2-11-2010

Bio-Creep in Non-Inferiority Clinical Trials

Siobhan P. Everson-Stewart
University of Washington - Seattle Campus, spes@uw.edu

Scott S. Emerson
University of Washington, semerson@u.washington.edu

Bio-Creep in Non-Inferiority Clinical Trials

Siobhan Everson-Stewart[1], Scott S. Emerson

Department of Biostatistics, Box 357232, University of Washington, Seattle, Washington, 98195, U.S.A.

Key Words: constancy assumption

Abstract

After a non-inferiority clinical trial, a new therapy may be accepted as effective, even if its treatment effect is slightly smaller than the current standard. It is therefore possible that, after a series of trials where the new therapy is slightly worse than the preceding drugs, an ineffective or harmful therapy might be incorrectly declared efficacious; this is known as "bio-creep." Several factors may influence the rate at which bio-creep occurs, including the distribution of the effects of the new agents being tested and how that changes over time, the choice of active comparator, the method used to model the variability of the estimate of the effect of the active comparator, and changes in the effect of the active comparator from one trial to the next (violations of the constancy assumption). We performed a simulation study to examine which of these factors might lead to bio-creep and found that bio-creep was rare, except when the constancy assumption was violated.

1 Introduction

When a therapy exists that has been proven to reduce the rate of mortality or major morbidity for a given condition, it is generally considered unethical to withhold this treatment from subjects as would happen in a placebo-controlled clinical trial. In these settings, then, investigational treatments are frequently tested against an active comparator. For registrational purposes, these new therapies are often not required to be more efficacious than other treatments on the market; they must merely be shown to have a positive treatment effect. Especially if an investigational treatment has advantages over the standard therapy, such as an easier mechanism of

[1]Corresponding author:
Siobhan Everson-Stewart
Department of Biostatistics, Box 357232
University of Washington
Seattle, WA 98195-7232

e-mail: spes@uw.edu
Tel: +1-206-616-0463
Fax: +1-206-543-0131

This research was supported in part by the NHLBI through training grant T32 HL007183 to the University of Washington.

delivery or an improved safety profile, some reduction in efficacy may be clinically acceptable. A clinical trial designed to demonstrate that the difference in the treatment effects of the investigational therapy and active comparator is within such an acceptable margin is called a non-inferiority clinical trial. However, this raises the concern that if each new therapy is perhaps slightly worse than the preceding drugs, after a series of non-inferiority trials, an ineffective or harmful therapy may falsely be deemed efficacious, a phenomenon known as "bio-creep" [1, 2].

Bio-creep has previously been mentioned as a theoretical possibility, but little exploration has been done to discover if and when it might occur in practice. In order to address these issues, we designed a simulation study to investigate what factors contribute to the occurrence of bio-creep, and to quantify how frequently it may be happening. There are several factors we believed may influence how often bio-creep occurs, including the true efficacy of the new therapies being tested, how the effect of a single drug changes from trial to trial, and the characteristics of the trials themselves. These traits can be loosely grouped into those describing the clinical setting in which the trial is being performed, such as the distribution of the effects of new therapies, and those related to trial methodology. In addition to exploring when bio-creep occurs, we also strove to describe how the situations where bio-creep occurred differed from those where it did not.

2 Defining Non-Inferiority

The definition of non-inferiority in any trial has two major components: the non-inferiority margin and the statistic being used. The non-inferiority margin can be thought of as the amount by which the true effect of the new therapy is allowed to be worse than that of the active control: when we can, with a reasonable degree of certainty, establish that the effect of the new therapy is not worse than this threshold, we consider it to be acceptably effective. This quantity is normally specified in the parameter space (i.e., in terms of the true effect of the treatment relative to the active control). This, along with information from the historical trials, intrinsically defines a corresponding margin in the sample space (i.e. the observed effect that would be declared as acceptably non-inferior). The choice of method used to combine the information from historical trials of the active comparator with that from the new trial is not as clinically interpretable as is the margin, but still importantly influences many of the operational characteristics of a trial.

2.1 Choosing the Non-Inferiority Margin

One of the greatest challenges of planning a non-inferiority clinical trial is selecting the non-inferiority margin. The ICH E-9 states that the non-inferiority margin should be the "largest difference that can be judged to be clinically acceptable, and should be smaller than the differences observed in superiority trials of the active comparator [3]." When setting the margin, one must consider the clinical significance of the specified decrease in efficacy, as well as what is known of the effect of active comparator. It is important that the non-inferiority margin not exceed the effect

of the active comparator, or else a treatment known to be ineffective would be considered "non-inferior" by definition.

This margin may be set on either an absolute or a relative scale. In the first case, one may demand that the effect of the new therapy as compared to placebo exceed some threshold (e.g. hazard ratio on placebo compared to experimental therapy must exceed 1.1), often chosen to be the minimally clinically important difference. In the second, the margin is defined relative to the efficacy of the active control; the experimental treatment must retain some percentage of the effect of the active comparator. As long as the active comparator retains some effect, using a relative margin ensures that an approved therapy must deliver some benefit. It is possible, however, for a new therapy to preserve the pre-specified percentage while falling below an absolute threshold defining the minimally clinically important difference.

Selecting a relative non-inferiority margin can be viewed not only as a way of ensuring that the efficacy of the investigational agent is acceptably close to that of the active control, but also as a tool to "discount" the effect of the active control in the current trial relative to its historical effect [4]. In their 2003 article, Wang and Hung express such a sentiment: "In order to be fairly certain that the new drug would have been superior to placebo had the placebo treatment been studied in the trial, it was decided that the new drug must be shown to preserve at least 50% of the control effect in this target population of the active-controlled trial [5]."

One can also use concerns about patient safety to argue for demanding that a certain percentage of the effect of the active control is retained. While, strictly speaking, efficacy requirements demand only that a new treatment be more effective than placebo, safety must also be considered. The full safety profile of a new therapy is seldom known at the time it is approved, and those receiving such treatment during a clinical trial are being exposed to unknown risks. In light of this uncertainty, it seems prudent to demand that a new treatment not be meaningfully worse than an existing treatment, as it would not be ethical to expose patients to a new treatment with an unknown safety profile if it is markedly worse than their existing choices. Demanding 50% retention of the effect of the active comparator, with the hope of actually preserving this fraction of the efficacy, is one way of ensuring that is not the case. Additionally, doing so also allows one to infer efficacy of the new treatment as compared to placebo, even if the effect of the active control is markedly reduced from what was observed historically [6, 7, 8].

