

On some statistical issues involving clinical trials in Oncology

Mark Rothmann

Presented November 9, 2005 at BASS XII in Savannah

* The views expressed in this talk are those of the author and not necessarily those of the FDA.

Topics

1. Issues to consider when constructing a non-inferiority analysis
2. On surrogate comparisons

Issues to consider when constructing a non-inferiority analysis

Preface

- Focus on two-arm active-control trials
- Emphasis on the speaker's experience with Oncology clinical trials
- Discuss some issues and steps involved with the design and analysis of non-inferiority trials

Purpose of a Non-inferiority comparison

There can be different reasons for a non-inferiority comparison

- It is unethical to do a placebo-controlled trial and the standard requirement is the demonstration of efficacy against a placebo.
- It is unethical to do a placebo-controlled trial and it is required to demonstrate that the test therapy has efficacy greater than some minimal amount.
- It is desired to demonstrate that the test therapy is efficacious and either an alternative to a standard therapy or it is better than a standard therapy.

Issues

There are many issues to consider when constructing a non-inferiority analysis. These issues include:

1. how to define the active-control effect;
2. how to estimate or model the active-control effect for the current trial;
3. the reproducibility of the active-control effect size with respect to the current active-controlled trial conditions;
4. how to design a trial consistent with a non-inferiority inference;
5. the hypothesis of interest;
6. whether the design of the active-control trial is retrospective or prospective with respect to the estimation of the active-control effect; and
7. the interpretation of the results.

Defining the Active-Control effect

If the aim is that drug/drug combination B is an alternative to drug/drug combination A, the efficacy of B may need to satisfy certain requirements:

1. It may be necessary (but perhaps not sufficient) that B must be superior (or concluded superior) to every drug, drug combination and regimen for which A is superior.
2. When each component of a drug combination B is regarded as “active,” it may also be necessary for such a drug combination to have more efficacy than any subset of the drugs in the drug combination.

For a non-active add-on (toxicity reducing add-on), this may not be necessary.

Modeling the active-control effect size

- Are there multiple historical trials that show consistent effects?
Care should be taken when considering the results from similarly designed trials and those trials which are not so similarly designed.
- Has the effect changed? Should the historical effect size estimator be “reduced” by some fraction to estimate the active-control effect size for the current active-control
- In settings where there has been some between trial variability in the effects of therapy, that between trial variability needs to be considered.

Design features should be consistent with a non-inferiority comparison

- For a mortality superiority trial it may be allowable for patients on the experimental arm to later receive the control therapy whereas patients on the control arm are prevented from receiving the experimental therapy.

This would compare a scenario where the experimental therapy would be approved for the indication with a scenario where the experimental therapy would not be approved (allowed) for the indication. For a superiority mortality trial, this is what is of interest to patients.

Design features should be consistent with a non-inferiority comparison

This feature, however, does not make sense for a non-inferiority trial where it is desired to show that the experimental treatment is not much worse than the active-control.

The active-control therapy is clearly not much worse than itself. Therefore, allowing the control therapy to be available to the patients in the experimental arm obscures a non-inferiority comparison.

A placebo (as the experimental therapy) could easily demonstrate non-inferiority to a beneficial control therapy, if patients on that arm were allowed to get the control therapy early enough in the trial.

Notation

T - Experimental Treatment

C - Active-Control

P - Placebo or other reference therapy

HR - Hazard Ratio

Hypotheses for a fixed difference

Example:

$$H_0: \theta_T - \theta_C = d_0 \quad \text{vs.} \quad H_a: \theta_T - \theta_C > d_0$$

where d_0 is some constant.

Definition of the Proportion of Effect Retained, δ

For “difference” measures (e.g., means)

$$\delta = \frac{(\theta_T - \theta_P)}{(\theta_C - \theta_P)} = \frac{(\theta_C - \theta_P) - (\theta_C - \theta_T)}{(\theta_C - \theta_P)}, \text{ provided } \theta_C - \theta_P > 0.$$

For relative measures (e.g., log-hazard ratio)

$$\delta = \frac{\theta_{P/T}}{\theta_{P/C}} = \frac{\theta_{P/C} - \theta_{T/C}}{\theta_{P/C}}, \text{ provided } \theta_{P/C} > 0.$$

The order of subscript for both definitions may need to be reversed depending on whether we are measuring positive or negative outcomes.

