A novel approach to sample size determination of clinical trials for rare diseases assuming symmetry

Emma Wang (e.y.wang@qmul.ac.uk)
Wolfson Institute of Preventive Medicine https://orcid.org/0000-0002-5200-1043
Bernard North
Exploristics
Peter Sasieni
King's College London

Research article

Keywords: Rare diseases, non-inferiority, simulations, survival, Poisson

Posted Date: November 5th, 2019

DOI: https://doi.org/10.21203/rs.2.16784/v1

A novel approach to sample size determination of clinical trials for rare diseases assuming symmetry

Emma Yu Wang[a*], Bernard North[b] and Peter Sasieni[c]

a. Centre for Cancer Prevention, Queen Mary University London. Wolfson Institute, Charterhouse Square, London EC1M 6BQ

b. Exploristics Ltd, Floor 4, 24 Linenhall Street, Belfast BT2 8BG

c. Innovation Hub, Cancer Centre, King's College London, Great Maze Pond, London SE1 9RT

*Author for correspondence: Emma Yu Wang, email: e.y.wang@qmul.ac.uk

Abstract

Background

Rare and uncommon diseases are difficult to study in clinical trials due to limited recruitment. If the incidence of the disease is very low, international collaboration can only solve the problem to a certain extent. A consequence is a disproportionately high number of deaths from rare diseases, due to unclear knowledge of the best way to treat patients suffering from these diseases. Hypothesis testing using the conventional Type I error in conjunction with the number of patients who can realistically be enrolled for a rare disease, would cause the trial to be severely underpowered.

Methods

Our proposed method recognises these pragmatic limitations and suggests a new testing procedure, wherein conclusion of efficacy of one arm is grounded in robust evidence of non-inferiority in the endpoint of interest, and reasonable evidence of superiority, over the other arm.

Results

Simulations were conducted to illustrate the gains in statistical power compared with conventional hypothesis testing in several statistical settings as well as the example of clinical trials for Merkel cell carcinoma, a rare skin tumour.

Conclusions

Our proposed analysis method enables conducting clinical trials for rare diseases, potentially leading to better standard of care for patients suffering from rare diseases.

Key words: Rare diseases, non-inferiority, simulations, survival, Poisson

1. Background

A cancer is defined as rare by Surveillance of Rare Cancers (RARECARE) if its incidence is less than 6 cases per 100,000 per year[1]. Gatta *et al*. found that collectively, rare cancers comprised of 22% cancer cases in Europe.

Patients with rare cancers have worse survival than patients with non-rare cancers: the 5-year-relative survival of patients diagnosed with a rare cancer in 1995-1999 was 47%, compared to 65% in patients diagnosed with common cancers[2]. It is suggested that this difference is due to uncommon cancers being understudied[3].

Clinical trials in the field of rare tumours are limited, and those that do exist are often imperfectly conducted, single-arm trials from which it is difficult to make satisfactory inferences. Recruitment is lower, and attrition is higher in trials of rare diseases, compared to common ones[4].

Often, patients with rare cancers are assigned similar treatment to patients of more common tumours at a proximous site. However, the optimal treatment for most rare diseases is still unclear.

There is clearly a need for more scientific exploration to discern which therapy, out of those currently used for treating patients with a particular rare cancer, is optimal. To provide a means of resolving the problem of clinical trials with low recruitment, we propose a novel method of comparing two treatments that will allow progress despite the limited numbers.

We will describe the motivation for the paper and briefly summarise some pre-existing solutions in **Section 2**. In **Section 3** we will describe the proposed testing procedure. Calculations illustrating the reduction in sample size by using our method in the case of a specific rare cancer are presented in **Section 4**, and results from simulating clinical trials for this rare disease are presented in **Section 5**. In **Section 6**, we discuss the quantitative results before concluding the paper.

2 The Rare Disease Problem

2.1 Low power of clinical trials for rare diseases using conventional hypothesis testing

Randomised Phase III trials are the optimal method for establishing best patient care. Using levels of Type I (α) and Type II (β) error that are common to Phase III trials (two-sided $\alpha = 5\%$ and $\beta = 20\%$[5]), it would not be possible to conduct a randomised controlled trial for a

rare disease as there would be insufficient numbers of patients. If the trial were to go ahead using the conventional Type I error, the power would be very low.

If the comparison of efficacy is between two treatments that are already used in patients with a given disease, then the clinical trial would require prohibitively high numbers of patients in order to conclude which is better, given that both have already demonstrated they are superior to placebo by a reasonable margin.

Scientists might consider it unethical to conduct small clinical trials[6], as one should only expose patients to a new regimen if the trial is sufficiently powered to answer the scientific question at hand. However, the focus of this paper is comparing therapies *already* used on patients, and thus, they are not untested treatments. This argument is given further support considering that even information from a modestly-powered clinical trial is better than no information at all, and it can be used to contribute knowledge in the field via a meta-analysis.

2.2 Alternative methods proposed for dealing with rare diseases

Some alternative methods of analysis to solve the problem of low occurrence of an endpoint of interest in rare diseases have been explored in the past. Suggested solutions include: conducting a one-arm trial using a historical control; use of a surrogate endpoint; conducting 1-sided hypothesis testing; accepting lower power; and incorporating Bayesian inference into the trial[7]. The benefits and shortcomings of each method are discussed by Billingham *et al.*[7]

As this paper concerns the comparison of two treatments already in routine use, a solution might be to conduct a non-inferiority trial comparing one treatment against the other. In non-inferiority trials, a novel treatment is examined against an active treatment to show it is not worse by a given margin, Δ. Non-inferiority tests are carried out in instances where the control treatment has already been proved to be efficacious[8]. In most cases there would be some other advantage, such as cost, side-effects or convenience, of the new treatment.

Whilst non-inferiority trials can substantially reduce the necessary sample size, there is a concern of bio-creep associated with them[9]. By accepting new treatments that are slightly worse (by $\frac{\Delta}{2}$ say), over several successive clinical trials in the disease of interest, the bar is set so low that the experimental treatment that is eventually accepted using non-inferiority rules ends up being ineffective or even harmful for the patient, compared to no treatment.

Furthermore, unless a treatment is substantially less toxic or cheaper, we are not setting out merely to conclude one arm is not worse than the other arm. Rather, we wish to obtain a reasonable idea of which arm leads to better outcomes, or, if we fail to conclude either arm is better than the other, for this to be grounded in reasonable evidence. That is to say, conclusion of efficacy using our proposed testing procedure means we are reasonably happy the proposed arm is better, or at the very least, equal to the arm with which it is being compared[10], and more likely than not, given the evidence from the trial, it is better.

In the next section we will propose two changes to conventional hypothesis testing in order to achieve required sample sizes for trials for rare diseases that are pragmatic yet can achieve the aims outlined above.

3. Methods

3.1 New solution: theoretical explanation of proposed testing procedure

In a conventional superiority trial, the null hypothesis is that the new treatment and the control have the same effect on the primary outcome[9]. In order to accept the new treatment, the clinical trial must demonstrate that it is highly likely to be better than the control. Consider the null and alternative hypotheses:

$H_o: \mu_t = \mu_c$ The two treatments are the same with respect to the mean response

$H_1: \mu_t \neq \mu_c$ The two treatments are different with respect to the mean response

where μ_c, μ_t represent the mean outcomes for the control and treatment groups, respectively. If the outcome is negative, for example death, then we seek $\mu_t < \mu_c$ and efficacy is declared only when the two-sided p-value is less than the pre-set Type I error level.

Our aim is to explore how to compare two treatment regimens already in use. Thus, in comparison to conventional hypothesis testing, there is no "control" as such, as neither treatment is the accepted standard of care. Hence, we are assuming symmetry of efficacy regarding the treatments. The two different treatments will be subscripted in the paper as A and B.

In our paper, we therefore propose a testing procedure that produces more achievable sample sizes. The testing procedure involves two parts: non-inferiority at the two-sided Type I error level of 5%, and superiority at the two-sided Type I error level of 50%.

