Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2015, Vol. 24(6) 909–919

DOI: 10.1177/0962280211432192
smm.sagepub.com

The choice of test in phase II cancer trials assessing continuous tumour shrinkage when complete responses are expected

James MS Wason and Adrian P Mander

Abstract

Traditionally, phase II cancer trials test a binary endpoint formed from a dichotomisation of the continuous change in tumour size. Directly testing the continuous endpoint provides considerable gains in power, although also results in several statistical issues. One such issue is when complete responses, i.e. complete tumour removal, are observed in multiple patients; this is a problem when normality is assumed. Using simulated data and a recently published phase II trial, we investigate how the choice of test affects the operating characteristics of the trial. We propose using parametric tests based on the censored normal distribution, comparing them to the *t*-test and Wilcoxon non-parametric test. The censored normal distribution fits the real dataset well, but simulations indicate its type-I error rate is inflated, and its power is only slightly higher than the *t*-test. The Wilcoxon test has deflated type I error. For two-arm designs, the differences are much smaller. We conclude that the *t*-test is suitable for use when complete responses are present, although positively skewed data can result in the non-parametric test having higher power.

Keywords

censored normal distribution, complete response, continuous tumour endpoint, phase II cancer trial

1 Introduction

Phase II cancer trials of cytotoxic drugs are conducted to test the anti-tumour activity of a novel compound. One-arm studies are still routinely used[1] for sample-size reasons, although randomised two-arm designs are becoming increasingly popular due to high subsequent failure rates of drugs which were successful in one-arm phase II trials.[2]

In early phase studies, the anti-tumour effect of a drug is traditionally measured by the change in the sum of diameters of target lesions. The clinical endpoint used, the tumour response rate, is

Hub for Trials Methodology Research, MRC Biostatistics Unit, Cambridge, UK.

Corresponding author:
James MS Wason, MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 0SR, UK.
Email: james.wason@mrc-bsu.cam.ac.uk

formed from a dichotomisation of this underlying continuous endpoint; patients experiencing a partial or complete response (at least a 30% reduction in the total diameter of target lesions) according to the response evaluation criteria in solid tumors (RECIST) criteria[3] are classed as treatment successes; others as treatment failures. In addition, patients in whom new lesions appear, or for whom non-target lesions grow beyond a certain percentage, are classed as treatment failures.

It is well known in the statistical literature[4–6] that using a dichotomised continuous variable as an endpoint leads to loss of power and therefore higher sample size requirements. The idea of directly using the continuous endpoint to design and analyse cancer trials was originally proposed by Lavin.[7] Karrison et al.[8] discuss issues in the design of such a trial, and recommend their use. Wason et al.[9] quantify the sample size savings for a two-stage trial design, with 50% reductions observed for a previously designed study. Despite the potential savings, the binary endpoint remains the standard endpoint.

One issue arising from using continuous endpoints is that of complete responses. Complete responses signify that lesions are no longer detected and, if one is using the percentage change, thus have the outcome -100%. If there are several patients experiencing complete responses, then multiple observed effects of treatment will be tied at -100%. Complete responses are observed regularly in several areas of cancer, including metastatic breast cancer,[10] recurrent ovarian cancer,[11] metastatic gastric cancer,[12] and carcinomas of unknown primary site.[13] Panageas[14] proposed modelling complete responses separately using a trinomial outcome, but this also discards information contained in the continuous tumour shrinkage. One could test the clinical binary endpoint, but use the continuous tumour shrinkage to improve the power using the method of Suissa.[15,16] This requires the distribution of the continuous endpoint to be known, and so may not perform well when there are complete responses.

When directly modelling the continuous shrinkages, the properties of any statistical method assuming normality of the data, such as the t-test, may deviate from asymptotic properties when there are complete responses. We briefly discuss this issue further in Section 2.1.1.

Karrison et al.[8] suggest that when complete responses are expected, non-parametric tests should be used. However, a well-fitting parametric model could result in power gains. In this article, we investigate a range of straightforward tests for hypothesis testing in phase II cancer trials. We compare the approaches to the Wilcoxon and t-tests, considering one-arm and randomised two-arm trials separately.

As a motivating example, we use a recently published phase II study in metastatic gastric cancer by Park et al.[12] The parametric approaches generally fit well to the observed data. To assess the type-I error rate and power of the various approaches, we use a broad range of simulation scenarios.