Hypotheses for a retention of pre-specified fraction of the active-control effect

We will assume that $\theta_C - \theta_P$ or $\theta_{P/C} > 0$. When testing whether the treatment maintains $100\delta_0\%$ of the effect of the active control, the hypotheses are:

$$H_0: \delta = \delta_0 \text{ vs. } H_a: \delta > \delta_0$$

For log-hazard ratios this translates to

$$H_0: \log \text{HR}(T/C) - (1-\delta_0)\log \text{HR}(P/C) = 0 \text{ vs.}$$

$$H_a: \log \text{HR}(T/C) - (1-\delta_0)\log \text{HR}(P/C) < 0$$

Comparisons to the placebo or reference therapy

- For some indications just being better than placebo, when there is a standard of care that is better placebo, may not be enough. As Dr. Temple* pointed out for pneumonia clinical trials for some indications those reasons that make it unethical to do a placebo controlled trial, are the same reasons that attribute to the unwillingness to have a test therapy lose too much effectiveness.

US Food and Drug Administration Division of Anti-Infective Drug Products Advisory Committee meeting transcript. Feb. 19-20, 2002 Available at:
<http://www.fda.gov/ohrms/dockets/ac/cder02.htm#Anti-Infective>

Comparisons to the placebo or reference therapy

- Also, “It does not seem prudent to regard a test therapy as approvable or effective, if the true test therapy vs. placebo survival hazard ratio is 0.99, particularly in light of an estimate of the standard therapy vs. placebo survival hazard ratio of 0.50, if such an estimate has reasonable precision.”
 - Rothmann, M. D. (2004) “Author’s reply” *Statistics in Medicine* vol. 23 no. 17 2774-2778

The Test procedure

There are several factors to consider when determining an appropriate test procedure including:

- the desired level of significance, uncertainty or type I error rate for the procedure;
- whether the design of the active-controlled trial was independent of any modeling of the active-control effect.
- whether to incorporate the uncertainty that the active-control is effective for the current trial (is essence an alpha adjustment);
- and whether to include more than one criteria that the experimental drug is effective.

Independent Design vs. Dependent Design

When testing the equality of means from two different populations, independent random samples are drawn from the two populations and a test is performed – for example, a t-test or a large sample normal test. For such a test procedure, there is no difference between simultaneously drawing the two independent samples or drawing the second sample after the first sample, if the sample size or design of the second sample is independent of the results from the first sample. No adjustment is needed to the type I error probabilities in such situations.

Independent Design vs. Dependent Design

However, when the sample size or design for the second sample is dependent on the results of the first sample, the type I error probability is altered for each specific possibility in the null hypothesis.

Dependent Design Case - Set up

Suppose $X \sim \text{Normal}(\mu_X, \sigma_X^2)$ where σ_X can be pre-set based on the value of Y which is distributed as $\text{Normal}(\mu_Y, \sigma_Y^2)$ where σ_Y is fixed.

The hypotheses of interest are:

$$H_0: \mu_X = \mu_Y \text{ vs. } H_1: \mu_X < \mu_Y.$$

In the non-inferiority setting involving hazard ratios we may have, $\mu_X = \log \text{HR}(T/C)$ and $\mu_Y = (1 - \delta_0) \times \log \text{HR}(P/C) + d_0$.

Test procedure

Consider the test that rejects H_0

$$\text{whenever } \frac{X - Y}{\sqrt{\sigma_X^2 + \sigma_Y^2}} < k$$

where σ_X is set after the value of Y is observed ($\sigma_X = \sigma(y)$, when $Y = y$).

The frequentist test is altered.

The Bayesian test is unaltered, but the frequentist interpretation changes.

Rothmann (2005) “Type I error probabilities based on design-stage strategies with applications to non-inferiority trials” *Journal of Biopharmaceutical Statistics* vol.15 no. 1 109-127.

Factors that influence the type I error rate of a test procedure

- When the design of the active-controlled trial is independent of the estimation of the active control effect, the type I error probability depends on the
 - test procedure and
 - how well the active-control effect was modeled.
- When the design of the active-controlled is dependent of the estimation of the active control effect, the type I error probability depends on the
 - the design strategy (how the design is dependent on the estimate of the active control effect)
 - the specific point in the null hypothesis
 - test procedure and
 - how well the active-control effect was modeled

Incorporating other relevant comparisons into the analysis

- The uncertainty that the active-control is effective can also be incorporated by making a respective adjustment to the significance level of the non-inferiority test or the event that the active-control is effective can be included when calculating a posterior probability that the experimental treatment is effective (Simon (1999) Bayesian design and analysis of active control clinical trials. *Biometrics* 55:484–487).
- Other possibilities of effectiveness can be included in the hypotheses and the test procedure that do not depend on the active-control being effective.

