledgement

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.conctc.2019.100360.

Disclosure

The author has no relevant conflicts of interest to declare.

References

[1] N.E. Breslow, Analysis of survival data under the proportional hazards model, Int. Stat. Rev. 43 (1975) 44–58.
[2] D.M. Finkelstein, A. Muzikansky, D.A. Schoenfeld, Comparing survival of a sample to that of a standard population, J. Natl. Cancer Inst. 95 (2003) 1434–1439.
[3] M. Kwak, S.H. Jung, Phase II clinical trials with time-to-event endpoints: optimal two-stage designs with one-sample log-rank test, Stat. Med. 33 (2014) 2004–2016.
[4] S.H. Jung, Randomized Phase II Cancer Clinical Trial, CRC Press: Chapman and Hall, 2013.
[5] X. Sun, P. Peng, D. Tu, Phase II cancer clinical trials with a one-sample log-rabk test and its corrections based on the Edgeworth expansion, Contemp. Clin. Trials 32 (2011) 108–113.
[6] J. Wu, Sample size calculation for the one-sample log-rank test, Pharmaceut. Stat. 14 (2015) 26–33.
[7] J.F. Lawless, Statistical Models and Methods for Lifetime Data, second ed., John Wiley and Sons, New York, 2003.
[8] SWOG: Cancer Research Network – Cancer Research and Biostatistics, Available at: https://stattools.crab.org/Calculators/oneArmSurvivalColored.html.
[9] B. Epstein, M. Sobel, Life testing, J. Am. Stat. Assoc. 48 (263) (1953) 486–582.
[10] S.C. Narula, F.S. Li, Sample size calculations in exponential life testing, Technometrics 17 (2) (1975) 229–231.
[11] PASS 14 Power Analysis and Sample Size Software, NCSS, LLC, Kaysville, Utah, USA, 2015 ncss.com/software/pass.