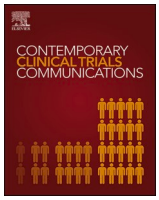
Contemporary Clinical Trials Communications 41 (2024) 101344

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

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



Sample size considerations for single-arm clinical trials with time-to-event endpoint using the gamma distribution

Junqiang Dai *, Jianghua He, Milind A. Phadnis

Department of Biostatistics and Data Science, University of Kansas Medical Center, Kansas City, KS, USA

ARTICLE INFO

Keywords:
Clinical trial
Single-arm
Survival
Sample size
Gamma distribution

ABSTRACT

Background: Time-to-event (TTE) endpoints are evaluated as the primary endpoint in single-arm clinical trials; however, limited options are available in statistical software for sample size calculation. In single-arm trials with TTE endpoints, the non-parametric log-rank test is commonly used. Parametric options for single-arm design assume survival times follow exponential distribution or Weibull distribution.

Methods: The exponential- or Weibull-distributed survival time assumption does not always reflect hazard pattern of real-life diseases. We therefore propose gamma distribution as an alternative parametric option for designing single-arm studies with TTE endpoints. We outline a sample size calculation approach using gamma distribution with a known shape parameter and explain how to extract the gamma shape estimate from previously published resources. In addition, we conduct simulations to assess the accuracy of the extracted gamma shape parameter and to explore the impact on sample size calculation when survival time distribution is misspecified.

Results: Our simulations show that if a previously published study (sample sizes $\geq$ 60 and censoring proportions $\leq$ 20 %) reported median and inter-quartile range of survival time, we can obtain a reasonably accurate gamma shape estimate, and use it to design new studies. When true survival time is Weibull-distributed, sample size calculation could be underestimated or overestimated depending on the hazard shape.

Conclusions: We show how to use gamma distribution in designing a single-arm trial, thereby offering more options beyond the exponential and Weibull. We provide a simulation-based assessment to ensure an accurate estimation of the gamma shape and recommend caution to avoid misspecification of the underlying distribution.

1. Introduction

In designing phase II trials, many methods are available for two-arm randomized designs with a dichotomous (yes/no) tumor response as a primary endpoint. With the rapid evolution in oncology drug development, this dominant paradigm has been challenged in two ways [1]. First, the previously accepted primary endpoint of dichotomous tumor response fails to predict survival benefits in many diseases such as lung, colon, breast, and renal cancers [2–4], or tumor response is difficult to measure in diseases such as glioblastoma and prostate cancer. Second, investigators cannot conduct the gold standard of randomized two-arm clinical trial design due to practical constraints, such as slow patient accrual in drug development for rare diseases. In such scenarios, the utilization of a single-arm phase II design with a time-to-event (TTE) endpoint could be considered, employing a natural history or historical control group. This approach, recognized by the FDA, acknowledges that a control group incorporating real-world evidence (RWE) can be deemed a valid comparison in evaluating treatment efficacy [5,6].

To design a single-arm study with a TTE endpoint, options for sample size calculation are very limited in literature and standard statistical software. In literature, the most common non-parametric methods for sample size calculation using the log-rank test and its weighted versions are proposed by several researchers (Breslow [7], Finkelstein et al. [8], Kwak and Jung [9], Jung [10] and Sun et al. [11]). A frequently used parametric sample size calculation method was proposed by Lawless [12], which assumes an exponentially distributed survival time, and it's adopted in the free web-based calculator by SWOG [13]. However, in commercial software like PASS [14] and nQuery [15], only two other approaches, a log-rank test proposed by Wu [16] and an exact parametric approach proposed by Phadnis [17], are implemented and available for sample size calculation for single-arm trials with TTE endpoints. Wu's [16] version of the log-rank test assumes survival times follow a Weibull distribution and calculates the sample size formula using the exact variance of the test statistic. Phadnis' [17] method also

* Corresponding author. 3901 Rainbow Blvd, Kansas City, KS, 66160, USA.
E-mail address: jdai2@kumc.edu (J. Dai).

https://doi.org/10.1016/j.conctc.2024.101344
Received 6 November 2023; Received in revised form 19 July 2024; Accepted 27 July 2024
Available online 2 August 2024

assumes Weibull distributed survival time and extends the exact parametric approach of Narula and Li [18] with a known point estimate of the Weibull shape parameter.

Compared to exponential distributed survival time assuming constant hazard, Weibull distribution provides the flexibility to model increasing or decreasing hazard patterns. However, there are limitations to using the Weibull distribution because the Weibull distribution allows hazards to increase to infinity (without an upper bound) or decrease to zero (without a lower bound) over time, and in real-life diseases, this hazard pattern may not hold. For example, pancreatic carcinoma, currently the third leading cause of cancer-related death in the USA, carries a dismal prognosis with a median survival time of 3–6 months in those untreated. Standard care has been found to have only a modest beneficial impact on advanced-stage patients [19]. The mortality rate among cancer patients is significantly increased with the presence of Venous thromboembolism (VTE). Some studies have shown evidence that pancreatic cancer patients who develop thromboembolism have worse survival [20,21] and PFS [22] compared to those without VTE. In the pancreatic cancer patient population, the survival functions for those with and without VTE differ significantly. When designing a single-arm trial with a TTE endpoint for pancreatic cancer patients without VTE or a disease where the risk stabilizes after a specific period, using a Weibull distribution, which allows hazard increases to infinity in a short study period, might not be appropriate.

Similar to Weibull distribution, gamma distribution provides the flexibility of modeling increasing, constant, and decreasing hazard scenarios. However, unlike Weibull, gamma distribution constrains the increasing hazard to approach a constant value rather than infinity and the decreasing hazard to approach a finite value instead of zero. When modeling data for diseases where the Weibull assumption of infinite hazard escalation is questionable, the gamma-distributed survival time may more accurately capture real-life phenomena than the Weibull distribution, since it caps the increasing or decreasing hazard to a finite constant.

This paper is focused on the following objectives: (1) Assuming gamma-distributed survival time, we provide a parametric sample size calculation formula adjusting for administrative censoring rates and probability of events occurring given pre-specified type I error; (2) We explain an approach of extracting a point estimate of gamma shape using previously published results, and assess the accuracy of the extracted gamma shape parameter estimate under various scenarios through simulations; (3) We explore the impact of misspecification of the survival time distribution on sample size calculation, by comparing calculated sample sizes using Weibull approach and gamma approach when underlying survival time is Weibull distributed.

2. Methods

2.1. Sample size formula

The probability density function (PDF) for a two-parameter gamma distribution is

$$f(t)=\frac{1}{\Gamma(k)\theta}\left(\frac{t}{\theta}\right)^{k-1}e^{-\frac{t}{\theta}};k,\theta>0,t\geq 0 \tag{1}$$

where θ is the scale parameter, and k is the shape parameter.

The shape parameter k determines the shape of the gamma hazard function: when $k>1$, the hazard increases over time to a constant; when $k<1$, the hazard decreases over time to a constant; when $k=1$ the hazard is a constant (reduces to exponential distribution). The median of the gamma distribution can be expressed as

$$M=\theta\gamma^{-1}\left(k,\frac{\Gamma(k)}{2}\right) \tag{2}$$

Here, $\gamma^{-1}\left(k,\frac{t}{\theta}\right)$ is the inverse of the lower incomplete gamma function, and it does not have a closed solution. However, if the gamma shape parameter is known and fixed, it's straightforward to show that testing the hypotheses $H_0:M=M_0$ versus $H_a:M=M_1$ is equivalent to testing $H_0:\theta=\theta_0$ versus $H_a:\theta=\theta_1$. Thus, sample size calculation for studies with TTE endpoint using gamma distribution with a known shape can be achieved in the following steps using iterative procedures.

(1) Find the number of events needed to test the hypotheses $H_0:M=M_0$ versus $H_a:M=M_1$. Narula and Li [18] have shown that, with gamma-distributed survival time and a known shape parameter k, given significance level α and type-II error rate β, calculation of the number of events E needed for testing the hypotheses $H_0:\theta=\theta_0$ versus $H_a:\theta=\theta_1$ reduces to solving for δ using

$$\delta=\chi^2_{1-\beta}(\nu)\Big/\chi^2_{\alpha}(\nu) \tag{3}$$

with $\delta=\theta_0/\theta_1$ and $\nu=2Ek$. Here $\chi^2_{1-\beta}(\nu)$ and $\chi^2_{\alpha}(\nu)$ are the $(100^*\ \beta)^{\text{th}}$ and $(100^*\ (1-a))^{\text{th}}$ percentile of the Chi-square distribution with ν degrees of freedom.