Posterior probability = the probability that the experimental treatment is non-inferior to an effective active-control or that the experimental treatment is better than both the placebo and the active-control.

Interpretation of the results

- Formally, since the active-controlled trial does not have a placebo or reference therapy arm, any inference that uses information outside of the trial is an extrapolated inference for that trial. Extrapolation can be very risky.

Formally exact cause-and-effect conclusions can not be drawn from cross trials comparisons.

Other issues involving non-inferiority trials

- The powering of studies for non-inferiority (selecting the alternative) or the “non-inferiority myth” and
- the meaning/lack of meaning of reproducibility
- Performing a non-inferiority analysis on a surrogate endpoint

A pair of non-inferiority myths

The following statements are false

1. A smaller sample size is needed for a superiority trial than for a non-inferiority trial
2. It is easier to demonstrate non-inferiority to a more efficacious therapy than to a less efficacious therapy

Non-inferiority myth: to do a non-inferiority trial or a superiority trial

- Less needs to be statistically ruled out in a non-inferiority comparison than for a superiority comparison.
- Why does lowering the bar for the demonstration of efficacy increase the overall sample size needed for a clinical trial? It doesn't.
- As Dr. Fleming* points out, non-inferiority trials with “scientifically rigorous margins” need not require very large sample sizes. Non-inferiority trials can be powered at alternatives where the test therapy is a little better than the standard therapy.
- Whether the decision is to design a superiority trial or a non-inferiority trial for a particular test therapy, the trial should be powered for a singular alternative. Sample size (event size) calculations should be compared at the same alternative.”

*US Food and Drug Administration Division of Anti-Infective Drug Products Advisory Committee meeting transcript. Feb. 19-20, 2002 Available at:
<http://www.fda.gov/ohrms/dockets/ac/cder02.htm#Anti-Infective>

“Non-inferiority myths”

- Suppose that a sponsor has two choices, C1 and C2 for the active-comparator for a trial and that C2 is better than C1 ($C2 > C1$).
- It is easier (more probable or requires a smaller size) to demonstrate that T is superior to C1 than to demonstrate that T is superior to C2.
- It is also easier to demonstrate that T is non-inferior to C1 than to demonstrate that T is non-inferior to C2.
- A non-inferiority trial of T vs. C1 powered at $T=C1$ will require a larger sample size than a non-inferiority trial of T vs. C2. However, $T=C1$ and $T=C2$ are different alternatives.

The meaning/lack of meaning of reproducibility; Three arm studies

- What does reproducibility mean for a non-inferiority inference?

the reproducibility of $(T-C) - (P-C)$,

OR

the reproducibility of $(T-C)$ and the reproducibility of $(P-C)$

To Summarize

- There are many issues to consider when designing a non-inferiority trial – including the intended purpose of the trial.
- Make sure that the design and conduct of the trial is consistent with the intended purpose of the trial
- Wisely select an appropriate alternative to power the study

On Surrogate Comparisons

Outline

1. Introductory slides
2. The possible relationships between a comparison of a potential surrogate endpoint and a clinical benefit endpoint
3. A linear regression approach with an application in a metastatic disease setting relating PFS and OS comparisons
4. Predictability of an interim analysis of OS on the final analysis of OS.

Notation

Clinical benefit endpoint (CBE): Overall Survival

Potential surrogate endpoint (PSE): PFS

Types of surrogate comparisons

- Surrogate for regular approval – The aim is to **conclude** clinical benefit based on a comparison on a surrogate endpoint
- Surrogate for accelerated approval - The aim is to **predict** (“*reasonably likely to predict*”) clinical benefit based on a comparison on a surrogate endpoint

Considerations

Meaning of a surrogate comparison

1. What is the definition of a surrogate comparison?/With respect to what concept is one comparison a potential surrogate for another comparison?

Regulatory Objective for a CBE per trial

- When [the distribution of] CBE is equal between arms, there is a 2.5% chance of claiming that the treatment arm is better than the control arm with respect to the CBE.

Primary property for a surrogate comparison

- When [the distribution of] CBE is equal between arms, there is a 2.5% chance of claiming that the treatment arm is better than the control arm with respect to the CBE.

