Published in final edited form as:

Biometrics. 2012 September ; 68(3): 983–989.

Median Tests for Censored Survival Data; a Contingency Table Approach

Shaowu Tang¹ and Jong-Hyeon Jeong^{1,*}

¹Department of Biostatistics, University of Pittsburgh, Pittsburgh, Pennsylvania 15261, U.S.A

Summary

The median failure time is often utilized to summarize survival data because it has a more straightforward interpretation for investigators in practice than the popular hazard function. However, existing methods for comparing median failure times for censored survival data either require estimation of the probability density function or involve complicated formulas to calculate the variance of the estimates. In this article, we modify a K -sample median test for censored survival data (Brookmeyer and Crowley, 1982, *Journal of the American Statistical Association* **77**, 433–440) through a simple contingency table approach where each cell counts the number of observations in each sample that are greater than the pooled median or vice versa. Under censoring, this approach would generate noninteger entries for the cells in the contingency table. We propose to construct a weighted asymptotic test statistic that aggregates dependent χ^2 -statistics formed at the nearest integer points to the original noninteger entries. We show that this statistic follows approximately a χ^2 -distribution with $k - 1$ degrees of freedom. For a small sample case, we propose a test statistic based on combined p -values from Fisher's exact tests, which follows a χ^2 -distribution with 2 degrees of freedom. Simulation studies are performed to show that the proposed method provides reasonable type I error probabilities and powers. The proposed method is illustrated with two real datasets from phase III breast cancer clinical trials.

Keywords

Censoring; Median failure time; Mood's median test; Quantile; Survival data

1. Introduction

In clinical research, investigators are often interested in estimating the mean or median time to occurrence of an event in the population under study because the mean or median failure time would be intuitively more interpretable than the popular hazard function-based results. For example, for the analysis of data from clinical trials, it would be more straightforward to summarize the efficacy of a new drug in terms of extending the mean or median time from

randomization to an event of interest or death by a certain number of years (Perry, Herndon, and Eaton, 1998; Hoskins, Swenerton, and Pike, 2002; Cunningham, Humblet, and Siena, 2004), because the hazard ratio only concerns the constant ratio of two instantaneous failure rates. Also in one of the real examples we use in Section 4, breast cancer patients in the experimental group were treated with tamoxifen (a hormonal therapy); therefore, investigators as well as patients would be interested in knowing how much longer a patient's remaining life years can be prolonged by taking tamoxifen, rather than reduction in the hazard ratio due to the hormonal therapy. Because the median is less sensitive to the outliers, however, the median failure time is often preferred as a natural choice to summarize a time-to-event distribution (Reid, 1981).

The asymptotic variance formulas for 2- or K -sample test statistics for the median involve the probability density function of the underlying distribution (Brookmeyer and Crowley, 1982; Wang and Hettmansperger, 1990). This means that to evaluate a test statistic for an observed dataset one needs to estimate the probability density function under censoring, which is not always straightforward. To avoid estimating the probability density function, Su and Wei (1993) proposed a nonparametric test statistic for comparing two median failure times, based on the minimum dispersion statistic (Basawa and Koul, 1988). However, the algorithm to minimize the dispersion statistic could be cumbersome and time consuming yet their simulation results indicated that the test was conservative in type I error probabilities, which would generally lead to reduction in powers.

In this article, we modify the Brookmeyer and Crowley's K -sample test for censored survival data (Brookmeyer and Crowley, 1982) through a simple contingency table approach, where each cell counts the number of observations in each sample that are greater than the pooled median or vice versa. In the method proposed in Brookmeyer and Crowley (1982), the variance formula for the test statistic involves the probability density function of the distribution of failure times, even though it disappears under the null hypothesis with balanced sample proportions. Furthermore, their standardized statistic involves the G -inverse that might not have the full rank, which led the authors to estimate the median from the extrapolated (continuous) version of the weighted Kaplan–Meier estimates that seems unnatural, especially for small sample cases. The proposed method is more straightforward in that the test statistic is a weighted sum of chi-square statistics from contingency tables at the nearest integer points to the original noninteger entries generated under censoring. This would make the implementation much simpler in practice for applied statisticians and medical investigators.

