Published in final edited form as:

Clin Trials. 2016 April ; 13(2): 188–198. doi:10.1177/1740774515614542.

Sample Size Under the Additive Hazards Model

Lee S. McDaniel¹, Menggang Yu², and Rick Chappell³

¹Biostatistics Program, School of Public Health, Louisiana State University Health Sciences Center, 2020 Gravier Street, 3rd Floor, New Orleans, LA 70112, lmcda4@lsuhsc.edu, phone: 504-568-5771

²Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, Wisconsin 53792, U.S.A.

³Department of Statistics and Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, Wisconsin 53706, U.S.A.

Abstract

Background—The additive hazards model can be easier to interpret and in some cases fits better than the proportional hazards model. However, sample size formulas for clinical trials with time to event outcomes are currently based on either the proportional hazards assumption or an assumption of constant hazards.

Aims—The goal is to provide sample size formulas for superiority and non-inferiority trials assuming an additive hazards model but no specific distribution, along with evaluations of the performance of the formulas.

Methods—Formulas are presented that determine the required sample size for a given scenario under the additive hazards model. Simulations are conducted to ensure that the formulas attain the desired power. For illustration, the non-inferiority sample size formula is applied to the calculations in the SPORTIF III trial of stroke prevention in atrial fibrillation.

Conclusions—Simulation results show that the sample size calculations lead to the correct power. Sample size is easily calculated using a tool that is available on the web at <http://leemcdaniel.github.io/samplesize.html>.

Keywords

Sample size; non-inferiority; time-to-event; additive hazards

Introduction

Additive hazards

The proportional hazards model is dominant in the analysis of clinical trials in which the primary outcome is the time to an event. If we are designing a superiority trial and the additive hazards model is more appropriate for the data than the proportional hazards model,

then sample size calculations may be less accurate if they are based on the proportional hazards assumption. So if we believe that the additive hazards model fits better, it is more appropriate to design the trial using an additive hazards assumption.

The additive hazards model has an even greater advantage in interpretability and this is an important factor in the design of non-inferiority trials. When designing a non-inferiority trial, it is necessary to define a non-inferiority margin and its scale. If the experimental treatment is not worse than the active control by this pre-specified margin, then the experimental treatment is considered non-inferior. An additive hazards margin can have greater appeal than a proportional hazards margin due to clinical relevance. For example, the designers of the Stroke Prevention Using Oral Thrombin Inhibitor in Atrial Fibrillation (SPORTIF III) trial defined a new treatment to be equivalent to the standard of care if it was proven to not increase event rates by 2% or more per year. This type of specification is intuitive to clinicians. Cox¹ also notes interpretation problems with the proportional hazards model due to the lack of a link to an underlying generating process. Aalen² discusses weaknesses of the proportional hazards model, including the fact that the “proportional intensity assumption is vulnerable to modifications in the number of covariates modeled.” The importance of considering the scale of the margin in a non-inferiority trial is discussed by Chappell³ and Garrett⁴.

After providing an introduction to non-inferiority trials and the specific additive hazards model we will be using, we present the first of our results: a sample size formula for a superiority trial under the additive hazards assumption that is not based on a distribution. Then we move on to the calculation of the sample size for a non-inferiority trial. Simulation results are provided, and the calculations are applied to the SPORTIF III trial. We conclude with a discussion of the results.

Non-inferiority trials

Non-inferiority trials are clinical trials with an active control with the intention of showing a new treatment is not much less effective than the active control. So a non-inferiority trial could be successful even if it is shown that the new treatment is at worst slightly inferior to the active control. We can still claim effectiveness of the new treatment due to prior knowledge about the active control. The known effectiveness of the active control supports the conclusion that the new treatment is effective.⁵

The non-inferiority margin plays an important part in the design of non-inferiority trials, as it represents the maximum loss of efficacy in the new treatment that we are willing to accept. We will denote this margin by $\delta > 0$. If a trial tests the difference in means, where a *higher mean is considered unfavorable*, then the corresponding non-inferiority hypotheses would be

$$H_0: \mu_B - \mu_A \geq \delta$$

$$H_1: \mu_B - \mu_A < \delta,$$

where μ_B is the mean for the experimental group and μ_A is the mean for the control group. Of course, hazards, log-hazards, or other quantities may be used instead of means.

The selection of δ is a contentious subject which can lead to problems such as degrading drug standards. A sensible rule is that δ should be substantially less than the effect size of the control drug. To ensure this, a percentage of the active control's effect size is frequently chosen. However, DeMets and Friedman⁶ argue that a percentage of the effect size is not an appropriate way to select the margin, and clinical concerns should be the overriding factor.