Considerations

Add-on trials of anti-cancer drugs vs. Substitution trials

2. The need to differentiate between add-on trials and substitution trials
 - Add-on trial: clear respective ordering of arms (A+B/A)
 - Substitution/Replacement trials: respective ordering of arms is arbitrary (A/B or B/A). Criterion of evaluating the relationship in comparisons of the PSE (PFS) to comparisons of the CBE (OS) needs to be “symmetric” (it should not matter whether the results of an arm are given as A/B or B/A).

A perfect surrogate endpoint for a superiority comparison

- The surrogate endpoint captures all of the influences on the clinical benefit endpoint:

Given any fixed outcome for the surrogate endpoint, the conditional distribution for the clinical benefit endpoint is identical across treatment arms (independence between the treatment arms and conditional distribution of the clinical benefit endpoint).

A perfect surrogate endpoint

For example, if PFS were a perfect surrogate for OS, the distributions of overall survival for those patients that had PFS = 10 would be identical across treatment arms.

Use of a perfect surrogate endpoint for a one-sided significance level of 0.025 will commit a type I error for a conclusion (prediction) on the clinical benefit endpoint 2.5% of the time

Other possible relationships between a potential SE and a CBE

Consider an add-on trial of an anti-cancer agent A vs. A + B.

1) Given any fixed value for the PSE (PFS), the conditional distribution for the CBE (OS) is better for arm A patients

Use of this surrogate endpoint at a one-sided 0.025 significance level will lead to type I errors in the conclusions (predictions) on the clinical benefit endpoint over 2.5% of the time.

Type I error rate for Scenario 1

An add-on trial of an anti-cancer agent A vs. A + B.

Suppose that the null hypothesis of equal CBE is true and scenario 1 holds then

When CBE has “ $A = A+B$ ” we have SE has “ $A < A+B$ ”

As the amount of information (number of PFS events) for comparing the SE increases, the power for concluding that “ $A < A+B$ ” for the SE increases and thus, the chance of falsely concluding/predicting that “ $A < A+B$ ” for the CBE increases.

Comment about an example to be discussed later

- Seven of eight trials of a particular metastatic disease that were studied (to be discussed later) empirically fell into scenario 1. That is, for fixed values for the of PFS, the conditional distribution for the OS is better for arm A patients than for arm A+B patients

Other possible relationships between a potential SE and a CBE

Consider an add-on trial of an anti-cancer agent A vs. A + B.

2) Given any fixed value for the potential SE (PFS), the conditional distribution for the CBE (OS) is better for arm A + B patients

Use of this surrogate endpoint at a one-sided 0.025 significance level will lead to type I errors in the conclusions (predictions) on the clinical benefit endpoint less than 2.5% of the time.

Type I error rate for Scenario 2

An add-on trial of an anti-cancer agent A vs. A + B.

Suppose that the null hypothesis of equal CBE is true and scenario 2 holds then

When CBE has “ $A = A+B$ ” we have SE has “ $A > A+B$ ”

As the amount of information (number of PFS events) for comparing the SE increases, the power for concluding that “ $A < A+B$ ” for the SE decreases and thus, the chance of falsely concluding/predicting that “ $A < A+B$ ” for the CBE decreases.

Usefulness of a scenario 2 PSE

- These PSE can be useful if data on the PSE matures more quickly than data on a CBE
- A positive conclusion on the PSE implies a positive conclusion on the CBE

Hypothetical known situation

Model: For each possible value, x , for the estimated difference in the PSE there is some modeled uncertainty, $F(0|x)$, that CBE favors the control arm ($F(-\delta|x)$ for NI comparisons). F based on empirical data, simulations and models.

Current trial: Based on the results for the current trial for PSE comparison, can determine a density, g , that summarizes the uncertainty in the PSE comparison ($g(y)=dG(y)/dy$, where $G(y)$ is the confidence coefficient that goes with $(-\infty,y)$).

Integration of the Model and the Results of the current trial

- The uncertainty (one-sided p-value/posterior probability) that CBE is better for the treatment arm is given by

$$\int_{-\infty}^{\infty} F(0 | x)g(x)dx \quad (\text{the posterior probability that CBE is better for the control arm})$$

A 95% uncertainty interval for the CBE difference can be found by solving

$$\int_{-\infty}^{\infty} F(l | x)g(x)dx = 0.025 \quad \text{and} \quad \int_{-\infty}^{\infty} F(u | x)g(x)dx = 0.975$$

Difficulties

- Frequentist: Do not have historical pairs of parameter values for (PSE,CBE).
- Does this matter? Previous studies will tell us whether certain pairs of parameter values can go together.
- In any case determining $F(|x)$ is quite a task.

