

Design and analysis of non-inferiority mortality trials in oncology^{‡,¶}

Mark Rothmann^{1,*}, Ning Li¹, Gang Chen¹, George Y. H. Chi¹,
Robert Temple² and Hsiao-Hui Tsou¹

¹*Division of Biometrics I, OB/OPaSS/CDER, Food and Drug Administration, HFD-710,
WOCII 5600 Fishers Lane, Rockville, MD 20857, U.S.A.*

²*Office of Drug Evaluation I, ORM/CDER, Food and Drug Administration, HFD-102,
WOCII 5600 Fishers Lane, Rockville, MD 20857, U.S.A.*

SUMMARY

The recent revision of the Declaration of Helsinki and the existence of many new therapies that affect survival or serious morbidity, and that therefore cannot be denied patients, have generated increased interest in active-control trials, particularly those intended to show equivalence or non-inferiority to the active-control. A non-inferiority hypothesis has historically been formulated in terms of a fixed margin. This margin was historically designed to exclude a ‘clinically meaningful difference’, but has become recognized that the margin must also be no larger than the assured effect of the control in the new study. Depending on how this ‘assured effect’ is determined or estimated, the selected margin may be very small, leading to very large sample sizes, especially when there is an added requirement that a loss of some specified fraction of the assured effect must be ruled out. In cases where it is appropriate, this paper proposes non-inferiority analyses that do not involve a fixed margin, but can be described as a two confidence interval procedure that compares the 95 per cent two-sided CI for the difference between the treatment and the control to a confidence interval for the control effect (based on a meta-analysis of historical data comparing the control to placebo) that is chosen to preserve a study-wide type I error rate of about 0.025 (similar to the usual standard for a superiority trial) for testing for retention of a prespecified fraction of the control effect. The approach assumes that the estimate of the historical active-control effect size is applicable in the current study. If there is reason to believe that this effect size is diminished (for example, improved concomitant therapies) the estimate of this historical effect could be reduced appropriately. The statistical methodology for testing this non-inferiority hypothesis is developed for a hazard ratio (rather than an absolute difference between treatments, because a hazard ratio seems likely to be less population dependent than the absolute difference). In the case of oncology, the hazard ratio is the usual way of comparing treatments with respect to time to event (time to progression or survival) endpoints. The proportional hazards assumption is regarded as reasonable (approximately holding). The testing procedures proposed are conditionally equivalent to two confidence interval procedures that relax the conservatism of two

* Correspondence to: Mark Rothmann, Division of Biometrics I, Center for Drug Evaluation and Research, FDA, WOCII 5600 Fishers Lane, Rockville, MD 20857, U.S.A.

† E-mail: rothmannm@cder.fda.gov

‡ This article is a US Government work and is in the public domain in the U.S.A.

¶ The views expressed in this article do not necessarily represent those of the U.S. Food and Drug Administration.

Contract/grant sponsor: Center for Drug Evaluation and Research; contract/grant number: RSR-01-14.

95 per cent confidence interval testing procedures and preserve the type I error rate at a one-sided 0.025 level. An application of this methodology to Xeloda, a recently approved drug for the treatment of metastatic colorectal cancers, is illustrated. Other methodologies are also described and assessed – including a point estimate procedure, a Bayesian procedure and two delta-method confidence interval procedures. Published in 2003 by John Wiley & Sons, Ltd.

KEY WORDS: non-inferiority trials; active control trials

1. INTRODUCTION

Although there has been recent active discussion of the appropriateness of placebo-controlled trials and active-controlled trials in various settings (Declaration of Helsinki [1], International Conference on Harmonization E-10 guidance document [2], Temple and Ellenberg [3], Ellenberg and Temple [4] and Rothman and Michels [5]), there is no doubt that when an available treatment affects survival or irreversible morbidity, a placebo-controlled trial cannot be carried out (ICH E-10 [2], Temple and Ellenberg [3]). There are many such situations, including a wide range of circumstances in which cardiovascular interventions, use of anti-infective drugs, and treatments for malignancy have been shown to have such effects. In those cases, unless a new treatment can be shown superior to an established therapy or can be studied as an ‘add-on’ to existing therapy (standard treatment plus new drug versus either standard treatment alone or a placebo), it must be studied in an equivalence or non-inferiority trial (ICH-E-10 [2], Temple and Ellenberg [3]). Non-inferiority trials, however, pose significant interpretation problems, not always recognized.

It might at first be thought that a new treatment no better than existing treatment is not very important, so that non-inferiority would be of little interest, but new treatments can have safety and tolerability advantages that are substantial. SSRI antidepressants, new ‘atypical’ antipsychotics, non-sedating anti-histamines, and essentially all treatments for hypertension since diuretics and reserpine are not more effective than the drugs that preceded them. They are, however, much better tolerated and in many cases clearly safer. New treatments may also provide alternatives for people who do not respond to available therapy and there may, of course, be cost/competitiveness reasons for having more than one agent, even if they are equally effective (Ellenberg and Temple [4]). In oncology, while there is always a search for greater effectiveness, new therapies may also provide more attractive regimens (oral versus parenteral, shorter infusions versus longer infusions, less frequent dosing), even if they are not more effective. It is therefore desirable to be able to establish the effectiveness of such treatments, and how to conduct a meaningful non-inferiority trial is therefore of interest.

‘Non-inferiority’ is to some extent a misnomer; non-inferiority studies in fact seek to place a limit on the possible inferiority of the new treatment compared to control. Assurances that all of the control effect is retained can only be accomplished by showing superiority, so that permitting some potential inferiority is essential to developing new drugs that are actually of equal effectiveness. If the new treatment offers a better toxicity profile compared to the active-control or standard treatment, then it may even be considered beneficial when some efficacy is actually lost. In such cases, the objective of an active control trial should be to demonstrate that the new treatment is effective, retains a substantial fraction of the effect of the control and at the same time has better toxicity and/or safety profile. More commonly,

even if the new therapy does not have a clear advantage over existing treatments, it may be acceptable to approve new treatments based on a showing that the trial can rule out the loss of more than a certain 'clinically important' amount of the active-control effect. The issue then comes down to the proper choice of 'how much worse than the active-control' is clinically acceptable.

The most critical issue in designing a non-inferiority study intended to demonstrate the effectiveness of a drug is the choice of the degree of inferiority, M (often called the non-inferiority margin) that must be ruled out statistically by the study, in order to conclude that the new drug is effective. The hypotheses can be expressed as

$$H_0 : C-T \geq M \text{ versus } H_1 : C-T < M$$

where $C-T$ is the degree of superiority of the control treatment (C) to test treatment (T) and M may represent the entire effect of the control (relative to placebo) or a fraction of that effect (for example, 50 per cent). These hypotheses can be written for a hazard ratio analysis as

$$H_0 : \text{HR}(T/C) \geq 1 + M \text{ versus } H_1 : \text{HR}(T/C) < 1 + M$$

For a fixed value of M , the null hypothesis is ordinarily rejected when the two-sided 95 per cent CI for $C-T$ lies entirely below M . The choice of M involves several considerations:

- (i) *The effect of the control in the new non-inferiority study.* This is not measured in the trial but must be known or estimated from previous studies of C , generally versus placebo (P). The effect represented by M should be no larger than the entire effect of the drug. If $C-T$ were larger than M , then a non-inferiority finding could represent a loss of all of the effect versus the control.
- (ii) *The fraction of the effect of the control drug that needs to be preserved.* This is a clinical judgement. If the desired retention is very high, for example, 90 per cent, as noted above, studies become very large or as a practical matter only drugs that are actually superior will be successful. On the other hand, as the usual reason for using a two-arm non-inferiority design is the need to avoid exposure of patients to inferior treatment (that is, a placebo) when effective treatment exists, loss of much of the control agent's effect is unacceptable.

The linkage between the 'clinically important' difference and the known effect of the control in the study is critical and both values must be considered together. Thus, if the effect of the control agent is very large and easily identified (for example, cure rates in many infections), the clinically relevant difference would ordinarily be much smaller than the control effect, so that a fixed value of M could be chosen largely on clinical grounds. That may not be true, however, if the effect of the control drug is small. In that case, a difference that might be considered clinically relevant could actually be larger than the whole effect of the control. Showing $C-T$ to be smaller than that value would not indicate effectiveness of the new agent.

In the past, many new oncologic drugs were compared with established therapy and an arbitrary clinical cutoff of an increased risk of 25 per cent was used; that is, ruling out a hazard ratio for survival of not more than 1.25 for the experimental drug versus control would be considered evidence of effectiveness. The problem is that if the effect of the control was not at least 1.25, but say 1.15, showing non-inferiority would not represent evidence of any effectiveness.

Although the theoretical formulation of the non-inferiority comparison is clear enough, how to estimate the effect of the control in the new study is not. First, translating the historical experience with the control agent to a new situation is inevitably difficult, sharing many of the problems associated with attempts to conduct historically controlled trials. Second, even if one is comfortable with applying the historical estimates to the present study, there are still difficult choices.

One approach to ‘estimating’ the effect of the control treatment that has been used for thrombolytics has been to examine a meta-analysis of placebo-controlled trials, then use the lower limit of the one-sided 95 per cent CI for the control effect to identify an effect size that the control treatment would be likely to exceed in the new study. Loss of more than half of this effect is then to be ‘ruled out’ by showing that the upper limit of the one-sided 95 per cent CI for the relative effect of the new treatment ($C-T$) in the non-inferiority trial does not include a difference (loss) as large as half the ‘estimated’ effect size. This method has been called a two 95 per cent confidence interval testing procedure and it is clearly conservative, because it is comparing two ‘statistically worst’ scenarios.

Alternatives to the 95–95 approach have been suggested by various authors (Hauck and Anderson [6], Holmgren [7], Simon [8], Koch and Tangen [9], Hasselblad and Kong [10] and Wiens [11]). They propose not to choose a margin based on the lower bound of a 95 per cent confidence interval for historical experience, an approach that is quite conservative, but to compare the difference between treatment and control by making different use of the estimated effect of the control with its corresponding standard error.