Figure 1, adapted from the U.S. Food and Drug Administration (FDA)⁵, shows possible outcomes of a non-inferiority trial. In it, outcomes (1), (4), and (5) are all situations where the new treatment is clearly shown to be non-inferior to the active control. In fact, case (5) demonstrates superiority of the experimental drug which implies non-inferiority. Possibilities (2) and (3) do not provide enough evidence to reject the hypothesis that the new treatment is inferior. Outcome (6) is also a situation in which the new treatment would be considered non-inferior, but with the added complication that the active control is shown to be superior to the new treatment.

If we want to design a non-inferiority trial with time to some event, denoted T , as the outcome, then a convenient way to specify the hypotheses is as a relationship between the hazard functions for the two treatments. If we wish to specify the hypotheses in terms of a hazard ratio, a simple formula exists.⁷ However it might be the case, as in SPORTIF III, that the hazard ratio is less interesting than the hazard difference. Our goal is to provide a score-type test and sample size formula based on a hazard difference for non-inferiority trials.

The specific set of hypotheses we will look at are

$$H_0: \lambda_B(t) - \lambda_A(t) \geq \delta \text{ for some } t \in [0, \tau]$$

$$H_1: \lambda_B(t) - \lambda_A(t) < \delta \forall t \in [0, \tau].$$

where $\lambda_B(t)$ is the hazard function for the experimental group, $\lambda_A(t)$ is the hazard function for the control group, δ is the non-inferiority margin, and τ is the time at the end of the study.

We define T as the failure time, C as the censoring time, and X , the minimum of T and C , as the observation time.

SPORTIF example

The SPORTIF III trial provides a motivating example for the current work. The new treatment ximelagatran was compared to the active control warfarin in the prevention of ischaemic strokes in patients with atrial fibrillation. Warfarin requires dose adjustment for each patient where ximelagatran does not, which makes ximelagatran a useful alternative. The event rate for the control group was estimated prior to the trial to be 3.1% per year and

the experimental treatment would be deemed non-inferior if the event rate for the experimental group was less than 5.1% per year.⁸

SPORTIF III was a randomized, open-label trial with 3410 patients from 259 hospitals, doctor's offices, or clinics. The primary endpoint was stroke or systemic embolism. Ximelagatran was found to be at least as effective as warfarin, with the point estimate of the absolute risk difference on a yearly basis of -0.7% and a 95% CI of [-1.4, 0.1].⁹

With the low event rates for stroke or systemic embolism, the exponential distribution provides a good approximation for analysis purposes. However, if the primary outcome had been more common, the exponential assumption may have led to incorrect conclusions, since a low event rate ensures that the exponential assumption is adequate.¹⁰ For example, 31% of patients had bleeding at some point in the trial. In this situation, the hazard rate is not so low that we can reasonably appeal to the exponential distribution, and the estimated hazard function is not close to constant as shown in Figure 2. So the exponential hazard assumption would not be appropriate for future trials with this endpoint.

Background

Aalen² presents an additive hazards model as an alternative to Cox's proportional hazards model, allowing for a different relationship between the hazard functions. The original additive hazards model is

$$\lambda(t) = \lambda_0(t) + \gamma(t)'Z(t). \quad (1)$$

where $Z(t)$ is a vector of covariates and $\gamma(t)$ is a vector of unknown coefficients. Lin and Ying¹¹ modify the model in (1) by removing the possibility of time-varying coefficients. So the new model is

$$\lambda(t) = \lambda_0(t) + \gamma'Z(t).$$

Consider n independent subjects such that the counting process $\{N_i(t); t \geq 0\}$ for the i th subject is 1 if the subject had an event in $[0, t]$, with $N_i(t)$ adapted to the filtration $\mathcal{F}_t$. Then

the hazard function for $N_i(t)$, defined as $\lim_{h \downarrow 0} \frac{1}{h} E(N_i(t+h) - N_i(t) | \mathcal{F}_t)$, is given by

$$Y_i(t)\lambda_i(t) = Y_i(t)\{\lambda_0(t) + \gamma'Z_i(t)\},$$

where $Y_i(t)$ is a 0–1 predictable at risk process that takes value 1 if the i th subject is at risk at time t and $Z_i(\cdot)$ is the covariate process for the i th subject. Lin and Ying¹¹ propose the following estimating function to estimate γ

$$U(\gamma) = \sum_{i=1}^n \int_0^\infty \{Z_i(t) - \bar{Z}(t)\} \{dN_i(t) - Y_i(t)\gamma'Z_i(t)dt\},$$

where

$$\bar{Z}(t) = \frac{\sum_{j=1}^n Y_j(t) Z_j(t)}{\sum_{j=1}^n Y_j(t)}.$$

In the case of a randomized two-arm non-inferiority trial we only need one covariate, a treatment indicator, which can reasonably be assumed to remain constant for the duration of the trial. So if we define Z_i to be 1 if the i th subject is on treatment B and 0 if treatment A , then the hazard function for the i th subject will be

$$\lambda_i(t) = \lambda_A(t) + \gamma Z_i \quad (2)$$

and the estimating function will be

$$U(\gamma) = \sum_{i=1}^n \int_0^\infty \{Z_i - \bar{Z}(t)\} \{dN_i(t) - Y_i(t) \gamma Z_i dt\}. \quad (3)$$

In what follows, $\lambda_A(\cdot)$ and $\lambda_B(\cdot)$ are the hazard functions for the control and experimental groups. Whenever $S(\cdot)$, $\lambda(\cdot)$, or $\Lambda(\cdot)$ are used without a subscript identifying the treatment or control group, it is because the function is the same for both groups.