The rest of the article is organized as follows: In Section 2, the Brookmeyer and Crowley's K -sample test is simplified in a contingency table setting, and new statistics are proposed to test the equality of the medians. In Section 3, extensive numerical studies are performed to assess the finite sample properties of the proposed test statistics. In Section 4, the proposed method is illustrated with breast cancer datasets from National Surgical Adjuvant Breast and Bowel Project (NSABP). We conclude with a brief remark in Section 5.

2. Median Test Statistics—Old and New

Assume that X_{ij} is the j th independent subject drawn from group i with $1 \leq i \leq k, 1 \leq j \leq n_i$, and θ_{pool} is the median from the pooled sample. Then for each group one can count the number of elements that are greater than $\hat{\theta}_{\text{pool}}$ for a median test (Mood, 1950). The results can be summarized into the following $2 \times k$ contingency table:

	Group 1	Group 2	...	Group k	Total
$> \theta_{\text{pool}}$	n_{11}	n_{21}	...	n_{k1}	m_1
$\leq \theta_{\text{pool}}$	n_{12}	n_{22}	...	n_{k2}	m_2
Total	n_1	n_2	...	n_k	N

Here $N = m_1 + m_2 = \sum_i n_i$. For each n_{is} , one can define the expected number μ_{is} as

$$\mu_{is} = \frac{n_i m_s}{N}, \text{ where } i = 1, 2, \dots, k; s = 1, 2.$$

Then a χ^2 -test statistic can be formed as

$$V = \sum_{i=1}^k \sum_{s=1}^2 \frac{(n_{is} - \mu_{is})^2}{\mu_{is}} \sim \chi_{k-1}^2. \quad (1)$$

It is well known that the chi-square approximation improves as μ_{is} increases, and $\mu_{is} \geq 5$ is usually sufficient for a good approximation.

Because the median test described above was designed only for complete data, in this section, we modify it for censored survival data. Let X be time to an event, which is a nonnegative random variable from a homogeneous population, and C be an independent censoring variable from an arbitrary distribution. Denote the corresponding survival function and density function as $S(x)$ and $f(x)$, respectively. Under censoring, one can only observe pairs of random variables (T, δ) , where $T = \min(X, C)$ and $\delta = 1$ for an event and $\delta = 0$ for a censored observation. The median of the distribution of X , θ , is defined as

$$\theta = \min \left\{ x : S(x) \leq \frac{1}{2} \right\}. \quad (2)$$

Now suppose that n_i independent observations are drawn from the i th population ($i = 1, 2, \dots, k$). Let T_{ij} be the j th observed survival time from the i th population ($1 \leq j \leq n_i$), δ_{ij} being an associated censoring indicator. One can estimate the pooled median failure time $\hat{\theta}_{\text{pool}}$ as,

$$\hat{\theta}_{\text{pool}} = \min \left\{ t : \hat{S}_{\text{pool}}(t) \leq \frac{1}{2} \right\},$$

where $\hat{S}_{\text{pool}}(t)$ is the weighted Kaplan–Meier estimates (Brookmeyer and Crowley, 1982) defined as

$$\hat{S}_{\text{pool}}(t) = \sum_{i=1}^k \lambda_i^N \hat{S}_i(t) \text{ with } \lambda_i^N = \frac{n_i}{N}, \quad (3)$$

where $\hat{S}_i(t)$ is the Kaplan–Meier estimate for the i th population (Kaplan and Meier, 1958).

Note that for $1 \leq i \leq k$, the entries in the above contingency table can be generalized for censored data as

$$n_{i1} = \sum_{j=1}^{n_i} \Pr(X_{ij} > \theta_{\text{pool}}),$$

which can be estimated by

$$\hat{n}_{i1} = \sum_{j=1}^{n_i} \Pr(X_{ij} > \hat{\theta}_{\text{pool}}),$$

where X_{ij} is the true failure time of the j th observation in the i th population, and $\Pr(X_{ij} > \hat{\theta}_{\text{pool}})$ estimates the probability that the true failure time of the j th observation is beyond $\hat{\theta}_{\text{pool}}$, which can be defined as the ‘pseudocount’ because it does not necessarily generate integer values. Specifically, for $i = 1, 2, \dots, k$ and $j = 1, \dots, n_i$, we have (Brookmeyer and Crowley, 1982)