2.2 Incorporating Uncertainty

It is important to incorporate the uncertainty associated with estimating the effect of the active comparator when selecting the statistical methods to be used, especially if the non-inferiority margin is defined as a proportion of the effect of that therapy estimated from historical data. Many available methods do take this increased variability into account, though their operating characteristics vary widely [9, 10]. For the purpose of this study, we used a more direct approach based on a putative placebo, sometimes known as a "synthesis" method [10, 11].

Another commonly used method is the 95-95 rule [12]. In this case, the non-inferiority margin is set at the lower bound of a 95% confidence interval for the

efficacy of the active control (as compared to placebo). As long as the 95% confidence interval for the effect of the experimental therapy (as compared to the active control) lies above this margin, the experimental therapy is declared non-inferior. If one wants to retain a fraction of the effect of the active comparator, say $X * 100\%$, then the margin is set at X times the lower bound of the confidence interval for the effect of the active comparator. This method is known to be conservative when the true effect of the active control is constant across trials [10].

2.3 Notation

If one believes that historical trials of the active comparator versus a placebo accurately estimate the expected efficacy of the active comparator in the new trial, it is very straightforward to estimate the efficacy of the experimental agent transitively (this is the aforementioned synthesis method). The estimate would be based on the information we have on three groups of subjects: the placebo group, denoted P; the group receiving the active comparator, or standard therapy, S; and those receiving the new treatment, N. In a proportional hazards model of a time-to-event analysis, we have $\lambda_P(t) = \lambda_S(t)\exp(\theta_{SP})$. Under this model, θ_{SP} is the log hazard ratio of the placebo group compared to the active treatment group, and we assume without loss of generality that $\theta_{SP} > 0$ implies that the standard therapy is efficacious. Although θ_{NP} is not directly estimable from a non-inferiority trial, combining the historical data with the data from the new trial gives us $\hat{\theta}_{NP} = \hat{\theta}_{NS} + \hat{\theta}_{SP}$. Inference about the efficacy of the new agent could then be based directly on $\hat{\theta}_{NP}$.

In the case where we suspect that the historical trial may not be directly relevant, we may be able to guard against a loss of efficacy by requiring that the investigational treatment retain a set percentage of the effect of the active control. This provides protection against a diminished effect of the active comparator, as long as that treatment is neither ineffective nor harmful in the subpopulation enrolled in the trial. For these simulations, we took this approach, with an arbitrary 50% retention threshold. In practice, this threshold should additionally take into account both the clinical setting and considerations of patient safety.

Our definition of non-inferiority is equivalent to testing $\theta_{NP} > \frac{1}{2}\theta_{SP}$, where θ_{NP} is the log hazard ratio of the experimental treatment group compared to the placebo group. We want to show:

$$\theta_{NP} = \theta_{NS} + \theta_{SP} > \frac{1}{2}\theta_{SP}$$

or, equivalently,

$$\tau \equiv \theta_{NS} + \frac{1}{2}\theta_{SP} > 0.$$

If a 95% confidence interval for τ lies entirely above 0, then non-inferiority can be concluded with a nominal one-sided 0.025 Type I error rate. Since $Var(\tau) = \sigma^2_{NS} + \frac{1}{4}\sigma^2_{SP}$, this confidence interval can be easily constructed using estimates from historical trials of the active comparator as well as the data from the current trial.

2.3.1 Example

As an example, consider the case of second-line chemotherapy in non-small-cell lung cancer. Docetaxel was established as effective in a placebo-controlled clinical trial [13]. Later, when pemetrexed was tested in this same indication, a non-inferiority trial was performed comparing pemetrexed to docetaxel [14]. From the earlier trial, we have $\exp\left(\hat{\theta}_{SP}\right) = 1.78$, with a 95% confidence interval of 1.14 to 2.86 [14]. Considering this historical data, what non-inferiority margin should be used? If we can establish that $\exp(\theta_{NS}) > \frac{1}{1.78} = 0.56$, it suggests that pemetrexed is more effective than placebo. However, the variability in the estimate of θ_{SP} must be taken into consideration. Additionally, we may want to demand that pemetrexed retain some portion of the effect of docetaxel before declaring non-inferiority. If we choose to use a synthesis method to account for the variability of $\hat{\theta}_{SP}$ while demanding 50% retention, we want to be confident that

$$\tau \equiv \theta_{NS} + \frac{1}{2}\theta_{SP} > 0.$$

We obtain $\hat{\theta}_{SP} = \log(1.78)$ and $\hat{\sigma}_{SP} = 0.23$ from the historical data. Since the trial comparing docetaxel to pemetrexed was designed to stop when $n = 400$ events had been observed, we know that $\sigma^2_{NS} \approx \frac{4}{400} = 0.01$. We can then calculate that pemetrexed will be declared non-inferior to docetaxel when $exp\left(\hat{\theta}_{NS}\right) \geq 1.01$, with a corresponding confidence interval of $(0.83, 1.23)$. In this setting, using a synthesis method with 50% retention implies a margin of 0.83 on the $\exp(\theta_{NS})$ scale.

Alternatively, a 95-95 rule could be utilized. If the aim of the study is to determine that pemetrexed would have been show to be superior to placebo had one been included, then the margin would be $\frac{1}{1.14} = 0.88$. We can also incorporate a demand for 50% retention while using this 95-95 method, yielding a margin of $\sqrt{(0.88)} = 0.94$ for $\exp(\theta_{NS})$. With the trial designed to stop when 400 events had been observed, this lower bound would only be exceeded when $\exp\left(\hat{\theta}_{NS}\right) > 1.14$.

3 Methods: A Simulation Study

We hoped to identify which factors can lead to increases in the rate of bio-creep.

1. We began by considering how bio-creep is affected by the distribution from which the effects of new agents are drawn. Naturally, if all agents tested in trials are effective, bio-creep will never occur.