Sample size calculations for superiority trials

We will first look at sample size calculations for a superiority trial with the alternative hypothesis of an additive hazards difference. We assume the Lin and Ying model, which in the superiority case reduces to

$$\lambda_B(t) = \lambda_A(t) + \gamma$$

and the hypotheses are

$$H_0: \gamma = 0 \quad \text{vs.} \quad H_1: \gamma \neq 0.$$

We further assume that treatment is properly randomized with each subject assigned to the experimental group with probability p . We include the independent censoring assumption (i.e. event times are independent of censoring times given Z), the assumption that all subjects are independent, and the assumption that the censoring distribution, which has distribution function G , is the same in the experimental and control groups. Finally, we assume that there are no ties in event times.

We first let n_B be the initial sample size of the treatment group,

$$Y_B(t) = \sum_{i=1}^n Z_i Y_i(t), \quad N_B(t) = \sum_{i=1}^n Z_i N_i(t) \text{ and}$$

$$M_B(t) = \sum_{i=1}^n Z_i \left\{ N_i(t) - \int_0^t Y_i(u) \lambda_B(u) du \right\},$$

with Z_i being the treatment indicator for the i th subject, is a martingale with respect to $\mathcal{F}_t$.

Then n_A is the initial sample size for the control group,

$$Y_A(t) = \sum_{i=1}^n (1 - Z_i) Y_i(t), N_A(t) = \sum_{i=1}^n (1 - Z_i) N_i(t), \text{ and}$$

$$M_A(t) = \sum_{i=1}^n (1 - Z_i) \left\{ N_i(t) - \int_0^t Y_i(u) \lambda_A(u) du \right\},$$

Finally, we define $Y(t) = Y_A(t) + Y_B(t)$, $N(t) = N_A(t) + N_B(t)$, and $n = n_A + n_B$.

The test statistic

Under the null hypothesis (i.e. $\gamma = 0$), Equation (3) becomes

$$U(0) = \int_0^\infty \{1 - \bar{Z}(t)\} dM_B(t) - \int_0^\infty \bar{Z}(t) dM_A(t)$$

The two martingale integrals above are uncorrelated (since there are no ties and $\bar{Z}(t)$ is predictable with respect to $\mathcal{F}_t$), the variance under the null is

$$\text{Var}_{H_0}(U(0)) = \int_0^\infty \{1 - \bar{Z}(t)\}^2 Y_B(t) \lambda(t) dt + \int_0^\infty \bar{Z}(t)^2 Y_A(t) \lambda(t) dt = \int_0^\infty [\{1 - \bar{Z}(t)\}^2 Y_B(t) + \bar{Z}(t)^2 Y_A(t)] \lambda(t) dt,$$

where $\lambda(t)$ is the hazard rate under the null hypothesis, i.e., the common hazard rate.

Under the null hypothesis, a consistent estimator for the cumulative hazard¹² is

$$\int_0^t \frac{dN(t)}{Y(t)} = \int_0^t \frac{dN_A(t) + dN_B(t)}{Y_A(t) + Y_B(t)},$$

so a natural variance estimator is

$$\widehat{\text{var}}_{H_0}(U(0)) = \int_0^\infty \frac{Y_A(t) Y_B(t)}{Y(t)^2} dN(t).$$

Note

$$\frac{Y_A(t)}{n_A} \rightarrow P(X \geq t | Z=0) = S(t)G(t),$$

where $S(t)$ is the common survival function, with $Y_B(t)/n_B$ converging to the same thing. So

$$\frac{Y(t)}{n} = \frac{n_A}{n} \frac{Y_A(t)}{n_A} + \frac{n_B}{n} \frac{Y_B(t)}{n_B} \rightarrow (1-p)S(t)G(t) + pS(t)G(t).$$