Since there is no substantial loss of choosing A over B, or vice versa, when they are equally efficacious, there does not exist an ethical imperative to control Type I error under the null hypothesis of equality, $\mu_A = \mu_B$, so stringently. We propose using a much larger value for the two-sided Type I error, 50% rather than 5%.

We also want the trial to be sufficiently well-powered to be confident that the arm that is being declared more efficacious is non-inferior to the other arm. We allow a threshold of indifference, Δ, and determine the required sample size using conventional levels of Type I error and power when this threshold is included. This modification requires an alteration of the conventional null hypothesis, to assume the two treatments are symmetric. Note that this implies that we are not considering non-inferiority when the two treatments are equally effective, but when one is superior to the other.

The symmetry element of the trial means we are interested in showing either that A is non-inferior to B (when in fact A is slightly better) or that B is non-inferior to A (when in fact B is slightly better). Consider two non-inferiority tests:

$$H_0: \mu_A - \mu_B \geq \Delta \qquad [A \; is \; worse \; than \; B]$$
$$H_1: \mu_A - \mu_B < \Delta \qquad [A \; is \; not \; worse \; than \; B]$$

and

$$H_0^*: \mu_A - \mu_B \leq -\Delta \qquad [B \; is \; worse \; than \; A]$$
$$H_1^*: \mu_A - \mu_B > -\Delta \qquad [B \; is \; not \; worse \; than \; A]$$

where Δ is the maximum level of difference one would be willing to accept before concluding inferiority. We set the Type I error so that the probability of rejecting H_0 when $\mu_B - \mu_A = \Delta$ is $\frac{\alpha}{2}$ and by symmetry, the probability of rejecting H_0^* when $\mu_A - \mu_B = \Delta$ is also $\frac{\alpha}{2}$.

3.2 Derivation of the sample size required under our testing procedure for a normally distributed endpoint

We will now demonstrate how these adaptations translate to changes in required sample sizes.

The generic equation for number of patients in each arm, for a two-arm trial, of a standard superiority test for normally distributed endpoints, when the null hypothesis assumes equal means, is:

$$n = \frac{\left(Z_{1-\alpha/2}\left(\sigma_{A_0}^2 + \sigma_{B_0}^2\right)^{1/2} + Z_{1-\beta}\left(\sigma_{A_1}^2 + \sigma_{B_1}^2\right)^{1/2}\right)^2}{\left(\mu_{A_1} - \mu_{A_0} - \mu_{B_1} + \mu_{B_0}\right)^2} \qquad (1)$$

where $Z\gamma$ is the standard normal deviate such that if $Z \sim N\ (0,1)$, $P(Z \leq Z_\gamma) = \gamma$, α and β are the Type I and II errors for the trial, $\sigma_{r_i}^2$ is the variance in arm r under hypothesis H_i and μ_{r_i} is the mean in arm r under hypothesis H_i.

Under the null hypothesis of equal means (which we use for superiority testing), $\mu_A = \mu_B = \bar{\mu}$. If the variance depends only on the means, then the two treatment arms have a common variance, $\sigma_{A_0}^2 = \sigma_{B_0}^2 = {\sigma_0}^2$ under the null.

Let δ represent the difference in endpoint of interest between A and B under the alternative hypothesis that the study is powered to detect ($\delta = \mu_{A_1} - \mu_{B_1}$). Substituting these values, the denominator of (1) becomes δ^2.

Increasing the one-sided Type I error from ${\alpha}/{2} = 0.025$ to ${\alpha^*}/{2} = 0.25$ when conducting the superiority test reduces the required sample size substantially. $Z_{1-\alpha/2}$, conventionally taken to be 1.96, is replaced with 0.6745 to give the required number of patients per arm as:

$$n = \frac{(\mathbf{0.6745}(2{\sigma_0}^2)^{1/2} + Z_{1-\beta}\left(\sigma_{A_1}^2 + \sigma_{B_1}^2\right)^{1/2})^2}{\delta^2} \qquad (2)$$

Equation (2) is the minimum sample size for superiority at $\alpha^* = 0.50$ with power $1 - \beta$.

For the non-inferiority test (Section 3.1), $\mu_{A_0} - \mu_{B_0} = \Delta$, (i.e., under the null hypothesis, the mean in arm A is greater than that in arm B by Δ) and $\mu_{B_1} - \mu_{A_1} = \Delta$ (i.e., under the alternative hypothesis, the mean in arm A is smaller than that in arm B by Δ). Thus, the value in the denominator of equation (1) is $(2\Delta)^2 = 4\Delta^2$. Note that the sample size for requirements for H_0 and H_1 is the same as for H_0* and H_1*.

The sample size for our second criterion is thus:

$$n = \frac{\left(1.96(2{\sigma_0}^2)^{1/2} + Z_{1-\beta}\left(\sigma_{A_1}^2 + \sigma_{B_1}^2\right)^{1/2}\right)^2}{\mathbf{4}\Delta^2} \qquad (3)$$

For moderate treatment effect sizes, our recommendation is to take $\Delta = \delta$ in equation (3). We also only consider using the same power for the non-inferiority and superiority tests.

Equation (3) yields a sample size that would generate reasonable evidence that A is not clinically worse than B. Both requirements (2) and (3) must be met to satisfy our definition of A being 'better' than B.

Thus, the bigger n from the two equations is taken. The adaptations to the conventional formula are highlighted in bold:

$$n = \max\left(\frac{\left(\boldsymbol{Z_{0.75}}(2{\sigma_0}^2)^{1/2} + Z_{1-\beta}\left(\sigma_{A_1}^2 + \sigma_{B_1}^2\right)^{1/2}\right)^2}{\delta^2}, \frac{\left(1.96(2{\sigma_0}^2)^{1/2} + Z_{1-\beta}\left(\sigma_{A_1}^2 + \sigma_{B_1}^2\right)^{1/2}\right)^2}{\boldsymbol{4}\delta^2}\right) \quad (4)$$

In order to compare the two criteria, we apply the simplifying assumption $\sigma^2 = {\sigma_A}^2 = {\sigma_B}^2$, where σ^2 is the variance of the mean amount of deaths across the entire trial and consider only 80% power. The required sample size is:

$$n = \max\left(\frac{4\sigma^2}{\delta^2}(\boldsymbol{0.6745 + 0.8416})^{\boldsymbol{2}}, \frac{4\sigma^2}{\delta^2}\frac{(\boldsymbol{1.96 + 0.8416})^{\boldsymbol{2}}}{\boldsymbol{4}}\right) \quad (5)$$

Comparing the two values emboldened, the superiority criterion (2) leads to the greater sample size, so for 80% power in this example, this rule dominates. It is only when power below 73% is sought, that the non-inferiority criteria (3) dominates when $\Delta = \delta$.

3.3 Adapting the normal formula to non-normally distributed endpoints

We will now apply discuss how to apply the formulae in Sections 3.1 to 3.3 to the asymptotic distribution of test statistics derived from endpoints with different underlying statistical distributions.

3.3.1 Binomially-distributed endpoints when the focus is on the difference in proportions

If the endpoint is binary we can use the normal approximation to the binomial to adapt the formulae. Here, $\mu_i = p_i$, where p_i is the probability of the outcome event in treatment group i.

The hypothesis test now pertains to the difference in proportion of occurrence of the endpoint. The treatment effect under exploration changes from $\mu_A - \mu_B$ to $p_A - p_B$, with the hypotheses outlined in 3.1 altered accordingly.

Under the null hypothesis, the proportion of events in the two arms are equal. We use $\bar{p} = \frac{p_{A_1} + p_{B_1}}{2}$ as the common proportion. Following this, the variances of arm A and arm B are equal: $\sigma^2 = \bar{p}(1-\bar{p})$. Under the alternative hypothesis of differing proportions by group, the variance in arm i is $\sigma_i^2 = p_{i_1}(1-p_{i_1})$.

3.3.2 Binomially-distributed endpoints when the focus is on the log odds ratio

The problem with using the normal distribution to approximate the difference in two binomial proportions is that the difference is constrained to be on the interval [-1,1] (and more so once one of the proportions is set). Approximating the log-odds by the normal distribution bypasses this problem.

