Designing exploratory cancer trials using change in tumour size as primary endpoint

Thomas Jaki[1], Valérie André[2], Ting-Li Su[1], John Whitehead[1]

[1]*Medical and Pharmaceutical Statistics Research Unit, Lancaster University, UK*

[2]*Eli Lilly and Company, Erlwood Manor, Sunninghill Road, Windlesham, UK*

27 September 2012

FOR SUBMISSION TO STATISTICS IN MEDICINE

Address for correspondence: Thomas Jaki, MPS Research Unit, Fylde College, Lancaster University, Lancaster LA1 4YF, UK, Tel: +44 1524 592318, email: jaki.thomas@gmail.com

Summary

In phase III cancer clinical trials, overall survival is commonly used as the definitive endpoint. In phase II clinical trials, however, more immediate endpoints are generally used such as incidence of complete or partial response within one or two months or progression-free survival (PFS). Due to the limited ability to detect change in overall survival with response, the inherent variability of PFS and the long wait for progression to be observed, more informative and immediate alternatives to overall survival are desirable in exploratory phase II trials. In this paper we show how comparative trials can be designed and analysed using change in tumour size as the primary endpoint. The test developed is based on the framework of score statistics and will formally incorporate the information of whether patients survive until the time at which change in tumour size is assessed. Using an example in non-small cell lung cancer, we show that the sample size requirements for a trial based on change in tumour size are favourable compared to alternative randomized trials and demonstrate that these conclusions are robust to our assumptions.

1. Introduction

In confirmatory clinical trials in cancer, overall survival (OS) is commonly used as the definitive endpoint. In early phase trials of cancers involving solid tumours, more immediate measures of tumour shrinkage such as incidence of complete or partial response within one or two months or progression-free survival (PFS) are used. Each of these endpoints has been criticised for different shortcomings. The value of tumour response has been questioned for drugs acting to stabilize tumour size and more generally after several instances where high response rates did not

translate into improved survival [1]. PFS on the other hand suffers from the long wait for progression to be observed and is heavily influenced by the assessment schedules [2]. Consequently more informative and immediate endpoints capable of predicting overall survival are desirable in exploratory (Phase II) trials.

A natural alternative to the endpoints above is to use the tumour size itself as the primary endpoint in the early stages of development. The RECIST guideline [3,4] recommends the use of the sum of the largest diameters of solid tumours as a measurement of the size of a tumour while various authors have, for different types of cancer, shown that tumour size is related to survival [5, 6]. Moreover it has been conjectured that change in tumour size can be used to aid decisions about proceeding to a confirmatory trial in non-small cell lung cancer [7] and approaches for designing and analysing trials using continuous tumour size measurements have been discussed [8,9]. More recently it has been demonstrated that using tumour size in 2-stage single armed cancer trials can reduce the required sample size by approximately one third [10]. It is important to note here that to determine tumour response or PFS, the tumour size has to be measured anyway. Using tumour size as primary endpoint does therefore not require any additional data to be collected beyond what is currently recorded.

In this manuscript we derive the methodology necessary to design and analyse a comparative trial based on change in tumour size (Section 2). The method is based on score statistics and, unlike previous work [8-10], formally incorporates whether a patient survives until the time of tumour size assessment. The methodology is constructed so that the effect on change in tumour size is a function of the treatment effect in terms of overall survival, and particular focus will be

given to designing such a trial. In Section 3 we compare the sample size requirements for different early phase cancer trial design alternatives in the context of a trial in non-small cell lung cancer. We relate the effects of treatment on change in tumour size and on PFS to the treatment effect on OS, using pre-existing data in which all of these measures are available to ensure a fair comparison between the methods. A thorough sensitivity analysis is included to evaluate the robustness of the sample size comparison. The manuscript concludes with a brief discussion and further directions.

2. Designing and analysing trials based on change in tumour size

For a trial based on change in tumour size, the extent of the tumour is measured immediately prior to treatment at time t_0, and after start of the treatment at time, t_1 (often one or two months later). The analysis of such a trial then comprises of two separate components: surviving until time t_1 (yes or no) and the change in tumour size for patients who survived until t_1. To establish whether a novel treatment yields an improvement over control we will model the treatment difference in respect of both components using a single parameter. To do so we require historical data on both tumour size and survival for each patient. As the dataset is used to relate both endpoints to overall survival the patients in this dataset should have received the same treatment as the control group of the forthcoming phase II study. As pointed out above, tumour size is currently collected in most phase II cancer trials as it is used to determine both tumour response and PFS, while survival time is a standard secondary endpoint in these studies. As a result a suitable historical dataset will often be available.

After the study has been completed, the following data are available for the i^{th} patient: x_i –

treatment indicator, 0 for control treatment (C) and 1 for experimental treatment (E); $\mathbf{z}_i$ – vector of prognostic factors (baseline tumour size, …); s_i – survival until t_1, 1 for no and 0 for yes; d_i – change in tumour size, cts (missing if $s_i = 0$). We denote the vector of the treatment indicators for all subjects by $\mathbf{x} \equiv (x_1, \ldots, x_n)$ and similarly define **s** and **d** as the vectors of survival indicators and changes in tumour size, respectively. The matrix of prognostic factors will be written as **z**. Note further that patients dropping out of the study will be given a survival indicator of one (i.e treated as deaths). This is to ensure that the resulting treatment effect estimate is pessimistic rather than overly optimistic and is in line with the standard way of handling drop-outs in a trial based on response.

The parameter expressing the difference between treatments in the proportions of patients dying before the second visit, θ_{psurv}, can be estimated using a generalized linear model relating the s_i to x_i and $\mathbf{z}_i$. An estimate for the treatment effect on change in tumour size, $\hat{\theta}_{\text{cts}}$, where θ_{cts} is the parameter expressing the difference in mean tumour size measurements, will be found using a linear regression model of d_i on x_i and $\mathbf{z}_i$ for all patients surviving until t_1. Specifically, $\theta_{\text{cts}} = \mu_E - \mu_C$ with μ_E and μ_C denoting the mean change in tumour size in the experimental and control group, respectively.