2. The changes in the effect of a single therapy across trials are also of great importance. Most non-inferiority methodology assumes that the treatment of any one drug is constant across trials; this is known as the "constancy assumption" [9, 12, 15]. We examined how violations of this assumption affect bio-creep. These changes in the efficacy of a therapeutic agent over time may result from variation in patient characteristics from one study to the next, differences in ancillary treatment, or other changes in the clinical setting. We

chose to imagine that all drugs are effective in one sub-population, called "susceptibles" here, and ineffective in the remainder of the population. We modeled this proportion as random, and then as steadily decreasing.

3. We also thought the way that distribution of the efficacy of new therapies changes across a series of trials could influence bio-creep, so we looked to see the consequences of allowing that distribution to shift over time.

4. We additionally examined two different methods for modeling the variance in $\hat{\theta}_{SP}$: ignoring it completely and accounting for in with the synthesis method. By doing so, we hoped to identify how much harm could be done by mis-specifying the variability of θ_{NP}.

5. Finally, we looked at two different methods for selecting the active control in the current trial. Our first approach was to select the previously approved therapy with the highest estimated treatment effect; in the second, we selected the active control that gave an ineffective new treatment the highest probability of being approved.

In each repetition of the study, 11 clinical trials were simulated: one of a new therapy against placebo and 10 non-inferiority trials of new agents against an active comparator. Every trial contained 500 subjects on both treatment arms. A time-to-event analysis was used, with event and censoring times generated using the exponential distribution. The baseline (placebo group) event rate was set at 0.25 per year; all groups had a censoring rate of 0.1 per year. A trial ended when 100, 376, or 500 events had been observed.

In the first, placebo-controlled trial, the true treatment effect of the experimental therapy, θ_{NP}, was $log(1.5) = 0.405$. The simulated data was used to estimate this treatment effect, as well as its variance. After the conclusion of this first trial, the experimental treatment became the new standard treatment and was used as the active comparator in the next trial. For this second trial, we imagined investigators selected a new experimental therapy from a population of treatments. Data for this second trial was simulated, and the results combined with those of the first trial to estimate τ as above. If the confidence interval for τ was entirely positive, then the new treatment was declared non-inferior to the active control. If, additionally, $\hat{\theta}_{NP} > \hat{\theta}_{SP}$, then in some simulations the new treatment became the standard therapy, and was used as the active comparator in the next trial. In this case, the new estimate of the treatment effect of the standard versus the placebo is $\hat{\theta}^*_{SP} = \hat{\theta}_{NP} = \hat{\theta}_{NS} + \hat{\theta}_{SP}$ and the estimated variance $\hat{\sigma}^{2*}_{SP} = \hat{\sigma}^2_{NP} = \hat{\sigma}^2_{NS} + \hat{\sigma}^2_{SP}$. Subsequent trials proceed as the second trial, with a randomly generated experimental therapy compared to the treatment considered the standard at the time. For each combination of parameters, 1000 repetitions were performed;

3.1 Study 1: Distribution of the Treatment Effect of New Agents

We modeled the true effect of the treatments being tested in the non-inferiority trials using a normal distribution with a mean of 0.405, 0.305, or 0.155, corresponding to

hazard ratios of 1.5, 1.36, and 1.17 respectively, and a standard deviation of 0.05, 0.10, or 0.50. These nine different combinations were selected in an attempt to mimic different scenarios that might occur in drug development. Centering the distribution of new drugs at the same point as the first approved therapy may approximate the scenario where companies attempt to develop a therapy similar to one their competitors are marketing. Similarly, by centering the distribution at a point slightly lower than the first drug, we hope to mimic the case where, by modifying an existing molecule in an attempt to reduce its side-effects, the efficacy is slightly reduced as well. The lowest of the means represents a more pessimistic view of drug development.

3.2 Study 2: Violations of the Constancy Assumption

To model possible treatment effect variation, we repeated the simulation study in a population where all therapies were effective in a proportion of the population and had no effect in the rest. As this proportion changed from trial to trial, the true effect of any given product, averaged over the study population, changed as well. One can easily imagine a series of trials of agents in a class, where some patients benefit from any of the drugs in the class, but for others the therapies are ineffective.

The susceptible proportion was generated in two different ways. First, for each trial, this proportion was randomly generated from a Beta(10,3) distribution; this distribution give treatment-susceptible percentages between 56.1% and 92.8% in 95% of the trials, with a mean of 76.9%. This corresponds to a situation where each trial draws from a single population of interest; differences in treatment effect are purely random. We also considered the case where the susceptible proportion declined steadily from one trial to the next, corresponding to a situation where susceptible patients are no longer interested in participating in research, as they have found a therapy effective for them. In this situation, the susceptible proportion started at 0.95 in the first trial, and declined in intervals of 0.05 to 0.45 in the final trial in each sequence. In both cases, the rest of the simulation proceeded as before.

3.3 Study 3: Choice of active comparator

While changing the susceptible proportion from trial to trial may increase the occurrence of bio-creep, many other trial characteristics affect the rate at which it occurs as well. First, we wanted to see how the choice of active comparator influenced bio-creep. We compared two different strategies: in the first, we used the approved therapy with the highest estimated efficacy as the active comparator, as in the first two studies; in the second, we used the approved drug that gave an ineffective therapy ($\theta_{NP} = 0$) the highest probability of being approved.

In order to select the standard that will make it "easiest" for an ineffective therapy to be approved, we need

$$Pr\left[\hat{\theta}_{NS} + \frac{1}{2}\hat{\theta}_{SP} - 1.96\sqrt{\hat{\sigma}^2_{NS} + \frac{1}{4}\sigma^2_{SP}} > 0 \;\middle|\; \theta_{NP} = 0, \hat{\theta}_{SP}, \hat{\sigma}^2_{SP}\right]. \qquad (1)$$

for any potential standard. By noting that in this case $\hat{\theta}_{NS} \sim N(-\theta_{SP}, \sigma^2_{NS})$ and $\hat{\theta}_{SP} = \theta_{SP} + b_{SP} + \epsilon_{SP}$, where b_{SP} is the bias of $\hat{\theta}_{SP}$ and $\epsilon_{SP} \sim N(0, \sigma^2_{SP})$, and estimating σ^2_{NS} with $\frac{4}{n}$, we can express this probability as

$$\int_{-\infty}^{\infty} \int_{\sqrt{\frac{4}{n}}\left(\frac{1}{2}\hat{\theta}_{SP} - b_{SP} - \epsilon_{SP} + 1.96\sqrt{\frac{4}{n} + \frac{1}{4}\hat{\sigma}^2_{SP}}\right)}^{\infty} \frac{1}{\hat{\sigma}_{SP}} \phi(x)\, \phi\left(\frac{\epsilon_{SP}}{\hat{\sigma}_{SP}}\right) dx\, d\epsilon_{SP}$$

After estimating b_{SP} by simulation, (1) was calculated for each potential active comparator; the approved therapy that maximizes this probability was then chosen as the new active control. This was done in the setting where the constancy assumption held.