Under the binomial log-odds, the statistic under exploration is $\mu_i = \log\left(\frac{p_i}{1-p_i}\right)$ and hypotheses testing explores the log-odds of proportions of event in the two treatment groups. The variance is calculated using the asymptotic variance $\sigma_i^2 = \frac{1}{p_i} + \frac{1}{1-p_i}$ and $\delta = \log\left(\frac{p_{A_1}(1-p_{B_1})}{p_{B_1}(1-p_{A_1})}\right)$.

3.3.3 Poisson-distributed endpoints

In the Poisson distribution, the statistic of interest is $\mu = ln(\lambda_i)$, where λ_i is the rate the event occurs on arm i. Note that $\sigma^2 = Var(\lambda_i) = \frac{1}{\lambda_i}$ and $\delta = \log\left(\frac{\lambda_A}{\lambda_B}\right)$.

3.3.4 Survival endpoints using the log-rank test

In this section, we will assume proportional hazards:

$$H_B(u) = \theta H_A(u), \quad \forall u$$

where $H_A(u)$ is the cumulative hazard at time u in arm A.

The hypotheses now compare survival of the two arms at all times t:

$$\mathrm{H}_0: \frac{H_A(t)}{H_B(t)} \geq \Delta \qquad [A \text{ is worse than } B]$$

$$\mathrm{H}_1: \frac{H_A(t)}{H_B(t)} < \Delta \qquad [A \text{ is not worse than } B]$$

and

$$\mathrm{H}_0{}^*: \frac{H_B(t)}{H_A(t)} \geq \Delta \qquad [B \text{ is worse than } A]$$

$$H_1{}^*: \frac{H_B(t)}{H_A(t)} < \Delta \qquad [B\ is\ not\ worse\ than\ A]$$

Applying our adaptations to the formula for comparison of survival using the log-rank test given by Freedman (1982)[11] for the superiority component, gives:

$$d_s = (Z_{\alpha^*/2} + Z_{1-\beta})^2 \left(\frac{1+\theta}{1-\theta}\right)^2 \quad (6)$$

Where d_s is the number of events required and θ is the postulated hazard ratio, when the superiority component of the proposed testing method entails increasing the value of the Type I error from $^{\alpha}/_{2} = 1.96$ to $^{\alpha^*}/_{2} = 0.6745$.

From this, the total required number of patients is calculated as:

$$N_s = \frac{2 \times d_s}{2 - S_A(\tau) - S_B(\tau)} \qquad (7)$$

where $1 - S_i(\tau)$ is the probability of having an event (being observed to die) in arm i by the end of the study as before.

To satisfy the non-inferiority part of the testing procedure, two changes are implemented to Freedman's formulae. First, $\theta^* = \theta \times \frac{1}{\Delta}$, to account for the altered null hypotheses, where Δ is the maximum harmful treatment effect that would be accepted. In this case where $\Delta = \frac{1}{\theta}$, disproving the null hypothesis would entail two magnitudes of the proposed hazard ratio being observed relative to an assumed level of inferiority. Throughout this paper, $\Delta = 1.25$ was used.

The survival corresponding to the treatment effect when including the non-inferiority criterion is then substituted for $S_A(\tau)$ in (7).

Thus, for the non-inferiority criteria:

$$d_{n-i} = \left(z\alpha_{/2} + z_{1-\beta}\right)^2 \left(\frac{1+\theta^*}{1-\theta^*}\right)^2$$

$$N_{n-i} = \frac{2 \times d_{n-i}}{2 - S_{A_W}(\tau) - S_B(\tau)} \qquad (8)$$

$S_{A_W}(\tau)$ is the survival that would be observed in arm A if the hazard ratio of A:B were Δ (i.e., A was worse than B rather than better). As before, the required sample size is chosen from the larger of the two required:

$$N = \max(N_s, N_{n-i})$$

4 Application of Proposed Testing Procedure to a Real Example of a Rare Cancer: Merkel Cell Carcinoma

4.1 Background: Merkel cell carcinoma

The method of analysis highlighted in the preceding section will now be applied to a trial of patients with Merkel cell carcinoma (MCC) to explore the gains in power, or reduction in sample size from adopting our proposed testing procedure.

MCC is a rare skin tumour that mainly occurs on the head and neck of Caucasians and is caused by exposure to sunlight[12]. It is less common than melanoma, but has much higher rates of metastasis and recurrence, and hence is much more fatal. In the United Kingdom between 1999 and 2008, only 1515 cases of MCC were diagnosed[13], rendering it a rare disease. By virtue of being a rare disease, the best treatment is unknown, and choice of treatment is often at the discretion of doctors (in consultation with the patient).

Currently, the most commonly used treatment for MCC is surgery followed by radiotherapy to kill off any remaining cancer cells[14]. It has been postulated that neoadjuvant radiotherapy (to reduce the size of the tumour to make it easier to remove during surgery) would be as effective, or better, for the survival of MCC patients. Indeed, this is the focus of the Rational MCC clinical trial[14], which is currently underway.

The scientific question of Rational MCC is analogous to comparing two existing treatments because the scientific question is not which *treatment* is better for the patient, but which *order* of the treatment is better. All patients are offered both treatments, but one arm receives radiotherapy before surgery and the other receives radiotherapy after surgery. There are therefore arguably no losses from recommending one treatment sequence above the other if the true differential treatment effect is small.

4.2. Application of proposed testing procedure using MCC

4.2.1.1. Illustration of new testing procedure using the example of MCC in a binomial setting

Death within 2 years of diagnosis can be considered as a binomial event, with patients either dying or being "cured". Thus, by following all patients for two years, we could estimate the proportion cured in each arm.

The Rational MCC brief suggested comparing these two treatments aimed to detect a treatment effect of a hazard ratio of 0.8[14]. An average event rate under the alternative hypothesis of 20% was assumed.

Translating this to absolute values gives p_{A_1}= 0.18026 and $p_{B_1} = 0.22$ (with these values $\frac{-log(1-p_{A_1})}{-log(1-p_{A_2})} = 0.80$). The difference in the proportion of deaths is thus 3.974%.

In the Type I error component of the sample size equation, we are assuming a common variance in the two arms, calculated by using the average proportion, $\bar{p} = \frac{p_A+p_B}{2} = \frac{0.18026+0.22}{2} = 0.20013$.

In order to have 80% power to conduct a 5% two-sided hypothesis test of a difference of 3.974% between arms using equation (7) would yield:

$$2 \times \left(\frac{1.96\sqrt{2 \times 0.20013 \times 0.79987} + 0.8416\sqrt{0.18026(0.81974) + 0.22(0.78)}}{0.03974} \right)^2$$

$$= 3180$$

A total sample size of 3180 patients would be required. Over a ten-year-period in the UK, less than half this amount of MCC diagnoses occurred. Aiming to recruit this magnitude of patients would make the trial unfeasible, even internationally. Using our proposal, this large sample size is substantially reduced.

In order to fulfil the superiority criteria, which dominates here, the upper 50% confidence interval of the difference, δ must fall below 0. This translates to a sample size calculation of:

$$2 \times \left(\frac{0.6745\sqrt{2 \times 0.20013 \times 0.79987} + 0.8416\sqrt{0.18026(0.81974) + 0.22(0.78)}}{0.03974} \right)^2$$

$$= 932$$

It is noted that the sample size requirements when assuming symmetry on the log-odds scale rather than in the difference in proportions are almost identical to those derived here.

4.2.1.2. Simulations of differences in proportions

The accuracy of the sample size equations derived for the various statistical distributions were assessed using simulations in R version 3.0.0 (https://cran.r-project.org/src/base/R-3/).

Simulations of numbers of events of deaths were made for arms A and B, where the average proportion of death in each arm were p_A and p_B. Hypothesis testing was then conducted on the difference in proportions, $p_B - p_A$. We present the results for the binomial setting below.

[insert Table 1]

Under the null scenario of equal p_i in both arms under the binomial distribution, the total Type I error was indeed 50% (i.e., the probability of concluding that either A is better or that B is better) when the sample size was 932 or greater.