For large enough samples, $\hat{\theta}_i \sim N\left\{\theta_i, se\left(\hat{\theta}_i\right)^2\right\}$ for i = psurv, cts and $\hat{\theta}_{\text{psurv}}$ and $\hat{\theta}_{\text{cts}}$ are independent provided prognostic factors, $\boldsymbol{z}$, have been accounted for in the estimation of the parameters. This follows because the joint density of $\boldsymbol{s}$ and $\boldsymbol{d}$ satisfies:

$$f(\boldsymbol{s}, \boldsymbol{d}|\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \boldsymbol{x}, \boldsymbol{z}) = f_1(\boldsymbol{s}|\boldsymbol{\beta_1}, \boldsymbol{x}, \boldsymbol{z}) f_2(\boldsymbol{d}|\boldsymbol{\beta_2}, \boldsymbol{s}, \boldsymbol{x}, \boldsymbol{z})$$

where β_1 are the parameters governing the generalized linear model of survival until t_1 and β_2 are

the parameters associated with the regression of changes in tumour size amongst the survivors. These two measures can now be put into the familiar framework of score tests by defining $Z_i = \hat{\theta}_i / se(\hat{\theta}_i)^2$ which for large n and small θ_i yields $Z_i \sim N(\theta_i V_i, V_i)$, where $V_i = 1/\{se(\hat{\theta}_i)\}^2$ is the information.

In order to obtain a single test statistic that can be used to establish whether a novel treatment yields an improvement over control, the two score tests above need to be related to a single parameter. As the objective of the treatment is to extend overall survival, this single parameter is taken to be θ_{OS}, the negative of the log hazard ratio in terms of overall survival.

As pointed out above, an estimate of the effect for survival until time t_1 can be obtained in terms of the log-odds ratio for a treated person relative to an untreated person by fitting a generalized linear model of s_i on x_i and $\mathbf{z}_i$. If the complementary log-log link is used in this model, then the survival parameter, θ_{psurv}, defined as

$$\theta_{psurv} = \log\{-\log(1 - p_C)\} - \log\{-\log(1 - p_E)\}$$

where p_C and p_E are the proportions of subjects dying at time t_1 on control and experimental treatment, respectively, is exactly equal to θ_{OS}, the log-hazard ratio for overall survival [11, Chapter 9]. Note that, even if the logit link is used this relationship holds approximately as p_E and p_C are typically small. Consequently we can write the distribution of the first score statistic as $Z_{psurv} \sim N(\theta_{OS} V_{psurv}, V_{psurv})$.

To find a similar relationship between overall survival and change in tumour size we use a Cox-proportional hazards model of the form

$$h(t \mid \mathbf{z}_i, cts) = h_0(t)e^{\boldsymbol{\beta}'\mathbf{z}_i + \beta_{cts} \times cts},\ t > t_1$$

where $\mathbf{z}_i$ is the vector of covariates and $\boldsymbol{\beta}$ the vector of associated coefficients using the existing historical dataset. Note that β_{cts} will be negative if a smaller tumour size corresponds to smaller hazard (see Eqn (2) below) and that the model will be based on only patients who survive to t_1. Provided that there are no interactions between treatment and the prognostic factors in the model, the hazard functions of the two treatment groups, will be related via $h_E(t \mid \mathbf{z}_i, cts) = e^{-\theta_{OS}} h_C(t \mid \mathbf{z}_i, cts)$, where h_E corresponds to the hazard function of the experimental group and h_C to that of the control group. A patient who would have a change in tumour size equal to cts if put on the control treatment, will therefore *typically* have change in tumour size of cts + θ_{cts} if put onto experimental treatment. It follows that, approximately,

$$\begin{aligned} h_E(t \mid \mathbf{z}_i, cts) &= h_0(t)e^{\boldsymbol{\beta}'\mathbf{z}_i + \beta_{cts}(cts + \theta_{cts})} \\ &= e^{\beta_{cts}\theta_{cts}} h_C(t \mid \mathbf{z}_i, cts). \end{aligned}$$

Note that in the above we are assuming that any treatment effect on overall survival (conditional on survival to t_1) is captured by the effect of treatment on change in tumour size. The change in tumour size parameter can therefore be expressed in terms of a hazard ratio for overall survival as $e^{-\theta_{OS}} = e^{\beta_{cts}\theta_{cts}}$ so that

$$\theta_{cts} = \frac{\theta_{OS}}{-\beta_{cts}}. \qquad (1)$$

The distribution of the score statistic relating to change in tumour size can be written as

$$Z_{cts} \sim N\left(\frac{\theta_{OS}}{-\beta_{cts}} V_{cts}, V_{cts}\right).$$

A natural choice to constructing the joint test statistic is then

$$Q = Z_{psurv} + \frac{1}{-\beta_{cts}} Z_{cts}.$$

The division by $-\beta_{\text{cts}}$ arises from the desire to construct a dimensionless test statistic, Q: Z_{cts}, carries forward the dimension of β_{cts} by construction resulting in different values of Z_{cts} if (for example) tumour size is measured in mm or in inches. Moreover the distribution of the test statistic then has the familiar structure,

$$Q \sim N\left\{\theta_{\text{OS}}\left(V_{\text{psurv}} + \frac{1}{\beta_{\text{cts}}{}^2} V_{\text{cts}}\right), \left(V_{\text{psurv}} + \frac{1}{\beta_{\text{cts}}{}^2} V_{\text{cts}}\right)\right\}.$$