$$\Pr(X_{ij} > \hat{\theta}_{\text{pool}}) = \begin{cases} 1, & \text{if } T_{ij} > \hat{\theta}_{\text{pool}} \text{ and } \delta_{ij} = 1, \\ 1, & \text{if } T_{ij} \geq \hat{\theta}_{\text{pool}} \text{ and } \delta_{ij} = 0, \\ 0, & \text{if } T_{ij} \leq \hat{\theta}_{\text{pool}} \text{ and } \delta_{ij} = 1, \\ \hat{q}_{ij}, & \text{if } T_{ij} < \hat{\theta}_{\text{pool}} \text{ and } \delta_{ij} = 0, \end{cases}$$

where $\hat{q}_{ij}$ satisfies

$$\hat{q}_{ij} = \Pr(X_{ij} > \hat{\theta}_{\text{pool}} | T_{ij} < \hat{\theta}_{\text{pool}}, \delta_{ij} = 0) = \Pr(X_{ij} > \hat{\theta}_{\text{pool}} | X_{ij} > T_{ij}) = \frac{\Pr(X_{ij} > \hat{\theta}_{\text{pool}})}{\Pr(X_{ij} > T_{ij})} = \frac{\hat{S}_i(\hat{\theta}_{\text{pool}})}{\hat{S}_i(T_{ij})}.$$

Now one can define

$$\hat{n}_{i2} = \sum_{j=1}^{n_i} \Pr(X_{ij} \leq \hat{\theta}_{\text{pool}}) = n_i - \hat{n}_{i1}.$$

Therefore, the $2 \times k$ contingency table for the modified median test can be displayed as follows:

	Group 1	Group 2	...	Group k	Total
$> \hat{\theta}_{\text{pool}}$	$n_{11}^{\hat{}}$	$n_{21}^{\hat{}}$	...	$n_{k1}^{\hat{}}$	$m_1^{\hat{}}$
$\hat{\theta}_{\text{pool}}$	$n_{12}^{\hat{}}$	$n_{22}^{\hat{}}$	...	$n_{k2}^{\hat{}}$	$m_2^{\hat{}}$
Total	n_1	n_2	...	n_k	N

Note that although $n_{i1}^{\hat{}} + n_{i2}^{\hat{}} = n_i$ holds for $i = 1, 2, \dots, k$, in general $n_{i1}^{\hat{}}$, $n_{i2}^{\hat{}}$, and $m_s^{\hat{}}$ are not integers. To generate $2 \times k$ contingency tables with neighboring integer entries, let us define

$$\tilde{n}_{i1} = \sup\{n \in \mathbf{N} : n \leq \hat{n}_{i1}\} \text{ and } \tilde{n}_{i2} = n_i - \tilde{n}_{i1},$$

where $\mathbf{N}$ is the set of positive integers. Obviously if $\tilde{n}_{i1}$ happens to be an integer, then $\tilde{n}_{i1} = n_{i1}^{\hat{}}$ and $\tilde{n}_{i2} = n_{i2}^{\hat{}}$. Otherwise we have

$$\tilde{n}_{i1} < \hat{n}_{i1} < \tilde{n}_{i1} + 1 \text{ and } \tilde{n}_{i2} - 1 < \hat{n}_{i2} < \tilde{n}_{i2}.$$

Given each set of the nearest integer points indexed by l ($l = 1, 2, \dots, 2^k$), one can construct the following $2 \times k$ contingency table:

	Group 1	Group 2	...	Group k	Total
$> \tilde{\theta}_{\text{pool}}^{(l)}$	$\tilde{n}_{11}^{(l)}$	$\tilde{n}_{21}^{(l)}$	...	$\tilde{n}_{k1}^{(l)}$	$\tilde{m}_1^{(l)}$
$\tilde{\theta}_{\text{pool}}^{(l)}$	$\tilde{n}_{12}^{(l)}$	$\tilde{n}_{22}^{(l)}$	...	$\tilde{n}_{k2}^{(l)}$	$\tilde{m}_2^{(l)}$
Total	n_1	n_2	...	n_k	N

where $\tilde{n}_{i1}^{(l)} = \tilde{n}_{i1}$ or $\tilde{n}_{i1} + 1$, $\tilde{n}_{i2}^{(l)} = n_i - \tilde{n}_{i1}^{(l)}$, and $\tilde{m}_s^{(l)} = \sum_{i=1}^k \tilde{n}_{is}^{(l)}$ ($s=1, 2$). This means that, for each i , there are only two choices for $\tilde{n}_{i1}^{(l)}$, so that one can construct 2^k such $2 \times k$ contingency tables.