3.4 Study 4: Modeling of variability

We next examined the choice of variance model. We compared using a synthesis method approach that appropriately modeled the variance in the estimated effect of the active comparator, as above, to one where the variability of the estimate of the active control was ignored.

3.5 Study 5: Trends in treatment effect

Finally, we looked at shifts in the distribution of θ_{NP} over time. In distinction from the first two factors examined, this is a characteristic of the study setting, and cannot be controlled by investigators. We compared the case where the distribution of new therapies centered at 0.405, 0.305, or 0.115, to that when it was instead centered around θ_{SP}, $\theta_{SP} - 0.10$, or $\theta_{SP} - 0.25$. These two factors were examined in the context where the proportion of "susceptible" subjects decreases steadily over time, with each trial stopped when 100 events had been observed.

4 Results

4.1 Study 1: Distribution of the Treatment Effect of New Agents

Several aspects of the distribution of the treatment effects of new agents influence the rate with which bio-creep occurs. As expected, when the mean of the distribution is lower, the rate of bio-creep increases. Perhaps counter-intuitively, however, the rate of bio-creep is highest at intermediate levels of variance. Additionally, increasing sample size also reduces the incidence of bio-creep.

Each simulation setting yields detailed information about the behavior of a series of non-inferiority trials. The case where mean(θ_{NP})= 0.305, SD(θ_{NP})= 0.10, and with 100 events will be reviewed in detail, and then a summary of the other settings will be provided. For this combination of trial parameters, an average of 3.25 products were approved for market after the series of 11 trials. The true hazard ratio of placebo as compared to a newly approved product ranged from 1.07 to 1.94. Table 1 gives the quantiles of the true effects of each newly approved product, by the number

of approved products, for the second through seventh products approved; the first product approved has a hazard ratio of 1.5 by design, and eight or more products were approved for marketing in so few replications as to make those summaries unreliable. First, it should be noted that about 95% of the products approved truly are "non-inferior" to the first approved product - that is, they preserved at least 50% of the treatment effect, corresponding to a HR of 1.22 or greater. Secondly, while no approved products were harmful, several had negligible treatment effects. Interestingly, the distribution of the approved products did not appear to change over time but remained relatively constant over the course of all of the trials.

An average of 3.0 new standards were adopted per sequence of trials. Since this is only marginally fewer than the number approved, it was rare that a product was approved for market and was not adopted as the new standard. Figure 1 shows the effect of the standard therapy over the 11 trials for the first 10 repetitions of the simulation for this case where mean(θ_{NP})= 0.305, SD(θ_{NP})= 0.10, and 100 events were observed. This figure illustrates that for some repetitions, the standard is constant over the duration of all trials; for other repetitions, the effect of the standard changes frequently. When the effect of the standard changes, rather than being monotone, those changes tend to oscillate between increases and decreases in efficacy. The distribution of the standard treatment at the end of trials 2 through 11 is presented in Table 2. Overall, the positive predictive value of the non-inferiority test was 96.4%, the negative predictive value was 78.8%, and the type-I error rate was 0.053, double the nominal value. No harmful treatments (HR<1) were approved for marketing, and ineffective treatments, defined as those with a hazard ratio of 1.10 or less, were approved in only 0.6% of the repetitions.

An overview of the results from Study 1, where the effect of each drug was constant over time, is given in Table 3. Each rate is the percentage of repetitions of the simulation in which a harmful or ineffective product was approved, not the percentage of approved treatments that were harmful or ineffective. Overall, harmful or ineffective products were approved in relatively few series of trials. Not surprisingly, the rate increased as the average efficacy of the products being tested decreased. Bio-creep also occurred more frequently in the lower-powered trials.

Figure 2 shows the median, central 50%, and central 90% of the effects of all approved therapies at the conclusion of each trial for the cases where $n = 100$. Notably, when SD(θ_{NP}) = 0.5 and hence products much superior to the initial therapy are being tested, the distribution of approved therapies shifts toward more beneficial drugs.

4.2 Study 2: Violations of the Constancy Assumption

We also wanted to see how often bio-creep occurred when the constancy assumption was violated. We first considered a violation of this assumption due to random changes in the proportion "susceptible" to the class of agents being tested. Table 4 gives the results from this simulation, where the effect of a single drug in the study population changed from trial to trial. Here, not surprisingly, bio-creep occurred much more frequently. For some of the treatment effect distributions, harmful products were approved in more than 3% of repetitions, a clearly unacceptable level of

error.

We next examined the rates of bio-creep that occur when the proportion of "susceptible" subjects decreases steadily over time. These results are presented in Table 5. Here, the rates of bio-creep are even higher than in the above sub-study, with ineffective products approved in up to 16.3% of repetitions and harmful products in as many as 4.9%. Interestingly, increasing sample size does not offer the same protection as before. While the rates of bio-creep did decrease with increasing sample size for all parameter settings, event with 500 events, as many as 10.2% of repetitions saw ineffective therapies approved.

4.3 What goes wrong?

We examined the characteristics of the repetitions where ineffective or harmful therapies were approved, and compared them to the repetitions where this did not occur. Hopefully, identifying the traits that led to this bio-creep can help us develop ways of preventing it in the future. This was done in the context of Study 2, where the class of the agents under study decreases steadily over time; we present detailed results for the case where mean(θ_{NP}) = 0.305, SD(θ_{NP}) = 0.10, and each trial was stopped when 100 events had been observed.

