Research Article

Statistics
in Medicine

Received 8 September 2009, Accepted 14 July 2010 Published online 30 August 2010 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4053

# Bio-creep in non-inferiority clinical trials

Siobhan Everson-Stewart*† and Scott S. Emerson

**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 account for 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. Copyright © 2010 John Wiley & Sons, Ltd.**

**Keywords:** bio-creep; non-inferiority clinical trials; constancy assumption

## 1. Introduction

Before any drug or biologic is approved for administration to patients, its efficacy must first be evaluated in a clinical trial. These studies, performed in human volunteers, are designed to evaluate the risk/benefit profile of a treatment. Well-conducted, randomized, placebo-controlled trials provide the strongest available evidence about the effectiveness and safety of a new agent.

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, 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 beneficial treatment effect over placebo. Especially if an investigational treatment has advantages over the standard therapy, such as an easier mechanism of 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.

Through these non-inferiority clinical trials, a new therapy may be approved even if it is less effective than its predecessor. This raises the possibility that, after a series of non-inferiority trials with each new drug being a little worse than its predecessor, an ineffective or harmful therapy may falsely be deemed efficacious. This phenomenon is 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 happen. There are several factors which 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.

*Department of Biostatistics, University of Washington, Seattle, WA 98195, U.S.A.*

*Correspondence to: Siobhan Everson-Stewart, Resuscitations Outcome Consortium Clinical Trial Center, University of Washington, 1107 NE 45th St. Suite 505, Seattle, WA 98105-4689, U.S.A.*

†*E-mail: spes@uw.edu*

## 2. Defining non-inferiority

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 does 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 the active control exceed some threshold (e.g. hazard ratio for experimental therapy to active control must exceed 0.9), often chosen based on 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. 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 per cent 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 retaining a certain percentage of the effect of the active control. Strictly interpreted, efficacy requirements demand only that a new treatment be more effective than placebo, but 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 clinically 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 that some percentage of the effect of the active comparator be retained, with the hope of actually preserving this fraction of the effect, is one way of ensuring that is not the case. Additionally, when some proportion of the effect of the active control is retained, one can 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 [6–8].

### *2.2. Incorporating uncertainty*

It is important to incorporate the uncertainty of the estimate of the effect of the active comparator when selecting the statistical methods to be used. Many available methods do take this increased variability into account, although their operating characteristics vary widely [9, 10]. For the purpose of this study, we used a 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 per cent confidence interval for the efficacy of the active control (as compared to placebo). As long as the 95 per cent 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 $\pi\times 100$ per cent, then the margin is set at $\pi$ 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 about 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 $\lambda_P(t)=\lambda_S(t)\exp(\theta_{PS})$. Under this model, $\theta_{PS}$ is the log hazard ratio of the placebo group compared to the active treatment group, and $\theta_{PS}>0$ implies that the standard therapy is efficacious. Although $\theta_{PN}$ is not directly estimable from a non-inferiority trial, combining the historical data with the data from the new trial gives us $\hat{\theta}_{PN}=\hat{\theta}_{PS}+\hat{\theta}_{SN}$. Inference about the efficacy of the new agent can then be based directly on $\hat{\theta}_{PN}$.

If one suspects 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 some protection against a diminished effect of the active comparator, as long as the decrease in efficacy is not large. For these simulations, we took this approach, with an arbitrary 50 per cent retention threshold.

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

$$\theta_{PN}=\theta_{PS}+\theta_{SN}>\tfrac{1}{2}\theta_{PS}$$

or, equivalently,

$$\tau\equiv\tfrac{1}{2}\theta_{PS}+\theta_{SN}>0.$$

If a 95 per cent confidence interval for $\tau$ lies entirely above 0, then non-inferiority can be concluded with a nominal one-sided 0.025 Type I error rate. As $\mathrm{Var}(\tau)=\sigma^2_{SN}+\frac{1}{4}\sigma^2_{PS}$, 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 non-inferiority trial and margins.* Consider the case of second-line chemotherapy in non-small-cell lung cancer. In this indication, Shepherd *et al.* established the effectiveness of docetaxel in a placebo-controlled clinical trial [13]. Later, pemetrexed was tested in a nominally similar patient population by Hanna and co-authors through a non-inferiority trial comparing pemetrexed to docetaxel [14]. From the earlier trial, we have $\exp(\hat{\theta}_{PS})=1.78$, with a 95 per cent confidence interval of 1.14 to 2.86 [14]. If one can establish that $\exp(\theta_{SN})>\frac{1}{1.78}=0.56$, it suggests that pemetrexed is more effective than placebo. However, the variability in the estimate of $\theta_{PS}$ must be taken into consideration. Additionally, one may want to demand that pemetrexed retain some portion of the effect of docetaxel before declaring non-inferiority. If one chose to use a synthesis method to account for the variability of $\hat{\theta}_{PS}$ while demanding 50 per cent retention, one wants to be confident that