Now we can say that

$$\begin{aligned}\frac{1}{n}\widehat{\text{Var}}_{H_0}(U(0)) &= \int_0^\infty \frac{n_A}{n} \frac{Y_A(t)}{n_A} \frac{n_B}{n} \frac{Y_B(t)}{n_B} \frac{n}{Y(t)} \lambda(t) dt \\ &\rightarrow \int_0^\infty (1-p)S(t)G(t)pS(t)G(t)\{S(t)G(t)\}^{-1}\lambda(t)dt \\ &= \int_0^\infty p(1-p)S(t)G(t)\lambda(t)dt\end{aligned}$$

and by the martingale central limit theorem,

$$n^{-1/2}U(0) \rightarrow N(0, \sigma^2),$$

where

$$\sigma^2 = \int_0^\infty p(1-p)S(t)G(t)\lambda(t)dt = \int_0^\infty p(1-p)S(t)G(t)\frac{f(t)}{S(t)}dt = p(1-p)d$$

and $d = \int_0^\infty f(t)G(t)dt$ is the expected event rate under the null.¹³

So we can test the null hypothesis with the statistic

$$\frac{U(0)}{\sqrt{\widehat{\text{Var}}(U(0))}} = \left(\sum_{i=1}^n \int_0^\infty \{Z_i - \bar{Z}(t)\} dN_i(t) \right) / \sqrt{\int_0^\infty \frac{Y_A(t)Y_B(t)}{Y(t)^2} dN(t)},$$

which is asymptotically standard normal under the null hypothesis.

Estimating equation under the alternative

To calculate the contribution of the estimating equation to the sample size, we consider alternatives satisfying the local alternative assumption. Specifically,

$$\lambda_B^{(n)}(t) = \lambda_A(t) + \gamma^{(n)}$$

where

$$\sqrt{n}\gamma^{(n)} \rightarrow \phi.$$

This assumption is common in sample size calculations.¹⁴

Then

$$\frac{1}{\sqrt{n}}U(0) = \int_0^\infty \frac{Y_A(t)dM_B^{(n)}(t) - Y_B(t)dM_A(t)}{\sqrt{n}Y(t)} \quad (4)$$

$$+ \int_0^\infty \frac{Y_A(t) \sqrt{n} \gamma^{(n)} Y_B(t)}{nY(t)} dt, \quad (5)$$

where

$$M_B^{(n)}(t) = \sum_{i=1}^n Z_i \left\{ N_i(t) - \int_0^t Y_i(u) \lambda_B^{(n)}(u) du \right\}.$$

Now (4) is a difference of two martingale integrals and converges in distribution to a $N(0, \sigma^2)$ random variable as a result of the local alternative. To see this, note that the expected value of (4) will still be 0. The variance of the estimating equation under the alternative will be

$$\frac{1}{n} \text{var}_{H_1}(U(0)) \sim \int_0^\infty \frac{p(1-p)S_A(t)S_B^{(n)}(t)G(t) \left\{ (1-p)S_A(t)(1+\gamma^{(n)}) + pS_B^{(n)}(t) \right\} \lambda_A(t)}{\left\{ pS_B^{(n)}(t) + (1-p)S_A(t) \right\}^2} dt, \quad (6)$$

where we use $\sim$ to denote asymptotic equivalence. Details are available in Section 1 of the supplementary material.

For (5),

$$\int_0^\infty \frac{Y_A(t) \sqrt{n} \gamma^{(n)} Y_B(t)}{nY(t)} dt \sim \int_0^\infty \frac{p(1-p)S_A(t)G(t)S_B^{(n)}(t)G(t)}{pS_B^{(n)}(t)G(t) + (1-p)S_A(t)G(t)} \sqrt{n} \gamma^{(n)} dt$$

Using the local alternative assumption and the assumption of the additive hazards model,

$$S_B^{(n)}(t) = S_A(t) e^{-\gamma^{(n)} t} \approx S_A(t) (1 - \gamma^{(n)} t) \rightarrow S(t),$$

where $S(t)$ is the common survival function. Then

$$\frac{1}{n} \text{var}_{H_1}(U(0)) \rightarrow \int_0^\infty p(1-p)S(t)G(t)\lambda(t)dt = p(1-p)d,$$

and (5) converges to

$$\int_0^\infty \frac{p(1-p)S(t)G(t)G(t)S(t)}{pS(t)G(t) + (1-p)S(t)G(t)} \phi dt = \int_0^\infty p(1-p)S(t)G(t)\phi dt.$$

Note that $\int_0^\infty S(t)G(t)dt = E(X)$, the expected observation time for a subject, which is

asymptotically the same for both groups due to the local alternative assumption. So $\frac{1}{\sqrt{n}}U(0)$ converges to a normal random variable with mean $p(1-p)\phi E(X)$ and variance $p(1-p)d$.

Sample size

In order to achieve the desired power, we need that

$$1 - \beta = P\left(U(0) > z_{1-\alpha/2} \sqrt{\text{Var}(U(0))}\right)$$

which leads to

$$n = \frac{(z_{1-\beta} + z_{1-\alpha/2})^2 d}{p(1-p)\gamma^2 E(X)^2}. \quad (7)$$

In Equation (7), we note that n is minimized when $p = 0.5$, meaning we have equal allocation.