It is the intent of this paper to discuss the objectives, the design considerations, the formulation of hypotheses, the statistical methodology, the critical underlying assumptions and the interpretation of active control non-inferiority trials. We will consider how to model the active-control effect, and define a test that maintains a reasonable, but not unusually stringent type I error probability. We will illustrate the approaches with non-inferiority analyses used in the recent approval of the drug Xeloda for the treatment of colorectal cancer. We will describe and discuss other non-inferiority procedures.

2. DESIGN, DESIGN ISSUES AND INTERPRETATION OF RESULTS

In this section we state hypotheses corresponding to a retention of a prespecified proportion of the active-control survival effect (versus a placebo or other reference therapy). The choice of retention proportion is a matter of judgement. We will also discuss modelling the active-control survival effect for the current trial, had the two arms in the current trial been the active-control and the placebo (or other reference therapy).

2.1. Hypotheses involving a retention of a proportion of the active-control effect

For oncology, a hazard ratio is regarded as the appropriate measure for comparing two arms for a time to event endpoint. The proportional hazards assumption is regarded as reasonable (approximately holding). For fixed $0 \leq \delta_0 \leq 1$, the hypothesis to be tested is whether the experimental treatment retains more than $100\delta_0$ per cent of the active-control effect on survival. For $HR(P/C) > 1$, the active-control survival effect may be defined by $HR(P/C) - 1$ or by $\log HR(P/C)$. These values represent, respectively, the increased risk or log-risk of an event using

the placebo or other reference therapy instead of the active-control. The difference in increased risk or log-risk of an event relative to the active-control, using the placebo or other reference therapy instead of the experimental treatment, is given, respectively, by $HR(P/C) - HR(T/C)$ or $\log HR(P/C) - \log HR(T/C)$. We have that $100\delta_0$ per cent of the active-control effect can be given, respectively, by $\delta_0 \times (HR(P/C) - 1)$ or by $\delta_0 \times \log HR(P/C)$. If the difference in increased risk or log-risk using the placebo or other reference therapy instead of the experimental treatment exceeds $100\delta_0$ per cent of the active-control effect, then the experimental treatment is said to retain more than $100\delta_0$ per cent of the active-control effect. Thus, when the active-control is effective (that is, $HR(P/C) > 1$), the null and alternative hypotheses can be expressed as (using some algebra)

$$H_0 : HR(T/C) \geq \delta_0 + (1 - \delta_0)HR(P/C) \text{ versus } H_1 : HR(T/C) < \delta_0 + (1 - \delta_0)HR(P/C) \quad (1a)$$

or, corresponding to a log scale, as

$$\begin{aligned} H_0 : \log HR(T/C) &\geq (1 - \delta_0) \log HR(P/C) \text{ versus} \\ H_1 : \log HR(T/C) &< (1 - \delta_0) \log HR(P/C) \end{aligned} \quad (1b)$$

Cases (1a) and (1b), respectively, correspond to the arithmetic and geometric definition of the proportion of active-control effect retained by the new treatment, which will be given later. Non-inferiority or effectiveness is demonstrated by the treatment, if it is shown that the treatment retains at least a δ_0 proportion of the active-control effect. For example, for $\delta_0 = 0.5$, rejecting an above null hypothesis means that the treatment preserves more than 50 per cent of the active-control effect. While inferences about $HR(T/C)$ are made from the non-inferiority trial, historical trials and changes in study conditions are used to make inferences about $HR(P/C)$.

In many oncologic drug trials, and in thrombolytic trials, $\delta_0 = 0.5$ has been used. Using $\delta_0 = 0.5$ is arbitrary and in practice may be chosen to limit the effectiveness of the control that can be lost and the selection of δ_0 should be on a case by case basis. One might ask why any loss of the control effect should be acceptable. In some cases, the answer might be that none would be, in which case the study must demonstrate superiority of the new drug to control. It must be noted, however, that estimates of the control effect are not precise, that as a practical matter, a new treatment even modestly worse than control is unlikely to be able to show non-inferiority in a study of reasonable size and that it is usually desirable to have therapeutic choices.

A higher proportion of the active-control effect to retain may also be required when the active-control is very effective, as with antibiotics or treatments for many leukaemias or lymphomas. In adjuvant settings, the effect of standard therapy tends to be quite a bit larger than in the corresponding metastatic setting (for example, colorectal cancer), thus a larger per cent retention might be specified in adjuvant settings.

Note that when $\delta_0 = 1$, the hypotheses are those for a superiority trial. When $\delta_0 = 0$, demonstration of greater than 0 per cent retention of the active-control survival effect is desired – this corresponds to demonstrating that the test drug is superior to the placebo/other reference therapy.

As noted above, the non-inferiority design depends on being able to state, with some assurance, what the effect of the active-control is in the new study even though there is no placebo group to help measure that effect directly.

Modelling the active-control effect for the current trial will be discussed in the next subsection.

2.2. *Modelling the active-control effect*

If there are no reliable historical studies for the active-control, an active-control non-inferiority trial cannot be used. When there are available well-controlled studies, it may be possible to estimate the active-control effect.

There are several considerations pertinent to ‘estimating’ the current active-control effect versus the placebo or other reference therapy, including the amount of data available to assess this effect, the size and consistency of the effect, and the similarity of the current trials to the past trials (including with respect to concomitant and/or subsequent therapies).

The weak link in any non-inferiority trial is the presumed active-control effect, which is not measured but estimated. This means all non-inferiority studies have elements of a cross-study comparison and depend on assumptions whose validity is uncertain. Modelling the active-control effect for the current trial ($HR(P/C)$) begins with availability of past studies comparing the active-control to placebo or other reference therapy. A meta-analysis using results from such studies provides a more reliable estimate of $HR(P/C)$ and information about the between-trial variability and heterogeneity. Important conditions of the current trial, including the active-control regimen, subsequent therapies, and patients’ disease and demographic characteristics, should be very similar to and central to those in the studies, which are included in the meta-analysis. Limitations of any meta-analysis should be recognized and the analysis treated cautiously.

Critical information from the trials in the meta-analysis may not be readily available. For example, a trial may only report the number of events and a log-rank test p -value, while the log-hazard ratio and the corresponding standard error estimate are what are needed. It may be possible, however, to extract the necessary information. Parmar *et al.* [12] introduced a series of methods to extract summary statistics from published literature for survival endpoints. These methods include the utilization of the p -value from the logrank test with the number of events, a Kaplan–Meier survival curve, and a hazard ratio from a Cox regression model. Even when these analyses are possible, it may be difficult to adjust for critical covariates, which may not be available. To obtain the overall active-control effect, a decision will need to be made (case by case) between using a fixed effects model or a random effects model for the meta-analysis, whichever is more appropriate. A random effects model (DerSimonian and Laird [13]) would utilize any between-trial variation.

Publication and selection biases are also major concerns in the determination of the active-control effect. If only favourable studies are included in the meta-analysis, the historical active-control survival effect will be overestimated, and the type I error probability associated with a non-inferiority finding (with respect to the null hypotheses in (1a) and (1b)) will be inflated.

We will present later an application involving the experimental oral treatment Xeloda, the active-control of infusional 5-fluorouracil with leucovorin (5-FU+LV) and the reference therapy of infusional 5-fluorouracil (5-FU). The drug Xeloda was approved in April 2001 for first-line metastatic colorectal cancer. For each of two trials, the efficacy analysis criterion

was a demonstration by Xeloda of a greater than 50 per cent retention of the survival effect of 5-FU+LV relative to 5-FU alone.

2.3. *Adjusting the active-control effect*

The effect of the active-control in the current trial may have changed (possibly reduced by some fraction). For example, changes in supportive care or subsequent therapies may dampen the active-control effect in the present study compared to the past. If there is reason to believe that the true log-hazard ratio of P versus C for the current study is different from the true mean (based on a random effects model, if used) log-hazard ratio of P versus C across those non-concurrent trials, then the estimator of the true mean log-hazard ratio of P versus C across those non-concurrent trials should be adjusted by some appropriate factor, $\theta > 0$. Let $\log \text{HR}(P2/C2)$ denote the true (theoretical) log-hazard ratio for the placebo (or other reference therapy) versus the active-control, had the two arms in the non-inferiority trial been the placebo (or other reference therapy) and the active-control and let $\log \text{HR}(P1/C1)$ denote the mean (theoretical) log-hazard ratio for the placebo or other reference therapy versus the active-control based on (estimated by) a meta-analysis model (or single-trial model). It may be reasonable to assume that

$$\text{HR}(P2/C2) - 1 = \theta[\text{HR}(P1/C1) - 1] \quad \text{for some fixed } 0 \leq \theta \leq 1 \text{ (or } \theta > 0)$$

where θ is the fraction of historical effect that is believed currently exists (in practice, $0 \leq \theta \leq 1$ may usually be relevant); thus, $\theta = 0$ means that there is no effect of the active-control in the current study and $\theta = 1$ means that no adjustment is needed. When the value of θ used is larger than the correct value of θ , procedures designed to maintain a desired type I error rate will have a larger type I error rate than desired (with respect to the null hypotheses in (1a) and (1b)). When the value of θ used is smaller than the correct value of θ , procedures designed to maintain a desired type I error rate will have a smaller type I error rate than desired (with respect to the null hypotheses in (1a) and (1b)).

Another possibility (on a log-scale) that assumes the active-control effect for the current trial is a fraction of the effect from the previous trials has

$$\log \text{HR}(P2/C2) = \theta \log \text{HR}(P1/C1) \quad \text{for some fixed } \theta \geq 0$$

2.4. *Standard error adjustment*

There may be situations to consider a higher value for the standard error than that given in a meta-analysis or from a single trial. When there are one or two historical trials that compare the active-control to placebo or other reference therapy, between-trial variability cannot be assessed. Not taking between-trial variability into consideration, when such variability exists, will lead to estimators of effects that seem to be more precise than they actually are. In practice, non-concurrent historical trials comparing the active-control to placebo or other reference therapy often have a considerably smaller number of patients than the current active-control trial being designed. These smaller trials tend to have some design differences, including varying patient populations (possibly due to regional differences and/or inclusion-exclusion differences). If all designs, patient populations, concurrent and subsequent care were identical, there would be no need to consider between-trial variability.