2 Methods

2.1 One-arm cancer trials

2.1.1 Notation and hypotheses

In a one-arm cancer trial, n patients receive treatment and are assessed for tumour change. The question of which continuous endpoint to use has not been extensively studied in cancer. The absolute change is not desirable because its variance depends on the baseline value, and the possible change in tumour size will be limited by the baseline value. Lavin[7] recommends using the logarithm of the ratio of the final tumour size to the baseline. This recommendation was based on one relatively small dataset in gastric cancer. Additionally, in the case of complete responses, the observed endpoint will be minus infinity, which makes fitting a parametric distribution problematic. One could specify a truncation point, with observations below being replaced by the truncation point. However, results

could be sensitive to the selection of the truncation point. For these reasons, we use the percentage change in tumour size. The central-limit theorem implies that the mean percentage change is asymptotically normally distributed, although convergence to normality may be slow, especially if a large proportion of observations are complete responses.

Let $X_1,\ldots, X_n$ be the percentage tumour changes for each patient. The hypotheses tested are H_0 : $\delta > \delta_0$ and H_1 : $\delta \leq \delta_0$, where δ is the mean tumour change. If $X_1,\ldots, X_n$ are normally distributed, then H_0 can be tested using a one-sample t-test. When several of the X_i's take the value -100, the normality assumption is violated, and the t-test may not be valid. The one-sample t-test has been shown to be robust to a non-normal distribution when skewness and kurtosis were low,[17] but many complete responses may result in a large skewness.

2.1.2 Wilcoxon test

One-sample non-parametric tests, such as the Wilcoxon-signed ranks test, are used to test differences in the medians rather than the means. This assumes that the underlying population distribution under the null is symmetric. Thus, for directly testing observed tumour size changes, the Wilcoxon test may not be valid, as the lower limit of -100% prevents symmetry when there are several complete responses.

2.1.3 Censored normal distribution

Let X be normally distributed with mean μ and variance σ^2. Then Y, defined as:

$$Y = \begin{cases} X & \text{if} \quad X > c \\ c & \text{if} \quad X \leq c \end{cases}, \tag{1}$$

is a left-censored normal random variable with location parameter μ, scale parameter σ^2 and censoring point c, denoted $CN(\mu, \sigma^2, c)$. We assume the distribution of percentage change in tumour size is $CN(\mu, \sigma^2, -100)$. This is a special case of a tobit model.[18]

The assumption of the model that the observed percentage change in tumour size can go beyond -100%, but is censored. This is wrong, but is a useful construct to fit a convenient model. For interpretation reasons, one may prefer to test the mean of the observed distribution, $m(\mu, \sigma)$, using an estimate of the standard error of m from the censored normal likelihood. It is also possible to test only the location parameter, μ, but this requires eliciting values for the mean and standard deviation of the change. Testing the mean just requires specification of a null value for the mean change.

The log-likelihood of independent identically distributed $CN(\mu, \sigma^2, -100)$ random variables $(Y_1,\ldots, Y_n)$, can be written as:[19]

$$\mathcal{L}(\mu,\sigma) = \log\binom{n}{n_c} + n_c \log\left(\Phi\left(\frac{-100-\mu}{\sigma}\right)\right) + (n-n_c)\log(\sigma\sqrt{2\pi}) - \frac{\sum_{i=1}^{n-n_c}(Y_{(i)}-\mu)^2}{2\sigma^2}, \tag{2}$$

where Φ is the cumulative distribution function of the standard normal distribution, n_c the number of truncated responses and $(Y_{(1)},\ldots, Y_{(n-n_c)})$ the order statistics of the untruncated observations.

By maximising (2), the maximum likelihood estimators $\hat{\mu}$ and $\hat{\sigma}$ can be obtained. The joint distribution distribution of $(\hat{\mu},\hat{\sigma})^T$ will be approximately normal with mean (μ, σ) and variance $\mathcal{I}(\hat{\mu},\hat{\sigma})^{-1}$, where $\mathcal{I}$, the Fisher information, is given in Cohen.[19]

In terms of μ and σ, the mean of the observed distribution is:

$$m(\mu,\sigma) = -100\Phi((-100-\mu)/\sigma) + (1-\Phi((-100-\mu)/\sigma))\left(\mu + \frac{\phi((-100-\mu)/\sigma)\sigma}{1-\Phi((-100-\mu)/\sigma)}\right), \tag{3}$$

where ϕ is the probability density function of the standard normal distribution.