As shown by the values in bold, 80% power to satisfy the non-inferiority criterion is reached with 800 patients when $\delta = 0.03974$. However, as stipulated, both the superiority and the non-inferiority criterions must be fulfilled for the results of a trial to count as 'efficacy' using our rules, and 80% for both these rules are met once the sample size reaches of 932.

Depending on the sample size, different rules dominate. When $n \leq 600$ in the alternative scenario, the non-inferiority criterion dominated. When $n = 670$, there were a small proportion (0.1%) of simulations which fulfilled non-inferiority criteria but not the superiority rule. This was because, when solving equation (6) when $\Delta = \delta$, the point at which superiority dominates non-inferiority was for powers of 72.9% or higher, which is exactly that at $n = 670$.

Intuitively, the smaller the treatment difference observed, the higher the probability of incorrectly concluding that the less effective treatment B is better than A, as shown in the case when the difference is 1%. However, if all clinicians believe the difference in event proportion between the two arms is in the vicinity of $\delta = 1\%$, conducting a clinical trial for such a clinically insignificant difference arguably represents a misuse of resources.

Table 1 shows that the normal approximation provides an excellent estimate of sample sizes required under the binomial distribution for our testing procedure. This was also true for simulations of the Poisson distribution (not presented).

4.2.2 Illustration of new testing procedure using the example of MCC in a survival data setting

Having demonstrated the reduction in sample size from using the new method in binomial setting theoretically and using simulations, we will now model the aforementioned MCC trial by comparing survival times between two treatment arms using the log-rank test.

In our simulated clinical trial, the endpoint of interest was death. The trial was powered based on expected number of deaths in an average follow-up of two years, but not every patient was to be followed up for the full two years.

To reflect a steady accrual over time, a recruitment period of two years was modelled, where patients were equally likely to be recruited at the start of the trial (year 0), 6 months, 1 year, 18 months or 2 years. All patients were then followed up for an additional year, meaning that the distribution of follow-up times was split into fifths between 3 years, 2.5 years, 2 years, 1.5 years and 1 year. This gave an average follow-up for two years. For simplicity, we assumed no drop-outs.

The theoretical sample size for superiority with $\alpha = 0.025$ and $\beta = 0.2$ is 3178. This is very similar to that estimated for the binary endpoint.

When implementing our proposed testing procedure under the assumption that $\Delta = \delta$, the superiority criteria (which translates to the requirement that the upper 50% CI of $\hat{\theta} < 1$) dominates, with a sample size of $n = 930$, so this is the sample size that is used.

Again, there is a notable reduction from the sample sizes required by conventional hypothesis testing. We demonstrate the gains in power from using our method diagrammatically:

[insert Figure 1]

In the case of an observed $\hat{\theta} = 0.6$ for the MCC example (with harm defined as $\Delta = 1.25$), the required sample size when seeking 80% for the non-inferiority component yields a greater sample size, 314, than the superiority criterion of 206, so that takes precedence.

By using the new rule, recruiting 314 patients can achieve 80% power when the observed treatment effect $\hat{\theta} = 0.6$. In conventional hypothesis testing (Figure 1, left-hand plot), the power is much lower under a sample size of 314, 47.9%.

5 Simulations and Results

5.1 Details of clinical trial simulations

Simulations of the survival of MCC patients were carried out in R. The steps of the simulations were:

1) Two separate arms (A and B) were simulated. In each arm, patients had exponentially distributed survival times which were independent and identically distributed. Treatment arm B had an annual rate of 0.12423, whereas treatment arm A had a rate of 0.12423θ. These values correspond to an average event rate at two years across the two arms of 20% when $\theta = 0.8$.
2) Uniform recruitment was assumed in both arms, with the distribution of follow-up outlined in 4.2.2.
3) The two sets of survival and censoring times were combined, with a binary indicator to indicate the treatment groups.
4) The log-rank test was applied to the simulated datasets to conduct survival analysis, with treatment as the sole covariate.
5) Steps (1)-(4) were repeated 10,000 times, in order to ascertain median values of $\hat{\theta}$ and $SE(\ln(\hat{\theta}))$. Power under the testing procedure outlined in Section 3.1, as well as power using conventional hypothesis testing at a 2-tailed α-level of 5%, were explored for each sample size and true hazard ratio combination.

5.2 Results

Table 2 summarises the proportion of the 10,000 simulations in which treatment A is selected using the conventional and new measures of efficacy for different combinations of true hazard ratios and sample sizes. The maximum margin of inferiority (for the hazard ratio) was given as $\Delta = 1.25$ in all cases.

[insert Table 2]

As Table 2 demonstrates, when the treatment effect is modest (HR = 0.8), $n = 932$ would give 80% power using our measure, whereas it would only yield 32.7% using conventional hypothesis testing.

When the treatment A is substantially worse than its competitor (HR = $\frac{5}{3}$, 1.5, $\frac{10}{7}$), conventional hypothesis testing produced none or next-to-no simulations which concluded efficacy of this treatment arm for all sample sizes presented. Under HR = $\frac{10}{7}$, our proposed testing procedure

did yield small probabilities of drawing the wrong conclusion for the smallest sample sizes (<1%), but very few or no erroneous conclusions where made when the hazard ratio was $\frac{5}{3}$.

5.3 Comparison with proposed analysis method of Rational MCC

The proposed method of analysis for Rational MCC was to use Bayesian posterior probabilities to guide decision-making. This method facilitates even smaller sample sizes than from using our proposed testing procedure.

The investigator's calculations were replicated using an uninformative prior for the treatment effect ($ln(HR) \sim N(0, 100^2)$) for the same range of sample sizes they presented in their statistical analysis plan, to explore the posterior probabilities under a range of treatment effects.

[insert Table 3a]

As the simulations show, the Type I error control is poor when making inferences using just the point estimates of the Bayesian posterior probabilities. When the sample size is $n = 100$ and the average treatment effect in the trial is a hazard ratio of 1.25, the posterior probability of the hazard ratio being less than 0.8 is 14.3%. Conventional hypothesis testing and our testing procedure yields probabilities of just 0.6% and 1.9%, respectively, of making such incorrect decisions based on the data observed.

Even when the sample size increases to $n = 300$ and there is a harmful effect of HR = 1.25, 16.5% of the simulated posterior distributions are $\widehat{HR} < 1$, where $\widehat{HR}$ is the median observed posterior HR. Using our testing procedure, the probability of drawing an incorrect conclusion of efficacy is 2.0%. Using just the point estimate of the Bayesian posterior probabilities give a large variability of outcomes, from which it is difficult to make robust inferences.

Therefore, whilst the method proposed by the investigators of Rational MCC does facilitate magnitudes of sample sizes that would be easier to recruit for given the low incidence of MCC, the risk of making incorrect inferences when using the point estimate of Bayesian posterior probabilities is large.

Bayesian analysis produces a posterior distribution of $ln(HR)$. Thus, a more reliable way of making inferences is to consider the <u>overall</u> posterior distribution, and to use a threshold for which the proportion of posterior densities falls below $\widehat{HR} < 1$ to define whether the treatment has satisfied efficacy as a result of the trial. To further concretise the validity of the inference, there existed the requirement that less than a given proportion (for example, 2.5%) of the

posterior distribution must fall in the area of harm, $\widehat{HR} > 1.25$. Figure 2 demonstrates these two criteria. Both the pink area of the diagram must be greater than a given value, γ, and the purple part less than another value, τ, in order to conclude the trial has demonstrated efficacy.

The investigators of Rational MCC stated that they expected the trial to recruit between 150 to 250 patients. Table 3b presents a comparison of the powers (the proportion of simulations which fulfil the outlined criterion) under these two sample sizes for the following testing procedures. The third criterion is the 'Bayesian power' in these scenarios, where frequentist power is applied to the values from Bayesian posterior probability.