There may be other reasons for using a higher estimate of the standard error than that from a meta-analysis. For example, the standard error may be magnified if there is uncertainty of whether or not the active-control effect for the current trial is similar to that of the non-concurrent trials. If the quality of the standard error (for $\log \hat{HR}(P1/C1)$) from the meta-analysis is not believed to be on par with the quality of the standard error (for $\log \hat{HR}(T/C2)$) from the current trial, it may be desirable to model the current active-control effect with a larger standard error than that from the meta-analysis.

3. METHODS OF ANALYSIS

In this section we give two definitions of the proportion of active-control effect retained, specify the hypotheses of interest, and provide corresponding large sample normal test statistics and two confidence interval procedures. In metastatic settings in oncology, where standards of therapy usually have small survival effect sizes, studies are powered at equivalent survival with less than desired power. The stopping rule for the current trial is independent of the results for estimating the active-control effect. We will make that assumption here.

3.1. Definitions of the proportion of active-control survival effect retained, δ

- (i) *Arithmetic definition.* For the current trial, the arithmetic definition of the proportion of the active-control survival effect retained by the experimental treatment, δ , is given by

$$\delta = \frac{HR(P2/C2) - HR(T/C2)}{HR(P2/C2) - 1} = \frac{1 - HR(T/P2)}{1 - HR(C2/P2)} = 1 - \frac{HR(T2/C2) - 1}{HR(P2/C2) - 1}$$

when $HR(P2/C2) > 1$. For example, if $HR(P2/C2) = 1.3$ and $HR(T/C2) = 1.2$, then $\delta = 1/3$.

- (ii) *Geometric definition.* For the current trial, the geometric definition of the proportion of the active-control survival effect retained by the experimental treatment, δ , is given by

$$\delta = \frac{\log HR(P2/C2) - \log HR(T/C2)}{\log HR(P2/C2)} = \frac{\log HR(P2/T)}{\log HR(P2/C2)}$$

when $HR(P2/C2) > 1$. For example, if $HR(P2/C2) = 1.3$ and $HR(T/C2) = 1.2$, then $\delta = 0.30508$.

The geometric definition is a more appropriate definition for a relative measure of the proportion of active-control effect retained than the arithmetic definition. Differences of hazard ratios do not have any meaning. A hazard ratio of 1 describes no association. Hazard ratios of 0.8 and 1.25 describe the same level of association (different direction), but have different absolute differences with 1.

The two definitions agree at $\delta = 0$ and $\delta = 1$. Otherwise, the values for δ for both definitions will always fall on the same sides of 0 and 1. When the value for the geometric δ is between 0 and 1, the value for the arithmetic δ can be anywhere between the value for the geometric δ and 1.

3.2. Hypotheses

We will assume that $HR(P2/C2) > 1$, that is, that the active-control is effective in the current study. When testing whether the treatment maintains $100\delta_0$ per cent ($0 < \delta_0 < 1$) of the effect of the active-control, the hypotheses of interest are

$$H_0 : \delta \leq \delta_0 \text{ versus } H_1 : \delta > \delta_0 \quad (2)$$

For the arithmetic and geometric definitions, the hypotheses in (2) reduce to, respectively

$$\begin{aligned} H_0 : HR(T/C2) &\geq \delta_0 + (1 - \delta_0)HR(P2/C2) \text{ versus} \\ H_1 : HR(T/C2) &< \delta_0 + (1 - \delta_0)HR(P2/C2) \end{aligned} \quad (3a)$$

and

$$\begin{aligned} H_0 : \log HR(T/C2) &\geq (1 - \delta_0) \log HR(P2/C2) \text{ versus} \\ H_1 : \log HR(T/C2) &< (1 - \delta_0) \log HR(P2/C2) \end{aligned} \quad (3b)$$

When it is reasonable to assume $HR(P2/C2) = HR(P1/C1)$, we will test these following surrogate hypotheses and extrapolate to the above hypotheses:

$$\begin{aligned} H_0 : HR(T/C2) &\geq \delta_0 + (1 - \delta_0)HR(P1/C1) \text{ versus} \\ H_1 : HR(T/C2) &< \delta_0 + (1 - \delta_0)HR(P1/C1) \end{aligned} \quad (4a)$$

and

$$\begin{aligned} H_0 : \log HR(T/C2) &\geq (1 - \delta_0) \log HR(P1/C1) \text{ versus} \\ H_1 : \log HR(T/C2) &< (1 - \delta_0) \log HR(P1/C1) \end{aligned} \quad (4b)$$

When $HR(P2/C2) = HR(P1/C1)$, these above pairs of hypotheses are the same. Note that the alternative hypothesis for the geometric definition can be expressed as

$$\frac{1}{2 - \delta_0} \log HR(C2/T) + \frac{1 - \delta_0}{2 - \delta_0} \log HR(P1/C1) > 0$$

Thus we are testing whether a weighted average is positive. When $\delta_0 = 0$, each log-hazard ratio is given equal weight, whereas when $\delta_0 = 1$ no weight is given to the historical log hazard ratio ($\delta_0 = 1$ corresponds to a superiority trial). The larger δ_0 is, the less weight is given to the historical log-hazard ratio.

Two examples for adjusting the active-control effect are given below:

Example (i). For the arithmetic definition, in such cases when it is reasonable to assume $HR(P2/C2) - 1 = \theta[HR(P1/C1) - 1]$ for some fixed $\theta > 0$, we will test these following surrogate hypotheses and extrapolate to the hypotheses in (3a):

$$H_0 : HR(T/C2) \geq \delta_0 + (1 - \delta_0)(1 + \theta[HR(P1/C1) - 1])$$

and

$$H_1 : \text{HR}(T/C2) < \delta_0 + (1 - \delta_0)(1 + \theta[\text{HR}(P1/C1) - 1])$$

When $\text{HR}(P2/C2) - 1 = \theta[\text{HR}(P1/C1) - 1]$, these hypotheses are the same as the hypotheses in (3a). When $\text{HR}(P2/C2) - 1 \geq \theta[\text{HR}(P1/C1) - 1]$ ($\text{HR}(P2/C2) - 1 < \theta[\text{HR}(P1/C1) - 1]$), the null hypothesis in (3a) is a subset (superset) of the null hypothesis above.

Example (ii). For the geometric definition, in such cases when it is reasonable to assume $\log \text{HR}(P2/C2) = \theta \log \text{HR}(P1/C1)$ for some fixed $\theta > 0$, we will test these following surrogate hypotheses and extrapolate to the hypotheses in (3b):

$$H_0 : \log \text{HR}(T/C2) \geq (1 - \delta_0)\theta \log \text{HR}(P1/C1) \text{ versus}$$

$$H_1 : \log \text{HR}(T/C2) < (1 - \delta_0)\theta \log \text{HR}(P1/C1)$$

When $\log \text{HR}(P2/C2) = \theta \log \text{HR}(P1/C1)$, these hypotheses are the same as the hypotheses in (3b).

3.3. Setting non-inferiority cutoffs based on no uncertainty of the current active-control effect

If there is no uncertainty of the active-control effect, that is, the exact value of $\text{HR}(P2/C2)$ is known, then the test procedure that rejects the null hypothesis and concludes non-inferiority when the upper limit for the $100(1 - 2\alpha)$ per cent two-sided confidence interval for $\text{HR}(T/C2)$ lies beneath $\delta_0 + (1 - \delta_0) \times \text{HR}(P2/C2)$ has a one-sided significance level of α . For example, for $\delta_0 = 0.5$, $\alpha = 0.025$ and the belief with complete certainty that $\text{HR}(P2/C2) = 1.14$, then non-inferiority would be concluded for $\delta_0 = 0.5$, when the upper limit for the 95 per cent two-sided confidence interval for $\text{HR}(T/C2)$ lies beneath 1.07. If instead we have $\delta_0 = 0$ and $\alpha = 0.001$ with again the certainty that $\text{HR}(P2/C2) = 1.14$, then non-inferiority would be concluded for $\delta_0 = 0$ (superiority of the experimental treatment to the placebo or other reference therapy), when the upper limit for the 99.8 per cent two-sided confidence interval for $\text{HR}(T/C2)$ lies beneath 1.14 (at a one-sided significance level of 0.001).

3.4. The 95 per cent CI lower limit method

Results in Section 3.4 to Section 3.9 will be given for only the geometric definition. Arguments are similar and expressions are more involved for the arithmetic definition. Results for the arithmetic definition are given in the Appendix.

A two 95 per cent two-sided confidence interval procedure using the geometric definition is as follows. The lower limit of the 95 per cent two-sided confidence interval for $\log \text{HR}(P1/C1)$ is determined based on some model. The non-inferiority cutoff is defined as $(1 - \delta_0)(\text{lower limit of 95 per cent CI for } \log \text{HR}(P1/C1))$, for some given $0 < \delta_0 < 1$. If the 95 per cent two-sided confidence interval for $\log \text{HR}(T/C2)$ from the current trial lies entirely beneath this cutoff, non-inferiority is inferred – a better than $100\delta_0$ per cent retention (based on the geometric definition) of the active-control effect by the experimental treatment is inferred. This method is referred to as the ‘95 per cent CI lower limit’ method or as the two 95 per cent

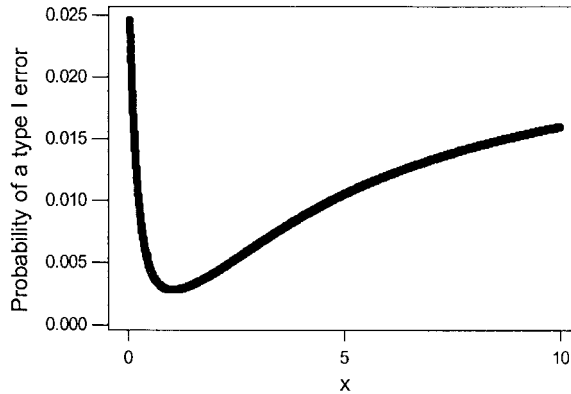


Figure 1. Type I error rate versus square root of the information ratio when using the 95 per cent CI lower limit to define the cutoff.

confidence interval method. This case has been represented as using a worst case standard to estimate the effect of the control ($\log \text{HR}(P2/C2)$) with no attributed uncertainty.