(2) Compute the required sample size by adjusting for administrative censoring with pre-specified study accrual and follow time. Assume patient accrual follows a uniform distribution with accrual time a and follow-up time f. The censoring distribution function $G(t)$ is

$$G(t)=\begin{cases}1, t\leq f\\ \dfrac{a+f-t}{a}, f\leq t\leq a+f\\ 0, \text{otherwise}\end{cases} \tag{4}$$

The probability of one subject having an event during the study can be expressed as

$$d=\int_0^{\infty}G(t)f_1(t)dt \tag{5}$$

where $f_1(t)$ is $f(t)$ with $\theta=\theta_1$. The required sample size is calculated as $n=E/d$.

2.2. Gamma shape parameter estimation from published sources

Parameters of the gamma distribution for fully observed data can be estimated using maximum likelihood estimation or method of moments with correction [23]. However, in many published studies, only a few survival quantiles (such as median and interquartile range) instead of the whole observed data are available. Maximum likelihood estimators and method of moment estimators may not result in accurate parameter estimates due to the small number of data points. To estimate the shape parameter based on a limited number of survival quantiles from a previously published resource, we can utilize iterative procedures to find a gamma distribution that best fits the published data, and the shape parameter of the best-fit gamma distribution is our shape estimate. For example, a previous study reported p survival quantiles, $s_{o_1}, s_{o_2}, \ldots, s_{o_p}$, and their survival times, $t_{o_1}, t_{o_2}, \ldots, t_{o_p}$. Suppose the above survival times correspond to survival quantiles, $s_{t_1}, s_{t_2}, \ldots, s_{t_p}$ in a theoretical distribution $gamma(k^*, \theta^*)$. Let $q_i=s_{o_i}-s_{ti}\ (i=1,2,\ldots,p)$ denote the distance between the published survival quantile s_{o_i} and theoretical quantile s_{ti} for survival time t_{o_i}. Iterative searching procedures could be initiated to search for the best-fit distribution $Gamma(k^*,\theta^*)$ with shape parameter k^*, such that $\sum_{i=1}^{p}q_i^2=\sum_{i=1}^{p}(s_{o_i}-s_{ti})^2$ is minimized. By implementing this algorithm, we can obtain an estimate of the gamma shape parameter $\widehat{k}=k^*$ using the survival quantiles from previous studies. For example, if

a published study reported 75 %, 50 %, and 25 % survival time as 1.7 months, 2.9 months, and 4.8 months. As shown in Fig. 1a, in gamma distribution with shape = 1, scale = 3, the survival probabilities (1-CDF) for the three survival times are 0.57, 0.38, and 0.2. The squared distance defined in section 2.2 is $q = (0.75-0.57)^2 + (0.5-0.38)^2 + (0.25-0.2)^2$. We can perform an iterative procedure to search for the shape and scale parameters with minimum squared distance q.

2.3. Assessing the accuracy of gamma shape estimate through simulations

The formula of sample size calculation using gamma distribution requires a known gamma shape parameter, this point estimate of which is usually obtained from published studies that either provide survival quantities (median and or inter-quartile range) or Kaplan Meier curves. As Phadnis [24] discussed regarding an accurate Weibull shape parameter estimated from historical studies, the accuracy of the gamma shape parameter estimate is critical as it impacts the sensitivity of subsequent sample size calculation in clinical trials with TTE endpoint. Underestimation or overestimation could happen if an imprecisely estimated gamma shape parameter is used in sample size calculation, and the inaccurate estimate further impacts study design and analysis. For example, if a gamma point estimate extracted from one historical study is $\widehat{k}_1 = 1.5$, and using another published study, we get $\widehat{k}_2 = 1.25$. To test the hypothesis of H_0 : *Median* = 2 *months vs.* H_1 : *Median* = 3 *months* (effect size of 1.5), assuming loss to follow-up is $r =$ 15 %, accrual and follow up time are both 12 months, with significance level at 0.05 and type II error of 0.2. Applying the sample size formula in section 2.1, the number of events needed (E), the probability of event occurring during study period (d), and final sample sizes (n) accounting for loss to follow-up will be 25, 0.995, and 30 respectively, when $\widehat{k}_1 = 1.5$. In the calculation using $\widehat{k}_2 = 1.25$, we will need 25 events, with probability of events occurring at 0.990, and a total of 36 samples adjusting for loss to follow-up (in both calculations, final sample size $n = \frac{E}{d*\left(1-\frac{r}{100}\right)}$). If the true shape is $k = 1.5$ but we inaccurately use $\widehat{k} = 1.25$, the calculated sample size will require six more or 20 % more subjects to be recruited, and this could be a significant practical issue for small-sized trials or trials with slow accrual. In the opposite scenario, if the true shape is $k = 1.25$, but we inaccurately use $\widehat{k} = 1.5$ for sample size calculation, the study will be underpowered to detect the prespecified effect size due to underestimated sample size.

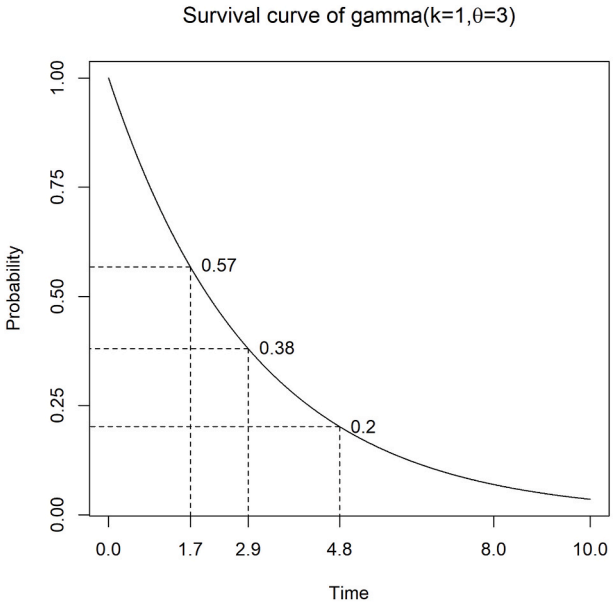


Fig. 1a. example of estimating gamma parameters.

An extensive simulation study is conducted to assess the accuracy of the estimator proposed in section 2.2. The event times are simulated from $gamma(k, \theta_E)$ using different values of gamma shape $k = 0.5, 0.75, 1, 1.25, 1.5$ (representing decreasing, constant, and increasing hazard patterns), and scale values of $\theta_E = 1,2,3,4,5$. To maintain a prespecified event rate m (or censoring rate $c/100 = 1 - m$), the censoring time is simulated from $gamma(1, \theta_C)$ with $\theta_C = \theta_E * \frac{m^{1/k}}{1-m^{1/k}}$ by applying the method discussed in Wan [25] (See Appendix (a)). Right-censoring mechanism with censoring rate $c =$ 0%, 10%, 20%, 30%, 40%, and sample sizes of $n = 25, 50, 100, 200, 500$ are considered simulation scenarios. For each combination of shape value k, scale value θ_E, censoring rate c, and sample size n, N = 10,000 data sets are generated, and Kaplan Meier estimates of survival quantiles are obtained. Next, we extract the point estimate $\widehat{k}$ of gamma shape for each simulation run (using the approach described in section 2.2) for five scenarios of 'number of information points (NIP)' followed.

(i) NIP = 2: Extracting k using the 25^{th} and 50^{th} percentiles of KM estimates
(ii) NIP = 2: Extracting k using the 25^{th} and 75^{th} percentiles of KM estimates
(iii) NIP = 3: Extracting k using the 25^{th}, 50^{th} and 75^{th} percentiles of KM estimates
(iv) NIP = 4: Extracting k using the 20^{th}, 40^{th}, 60^{th}, and 80^{th} percentiles of KM estimates
(v) NIP = 5: Extracting k using the 17^{th}, 34^{th}, 50^{th}, 67^{th}, and 84^{th} percentiles of KM estimates

In practice, NIP scenarios (i) – (iii) are most likely to be observed, NIP scenarios (iv) and (v) are for exploratory purposes to assess how the accuracy of $\widehat{k}$ changes as NIP increases. Each survival quantile is obtained by applying interpolation with the two nearest KM estimates. For example, when $c =$ 0%, for sample size $n = 25$, $n = 27$, and $n = 28$, the 25^{th} percentile of a KM curve (or 75^{th} percentile of survival) will be the time of the 7^{th} event, time of the 7^{th} event, and time between the 7^{th} and 8^{th} event, respectively. However, the exact survival percentile for the 7^{th} event is 72 % when $n = 25$, the exact survival percentile for the 7^{th} event is 74.07 % when $n = 27$, and the exact survival percentile for the 7^{th} event is 75 % when $n = 28$ (See Appendix (b)). In all three sample size cases, the times of the 25^{th} percentile of KM estimates are not precisely the 75^{th} percentile of survival, which leads to an inaccurate estimation of the gamma shape parameter. The purpose of interpolation is to reduce the inaccuracy introduced by non-parametric estimation.