The mean can be estimated by $m(\hat{\mu}, \hat{\sigma})$ and its standard error estimated by the delta method, i.e.:

$$\mathrm{Var}(m(\hat{\mu}, \hat{\sigma})) \approx \bigtriangledown m(\hat{\mu}, \hat{\sigma})^T \mathcal{I}(\hat{\mu}, \hat{\sigma})^{-1} \bigtriangledown m(\hat{\mu}, \hat{\sigma}), \tag{4}$$

where ∇ is the vector of partial derivatives of $m(\mu, \sigma)$ with respect to μ and σ.

For a one-sided test of size α, a confidence interval with coverage $1-2\alpha$ can be obtained by: $(m(\hat{\mu}, \hat{\sigma}) - \Phi^{-1}(1-\alpha)\sqrt{\mathrm{Var}(m(\hat{\mu}, \hat{\sigma}))}, m(\hat{\mu}, \hat{\sigma}) + \Phi^{-1}(1-\alpha)\sqrt{\mathrm{Var}(m(\hat{\mu}, \hat{\sigma}))})$. A one-sided test of size α is to reject inferiority of the new treatment if the upper point of the confidence interval is below the value supposed for the mean under the null hypothesis.

This procedure makes two assumptions. The first is that the non complete-responses can be modelled as a normal distribution truncated below at -100%, and the second that two parameters are sufficient to model both the truncated component of the distribution and the probability of complete response. If the first assumption is incorrect, it may be possible to transform the data so that the normality assumption is valid. Inferences from tobit models are sensitive to deviations from the assumption of normality.[20] Thus, if the second assumption is incorrect, use of the model described in this section could lead to invalid inferences.

2.1.4 Unrestricted probability of complete response

A more flexible distribution is a mixture distribution of a point mass, p, at -100% and a normal distribution truncated below at -100%. The mixture probability is not constrained as it was previously. This is similar to a zero-inflated model with continuous data (the literature on this subject is discussed in Mahmud et al.[21] and Li et al.[22]).

The log-likelihood of this distribution is:

$$\begin{aligned}\mathcal{L}(\mu, \sigma, p) = \log \binom{n}{n_c} + n_c \log(p) + (n - n_c)\log(1-p) - \frac{(n-n_c)}{2}\log(2\sigma^2) \\ - (n-n_c)\log\left(1 - \Phi\left(\frac{-100-\mu}{\sigma}\right)\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n-n_c}(Y_{(i)} - \mu)^2.\end{aligned} \tag{5}$$

Since the log-likelihood decomposes into terms involving p, and terms not involving p, the MLE of p can be estimated immediately as $\frac{n_c}{n}$.

The mean of the observed data will be:

$$m_u(\mu, \sigma, p) = -100p + (1-p)\left(\mu + \frac{\phi((-100-\mu)/\sigma)\sigma}{1 - \Phi((-100-\mu)/\sigma)}\right), \tag{6}$$

from which the hypothesis of inferiority can be tested in a similar manner to before.

2.2 Randomised two-arm cancer trials

2.2.1 Notation

In a two-arm cancer trial, n_1 patients are randomised to the control treatment, and n_2 to the treatment being tested. All patients are assessed for tumour size change. The parameter tested is the mean difference between control and intervention groups, δ, with hypotheses $H_0 : \delta \leq \delta_0$ and $H_1 : \delta > \delta_0$. The trial is powered at $\delta = \delta_1$.

2.2.2 Two-sample parametric tests

The two proposed parametric methods are straightforward to extend to a two-arm study. The estimated difference between arms, together with its standard error can be used to form a confidence interval for the difference in means and a one-sided test of H_0 performed.

2.3 Simulation study

To compare the various approaches, we performed simulation studies under a range of scenarios:

(1) All patients have normally distributed tumour size changes with all changes of below -100% being set to -100%.
(2) Observed tumour changes follow a skew-normal random variable,[28] with all changes of below -100% being set to -100%. The shape parameter is varied to change the direction and magnitude of the skewness.
(3) For each patient, a Bernoulli random variable and a normal random variable truncated below at -100% are generated. If the Bernouli variable is 1, then the patient is a complete response, otherwise they have the continuous component as their observed change.