As can be seen in Table 5, ineffective therapies were approved in 1.9% of repetitions, and harmful therapies in 0.2%. No more than one harmful or ineffective therapy was approved in any one repetition; on average, 0.32% of approved therapies were ineffective, and 0.02% were harmful.

Ineffective and harmful therapies were more likely to be approved in repetitions with many approved therapies and standards. In repetitions where no ineffective therapies were approved, a mean of 3.6 treatments were on the market, with an average of 3.2 standards. In contrast, in repetition with ineffective therapies were approved, the mean number of treatments available was 6.9, with a mean of 5.7 standards. The contrast is even more striking when examining the repetitions where a harmful therapy was approved: there, the mean number of approved treatments was 8.5, with an average of 6.0 standards. Of course, those numbers are not be estimated very precisely, since harmful therapies were approved in just 2/1000 repetitions. In all of these repetitions, harmful and ineffective therapies were not approved until later in the chain of trials. The two harmful therapies appeared in trials 8 and 11. The first ineffective therapy was approved in a trial 5, and the mode (7/19) was at trial 11. The median trial number where the first ineffective therapy was observed was 9.

It is noteworthy that the mean estimated effect of the first treatment was much larger (-0.658) in repetitions that eventually led to the approval of ineffective therapies than in those that did not (-0.553), and even higher in those that yielded a harmful therapy (mean=-0.772). It seems that when the efficacy of the first therapy, the "anchor" to the chain of non-inferiority trials, is over-estimated, it is much easier for new therapies to be approved, including those that are ineffective and harmful.

4.4 Study 3: Choice of active comparator

We began by examining the strategy used to selecting the active comparator. First, we used the approved therapy with the highest estimated efficacy as the active comparator. For the next set of simulations, we selected the therapy that, if it was chosen as the new standard, would result in the highest chance of a completely ineffective therapy ($\theta_{NP} = 0$) being approved. This was done in the setting where constancy holds. In Table 6, these results are denoted by the columns marked "Best" and "Easiest". The rates of approval of ineffective and harmful therapies are similar for the two procedures, suggesting that it is difficult to chose the active comparator in a way that "games" the system. The active comparator that provides the easiest path to approval would have low θ_{SP}, high b_{SP}, and low σ^2_{SP}. With θ_{SP} unknown, however, those factors can't be optimized simultaneously. High values of $\hat{\theta}_{SP}$ can suggest both high θ_{SP} and high b_{SP}; therapies approved later in a chain of non-inferiority trials will have both high b_{SP} and high σ^2_{SP}. With these factors off-setting one another, the method used for selecting the active control does not appear to greatly affect the rates of bio-creep.

4.5 Study 4: Modeling of variability

We next examined the choice of variance model. We compared a synthesis method approach to one where the variability of the estimate of the active control was ignored. This was done when the "susceptible" proportion of the study population decreases steadily from trial to trial. As expected, using the incorrect variance model led to disastrous results, as can be seen in the first two columns of Table 7. In the worst case setting, ineffective therapies were approved in as many as 48% of repetitions, and harmful therapies in 36%.

4.6 Study 5: Trends in treatment effect

Finally, we wanted to see how changes in the distribution of θ_{NP} over time would affect the rates of bio-creep. When we allowed this distribution to change over time, θ_N was centered around either θ_{SP}, $\theta_{SP} - 0.10$, or $\theta_{SP} - 0.25$, rather than 0.405, 0.305, or 0.115. Again, the "susceptible" proportion of the study population decreased steadily from trial to trial. Results are given in Table 7 for the cases when the variance model is correct or incorrect. Allowing the distribution of θ_{NP} to change over time affected the bio-creep rates differently, depending on the variance in the distribution of θ_{NP}. At the highest standard deviation studied, 0.50, shifting the mean of θ_{NP} reduced the rates of bio-creep. At the other two variance levels, however, shifting the mean increased the rate of bio-creep, sometimes drastically. For example, when mean(θ_{NP}) = 0.305 and SD(θ_{NP}) = 0.05, when ignoring the variability of θ_{SP}, but with a fixed mean of the distribution of θ_{NP}, no ineffective therapies were approved. When the mean shifted over time, ineffective therapies were approved in 44% of the repetitions. Even when the variance model was correctly specified, ineffective therapies were approved in 23% of the repetitions.

5 Discussion

These results demonstrate what should be inherently apparent: it is imperative to use methodology appropriate for the non-inferiority setting when conducting a non-inferiority clinical trial. The variability inherent to estimating the effect of the active comparator must be considered and accounted for; failing to do so results in disastrously high rates of bio-creep.

Additional thought should be given to selecting the active comparator. We examined how this choice affected the rates of bio-creep for two different strategies: using the therapy with the best estimated treatment effect and that giving an ineffective therapy the highest chance of being approved. Naturally, there are many other possibility procedures that could be used. For example, a drug developer may think that its new product stands the best chance against the approved drug with the lowest estimated treatment effect. Of course, this will lead to a tighter margin than using a therapy with a higher estimated effect. For the purpose of this simulation study, we adopted a fixed policy for the selection of the active control. In reality, this decision should be made only after a careful evaluation of the available therapies. The estimated effects of the available treatments, and how those effects have appeared over time, need to be considered along with clinical practice. There may be ethical issue to entertain as well: are any of the available therapies known to be superior to the others? It may not be ethical to deny subjects such a therapy.

The evidence supporting the efficacy of the active comparator must also be evaluated. Since the first, placebo-controlled trial anchors the chain of non-inferiority trials which follow, any problems with this first trial can lead to trouble later. If the effect of that first approved drug just beats placebo, doing a trial with 50% retention is tantamount to performing a superiority trial. Spuriously high estimates of the effect of that first therapy greatly increase the chance of an ineffective or harmful treatment being approved in subsequent trials.

We examined the factors influencing and the rates of bio-creep using one definition of non-inferiority. Clearly, selecting a higher percentage retention will lead to lower rates of bio-creep, with a corresponding loss of power to approve beneficial therapies. Or one may instead be interested in the rates of bio-creep seen using a fixed margin. We expect the results to be similar to what we observed, with some variation depending on the population from which new treatment effects are generated. Other statistics could be used instead of the synthesis method approach utilized here. For example, we would expect a 95-95 approach, with its built in conservatism, to result in lower rates of bio-creep, again at the cost of a reduction in power.