After a χ_{k-1}^2 - statistic, V_l ($l = 1, 2, \dots, 2^k$), is formed for each of 2^k contingency tables, one can aggregate those 2^k test statistics by assigning some weights to V_l 's. To this end, for $i = 1, 2, \dots, k$, let us define

$$\hat{n}_{i1} = \tilde{n}_{i1} + \eta_i \text{ with } 0 \leq \eta_i < 1,$$

and

$$\varpi_i^{(l)} = \begin{cases} 1 - \eta_i, & \text{if } \tilde{n}_{i1}^{(l)} = \tilde{n}_{i1}, \\ \eta_i, & \text{if } \tilde{n}_{i1}^{(l)} = \tilde{n}_{i1} + 1. \end{cases}$$

and therefore the weight corresponding to the l th $2 \times k$ contingency table is defined as

$$\varpi^{(l)} = \prod_{i=1}^k \varpi_i^{(l)}. \quad (4)$$

For example, for $k = 2$, the four weights can be given as

$$\varpi^{(1)} = (1 - \eta_1)(1 - \eta_2), \varpi^{(2)} = \eta_1(1 - \eta_2), \varpi^{(3)} = \eta_2(1 - \eta_1), \varpi^{(4)} = \eta_1\eta_2. \quad (5)$$

In this case, if both η_1 and η_2 are small, then more weight is assigned to V_1 , and if both of them are large, then more weight is assigned to V_4 .

Therefore a way of aggregating information from 2^k subtables would be through a weighted test statistic defined as

$$U = \sum_{l=1}^{2^k} \varpi^{(l)} V_l. \quad (6)$$

It will be shown that U follows a χ^2 -distribution with $(k - 1)$ degrees of freedom. Proofs of the following lemma and theorem are provided in the Web Appendix.

Lemma 1. *Under the null hypothesis of k equal medians, as $n_1, n_2, \dots, n_k \rightarrow \infty$,*

$$\text{Corr}(Z_g, Z_h) = 1,$$

where $Z_l = (\tilde{n}_{11}^{(l)}, \tilde{n}_{21}^{(l)}, \dots, \tilde{n}_{k1}^{(l)}, \tilde{n}_{12}^{(l)}, \tilde{n}_{22}^{(l)}, \dots, \tilde{n}_{k2}^{(l)})$, $l = 1, 2, \dots, 2^k$

Theorem 1. *Suppose that V_l ($l = 1, 2, \dots, 2^k$) follows a χ^2 -distribution with $(k - 1)$ degrees of freedom. Then under the null hypothesis of k equal medians, the statistic*

$U = \sum_{l=1}^{2^k} \varpi^{(l)} V_l$, where $\sum_{l=1}^{2^k} \varpi^{(l)} = 1$, *approximately follows a χ^2 -distribution with $(k - 1)$ degrees of freedom.*

When μ_{is} in (1) is small, we can adopt an alternative approach to directly combining information through p -values. Assuming that p_l is a p -value associated with V_l ($l = 1, 2, \dots, 2^k$) computed from the Fisher's exact test, one can define the weighted p -value associated with the statistic U as

$$Q = \sum_{l=1}^{2^k} \varpi^{(l)} Q_l = \sum_{l=1}^{2^k} \varpi^{(l)} \{-2\log(p_l)\},$$

where $Q_l = -2 \log(p_l)$ follows a χ^2 -distribution with 2 degrees of freedom (Fisher, 1932) under the null hypothesis. Similarly as in Theorem 2, one can show that the statistic Q follows a χ^2 -distribution with 2 degrees of freedom under the null hypothesis.

3. Simulation Study—Type I Errors and Powers

Extensive simulation studies have been performed to study type I errors and powers of the proposed method at various nominal levels of α , sample sizes of n_i and censoring proportions. Similarly as in Su and Wei (1993), two scenarios are used to generate failure times to investigate type I error probabilities as follows:

- i. $S_1(t) = S_2(t) = \exp(-t)$.
- ii. $S_1(t) = \exp(-t)$ and $S_2(t) = 1 - \Phi(\log(1.44t))$.