For each scenario, we examine a range of parameters, assessing the type-I error rate and power of each method at significance level 5%. For each parameter combination, we generate 250000 replicate datasets under the null, and another 250 000 under the alternative. With this many replicates, the Monte-Carlo standard error of the estimated type-I error rate is 0.0004.

For two-arm trials, the null hypothesis tested is no difference in means (t-test, censored-normal, mixture model) or medians (Wilcox) between arms. For one-arm trials, the null hypothesis tested is whether the mean/median is equal to a specified null value. For example, for the first example, the theoretical mean is given in Equation (3), and if the probability of complete response is less than 50%, the median is the location parameter, μ.

3 Results

3.1 Properties of tests when observed data is censored normal

First, we examine the performance of the methods under scenario 1, i.e. that the percentage tumour size changes are normally distributed, and complete responses are those patients whose normal random variable is below -100%.

Parameters varied were μ_0, the location parameter of the censored normal distribution under the null, μ_1, the location parameter under the alternative and σ, the shape parameter of the censored normal distribution. For two-arm trials, under the null hypothesis, both treatments have location parameter μ_0, and under the alternative, the control treatment has location parameter μ_0, with the new treatment having location parameter μ_1. The shape parameter, σ, is assumed to be the same in both arms under both hypotheses.

Values of μ_0 considered were $\{-60, -70, -80\}$, with μ_1 being $\mu_0 - 10$. Values of σ considered were $\{20, 30\}$. For each combination of (μ_0, μ_1, σ), the sample size n was chosen as the minimum number that gave above 80% power at a 5% one-sided significance level using a t-test were the distribution not censored at -100%. For single-arm trials, this gives a sample size of 25 and 56 for $\sigma = 20, 30$ respectively. For randomised trials, it gives a sample size per arm of 50 and 112, respectively. This covers a range of sample sizes that one might expect to see in a well-powered phase II trial.

Table 1 shows the type-I error rate and power of each of the methods proposed for single-arm trials. The type-I error rate is generally higher than nominal for the non-parametric tests and lower than nominal for the parametric tests.

The Wilcoxon test performs well when the probability of complete response is low. The power is only slightly less than that of the parametric methods. As the probability of complete response increases, the type-I error rate and power decrease. The decrease is worse for the larger sample size. Thus, the Wilcoxon-signed rank test should not be applied directly to single-arm data with many complete responses.

The two parametric tests proposed in this article have a higher inflation in type-I error rate than the *t*-test. This is because the parametric tests rely on asymptotics: for the distribution of the maximum likelihood estimators and also for the delta method approximation of the standard error of the observed mean. The inflation is lower for $\sigma = 30$, i.e. when the sample size is higher. There does not appear to be much difference in allowing the complete response to be unconstrained, with a very slight decrease in type-I error rate and power.

The *t*-test generally performs well. It has the lowest deviation from the nominal significance level of all the methods, and does not lose much power compared to the two more sophisticated parametric methods. It is also much simpler to apply.

Table 2 provides results for randomised trials. For this set of simulations, there is very little difference in the performance of each method for randomised trials. All methods have close to nominal type-I error rate, although the two parametric methods are still somewhat inflated.

All results in Table 2 assume the same variance in each arm. This may not be the case, but the only method which explicitly assumes it is the two-sample *t*-test. In the case of both the variance and the sample size differing in each arm, the type-I error rate and power can be affected.[24,25] We also considered using a *t*-test which allows unequal variances. This showed no loss in power, therefore we recommend it for its additional robustness.

3.2 Properties of tests when observed data is censored skew-normal

Next, we simulated datasets where the outcomes were censored skew-normal random variables. We only considered the case where the mean change under the null and alternative are −70% and −80%, respectively, and the standard deviation of the underlying skew-normal random variables is 30. The shape parameter, was varied between −10 and 10 in increments of 1.25. For each value of

Table 1. Type-I error rate and power under simulation scenario 1 for single-arm trials.