Thus, the power for these three were recorded:

1) Conventional hypothesis testing at the two-sided 5% level: $\widehat{HR} < 1$ and p-value < 0.05
2) The testing procedure outlined in this paper, where the upper 50% CI of $\widehat{HR} < 1$ and upper 95% CI of $\widehat{HR} < 1.25$
3) The Bayesian posterior mean $Pr(HR < 1 \mid \mathcal{D}) > \gamma$, where $\gamma = 80\%, 95\%$ and a low proportion of the posterior density falling in the region of harm, $Pr(HR > 1.25 \mid \mathcal{D}) < 0.025$.

[insert Table 3b]

When the more stringent requirement of $\gamma = 0.95$ is used, this controls the probability of the yielding incorrect inferences from the posterior distribution much better than by merely inspecting the point estimate of the posterior probabilities. For example, when $n = 150$ and HR = 1.25, the proportion of simulations where $Pr(HR < 1) > 0.95$ is just 1.2%. This compares favourably to our testing procedure which gives 2.2% probability of incorrectly concluding the treatment is efficacious.

This increased robustness of decision-making with $\gamma = 0.95$ comes at the cost of reduced power when the treatment is truly efficacious. When $n = 250$ and the true HR (or likelihood) = 0.8, the percentage of simulations $Pr(HR < 1) > 0.95$ is just 20.0%. By comparison, our proposed testing procedure produces 34.7% of simulations that conclude efficacy.

A higher "Bayesian power" is obtained using the more lenient requirement of $\gamma = 0.8$, with powers observed when the treatment is efficacious that are higher than all the other testing procedures presented. However, this comes at a cost of the increased chance of concluding efficacy (4.7% under n = 150 and 4.2% under n = 250) when the treatment is harmful.

As the powers yielded using our testing procedure lie between the powers observed under the Bayesian method when the value of $\gamma = 0.8, 0.95$, it follows that a value of γ between 0.8 - 0.95 will result in power very similar to those observed using our testing procedure. We explored the power when γ was calibrated to the value that was the complement of the power observed under our testing procedure, when there was no treatment effect (HR = 1). This was $\gamma = 0.9142$ for n = 150 and $\gamma = 0.8746$ for n = 250.

As can be seen, by calibrating a value of γ, decision-making using Bayesian posterior ln(HR)s performs better than our testing procedure at both restricting the probability of making incorrect inferences when the treatment is harmful and concluding efficacy when there is a beneficial treatment effect.

Thus, a potential extension to our proposed testing procedure is using it in conjunction with Bayesian hypothesis testing using an uninformative prior.

Simulations were conducted under HR = 0.37 for a sample size of 150 and HR of = 0.53 for a sample size of 250. In both instances, these are the treatment effects that would theoretically yield 80% power using the formulae presented in Appendix I. The observed power was in the vicinity of 78%, showing that the sample size formulae fares less well for the smallest sample sizes, something which will be explored next.

5.4 Accuracy of the Sample Size Calculator

In Table 4, we briefly digress from MCC and present results from simulations of various clinical trials of different underlying event rates, conducted with differing powers to assess how accurately the sample size equation presented in the appendix works. In all of the examples presented, $\Delta = 1.25$.

The difference in power from using the new rule and conventional hypothesis testing is presented a percentage.

[insert Table 4]

As Table 4 demonstrates, the gains in power for survival data are considerable compared to from using conventional hypothesis testing, with the increase in power as high as 47.3% in some examples. Encouragingly, the modifications to the well-known sample size formulae for survival data approximates the required sample size well for our test procedure, with powers observed in simulations closely matching the theoretical power using the formulae.

The largest differences between theoretical and observed power occur under the most prominent treatment effects, for the smallest sample sizes. This is because the more extreme treatment effects are more likely to yield data that is non-normal, and disobey asymptotic theory. However, the magnitude of treatment effects that are expected to be observed using our testing procedure are unlikely to be in the vicinity of HR < 0.6. Were the one treatment that much better than the other, it is probable that clinicians would have already noticed and altered policy already.

Depending on the underlying event rate, treatment effect and power sought, different rules dominate. This can be explained by investigating equation (6), and which values yield a larger sample size, depending on the postulated treatment effect.

6 Discussion

6.1 Minimum sample size required for a clinical trial analysed using our proposed testing procedure

We saw that the Bayesian design for the MCC Rational trial results in a very small sample size. From a frequentist perspective, we are worried that at the end of the trial, someone might believe that treatment A was clinically superior to treatment B when in fact it is clinically inferior, thus making a poor decision. We suggest that, power aside, we should always require that the sample size is sufficient so that the probability of such an error is no more than 2.5%.

We restrict the probability of concluding an efficacious effect in arm B ($\mu_A - \mu_B \geq \delta$) when the true effect is the other way around ($\mu_B - \mu_A \geq \delta$) to ≤ 2.5%. This is defined as the probability of making an incorrect decision in favour of B.

The sample size required to ensure the probability of making an incorrect decision is below 2.5% is:

$$\frac{2({\sigma_{A_0}}^2 + {\sigma_{B_0}}^2)}{(2\delta)^2}(Z_{0.975})^2 \qquad (8)$$

A simplifying approximation of $\bar{\sigma} = \sigma_{A_0} = \sigma_{B_0}$ changes the formula to:

$$\frac{4\bar{\sigma}^2}{(2\delta)^2}(Z_{0.975})^2 \qquad (9)$$

Negligible differences arising from averaging variances aside, (9) will always be less than or equal to the second part of (5) when the Type I error is 0.025, which the trial will be powered

to be. This means the minimum sample size is implicitly covered by fulfilling the non-inferiority component of the two criteria. Equation (9) ensures a Type 1 error of 2.5% (one sided) (i.e. the error of selecting B) when the null hypothesis is that B is worse by an amount 2δ.

Under $n = 200$, which is less than the minimum sample size (Table 1) in the binomial setting, the probability of incorrectly concluding B is better than A is 2.7%, illustrating the need for a minimum sample size. The observed probability of drawing an incorrect conclusion that B is better than A under the minimum required sample size of 352 is 2.5%, showing that formula (8) (and its variations in other settings) is accurate at limiting the probability of drawing an incorrect decision.

However, merely using the minimum sample size does not yield high enough power to confidently make inferences from. Thus, if the trial is powered reasonably, the minimum sample size will be subsumed in the power calculation.

As demonstrated in Tables 1 to 3, the minimum sample size required will automatically be fulfilled by recruiting for reasonable levels of power to detect a moderate treatment effect (power > 50%, HR ≈ 0.8). It is nonetheless recommended to state a minimum value in the statistical analysis plan, below which a clinical trial would be unethical to conduct. This is due to the elevated probability of making incorrect conclusions under small sample sizes, and making the wrong conclusion due to small sample sizes is a scenario we strongly wish to avoid.

In Table 2, the results of simulations of survival data, sample sizes below the minimum sample size, n = 200 and n = 250 yield incorrect conclusions of 2.7% and 2.6%, demonstrating that, as in the binomial example, it is important to control in the sample size calculation for the possibility of drawing a false conclusion. Under the minimum sample size for survival-distributed endpoints, $n = 350$, a 2.5% probability of drawing the incorrect conclusions in simulations is observed. This shows the minimum sample size controls this error reasonably well.

If the calculated number of patients for 80% power using our measure are recruited, the probability of making a wrong decision falls to 1.0%, confirming that the probability of making a wrong decision will be subsumed in the overall sample size calculation.

6.2 Evaluation of using our proposed testing procedure

By imposing some constraints on the sample size, the sample size generated is attainable and pragmatic given the rarity of the disease, whilst being large enough so that if one treatment was truly worse, the probability of declaring it is superior to the other arm is restricted to be below 2.5%. The rules to conclude efficacy are considerably stricter than non-inferiority trials, and thus conclusion that one arm is 'better' than the other is more robust.

Intuitively, and as Tables 1 to 4 demonstrate, using our easier to obtain measure of efficacy yields much greater power in all combinations of underlying event rate and hazard ratios. Depending on the occurrence of the endpoint in interest, gains in power ranged from 23.0% to 47.3% in the examples presented. The gains in power were most pronounced in examples where there were subtle treatment effects, which are the kinds of magnitude of effect to realistically expect, or hope for, when comparing treatments currently in use for rare diseases. If the treatment effect was more prominent than this, it is probable that clinicians would have noticed in previous observational studies and altered policy already.