The null hypothesis of no difference between control and experimental treatment is then rejected if Q exceeds the critical value *u*. At the design stage, the critical value, *u*, and the sample size, *n*, can be determined through solving

$$\left(V_{\text{psurv}} + \frac{1}{\beta_{\text{cts}}{}^2} V_{\text{cts}}\right) = \left(\frac{z_{1-\alpha} + z_{1-\beta}}{\theta_{\text{OS}}}\right)^2$$

and

$$u = z_{1-\alpha} \sqrt{V_{\text{psurv}} + \frac{1}{\beta_{\text{cts}}{}^2} V_{\text{cts}}}.$$

Using a R:1 randomization between experimental and control treatment and denoting the total sample size of the trial by *n* gives $n_E = Rn/(R+1)$ and $n_C = n/(R+1)$ which are the sample sizes of the experimental and control arm, respectively. Using the complementary log-log link the pair (Z_{psurv}, V_{psurv}) can, for large *n*, be found as

$$Z_{\text{psurv}} = \frac{nRq}{(R+1)(\hat{p}_C + R\hat{p}_E)} (\hat{p}_C - \hat{p}_E)$$

$$V_{\text{psurv}} = \frac{nRq^2(1-\bar{p})}{(R+1)^2\bar{p}}$$

[e.g. 12, p. 42]. The quantity $\hat{p}_C - \hat{p}_E$ is the change in proportion of patients dying before t_1 on control versus experimental treatment, $\bar{p}$ is the average proportion of patients dying before t_1 and $q = -\log(1-\bar{p})$. Note that, although we have allowed for covariates when we derived the

relationship between survival and change in tumour size, the above expression does not include them. When finding the relationship $\theta_{cts} = \theta_{OS}/(-\beta_{cts})$ based on existing historical dataset, inclusion of covariates ensures that the effect, θ_{cts}, can be estimated more precisely. When designing a new trial, however, the nature of the covariate structure of the dataset to be collected is unknown and hence proceeding without such a structure is sensible. For the purpose of analysing a trial in which case covariate adjustment is likely to be of interest, expressions for Z_{psurv} and V_{psurv} can be found for example in [12].

For the analysis of change in tumour size for patients surviving past t_1 we assume that the change in tumour size measure is normally distributed with equal variance, σ^2, in both arms. The parameter of interest can then be defined as $\theta_{cts} = \mu_E - \mu_C$ with μ_E and μ_C denoting the mean change in tumour size in the experimental and control group, respectively. The sampling distribution of the estimate obtained by taking the difference in sample means is then

$$\hat{\theta}_{cts} \sim N\left\{\mu_E - \mu_C, \frac{(R+1)^2\sigma^2}{Rn}\right\},$$

so that the corresponding score statistic, obtained without covariate adjustment for the purpose of designing the study, is $Z_{cts} = \hat{\theta}_{cts}/\{se(\hat{\theta}_{cts})\}^2$, and the information is $V_{cts} = nR/\{\sigma^2(R+1)^2\}$. Combining the results we can determine the necessary sample size by solving

$$\left(V_{psurv} + \frac{1}{\beta_{cts}{}^2} V_{cts}\right) = \left(\frac{z_{1-\alpha} + z_{1-\beta}}{\theta_{OS}}\right)^2$$

$$\frac{nRq^2(1-\bar{p})}{(R+1)^2\bar{p}} + \frac{1}{\beta_{cts}{}^2}\frac{nR}{\sigma^2(R+1)^2} = \left(\frac{z_{1-\alpha} + z_{1-\beta}}{\theta_{OS}}\right)^2$$

which after rearranging gives

$$n = \frac{(R+1)^2}{R}\left(\frac{z_{1-\alpha} + z_{1-\beta}}{\theta_{OS}}\right)^2 \frac{\bar{p}\beta_{cts}{}^2\sigma^2}{q^2(1-\bar{p})\beta_{cts}{}^2\sigma^2 + \bar{p}}.$$

3. Comparing sample size requirement for trials using different endpoints

Having described how a trial based on tumour size can be designed and analysed, we now compare the sample size requirements of such a trial to alternative Phase II endpoints in the context of a particular therapeutic area. We focus on comparisons with trials based on tumour response and progression-free survival, and for completeness include some comparison with single-armed trials based only on tumour response.

The main challenge in making the comparison is to ensure comparable sizes of treatment effect. As illustrated in Figure 1, we will anchor our comparisons around meaningful differences in overall survival which can be translated into the relevant effects for surviving to t_1 and change in tumour size as described in the previous section. Then in Section 3.2 we analytically derive the effect size in terms of tumour response from these measures, while a proportional hazards model is used to relate PFS and change in tumour size.

~~~ FIGURE 1 ~~~

Using the equivalent effect sizes, the various sample size requirements will be compared (Section 3.3) and the sensitivity to assumptions made when deriving equivalent effect sizes evaluated (Section 3.4).

3.1. Dataset
~~~

The following comparison is based upon a dataset of 288 patients with non-small cell lung cancer (NSCLC). The data are the control arm, receiving standard of care for NSCLC, of a recent randomized phase III trial. The dataset includes the main variables overall survival and progression-free survival as well as sum of the largest diameter of up to 10 lesions at the initial visit, S_0, and after 8 weeks, S_8, defined according to RECIST 1.0 [3]. The change in tumour size for an individual patient is defined as

$$\text{cts} = \log\{(1 + S_0)/(1 + S_8)\} = \log(1 + S_0) - \log(1 + S_8), \qquad (2)$$

being the shifted log-difference of tumour size. This definition was chosen so that cts is positive for a positive outcome (i.e. a reduction in tumour size) and because the distribution of cts defined in this way was found to be approximately normal in the dataset studied. The shift by 1 caters for patients with a tumour size of zero (complete response). Note that the methods described in this paper are also applicable to other measures of change in tumour size, provided that they can be modelled as normally distributed, or for large sample sizes an asymptotic normal distribution can be assumed for the test statistic.