$$\tau\equiv\theta_{SN}+\tfrac{1}{2}\theta_{PS}>0.$$

One may obtain $\hat{\theta}_{PS}=\log(1.78)$ and $\hat{\sigma}_{PS}=0.23$ from the historical data. As the trial comparing docetaxel to pemetrexed was designed to stop when $n=400$ events had been observed, one can infer that $\sigma^2_{SN}\approx\frac{4}{400}=0.01$ [15]. One can then calculate that pemetrexed will be declared non-inferior to docetaxel when $\exp(\hat{\theta}_{SN})\geqslant 1.01$, with a corresponding confidence interval of (0.83, 1.23). In this setting, using a synthesis method with 50 per cent retention implies a margin of 0.83 on the $\exp(\theta_{SN})$ 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 per cent retention while using this 95–95 method, yielding a margin of $\sqrt{(0.88)}=0.94$ for $\exp(\theta_{SN})$. With the trial designed to stop when 400 events had been observed, this lower bound would only be exceeded when $\exp(\hat{\theta}_{SN})>1.14$.

## 3. Methods: a simulation study

We aimed to identify the factors that can lead to increases in the rate of bio-creep. We began in Study 1 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. Similarly, if all new agents tested are harmful, then any new therapy that is approved can be thought of as an occurrence of bio-creep. We focused on the intermediate case, where the efficacy of the new therapies being tested ranges from harmful to beneficial.

Next, in Study 2, we explored changes in the effect of a single therapy across trials. Most non-inferiority methodologies assume that the treatment of any one drug is constant across trials; this is known as the *constancy assumption* [9, 12, 16]. We examined how violations of this assumption affect bio-creep. These changes in the efficacy of a therapeutic agent over time may result from variations in patient characteristics from one study to the next, differences in ancillary treatment, or other changes in the clinical setting. We assumed that all drugs are effective in one subpopulation, called *susceptibles*, and ineffective in the remainder of the population. We first sampled this proportion for each trial from a Beta distribution. In a second set of simulations, we set this proportion to 0.95 in the first trial, and then decreased it by 0.05 for each subsequent trial.

In Study 3, we investigated 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. It is feasible that an unethical sponsor might take such an approach, and we wanted to see what the impact of this worst case scenario would be.

We were also interested in the impact of ignoring the variability in $\hat{\theta}_{PS}$, rather than using the synthesis method to account for it. This was explored in Study 4.

Finally, in Study 5, we investigated the consequences of allowing the distribution of the efficacy of new therapies to change across the series of trials.

For all studies except Study 4, we used a synthesis method approach with a 50 per cent retention threshold as detailed in Section 2.3. For each repetition of a study, a series of 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 $\lambda_P = 0.25$; all groups had a censoring rate of 0.1 per year. A trial ended when $n$ events had been observed, where $n = 100$, 376, or 500.

In the first trial, the true treatment effect of the experimental therapy was $\lambda_1 = 0.25/1.5$: the hazard ratio of placebo compared to the first drug was 1.5 and $\theta_1 = \log(1.5) = 0.405$. The Cox proportional hazards model, $\lambda_P(t) = \lambda_1(t)\exp(\theta_{P1})$, was fit to obtain $\hat{\theta}_{P1}$ and $\hat{\sigma}_{P1}$. When the effect of the experimental therapy was significantly greater than zero (i.e. $\hat{\theta}_{P1} - 1.96\hat{\sigma}_{P1} > 0$), then Drug 1 became the active control, and we continued to Trial 2. Otherwise, the first study was repeated (with the same experimental agent) until a significant result was achieved. At that time, the most recent trial results were used to obtain $\hat{\theta}_{P1}$ and $\hat{\sigma}_{P1}$; previous results were discarded. This process was used to mimic the bias introduced as a result of only positive trials being used to establish a standard therapy. Consequently, we expect the observed effect of the active control to tend to be lower in the new trial than was seen in the historical trial due to regression to the mean.

For all other treatments, the event rate $\lambda_i = 0.25/\exp(\theta_i)$, with $\theta_i$ generated independently from a normal distribution with mean $\mu$ and standard deviation $\sigma$, $i = 2, \ldots, 11$. The second trial began by randomly selecting $\theta_2$. The event and censoring times were simulated as in trial 1, stopping the trial when $n$ events had been observed. As before, a Cox proportional hazards model, $\lambda_1(t) = \lambda_2(t)\exp(\theta_{12})$, was fit to obtain $\hat{\theta}_{12}$ and $\hat{\sigma}_{12}$. These results were combined with $\hat{\theta}_{P1}$ and $\hat{\sigma}_{P1}$ to get a 95 per cent confidence interval for $\tau$ with the 50 per cent retention criterion as above. When the interval for $\tau$ fell entirely above zero, Drug 2 was approved. If in addition, $\hat{\theta}_{12} > 0$, then Drug 2 became the new standard, with an estimated effect of $\hat{\theta}^*_{PS} = \hat{\theta}_{P2} = \hat{\theta}_{12} + \hat{\theta}_{P1}$, and estimated standard deviation of $\hat{\sigma}^*_{PS} = \hat{\sigma}_{P2} = \sqrt{\hat{\sigma}^2_{12} + \hat{\sigma}^2_{P1}}$.