				Type-I error rate				Power			
Parameters	n	$\mathbb{P}(CR\|H_0)$	$\mathbb{P}(CR\|H_1)$	W	T	CN	CNU	W	T	CN	CNU
(60,70,20)	25	0.023	0.067	0.048	0.053	0.063	0.062	0.758	0.784	0.812	0.811
(60,70,30)	56	0.091	0.159	0.040	0.055	0.060	0.059	0.736	0.789	0.802	0.799
(70,80,20)	25	0.067	0.159	0.045	0.056	0.067	0.067	0.744	0.785	0.813	0.810
(70,80,30)	56	0.159	0.252	0.027	0.058	0.064	0.063	0.648	0.780	0.796	0.790
(80,90,20)	25	0.159	0.309	0.035	0.064	0.076	0.074	0.673	0.777	0.806	0.802
(80,90,30)	56	0.252	0.369	0.010	0.063	0.069	0.067	0.435	0.766	0.783	0.775

Method abbreviations: W – Wilcoxon signed rank test; T – *t*-test; CN – censored normal; and CNU – censored normal with unrestricted probability of complete response.

Table 2. Type-I error rate and power under simulation scenario 1 for randomised trials.

				Type-I error rate					Power				
Parameters	$2n$	$\mathbb{P}(CR\|H_0)$	$\mathbb{P}(CR\|H_1)$	W	TS	TD	CN	CNU	W	TS	TD	CN	CNU
(60,70,20)	100	0.023	0.067	0.049	0.049	0.049	0.053	0.054	0.778	0.794	0.794	0.806	0.807
(60,70,30)	224	0.091	0.159	0.051	0.050	0.050	0.052	0.053	0.778	0.784	0.784	0.792	0.791
(70,80,20)	100	0.067	0.159	0.050	0.050	0.050	0.054	0.055	0.777	0.786	0.786	0.799	0.798
(70,80,30)	224	0.159	0.252	0.050	0.050	0.050	0.052	0.053	0.773	0.771	0.771	0.780	0.779
(80,90,20)	100	0.159	0.309	0.050	0.050	0.050	0.054	0.055	0.769	0.766	0.766	0.781	0.779
(80,90,30)	224	0.252	0.369	0.050	0.050	0.050	0.052	0.053	0.759	0.748	0.748	0.758	0.756

Method abbreviations: W – Wilcoxon rank-sum test; TS – *t*-test assuming same variance between arms; TD – *t*-test allowing difference variances; CN – censored normal; CNU – censored normal with unrestricted probability of complete response.

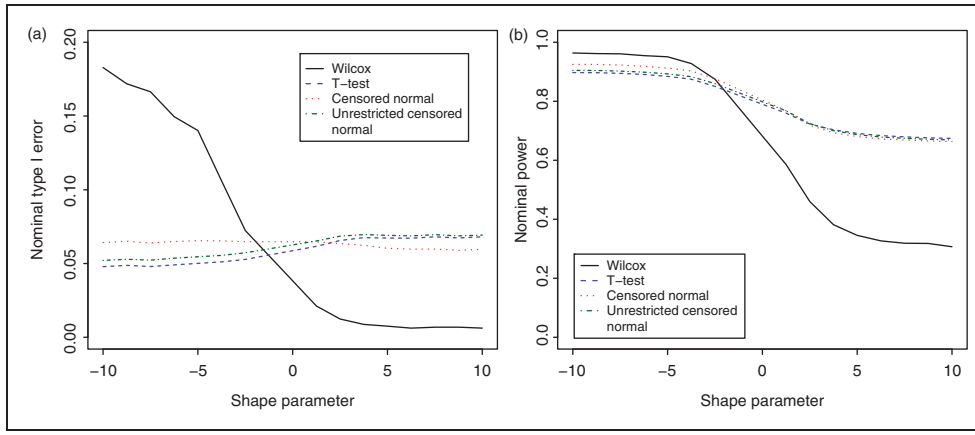

Figure 1. Type-I error rate and power of the different approaches as the shape parameter changes, for simulation scenario 2.

the shape parameter, location and scale parameters were found under the null and alternative such that the means and standard deviation were as specified.

Figure 1(a) and (b) show the type-I error rate and power, respectively, of the Wilcoxon test and the three parametric approaches, as the shape parameter varies, for single-arm trials.

The Wilcoxon test is the most sensitive to non-zero skewness of the four approaches. The type-I error rate is very large for large negative skewness, and very low for large positive skewness. This may be partly due to deviations from the assumption of a symmetric distribution. As in the first scenario, there does not seem to be any advantage in using the more sophisticated parametric procedures over that of using the *t*-test. There is a difference in the type-I error rate and power, but no single technique shows a consistent advantage.