3.4.1. Type I error rate. Throughout this paper, unless otherwise stated, we will assume that $\text{HR}(P2/C2) = \text{HR}(P1/C1)$, that is, that $\theta = 1$. For a moderate or large prespecified number of events, the estimator of $\log \text{HR}(T/C2)$ will have an approximate normal distribution. Also, we use Φ to denote the standard normal distribution function. For simplification we introduce the following notation (SE = 'standard error'):

$$s_1 = \hat{\text{SE}}(\log \hat{\text{HR}}(T/C2)) \quad \text{and} \quad s_2 = \hat{\text{SE}}(\log \hat{\text{HR}}(P1/C1))$$

Using normal approximation, the probability of rejecting H_0 in (4b) at the boundary of H_0 (where the maximum probability is attained), is estimated by (approximately equal to)

$$\alpha = P(\text{Reject } H_0 \mid \text{bdry}(H_0)) \approx \Phi\left(\frac{-1.96 - 1.96x}{\sqrt{1+x^2}}\right) \quad (5)$$

where $x = (1 - \delta_0)s_2/s_1 > 0$. If the current active-control effect is correctly modelled ($\text{HR}(P2/C2) = \text{HR}(P1/C1)$), the expression in (5) gives an approximate probability of rejecting H_0 in (3b) at the boundary of H_0 . The value of x in (5) represents the square root of the information ratio between estimating $(1 - \delta_0)\log \text{HR}(P1/C1)$ and estimating $\log \text{HR}(T/C2)$. The approximate type I error probability, α , decreases from 0.025 near $x = 0$, to a minimum of $0.0027869 = \Phi(-1.96\sqrt{2})$ at $x = 1$ and then increases asymptotically towards 0.025 as $x \rightarrow \infty$ (see Figure 1). From Figure 1, we see for fixed estimates of the log-hazard ratio of the placebo (or other reference therapy) versus the active-control and corresponding standard errors, the approximate probability of a type I error (and hence, power at alternatives near the null hypothesis) is not monotone as the standard error for the log-hazard ratio of the test drug versus the active-control increases.

3.5. The point estimate method

The non-inferiority cutoff is given as $\delta_0 + (1 - \delta_0)(\text{estimate of } \log \text{HR}(P1/C1))$, for some given $0 < \delta_0 < 1$. If the 95 per cent two-sided confidence interval for $\log \text{HR}(T/C2)$ from the current trial lies entirely beneath this cutoff, non-inferiority is inferred. We will refer to this method as the 'point estimate method.' This case uses the point estimate for $\text{HR}(P1/C1)$ as the actual value, removing all uncertainty attached to the point estimate. Using normal approximation, the probability of rejecting H_0 in (4b) at the boundary of H_0 (where the maximum probability is attained), is estimated by (approximately equal to)

$$\alpha = \Phi \left(\frac{-1.96s_1}{\sqrt{\{s_1^2 + (1 - \delta_0)^2 s_2^2\}}} \right) \geq 0.025 \quad (6)$$

Note that α decreases as s_1/s_2 increases. When $s_1/s_2 \rightarrow \infty$, $\alpha \rightarrow 0.025$ and as $s_1/s_2 \rightarrow 0$, $\alpha \rightarrow 0.5$. Thus, the point estimate method can have a type I error rate that is anywhere between 0.025 and 0.5. Such rates are too high and unacceptable. Note also that for fixed s_1/s_2 , as δ_0 goes from 0 to 1 (a superiority trial), α decreases towards 0.025.

All these type I error probability results analogously hold when $\text{HR}(P2/C2)$ is correctly modelled via an adjustment to the estimator of $\text{HR}(P1/C1)$ (that is, in the case where $\log \text{HR}(P2/C2) = \theta \times \log \text{HR}(P1/C1)$).

3.6. Test statistic

The following is an asymptotic standard normal (at the boundary of the null hypothesis) test statistic which can be used to test those hypotheses in (4b):

$$Z^* = \frac{\log \hat{\text{HR}}(T/C2) - (1 - \delta_0) \log \hat{\text{HR}}(P1/C1)}{\sqrt{\{s_1^2 + (1 - \delta_0)^2 s_2^2\}}}$$

Under certain assumptions, this statistic quickly approaches a normal distribution. Comparing Z^* with -1.96 leads to an approximate one-sided 0.025 level test for testing the hypotheses in (4b) (and also in (3b), if $\text{HR}(P2/C2) = \text{HR}(P1/C1)$). Note that for fixed θ , for the adjustment of $\log \hat{\text{HR}}(P2/C2) = \theta \times \log(\hat{\text{HR}}(P1/C1))$, Z^* has the same calculated value among different pairs of (δ_0, θ) for which $(1 - \delta_0)\theta$ is the same.

A test statistic based on the arithmetic definition is given in the Appendix, which is similar to a test statistic Holmgren [7] found for relative risks.

3.7. Two CI procedures that are analogous to using test statistics

In this section we will express the large sample normal tests as two confidence interval procedures. Non-inferiority will be inferred if the upper limit of the 95 per cent CI for $\log \text{HR}(T/C2)$ lies entirely below $(1 - \delta_0)(\text{lower limit of the } 100\gamma \text{ per cent CI for } \log \text{HR}(P1/C1))$ for some fixed value of $0 \leq \gamma < 1$. Of particular interest is the value of γ that gives an approximate type I error probability of 0.025. The results in this section will be used in the next section to determine non-inferiority cutoffs, for tests, at the design stage.

Non-inferiority will be inferred when

$$\log \hat{\text{HR}}(T/C2) - (1 - \delta_0) \log \hat{\text{HR}}(P1/C1) < -1.96s_1 - (1 - \delta_0)\Phi^{-1} \left(\frac{1 + \gamma}{2} \right) s_2 \quad (7)$$

The estimator $\log \widehat{HR}(T/C2)$ has an asymptotic normal distribution and under certain assumptions $\log \widehat{HR}(P1/C1)$ has an asymptotic normal distribution. In such cases, at the boundary of the null hypothesis, the left-hand side of the above expression has an approximate normal distribution with mean 0 and a standard deviation estimated (from a consistent estimator) by $\sqrt{\{s_1^2 + (1 - \delta_0)^2 s_2^2\}}$. Using normal approximation gives the following estimate (approximation) to the probability of a type I error at the boundary of the null hypothesis (where the maximum probability is attained):

$$\alpha = P(\text{Reject } H_0 \mid \text{bdry}(H_0)) \approx \Phi \left(\frac{-1.96 - \Phi^{-1}(\frac{1+\gamma}{2})x}{\sqrt{(1+x^2)}} \right) \quad (8)$$

where $x = (1 - \delta_0)s_2/s_1 > 0$. We see that α is a decreasing function in x over $(0, \Phi^{-1}((1 + \gamma)/2)/1.96)$ and an increasing function in x over $(\Phi^{-1}((1 + \gamma)/2)/1.96, \infty)$. Also, note that, as $x \rightarrow 0, \alpha \rightarrow 0.025$ (regardless of the value of γ) and as $x \rightarrow \infty, \alpha \rightarrow (1 - \gamma)/2$.

Setting the expression in (8) equal to 0.025 and solving for γ gives

$$\gamma = 2\Phi \left(\frac{1.96\sqrt{(1+x^2)} - 1.96}{x} \right) - 1 \quad (9)$$

Using this value of γ leads to a two confidence interval test procedure that has an approximate 0.025 type I error probability (has the same test decision as comparing Z^* to -1.96). Hauck and Anderson [6] did something similar to this for means. From expression (9), we see that γ is a decreasing function in x from $(0, \infty)$ onto $(0, 0.95)$. For example, as the proportion of active-control effect, δ_0 , approaches 1 (a superiority requirement), less weight is given to the historical data and the value of γ needed for a one-sided 0.025 level test approaches 0 (the point estimate of the historical active-control effect is used to represent the active-control effect).

Example (iii). We will consider the metastatic colorectal cancer case involving the test drug Xeloda, the active-control of 5-fluorouracil with leucovorin (5-FU+LV) and the reference therapy of 5-fluorouracil alone (5-FU). To estimate the effect of 5-FU+LV on survival, the FDA [14] conducted a ten-study random effects meta-analysis of randomized comparisons of 5-FU alone (5-FU alone was considered to have no survival effect) versus 5-FU+LV. The results gave $\log \widehat{HR}(P1/C1) = 0.23411$ and $\widehat{SE}(\log \widehat{HR}(P1/C1)) = 0.07501$ (from the information given in Reference [14], the fixed effects estimate of the standard error is 0.05327). Also, from this review we have for each study comparing Xeloda (T) with 5-FU+LV ($C2$) that $\widehat{SE}(\log \widehat{HR}(T/C2)) \approx 0.0867$. For $\delta_0 = 0.5$, when the expression in (9) is set equal to 0.025, the solution for γ is 0.315 ($z = 0.406$). The lower limit of the 31.5 per cent CI for $\log \widehat{HR}(5-FU/5-FU+LV)$ based on the ten-paper meta-analysis is 0.204. Again, assuming $\widehat{HR}(P2/C2) = \widehat{HR}(P1/C1)$ (that is, $\theta = 1$), non-inferiority at 50 per cent of the active-control effect retained would be inferred had the upper limit of the 95 per cent CI for $\widehat{HR}(\text{Xeloda}/5-FU+LV)$ been below 1.107 ($\exp(0.5 \times 0.204)$). For these two studies, the 95 per cent two-sided confidence interval upper limit for $\widehat{HR}(\text{Xeloda}/5-FU+LV)$ were 1.181 and 1.089 (see Table III in Section 4). One of these two studies had its 95 per cent two-sided confidence interval upper limit below the non-inferiority cutoff; the other study had its 95 per cent two-sided confidence interval upper limit substantially above the non-inferiority cutoff.