From the above results, it is apparent that violations of the constancy assumption are the largest potential source of bio-creep. When this assumption holds, as long as appropriate analysis techniques are used in conjunction with a reasonable non-inferiority margin, bio-creep appears to be rare (at least in settings of our simulations). However, when this assumption is violated, rates of bio-creep can be disturbingly high, even when proper methodology is used. Since this assumption is largely untested (some might even say untestable), these results are quite concerning. Further consideration should be given to techniques for identifying these

violations, and methods for handling them when they are detected.

References

[1] D'Agostino Sr RB, Massaro JM, Sullivan LM. Non-inferiority trials: Design concepts and issues – The encounters of academic consultants in statistics. *Statistics in Medicine* 2003; **22**(2):169–186.

[2] Fleming TR. Current issues in non-inferiority trials (Pkg: P317-342). *Statistics in Medicine* 2008; **27**(3):317–332.

[3] E-9 I. *International Conference on Harmonisation - Statistical Principles for Clinical Trials*. Published in the Federal Register of 16 September 1998 (63 FR 49583).

[4] Snapinn SM. Alternatives for discounting in the analysis of noninferiority trials. *Journal of Biopharmaceutical Statistics* 2004; **14**(2):263–273.

[5] Wang SJ, Hung HMJ. Assessing treatment efficacy in noninferiority trials. *Controlled Clinical Trials* 2003; **24**:147–155.

[6] Holmgren E. 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.

[7] Memorandum C. Summary of CBER considerations on selected aspects of active controlled trial design and analysis for the evaluation of thrombolytics in actue mi June 1999.

[8] Laster LL, Johnson MF. Non-inferiority trials: the 'at least as good as' criterion. *Statistics in Medicine* 2003; **22**:187–200.

[9] Hung HMJ, Wang SJ, Tsong Y, Lawrence J, O'Neil RT. Some fundamental issues with non-inferiority testing in active controlled trials. *Statistics in Medicine* 2003; **22**(2):213–225.

[10] Wang SJ, Hung HMJ. TACT method for non-inferiority testing in active controlled trials. *Statistics in Medicine* 2003; **22**(2):227–238.

[11] Hasselblad V, Kong DF. Statistical methods for comparison to placebo in active-control trials. *Drug Information Journal* 2001; **35**:435–449.

[12] Rothmann M, Li N, Chen G, Chi GYH, Temple R, Tsou HH. Design and analysis of non-inferiority mortality trials in oncology. *Statistics in Medicine* 2003; **22**(2):239–264.

[13] Shepherd F, Dancey J, Ramlau R, Mattson K, Gralla R, O'Rourke M, Levitan N, Gressot L, Vincent M, Burkes R, *et al.*. Prospective randomized trial of docetaxel versus best supportive care in patients with non-small-cell lung cancer previously treated with platinum-based chemotherapy. *Journal of Clinical Oncology* 2000; **18**:2095–2103.

[14] Hanna N, Shepherd F, Fossella F, Pereira J, De Marinis F, von Pawel J, Gatzemeier U, Tsao T, Pless M, Muller T, *et al.*. Randomized phase III trial of pemetrexed versus docetaxel in patients with non-small-cell lung cancer previously treated with chemotherapy. *Journal of Clinical Oncology* MAY 1 2004; **22**(9):1589–1597.

[15] Fisher M Lloyd Dand Gent, Buller HR. Active-control trials: How would a new agent compare with placebo? a method illustrated clopidogrel, aspirin, and placebo. *American Heart Journal* 2001; **141**:26–32.

Number of Approved Products	Min.	5th %-tile	25th %-tile	Median	75th %-tile	95th %-tile	Max.
2	1.07	1.27	1.37	1.46	1.54	1.70	1.94
3	1.12	1.23	1.34	1.42	1.51	1.65	1.85
4	1.03	1.19	1.32	1.41	1.50	1.64	1.82
5	1.12	1.20	1.32	1.40	1.51	1.65	1.92
6	1.11	1.19	1.30	1.39	1.50	1.63	1.82
7	1.09	1.19	1.33	1.38	1.48	1.63	1.79

Table 1: True effects of products on market by number approved from simulation where mean(θ_{NP}) = 0.305, SD(θ_{NP}) = 0.10, and 100 events were observed.