What is done in practice

- Plot the pairs of the estimates for the PSE and CBE differences, and make an inference (or fail to make an inference) on the relationship. For example, a regression line may be determined.
- So we are using joint estimates to make inferences on the universal behavior of joint estimates. But did any of these pairs of estimates come from the regression-line?

What is done in a meta-analysis

- Typically in a meta-analysis, each study has an inference about a study parameter and the meta-analysis makes an inference about a parameter (common study parameter value or the average parameter value). We can also test for heterogeneity.
- This is a meta-analysis. Considerations, tests, and models used for a meta-analysis can be applied here.

Can Do

- For each individual study determine the joint relationship for the estimates of the differences in PSE and CBE

Relating a PFS comparison with the OS comparison

- For each of eight add-on studies of a particular metastatic disease, determine the study-specific regression line

$$y = \alpha + \beta x \text{ where}$$

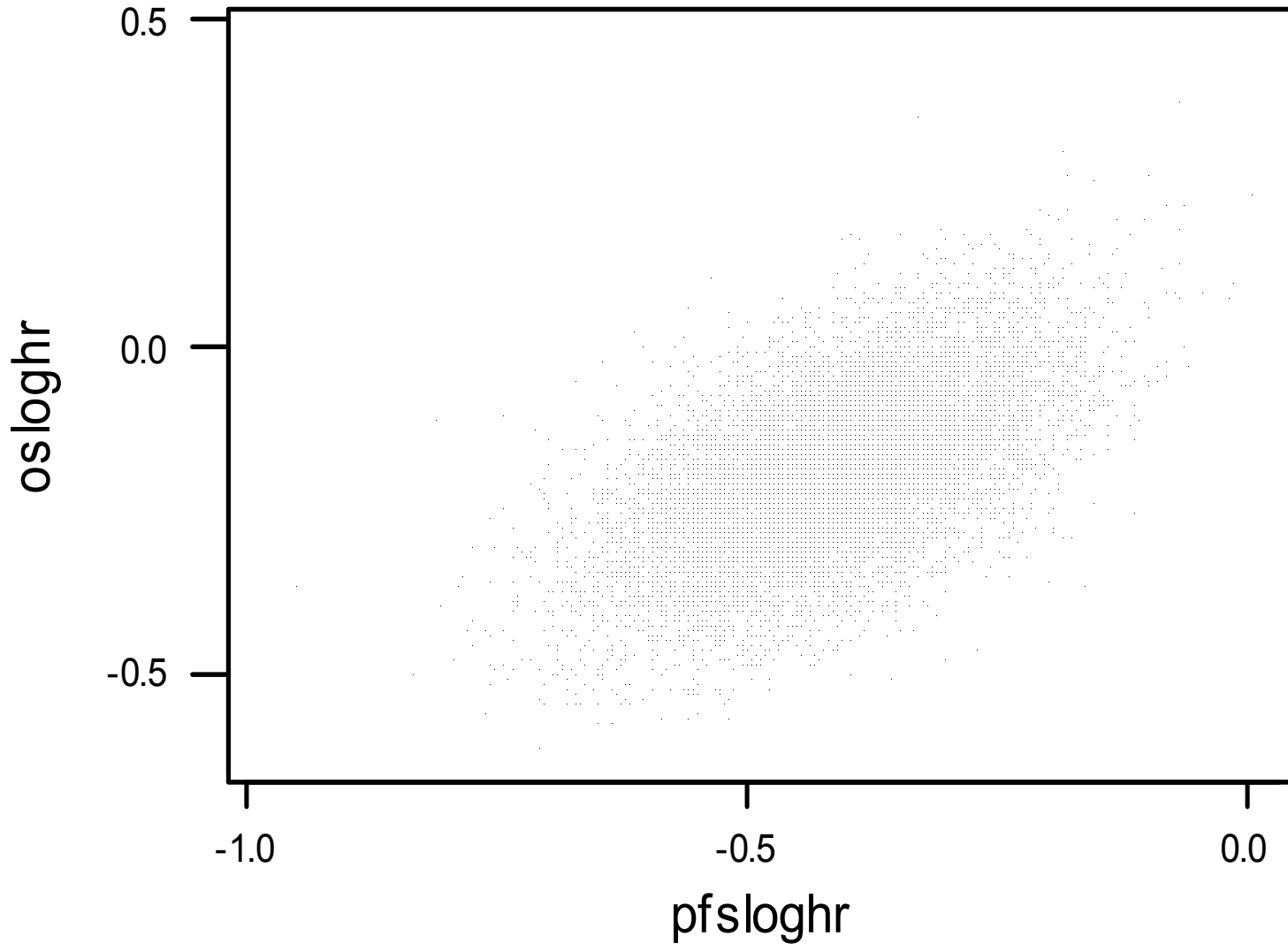
x: OS logHR(A+B/B) estimate

y: PFS logHR(A+B/B) estimate


Bootstrap

- (i) This is done via many simulations where a single simulation bootstraps from each arm a sample of the original sizes and then compute the PFS log HR and the OS log HR.
- (ii) The standard errors for the estimators of α and β can be determined by many simulations where for one simulation first bootstrap from each arm a sample of the original sizes. Use these samples as the original observed samples. Then do (i) to come up with estimates of α and β .

An example



Sample Distributions

- Sample Distributions were all approximately bivariate normal (calculated probabilities  relative frequencies)
- Calculated correlation coefficients (increasing order):

0.47, 0.49, 0.50, 0.50, 0.55, 0.59, 0.59, 0.67

Sample Distributions

- Ordered values for the slopes (rounded to the nearest 0.05):

0.45, 0.55, 0.55, 0.65, 0.65, 0.65, 0.75, 0.80

With standard errors ranging from 0.052 to 0.105

Sample Distributions

- Ordered values for the intercepts (rounded to the nearest 0.05):

-0.15, -0.10, -0.05, -0.05, 0.10, 0.10, 0.35,
0.35

With standard errors ranging from 0.089 to
0.314

Plot of regression lines

Test for Heterogeneity

- Test of slopes – p-value = 0.0002
- Test of intercepts – p-value = 0.035