Each trial had less than 25 per cent power to infer better than 50 per cent retention of the 5-FU+LV survival effect. Note that for the arithmetic definition, the non-inferiority cutoff is 1.111 based on the lower limit of the 34.9 per cent CI for HR(5-FU/5-FU+LV) of 1.222 (found by setting the expression in (A4) equal to 0.025 and solving for γ).

Retaining better than 0 per cent of the 5-FU+LV survival effect is equivalent to having better survival than 5-FU. For $\delta_0 = 0$, when the expression in (9) is set equal to 0.025, the solution for γ is 0.535 ($z = 0.730$). The lower limit of the 53.5 per cent CI for log HR(5-FU/5-FU+LV) based on the ten-paper meta-analysis is 0.179. Non-inferiority at better than 0 per cent of the active-control effect retained would be inferred for a given trial had the upper limit of the 95 per cent CI for HR(Xeloda/5-FU + LV) been below 1.196 ($\exp(0.179)$). Both Xeloda studies met that criterion.

3.8. Determining a non-inferiority cutoff based on a stopping rule of a fixed number of events

In this section we will determine non-inferiority cutoffs for tests at the design stage. These cutoffs are such that the test, which infers non-inferiority when the 95 per cent two-sided confidence interval for HR($T/C2$) from the current trial lies entirely beneath the cutoff, is approximately the same as the test, which uses a test statistic.

When the meta-analysis (or model for HR($P2/C2$)) is done at the design stage and s_1 is approximately known at the design stage, the value (or approximate value) of the 100 γ per cent CI lower limit for HR($P1/C1$) or log HR($P1/C1$) (and thus the corresponding non-inferiority cutoff) can be determined at the design stage.

Let n_1 denote the number of events in the active-control arm, n_2 denote to the number of events in the treatment arm and E denote expected values. The asymptotic standard error for the log-hazard ratio is given by $\sqrt{\{1/E(n_1) + 1/E(n_2)\}}$ (see Fleming and Harrington [15]), which is at least $2/\sqrt{n}$, where $n = n_1 + n_2$. For oncology, the stopping rule for the final analysis of time to event endpoints tends to be based on a prespecified fixed number of events. For a 1:1 randomization with n as the prespecified total number of events at stopping, $2/\sqrt{n}$ represents not only a lower bound, but also an approximation for the asymptotic standard deviation of log HR($T/C2$). Using this approximate lower bound as the standard error at the design stage allows us to algebraically 'equate' use of a test statistic with a prespecified two CI procedure. The standard error is not as well approximated at the design stage for many other endpoints/measures (for example, difference in proportions, (log) relative risks and (log) odds ratio).

For a test with a fixed one-sided level α and a 1:1 randomization, the corresponding two confidence interval procedure that rejects the null hypothesis when the 100(1 - 2 α) per cent two-sided confidence interval for HR(T/C) lies entirely beneath the cutoff, k , has k given by

$$\log k = \Phi^{-1}(1 - \alpha) \times 2/\sqrt{n} + (1 - \delta_0) \log \hat{R}(P1/C1) + \Phi^{-1}(\alpha) \sqrt{\{4/n + (1 - \delta_0)^2 s_2^2\}} \quad (10)$$

This follows from first setting $\log k = (1 - \delta_0)(\log(\hat{R}(P1/C1)) - \Phi^{-1}((1 + \gamma)/2)s_2)$ and then, using (8) with $\Phi^{-1}(1 - \alpha)$ replacing 1.96 and $2/\sqrt{n}$ substituted for s_1 to solve for $\Phi^{-1}((1 + \gamma)/2)s_2$.

Example (iv). Consider the case where $\delta_0 = 0.5$, there is a 1:1 randomization, and the analysis will occur after 1000 events have occurred (0.063246 is used for s_1) and a one-sided 0.025 significance level is desired. Table I gives the calculated cutoffs and

Table I. Calculated cutoffs for various normal models for $\log \text{HR}(P1/C1)$.

$\log \hat{\text{HR}}(P1/C1)$	SE	Cutoff	100γ per cent
0.234	0.075	1.102	40.9
0.211	0.0675	1.093	37.6
0.234	0.09	1.093	46.9

corresponding values of γ for normal models for $\log \text{HR}(P2/C2)$ for three cases of estimates of $\log \text{HR}(P2/C2)$ and corresponding standard errors.

Row 1 of Table I represents the results of a meta-analysis of 5-FU versus 5-FU+LV meta-analysis on survival (with no adjustment; see example (iii)). Row 2 models $\log \text{HR}(P2/C2)$ by $0.9 \times$ meta-analysis estimator ($\log \hat{\text{HR}}(P1/C1)$) and row 3 models $\log \text{HR}(P2/C2)$ by the meta-analysis estimator plus some additional variability. From Table I we see that the non-inferiority cutoffs are fairly similar in these three cases.

There are different forms for approximations for s_1^2 for different randomization ratios – for example, in a 2:1 treatment versus control randomization, s_1^2 can be closely approximated by the value (which is not a lower bound) $4.5/n$. For every randomization ratio, $4/n$ is a lower bound for s_1^2 . We see from expression (10) that the smaller the value substituted for s_1 , the smaller the calculated cutoff and thus the smaller the chance of rejecting the null hypothesis. Thus, the value of k in expression (10) would yield a two confidence interval procedure with a slightly smaller type I error rate than 0.025.

The larger the prespecified number of events, n , the smaller the calculated ‘non-inferiority cutoff’, k . A more conservative ‘non-inferiority cutoff’ (leading to a smaller than 0.025 type I error rate) can be determined by choosing a design-stage substitute for s_1^2 that is smaller than $4/n$. An universal non-inferiority cutoff (leading to a smaller than 0.025 type I error rate in every case), can be determined by choosing a design-stage substitute for s_1^2 of $4/\text{maximum possible } n$.

3.9. Design-stage power and event-size determination

We will define the design-stage power as the probability of rejecting the null hypothesis at some alternative conditioned on the model of the active-control effect (the estimate and corresponding standard error) being known. Because of this condition, such power depends on the alternative selected only through the value of $\text{HR}(T/C2)$. Suppose that the final analysis cutoff date is the time of the n th event (n to be determined). Let k be the corresponding cutoff for an approximate (one-sided) size α -level test (the cutoff that equates the two confidence interval test procedure with comparing Z^* with $\Phi^{-1}(\alpha)$) and let $1 - \beta$ be the design-stage power at $\log \text{HR}(T/C2)$, where $\log k > \log \text{HR}(T/C2)$.

In addition to equation (14), k, α, β and n are related through the sample size calculation equation based on Wald’s test statistic given in (11) below:

$$n(\log k - \log \text{HR}(T/C2))^2 = 4 \times (\Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta))^2, \\ \text{provided } \log k > \log \text{HR}(T/C2) \quad (11)$$

Table II. Number of events needed for 80 per cent power.

HR($T/C2$)	Number of events		
	Geometric	Arithmetic	Holmgren
1	4800	4816	19803
0.95	1505	1466	1855
0.9	750	728	810
0.85	446	433	460
0.8	291	284	295

Substituting the RHS of (10) for $\log k$ in equation (11) gives:

$$n(\Phi^{-1}(1-\alpha) \times 2/\sqrt{n} + (1-\delta_0) \log \hat{HR}(P1/C1) + \Phi^{-1}(\alpha) \sqrt{\{4/n + (1-\delta_0)^2 s_2^2\}} - \log \text{HR}(T/C2))^2 = 4 \times (\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta))^2$$

Since $1-\beta > \alpha$ (that is, $\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta) > 0$), the above equation reduces to

$$\frac{2^* \Phi^{-1}(1-\beta)}{\sqrt{n}} = (1-\delta_0) \log \hat{HR}(P1/C1) - \log \text{HR}(T/C2) + \Phi^{-1}(\alpha) \sqrt{\left\{ \frac{4}{n} + (1-\delta_0)^2 s_2^2 \right\}} \quad (12)$$

provided $\log k > \log \text{HR}(T/C2)$. When all other quantities are known in equation (12), this expression can be solved for n . An analogous sample size equation using the arithmetic definition is given in the Appendix. That equation is quite different (explanation in the Appendix) from that in Holmgren [7] (modifying Holmgren's equation from involving relative risks to a hazard ratio), which is also given in the Appendix.

Note that for $\delta_0 = 1$ (a superiority test), equation (12) and the equations in the Appendix all reduce to

$$n = \frac{4(\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta))^2}{(\log \text{HR}(T/C2))^2}$$

which is the standard event size calculation using Wald's test.

Example (v). We refer back to example (iii), which involved first line metastatic colorectal cancer trials using the active-control of 5-FU+LV. Again, we have $\log \hat{HR}(P1/C1) = 0.23411$ and $\text{SE}(\log \hat{HR}(C1/P1)) = 0.07501$. Let $\alpha = 0.025$, $\beta = 0.2$ and $\delta_0 = 0.5$. For both definitions and also Holmgren's equation, Table II gives the number of events needed for various values of $\text{HR}(T/C2)$.

For first-line metastatic colorectal cancer, we see from Table II that unless a treatment is actually somewhat superior to 5-FU+LV in its effect on survival, it may not be feasible to do a non-inferiority survival trial with an active-control of 5-FU+LV. The reversal in the first row ($4816 > 4800$) is due a greater deviation from normality for Z_2 compared to Z .

Table III. Summary of relevant survival descriptive statistics.

Study	HR(X/5-FU+LV)	log HR	SE(log HR)	95 per cent CI
SO14695	1.00	-0.0036	0.0868	(0.844, 1.181)
SO14796	0.92	-0.0844	0.0867	(0.775, 1.089)

Table IV. Xeloda non-inferiority analysis results.