Trials 3–11 proceeded as Trial 2. We looked for approval of harmful (HR<1) and ineffective (HR<1.1) therapies. For each combination of parameters, 1000 repetitions of the sequence of 11 trials 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 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 that 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 on 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 gave treatment-susceptible percentages between 56.1 and 92.8 per cent in 95 per cent of the trials, with a mean of 76.9 per cent. This corresponds to a situation where each trial draws from a single population of interest; differences in treatment effect are purely random. Second, we considered the case where the susceptible proportion declined steadily from one trial to the next. 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. This was designed to correspond to a situation where susceptible patients are no longer interested in participating in the research, as they have found a therapy that is effective for them.

### 3.3. Study 3: *choice of active comparator*

To see how the choice of active comparator influenced bio-creep, we compared two different strategies. First, we used the approved therapy with the highest estimated efficacy as the active comparator, as in Studies 1 and 2. Second, we used the approved drug that gave an ineffective therapy ($\theta_{PN}=0$) 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

$$Pr[\hat{\theta}_{SN}+\tfrac{1}{2}\hat{\theta}_{PS}-1.96\sqrt{\hat{\sigma}^2_{SN}+\tfrac{1}{4}\sigma^2_{PS}}>0|\theta_{PN}=0,\hat{\theta}_{PS},\hat{\sigma}^2_{PS}] \tag{1}$$

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

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

The bias of the first approved drug was estimated by averaging the difference between the actual and estimated effect of that drug across 1000 simulations. This was repeated for the second approved drug, the third approved drug, and so on. Using this as an estimate of $b_{PS}$, (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: *accounting for variability*

We next examined the impact of ignoring the variability of $\hat{\theta}_{PS}$. We compared the synthesis method approach that incorporated the variance of the estimated effect of the active comparator, as above, to one where the variability of the estimate of the active control was ignored. This was done where the proportion of susceptible subjects decreased steadily over time, with each trial stopping when 100 events had been observed.

### 3.5. Study 5: *trends in treatment effect*

Finally, we compared the case where the distribution of new therapies was centered at 0.405, 0.305, or 0.115, to that when it was instead centered around $\theta_{PS}$, $\theta_{PS}-0.10$, or $\theta_{PS}-0.25$. As in Study 4, we looked at this factor in a setting where the proportion of susceptible subjects decreased steadily over time and each trial stopped when 100 events had been observed.

## 4. Results

### 4.1. Study 1: *distribution of the treatment effect of new agents*

Each simulation setting yields detailed information about the behavior of a series of non-inferiority trials. The case where mean($\theta_{PN}$)$=0.305$, SD($\theta_{PN}$)$=0.10$, and $n=100$ 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 I 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

**Table I.** True effects of products on the market by number approved from simulation, where mean($\theta_{PN}$)=0.305, SD(($\theta_{PN}$)=0.10, and 100 events were observed.

| Number of approved products | Min. | 5th percentile | 25th percentile | Median | 75th percentile | 95th percentile | 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 |

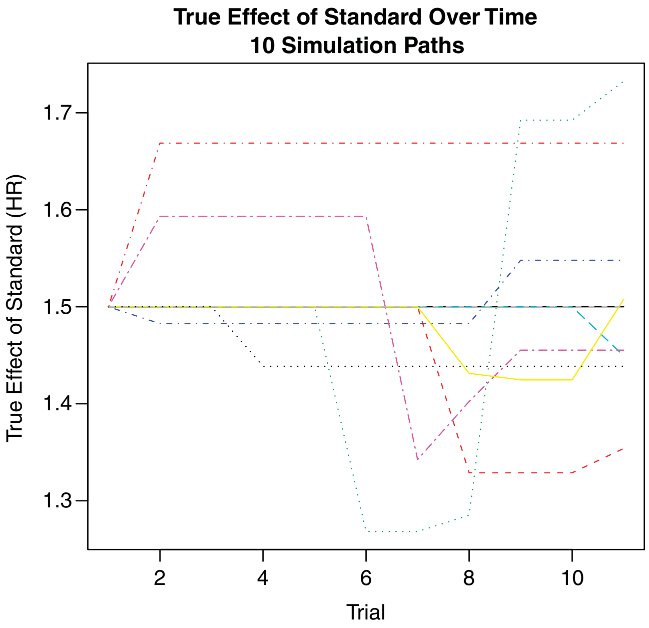

**Figure 1.** Plot of the effect of the standard at the end of each trial, for the first 10 repetitions of the simulation. Parameters set to mean($\hat{\theta}_{PN}$)=0.305, SD($\hat{\theta}_{PN}$)=0.10, and 100 observed events.