The following covariates were included as prognostic factors in the models used to find equivalent effect sizes: sum of the largest diameter of up to 10 lesions at visit 0 (which will be referred to as baseline), ECOG-score [13] and gender. Other variables such as age, though available, were not used in the models as they did not contribute significantly to their fit. Patients who had either an ECOG score greater than 1 or did not have a baseline tumour size measurement were excluded, leaving 225 patients for model building and relating effect sizes.

3.2. Effect sizes

In Section 2 we discussed how effect sizes for change in tumour size based on a desired effect in terms of hazard ratio for overall survival can be found. In particular it was argued that the effect on surviving to 8 weeks is equivalent to the hazard ratio for overall survival, θ_{OS}, while the effect in change in tumour size is found to be $\theta_{cts} = \theta_{OS}/(-\beta_{cts}) = \theta_{OS}/2.2159$ in the dataset on NSCLC. A hazard ratio of 0.8 (θ_{OS} = 0.2231) therefore corresponds to an average change in tumour size of 0.10. In this section we will relate the effect sizes of response and PFS to overall survival.

To find the effect in response for a given effect in overall survival, we consider the main criterion for partial or complete response as given in RECIST [4] states that a patient is said to have responded if a reduction in the sum of diameters of target lesions of at least 30% over baseline is achieved. In our investigation, a patient is therefore classed as having responded if $(S_0 - S_8)/S_0 > 0.3$. All other patients, including patients whose tumour size evaluation after 8 weeks is missing, are considered non-responders. Note that the definition of response used is a slight simplification of the RECIST criteria as it does not incorporate an increase in the number of lesions. This simplification is expected to have little influence on the sample size for a trial based on response as the difference between the response rates will likely remain the same. In addition no patients with a sufficiently large reduction in tumour size had an increase in the number of lesions in the NSCLC dataset so that this simplification did not have an impact on the numbers in our example.

Using the above definition, we can estimate the proportion of patients responding on control from the data. To estimate the response rate under the experimental treatment we will use some

generated data under the desired effect on overall survival. In particular we start by considering the parameter θ_{cts} which is defined as the expected difference in cts values between the two treatments, so that $\theta_{cts} = E\{\log(1 + S_{8C}) - \log(1 + S_{8E})\}$, where S_{8E} is the sum of largest diameters 8 weeks after start of treatment for a patient in the experimental group and S_{8C} is the corresponding quantity for the control group. The expected tumour size 8 weeks after start of the experimental treatment with effect θ_{cts} can be found as $E(S_{8E}) = \exp\{\log(1 + S_{8C}) - \theta_{cts}\} - 1$. The response rate under experimental treatment can then be found from the expected week 8 tumour size by applying the definition of response for each subject and finding the proportion of subjects that are expected to respond.

To find an effect size for PFS based on the desired effect in overall survival we use the fact that we have already linked the change in tumour size measure to overall survival. The link between PFS and cts will be established using a proportional hazards model, using PFS rather than overall survival as dependent variable. The model will use the same covariates as previously but parameters will be denoted by γ to reflect that PFS is used instead of overall survival. Using the same arguments as in Section 2 we obtain

$$h_E(t) = h_0(t)e^{\boldsymbol{\gamma}'\mathbf{z}_i + \gamma_{cts} \times (cts + \theta_{cts})} = e^{\gamma_{cts}\theta_{cts}} h_C(t),$$

and hence the worthwhile treatment effect for PFS in terms of the negative of the log-hazard ratio, θ_{PFS}, can be expressed in terms of a worthwhile effect in change in tumour size as $\theta_{PFS} = -\gamma_{cts}\theta_{cts}$.

For the NSCLC dataset this relationship becomes $\theta_{PFS} = 2.3857\theta_{cts} = (2.3857/2.2159)\,\theta_{OS}$. A hazard ratio of 0.8 in overall survival therefore corresponds to a change in tumour size of 0.10 as

argued previously, and the corresponding hazard ratio for PFS is found as 0.79. This shows that the hazard ratios for progression-free survival and overall survival are empirically quite similar and suggests that the effects in terms of PFS and OS should be similar.

To illustrate the relationships of the different effect size measures used in the different trial designs some examples are given below. The change in median overall survival time from 8 months (the median OS in the dataset) will be taken as the baseline effect size. The hazard ratio is obtained by creating a "pseudo-treatment group" which is an exact copy of the original data where OS is shifted by the improvement in median OS. So for example one month is added to the survival times to represent an experimental treatment that increases the median survival time from 8 to 9 months. The original dataset and the artificial dataset are then analysed jointly to derive the hazard ratio of 0.8406. This approach also allows the estimation of the change in proportion of patients dying prior to the second assessment of tumour size in the same manner. Table 1 provides the effect sizes for different changes in median OS. These effect sizes will subsequently be used for sample size comparisons in Section 3.3.

~~~ Table 1 ~~~

3.3. Sample size requirements

We now turn to comparing the sample sizes for the different endpoints using the equivalent effect sizes derived in the previous section. For each of the designs the hypotheses H_0: $\theta = \theta_0$ and H_A: $\theta > \theta_0$, where θ is the effect parameter for the design, are to be tested. The calculations will be made to satisfy the type-I and type-II error constraints, $P(reject\ H_0|\theta = 0) \leq \alpha$ and
~~~

$P(reject\ H_0|\theta = \theta_1) \geq 1 - \beta$, where θ_1 is an effect we wish to detect if present. Throughout we will denote the 100 η % percentile of the standard normal distribution by z_η. For comparative trials we will allow for R:1 randomization between experimental and control treatment and denote the total sample size of the trial by n.

The necessary sample size for a trial based on change in tumour size is derived as discussed in Section 2. The sample size and critical value for a single armed trial based on tumour response are found by solving

$$P(S \geq u|p = p_0) = \sum_{k=u}^{n} \binom{n}{k} {p_0}^k (1 - p_0)^{n-k} \leq \alpha$$

and

$$P(S \geq u|p = p_1) = \sum_{k=u}^{n} \binom{n}{k} {p_1}^k (1 - p_1)^{n-k} \geq 1 - \beta.$$

Since these inequalities will be satisfied for a number of different combinations of *n* and *u*, the smallest sample size satisfying the above inequalities will be reported.