By assuming symmetry between treatments, a difference about this manner of determining efficacy from clinical trial data is that rather than having two outcomes at the end of the trial, there are three. But failing to conclude either A or B are better than each other using our testing procedure need not be regarded as a "failed" clinical trial, rather, a conclusion that there was reasonable evidence that neither treatment was substantially better or worse than the other, and whichever selected could be left up to the doctor or patients' discretion.

Overall, Table 4 illustrates good concordance between the theoretical sample sizes for a given power desired from the equation and the power observed for that given sample size in 10,000 simulations, showing that the adapted formulae for sample size under the log-rank test can be used to approximate the number of events when calculating sample sizes for survival data when the ratio is equal between treatment groups.

MCC occurs when several specific mutations are present[13]. Thus, the analysis method proposed by this paper has potential applications in the field of stratified medicine using treatments targeting pathways affected by specific mutations. Even common cancers have stratum that are rare such as basal-type and papillary breast cancer[15]. The increasing volume of common cancers being sub-classified as rarer ones by molecular markers means that our novel measure can be easily extrapolated in the setting of common cancers by mutation status.

Our method of comparing the efficacy of two active treatments against each other is attractive to the patient for two reasons. Firstly, they are guaranteed to be randomised to a therapy in

existing use, rather than a placebo. By offering two active treatments, recruitment should be higher, which increases the numbers of patients willing to enter the clinical trial. This is crucial in diseases where the absolute number of patients suffering from it are low.

Furthermore, the unambiguous rules for concluding efficacy lead to straightforward, easy to understand, interpretation. If A has been shown to be better than B using our proposed testing procedure, a clinician can prescribe A to the patient, saying '*on the balance of probability, Treatment A is better*'. For patients who may not have the grasp of Type I and Type II error, and simply wish to receive the treatment which will not harm them and might be better than the other, this could be recommendation enough.

The main limitation of this new method is that it cannot be used to test the efficacy of new treatments. Our rule can only be used to compare treatments whose safety profiles have been well-demonstrated in the past, or have been used historically with acceptable safety profiles, thus *justifying* the permissive hypotheses testing. If a completely experimental treatment is being assessed, it would be prudent to use conventional rules for hypothesis testing, so a trial rigorously demonstrates the treatment is better than placebo, before introducing it to the general patient population.

7. Conclusion

By assuming symmetry of the efficacy of treatments, our measure yields pragmatic sample sizes.

For exploration of comparison of two active treatments, we strongly recommend the use of this novel measure proposed in order to reach a fairly robust conclusion about the best method of treating patients with rare tumours. Used in tandem with ever-increasing research in the field of rare cancers, this analysis method will help expedite the process of increasing quality of treatment for patients of rare cancers.

Declarations

Not applicable.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The simulations used for this paper are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

EYW was supported by a studentship from Cancer Research UK.

Authors' contributions

EYW carried out the analysis and wrote the manuscript. PS' ideas and guidance formed the foundation for the manuscript. BN helped with the theory and the coding for the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Tables

Table 1. Results of 10,000 simulations of binomially-distributed occurrences of death and the power to conclude A, B, and individual components of A's power using the proposed testing procedure ($\Delta = \delta = 0.03974$)

n	p_A	p_B	δ	Probability of concluding A better	Probability of concluding B better	% that fulfil A superiority	% that fulfil A non-inferiority
200	0.18026	0.22	0.03974	30.2%	2.7%	52.2%	30.2%
300	0.18026	0.22	0.03974	41.4%	2.3%	58.4%	41.4%
352	0.18026	0.22	0.03974	46.0%	2.5%	59.2%	46.0%
400	0.18026	0.22	0.03974	51.8%	2.5%	62.4%	51.8%
502	0.18026	0.22	0.03974	61.0%	2.5%	67.1%	61.0%
600	0.18026	0.22	0.03974	69.5%	2.4%	71.9%	69.6%
670	0.18026	0.22	0.03974	73.4%	2.4%	73.4%	74.4%
800	0.18026	0.22	0.03974	77.2%	2.2%	77.2%	**80.4%**
872	0.18026	0.22	0.03974	78.6%	1.5%	78.6%	83.7%
932	0.18026	0.22	0.03974	**79.6%**	1.4%	79.6%	86.1%
3180	0.18026	0.22	0.03974	98.3%	0.0%	98.3%	100.0%
200	0.18026	0.2	0.01974	19.8%	5.7%	38.5%	19.8%
400	0.18026	0.2	0.01974	33.4%	7.7%	43.3%	33.4%
600	0.18026	0.2	0.01974	47.5%	8.9%	49.0%	47.5%
800	0.18026	0.2	0.01974	51.7%	8.1%	51.7%	57.7%
1000	0.18026	0.2	0.01974	54.3%	7.2%	54.3%	67.1%
3180	0.18026	0.2	0.01974	77.4%	1.7%	77.4%	99.0%
200	0.19	0.2	0.01	15.2%	8.0%	32.0%	15.2%
400	0.19	0.2	0.01	24.1%	11.8%	33.6%	24.1%
600	0.19	0.2	0.01	34.9%	14.8%	37.0%	34.9%
800	0.19	0.2	0.01	38.7%	15.0%	38.7%	43.5%
1000	0.19	0.2	0.01	39.2%	13.9%	39.2%	50.9%
1500	0.19	0.2	0.01	41.9%	12.4%	41.9%	67.6%
3180	0.19	0.2	0.01	51.4%	8.1%	51.4%	94.3%
400	0.22	0.22	0	16.1%	16.2%	25.5%	16.1%
600	0.22	0.22	0	22.8%	21.5%	26.5%	22.8%
796	0.22	0.22	0	25.4%	25.5%	25.4%	27.4%
932	0.22	0.22	0	24.9%	25.1%	24.9%	31.0%
1500	0.22	0.22	0	25.1%.	25.0%	25.1%	46.0%
352	0.25974	0.22	-0.03974	2.4%	42.2%	5.6%	2.4%

Table 2. Power from 10,000 clinical trials to declare efficacy of A using new and conventional testing procedures against eight different true hazard ratios of A vs B

			Sample size				
		Power using	200	250	350	500	932
True HR of A:B	$\frac{5}{3}$	*New measure*	0.1%	0.1%	0.1%	0.0%	0.0%
		Conventional	0.0%	0.0%	0.0%	0.0%	0.0%
	1.5	*New measure*	0.4%	0.4%	0.3%	0.1%	0.0%
		Conventional	0.0%	0.0%	0.0%	0.0%	0.0%
	$\frac{10}{7}$	*New measure*	0.7%	0.7%	0.6%	0.3%	0.0%
		Conventional	0.1%	0.1%	0.0%	0.0%	0.0%
	1.25	*New measure*	2.7%	2.6%	2.6%	2.5%	1.0%
		Conventional	0.2%	0.3%	0.2%	0.1%	0.0%
	1	*New measure*	11.7%	12.7%	16.3%	20.9%	24.6%
		Conventional	2.4%	2.6%	2.6%	2.5%	2.5%
	0.8	*New measure*	29.1%	34.2%	46.0%	60.2%	80.4%
		Conventional	10.5%	11.9%	15.2%	19.7%	32.7%
	0.7	*New measure*	43.2%	50.5%	65.6%	79.9%	95.2%
		Conventional	19.3%	22.2%	29.9%	40.1%	65.4%
	0.6	*New measure*	59.0%	68.4%	82.1%	92.7%	99.6%
		Conventional	32.2%	38.3%	51.3%	66.6%	90.2%

Table 3a. Comparison between inferences from Bayesian posterior probabilities, conventional hypothesis testing and our proposed testing procedure