Study	Value for Z^*	Value for Z_2^*	Solutions for δ_0 to	
			$Z^* = -1.96$	$Z_2^* = -1.96$
SO14695	-1.276	-1.323	0.091	0.095
SO14796	-2.133	-2.163	0.590	0.611

4. APPLICATION: TWO XELODA VS 5-FU+LV TRIALS IN METASTATIC COLORECTAL CANCER

Xeloda is an orally administrated drug containing 5'-deoxy-5-fluorouridine that is converted to 5-FU. 5-FU+LV can only be administrated through intravenous infusion and has an approved regimen for first-line metastatic colorectal cancer.

There were two randomized trials of about 600 patients each comparing Xeloda with 5-FU+LV. For each trial, the efficacy criterion was a demonstration that Xeloda had a greater than 50 per cent retention (using the arithmetic definition) of the survival effect of 5-FU+LV relative to 5-FU alone. Although 5-FU alone may have a slightly better survival effect than a true placebo, there are no studies documenting this. There have, however, been ample studies comparing 5-FU+LV to 5-FU as first-line treatment of metastatic colorectal cancer and the difference between their treatments can be considered a conservative estimate of the effect of 5-FU+LV compared to placebo. Table III summarizes survival results for each trial (FDA [14]) based on intent-to-treat populations. Given for each case in Table III are the estimate for the Xeloda versus 5-FU+LV hazard ratio, the log-hazard ratio estimate with corresponding standard error and the 95 per cent confidence interval for the true Xeloda versus 5-FU+LV hazard ratio.

As given in example (iii), we have the estimate of $\log \text{HR}(5\text{-FU}/5\text{-FU+LV})$ as 0.23411 with corresponding standard error of 0.07501. Table IV gives the analysis results under the assumption that the survival effect of 5-FU+LV compared to 5-FU has not changed, that is, $\text{HR}(P1/C1) = \text{HR}(P2/C2)$. The calculated values for test statistics are given for each definition of the proportion of effect retained – geometric (Z^*) and arithmetic (Z_2^*). Also, for each definition of the proportion of effect retained, that proportion which gives a test statistic value of -1.96 (would just result in statistical significance) is given. Setting $Z = -1.96$ and solving for δ_0 is a direct example of a Fieller approach for finding the lower limit of a two-sided 95 per cent confidence interval for a ratio (Read [16]).

We see that a greater than 50 per cent retention of the 5-FU+LV survival effect was inferred for trial SO14796 but not for trial SO14695. For the arithmetic (geometric) definition, the largest per cent of the 5-FU+LV survival effect retained by Xeloda that can be statistically

Table V. Models using Z_2^* and a 0.025 significance level that gives the same conclusion for these Xeloda trials as the 95 per cent CI lower limit method.

Model	log HR($P2/C2$) Estimate	Standard error	δ_0	'Cutoff'
1	0.08709	0	0.5	1.0455
2	0.09224	0.02955	0.5	1.0455
3	0.23411	0.07501	0.815	1.0455
4	0.23411	0.10501	0.7995	1.0455

concluded is 9.5 per cent and 61 per cent (9.1 per cent and 59 per cent), respectively, for studies SO14695 and SO14796.

The robustness of demonstrating at least 50 per cent retention by Xeloda of the 5-FU+LV survival effect for study SO14796 can be examined by determining what adjustments could have been made to the estimation of the 5-FU+LV survival effect and still conclude at least 50 per cent retention by Xeloda of the 5-FU+LV survival effect. In the arithmetic definition case, for the adjustment of $\text{HR}(P2/C2) - 1 = \theta[\text{HR}(P1/C1) - 1]$, statistical significance would have just been reached ($Z_2^* = -1.96$), for any pair of (δ_0, θ) , for which, $(1 - \delta_0)\theta = 0.389$ – for example, $\delta_0 = 0.5$ and $\theta = 0.778$. Thus, for study SO14796, a statistically significant result would have occurred, even if we had reduced the 5-FU+LV historical survival effect by 22 per cent. For the geometric definition case, for study SO14796, a statistically significant result would have occurred, even if we had reduced the 5-FU+LV historical survival effect by 18 per cent.

4.1. Comparison to the 95 per cent CI lower limit method

The 95 per cent CI lower limit method (arithmetic definition – see Appendix) gives a non-inferiority cutoff of 1.0455. Table V gives various normal models for the current placebo versus active-control hazard ratio for which comparing Z_2^* with -1.96 gives the same conclusion as comparing a 95 per cent two-sided confidence interval for $\text{HR}(T/C2)$ with 1.0455. Models 1–4 would give non-inferiority cutoffs of 1.0455 based on equation (A5) in the Appendix. In this setting, these four models give different interpretations for using a cutoff of 1.0455.

Models 1 and 2 test for more than 50 per cent retention of the active-control effect. Model 1 uses the 95 per cent CI lower limit of $\log \text{HR}(P1/C1)$ as an estimate of $\log \text{HR}(P2/C2)$ with no uncertainty. Model 2 reduces the estimate of $\log \text{HR}(P1/C1)$ by 60.6 per cent to estimate $\log \text{HR}(P2/C2)$ ($\log \text{HR}(P1/C1) = (1 - 0.606) \times \log \text{HR}(P2/C2)$). Model 3 estimates $\log \text{HR}(P2/C2)$ with the estimate of $\log \text{HR}(P1/C1)$ with no adjustment to the estimate or its uncertainty – but tests for a more than 81.5 per cent retention of the active-control effect. Model 4 estimates $\log \text{HR}(P2/C2)$ with the estimate of $\log \text{HR}(P1/C1)$ with an addition of 0.03 to the standard error and tests for a more than 79.95 per cent retention of the active-control effect.

5. OTHER TEST PROCEDURES

In this section we will discuss other test procedures used for hypotheses based on the proportion of effect retained. Simon's method is a Bayesian procedure, in which the test-

ing decision is based on a calculated posterior probability. Hassalblad and Kong's procedure and an analogous arithmetic δ CI procedure are delta-method (Taylor's theorem) confidence interval approaches.

5.1. Simon's method with non-informative priors

Normal posterior probability distributions (or a posterior bivariate normal distribution) are determined from non-informative priors (Simon [8]). A posterior probability is found for the event that both $\log \text{HR}(T/C2) < (1 - \delta_0) \log \text{HR}(P1/C1)$ and $\log \text{HR}(P1/C1) > 0$. This incorporates the uncertainty of the active-control being effective (better than the placebo or other reference therapy). Non-inferiority would be inferred, if this probability is greater than 0.975 (this is similar to a one-sided 0.025 level test).

Note that the posterior probability of $\log \text{HR}(T/C2) < (1 - \delta_0) \log \text{HR}(P1/C1)$ is one minus the one-sided p -value using the test statistic Z^* . Thus, if the posterior probability (with non-informative priors) that $\log \text{HR}(P1/C1) > 0$ is very close to 1, tests based on Z^* (p -values $< \alpha$) and Simon's method with non-informative priors (posterior probability $> 1 - \alpha$) will be approximately the same. When the posterior probability that $\log \text{HR}(P1/C1) > 0$ is between 0.975 and 1, either the posterior probability threshold of 0.975 or the critical value of -1.96 for Z^* can be changed, so that Simon's method and the test using the statistic Z^* will be approximately equal.

5.2. Hasselblad and Kong's procedure

The following procedure is given in Hassalblad and Kong [10]. The proportion of historical effect of the active-control retained is estimated by

$$\hat{\delta} = \frac{\log \hat{\text{HR}}(P1/C1) - \log \hat{\text{HR}}(T/C2)}{\log \hat{\text{HR}}(P1/C1)} = 1 - \frac{\log \hat{\text{HR}}(T/C2)}{\log \hat{\text{HR}}(P1/C1)}$$

A '95 per cent' two-sided confidence interval is calculated using a normal distribution for $\hat{\delta}$ with estimated standard deviation (error) given by

$$\sqrt{\left\{ \left(\frac{\log \hat{\text{HR}}(T/C2)}{\log \hat{\text{HR}}(P1/C1)} \right)^2 \left(\frac{s_1^2}{(\log \hat{\text{HR}}(T/C2))^2} + \frac{s_2^2}{(\log \hat{\text{HR}}(P1/C1))^2} \right) \right\}}$$

If this confidence interval lies above δ_0 , non-inferiority is inferred.

The above estimator of δ is one minus a ratio of two independent quantities, each of which has an approximate normal distribution. Such a quantity has an approximate normal distribution only in limited cases (when the denominator behaves as approximately a non-zero constant with respect to the numerator).

5.3. Analogous arithmetic δ CI procedure

The proportion of historical effect of the active-control retained is estimated by

$$\hat{\delta} = \frac{\hat{\text{HR}}(P1/C1) - \hat{\text{HR}}(T/C2)}{\hat{\text{HR}}(P1/C1) - 1} = 1 - \frac{\hat{\text{HR}}(T/C2) - 1}{\hat{\text{HR}}(P1/C1) - 1}$$

Table VI. Simulated probabilities using the Hassalblad and Kong procedure and an analogous arithmetic δ CI procedure.

δ_0	HR(P1/C1)	Hassalblad and Kong		Arithmetic δ CI procedure	
		Probability '95 per cent CI' > δ_0	Probability '95 per cent CI' < δ_0	Probability '95 per cent CI' > δ_0	Probability '95 per cent CI' < δ_0
0	1.25	0.0715	0	0.0879	0
0.5	1.25	0.0340	0	0.0540	0
1	1.25	0.0025	0.0026	0.0027	0.0004
0	1.5	0.0560	2×10^{-6}	0.0691	0
0.5	1.5	0.0386	0.0004	0.0604	6×10^{-6}
1	1.5	0.0115	0.0114	0.0137	0.0026

A '95 per cent' two-sided confidence interval is calculated using a normal distribution for $\hat{\delta}$ with estimated standard deviation (error) given by

$$\sqrt{\left\{ \left(\frac{\widehat{\text{HR}}(T/C2)}{\widehat{\text{HR}}(P1/C1) - 1} \right)^2 s_1^2 + \left(\frac{(\widehat{\text{HR}}(T/C2) - 1)\widehat{\text{HR}}(P1/C1)}{(\widehat{\text{HR}}(P1/C1) - 1)^2} \right)^2 s_2^2 \right\}}$$

If this confidence interval lies above δ_0 , non-inferiority is inferred.