For a comparative trial using tumour response, the approximation to a normal distribution is used [14] so that the critical value and sample size are found as

$$n = \left\{ \frac{z_{1-\alpha}\sqrt{\frac{\bar{p}(1-\bar{p})(R+1)^2}{R}} + z_{1-\beta}\sqrt{\frac{p_E(1-p_E)(R+1)}{Rn} + p_C(1-p_C)(R+1)}}{\delta} \right\}^2$$

and

$$u = z_{1-\alpha}\sqrt{\frac{\bar{p}(1-\bar{p})(R+1)^2}{Rn}}.$$

For the design of a trial based on progression-free survival, further information such as the follow up time of patients is necessary for finding sample sizes. To avoid additional assumptions we will base the design used here on the number of events required. The sample size will be the same as the required number of events if all subjects do progress during the study duration and will otherwise be larger. Under the proportional hazards model and when θ_{PFS} is small and the information is large, the required number of events, *e*, is approximately

$$e \approx \frac{(R+1)^2}{R}\left(\frac{z_{1-\alpha}+z_{1-\beta}}{\theta_{PFS}}\right)^2$$

as shown in [15].

Figure 2 compares sample sizes for the different design proposals over a range of improvements in median survival time for equal allocation between experimental treatment and control, a type I error of 0.1 (one-sided) and a power of 80%. Tables 2-3 provide the sample sizes for the four different design proposals using the effect sizes given in Table 1 above for different type I and type II errors.

~~~ Figure 2 ~~~

~~~ Table 2 and 3 ~~~

The results clearly show that a trial based on change in tumour size requires the smallest sample size among the alternative approaches for all situations considered. Moreover, the sample size required for a comparative trial based on change in tumour size is very similar to the requirements for a single-armed trial based on tumour response even though the latter depends on a historical response rate for the control group which may be inaccurate for the patient

population under study. Since the required sample sizes for a trial based on change in tumour size can be relatively small, the intended type I error rate and power are potentially inaccurate as they are based on asymptotic results. Tables 2 and 3 provide reassurance based on 100,000 replicate simulation runs. Throughout the tables, the empirical type I error and power are close to their nominal levels, indicating the approximations used are satisfactory.

3.4. Sensitivity analysis

In order to obtain more insight into the relationship between the sample size requirements and to investigate the robustness of the conclusions drawn, a sensitivity analysis has been undertaken. A key ingredient to our comparison in Section 3.3 is that the relationship between the different effect measures has been captured accurately. Consequently we now consider alterations to these relationships and investigate the implications for the required sample sizes for the different endpoints. The results shown are based on a hazard ratio for overall survival of 0.8406, a one-sided type I error of 0.1 and 80% power using a 1:1 allocation. Other choices of randomization, type I error and power have also been studied, but details are omitted as they just confirm the patterns shown here.

The first alteration looks at the effect of changing the relationship between overall survival and change in tumour size by modifying β_{cts} in Equation (1). Weakening this relationship (moving the negative value of β_{cts} closer to 0) means that it is necessary to find a larger effect on change in tumour size to anticipate a given effect on overall survival and vice versa. As the effect sizes for PFS and tumour response depend directly on the effect on change in tumour size, the sample sizes and the number of events required for the other endpoints will change accordingly (Figure 3). It is clear from the graph, however, that the order of required sample sizes of the endpoints is

maintained although the relative differences change showing that, the required sample sizes for a trial based on change in tumour size are smaller than for the alternative designs considered. Similarly the required sample size remains comparable to that for a single-arm trial.

~~~ Figure 3 ~~~

The second modification looks at the effect of changing the relationship between overall survival and progression free survival by modifying γ_{cts}. Weakening this link means that information on PFS does not contain the same information about overall survival as change in tumour size and leads therefore to a smaller effect size to be looked for. Since this modification has no bearing on the effect sizes used for change in tumour size and response, Figure 4 shows unchanged sample sizes for all endpoints other than PFS for which the number of events required increase. Even changing γ_{cts} from the estimated value of 2.3857 to 3.5 does result in a smaller required event number than the sample size required for change in tumour size showing that a trial based on cts is still more powerful than a trial based on PFS.

~~~ Figure 4 ~~~

The final modification changed the effect sought in terms of tumour response while keeping all other effect sizes constant. The effect sizes considered are chosen to ensure that only meaningful situations are considered (i.e. $p_1 > p_0$). As in the previous situation this adjustment only has implications on the sample size required for trials based on tumour response as seen in Figure 5. It can be seen that more than doubling the effect size is required to find that the sample size for a randomized trial based on tumour response is smaller than for a trial based on change in tumour

size, while only small reductions in treatment effect lead to markedly increased sample sizes for trials based on tumour response. The conclusions about the merits of using change in tumour size therefore appear to be robust.

~~~ Figure 5 ~~~

4. Discussion

In this manuscript a flexible approach for designing and analysing comparative cancer trials with change in tumour size as the primary endpoint, taking account of all information available, has been described. The model expressing a normally distributed measure of change in tumour size in terms of treatment and other prognostics factors has been related to a model expressing the hazard of death to the same set of covariates. In particular, the parameter representing the advantage of the experimental treatment over control in terms of change in tumour size has been related to that expressing advantage in terms of overall survival. This has been achieved through the analysis of a pre-existing dataset concerning patients treated in a similar way to the trial of interest. One of the key assumptions made is that the effect of treatment on overall survival is captured by the effect of treatment on change in tumour size. This implies that, just as in the situation when PFS or response rate is used, we are looking at situations in which change in tumour size is highly predictive of overall survival.

The relationships identified are used to compare the sample sizes required for phase II studies based on different patient outcomes, under magnitudes of treatment effect that are as far as possible consistent with one another. The approach described is particularly suitable in the design of phase II trials, to ensure that the sample size used is sufficient to detect a change in tumour size effect that is consistent with the magnitude of hazard ratio that would be of interest in a subsequent phase III trial. The use of historical data to form a link between treatment effects
~~~

on different endpoints is similar in principle to the approach taken elsewhere for the design of stroke trials [16].