In scenarios where simulated data has small sample sizes and high censoring rates, we cannot obtain some KM estimates of low survival quantiles. To address this issue, the simulation is modified to use the closest step boundary to replace the unavailable KM estimates. For instance, with a 40 % censoring rate, sample size $n = 25$, $NIP = 5$, we are able to extract the KM estimates of the 17^{th}, 34^{th}, 50^{th}, and 67^{th} percentiles but not the 84^{th} percentile. In this case, the closest step boundary to the KM estimate of 84^{th} percentile is used (e.g., using the KM estimate of 75^{th} percentile instead of the 84^{th} percentile).

Denote $\widehat{k}_{(i)}$ as the estimate of k from the i^{th} simulation $(i = 1, 2, \ldots, N)$ and $\widehat{k}_{avg} = \frac{1}{N}\sum_{i=1}^{N}\widehat{k}_{(i)}$ is the average of all $\widehat{k}_{(i)}$ values. The average relative bias (ARB) is considered the primary criterion for assessing the accuracy of $\widehat{k}_{avg}$ since sample size calculations depend on a reasonably accurate estimate of the shape parameter. Other metrics include root mean squared error (RMSE), scaled root mean squared error (SRMSE), and coefficient of variation (CV). We also compute the average maximum likelihood estimate of shape, $\widehat{k}_{mle}$, using the completed simulated data set and evaluate the bias of the estimate relative to the $\widehat{k}_{mle}$ (RARB). The

definitions are followed.

(a) The average relative bias of the estimator $\widehat{k}$ is defined as:

$$ARB = \frac{1}{N}\sum_{i=1}^{N}\frac{\widehat{k}_{(i)} - k}{k} = \frac{\widehat{k}_{avg} - k}{k} \tag{6}$$

(b) Root mean squared error (RMSE) of the estimator $\widehat{k}$ is defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{k}_{(i)} - k\right)^2} \tag{7}$$

(c) Scaled root mean squared error (SRMSE) of the estimator $\widehat{k}$ is defined as:

$$SRMSE = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{k}_{(i)} - k\right)^2}}{k} \tag{8}$$

(d) The coefficient of variation of $\widehat{k}$ is defined as:

$$CV = \sqrt{\frac{var(\widehat{k})}{\widehat{k}}} = \sqrt{\frac{\frac{1}{N-1}\sum_{i=1}^{N}\left(\widehat{k}_{(i)} - \widehat{k}_{avg}\right)^2}{\widehat{k}}} \tag{9}$$

(e) The bias of $\widehat{k}$ relative to the maximum likelihood estimate $\widehat{k}_{MLE}$ is calculated as:

$$RARB = \frac{\widehat{k}_{avg} - \widehat{k}_{mle}}{\widehat{k}_{mle}} \tag{10}$$

2.4. Misspecification of the underlying distribution

Our sample size calculation assumes gamma-distributed survival time with a known shape parameter. Specifically, the proposed test statistic follows a chi-square distribution leading to - (i) calculating the number of events (for prespecified values of power, type I error, and effect size) and (ii) calculating the probability of an event in the allotted study time. Subsequently, (i) and (ii) are used to calculate the required sample size. However, it is possible that for a clinical trial of interest, survival times may not adhere to the gamma distribution. In such scenarios, it is important to assess the 'cost of misspecification' such as overestimation or underestimation of the total sample size. Since commercial software (like nQuery [15]or PASS [14]) provides options for designing single-arm trials using the Weibull distribution (with exponential as a special case), we decide to study the 'misspecification effect' assuming that true survival distribution is Weibull. That is, we ask the question: "What if the proposed method intended for gamma-distributed survival times is used when the true survival distribution is Weibull?" This is done in the following steps: (i) Survival times are simulated from the Weibull distribution. These survival times represent data from a previous trial; (ii) These data are analyzed by a statistician using the Weibull distribution, however, only the summary statistics (say 3 NIP - often the median and IQR) are published; (iii) Our task is to use these limited NIP to calculate the sample size required for designing a new study using the proposed method. With the true Weibull distribution parameters, we also calculate the "true" sample size using the methods of Wu [16] and Phadnis [17]. Comparing this "true" sample size to the sample size obtained in (iii) allows us to study the effect of misspecification. The Weibull shape parameter estimate can be obtained by performing the median rank regression with at least two Weibull survival times and their corresponding survival probabilities [26] (See Appendix (c)).

We perform all the calculations and simulations with statistical software R [27] (version 3.6), using the high-performance computing (HPC) facilities in the Center for Research Computing at the University of Kansas.

3. Results

A real-life single-arm clinical trial about cholangiocarcinoma was discussed by Phadnis [17] and Waleed [28], which was designed using the Weibull distribution. Chemotherapy-refractory advanced metastatic biliary cholangiocarcinoma is a rare but aggressive neoplasm with a median PFS of 2.5 months and IQR of around 2–5 months [29]. The hypothesis used by Waleed [28] is $H_0 : M \leq 2.5$ months versus $H_1 : M >$ 3.75 months, where M is the median PFS time. In this manuscript, we are using the hypotheses in the context of utilizing the gamma distribution.

3.1. Sample size calculation results using the gamma distribution

Suppose researchers want to calculate the required sample size for testing the hypotheses below with a one-sided type I error of 5 % and power of 80 %, $H_0 : Median = 2.5$ months versus $H_a : Median = 3.75$ months. We performed sample size calculation using the method in section 2.1 with combinations of different settings of gamma shape values, accrual time, and follow-up time. In addition to sample size calculation, we want to check whether the empirical type I error is controlled, as well as whether the empirical power is preserved. Simulations (N = 10,000) were conducted to compute the empirical type I error and the empirical power with the calculated sample sizes.

Table 1 presents the results of sample size calculation and the results of assessing the empirical type I error and type I power. The columns in Table 1 are simulation settings of true gamma shape (k), accrual time in months (a), follow-up time in months (f), total sample size calculated using the gamma approach (n), the average number of events observed under H_0, empirical type I error, the average number of events observed under H_1, and the empirical power. From Table 1, we see that the calculated sample size increases as accrual time a and follow-up time f increases. In all the simulation scenarios, empirical type I error is well controlled (very close to the nominal level of 0.05) within the simulation error margin. Regarding the empirical power, we only observe slightly below the target power of 80 % when the gamma shape equals 0.5, particularly with the small accrual and follow-up times. Across all other simulation scenarios, the empirical power remains consistent and preserved.

3.2. Assessing the accuracy of the gamma shape parameter

Our simulation results of assessing the accuracy of the gamma shape parameter are shown in Table 2. The first four columns are censoring rate (c), the true value of gamma shape parameter in the simulation (k), varying sample sizes (n), and the average of maximum likelihood estimate of shape parameter over 10,000 simulations ($\widehat{k}_{mle}$). The remainder columns are the average estimated shape parameter over 10,000 simulations ($\widehat{k}_{avg}$) and ARB of 5 different NIP scenarios mentioned in section 2.3. Other criteria, such as root mean squared error, scaled root mean squared error, and coefficient variation, are also evaluated for $\widehat{k}_{avg}$, and the results are presented in Supplementary Table 1 to Supplementary Table 5. We set a threshold value of 5 % as the "maximum permissible value of ARB" to ensure reasonable accuracy (in real life, this threshold may be relaxed to a higher value, such as 10 %). Our observations are summarized below.

(i) Three trends of ARB of point estimate $\widehat{k}_{avg}$ are observed for increasing, constant, and decreasing hazard patterns. First, the ARB of $\widehat{k}_{avg}$ decreases as the sample size n increases. For example, for a decreasing hazard $k = 0.5$ with $c = 0$ %, for $n = 25$, 50, 100, 200, and 500, the ARBs of $\widehat{k}_{avg}$ with $NIP = 2$ (25th and 50th

Table 1
Sample size calculated using gamma distribution and evaluation of empirical type I error and empirical power for cholangiocarcinoma study with $H_0 : M \leq 2.5$ months vs. $H_1 : M > 3.75$ months. 10,000 simulations with nominal type I error 5 %, target power 80 %.