This estimator of δ will have an approximate normal distribution only in limited cases (when $\hat{\sigma}_1^2$ is rather small and the denominator behaves as approximately a non-zero constant with respect to the numerator).

We scrutinized these delta-method CI procedures with simulations (see below).

5.4. Simulation results

We considered the following model (we are assuming $\theta = 1$). Suppose $\log \widehat{\text{HR}}(P1/C1) \sim N(\log \text{HR}(P1/C1), \sigma_1^2)$, and $\log \widehat{\text{HR}}(T/C2) \sim N(\log \text{HR}(T/C2), \sigma_2^2)$ are independent where $\log \text{HR}(T/C2) = (1 - \delta_0) \log \text{HR}(P1/C1)$ for the Hassalblad and Kong procedure simulations, and $\text{HR}(T/C2) = \delta_0 + (1 - \delta_0)\text{HR}(P1/C1)$ for the analogous arithmetic δ CI procedure method simulations. For each case we will use $\sigma_1 = \sigma_2 = 0.1$. Each case involves 500000 simulations. The proportion of times the calculated interval lies entirely above δ_0 and entirely below δ_0 are determined and given, respectively, for the Hassalblad and Kong procedure and the arithmetic δ CI procedure in Table VI.

For both procedures, in these cases the simulated type I error probability was much larger than 0.025 when δ_0 equalled 0, moderately larger than 0.025 when δ_0 equalled 0.5 and much smaller than 0.025 when δ_0 equalled 1. All of these simulated type I error probabilities were larger for arithmetic δ CI procedure than those in analogous cases were for the Hassalblad's and Kong's procedure. The simulated probabilities that the '95 per cent CI for δ ' lies entirely below δ was very small in every case.

6. SUMMARY AND DISCUSSION

The effectiveness of a drug is unambiguously documented by showing superiority to a control treatment (placebo or alternative therapy). In some cases, however, a placebo cannot be

used and superiority is either not expected or too difficult to show. This paper considered non-inferiority studies where a fixed proportion of the active-control survival effect is to be demonstrably retained. A two confidence interval approach is described that can be established at the design stage, along with identification of a specified number of events, that is equivalent to using a large sample normal test statistic and that preserves roughly the desired type I error rate. The non-inferiority test procedure must utilize the effect of the active-control and consider whether this effect is changed under present conditions.

In any non-inferiority study, which depends on cross-study comparisons for interpretation, there may be better validity if all trials (current trial and those comparing to the placebo or other reference therapy) adjust for the same meaningful covariates. Factors that can influence the type I error probability or extrapolation of results include: the validity of historical data (selection and publication bias etc.); changes in the effect of the active-control; the criteria for statistical significance; the variances of those log-hazard ratio estimators and the desired proportion of active-control effect to retain. The weight given to the historical trials in such comparisons becomes smaller as the proportion of the active-control effect to retain increases towards 1 and thus, the smaller the influence of these factors (for 100 per cent retention of the active-control effect, we have a superiority trial and all historical data are ignored).

When there is an active-control effect, maintaining an appropriate type I error probability under reasonable assumptions should be desired. For the 95 per cent CI lower limit method, the type I error probability is arbitrary under any assumptions and depends on a ratio of standard errors (placebo or other reference therapy versus active-control:treatment versus active-control). The 95 per cent CI lower limit method does not directly address biases, but rather expects that the statistically worst case solves such problems. A less conservative approach is described above in which the historical results and desired proportion of retention are used with a narrower confidence interval of the active-control versus placebo (or other reference therapy) log-hazard ratio, to define a non-inferiority cutoff that is based on maintaining a type I error rate of 0.025.

Active-control non-inferiority trials should be considered on a case-by-case basis. We need to have assurance that the active-control effect exists in the current patient population. We need to assess whether the current effect size, if an effect exists, has diminished. We should have multiple historical, randomized placebo-controlled studies that show relatively consistent results. Careful attention should be paid to the conduct and analysis of such a study.

APPENDIX: ARITHMETIC DEFINITION CASE

A1. 'Ninety-five per cent lower limit' method

Non-inferiority is inferred if the 95 per cent two-sided confidence interval for $HR(T/C2)$ from the current trial lies entirely beneath the cutoff of $\delta_0 + (1 - \delta_0)$ (lower limit of 95 per cent CI for $HR(P1/C1)$), for some given $0 < \delta_0 < 1$.

A1.1. Type I error rate: for simplification we set

$$r = \frac{(1 - \delta_0)\hat{HR}(P1/C1)}{\delta_0 + (1 - \delta_0)\hat{HR}(P1/C1)}$$

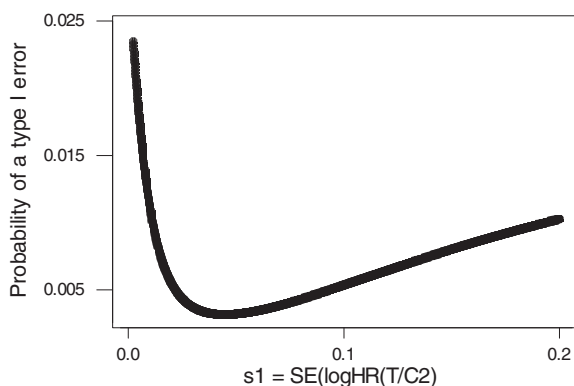


Figure A1. Probability of a type I error versus $s_1 = \text{SE}(\log \text{HR}(T/C2))$.

By the delta method, the approximate variance of $\log(\delta_0 + (1 - \delta_0)\hat{\text{HR}}(P1/C1))$ is given by $r^2 s_2^2$.

Using normal approximation, the probability of rejecting H_0 in (4a) at the boundary of H_0 (where the maximum probability is attained), is estimated by (approximately equal to)

$$\alpha = P(\text{Reject } H_0 \mid \text{bdry}(H_0)) \approx \Phi \left(\frac{-1.96s_1 + \log \frac{\delta_0 + (1 - \delta_0)\hat{\text{HR}}(P1/C1)e^{-1.96s_2}}{\delta_0 + (1 - \delta_0)\hat{\text{HR}}(P1/C1)}}{\sqrt{(s_1^2 + r^2 s_2^2)}} \right) \quad (\text{A1})$$

This probability will be somewhere between roughly 0.002 and 0.025. If the current active-control effect is correctly modelled ($\text{HR}(P2/C2) = \text{HR}(P1/C1)$), the expression in (A1) gives an approximate probability of rejecting H_0 in (3a) at the boundary of H_0 .

Example (vi). Consider a case where $\delta_0 = 0.5$, $\log \hat{\text{HR}}(P1/C1) = 0.3$ and $\text{SE}(\log \hat{\text{HR}}(P1/C1)) = 0.075$. Figure A1 is a graph of the approximate type I error probability versus $\text{SE}(\log \hat{\text{HR}}(T/C2))$ (which ranges from 0.002 to 0.2) using expression (A1).

We see from Figure A1 that the probability of a type I error is close to 0.025 when $s_1 \approx 0$ reaches a minimum of about 0.002 when $s_1 \approx 0.21$ and then increases.

A2. Point estimate method

Non-inferiority is inferred if the 95 per cent two-sided confidence interval for $\text{HR}(T/C2)$ from the current trial lies entirely beneath the cutoff of $\delta_0 + (1 - \delta_0)(\text{estimate of } \text{HR}(P1/C1))$, for some given $0 < \delta_0 < 1$. Using normal approximation, the probability of rejecting H_0 in (4a) at the boundary of H_0 (where the maximum probability is attained), is estimated by (approximately equal to)

$$\alpha = \Phi \left(\frac{-1.96s_1}{\sqrt{(s_1^2 + r^2 s_2^2)}} \right) \geq 0.025 \quad (\text{A2})$$

Note that, as $rs_2 \rightarrow 0$, $\alpha \rightarrow 0.025$ and as $s_1/s_2 \rightarrow 0$, $\alpha \rightarrow 0.5$. For fixed s_1/s_2 , as δ_0 goes from 0 to 1, α decreases towards 0.025. These type I error probability results analogously hold

when $\text{HR}(P2/C2)$ is correctly modelled via an adjustment to estimating $\text{HR}(P1/C1)$ (that is, $\text{HR}(P2/C2) - 1 = \theta \times [\text{HR}(P1/C1) - 1]$). Respectively, the estimator of $\log \text{HR}(P2/C2)$ and corresponding standard error are $\log(1 + \theta \times [\hat{\text{HR}}(P1/C1) - 1])$ and $(\theta \times \hat{\text{HR}}(P1/C1))/(1 + \theta \times [\hat{\text{HR}}(P1/C1) - 1]) \times s_2$.

A3. Test statistic

The following test statistic has an asymptotic standard normal distribution (at the boundary of the null hypothesis) under certain conditions:

$$Z_2^* = \frac{\log \hat{\text{HR}}(T/C2) - \log(\delta_0 + (1 - \delta_0)\hat{\text{HR}}(P1/C1))}{\sqrt{(s_1^2 + r^2 s_2^2)}}$$

Comparing Z_2^* with -1.96 leads to an approximate one-sided 0.025 level test for testing the hypotheses in (4a) (and also in (3a) if $\text{HR}(P2/C2) = \text{HR}(P1/C1)$). Holmgren [7] found an analogous statistic for relative risk. Note, for $\delta_0 = 0$ and also $\delta_0 = 1, Z_2^* = Z^*$. For fixed θ , when $\text{HR}(P2/C2) - 1$ is modelled by $\theta \times [\text{HR}(P1/C1) - 1]$, Z_2^* has the same calculated value among different pairs of (δ_0, θ) for which $(1 - \delta_0)\theta$ is the same (see the substitutions mentioned in Section A2). An approximate standard normal distribution for Z_2^* has been verified by simulations under various sets of practical values for $\text{HR}(P1/C1)$ and δ_0 .