In addition to specifying the treatment difference, γ we also need to elicit d and $E(X)$. We follow the example of Schoenfeld¹³ in elicitation of parameters. If we have prior data, an estimate of the survival function and the censoring distribution can be made and used directly in $E(X) = \int_0^\infty S(t)G(t)dt$ and $d = 1 + \int_0^\infty S(t)dG(t)$. Alternatively, it might be possible to directly estimate $E(X)$ and d without estimating the full distributions. For further discussion on elicitation, see Schoenfeld¹³ and Jung et. al.⁷.

It is worth comparing the superiority sample size formula given above, Equation (7), to the sample size formula associated with the log-rank test, as derived by Schoenfeld¹³

$$n = \frac{(z_{1-\beta} + z_{1-\alpha/2})^2}{p(1-p)d \log^2 \Delta}. \quad (8)$$

where Δ is the hazard ratio under the alternative. We note that Equation (7) differs in two important ways.

First, Equation (7) explicitly depends on the observation time while Equation (8) does not. There is a balance in both Equation (7) and Equation (8) between sample size and trial length. In the proportional hazards case, the goal is to accrue events, which can be accomplished with a longer trial, but may be more or less effective depending on the underlying hazard. If the hazard rate is decreasing, then a longer trial will yield a relatively small increase in d , so the increase in power will be relatively small. The reverse holds in the additive hazards sample size formula. In practical terms, this means that the underlying hazard rate must be considered in both cases when deciding whether to have a longer trial or to have a larger sample size, and possibly open more centers. If the hazard rate is decreasing, opening more centers will be more attractive under the proportional hazards model and increasing trial length will be more attractive under the additive hazards model.

The second point is the conclusion that the sample size in Equation (7) decreases as the event probability, d , decreases. This is analogous to the distinction between comparing two proportions by estimating differences or ratios in the proportions. When testing the rate

difference, the required sample size is small for low event rates, while the required sample size for testing the rate ratio is large for low event rates.

Small sample correction

Equation (7) is sensitive to violations of the local alternative assumption. Specifically, when $S_A(t)$ and $S_B(t)$ are assumed to be the same, then $\int S_A(t)G(t)dt$ may significantly underestimate the amount of time on trial leading to excess power. A similar correction should be made for d .

Therefore, we suggest replacing $E(X)$ in (7) with

$$E(\tilde{X}) = \sqrt{E(X_A)E(X_B)}$$

where $E(X_A)$ is the expected observation time for group A and $E(X_B)$ is the expected observation time for group B. Similarly, we suggest replacing d with

$$\tilde{d} = \sqrt{d_A d_B}$$

where d_A is the event rate in group A and d_B is the event rate in group B. Then the updated sample size formula will be

$$n = \frac{(z_{1-\beta} + z_{1-\alpha/2})^2 \tilde{d}}{p(1-p)\gamma^2 E(\tilde{X})^2} \quad (9)$$

This form of correction has been found to work well in simulation studies.

Sample size calculations for non-inferiority trials

For non-inferiority trials, we will make all the same assumptions as we did for superiority trials, except for the local alternative assumption. Due to the assumption of Lin and Ying's model, our hypotheses from the Non-inferiority trials section are reduced to

$$H_0: \lambda_B(t) - \lambda_A(t) \geq \delta$$

$$H_1: \lambda_B(t) - \lambda_A(t) < \delta.$$

For the purposes of power calculation, we will substitute the specific alternative of equivalence between groups for the assumption of a local alternative. That is, power will be calculated under the alternative hypothesis that $\lambda_A(t) = \lambda_B(t)$.

Estimating equation

The estimating equation at the boundary of the null hypothesis will be

$$U(\delta) = \int_0^\infty \frac{Y_A(t)}{Y(t)} dN_B(t) - \int_0^\infty \frac{Y_A(t)}{Y(t)} \delta Y_B(t) dt - \int_0^\infty \frac{Y_B(t)}{Y(t)} dN_A(t).$$

See Section 2 of the supplementary material for details.

Since the two martingales, M_B and M_A , are uncorrelated due to the independence assumption, we can get the variance of the estimating equation at the boundary of the null as

$$\begin{aligned} \text{Var}(U(\delta)) &= \int_0^\infty \{1 - \bar{Z}(t)\}^2 Y_B(t) \lambda_B(t) dt + \int_0^\infty \bar{Z}(t)^2 Y_A(t) \lambda_A(t) dt \\ &= \int_0^\infty \frac{Y_A(t)^2}{Y(t)^2} Y_B(t) \lambda_B(t) dt + \int_0^\infty \frac{Y_B(t)^2}{Y(t)^2} Y_A(t) \lambda_A(t) dt. \end{aligned}$$

Estimate of the variance under the null

In the non-inferiority setting, our null hypothesis is quite different from the superiority setting. Now we have that $\delta = 0$ and $S_A = S_B$.