Comparing the sample sizes for the design based on change in tumour size to those for alternative phase II designs using equivalent effect sizes shows that the former approach has the smallest sample size. Furthermore it gives sample sizes that are only slightly larger than those for a single armed trial based on tumour response. A sensitivity analysis shows that these conclusions are highly robust to assumptions made in the derivation of the equivalent effect sizes.

Since tumour measurements are already routinely made in order to ascertain response and progression-free survival, no change in current practice and no additional data are necessary in order to use change in tumour size as the primary endpoint in exploratory studies. If tumour size is assessed regularly, a repeated measure approach could be used, and exploration of this option forms the basis of future work. To implement the design approach described here to other cancers, it will be necessary to confirm that change in tumour size does indeed translate to an increase in overall survival for that particular cancer, and to estimate the magnitude of the relationship.

Acknowledgments

This report is independent research arising in part from Dr. Jaki's Career Development Fellowship (NIHR-CDF-2010-03-32) supported by the National Institute for Health Research. The views expressed in this publication are those of the authors and not necessarily those of the

NHS, the National Institute for Health Research or the Department of Health.

References

1. Dhani H, Tu D, Sargent DJ, Seymour L, Moore MJ. Alternate endpoints for screening Phase II studies. *Clinical Cancer Research* 2009; **15**:1873-1882.
2. Panageas KS, Ben-Porat L, Dickler MN, Chapman PB, Schrag D. When You Look Matters: The Effect of Assessment Schedule on Progression-Free Survival. *Journal of the National Cancer Institute* 2007; **99**:428-432.
3. Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, Verweij J, Van Glabbeke M, Van Oosterom AT, Christian MC, Gwyther SG. New guidelines to evaluate the response to treatment in solid tumors. *Journal of the National Cancer Institute* 2000; **92**:205-216.
4. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, Dancey J, Arbuck S, Gwyther SG, Mooney M, Rubinstein L, Shankar L, Dodd L, Kaplan R, Lacombe D, Verweij J. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer* 2009; **45**:228-247.
5. Donegan WL. Tumor-Related Prognostic Factors for Breast Cancer. *A Cancer Journal for Clinicians* 1997; **47**:28-51.
6. Port JL, Kent MS, Korst RJ, Libby D, Pasmantier M, Altorki NK. Tumor Size Predicts Survival Within Stage IA Non-Small Cell Lung Cancer. *Chest* 2003; **124**:1828-1833.
7. Wang Y, Sung C, Dartois C, Ramchandani R, Booth BP, Rock R, Gobburu J. Elucidation of Relationship Between Tumor Size and Survival in Non-Small-Cell Lung Cancer Patients Can Aid Early Decision Making in Clinical Drug Development. *Clinical*

Pharmacology & Therapeutics 2009; **86**:167-174.

8. Lavin PT. An alternative model for the evaluation of antitumor activity. *Cancer clinical trials* 1981; **4**:451-457.
9. Karrison TG, Maitland ML, Stadler WM, Ratain MJ. 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. *Journal of the National Cancer Institute* 2007; **99**:1455-1461.
10. Wason JMS, Mander AP, Eisen TG. Reducing sample size in two-stage phase II cancer trials by using continuous tumour shrinkage end-points. *European Journal of Cancer* 2011; **47**:983-989.
11. Collett, D. *Modelling Survival Data in Medical Research*. Chapman and Hall/CRC: Boca Raton. 2nd Edition. 2003.
12. Whitehead J. *The design and analysis of sequential clinical trials*. John Wiley & Sohns Ltd: Chichester. 2nd Edition. 1997.
13. Oken MM, Creech RH, Tormey DC, Horton J, Davis TE, McFadden ET, Carbone PP. Toxicity and Response Criteria of the Eastern Cooperative Oncology Group. *American Journal of Clinical Oncology* 1982; 5:649-655.
14. Machin D, Campbell MJ, Fayers PM, Pinol APY. *Sample Size Tables for Clinical Studies.* Oxford, U.K.: Blackwell Science Ltd. 2nd Edition. 1997.
15. Schoenfeld DA. Sample-Size Formula for the Proportional-Hazards Regression Model. *Biometrics* 1983; **39**:499-503.

16. Whitehead J, Bolland K, Valdés-Márquez E, Lihic A, Ali M, Lees K. Using historical lesion volume data in the design of a new phase II clinical trial in stroke. *Stroke* 2009; **40**:1347-1352.

Figure 1: Schematic overview of comparison.

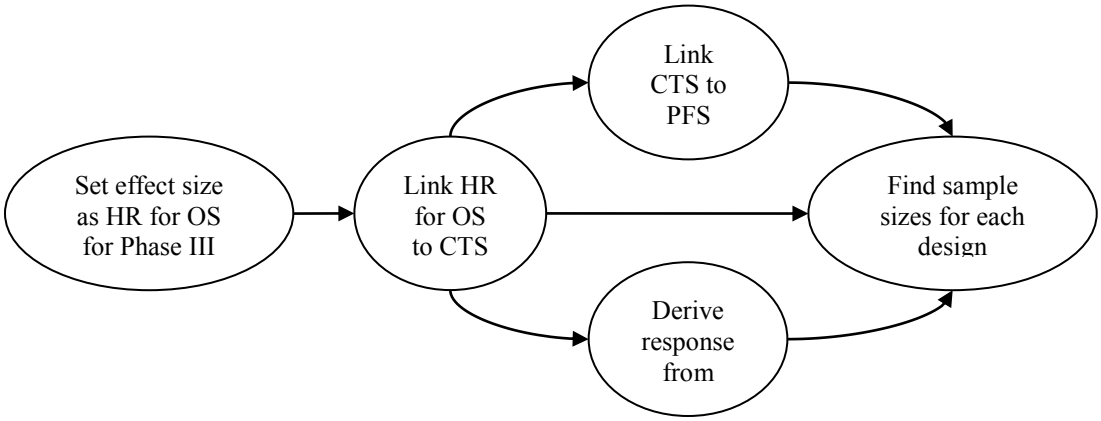

Figure 2: Required sample sizes and number of events for different design proposals for various improvements of overall survival time. Equal randomization between experimental and control, one-sided type I error is 0.1 and power is 0.8.