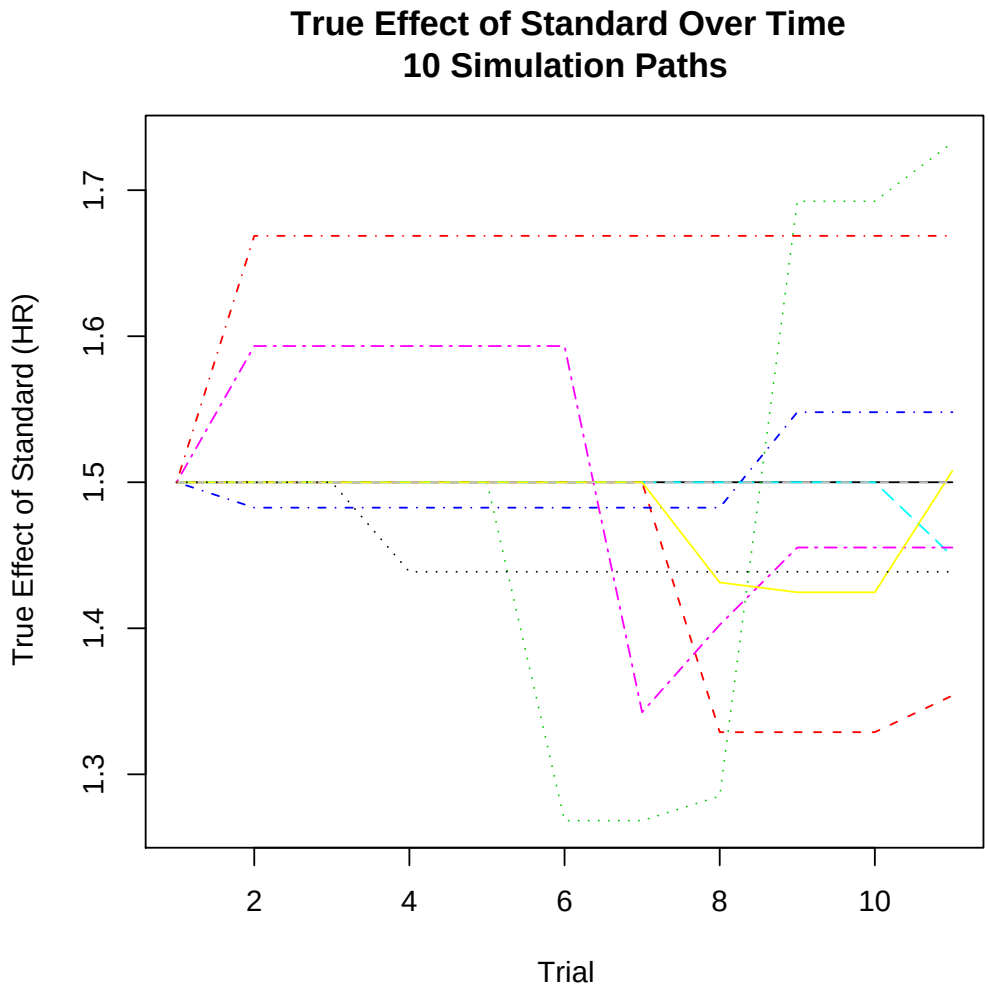

Figure 1: Plot of the effect of the standard at the end of each trial, for the first 10 repetitions of the simulation. Parameters set to mean($\hat{\theta}_{NP}$) = 0.305, SD($\hat{\theta}_{NP}$) = 0.10, and 100 observed events.

Number of Approved Products	Min.	5th %-tile	25th %-tile	Median	75th %-tile	95th %-tile	Max.
2	1.19	1.46	1.50	1.50	1.50	1.50	1.76
3	1.07	1.38	1.50	1.50	1.50	1.59	1.94
4	1.13	1.32	1.50	1.50	1.50	1.62	1.94
5	1.13	1.33	1.50	1.50	1.50	1.64	1.94
6	1.05	1.32	1.48	1.50	1.50	1.66	1.94
7	1.03	1.32	1.46	1.50	1.50	1.67	1.94
8	1.12	1.30	1.44	1.50	1.50	1.67	1.94
9	1.09	1.30	1.44	1.50	1.52	1.69	1.94
10	1.10	1.28	1.43	1.50	1.52	1.69	1.94
11	1.12	1.29	1.43	1.50	1.54	1.70	1.94

Table 2: True effects of standard therapy at the conclusion of each trial, from simulation where mean(θ_{NP}) = 0.305, SD(θ_{NP}) = 0.10, and 100 events were observed.

mean(θ_{NP})	SD(θ_{NP})		n=100	n=376	n=500
0.405	0.05	% Ineffective	0	0	0
		% Harmful	0	0	0
0.305	0.05	% Ineffective	0	0	0
		% Harmful	0	0	0
0.155	0.05	% Ineffective	4.5	1.5	0.9
		% Harmful	0	0	0
0.405	0.10	% Ineffective	0	0	0
		% Harmful	0	0	0
0.305	0.10	% Ineffective	0.6	0	0
		% Harmful	0	0	0
0.155	0.10	% Ineffective	6.0	2.3	1.0
		% Harmful	1.0	0.3	0.1
0.405	0.50	% Ineffective	0.6	0	0
		% Harmful	0.3	0	0
0.305	0.50	% Ineffective	0.5	0	0
		% Harmful	0.1	0	0
0.155	0.50	% Ineffective	1.1	0	0
		% Harmful	0.3	0	0

Table 3: Results from Study 1.