Following Lin and Ying, natural estimates for the cumulative hazards at the boundary of the null are

$$\hat{\Lambda}_A(t) = \int_0^t \frac{dN(u) - \delta Y_B(u)}{Y(u)} du$$

and

$$\hat{\Lambda}_B(t) = \int_0^t \frac{dN(u) + \delta Y_A(u)}{Y(u)} du.$$

Then the variance of the estimating equation can be estimated by

$$\widehat{\text{Var}}(U(\delta)) = \int_0^\infty \frac{Y_B(t)Y_A(t)}{Y(t)^2} \{dN(t) + \delta(Y_A(t) - Y_B(t))dt\}. \quad (10)$$

Details of the derivation are in Section 3 of the supplementary material.

Sample size

To calculate the sample size, we follow the techniques of Jung et. al.⁷, who examined a proportional hazards non-inferiority hypotheses. First we can write the test statistic as

$$\frac{U(\delta)}{\sigma_n(\delta)} = \frac{U(0)}{\sigma_n(0)} \frac{\sigma_n(0)}{\sigma_n(\delta)} + \frac{U(\delta) - U(0)}{\sigma_n(\delta)}$$

where $\sigma_n^2(\delta) = \widehat{\text{var}}(U(\delta))$ (the quantity given in (10)),

$$U(0) = \int_0^\infty \frac{Y_A(t)}{Y(t)} dN_B(t) - \int_0^\infty \frac{Y_B(t)}{Y(t)} dN_A(t),$$

$$\sigma_n^2(0) = \int_0^\infty \frac{Y_B(t)Y_A(t)}{Y(t)^2} dN(t),$$

and $U(0)/\sigma_n(0)$ is asymptotically standard normal under the alternative hypothesis. This means that the test statistic will be asymptotically normal with a mean and standard deviation we need to calculate.

First we note that

$$U(\delta) - U(0) = - \int_0^\infty \frac{Y_A(t)}{Y(t)} \delta Y_B(t) dt$$

and since, under the alternative hypothesis,

$$\frac{Y_A(t)}{Y(t)} = \frac{n_A}{n} \frac{Y_A(t)}{n_A} \frac{n}{Y(t)} \rightarrow (1-p) \frac{S(t)G(t)}{S(t)G(t)} = 1-p$$

and

$$\frac{1}{n} \int_0^\infty Y_B(t) dt = \frac{n_B}{n} = \int_0^\infty \frac{Y_B(t)}{n_B} dt \rightarrow p \int_0^\infty S(t)G(t) dt,$$

where $S(t)$ is the common survival distribution, which is the same for both groups assuming the alternative hypothesis holds. So, as a whole,

$$\frac{U(\delta) - U(0)}{n} \rightarrow -p(1-p)\delta E(X),$$

where $E(X) = \int_0^\infty S(t)G(t) dt$.

Note that $\sigma_n^2(0)$ is asymptotically equivalent to

$$p(1-p) \int_0^\infty dN(t) = p(1-p)D,$$

where D is the total number of events in the trial under the alternative hypothesis of equivalence.

Recall that $\sigma_n^2(\delta)$ has a similar form to $\sigma_n^2(0)$, so $\sigma_n^2(\delta)$ is asymptotically equivalent to

$$p(1-p)D + p(1-p)\delta \int_0^\infty \{Y_A(t) - Y_B(t)\} dt$$

which is asymptotically equivalent to

$$p(1-p)\{D + n\delta E(X)(1-2p)\}.$$

So the test statistic $U(\delta)/\sigma_n(\delta)$ is asymptotically equivalent to a normal distribution with mean

$$-\frac{p(1-p)n\delta E(X)}{\sqrt{p(1-p)\{D + n\delta E(X)(1-2p)\}}}$$

and variance

$$\frac{D}{D + n\delta E(X)(1-2p)}.$$

The sample size formula needs to satisfy the equation

$$-z_{1-\beta} = \frac{z_{1-\alpha} - \left[-\frac{p(1-p)n\delta E(X)}{\sqrt{p(1-p)\{D + n\delta E(X)(1-2p)\}}} \right]}{\sqrt{\frac{D}{D + n\delta E(X)(1-2p)}}} = \frac{z_{1-\alpha} \sqrt{D + n\delta E(X)(1-2p)}}{\sqrt{D}} + \frac{\sqrt{p(1-p)n\delta E(X)}}{\sqrt{D}}.$$

We will define $d = D/n$ to be the proportion of events. So multiplying both sides by $\sqrt{D}$ and dividing both sides by $\sqrt{n}$,

$$-z_{1-\beta} \sqrt{d} = z_{1-\alpha} \sqrt{d + \delta E(X)(1-2p)} + \sqrt{p(1-p)n\delta E(X)}.$$

The sample size is therefore

$$n = \frac{z_{1-\beta}^2 d + z_{1-\alpha}^2 \{d + \delta E(X)(1-2p)\} + 2z_{1-\alpha} z_{1-\beta} \sqrt{d} \sqrt{d + \delta E(X)(1-2p)}}{p(1-p)\delta^2 E(X)^2}. \quad (11)$$

We do not need to use a small sample correction analogous to the one in the Small sample correction section because no local alternative assumption has been made.