1.5 by design, and eight or more products were approved for marketing in few replications, making those summaries unreliable. First, it should be noted that about 95 per cent of the products approved truly are 'non-inferior' to the first approved product—that is, they preserved at least 50 per cent of the treatment effect, corresponding to an 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. As 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($\theta_{PN}$)=0.305, SD($\theta_{PN}$)=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, these changes tend to oscillate between increases and decreases in efficacy. The distribution of the standard treatment at the end of trials 2–11 is presented in Table II. 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 per cent 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 III. Each rate is the percentage of repetitions of the simulation in which a harmful or ineffective product was approved, and 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 per cent, and central 90 per cent of the effects of all approved therapies at the conclusion of each trial for the cases where $n=100$. Notably, when SD($\theta_{PN}$)=0.5 and hence products much

**Table II**. True effects of standard therapy at the conclusion of each trial, from simulation where mean($\theta_{PN}$)=0.305, SD($\theta_{PN}$)=0.10, and 100 events were observed.

| Number of approved products | Min. | 5th percentile | 25th percentile | Median | 75th percentile | 95th percentile | 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 III**. Results from Study 1.

| Mean($\theta_{PN}$) | SD($\theta_{PN}$) | | $n=100$ | $n=376$ | $n=500$ |
|---|---|---|---|---|---|
| 0.405 | 0.05 | Per cent ineffective | 0 | 0 | 0 |
| | | Per cent harmful | 0 | 0 | 0 |
| 0.305 | 0.05 | Per cent ineffective | 0 | 0 | 0 |
| | | Per cent harmful | 0 | 0 | 0 |
| 0.155 | 0.05 | Per cent ineffective | 4.5 | 1.5 | 0.9 |
| | | Per cent harmful | 0 | 0 | 0 |
| 0.405 | 0.10 | Per cent ineffective | 0 | 0 | 0 |
| | | Per cent harmful | 0 | 0 | 0 |
| 0.305 | 0.10 | Per cent ineffective | 0.6 | 0 | 0 |
| | | Per cent harmful | 0 | 0 | 0 |
| 0.155 | 0.10 | Per cent ineffective | 6.0 | 2.3 | 1.0 |
| | | Per cent harmful | 1.0 | 0.3 | 0.1 |
| 0.405 | 0.50 | Per cent ineffective | 0.6 | 0 | 0 |
| | | Per cent harmful | 0.3 | 0 | 0 |
| 0.305 | 0.50 | Per cent ineffective | 0.5 | 0 | 0 |
| | | Per cent harmful | 0.1 | 0 | 0 |
| 0.155 | 0.50 | Per cent ineffective | 1.1 | 0 | 0 |
| | | Per cent harmful | 0.3 | 0 | 0 |

superior to the initial therapy are being tested, the distribution of approved therapies shifts toward more beneficial drugs.

Because the first trial was repeated until a significant result was achieved, we expect the first estimate of $\theta_{PS}$ to be positively biased. For $n=100$, this is true. Over the 1000 simulations, mean($\hat{\theta}_{PS}$)=0.565, well above the true value of 0.405. With the larger sample sizes, this bias is negligible. When $n=376$, mean($\hat{\theta}_{PS}$)=0.408 and for $n=500$, mean($\hat{\theta}_{PS}$)=0.407. At the smaller sample size, then, we expect some regression to the mean; the apparent treatment effect of the active control may be lower in the second trial than in the first.

Several aspects of the distribution of the treatment effects of new agents influenced the rate at which bio-creep occurred. As expected, when the mean of the distribution was lower, the rate of bio-creep increased. Perhaps counter-intuitively, however, the rate of bio-creep was highest at intermediate levels of variance. With the smallest variance, very few ineffective or harmful therapies were tested, making bio-creep rare. With the highest variance, an extremely effective therapy was often approved, making it difficult for ineffective drugs to make it to the market. Additionally, increasing the sample size reduced the incidence of bio-creep.

### *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 IV 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 than when constancy held. For some of the treatment effect distributions, harmful products were approved in more than 3 per cent of repetitions, a clearly unacceptable level of error.