Shape k	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_1	Empirical power
0.5	3	3	137	86.39	0.0504	73.60	0.7322
	3	6	111	83.88	0.0542	73.05	0.7717
	3	12	92	81.19	0.0514	73.56	0.7872
	6	3	123	85.25	0.0482	73.52	0.7482
	6	6	105	83.41	0.0505	73.42	0.7791
	6	12	89	80.07	0.0500	73.10	0.7905
	12	3	107	83.01	0.0506	73.16	0.7725
	12	6	97	82.15	0.0452	73.77	0.772
	12	12	86	79.50	0.0518	73.52	0.8009
0.75	3	3	90	60.50	0.0453	49.39	0.7936
	3	6	70	57.75	0.0544	49.61	0.8177
	3	12	57	53.97	0.0538	49.67	0.8154
	6	3	78	58.41	0.0444	49.05	0.8098
	6	6	65	56.19	0.0518	49.28	0.8162
	6	12	55	52.73	0.0547	49.11	0.814
	12	3	67	56.08	0.0505	49.19	0.8126
	12	6	60	54.69	0.0482	49.51	0.8045
	12	12	54	52.55	0.0531	49.79	0.8136
1	3	3	67	47.23	0.0446	37.54	0.8392
	3	6	50	43.62	0.0451	37.39	0.8239
	3	12	41	40.00	0.0514	37.60	0.8199
	6	3	57	44.96	0.0456	37.32	0.8308
	6	6	47	42.69	0.0449	37.68	0.8248
	6	12	40	39.30	0.0480	37.38	0.8215
	12	3	49	42.85	0.0474	37.77	0.8277
	12	6	43	40.65	0.0481	37.35	0.825
	12	12	39	38.59	0.0475	37.30	0.8197
1.25	3	3	53	38.69	0.0525	30.02	0.8657
	3	6	39	35.25	0.0453	30.30	0.8465
	3	12	32	31.63	0.0466	30.22	0.8267
	6	3	45	36.78	0.0488	30.21	0.8558
	6	6	37	34.61	0.0472	30.81	0.8485
	6	12	32	31.75	0.0475	30.67	0.8309
	12	3	38	34.11	0.0481	30.17	0.843
	12	6	34	32.76	0.0456	30.45	0.8286
	12	12	31	30.86	0.0481	30.20	0.8209
1.5	3	3	44	33.10	0.0485	25.25	0.8781
	3	6	32	29.66	0.0485	25.67	0.8592
	3	12	26	25.86	0.0546	25.02	0.8293
	6	3	37	31.07	0.0517	25.44	0.8624
	6	6	30	28.59	0.0517	25.73	0.8535
	6	12	26	25.91	0.0545	25.30	0.8296
	12	3	31	28.30	0.0530	25.21	0.8503
	12	6	28	27.30	0.0509	25.64	0.8445
	12	12	26	25.95	0.0533	25.59	0.8285

percentile) are 0.2, 0.096, 0.045, 0.025, and 0.01, respectively. For constant hazard $k = 1$ and increasing hazard $k = 1.5$, the same trend is observed in fixed c and *NIP* scenarios. Second, the ARB of $\widehat{k}_{avg}$ decreases as *NIP* increases. For instance, in decreasing hazard scenario $k = 0.5$, $c = 0$ %, and $n = 25$, for $NIP = 2$ (25th and 75th percentile), 3, and 4, the ARBs of $\widehat{k}_{avg}$ are 0.091, 0.065 and 0.044. For the same settings of n and c, a similar pattern is also shown when the hazard is constant or increasing. Last, the ARB of $\widehat{k}_{avg}$ increases as c increases. In Table 2, in decreasing hazard $k = 0.5$ and $n = 25$, for $c = 0$ %, 20 % and 40 %, the ARB of $\widehat{k}_{avg}$ using 2 *NIP* (25th and 50th percentile) are 0.2, 0.231, and 0.292, respectively. Given the same n and *NIP*, similar patterns are also shown in other hazard patterns.

(ii) The maximum likelihood estimate of shape parameter $\widehat{k}_{mle}$ is closer to k, compared to $\widehat{k}_{avg}$ with $NIP = 2$ (25th and 50th percentiles, or 25th and 75th percentiles), when n is small and $c > 0$ %. For example, when $c = 20$ %, $k = 0.5$, and $n = 25$, the average relative bias of $\widehat{k}_{mle}$ (MARB) is 0.098, the ARB of $\widehat{k}_{avg}$ with $NIP = 2$ (25th and 50th percentiles, and 25th and 75th percentiles) are 0.231 and 0.119. This is because $\widehat{k}_{mle}$ is computed using all the sample points, and $\widehat{k}_{avg}$ only uses two information points. $\widehat{k}_{avg}$ gets closer to even smaller than $\widehat{k}_{mle}$ as n gets larger or *NIP* gets larger.

(iii) More available information points result in a more accurate point estimate of k, but the marginal benefit of additional information points on accuracy reduces when $NIP \geq 3$. For example, when $n = 25$ and $c = 0$ %, for $NIP = 2$ (25th and 75th percentile), 3, 4, and 5, the ARBs of $\widehat{k}_{avg}$ are 0.091, 0.065, 0.044, and 0.06, respectively. In a large sample size scenario, the marginal benefit becomes very small. For instance, when $k = 0.5$, $c = 0$ % and $n = 200$, for $NIP = 2$ (25th and 75th percentile), 3, 4, and 5, the ARBs of $\widehat{k}_{avg}$ are 0.012, 0.008, 0.006, and 0.005, respectively. We also notice that in the case of only two information points reported, the point estimate using wider-ranged *NIP* is more accurate than using narrower-ranged *NIP*. For example, in Table 2, for $k = 0.5$, $c = 0$ %, and $n = 25$, $NIP = 2$ (25th and 50th percentile) returns

Table 2
Simulation results for all NIP scenarios under various censoring proportions. ([a]NIP = 2 is 25th and 50th, [b]NIP = 2 is 25th and 75th, NIP = 3 is 25th, 50th and 75th percentile, NIP = 4 is 20th, 40th, 60th, and 80th percentile, NIP = 5 is 17th, 34th, 50th, 67th and 84th percentile).