Note that, if $p = 0.5$, then

$$n = \frac{4(z_{1-\alpha} + z_{1-\beta})^2 d}{\delta^2 E(X)^2}.$$

We also note that, if the non-inferiority margin is very large and a larger proportion of subjects are randomized to the experimental group, Equation (11) may not have a real solution. This will be the case if $d + \delta E(X) < 2p\delta E(X)$. Elicitation of the parameters can be done in the same way as for a superiority trial.

The counterpart to the non-inferiority sample size formula given above, Equation (11), for the log-rank test, due to Jung et. al.⁷ is

$$n = \frac{\left\{ \sqrt{\Delta} z_{1-\alpha} + (1 - p + p\Delta) z_{1-\beta} \right\}^2}{p(1-p)d(\Delta-1)^2}, \quad (12)$$

where Δ is the multiplicative non-inferiority margin. Equation (11) differs from Equation (12) in one key respect aside from those mentioned for the superiority sample size formula. When events are undesirable, the optimal allocation for subjects using the log-rank equation slightly favors the control group, while the optimal allocation for subjects using the additive hazards equation slightly favors the experimental group. In most cases, the difference in sample size between equal allocation and the optimal allocation is quite small.

Simulations

To investigate the performance of our sample size calculations, we rely on simulation. We calculate sample size and simulate power and type I error for 27 different scenarios. Weibull distributions with increasing and decreasing hazards are studied (shape parameter 2 and 1/2, respectively), along with a log-logistic distribution with shape parameter 3, which has a hazard rate that first increases and then decreases. Scale parameters for all distributions are set to 1. Censoring is administrative (all subjects are censored at the same time) and the censoring time was chosen so that the event rate would be either 1/4, 1/2, or 3/4. Allocation to the experimental group was one of 1/3, 1/2, or 2/3. Expected observation time is calculated based on these settings.

Superiority

We test the performance of Equation (9) by generating failure times from the experimental group according to the hazard function $\lambda_B(t)$ along with failure times from the control group according to the hazard function $\lambda_B(t) = \lambda_A(t) + \gamma$, where $\gamma > 0$. For these simulations, γ is calculated to provide a total sample size of 1,000, given the levels of z_α , z_β , and d_A , the event rate for the control group. All observations are censored at time C , which is calculated to provide an event rate of d_A for the control group.

We note that, given C and the failure time distribution, the expected observation times for the control and experimental groups can be calculated. We investigate the performance with control group event rates of 25%, 50%, and 75%. Type I error is set at $\alpha = 0.05$ and power at 90%.

Figure 3 shows the simulated power for each scenario. The power is very close to the target power, suggesting that Equation (9) performs well in these scenarios. Results for further

simulations, with scale parameters of 1/2 and 2 for each distribution are available as supplementary material, as are simulations with a sample size of 200.

Non-inferiority

The non-inferiority margin, δ , is set to ensure that the calculated sample size from Equation (11) is $n = 1000$ with type I error at $\alpha = 0.025$ and power targeted at 90%. We note that, for the power simulations, the alternative hypothesis holds, meaning that the two groups have the same failure time distributions. Censoring is the same for all subjects at time C , which is calculated to yield the required event rate, d . The expected observation time, $E(X)$, is calculated given the knowledge of the distribution and d .

Figure 4 gives the simulated power for each scenario. The power appears to be very close to the target of 90% in all cases, with none of the cases deviating by more than two percentage points. Results for further simulations, with scale parameters of 1/2 and 2 for each distribution are available as supplementary material.

Application to SPORTIF III

We now apply our sample size formula to the specifications from the SPORTIF III trial. Sample size in SPORTIF III was calculated assuming exponential survival times. The protocol specified a primary event rate of 3.1% per year for both treatment groups. The hazard rate here is $\lambda = 0.031$, which is used for the exponential sample size calculation.

The specified power was $1 - \beta = 0.9$ with a one-sided type I error of $\alpha = 0.025$. The non-inferiority margin was a difference of $\delta = 0.02$ per year and the average observation time was expected to be $E(X) = 1.5$ years. The probability of an event, assuming that a constant event rate is nearly correct, is about $d = 0.031 * 1.5 = 0.0465$. Allocation was equal, so the probability of being in the experimental group was $p = 0.5$.