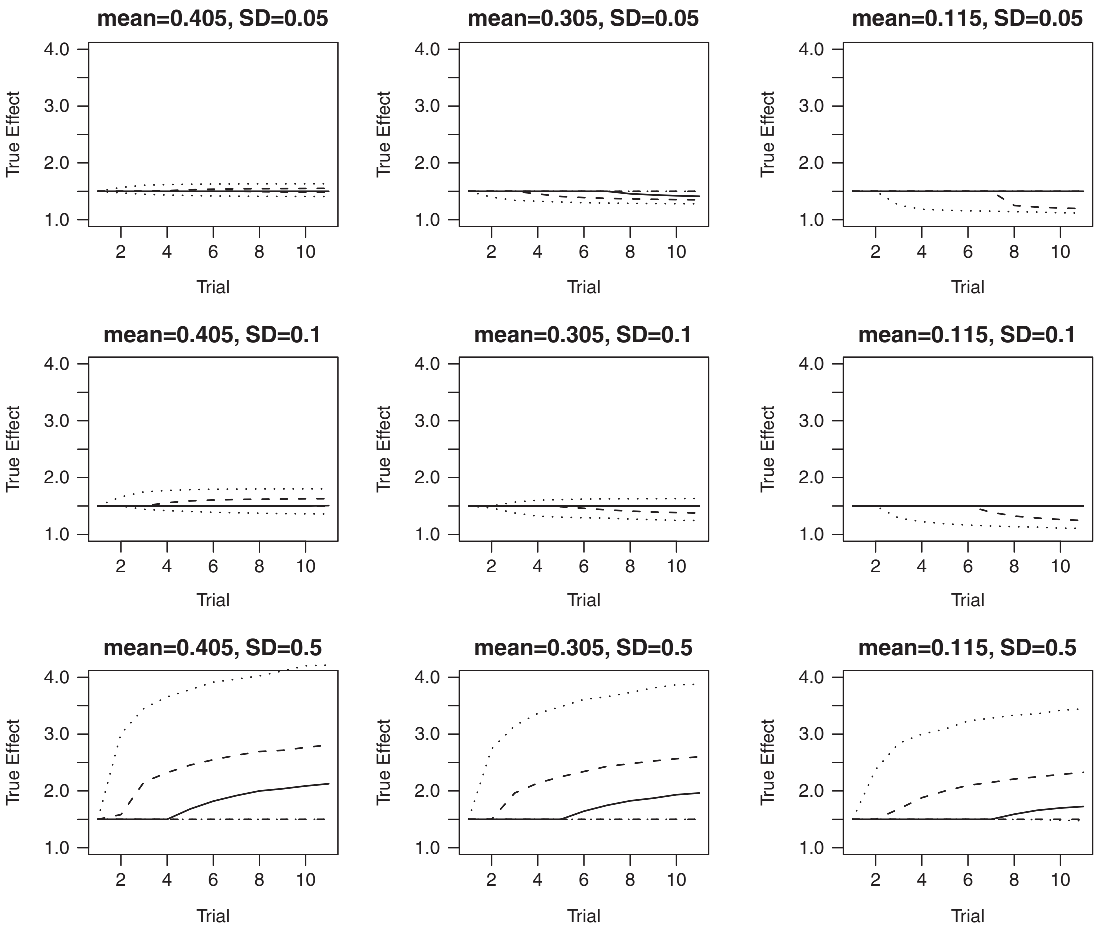

**Figure 2**. Median, central 50 per cent, and central 90 per cent of approved therapies over time. Each trial had 100 observed events.

**Table IV**. Results from Study 2 when susceptible proportion changes randomly.

| Mean($\theta_{PN}$) | SD($\theta_{PN}$) | | $n=100$ | $n=376$ | $n=500$ |
|---|---|---|---|---|---|
| 0.405 | 0.05 | Per cent ineffective | 0 | 0 | 0 |
| | | Per cent harmful | 0 | 0 | 0 |
| 0.305 | 0.05 | Per cent ineffective | 0 | 0 | 0 |
| | | Per cent harmful | 0 | 0 | 0 |
| 0.155 | 0.05 | Per cent ineffective | 8.9 | 2.8 | 2.2 |
| | | Per cent harmful | 0.2 | 0 | 0 |
| 0.405 | 0.10 | Per cent ineffective | 0.2 | 0 | 0 |
| | | Per cent harmful | 0 | 0 | 0 |
| 0.305 | 0.10 | Per cent ineffective | 1.4 | 0.3 | 0.4 |
| | | Per cent harmful | 0 | 0 | 0 |
| 0.155 | 0.10 | Per cent ineffective | 13.3 | 3.4 | 3.5 |
| | | Per cent harmful | 3.1 | 0.5 | 0.7 |
| 0.405 | 0.50 | Per cent ineffective | 3.1 | 0.4 | 0.3 |
| | | Per cent harmful | 1.2 | 0.3 | 0.1 |
| 0.305 | 0.50 | Per cent ineffective | 4.0 | 0.5 | 0.4 |
| | | Per cent harmful | 1.7 | 0.2 | 0.3 |
| 0.155 | 0.50 | Per cent ineffective | 4.5 | 0.4 | 0.5 |
| | | Per cent harmful | 2.7 | 0.1 | 0.1 |