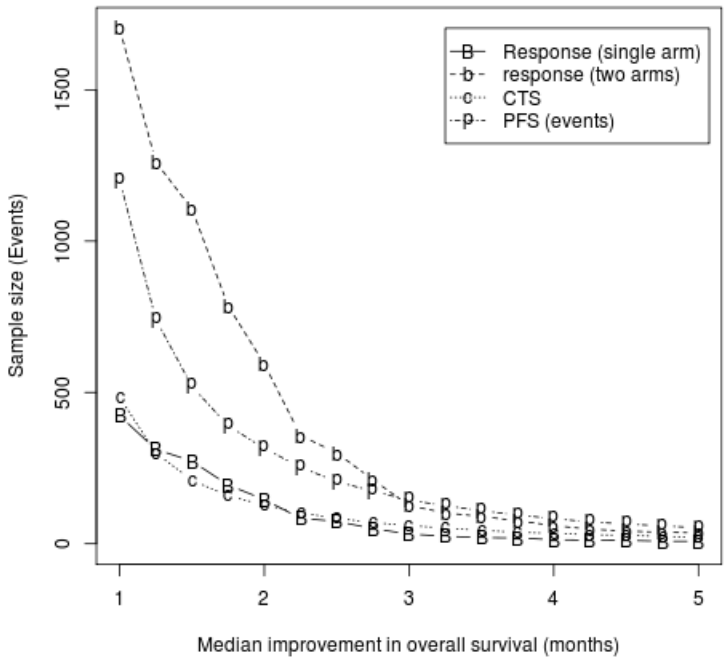

Figure 3: Sample size and number of events required when altering β_{cts}.

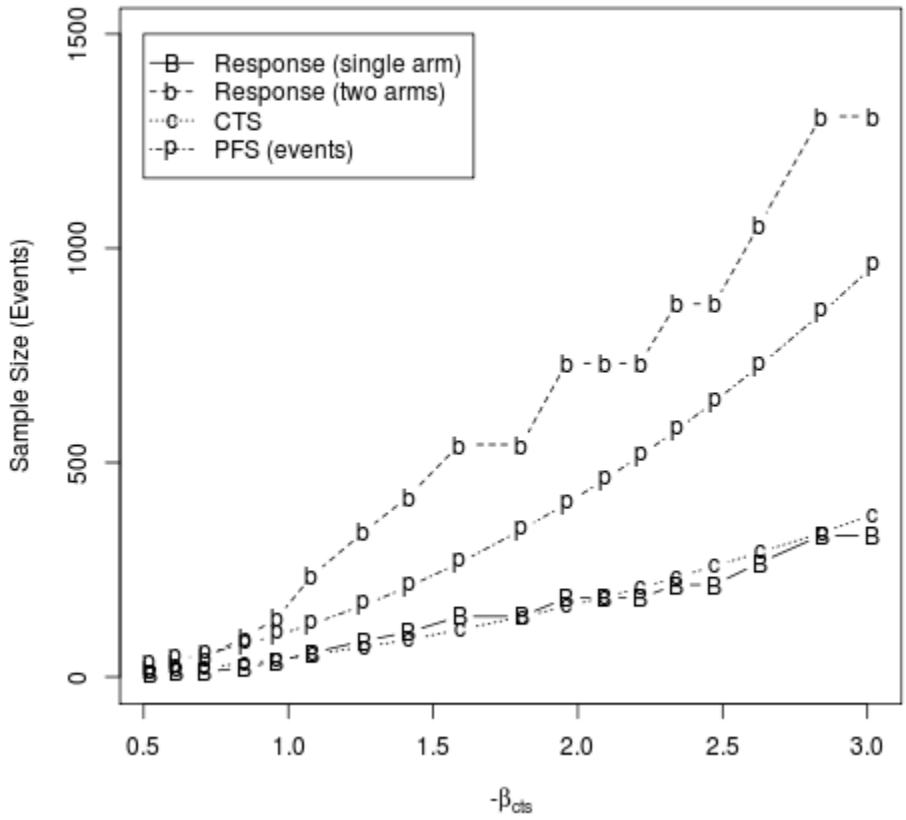

Figure 4: Sample size and number of events required when altering γ_{cts}.