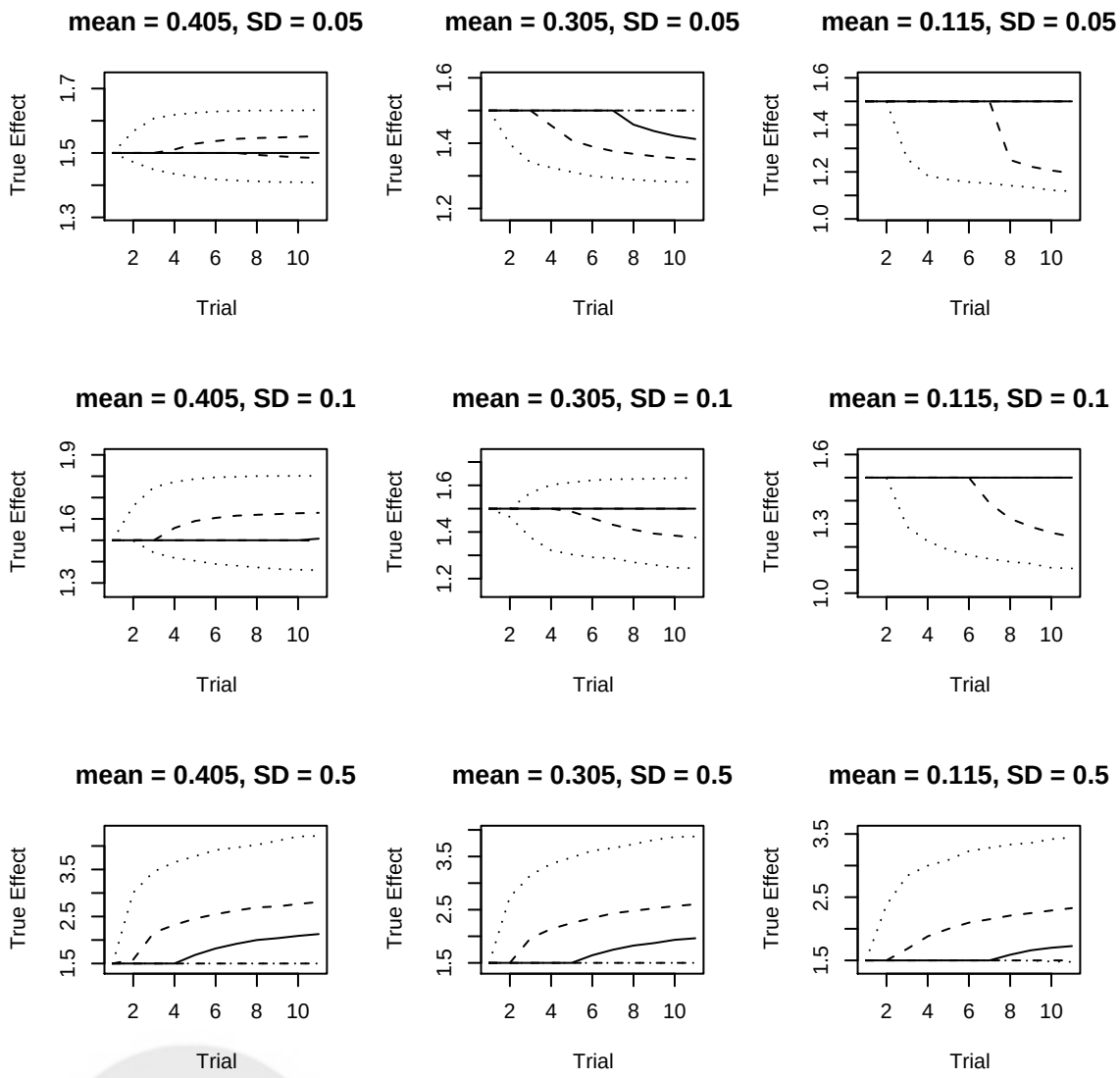

Figure 2: Median, central 50% and central 90% of approved therapies over time. Each trial had 100 observed events.

mean(θ_{NP})	SD(θ_{NP})		n=100	n=376	n=500
0.405	0.05	% Ineffective	0	0	0
		% Harmful	0	0	0
0.305	0.05	% Ineffective	0	0	0
		% Harmful	0	0	0
0.155	0.05	% Ineffective	8.9	2.8	2.2
		% Harmful	0.2	0	0
0.405	0.10	% Ineffective	0.2	0	0
		% Harmful	0	0	0
0.305	0.10	% Ineffective	1.4	0.3	0.4
		% Harmful	0	0	0
0.155	0.10	% Ineffective	13.3	3.4	3.5
		% Harmful	3.1	0.5	0.7
0.405	0.50	% Ineffective	3.1	0.4	0.3
		% Harmful	1.2	0.3	0.1
0.305	0.50	% Ineffective	4.0	0.5	0.4
		% Harmful	1.7	0.2	0.3
0.155	0.50	% Ineffective	4.5	0.4	0.5
		% Harmful	2.7	0.1	0.1

Table 4: Results from Study 2 when susceptible proportion changes randomly.

mean(θ_{NP})	SD(θ_{NP})		n=100	n=376	n=500
0.405	0.05	% Ineffective	0	0	0
		% Harmful	0	0	0
0.305	0.05	% Ineffective	0	0	0
		% Harmful	0	0	0
0.155	0.05	% Ineffective	8.3	6.1	4.7
		% Harmful	0.2	0	0
0.405	0.10	% Ineffective	0.2	0	0.2
		% Harmful	0	0	0.1
0.305	0.10	% Ineffective	1.9	0.6	1.3
		% Harmful	0.2	0	0.1
0.155	0.10	% Ineffective	16.3	10.6	10.2
		% Harmful	3.0	1.7	1.2
0.405	0.50	% Ineffective	6.6	2.5	1.8
		% Harmful	4.3	0.8	1.2
0.305	0.50	% Ineffective	7.1	2.4	1.9
		% Harmful	4.1	1.2	0.5
0.155	0.50	% Ineffective	8.2	2.7	1.7
		% Harmful	4.9	0.9	0.9

Table 5: Results from Study 2 when susceptible proportion decreases steadily over time.

mean(θ_{NP})	SD(θ_{NP})		Best	Easiest
0.405	0.05	% Ineffective	0	0
		% Harmful	0	0
0.305	0.05	% Ineffective	0	0
		% Harmful	0	0
0.155	0.05	% Ineffective	4.5	4.5
		% Harmful	0	0
0.405	0.10	% Ineffective	0	0
		% Harmful	0	0
0.305	0.10	% Ineffective	0.6	0.5
		% Harmful	0	0
0.155	0.10	% Ineffective	6.0	6.0
		% Harmful	1.0	1.2
0.405	0.50	% Ineffective	0.6	0.2
		% Harmful	0.3	0.2
0.305	0.50	% Ineffective	0.5	0.2
		% Harmful	0.1	0
0.155	0.50	% Ineffective	1.1	1.2
		% Harmful	0.3	0.3

Table 6: Comparison of the rates of bio-creep when either 1) the approved product with the best estimated treatment effect or 2) the one most likely to lead to the approval of an ineffective therapy is chosen as the active control for subsequent trials.

mean(θ_{NP})	SD(θ_{NP})	Variance Model Mean Shifts	C N	I N	C Y	I Y
0.405	0.05	% Ineffective	0	0	0.2	0.3
		% Harmful	0	0	0	0
0.405	0.10	% Ineffective	0.2	0.2	1.5	3.6
		% Harmful	0	0	0.6	1.5
0.405	0.50	% Ineffective	6.6	17.4	2.3	4.6
		% Harmful	4.3	10.5	1.5	3.4
0.305	0.05	% Ineffective	0	0	23.3	44.0
		% Harmful	0	0	12.7	28.2
0.305	0.10	% Ineffective	1.9	3.3	19.2	35.2
		% Harmful	0.2	0.2	9.9	21.6
0.305	0.50	% Ineffective	7.1	18.5	3.1	8.5
		% Harmful	4.1	11.7	2.0	5.4
0.155	0.05	% Ineffective	8.3	17.4	21.3	38.5
		% Harmful	0.2	0.2	18.8	35.9
0.155	0.10	% Ineffective	16.3	30.7	23.3	41.9
		% Harmful	3.0	7.9	15.6	33.0
0.155	0.50	% Ineffective	8.2	19.2	6.9	13.8
		% Harmful	4.9	13.5	3.9	9.3

Table 7: Results of "Worse Case" Scenario Simulation. In each setting, the "susceptible" proportion decreases steadily from 0.95 in the first trial to 0.45 in the last trial. Each trial stopped when 100 events had been observed.