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 V. The rates of bio-creep are even higher than seen in Table IV, with ineffective products approved in up to 16.3 per cent of repetitions and harmful products in as many as 4.9 per cent. Interestingly, increasing sample size does not offer the same protection as before. While the rates of bio-creep did decrease with

**Table V**. Results from Study 2 when susceptible proportion decreases steadily over time.

| Mean($\theta_{PN}$) | SD($\theta_{PN}$) | | $n=100$ | $n=376$ | $n=500$ |
|---|---|---|---|---|---|
| 0.405 | 0.05 | Per cent ineffective | 0 | 0 | 0 |
| | | Per cent harmful | 0 | 0 | 0 |
| 0.305 | 0.05 | Per cent ineffective | 0 | 0 | 0 |
| | | Per cent harmful | 0 | 0 | 0 |
| 0.155 | 0.05 | Per cent ineffective | 8.3 | 6.1 | 4.7 |
| | | Per cent harmful | 0.2 | 0 | 0 |
| 0.405 | 0.10 | Per cent ineffective | 0.2 | 0 | 0.2 |
| | | Per cent harmful | 0 | 0 | 0.1 |
| 0.305 | 0.10 | Per cent ineffective | 1.9 | 0.6 | 1.3 |
| | | Per cent harmful | 0.2 | 0 | 0.1 |
| 0.155 | 0.10 | Per cent ineffective | 16.3 | 10.6 | 10.2 |
| | | Per cent harmful | 3.0 | 1.7 | 1.2 |
| 0.405 | 0.50 | Per cent ineffective | 6.6 | 2.5 | 1.8 |
| | | Per cent harmful | 4.3 | 0.8 | 1.2 |
| 0.305 | 0.50 | Per cent ineffective | 7.1 | 2.4 | 1.9 |
| | | Per cent harmful | 4.1 | 1.2 | 0.5 |
| 0.155 | 0.50 | Per cent ineffective | 8.2 | 2.7 | 1.7 |
| | | Per cent harmful | 4.9 | 0.9 | 0.9 |

increasing sample size for all parameter settings, even with 500 events as many as 10.2 per cent of repetitions saw ineffective therapies approved.

*4.2.1. 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. This was done in the context of Study 2, where the effectiveness of the class of the agents under study decreased steadily over time; we present detailed results for the case where mean($\theta_{PN}$)$=0.305$, SD($\theta_{PN}$)$=0.10$, and each trial was stopped when 100 events had been observed.

As can be seen in Table V, ineffective therapies were approved in 1.9 per cent of repetitions, and harmful therapies in 0.2 per cent. No more than one harmful or ineffective therapy was approved in any one repetition; on average, 0.32 per cent of approved therapies were ineffective, and 0.02 per cent 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 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 estimated very precisely, since harmful therapies were approved in only 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 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 log hazard ratio 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.

### *4.3. Study 3*: *choice of active comparator*

Two strategies used to select the active comparator were examined. First, we used the approved therapy with the highest estimated efficacy as the active comparator; this case is denoted 'Best'. For the next set of simulations, we selected the therapy which, if it was chosen as the new standard, would result in the highest chance of a completely ineffective therapy ($\theta_{PN}=0$) being approved, marked as 'Easiest'. This was done in the setting where constancy holds. Table VI presents these results. The rates of approval of ineffective and harmful therapies are similar for the two procedures, suggesting that it is difficult to choose the active comparator in a way that 'games' the system, at least in the setting we studied. The active comparator that provides the easiest path to approval would have low $\theta_{PS}$, high $b_{PS}$, and low $\sigma^2_{PS}$. With $\theta_{PS}$ unknown, however, those factors cannot be optimized simultaneously. High values of $\hat{\theta}_{PS}$ can suggest both high $\theta_{PS}$ and high $b_{PS}$; therapies approved later in a chain of non-inferiority trials will have both high $b_{PS}$ and high $\sigma^2_{PS}$. With these factors off-setting one another, in these simulation settings the method used for selecting the