A3.1. Assessment of the normality of Z_2 . We modelled $\log \hat{\text{HR}}(P1/C1)$ and $\log \hat{\text{HR}}(T/C2)$ as having independent normal distributions. Simulations were performed to assess the distribution of Z_2^* when $\delta_0 = 0.5$. At δ_0 equal to 0 or 1, Z_2^* has a standard normal distribution. Thus, the closer δ_0 is to 0 or 1, the closer the distribution of Z_2^* should resemble a standard normal distribution. We performed 500000 simulations each for various practical models. Under these practical models, simulations showed that the distribution of Z_2^* is close to the distribution of a standard normal. We provide simulation results for two models. Models

$\log \hat{\text{HR}}(P1/C1) \sim N(\log \text{HR}(P1/C1), \sigma_1^2)$ and $\log \hat{\text{HR}}(T/C2) \sim N(\log \text{HR}(T/C2), \sigma_2^2)$ are independent where $\text{HR}(T/C2) = \delta_0 + (1 - \delta_0)\text{HR}(P1/C1)$. For these examples, $\sigma_1 = \sigma_2 = 0.1$. Model 1 has $\text{HR}(P1/C1) = 1.25$ and model 2 has $\text{HR}(P1/C1) = 1.5$. Each case involves 500000 simulations and results are summarized in Table A1.

A4. Two CI procedures and non-inferiority cutoffs

Consider the two confidence interval test procedure. For some given $0 < \delta_0 < 1$, non-inferiority is inferred if the 95 per cent two-sided confidence interval for $\text{HR}(T/C2)$ from the current trial lies entirely beneath the cutoff of $\delta_0 + (1 - \delta_0)$ (lower limit of the 100 γ per cent CI of $\text{HR}(P1/C1)$) for some $0 \leq \gamma < 1$. Non-inferiority will be inferred when

$$\log \hat{\text{HR}}(T/C2) - \log \left(\delta_0 + (1 - \delta_0) e^{\log \hat{\text{HR}}(P1/C1) - \Phi^{-1}(\frac{1+\gamma}{2}) s_2} \right) < -1.96 s_1 \quad (\text{A3})$$

Making use of the asymptotic normal distribution of the left-hand side (delta method) at the boundary of H_0 in (4a), the probability of rejecting H_0 in (4a) is estimated by (approximately equal to)

$$\alpha = P(\text{Reject } H_0 \mid \text{bdry}(H_0))$$

Table A1. Simulation results involving Z_2^* .

Percentiles	Model		Standard normal
	1	2	
N	500000	500000	
Proportion below -1.96	0.02485	0.02533	0.025
Proportion above 1.96	0.02488	0.02545	0.025
1st	-2.319	-2.326	-2.326
2.5th	-1.957	-1.963	-1.960
5th	-1.642	-1.651	-1.645
10th	-1.286	-1.289	-1.282
25th	-0.679	-0.683	-0.674
50th	-0.008	-0.008	0
75th	0.664	0.668	0.674
90th	1.272	1.278	1.282
95th	1.638	1.649	1.645
97.5th	1.958	1.967	1.960
99th	2.331	2.329	2.326
Mean	-0.006	-0.006	0
Standard deviation	0.998	1.001	1

$$\approx \Phi \left(\frac{-1.96s_1 + \log \frac{\delta_0 + (1 - \delta_0)\hat{HR}(P1/C1)e^{-\Phi^{-1}(\frac{1+\gamma}{2})s_2}}{\delta_0 + (1 - \delta_0)\hat{HR}(P1/C1)}}{\sqrt{(s_1^2 + r^2s_2^2)}} \right) \quad (\text{A4})$$

The expression in (A4) gives a close approximation to the true value that is slightly higher than the true value. Setting the expression in (A4) equal to 0.025 and solving for γ leads to a two confidence interval test procedure that has an approximate 0.025 type I error probability (this is the same test as comparing Z_2^* to -1.96). For a prespecified value of δ_0 , the solution for γ depends mostly on the two standard deviation estimates and also the estimate of $\text{HR}(P1/C1)$. Note that, as $s_2 \rightarrow 0, \alpha \rightarrow 0.025$ (regardless of the value of γ) and as $s_1 \rightarrow 0, \alpha \rightarrow (1 - \gamma)/2$ (going back to expression (A3) and simplifying).

For a test with a fixed one-sided level α and a 1:1 randomization ratio, the corresponding two confidence interval procedure which rejects the null hypothesis when the $100(1 - 2\alpha)$ per cent two-sided confidence interval for $\text{HR}(T/C)$ lies entirely beneath the ‘cutoff’, k , has k given by

$$\log k = \Phi^{-1}(1 - \alpha) \times 2/\sqrt{n} + \log(\delta_0 + (1 - \delta_0)\hat{HR}(P1/C1)) + \Phi^{-1}(\alpha)\sqrt{(4/n + r^2s_2^2)} \quad (\text{A5})$$

This follows from first setting $\log k = \log(\delta_0 + (1 - \delta_0)\hat{HR}(P1/C1)e^{-\Phi^{-1}(\frac{1+\gamma}{2})s_2})$ and then using (A4) with $\Phi^{-1}(1 - \alpha)$ replacing 1.96 and $2/\sqrt{n}$ substituted for s_1 to solve for $\log(\delta_0 + (1 - \delta_0)\hat{HR}(P1/C1)e^{-\Phi^{-1}(\frac{1+\gamma}{2})s_2})$.

A5. Design-stage power and event-size determination

Since $1 - \beta > \alpha$ (that is, $\Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta) > 0$), substituting the RHS of (A5) for $\log k$ in equation (11) gives

$$n(\Phi^{-1}(1 - \alpha) \times 2/\sqrt{n} + \log(\delta_0 + (1 - \delta_0)\hat{HR}(P1/C1))) \\ + \Phi^{-1}(\alpha)\sqrt{(4/n + r^2s_2^2) - \log HR(T/C2)}^2 = 4(\Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta))^2$$

Reducing the above expression gives

$$2*\Phi^{-1}(1 - \beta)/\sqrt{n} = \log \left(\frac{\delta_0 + (1 - \delta_0)\hat{HR}(P1/C1)}{HR(T/C2)} \right) + \Phi^{-1}(\alpha)\sqrt{(4/n + r^2s_2^2)} \quad (A6)$$

provided $\log k > \log HR(T/C2)$. When all other quantities are known in equation (A6), this expression can be solved for n .

Equation (A6) is quite different from that in Holmgren [7]. Modifying Holmgren's sample size equation from relative risks to a hazard ratio and substituting $2/\sqrt{n}$ for s_1 , gives the equation

$$n = 4/\{[(\log(HR(T/C2)/(\delta_0 + (1 - \delta_0)HR(P2/C2))))/ \\ (\Phi^{-1}(1 - \beta) + \Phi^{-1}(1 - \alpha))^2] - r^2s_2^2\}$$

Holmgren defines an unconditional power (type II error probability) at alternatives analogous to values of $\log HR(T/C2) - \log(\delta_0 + (1 - \delta_0)HR(P1/C1))$. In Holmgren's equation, $\hat{HR}(P1/C1)$ is treated as unknown at the design stage, but its corresponding standard error is treated as known.

ACKNOWLEDGEMENTS

This research is funded by the Review Science Research Grant, RSR-01-14, from the Center for Drug Evaluation and Research, U.S. Food and Drug Administration. The authors would like to thank the referees for their helpful comments, which have improved the presentation of this paper.

REFERENCES

1. World Medical Association. Declaration of Helsinki. Ethical principles for medical research involving human subjects. *Journal of the American Medical Association* 2000; **284**:3043–3045.
2. International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). E-10: Guidance on choice of control group in clinical trials. *Federal Register* **64**:51767–51780.
3. Temple R, Ellenberg SS. Placebo controlled trials and active-control trials in the evaluation of new treatments. Part 1: Ethical and scientific issues. *Annals of Internal Medicine* 2000; **133**:455–463.
4. Ellenberg SS, Temple R. Placebo controlled trials and active-control trials in the evaluation of new treatments. Part 2: Practical issues and specific cases. *Annals of Internal Medicine* 2000; **133**:464–470.
5. Rothman KJ, Michels KB. The continued unethical use of placebo controls. *New England Journal of Medicine* 1994; **331**:394–398.
6. Hauck W, Anderson S. Some issues in the design and analysis of equivalence trials. *Drug Information Journal* 1999; **33**:109–118.

7. Holmgren EB. Establishing equivalence by showing that a specified percentage of the effect of the active control over placebo is maintained. *Journal of Biopharmaceutical Statistics* 1999; **9**(4):651–659.
8. Simon R. Bayesian design and analysis of active control clinical trials. *Biometrics* 1999; **55**:484–487.
9. Koch GG, Tangen CM. Nonparametric analysis of covariance and its role in non-inferiority clinical trials. *Drug Information Journal* 1999; **33**:1145–1159.
10. Hasselblad V, Kong DF. Statistical methods for comparison to placebo in active-control trials. *Drug Information Journal* 2001; **35**:435–449.
11. Wiens B. Choosing an equivalence limit for non-inferiority or equivalence studies. *Controlled Clinical Trials* 2002; 1–14.
12. Parmar MKB, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Statistics in Medicine* 1998; **17**:2815–2834.
13. DerSimonian R, Laird N. Meta-analyses in clinical trials. *Controlled Clinical Trials* 1986; **7**:177–188.
14. FDA Medical-Statistical review for Xeloda (NDA 20-896) dated 23 April, 2001.
15. Fleming TR, Harrington D. *Counting Processes and Survival Analysis*. Wiley: New York; 1991.
16. Read CB. Fieller's theorem. In *Encyclopedia of Statistical Sciences*, Kotz S, Johnson N (eds). Wiley: New York, 1983; 86–88.