In case 1, both samples are taken from the same distribution, so that the medians are equal. In case 2, two samples are taken from different distributions but with the same medians down to two decimal places. Censoring times are generated from $Uniform(0, C_i)$, where C_i determines the censoring proportions for different distributions $S_i(t)$, $i = 1, 2$. We have chosen C_i 's such that the censoring proportions are close to those reported in Su and Wei (1993) for comparison. For the sample sizes of 30, 50, and 100 for each group and various scenarios of censoring proportions, 1000 iterations were carried out and the proportion of cases that gave significant difference in median failure times between two samples was evaluated at a given nominal level. The simulation results are summarized in Tables 1 and 2 for various significance levels of 5%, 10%, 15%, and 20%. One can notice that the proposed approach tends to be less conservative, compared with Su and Wei's method (Table 1, Su and Wei, 1993).

Power analysis has been also performed for various sample sizes and censoring proportions by increasing the median differences at a significance level of 0.05. Data were generated similarly as before from two distributions $S_1(t) = \exp(-t)$ and $S_2(t) = \exp(-t + t^*)$, where t^* is the median difference between the two exponential distributions. Table 3 summarizes the proportions of the proposed test statistic U being larger than the upper fifth percentile of the χ^2 -distribution with 1 degree of freedom, which quickly increase with larger median differences and sample sizes, as expected.

We have also compared powers of the proposed median tests, chi-square and p -value based (CombP), with other existing tests, log rank, Gehan (1965), and Brookmeyer and Crowley tests, by using the same simulation scenario as in Brookmeyer and Crowley (1982) for the double exponential case. In more details, for the baseline group, 50 failure times were generated from the double exponential distribution with the median of 100, and another 50 failure times were generated for the experimental group from the same distribution by decreasing the medians, i.e., 100, 99.8, 99.6, 99.4, and 99.2, so that the median differences

would be 0, 0.2, 0.4, 0.6, and 0.8, respectively. Censoring times were also generated from the double exponential distribution with the median of 100.5. Figure 1 shows a Kaplan–Meier plot for one of the realizations of simulated data when the median difference is 0.6, which features with nonproportional hazards ($p = 0.027$). In fact, we noticed that the double exponential scenario generates a mix of proportional and nonproportional hazards data. Figure 2 shows that all of the compared tests have reasonable performance at the significance level of 0.05 under the null hypothesis, even though the proposed methods tend to be slightly conservative. We notice that the powers of the proposed tests are close to ones from Brookmeyer and Crowley (1982), especially for a larger difference in medians. The log-rank test is performing poorly due to the nonproportionality in this case compared to other tests, especially as the difference in the medians increases. In fact, we note that this type of comparison could be misleading because the *general* test such as log rank or Gehan compares the overall failure time distributions whereas the quantile test such as the median test compares a specific percentile. For example, the median test might perform better when there is a moderate *middle* difference between two failure time distributions, while the general tests might perform better when there is an *early* or a *late* difference between two distributions but the maximum difference occurs before or after the median failure time. Therefore we believe that the comparison with Su and Wei's and Brookmeyer and Crowley's test statistics would be more direct and meaningful even though the latter was unnaturally based on the extrapolated version of the weighted Kaplan–Meier estimates, as pointed out in Section 1. Therefore our simulation results indicate that both of the proposed methods are less conservative than Su and Wei's and are comparable to Brookmeyer and Crowley's, and the simple p -value-based method might be more powerful than other methods especially for a small sample case.

4. Application to NSABP Data

In this section, we apply the proposed approach to two datasets from the NSABP studies (B-04 and B-14). The B-04 study (Fisher, Jeong, and Anderson, 2002) was designed to compare radical mastectomy with a less extensive surgery (total mastectomy) with or without radiation therapy. A total of 1079 women with clinically negative axillary nodes had either radical mastectomy or total mastectomy without axillary dissection if their nodes became positive. A total of 586 women with clinically positive axillary nodes experienced either radical mastectomy or total mastectomy without axillary dissection but with postoperative irradiation. The censoring proportion in node-negative patients is about 27%, while the censoring proportion in node-positive patients is about 16%. The second dataset comes from the NSABP B-14 study, where patients with primary breast cancer, negative axillary nodes, and estrogen receptor positive tumors were randomized to receive either tamoxifen (a hormonal therapy) or placebo following surgery. The trial itself is described in details in the literature (Fisher, Costantino, and Redmond, 1989). To demonstrate a small sample case, only 68 eligible patients with tumor size greater than 5 cm (30 from placebo group; 38 from tamoxifen group) were included in the second analysis. Comparison groups were nodal status (negative versus positive) in B-04 and treatment (placebo versus tamoxifen) in B-14.