Now the sample size using the additive hazards formula is

$$n = \frac{4 \times (z_{1-0.025} + z_{1-0.1})^2 \times 0.0465}{0.02^2 \times 1.5^2} \approx 2172$$

which is the same as the sample size using the exponential assumption, but here the distributional assumptions are relaxed. In general, the sample size using the additive hazards formula will be the same as the sample size using the exponential assumption when allocation is equal between the groups.

A tool to calculate the sample size is available at <http://leemcdaniel.github.io/samplesize.html>. Figure 5 shows a screenshot demonstrating its use.

Discussion

We have developed sample size formulas for superiority and non-inferiority trials with the specific alternative of equality using the additive hazards model. In a non-inferiority trial, allocating slightly more subjects to the experimental group leads to a lower sample size.

These sample size formulas do not depend on parametric assumptions, as is the case for the exponential hazard approach. The mean observation time needs to be specified, but this may be less of a burden than defining the full distribution.

Simulation results demonstrate that our superiority and non-inferiority sample size calculations result in the correct power. Calculations of the sample size using data from the SPORTIF III trial were conducted.

A tool to calculate sample size in a non-inferiority trial under the additive hazards assumption is available on the web at <http://leemcdaniel.github.io/samplesize.html>. The calculator asks for the non-inferiority margin (or hazard difference in the case of a superiority trial), the event probability, and the expected observation time and provides sample sizes for a range of allocation proportions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Cox, DR. Some remarks on the analysis of survival data. Proceedings of the first Seattle symposium in biostatistics: survival analysis; 20 – 21 November 1995; Seattle, United States. New York: Springer-Verlag; p. 1-9.
2. Aalen OO. Linear regression model for the analysis of life times. *Statistics in Medicine*. 1989; 8:907–925. [PubMed: 2678347]
3. Chappell, R. Non-inferiority trials. In: Ravina, B., editor. *Clinical trials in neurology: design, conduct, analysis*. Cambridge: Cambridge University Press; 2012. p. 135-147.
4. Garrett AD. Therapeutic equivalence: fallacies and falsification. *Statistics in Medicine*. 2003; 22:741–762. [PubMed: 12587103]
5. Food and Drug Administration. [2010, accessed 4 May 2015] Guidance for industry non-inferiority clinical trials. <http://www.fda.gov/downloads/Drugs/.../Guidances/UCM202140.pdf>
6. DeMets DL, Friedman L. Some thoughts on challenges for noninferiority study designs. *Drug Information Journal*. 2012; 46(4):420–427.
7. Jung S, Kang SJ, McCall LM, et al. Sample size computation for two-sample noninferiority log-rank test. *Journal of Biopharmaceutical Statistics*. 2005; 15:969–979. [PubMed: 16279355]
8. AstraZeneca. Report. UK: Leicestershire; Efficacy and safety study of the oral direct thrombin inhibitor H 376/95 compared with dose-adjusted warfarin in the prevention of stroke and systemic embolic events in patients with atrial fibrillation (SPORTIF III). [8 March 2000]
9. Olsson SB. Executive Steering Committee of the SPORTIF III Investigators. Stroke prevention with the oral direct thrombin inhibitor ximelagatran compared with warfarin in patients with non-valvular atrial fibrillation (SPORTIF III): randomised controlled trial. *Lancet*. 2003; 362:1691–1698. [PubMed: 14643116]
10. Durrett, R. *Probability: theory and examples*. 4th. Cambridge: Cambridge University Press; 2010. p. 126-135.
11. Lin DY, Ying Z. Semiparametric analysis of the additive risk model. *Biometrika*. 1994; 81(1):61–71.
12. Aalen O. Nonparametric inference for a family of counting processes. *The Annals of Statistics*. 1978; 6(4):701–726.
13. Schoenfeld DA. Sample-size formula for the proportional-hazards regression. *Biometrics*. 1983; 39(2):499–503. [PubMed: 6354290]

14. Chow, SC.; Shao, J.; Wang, H. Sample size calculations in clinical research. 2nd. Boca Raton: Chapman & Hall; 2008. p. 185

Figure 1.
Sample point estimates and 95% CIs for $\mu_B - \mu_A$.

Figure 2.
hazard rates for bleeding in SPORTIF III.

Figure 3.

Simulated power for the scenarios; target is 90% power. The bars around the points are 95% confidence intervals based on the Monte Carlo standard error, with the simulations based on 10,000 replicates. The event rates are for the control group, and were set at 25%, 50%, and 75%.

Figure 4.

Simulated power for the scenarios; target is 90% power. The bars around the points are 95% confidence intervals based on the Monte Carlo standard error, with the simulations based on 10,000 replicates.

Figure 5.

The sample size calculation tool. Parameters can be easily modified using the sliders on the right side of the page.