c	k	n	$\widehat{k}_{mle}$	*MARB*	[a] *NIP* = 2		[b] *NIP* = 2		*NIP* = 3		*NIP* = 4		*NIP* = 5	
					$\widehat{k}_{avg}$	ARB	$\widehat{k}_{avg}$	ARB	$\widehat{k}_{avg}$	ARB	$\widehat{k}_{avg}$	ARB	$\widehat{k}_{avg}$	ARB
0 %	0.5	25	0.549	0.098	0.600	0.200	0.545	0.091	0.533	0.065	0.522	0.044	0.530	0.060
		50	0.523	0.045	0.548	0.096	0.522	0.043	0.515	0.029	0.510	0.021	0.511	0.022
		100	0.511	0.021	0.523	0.045	0.510	0.019	0.506	0.012	0.504	0.008	0.504	0.009
		200	0.506	0.011	0.512	0.025	0.506	0.012	0.504	0.008	0.503	0.006	0.503	0.005
		500	0.503	0.005	0.505	0.010	0.502	0.005	0.502	0.003	0.501	0.002	0.501	0.002
	0.75	25	0.827	0.103	0.935	0.247	0.831	0.108	0.809	0.079	0.793	0.057	0.800	0.067
		50	0.786	0.048	0.838	0.117	0.788	0.050	0.775	0.034	0.768	0.024	0.769	0.025
		100	0.768	0.023	0.789	0.051	0.767	0.023	0.761	0.015	0.759	0.012	0.758	0.011
		200	0.759	0.011	0.769	0.026	0.759	0.012	0.756	0.008	0.754	0.006	0.754	0.006
		500	0.754	0.005	0.759	0.012	0.755	0.006	0.754	0.005	0.753	0.004	0.753	0.004
	1	25	1.109	0.109	1.286	0.286	1.126	0.126	1.094	0.094	1.069	0.069	1.079	0.079
		50	1.049	0.049	1.109	0.109	1.048	0.048	1.032	0.032	1.024	0.024	1.026	0.026
		100	1.028	0.028	1.060	0.060	1.031	0.031	1.023	0.023	1.018	0.018	1.019	0.019
		200	1.014	0.014	1.033	0.033	1.017	0.017	1.013	0.013	1.009	0.009	1.009	0.009
		500	1.005	0.005	1.013	0.013	1.006	0.006	1.004	0.004	1.004	0.004	1.003	0.003
	1.25	25	1.397	0.118	1.648	0.319	1.418	0.135	1.376	0.101	1.349	0.080	1.363	0.090
		50	1.318	0.055	1.423	0.138	1.328	0.062	1.306	0.045	1.293	0.035	1.293	0.034
		100	1.283	0.027	1.328	0.063	1.290	0.032	1.280	0.024	1.273	0.018	1.271	0.017
		200	1.267	0.014	1.292	0.034	1.271	0.017	1.265	0.012	1.261	0.009	1.261	0.009
		500	1.255	0.004	1.264	0.011	1.257	0.005	1.254	0.003	1.253	0.002	1.253	0.003
	1.5	25	1.677	0.118	2.034	0.356	1.716	0.144	1.665	0.110	1.622	0.082	1.632	0.088
		50	1.583	0.055	1.750	0.166	1.600	0.066	1.571	0.048	1.556	0.037	1.555	0.037
		100	1.541	0.028	1.621	0.080	1.550	0.034	1.536	0.024	1.527	0.018	1.526	0.017
		200	1.520	0.013	1.550	0.033	1.525	0.017	1.519	0.012	1.513	0.009	1.513	0.009
		500	1.509	0.006	1.520	0.013	1.510	0.007	1.507	0.005	1.507	0.005	1.506	0.004
20 %	0.5	25	0.549	0.098	0.615	0.231	0.560	0.119	0.546	0.091	0.536	0.071	0.547	0.094
		50	0.523	0.045	0.551	0.101	0.528	0.056	0.521	0.042	0.516	0.033	0.518	0.036
		100	0.511	0.021	0.525	0.049	0.512	0.024	0.508	0.016	0.506	0.012	0.507	0.014
		200	0.506	0.011	0.513	0.026	0.507	0.014	0.505	0.010	0.504	0.008	0.504	0.008
		500	0.503	0.005	0.505	0.010	0.503	0.006	0.502	0.004	0.502	0.003	0.502	0.004
	0.75	25	0.827	0.103	0.976	0.301	0.858	0.144	0.833	0.110	0.813	0.083	0.828	0.104
		50	0.786	0.048	0.848	0.131	0.799	0.066	0.786	0.048	0.777	0.036	0.781	0.041
		100	0.768	0.023	0.795	0.060	0.773	0.031	0.767	0.022	0.764	0.018	0.763	0.018
		200	0.759	0.011	0.772	0.029	0.762	0.015	0.758	0.011	0.757	0.009	0.757	0.009
		500	0.754	0.005	0.760	0.013	0.755	0.007	0.754	0.006	0.754	0.005	0.754	0.005
	1	25	1.109	0.109	1.341	0.341	1.160	0.160	1.121	0.121	1.093	0.093	1.115	0.115
		50	1.049	0.049	1.123	0.123	1.066	0.066	1.048	0.048	1.037	0.037	1.042	0.042
		100	1.028	0.028	1.064	0.064	1.037	0.037	1.028	0.028	1.024	0.024	1.025	0.025
		200	1.014	0.014	1.037	0.037	1.022	0.022	1.017	0.017	1.013	0.013	1.013	0.013
		500	1.005	0.005	1.015	0.015	1.008	0.008	1.006	0.006	1.005	0.005	1.004	0.004
	1.25	25	1.397	0.118	1.748	0.398	1.472	0.178	1.420	0.136	1.382	0.106	1.406	0.125
		50	1.318	0.055	1.444	0.155	1.352	0.081	1.327	0.061	1.309	0.048	1.313	0.051
		100	1.283	0.027	1.341	0.073	1.300	0.040	1.287	0.030	1.278	0.023	1.278	0.023
		200	1.267	0.014	1.296	0.037	1.277	0.021	1.270	0.016	1.265	0.012	1.266	0.013
		500	1.255	0.004	1.267	0.014	1.259	0.007	1.256	0.005	1.254	0.004	1.255	0.004
	1.5	25	1.677	0.118	2.126	0.417	1.782	0.188	1.720	0.146	1.664	0.109	1.691	0.127
		50	1.583	0.055	1.783	0.188	1.632	0.088	1.600	0.067	1.578	0.052	1.579	0.053
		100	1.541	0.028	1.634	0.089	1.565	0.043	1.549	0.032	1.539	0.026	1.538	0.025
		200	1.520	0.013	1.559	0.039	1.532	0.021	1.524	0.016	1.519	0.013	1.519	0.012
		500	1.509	0.006	1.524	0.016	1.513	0.009	1.510	0.007	1.509	0.006	1.508	0.006
40 %	0.5	25	0.549	0.098	0.646	0.292	0.619	0.238	0.606	0.212	0.601	0.203	0.621	0.242
		50	0.523	0.045	0.560	0.120	0.555	0.110	0.549	0.098	0.548	0.096	0.555	0.111
		100	0.511	0.021	0.529	0.058	0.524	0.049	0.521	0.042	0.522	0.044	0.527	0.054
		200	0.506	0.011	0.515	0.031	0.513	0.025	0.511	0.022	0.511	0.021	0.514	0.027
		500	0.503	0.005	0.506	0.012	0.505	0.010	0.504	0.009	0.504	0.009	0.505	0.010
	0.75	25	0.827	0.103	1.041	0.388	0.944	0.259	0.916	0.222	0.900	0.200	0.939	0.252
		50	0.786	0.048	0.869	0.159	0.835	0.114	0.821	0.094	0.814	0.086	0.823	0.097
		100	0.768	0.023	0.804	0.072	0.790	0.053	0.782	0.043	0.781	0.041	0.785	0.046
		200	0.759	0.011	0.776	0.034	0.771	0.027	0.767	0.023	0.765	0.021	0.767	0.023
		500	0.754	0.005	0.762	0.016	0.758	0.011	0.757	0.009	0.756	0.008	0.757	0.009
	1	25	1.109	0.109	1.443	0.443	1.286	0.286	1.242	0.242	1.212	0.212	1.262	0.262
		50	1.049	0.049	1.171	0.171	1.116	0.116	1.094	0.094	1.083	0.083	1.096	0.096
		100	1.028	0.028	1.085	0.085	1.059	0.059	1.048	0.048	1.043	0.043	1.047	0.047
		200	1.014	0.014	1.044	0.044	1.032	0.032	1.026	0.026	1.022	0.022	1.024	0.024
		500	1.005	0.005	1.018	0.018	1.012	0.012	1.009	0.009	1.009	0.009	1.009	0.009
	1.25	25	1.397	0.118	1.854	0.483	1.604	0.283	1.541	0.233	1.497	0.198	1.561	0.249
		50	1.318	0.055	1.516	0.213	1.404	0.123	1.372	0.098	1.358	0.087	1.368	0.094
		100	1.283	0.027	1.360	0.088	1.323	0.059	1.308	0.046	1.300	0.040	1.304	0.043
		200	1.267	0.014	1.302	0.042	1.286	0.029	1.279	0.023	1.274	0.019	1.277	0.021
		500	1.255	0.004	1.273	0.018	1.264	0.011	1.260	0.008	1.258	0.006	1.259	0.007

(*continued on next page*)

Table 2 (*continued*)

c	k	n	$\widehat{k}_{mle}$	*MARB*	[a] *NIP* = 2		[b] *NIP* = 2		*NIP* = 3		*NIP* = 4		*NIP* = 5	
					$\widehat{k}_{avg}$	ARB	$\widehat{k}_{avg}$	ARB	$\widehat{k}_{avg}$	ARB	$\widehat{k}_{avg}$	ARB	$\widehat{k}_{avg}$	ARB
	1.5	25	1.677	0.118	2.395	0.596	1.977	0.318	1.897	0.265	1.835	0.224	1.910	0.273
		50	1.583	0.055	1.893	0.262	1.704	0.136	1.663	0.109	1.641	0.094	1.655	0.103
		100	1.541	0.028	1.671	0.114	1.593	0.062	1.571	0.048	1.563	0.042	1.567	0.044
		200	1.520	0.013	1.580	0.053	1.548	0.032	1.538	0.026	1.533	0.022	1.535	0.023
		500	1.509	0.006	1.530	0.020	1.519	0.013	1.515	0.010	1.514	0.009	1.514	0.009

Note: c is the proportion of censored samples, k is the true gamma shape, n is the simulated sample size, $\widehat{k}_{mle}$ is the maximum likelihood estimate of gamma shape, *MARB* is the average relative bias of the $\widehat{k}_{mle}$, $\widehat{k}$ and ARB are the gamma shape estimate using available information points and its average relative bias.

$\widehat{k}_{avg} = 0.6$ and ARB $= 0.2$, and *NIP* $= 2$ (25th and 75th percentile) returns $\widehat{k}_{avg} = 0.545$ and ARB $= 0.091$.

The above results demonstrate how ARB changes for different settings of censoring proportion, sample size, and *NIP*. The same trends using the other assessing metrics (RMSE, SRMSE, CV) are also found in Supplementary Table 1 to Supplementary Table 5. From a practical point of view, we want to obtain an accurate point estimate of the gamma shape parameter from a published study and use this shape estimate in the subsequent sample size calculation. Thus, we need answers to the following two questions: (1) if only a few survival quantiles are reported in a previous study, how large the sample size should be for us to extract a reliable gamma shape parameter? And (2) if a historical trial has a fixed sample size, what is the minimum number of information points that we need to obtain an accurate gamma shape parameter?

To answer the questions, we present Fig. 1b and 2. Fig. 1b plots the ARB, on the y-axis, versus sample size for *NIP* = 2 (25th and 75th percentile), on the x-axis, with $k = 1.5$ and different combinations of censoring scenarios and acceptable ARB thresholds. Overall, a decreasing trend of ARB as sample size n increases is clearly exhibited in Fig. 1b. It is shown in Fig. 1b that to keep the ARB below the acceptable threshold of 5 %, the minimal sample sizes required are 65 ($c = 0$ %), 85 ($c = 20$ %), and 126 ($c = 40$ %). If we relax the threshold of ARB to 10 %, then the minimum sample sizes required are 28 ($c = 0$ %), 37 ($c = 20$ %), and 59 ($c = 40$ %).