As in the previous simulation scenario, there is much less difference between the techniques for randomised trials (data not shown). Again, the two censored normal approaches had higher deviations from the nominal significance level – lower than nominal for negative skewness and higher than nominal for zero to positive skewness. An observation of interest was that the power

of the Wilcoxon test was generally lowest at zero skewness. The power increased as the skewness deviated from zero. The parametric techniques have highest power for large negative skewness, with the power decreasing monotonically as the shape parameter increased. This is likely due to the probability of complete response increasing as the skewness goes from positive to negative.

3.3 Properties of tests when complete response probability is unconstrained

We next consider the third scenario – where the probability of complete response is independent of the continuous tumour change.

The scale parameter for the truncated normal component is set to be 30, with the location parameters under the null and alternative being chosen so that the mean tumour change, including complete responses, is -70 and -80, respectively.

We examined the case where the difference between p_1, the probability of complete response under the alternative, and p_0, the probability under the null, was constant. The difference was fixed at the difference when determined by the tail-probabilities of the normal distribution. With means -70, -80 and standard deviation 30, p_0 and p_1 are equal to 0.159 and 0.252, respectively. Thus, we fix the difference between p_0 and p_1 at 0.093, but vary p_0. Values of p_0 considered were $\{0.05, 0.1, 0.15, 0.2, 0.25\}$.

In single-arm cases, we found that the type-I error rate of the Wilcoxon test increased modestly with p_0, from 0.016 at $p_0 = 0.05$ to 0.022 at $p_0 = 0.25$. The type-I error rate of the restricted censored normal approach also increased at about the same relative rate - from 0.055 to 0.073. However, both the t-test and the unrestricted censored normal approach maintained their type-I error rate at around 0.06 and 0.065, respectively as p_0 increased. The power showed a large gap, implying again that for single-arm trials it would be a mistake to apply the Wilcoxon test to datasets with complete responses.

A surprising result was that the unrestricted censored normal test showed no power advantage. Its sole advantage was that the type-I error rate was closer to nominal. Even then the t-test showed less inflation in type-I error rate and around the same power.

For randomised trials, the only observation of note was that as p_0 increased, the power of the Wilcoxon test dropped considerably more than the parametric approaches. At $p_0 = 0.05$, the gap was around 0, but at $p_0 = 0.25$, it was 5%. Again, there was no noticeable advantage in using the censored normal approaches over using the t-test. Of the two censored approaches, the deviation from 5% significance level was lower using the unconstrained probability of complete response.

3.4 Case study

As a case study, we use published data from Park et al.,[12] a phase II study in metastatic gastric cancer. The trial had one arm, the treatment being a triplet regimen of S-1, irinotecan and oxaliplatin. A total of 41 patients were assessed, 7 of whom were complete responses.

As currently carried out, phase II cancer trials do not test hypotheses about the continuous change, so a null value being tested is not available. Instead, we compare the confidence intervals for the observed mean from the t-test and the two censored normal approaches as well as the confidence interval for the observed median from the Wilcoxon test.

First, we compare the fit of the two censored normal approaches to the observed data. Figure 2 shows the empirical cumulative density function and the model fit from each approach.

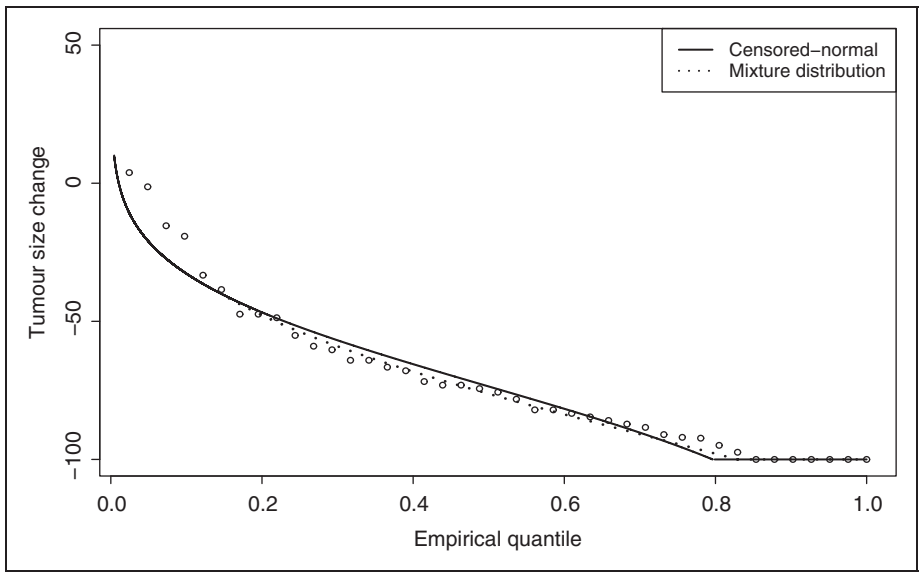

Figure 2. Observed quantiles of tumour change data from case-study, and fitted quantiles from the censored normal and unrestricted censored normal models.