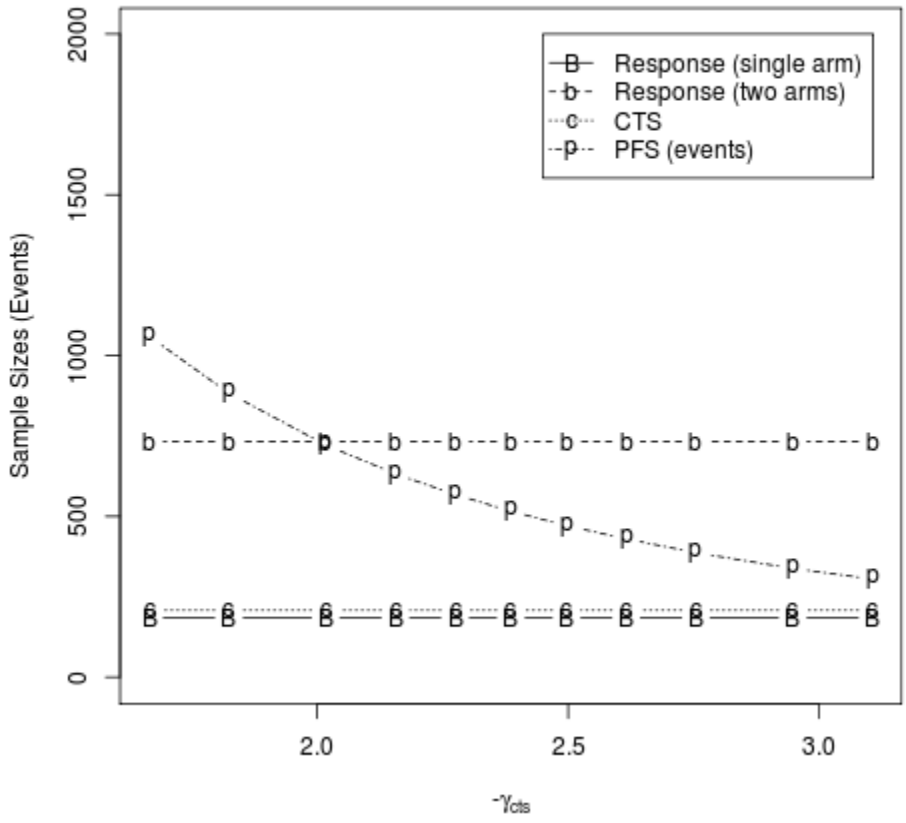

Figure 5: Sample size and number of events required when altering the effect for response. Vertical line corresponds to estimate based on provided data.

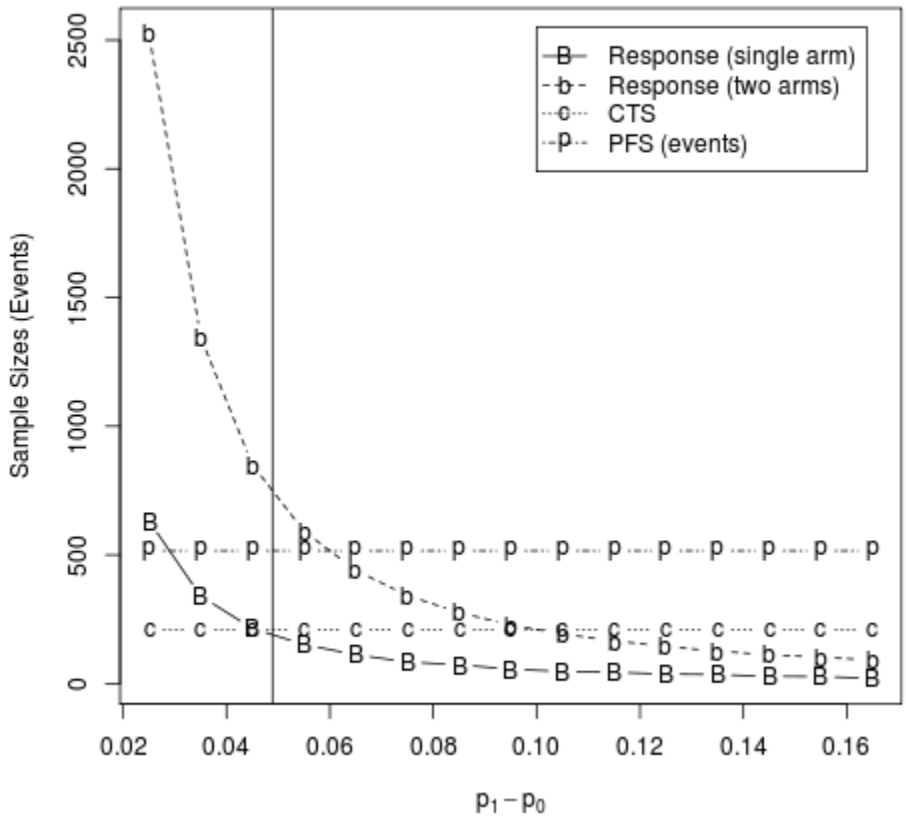

Table 1: Relationships between different effect size measures.

	Target effect size			
Median OS	Hazard ratio (OS)	Increase in proportion of responses	Change in tumour size	Hazard ratio (PFS)
8 to 9 months	0.8406	0.0489	0.0784	0.8295
8 to 10 months	0.7121	0.0889	0.1532	0.6938
8 to 11 months	0.6054	0.2267	0.2265	0.5826

Table 2: Total sample sizes for different designs and effect sizes with $\alpha = 0.1$ and $1 - \beta = 0.8$ using a 1:1 randomisation.

θ	Response		CTS	PFS*
	Single arm	comparative		
0.8406	184 (0.097;0.808)	732	208 [0.1075;0.7902]	516
0.7121	64 (0.088;0.801)	256	56 [0.1099;0.7976]	135
0.6054	13 (0.091;0.821)	54	26 [0.1084;0.8010]	62

(.,.)… achieved type I error and power

[.,.]… empirical type I error and power based on 100,000 simulations

θ… desired hazard ratio of overall survival times

* … number of events

Table 3: Total sample sizes for different designs and effect sizes with $\alpha=0.025$ and $1-\beta=0.9$ using a 1:1

randomisation.

θ	Response		CTS	PFS*
	Single arm	comparative		
0.8406	423 (0.023;0.901)	1704	484 [0.0293;0.8885]	1202
0.7121	146 (0.022;0.901)	594	128 [0.0295;0.8925]	315
0.6054	32 (0.016;0.910)	126	60 [0.0293;0.8998]	144

(.,.)… achieved type I error and power

[.,.]… empirical type I error and power based on 100,000 simulations

Θ… desired hazard ratio of overall survival times

* … number of events