Fig. 2 presents the results of *NIP* needed, on the x-axis, versus the minimum sample size required, on the y-axis, for different combinations of censoring cases and ARB thresholds with $k = 0.5$ and different θ values. We have two main findings. First, for the same c and θ, the minimum sample size needed to achieve a certain ARB decreases as *NIP* increases. For example, when $\theta = 1$ and $c = 20$ %, to achieve a 5 % ARB,

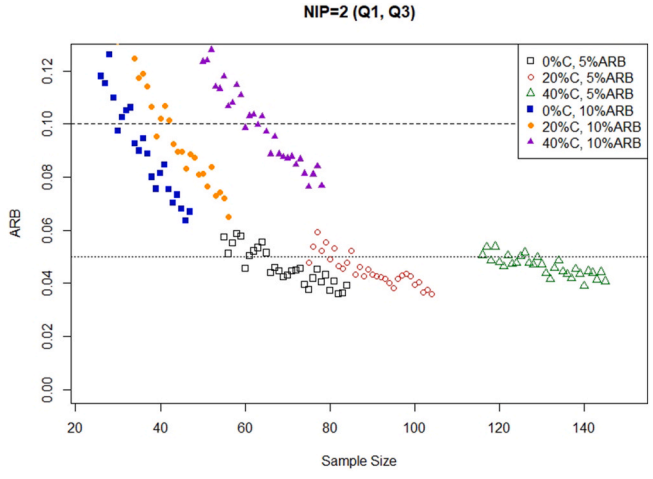


Fig. 1b. ARB vs Sample size for different censoring rates using 2 NIP (25th and 75th of KM estimates, $k = 1.25, \theta = 3$, *NIP* $= 2$).

the minimum sample size required for *NIP* = 2 (25th and 50th percentile), 2 (25th and 75th percentile), 3, 4, and 5, are 102, 60, 47, 37 and 37, respectively. Second, in the same *NIP* scenario, to achieve the same ARB threshold, a larger c corresponds to more sample sizes needed. Fixing k to reach the same ARB threshold, the minimum needed sample size for the same *NIP* scenario is similar when scale parameter θ changes. In Fig. 2, we can see that, with $k = 0.5$, $c = 20\%$, ARB threshold of 5 %, results for $\theta = 1$, 2, 3, and 4 under *NIP* = 2 (25th and 75th) are 60, 56, 60, and 55. Supplementary Fig. 1 and Supplementary Fig. 2 show similar results for constant hazard and increasing hazard patterns.

These results in Fig. 1b and 2 and Supplementary Figs. 1 and 2 can provide valuable information for statisticians when designing a single-arm study with a TTE endpoint. In a practical application, when the gamma assumption seems valid, a statistician can plan to calculate the sample size for such a trial using our proposed method. Our simulation result suggests that the statistician can use the reported median and IQR of event time from a previously published study (at least 60 samples and $c \leq 20\%$) to extract a reasonably accurate gamma shape parameter estimate and plug it into the sample size formula to calculate the required sample size.

3.3. Assessing the impact of misspecification

In real life, our parametric distributed survival time assumption might not hold by only looking at a few survival quantiles or a KM plot. When misspecification of survival distribution occurs, it is important to know how off-target our sample size calculation is and what consequence that could lead to. To explore that, we simulate Weibull distributed survival times, extract three survival quantiles (median and IQR) of KM estimates using the simulated data, and estimate gamma parameters with the survival quantiles. We present the sample size calculated using both true Weibull parameters and estimated gamma parameters to show the impact of misspecification. We calculated the sample sizes using Weibull and gamma approaches for the following six hypotheses testing: (1) H_0 : *Median* $= 2.5$ months vs. H_1 : *Median* $= 3$ months; (2) H_0 : *Median* $= 2.5$ months vs. H_1 : *Median* $= 3.5$ months; (3) H_0 : *Median* $= 2.5$ months vs. H_1 : *Median* $= 3.75$ months; (4) H_0 : *Median* $= 2.5$ months vs. H_1 : *Median* $= 4$ months; (5) H_0 : *Median* $= 2.5$ months vs. H_1 : *Median* $= 4.5$ months; (6) H_0 : *Median* $= 2.5$ months vs. H_1 : *Median* $= 5$ months. All the calculations assume 18 months of accrual time and 18 months of follow-up time, with type I error $= 0.05$ and power $= 90$ %.

Table 3 and Fig. 3 present the results of the sample size calculated with true Weibull parameters compared to the sample size calculated using the estimates of the gamma parameters (assuming gamma-distributed survival time). In Table 3, the first column is the true Weibull shape parameter used to simulate data and calculate sample size using the Weibull approach. The second column presents the approach used for calculation. The remaining six columns are the results of calculated sample sizes for testing the six sets of hypotheses above. In Fig. 3, the alternative hypothesized median times are displayed on the x-axis, and the y-axis presents the calculated samples.

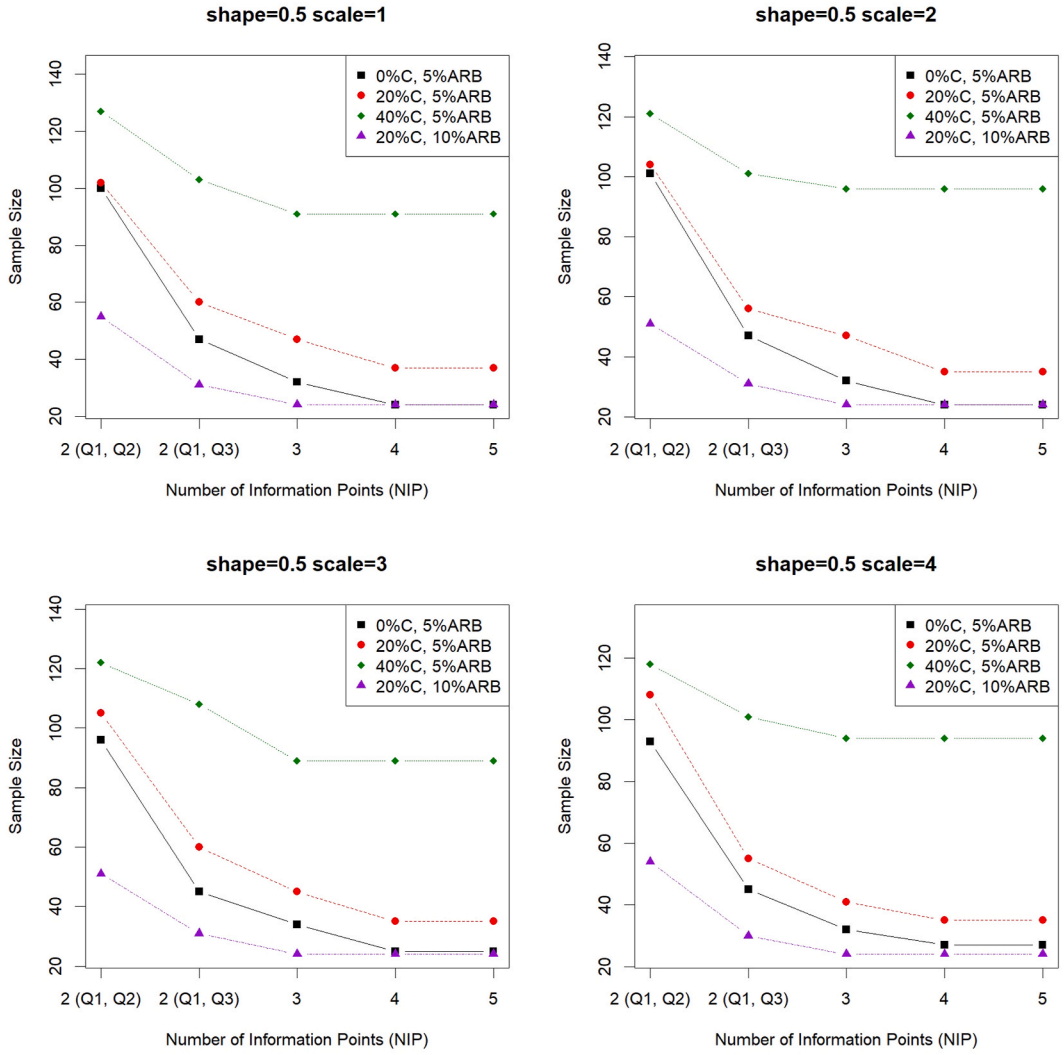


Fig. 2. Sample size vs NIP for different censoring rates and ARB thresholds ($k = 0.5$).

Table 3
Sample size calculation comparison using Weibull and gamma with various effect sizes (assuming accrual is 18 and follow up is 18, type I error = 0.05 and power = 90 %.