(p-value of 0.91 that similar comparison have the same intercept)

Relating an early analysis of OS to the final analysis of OS

r_1 = number of events at the interim analysis

r_2 = number of events at the final analysis

L_1 = interim analysis log hazard ratio estimator

L_2 = final analysis log hazard ratio estimator

Correlation between interim and final analyses of OS

Under the hypothesis of equal survival distributions, we have that the correlation of L_1 and L_2 is

$$\sqrt{\frac{r_1}{r_2}}$$

- For $r_1 = 250$ and $r_2 = 1000$, the correlation is 0.5
- For $r_1 = 447$ and $r_2 = 1000$, the correlation is 0.67

Regression line between interim and final analyses of OS

Under the assumption of proportional hazards and for a 1:1 randomization, we have that

(L_1, L_2) has an approximate bivariate normal distribution with means (θ, θ) , variances approximately $(4/r_1, 4/r_2)$ and correlation (in practice) $\otimes (r_1/r_2)^{0.5}$. Thus,

$E(L_2 | L_1) \otimes (1-\beta)\theta + \beta L_1$ where $\beta = r_1/r_2$.

Under the hypothesis that $\theta = 0$ and $E(L_2 | L_1) \otimes \beta L_1$.

Summary on predicting the final OS comparison

- For the disease setting studied, in general early analyses on OS would do better than a final analysis of PFS in predicting the final OS comparison.

Study Design example for drawing conclusions on the OS comparison for an add-on trial

Do an analysis on PFS at the one-sided 0.005 level. If significant conclude that OS is better for the A+B arm

Also conclude OS is better for the A+B arm, if the final (later) analysis on OS is significant at the one-sided 0.02 level.

If scenario 1 clearly holds this will inflate the type 1 error rate for the OS comparison.

Study Design example for accelerated approval for an add-on trial

Do an analysis on PFS at the one-sided 0.005 level. If significant *predict* that OS is better for the A+B arm

Also conclude OS is better for the A+B arm, if the final (later) analysis on OS is significant at the one-sided 0.02 level.

- How does one incorporate *predicting* an alternative hypothesis into an (type 1) error calculation.

Closing remarks

- Caution should be taken when considering a surrogate comparison
- The predictability of a PFS comparison on an OS comparison should be studied/understood
- Early analyses on OS are more predictive of the final OS comparison than is a final analysis on PFS