**Table VI**. 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($\theta_{PN}$) | SD($\theta_{PN}$) | | Best | Easiest |
|---|---|---|---|---|
| 0.405 | 0.05 | Per cent ineffective | 0 | 0 |
| | | Per cent harmful | 0 | 0 |
| 0.305 | 0.05 | Per cent ineffective | 0 | 0 |
| | | Per cent harmful | 0 | 0 |
| 0.155 | 0.05 | Per cent ineffective | 4.5 | 4.5 |
| | | Per cent harmful | 0 | 0 |
| 0.405 | 0.10 | Per cent ineffective | 0 | 0 |
| | | Per cent harmful | 0 | 0 |
| 0.305 | 0.10 | Per cent ineffective | 0.6 | 0.5 |
| | | Per cent harmful | 0 | 0 |
| 0.155 | 0.10 | Per cent ineffective | 6.0 | 6.0 |
| | | Per cent harmful | 1.0 | 1.2 |
| 0.405 | 0.50 | Per cent ineffective | 0.6 | 0.2 |
| | | Per cent harmful | 0.3 | 0.2 |
| 0.305 | 0.50 | Per cent ineffective | 0.5 | 0.2 |
| | | Per cent harmful | 0.1 | 0 |
| 0.155 | 0.50 | Per cent ineffective | 1.1 | 1.2 |
| | | Per cent harmful | 0.3 | 0.3 |

**Table VII**. 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. We implemented the synthesis method to account for the variability of $\hat{\theta}_{PS}$ ('C') or ignored it ('I'). The mean of $\theta_{PN}$ either shifted over time ('Y') or was fixed ('N').

| | | Mean shifts: | N | N | Y | Y |
|---|---|---|---|---|---|---|
| Mean($\theta_{PN}$) | SD($\theta_{PN}$) | Variance: $\hat{\theta}_{PS}$ | C | I | C | I |
| 0.405 | 0.05 | Per cent ineffective | 0 | 0 | 0.2 | 0.3 |
| | | Per cent harmful | 0 | 0 | 0 | 0 |
| 0.405 | 0.10 | Per cent ineffective | 0.2 | 0.2 | 1.5 | 3.6 |
| | | Per cent harmful | 0 | 0 | 0.6 | 1.5 |
| 0.405 | 0.50 | Per cent ineffective | 6.6 | 17.4 | 2.3 | 4.6 |
| | | Per cent harmful | 4.3 | 10.5 | 1.5 | 3.4 |
| 0.305 | 0.05 | Per cent ineffective | 0 | 0 | 23.3 | 44.0 |
| | | Per cent harmful | 0 | 0 | 12.7 | 28.2 |
| 0.305 | 0.10 | Per cent ineffective | 1.9 | 3.3 | 19.2 | 35.2 |
| | | Per cent harmful | 0.2 | 0.2 | 9.9 | 21.6 |
| 0.305 | 0.50 | Per cent ineffective | 7.1 | 18.5 | 3.1 | 8.5 |
| | | Per cent harmful | 4.1 | 11.7 | 2.0 | 5.4 |
| 0.155 | 0.05 | Per cent ineffective | 8.3 | 17.4 | 21.3 | 38.5 |
| | | Per cent harmful | 0.2 | 0.2 | 18.8 | 35.9 |
| 0.155 | 0.10 | Per cent ineffective | 16.3 | 30.7 | 23.3 | 41.9 |
| | | Per cent harmful | 3.0 | 7.9 | 15.6 | 33.0 |
| 0.155 | 0.50 | Per cent ineffective | 8.2 | 19.2 | 6.9 | 13.8 |
| | | Per cent harmful | 4.9 | 13.5 | 3.9 | 9.3 |

active control does not appear to affect the rates of bio-creep to the same degree as do violations of the constancy assumption.

### *4.4. Study 4: accounting for variability*

We next examined the impact of ignoring the variability of $\hat{\theta}_{PS}$. We compared this with the synthesis method approach used in Studies 1–3. This was done with the susceptible proportion of the study population decreasing steadily from trial to trial. As expected, ignoring the variability of $\hat{\theta}_{PS}$ led to disastrous results, as can be seen in Table VII. In the worst case setting, ineffective therapies were approved in as many as 31 per cent of repetitions, and harmful therapies in 14 per cent.

### 4.5. *Study 5: trends in treatment effect*

We investigated how changes in the distribution of $\theta_{PN}$ over time affect the chance of bio-creep occurring. When we allowed this distribution to change over time, $\theta_{PN}$ was centered around either $\theta_{PS}$, $\theta_{PS}-0.10$, or $\theta_{PS}-0.25$, 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 VII for simulations where the synthesis method was used, and when the variability of $\hat{\theta}_{PS}$ was ignored. Allowing the distribution of $\theta_{PN}$ to change over time affected the bio-creep rates differently, depending on the variance in the distribution of $\theta_{PN}$. At the highest standard deviation studied, 0.50, shifting the mean of $\theta_{PN}$ 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$(\theta_{PN})=0.305$ and SD$(\theta_{PN})=0.05$, when ignoring the variability of $\theta_{PS}$, but with a fixed mean of the distribution of $\theta_{PN}$, no ineffective therapies were approved. When the mean shifted over time, ineffective therapies were approved in 44 per cent of the repetitions. Even when the variance of the active control was accounted for correctly, ineffective therapies were approved in 23 per cent of the repetitions.