A 2×2 table after calculating the sum of the pseudocounts for each cell in the B-04 dataset is given as follows:

	Node negative	Node positive	Total
$> \hat{\theta}_{\text{pool}}$	608.89	223.57	832.46
$\hat{\theta}_{\text{pool}}$	470.11	362.43	832.54
Total	1079	586	1665

With $\eta_1 = 0.89$ and $\eta_2 = 0.57$, equation (5) gives the weights as $\varpi_1 = 0.047$, $\varpi_2 = 0.383$, $\varpi_3 = 0.063$, and $\varpi_4 = 0.507$. Therefore, the four 2×2 contingency tables with neighboring integer entries are given by

Case 1:

	Node negative	Node positive	Total
$> \hat{\theta}_{\text{pool}}$	608	223	831
$\hat{\theta}_{\text{pool}}$	471	363	834
Total	1079	586	1665

Case 2:

	Node negative	Node positive	Total
$> \hat{\theta}_{\text{pool}}$	609	223	832
$\hat{\theta}_{\text{pool}}$	470	363	833
Total	1079	586	1665

Case 3:

	Node negative	Node positive	Total
$> \hat{\theta}_{\text{pool}}$	608	224	832
$\hat{\theta}_{\text{pool}}$	471	362	833
Total	1079	586	1665

Case 4

	Node negative	Node positive	Total
$> \hat{\theta}_{\text{pool}}$	609	224	833
$\hat{\theta}_{\text{pool}}$	470	362	832
Total	1079	586	1665

The four corresponding test statistics are $V_1 = 50.84$, $V_2 = 51.35$, $V_3 = 49.89$, and $V_4 = 50.40$, and equation (6) gives the combined test statistic $U = 50.75$ with a p -value of 1.05×10^{-12} , which implies a significant difference in mean failure times between node-negative and node-positive women, consistent with the results from Jeong, Jung, and Costantino

(2008). Note that the value of the combined test statistic U for the B-04 data is similar to those of V_i , $i = 1, 2, 3, 4$ for this large sample case. The following table compares the results with other methods, including the combined p -value approach (CombP):

Method	Test statistic	p -value
Log rank	54.8	1.34×10^{-13}
Gehan	69.1	1.11×10^{-16}
Chi square (proposed)	50.75	1.05×10^{-12}
CombP (proposed)	54.63	1.37×10^{-12}

In this example, even if statistical conclusions are all the same, the Gehan's test gives the most significant p -value due to the nonproportionality of hazards between node-negative and node-positive patients (p -value from the nonproportionality test = 0.000165).

Now let us consider the NSABP B-14 dataset. The 2×2 contingency table containing the sum of pseudocounts from each cell is given by

	Placebo	Tamoxifen	Total
$> \hat{\theta}_{\text{pool}}$	12.84	21.21	34.05
$\hat{\theta}_{\text{pool}}$	17.16	16.79	33.95
Total	30	38	68

Because $\eta_1 = 0.84$ and $\eta_2 = 0.21$ in this small sample example, the weights are calculated as $\varpi_1 = 0.13$, $\varpi_2 = 0.66$, $\varpi_3 = 0.03$, and $\varpi_4 = 0.18$. Therefore, the four 2×2 contingency tables with neighboring integer entries are given by

Case 1:

	Placebo	Tamoxifen	Total
$> \hat{\theta}_{\text{pool}}$	12	21	33
$\hat{\theta}_{\text{pool}}$	18	17	35
Total	30	38	68

Case 2:

	Placebo	Tamoxifen	Total
$> \hat{\theta}_{\text{pool}}$	13	21	34
$\hat{\theta}_{\text{pool}}$	17	17	34
Total	30	38	68

Case 3:

	Placebo	Tamoxifen	Total
$> \hat{\theta}_{\text{pool}}$	12	22	34
$\hat{\theta}_{\text{pool}}$	18	16	34
Total	30	38	68

Case 4:

	Placebo	Tamoxifen	Total
$> \hat{\theta}_{\text{pool}}$	13	22	35
$\hat{\theta}_{\text{pool}}$	17	16	33
Total	30	38	68

The four corresponding test statistics are $V_1 = 1.56$, $V_2 = 0.95$, $V_3 = 2.15$, and $V_4 = 1.42$ and hence the combined test statistic is $U = 1.15$ with $\chi_1^2 p$ -value of 0.28, which implies the difference in median failure times between the two treatment groups among high-risk patients is not significant. The following table presents a comparison with other methods as before:

Method	Test statistic	p -value
Log rank	0.7	0.411
Gehan	1.2	0.269
Chi square (proposed)	1.15	0.28
CombP (proposed)	1.88	0.390

In this small sample example, the assumption of proportional hazards holds ($p=0.16$), so that the proposed combined p -value approach provides a similar result as the log-rank test, which is an optimal test under proportional hazards.