		Posterior probability that HR is				Power to conclude efficacy	
n	**True HR**	**< 0.80**	**< 1.00**	**> 1.20**	**> 1.25**	**Conventional**	**New method**
100	1.5	5.3%	14.5%	72.7%	69.4%	0.1%	1.0%
	1.25	14.3%	29.6%	54.0%	50.0%	0.6%	1.9%
	1	30.7%	49.9%	34.1%	30.9%	2.5%	7.6%
	0.8	50.6%	70.1%	17.8%	15.7%	7.0%	17.2%
	0.6	75.3%	87.1%	6.6%	5.6%	17.0%	32.3%
		Posterior probability that HR is				**Power to conclude efficacy**	
n	**True HR**	**< 0.80**	**< 1.00**	**> 1.20**	**> 1.25**	**Conventional**	**New method**
150	1.5	2.5%	10.4%	75.5%	71.2%	0.1%	0.5%
	1.25	8.9%	24.9%	55.4%	50.2%	0.4%	2.0%
	1	26.8%	50.4%	29.9%	26.1%	2.0%	8.8%
	0.8	50.4%	73.6%	13.0%	10.8%	7.3%	21.8%
	0.6	78.1%	91.0%	3.5%	2.8%	22.8%	43.7%
		Posterior probability that HR is				**Power to conclude efficacy**	
n	**True HR**	**< 0.80**	**< 1.00**	**> 1.20**	**> 1.25**	**Conventional**	**New method**
200	1.5	1.2%	7.1%	79.0%	74.4%	0.0%	0.3%
	1.25	5.8%	21.2%	56.5%	50.2%	0.4%	2.5%
	1	23.8%	50.3%	27.1%	22.8%	2.6%	10.4%
	0.8	50.8%	76.4%	10.0%	7.9%	9.9%	28.8%
	0.6	80.7%	93.7%	1.9%	1.4%	31.1%	57.9%
		Posterior probability that HR is				**Power to conclude efficacy**	
n	**True HR**	**< 0.80**	**< 1.00**	**> 1.20**	**> 1.25**	**Conventional**	**New method**
250	1.5	0.5%	5.0%	82.0%	77.3%	0.0%	0.5%
	1.25	4.1%	19.0%	56.8%	50.1%	0.2%	2.3%
	1	20.9%	50.2%	25.2%	20.7%	2.6%	12.7%
	0.8	49.8%	78.2%	7.6%	5.8%	11.0%	33.4%
	0.6	83.9%	95.7%	1.0%	0.7%	37.9%	67.4%
		Posterior probability that HR is				**Power to conclude efficacy**	
n	**True HR**	**< 0.80**	**< 1.00**	**> 1.20**	**> 1.25**	**Conventional**	**New method**
300	1.5	0.3%	3.5%	84.3%	79.4%	0.0%	0.2%
	1.25	2.8%	16.5%	58.2%	50.0%	0.2%	2.0%
	1	18.4%	50.0%	23.1%	18.3%	2.4%	13.5%
	0.8	50.2%	80.5%	5.9%	4.3%	14.2%	41.4%
	0.6	85.7%	96.9%	0.6%	0.4%	45.0%	76.0%

Table 3b: Bayesian posterior probabilities and frequentist powers using Bayesian decision rules under $n = 150$ and $n = 250$ using an uninformative prior for the treatment effect

Note that in the following two tables, a frequentist power is observed, using decision rules based on Bayesian posterior probabilities. 'Power' refers to the percentage of simulations which fulfil $P(HR < 1) > \gamma$. In the HR = 1.5 and 1.25 scenarios, the 'power' is thus the probability of drawing an incorrect inference.

n = 150

	Power under testing procedure				
			Bayesian: Prob (HR < 1) > γ*		
HR	$\widehat{HR}$ < 1 and p-value < 0.05	New	γ = 80%	γ = 95%	γ = 91.42%
1.5	0.1%	0.6%	1.3%	0.2%	0.4%
1.25	0.4%	2.2%	4.7%	1.2%	1.9%
1	2.1%	8.6%	15.8%	5.1%	8.5%
0.8	7.8%	21.6%	33.0%	15.4%	22.1%
0.6	23.3%	45.5%	61.1%	37.5%	47.5%
0.4	54.1%	74.4%	86.1%	70.6%	77.3%
0.37	59.9%	78.2%	88.1%	75.3%	80.6%

n = 250

	Power under testing procedure				
			Bayesian: Prob (HR < 1) > γ*		
HR	$\widehat{HR}$ < 1 and p-value < 0.05	New	γ = 80%	γ = 95%	γ = 87.46%
1.5	0.0%	0.2%	0.6%	0.0%	0.2%
1.25	0.2%	2.5%	4.2%	0.6%	2.1%
1	2.6%	12.5%	20.0%	5.1%	12.2%
0.8	11.3%	34.7%	46.5%	20.0%	35.3%
0.6	38.0%	68.0%	79.1%	53.3%	69.2%
0.53	51.4%	78.4%	86.5%	66.4%	79.2%

*In addition to the posterior probability outlined, Prob (HR > 1.25) < 0.025 needs to be fulfilled to conclude efficacy.

Table 4. Comparison between theoretical power and observed power in 10,000 simulations

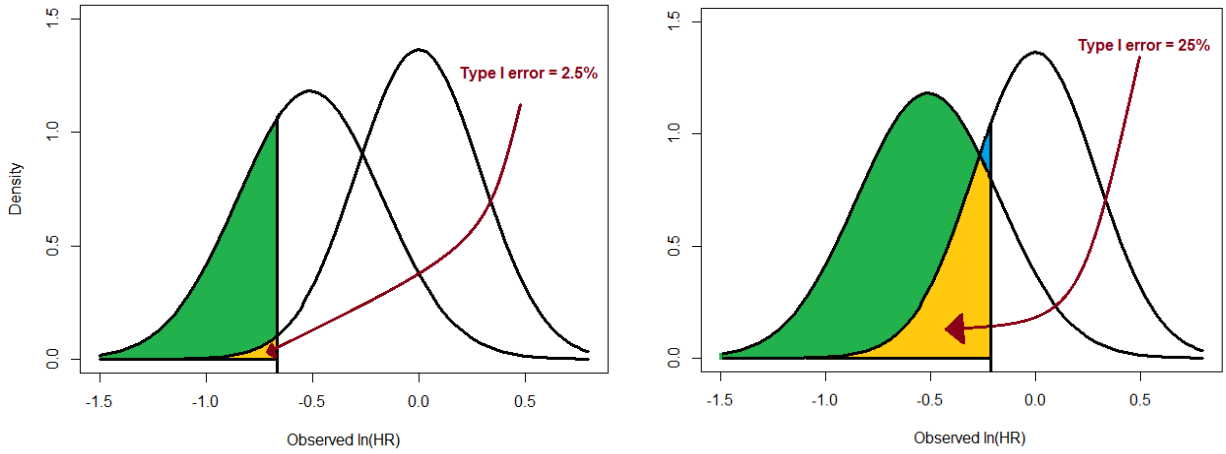
Event rate	True hazard ratio of A:B	Power sought	Theoretical sample size	Power attained	Sample size needed for required	% difference	% overpowered	$SE(ln(\hat{\theta}))$ model	$SE(ln(\hat{\theta}))$ actual	Power conventional	Gain in power
0.1	0.5	90%	656	89.5%	666	-1.5%	-0.6%	0.300	0.283	66.5%	23.0%
0.1	0.6	80%	728	80.3%	726	0.3%	0.3%	0.269	0.260	48.5%	31.8%
0.1	0.65	90%	1206	89.6%	1208	-0.2%	-0.4%	0.204	0.200	56.6%	33.1%
0.1	0.7	80%	1130	80.4%	1118	1.1%	0.4%	0.206	0.204	40.7%	39.7%
0.2	0.5	90%	328	89.5%	334	-1.8%	-0.5%	0.297	0.283	67.1%	22.4%
0.2	0.6	80%	346	78.1%	356	-2.9%	-2.4%	0.274	0.267	46.9%	31.2%
0.2	0.6	90%	462	88.6%	480	-3.9%	-1.6%	0.237	0.231	59.0%	29.6%
0.2	0.8	80%	1026	79.6%	1032	-0.6%	-0.5%	0.147	0.147	32.5%	47.1%
0.2	0.8	90%	1706	90.0%	1706	0.0%	0.0%	0.114	0.114	49.5%	40.5%
0.2	0.9	90%	7234	90.1%	7234	0.0%	0.1%	0.054	0.054	49.9%	40.1%
0.3	0.5	70%	130	70.3%	130	0.0%	0.4%	0.381	0.365	46.3%	24.0%
0.3	0.65	90%	396	90.4%	394	0.5%	0.5%	0.203	0.200	56.3%	34.2%
0.3	0.8	80%	680	80.4%	680	0.0%	0.5%	0.147	0.146	33.1%	47.3%
0.4	0.6	90%	232	89.4%	238	-2.6%	-0.7%	0.233	0.228	60.2%	29.3%
0.4	0.7	90%	352	89.7%	360	-2.3%	-0.3%	0.182	0.180	51.5%	38.2%
0.4	0.8	90%	844	90.4%	840	0.5%	0.5%	0.114	0.114	50.1%	40.3%
0.4	0.9	90%	3596	90.4%	3596	0.0%	0.5%	0.054	0.054	49.9%	40.6%
0.5	0.5	80%	98	81.0%	98	0.0%	1.2%	0.332	0.322	56.9%	24.1%
0.5	0.8	80%	404	80.1%	404	0.0%	0.1%	0.147	0.146	32.5%	47.6%
0.5	0.8	90%	670	89.5%	672	-0.3%	-0.6%	0.114	0.114	49.7%	39.8%
0.5	0.9	80%	1722	79.7%	1728	-0.3%	-0.4%	0.069	0.069	32.6%	47.1%
0.5	0.9	90%	2866	90.2%	2860	0.2%	0.3%	0.054	0.054	49.7%	40.6%
0.6	0.7	80%	172	79.4%	174	-1.2%	-0.8%	0.210	0.209	39.5%	39.9%
0.6	0.8	80%	334	79.7%	336	-0.6%	-0.3%	0.147	0.146	33.6%	46.1%