## 5. Discussion

These results demonstrate the importance of using the methodology appropriate for the non-inferiority setting when conducting a non-inferiority clinical trial, although high rates of bio-creep are possible even with the correct methodology. The variability inherent in estimating the effect of the active comparator should be considered and accounted for; failing to do so resulted in higher rates of bio-creep in the simulations settings considered here.

We examined how the choice of active comparator 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 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 this setting, we were not able to find a strategy for selecting the active control that greatly influenced the rates of bio-creep.

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 test of the effect of the first approved drug versus placebo is just barely significant, doing a trial with 50 per cent 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. Of course, in practice, one can never know whether an estimate of treatment effect is high because the therapy works, or due to chance. Comparing the observed performance of a drug in historical trials with its observed effect in the current trial may provide some insight into this question.

There are many different factors that may influence the incidence of bio-creep; we examined only a few here. For example, we defined non-inferiority as 50 per cent retention. Clearly, selecting a higher percentage retention will lead to lower rates of bio-creep, but with a corresponding loss of power to approve effective therapies. One may also be interested in the rates of bio-creep seen using a fixed margin rather than the percent retention formulation used here. Other methods could be used instead of the synthesis method approach we selected. For example, we expect a 95–95 approach, with its built-in conservatism, may result in lower rates of bio-creep, again at the cost of a reduction in power.

From the results of Studies 1–5, it is apparent that violations of the constancy assumption are one of the most important potential source of bio-creep. When the constancy assumption held, as long as appropriate analysis techniques were used in conjunction with a reasonable non-inferiority margin, bio-creep appeared to be rare in the settings used in our simulations. However, when this assumption was violated, rates of bio-creep could be high. As this assumption is typically untested, these results are of concern. Further consideration should be given to developing techniques to identify such violations, and for handling them when they are detected.

Even with a constant underlying treatment effect, the bias in the estimate of the first drug resulting from the screening process could lead to regression to the mean. That is, there might be differences in the observed effect of the first active control from one trial to the next. With the reasonably efficacious first drug and large sample sizes we used, this should not be much of a problem. However, with a less effective first drug, it is plausible that this could happen more frequently.

## Acknowledgements

This research was supported in part by the NHLBI through training grant T32 HL007183-33 to the University of Washington.

## 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**:169–186. DOI: 10.1002/sim.1425.
2. Fleming TR. Current issues in non-inferiority trials. *Statistics in Medicine* 2008; **27**:317–332. DOI: 10.1002/sim.2855.
3. ICH E-9. *International Conference on Harmonisation—Statistical Principles for Clinical Trials*. 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**:263–273. DOI: 10.1081/BIP-120037178.
5. Wang SJ, Hung HMJ. Assessing treatment efficacy in noninferiority trials. *Controlled Clinical Trials* 2003; **24**:147–155. DOI: 10.1016/S0197-2456(02)00304-5.
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**:651–659. DOI: 10.1081/BIP-100101201.
7. CBER/FDA Memorandum. 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. DOI: 10.1002/sim.1137.
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**:213–225. DOI: 10.1002/sim.1315.
10. Wang SJ, Hung HMJ. TACT method for non-inferiority testing in active controlled trials. *Statistics in Medicine* 2003; **22**:227–238. DOI: 10.1002/sim.1316.
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 H-H. Design and analysis of non-inferiority mortality trials in oncology. *Statistics in Medicine* 2003; **22**:239–264. DOI: 10.1002/sim.1400.
13. Shepherd FA, Dancey J, Ramlau R, Mattson K, Gralla R, O'Rourke M, Levitan N, Gressot L, Vincent M, Burkes R, Coughlin S, Kim Y, Berille J. 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 FA, Fossella FV, Pereira JR, De Marinis F, von Pawel J, Gatzemeier U, Tsao TCY, Pless M, Muller T, Lim HL, Desch C, Szondy K, Gervais R, Shaharyar Manegold C, Paul S, Paoletti P, Einhorn L, Bunn PA. Randomized phase III trial of pemetrexed versus docetaxel in patients with non-small-cell lung cancer previously treated with chemotherapy. *Journal of Clinical Oncology* 2004; **22**:1589–1597. DOI: 10.1200/JCO.2004.08.163.
15. Fleming TR, Harrington DP. *Counting Processes and Survival Analysis*. Wiley: New York, 1991.
16. Fisher LD. Self-designing clinical trials. *Statistics in Medicine* 1998; **17**:1551–1562. DOI: 10.1002/(SICI)1097-0258(19980730)17:14⟨1551::AID-SIM868⟩3.0.CO;2-E.