Table 3. Summary of model fit for the two censored normal procedures

	Restricted censored normal	Unrestricted censored normal
Log-likelihood	−156.1	−155.3
AIC	316.3	316.6
$\mathbb{P}$(CR) (standard error)	0.203 (0.063)	0.171 (0.059)

Figure 2 shows little noticeable difference in the fit of the two censored normal approaches. Both tend to fit well in the middle of the distribution, but not at the left tail. The unrestricted model fits better in the right tail, although the restricted model appears to model the probability of complete response sufficiently well.

Table 3 shows the log-likelihood, Akaike information criterion (AIC) and fitted probability of complete response for each model.

In terms of the log-likelihood and the AIC, there is little difference between the fit of the models for this dataset. There is a small difference in the fitted probability of complete response, but the standard error of the estimate for the restricted model is 0.063, so the confidence interval includes 0.171, the fitted probability of complete response under the unrestricted model.

The standard errors of the mean tumour change for the restricted and unrestricted models are 4.11 and 4.28, respectively using the delta method. Similar results were found using 5000 bootstrap samples (standard errors of 4.26 and 4.23, respectively).

The 95% confidence intervals for the mean tumour change are (−77.7%, −61.6%) from the restricted censored normal approach, (−78.8%, −62.1%) from the unrestricted approach and (−79.30%, −61.8%) from the *t*-test (i.e. by assuming the data is normally distributed). The 95% confidence interval for the median, from the Wilcoxon test, is (−82.1%, −64.1%). The confidence

intervals for the mean do not vary greatly between the methods, as one would expect from the simulation results. However, the width of the confidence interval from the *t*-test is slightly wider.

4 Discussion

In this article, we have considered which test to use for comparing tumour size changes (either between two groups, or to historical data) in phase II cancer trials when complete responses are expected. For randomised two-arm designs, there appears to be little difference in the performance of each test examined. Using the Wilcoxon rank-sum test provides equally good power to using the *t*-test. This agrees with earlier work on the two-sample case by Lachenbruch.[26,27] The two more sophisticated parametric approaches we examined have slightly inflated type-I error rates.

For one-arm designs, which are common in phase II cancer trials, there is more of a difference. The Wilcoxon-signed rank test has a lower than nominal type-I error rate, and a loss in power. The *t*-test shows a slight inflation in type-I error rate. However, the degree of inflation is lower than that of the two parametric approaches. Although the two parametric approaches provide a slight increase in power, this increase is less than 1%. Because of the extra complexity and the small gain, we would still recommend the *t*-test for use in practice.

One factor that makes a difference is the skewness of the data. In single-arm trials, negative skewness resulted in the Wilcoxon test having the lowest power of the approaches, whereas positive skewness meant it had the highest. For two-arm trials, the difference was again much smaller.

The censored parametric methods may be improved by testing the underlying location parameter instead of the observed mean. For single-arm trials, this would be difficult because it would require a null value for the location parameter to be specified before the trial. For two-arm trials, this is not required. It may still be not desirable however, as a difference in location parameter between arms would not be as interpretable as a difference in observed means, and would be more difficult to communicate to non-statisticians.

This article has considered cytotoxic agents, which are designed to shrink the tumour. More recently, cytostatic agents have been introduced, which are designed to control the tumour growth. For trials of cytostatic agents, complete responses are unlikely, and so the conclusions presented in this article are less relevant. However, cytotoxic drugs continue to be developed and tested.

The methods in this article assume that the only endpoint of interest is the change in size of target lesions. Real trials are often more complex, with death and toxicity occuring, which causes patients to drop out. This study presents important conclusions about modelling target tumour size when complete responses are present. Incorporating these conclusions into more complicated models that allow for death and toxicity is the subject of ongoing work.

Funding

The authors are funded by the medical research council, grant MC_US_A030_0035.