5. Conclusion

In this article, the Brookmeyer and Crowley's median test (Brookmeyer and Crowley, 1982) has been modified through a contingency table approach. The proposed method is simple, is easy to be implemented, and does not involve estimation of the probability density function to evaluate the variance of the test statistic. The interpretation of the results based the median failure times is generally more straightforward and specific than ones from the hazard function approach. Our simulation results indicate that the proposed method is less conservative, compared to the results in Su and Wei (1993), and comparable to the Brookmeyer and Crowley's test. The proposed method can be easily extended to adjust for other important (categorical) covariates by stratification. That is, a test statistic can be formed in each subcategory of other covariates and then a stratified test statistic can be formed by combining them. This article considered only inference on the median, but the results can be easily generalized to any quantiles.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We are thankful for thoughtful comments from a coeditor, an associate editor, and a referee, which improved the clarity of the article as well as broadened the scope of it substantially. This research was supported in part by National Health Institute (NIH) grants 5-U10-CA69974-09 and 5-U10-CA69651-11.

REFERENCES

- Basawa, Iv; Koul, HL. Large-sample statistics based on quadratic dispersion. *International Statistical Review*. 1988; 56:199–219.
- Brookmeyer R, Crowley J. A K -sample median test for censored data. *Journal of the American Statistical Association*. 1982; 77:433–440.
- Cunningham D, Humblet Y, Siena S. Cetuximab monotherapy and cetuximab plus irinotecan in irinotecan-refractory metastatic colorectal cancer. *New England Journal of Medicine*. 2004; 351:337–345. [PubMed: 15269313]
- Fisher, RA. *Statistical Methods for Research Workers*. London: Oliver and Boyd; 1932.
- Fisher B, Costantino J, Redmond C. A randomized clinical trial evaluating tamoxifen in the treatment of patients with node negative breast cancer who have estrogen-receptor-positive tumors. *New England Journal of Medicine*. 1989; 320:479–484. [PubMed: 2644532]
- Fisher B, Jeong J, Anderson S. Twenty-five year findings from a randomized clinical trial comparing radical mastectomy with total mastectomy and with total mastectomy followed by radiation therapy. *New England Journal of Medicine*. 2002; 347:567–575. [PubMed: 12192016]
- Gehan EA. A generalized Wilcoxon test for comparing arbitrarily singly censored samples. *Biometrika*. 1965; 52:203–223. [PubMed: 14341275]
- Hoskins PJ, Swenerton KD, Pike JA. Paclitaxel and carboplatin, alone or with irradiation, in advanced or recurrent endometrial cancer: A phase II study. *Journal of Clinical Oncology*. 2002; 19:4048–4053. [PubMed: 11600606]
- Jeong J-H, Jung SH, Costantino JP. Nonparametric inference on median residual life function. *Biometrics*. 2008; 64:157–163. [PubMed: 17501936]
- Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*. 1958; 53:457–481.
- Mood, AM. *Introduction to the Theory of Statistics*. New York: McGraw-Hill; 1950.
- Perry MC, Herndon JE III, Eaton WL. Thoracic radiation therapy added to chemotherapy for small-cell lung cancer: An update of cancer and leukemia group B study 8083. *Journal of Clinical Oncology*. 1998; 16:2466–2467. [PubMed: 9667265]
- Reid N. Estimating the median survival time. *Biometrika*. 1981; 68:601–608.
- Scatterthwaite FE. An approximate distribution of estimates of variance components. *Biometrics Bulletin*. 1946; 2:110–114. [PubMed: 20287815]
- Su JQ, Wei LJ. Nonparametric estimation for the difference or ratio of median failure times. *Biometrics*. 1993; 49:603–607. [PubMed: 8369391]
- Wang J-L, Hettmansperger TP. Two-sample inference for median survival times based on one-sample procedures for censored survival data. *Journal of the American Statistical Association*. 1990; 85:529–536.