Figures

Figure 1. Graphical illustration of efficacy as defined using conventional and new definitions

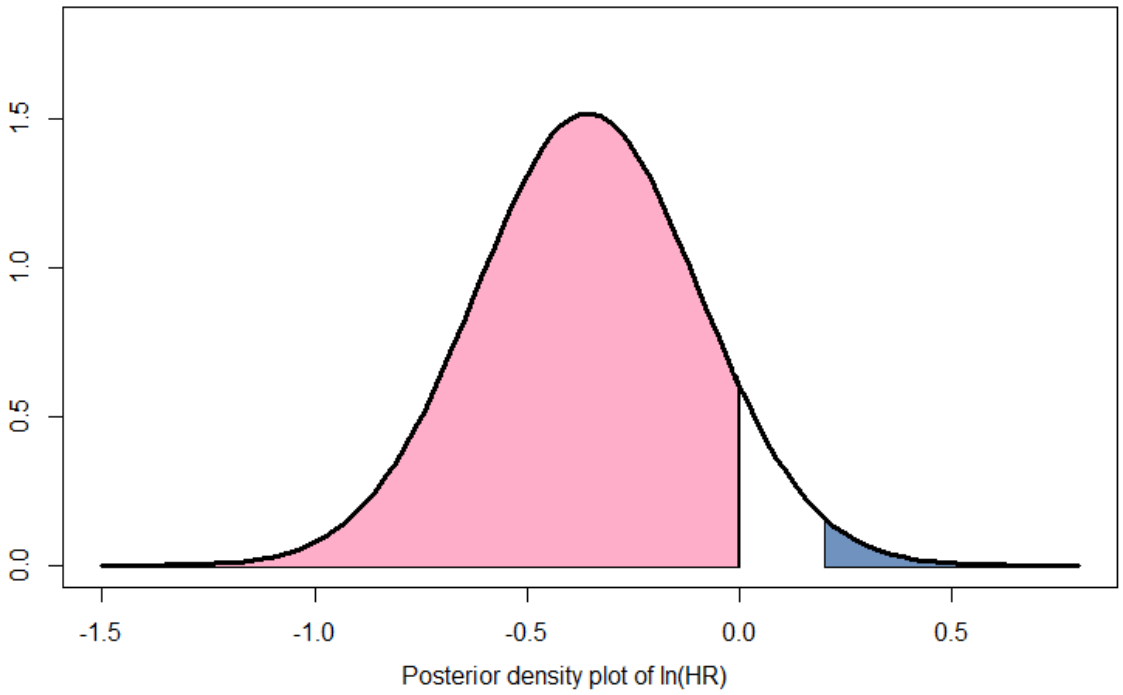
Left graph: Graphical illustration of power using conventional measure

Right graph: Graphical illustration of power using new measure



Power: green and yellow area; Type I error: yellow and blue areas.

Figure 2. Graphical demonstration of posterior probability requirements using when conducting hypothesis testing using Bayesian analysis


0.0
0.5
1.0
1.5
-1.5
-1.0
-0.5
0.0
0.5
Posterior density plot of ln(HR)

[1] Gatta G, van der Zwan JM, Casali PG, Siesling S, Dei Tos AP, Kunkler I, Otter R, Licitra L, Mallone S, Tavilla A, Trama A, Capocaccia R; RARECARE working group. Rare cancers are not so rare: the rare cancer burden in Europe. Eur J Cancer 2011; **47**(17): 2493-2511.
[2] Komatsubara KM, Carvajal RD. The promise and challenges of rare cancer research. The Lancet Oncology 2016; **17**(2): 136-138.
[3] Greenlee RT, Goodman MT, Lynch CF, Platz CE, Havener LA, Howe HL. The Occurrence of Rare Cancers in U.S. Adults, 1995-2004. Public Health Rep. 2010 Jan-Feb;125(1):28-43.
[4] Gupta S, Faughnan ME, Tomlinson GA, Bayoumi AM. A framework for applying unfamiliar trial designs in studies of rare diseases. J Clin Epidemiol 2011; **64**(10): 1085-1094.
[5] Wang D, Bakhai A. Clinical Trials: A Practical Guide to Design, Analysis, and Reporting. London: Remedica; 2005.
[6] Senn S. Statistical issues in Drug Development. West Sussex: John Wiley & Sons Ltd; 2007.
[7] Billingham L, Malottki K, Steven N. Research methods to change clinical practice for patients with rare cancers. The Lancet Oncology 2016; **17**(2): e70-e80.
[8] De Muth JE. Basic Statistics and Pharmaceutical Statistics Applications. Third edition. Florida: CRC Press; 2014.
[9] Everson-Stewart S, Emerson SS. Bio-creep in non-inferiority clinical trials. Stat Med 2010; **29**(27): 2769-2780.
[10] Flight L, Julious SA. Practical guide to sample size calculations: non-inferiority and equivalence trials. Pharm Stat 2016; **15**(1): 80-89.
[11] Freedman LS. Tables of the Number of Patients Required in Clinical Trials Using the Logrank Test. Statistics in Medicine 1982; **1**: 121-129.
[12] Nghiem PT, Bhatia S, Lipson EJ, Kudchadkar RR, Miller NJ, Lakshmanan A, Berry S, Chartash EK, Daud A, Fling SP, Friedlander PA, Kluger HM, *et al*. PD-1 Blockade with Pembrolizumab in Advanced Merkel-Cell Carcinoma. N Engl J Med 2016; **374**(26): 2542-2552.
[13] *Rare skin cancer in England*. National Cancer Intelligence Network. Available at: http://www.ncin.org.uk/publications/data_briefings/rareskincancer [last accessed 17 February 2019]
[14] Rational MCC Protocol Version 2.0. University Birmingham. Available online: https://njl-admin.nihr.ac.uk/document/download/2005841 [last accessed 17 February 2019]
[15] Cancer Research UK. Rare types of breast cancer. 16 October 2017. http://www.cancerresearchuk.org/about-cancer/breast-cancer/stages-types-grades/types/rare-types-breast-cancer [last accessed 17 February 2019]