References

1. Lee J and Feng L. Randomized phase II designs in cancer clinical trials: current status and future directions. *J Clin Oncol* 2005; **23**: 4450–4457.
2. Tang H, Foster N, Grothey A, Ansell S, Goldberg R and Sargent D. Comparison of error rates in single-arm versus randomized phase II cancer clinical trials. *J Clin Oncol* 2010; **28**: 1936–1941.
3. Eisenhauer E, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer* 2009; **45**: 228–247.
4. MacCallum R, Zhang S, Preacher K and Rucker D. On the practice of dichotomization of quantitative variables. *Psychol Meth* 2002; **7**: 19–40.

5. Royston P, Altman D and Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med* 2006; **25**: 127–141.
6. Altman D and Royston P. The cost of dichotomising continuous variables. *BMJ* 2006; **332**: 1080.
7. Lavin P. An alternative model for the evaluation of antitumor activity. *Cancer Clin Trials* 1981; **4**: 451–457.
8. Karrison T, Maitland M, Stadler W and Ratain M. Design of phase II cancer trials using a continuous endpoint of change in tumour size: application to a study of sorafenib and erlotinib in non-small-cell lung cancer. *JNCI* 2007; **99**: 1455–1461.
9. Wason J, Mander A and Eisen T. Reducing sample sizes in two-stage phase II cancer trials by using continuous tumour shrinkage endpoints. *Eur J Cancer* 2011; **47**: 983–989.
10. Seidman A, Hudis C, Albanel J, et al. Dose-dense therapy with weekly 1-hour paclitaxel infusions in the treatment of metastatic breast cancer. *J Clin Oncol* 1998; **16**: 3353–3361.
11. Audeh M, Carmichael J, Penson R, et al. Oral poly(ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and recurrent ovarian cancer: a proof-of-concept trial. *Lancet* 2010; **376**: 245–251.
12. Park S, Yong SY, Rhee J, et al. Phase II study of a triplet regimen of S-1 combined with irinotecan and oxaliplatin in patients with metastatic gastric cancer: clinical and pharmacogenetic results. *Ann Oncol* 2011; **22**: 890–896.
13. Moller A, Pedersen K, Gothelf A and Daugaard G. Paclitaxel, cisplatin and gemcitabine in treatment of carcinomas of unknown primary site, a phase II study. *Acta Oncol* 2010; **49**: 423–430.
14. Panageas K, Smith A, Gonen M and Chapman P. An optimal two-stage phase II design utilizing complete response information separately. *Controlled Clin Trials* 2002; **23**: 367–379.
15. Suissa S. Binary methods for continuous outcomes: a parametric alternative. *J Clin Epidemiol* 1991; **44**: 241–248.
16. Suissa S and Blais L. Binary regression with continuous outcomes. *Stat Med* 1995; **14**: 247–255.
17. Cressie N, Sheffeld L and Whitford H. Use of the one sample *t*-test in the real world. *J Chron Dis* 1984; **37**: 107–114.
18. Tobin J. Estimation of relationships for limited dependent variables. *Econometrica* 1958; **26**: 24–36.
19. Cohen A. Simplified estimators for the normal distribution when samples are singly censored or truncated. *Technometrics* 1959; **1**: 217–237.
20. Amemiya T. Tobit models: a survey. *J Econometrics* 1984; **24**: 3–61.
21. Mahmud S, Lou W and Johnston N. A probit- log- skew-normal mixture model for repeated measures data with excess zeros, with application to a cohort study of paediatric respiratory symptoms. *BMC Med Res Method* 2010; **10**: 55.
22. Li N, Elashoff D, Robbins W and Xun L. A hierarchical zero-inated log-normal model for skewed responses. *Stat Meth Med Res* 2011; **20**: 175–189.
23. Azzalini A. A class of distributions which includes the normal ones. *Scand J Stat* 1985; **12**: 171–178.
24. Boneau C. The effects of violation of assumptions underlying the t test. *Psychol Bull* 1960; **57**: 49–64.
25. Zimmerman D. Comparative power of the student t test and Mann-Whitney u-test for unequal sample sizes and variances. *J Exp Edu* 1987; **55**: 171–174.
26. Lachenbruch P. Comparisons of two-part models with competitors. *Stat Med* 2001; **20**: 1215–1234.
27. Lachenbruch P. Analysis of data with excess zeros. *Stat Meth Med Res* 2002; **11**: 297–302.