Figure 1.

Kaplan–Meier estimates of a simulated dataset from the double exponential distribution as in Brookmeyer and Crowley (1982). This figure appears in color in the electronic version of this article.

Figure 2.

Power comparison among the proposed methods (chi-square and CombP tests) and other methods (log-rank and Gehan's tests) based on simulated datasets from the double exponential distribution as in Brookmeyer and Crowley (1982). This figure appears in color in the electronic version of this article.

Table 1

Empirical type I error probabilities for case I $S_1(t) = S_2(t) = \exp(-t)$

n_i	α	Mean censoring proportions							
		0.43,0.43	0.28,0.28	0.1,0.1	0.01,0.01	0.1,0.28	0.1,0.43		
30	0.05	0.020	0.043	0.050	0.072	0.050	0.040		
	0.1	0.066	0.064	0.083	0.086	0.091	0.084		
	0.15	0.105	0.120	0.156	0.210	0.159	0.124		
50	0.2	0.150	0.184	0.203	0.188	0.192	0.187		
	0.05	0.025	0.035	0.048	0.042	0.055	0.049		
	0.1	0.066	0.092	0.085	0.095	0.082	0.078		
100	0.15	0.103	0.134	0.136	0.125	0.154	0.129		
	0.2	0.159	0.173	0.181	0.206	0.178	0.184		
	0.05	0.027	0.034	0.037	0.050	0.040	0.045		
	0.1	0.077	0.082	0.097	0.108	0.107	0.081		
	0.15	0.110	0.111	0.156	0.132	0.126	0.126		
	0.2	0.155	0.177	0.192	0.214	0.184	0.193		

Table 2

Empirical type I error probabilities for case II $S_1(t) = \exp(-t)$, $S_2(t) = 1 - \Phi(\log(1.44t))$

n_i	α	Mean censoring proportions							
		0.43,0.43	0.28,0.28	0.1,0.1	0.01,0.01	0.1,0.28	0.1,0.43		
30	0.05	0.032	0.044	0.047	0.070	0.043	0.038		
	0.10	0.085	0.093	0.094	0.076	0.093	0.096		
	0.15	0.109	0.143	0.161	0.185	0.122	0.138		
	0.20	0.163	0.191	0.187	0.187	0.212	0.165		
50	0.05	0.030	0.034	0.050	0.046	0.038	0.037		
	0.10	0.077	0.076	0.091	0.100	0.091	0.079		
	0.15	0.108	0.119	0.151	0.125	0.135	0.126		
	0.20	0.157	0.172	0.207	0.197	0.197	0.175		
100	0.05	0.035	0.040	0.044	0.051	0.035	0.052		
	0.10	0.063	0.074	0.096	0.113	0.098	0.076		
	0.15	0.108	0.130	0.140	0.135	0.138	0.114		
	0.20	0.168	0.188	0.189	0.180	0.175	0.177		

Table 3

Empirical powers for $S_1(t) = 1 - \exp(-t)$ and $S_2(t) = 1 - \exp(-t + t^*)$, where t^* is the median difference; significance level = 0.05

n_i	t^*	Mean censoring proportions							
		0.43,0.43	0.28,0.28	0.1,0.1	0.01,0.01	0.1,0.28	0.1,0.43		
50	0.1	0.051	0.076	0.077	0.080	0.078	0.062		
	0.2	0.123	0.126	0.167	0.169	0.168	0.158		
	0.3	0.248	0.291	0.313	0.323	0.312	0.260		
	0.4	0.383	0.448	0.474	0.481	0.450	0.434		
	0.5	0.571	0.636	0.669	0.676	0.653	0.608		
100	0.1	0.091	0.089	0.117	0.134	0.101	0.086		
	0.2	0.243	0.281	0.284	0.305	0.266	0.258		
	0.3	0.468	0.522	0.532	0.565	0.558	0.517		
	0.4	0.717	0.753	0.811	0.812	0.788	0.775		
	0.5	0.892	0.913	0.927	0.937	0.919	0.916		
200	0.1	0.122	0.141	0.169	0.185	0.168	0.128		
	0.2	0.478	0.494	0.496	0.501	0.491	0.486		
	0.3	0.835	0.848	0.847	0.855	0.838	0.831		
	0.4	0.953	0.976	0.977	0.978	0.975	0.970		
	0.5	0.997	0.998	0.998	0.999	0.998	0.998		