Weibull shape β	Approach	M = 2.5 vs. M = 3	M = 2.5 vs. M = 3.5	M = 2.5 vs. M = 3.75	M = 2.5 vs. M = 4	M = 2.5 vs. M = 4.5	M = 2.5 vs. M = 5
0.5	Weibull	1151	337	233	173	111	80
	gamma	728	214	147	110	71	51
0.75	Weibull	465	137	94	71	45	33
	gamma	388	114	79	59	38	28
1	Weibull	258	76	53	40	25	18
	gamma	250	74	51	39	24	18
1.25	Weibull	164	48	34	25	16	12
	gamma	173	51	35	26	17	12
1.5	Weibull	114	34	23	18	12	8
	gamma	127	38	26	19	13	9

Note: Median survival and IQR were simulated from Weibull with true shape β. Sample size calculations were performed using the Weibull and gamma approach for the six effect sizes above.

Gamma parameters are estimated using $NIP = 3$ (25th, 50th, and 75th percentile) from simulated data of $n = 100$, $c = 20$ %. From Table 3, we see that when the hazard is decreasing, the sample size calculated using Weibull is larger than the sample size calculated using gamma. When the hazard is constant, sample size calculated using Weibull and gamma are very close but still different. The difference is introduced by the bias of the gamma parameter estimates using three information points. In the increasing hazard scenario, the sample size calculated using Weibull is smaller than the sample size calculated using gamma.

For example, to test $H_0 : Median = 2.5$ vs. $H_1 : Median = 3.75$, in decreasing hazard case where true Weibull shape is 0.5, the sample size calculated using Weibull parameters and estimated gamma parameters are 233 and 147 in Table 3. As shown in Fig. 3 top left panel, the green dots of the 'correct' Weibull sample sizes are above the red dots of the gamma sample sizes. In the decreasing hazard scenario (Weibull shape <1), the true hazard decreases faster in the Weibull compared to the assumed gamma. Then the calculation using gamma will result in an underestimated sample size. Analogously, in the increasing hazard scenario, (Weibull shape >1), the true hazard increases faster in the Weibull compared to the assumed gamma. Then the calculation using gamma will result in an overestimated sample size.

4. Conclusion and discussion

In this paper, we propose a new parametric approach to calculate the

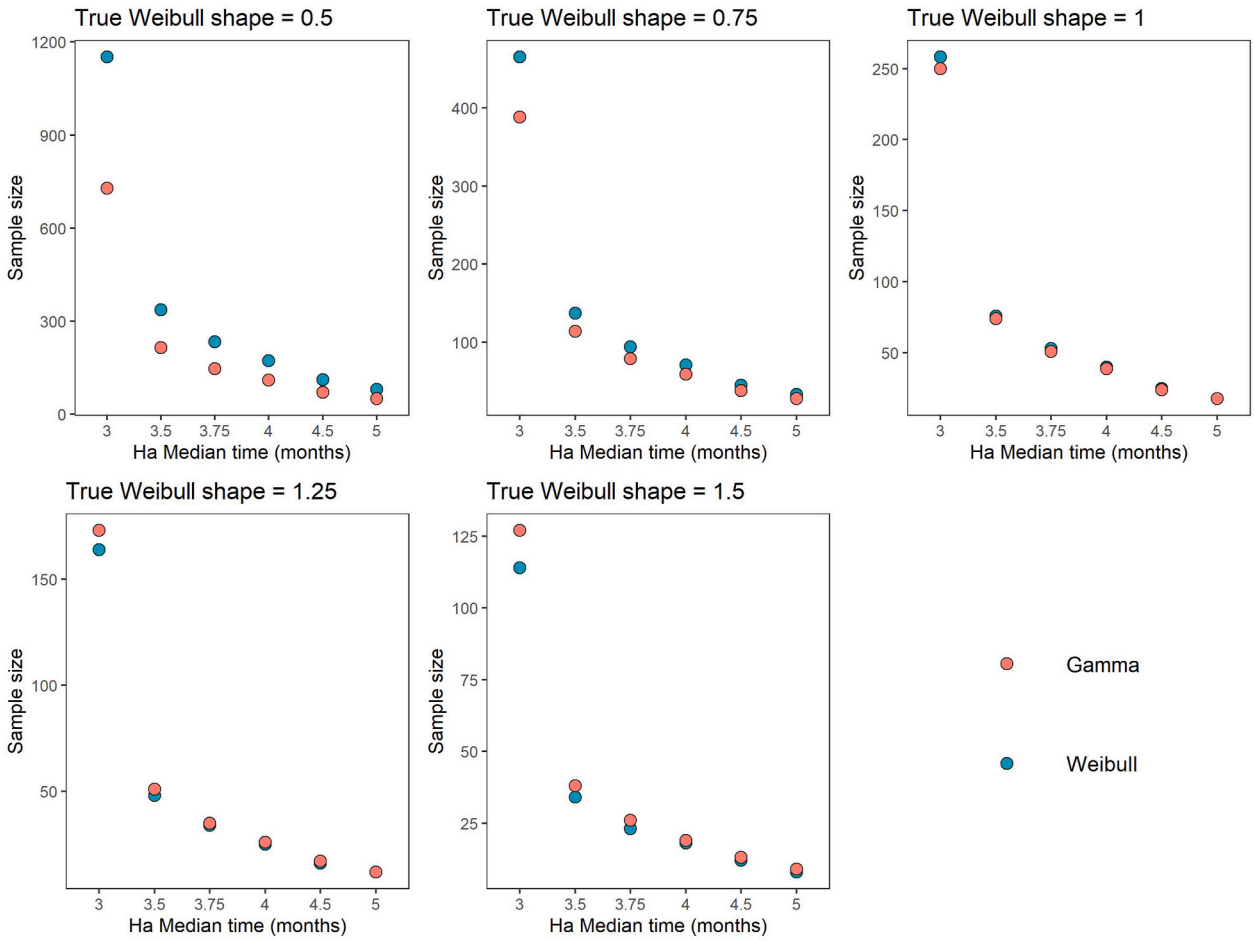


Fig. 3. Comparison of sample sizes using Weibull approach and gamma approach. Assuming true distribution is Weibull with different shape values and null median time of 2.5 months. Sample size calculated for various settings of alternative median times.

sample size for single-arm studies with a time-to-event endpoint. This approach assumes that survival time follows a gamma distribution, calculates the number of events needed using exact parametric test statistics with a known gamma shape parameter, adjusts for administrative censoring assuming uniform patient accrual to calculate the total number of subjects (sample size) that need to be recruited in the study.

The motivation for proposing this sample size calculation approach is that very limited parametric options are available in standard statistical software, and the assumptions regarding hazard behavior of existing parametric options do not always reflect the real-life phenomenon for some diseases, e.g., pancreatic cancer without the presence of VTE. In scenarios such as hazard increases to a finite constant but not infinity or hazard decreases to a constant but not zero, the proposed sample size calculation approach using gamma distribution can provide an alternative option that might be closer to the truth compared to the existing options.

To use the gamma approach for sample size calculation, a known shape parameter estimate is needed, and this point estimate is often extracted using published resources. Our simulation study provides the results of assessing the accuracy of the gamma shape parameter estimated from published data. These results facilitate statisticians to make informed decisions when designing a single-arm trial with a time-to-event endpoint under the gamma distribution framework. In cases where the gamma-distributed event time assumption is valid; our results enable statisticians to decide whether a historical study has a large enough sample size to extract an accurate gamma shape parameter from. Our results suggest that, in a decreasing hazard scenario, with at least three information points ($NIP \geq 3$), a published study with $n = 60$ and $c \leq 20$ % will be sufficient to provide a gamma shape estimate with ARB <5 %. Since our simulation uses interpolated information points from the Kaplan-Meier curve, we suggest that for the same setting as above, a statistician should use a published study with more than 60 sample sizes to obtain a reasonably accurate gamma shape estimate.

To our knowledge, in phase II clinical trials, the researchers often use the historical result as the null hypothesis and try to show that the new treatment performs better than the results of previously conducted trials. It is not uncommon that historical trials have small sample sizes and relatively high censoring rates. In such scenarios, our simulation result suggests more information points from the KM curve are needed to provide an accurate gamma shape estimate (ARB $< 5\%$). In extreme cases where insufficient information points are available from a previous study with small sample size and high censoring rate, a statistician may consider a relaxed threshold of ARB ≤ 10 % instead of ARB ≤ 5 % or use a more conservative Gamma shape parameter estimate for sample size calculation using the proposed approach.

Our study results also show that when the underlying distribution is misspecified, i.e., true event time is Weibull distributed but researchers incorrectly assume gamma distribution, the proposed sample size calculation approach will underestimate the required sample size when the true hazard decreases faster than assumed and overestimate the required sample size when the true hazard increases faster than assumed. Statisticians should be cautious when designing clinical trials assuming gamma-distributed survival time, and our method should not be used if the assumption of gamma distribution is inappropriate.

We have introduced a novel approach to determine the sample size for fixed-sample single-arm design utilizing the gamma distribution. In

the realm of potential future advancements, we recognize that integrating adaptive elements into a trial design could enhance the trial's flexibility, enable interim evaluation of the treatment effectiveness, and improve decision-making for researchers. Stochastic curtailment methods are often utilized in the interim analysis to detect early evidence of efficacy or futility for curtailment. Beyond the proposed fixed design method, we aim to incorporate futility-stopping rules using stochastic curtailment methods, like conditional power, predictive power, and Bayesian predictive probability to this the fixed design to optimize its use in designing single-arm trials with time-to-event endpoints.

Disclaimer

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Junqiang Dai: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jianghua He:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Conceptualization. **Milind A. Phadnis:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT and Grammarly in order to improve the clarity, conciseness, and overall readability of the content. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was performed at the HPC facilities operated by the Center for Research Computing at the University of Kansas supported in part through the National Science Foundation MRI Award #2117449.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.conctc.2024.101344.

Appendix

(a)

For each individual subject, assume the event time T and censoring time C are independent.

Let event time T $\sim$ $gamma(k,\theta_E)$, shape is k and scale is θ_E. The pdf of t is $f(t) = \frac{1}{\Gamma(k)\theta_E}\left(\frac{t}{\theta_E}\right)^{k-1} e^{-\frac{t}{k}}$.

censoring time C $\sim$ $gamma(1,\theta_C)$, shape is 1 and scale is θ_C. The pdf of t is $f(c) = \frac{1}{\theta_c}e^{-t}$.

Then the probability of T is censored by C is

$$P(T>C) = E[P(T>C|T)]$$

$$= E_T\left\{1-e^{-\frac{T}{\theta_C}}\right\} = \int_0^\infty f(t)\left(1-e^{-\frac{t}{\theta_C}}\right)dt = \frac{1}{\Gamma(k)\theta_E}\int_0^\infty \left(\frac{t}{\theta_E}\right)^{k-1} e^{-\frac{t}{k}}\left(1-e^{-\frac{t}{\theta_C}}\right)dt$$

Rearrange the integrand:

$$P(T>C) = 1-\frac{1}{\Gamma(k)\theta_E}\int_0^\infty \left(\frac{t}{\theta_E}\right)^{k-1} e^{-t\left(\frac{1}{\theta_E}+\frac{1}{\theta_c}\right)}dt = 1-\frac{1}{\Gamma(k)\theta_E^k}\int_0^\infty t^k e^{-t\left(\frac{1}{\theta_E}+\frac{1}{\theta_c}\right)}t^{-1}dt$$

Let $\lambda = \frac{1}{\theta_E}+\frac{1}{\theta_c}$, change of variables

$$u=\lambda t, t=\frac{u}{\lambda}, dt=\frac{1}{\lambda}du, t^{-1}dt=\left(\frac{u}{\lambda}\right)^{-1}\frac{1}{\lambda}du = u^{-1}du$$

Then

$$P(T>C) = 1-\frac{1}{\Gamma(k)\theta_E^k}\int_0^\infty \left(\frac{u}{\lambda}\right)^k e^{-u}u^{-1}du = 1-\frac{1}{\Gamma(k)\theta_E^k\lambda^k}\int_0^\infty e^{-u}u^{-1}du = 1-\frac{1}{\theta_E^k\lambda^k} = 1-\frac{1}{\left(1+\frac{\theta_e}{\theta_c}\right)^k}$$

To maintain a prespecified event rate of $m = 1-\ P(T>C) = \frac{1}{\left(1+\frac{\theta_e}{\theta_c}\right)^k}$.

We obtain $\theta_C = \theta_E * \frac{m^{1/k}}{1-m^{1/k}}$.

(b)

In the scenario where no censoring and all events observed. The Kaplan Meier estimate of 75 % survival time is the time of the event when survival percentile first dropped below 75 %.

If there are 25 total subjects, and all events are observed. When the 6th event happens, the KM estimate of survival probability is 0.76, when the 7th event happens, the KM estimate of survival probability is 0.72. So, the 75 % survival time is the time when the 7th event happens.

Similarly, when there are 27 total subjects, and all events are observed. When the 6th event happens, the KM estimate of survival probability is 0.7778, when the 7th event happens, the KM estimate of survival probability is 0.7407. So, the 75 % survival time is the time when the 7th event happens.

(c)

Using the median rank procedure to obtain the Weibull shape estimate. Denote survival probability as $S(t)$, survival time as t, shape parameter β, and scale parameter θ. $\log\left[-\log(S(t)\right] = \beta\log(t) - \beta\log(\theta)$.

With two or more sets of S(t) and t, β can be estimated by fitting a least squares regression line, the estimate of Weibull shape β is the slope of the regression line.

References

[1] L. Rubinstein, et al., Randomized phase II designs, Clin. Cancer Res. 15 (6) (2009) 1883–1890.

[2] T. Burzykowski, et al., Evaluation of tumor response, disease control, progression-free survival, and time to progression as potential surrogate end points in metastatic breast cancer, J. Clin. Oncol. 26 (12) (2008) 1987–1992.

[3] J. Goffin, et al., Objective responses in patients with malignant melanoma or renal cell cancer in early clinical studies do not predict regulatory approval, Clin. Cancer Res. 11 (16) (2005) 5928–5934.

[4] M. Buyse, et al., Relation between tumour response to first-line chemotherapy and survival in advanced colorectal cancer: a meta-analysis. Meta-Analysis Group in Cancer, Lancet 356 (9227) (2000) 373–378.

[5] Real-World Evidence. [cited 2024; Available from: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence.

[6] J. Corrigan-Curay, L. Sacks, J. Woodcock, Real-world evidence and real-world data for evaluating drug safety and effectiveness, JAMA 320 (9) (2018) 867–868.

[7] N.E. Breslow, Analysis of survival data under the proportional hazards model, Int. Stat. Rev./Rev. Int. Stat. 43 (1) (1975) 45–57.

[8] D.M. Finkelstein, A. Muzikansky, D.A. Schoenfeld, Comparing survival of a sample to that of a standard population, J. Natl. Cancer Inst. 95 (19) (2003) 1434–1439.

[9] 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 (12) (2014) 2004–2016.

[10] S.-H. Jung, Randomized Phase II Cancer Clinical Trials, CRC Press, 2013.

[11] X. Sun, P. Peng, D. Tu, Phase II cancer clinical trials with a one-sample log-rank test and its corrections based on the Edgeworth expansion, Contemp. Clin. Trials 32 (1) (2011) 108–113.

[12] J.F. Lawless, Statistical Models and Methods for Lifetime Data, vol. 362, John Wiley & Sons, 2011.

[13] SWOG. One Arm Survival. [cited 2023; Available from: https://stattools.crab.org/Calculators/oneNonParametricSurvival.htm.

[14] PASS, PASS 2021 Power Analysis and Sample Size Software, LLC, NCSS, 2021.

[15] nQuery, Sample Size And Power Calculation, Statsols (Statistical Solutions Ltd), 2017.

[16] J. Wu, Sample size calculation for the one-sample log-rank test, Pharmaceut. Stat. 14 (1) (2015) 26–33.

[17] M.A. Phadnis, 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? Contemp. Clin. Trials. Commun 15 (2019) 100360.

[18] S.C. Narula, F.S. Li, Sample size calculations in exponential life testing, Technometrics 17 (2) (1975) 229–231.

[19] A. Vincent, et al., Pancreatic cancer, Lancet 378 (9791) (2011) 607–620.

[20] H.T. Sørensen, et al., Prognosis of cancers associated with venous thromboembolism, N. Engl. J. Med. 343 (25) (2000) 1846–1850.

[21] H.K. Chew, et al., Incidence of venous thromboembolism and its effect on survival among patients with common cancers, Arch. Intern. Med. 166 (4) (2006) 458–464.

[22] M. Mandalà, et al., Venous thromboembolism predicts poor prognosis in irresectable pancreatic cancer patients, Ann. Oncol. 18 (10) (2007) 1660–1665.

[23] H. Lilliefors, Reducing the Bias of Estimators of Parameters for the Erlang and Gamma Distribution, Unpublished manuscript, 1971.

[24] M.A. Phadnis, et al., Assessing accuracy of Weibull shape parameter estimate from historical studies for subsequent sample size calculation in clinical trials with time-to-event outcome, Contemp. Clin. Trials. Commun 17 (2020) 100548.

[25] F. Wan, Simulating survival data with predefined censoring rates for proportional hazards models, Stat. Med. 36 (5) (2017) 838–854.

[26] O. Denisa, F. Laura, The evaluation of median-rank regression and maximum likelihood estimation techniques for a two-parameter Weibull distribution, Qual. Eng. 22 (4) (2010) 256–272.

[27] R.C. Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2020.

[28] M. Waleed, J. He, M.A. Phadnis, Some design considerations incorporating early futility for single-arm clinical trials with time-to-event primary endpoints using Weibull distribution, Pharmaceut. Stat. 20 (3) (2021) 610–644.

[29] J.E. Rogers, et al., Second-line systemic treatment for advanced cholangiocarcinoma, J. Gastrointest. Oncol. 5 (6) (2014) 408.