HHS Public Access
Author manuscript
Stat Biopharm Res. Author manuscript; available in PMC 2016 January 01.

Published in final edited form as:
Stat Biopharm Res. 2015 January 1; 7(1): 12–24. doi:10.1080/19466315.2014.1002627.

A simulation study evaluating bio-creep risk in serial non-inferiority clinical trials for preservation of effect

K. Odem-Davis and **T. R. Fleming**
University of Washington, Seattle, WA

Abstract

In non-inferiority trials, acceptable efficacy of an experimental treatment is established by ruling out some defined level of reduced effect relative to an effective active control standard. Serial use of non-inferiority trials may lead to newly approved therapies that provide meaningfully reduced levels of benefit; this phenomenon is called bio-creep. Simulations were designed to facilitate understanding of bio-creep risk when approval of an experimental treatment with efficacy less than some proportion of the effect of the active control treatment would constitute harm, such as when new antibiotics that are meaningfully less effective than the most effective current antibiotic would be used for treatment of Community-Acquired Bacterial Pneumonia. In this setting, risk of approval of insufficiently effective therapies may be great, even when the standard treatment effect satisfies constancy across trials. Modifiable factors contributing to this manifestation of bio-creep included the active control selection method, the non-inferiority margin, and bias in the active control effect estimate. Therefore, when non-inferiority testing is performed, the best available treatment should be used as the standard, and margins should be based on the estimated effect of this control, accounting for the variability and for likely sources of bias in this estimate, and addressing the importance of preservation of some portion of the standard's effect.

1 Introduction

Evaluation of the efficacy of an experimental treatment (Experimental) in the setting of an active controlled clinical trial based on comparison to an active control standard treatment (Standard) may be based on ensuring that Experimental is at least not "too much" worse than Standard in exchange for improvements in side effects, ease of administration, or cost. Trials designed to make this assessment are called non-inferiority (NI) trials and the level of acceptable loss is the NI margin, Δ.

NI trial design and conduct have been scrutinized in recent years. In response to public concerns about the inappropriate use of NI trials, the United States Food and Drug Administration (FDA) issued a draft guidance document on NI in March 2010, and the Government Accountability Office (GAO) released a report in July 2010 assessing the FDA's evaluation of NI trial results as evidence for new drug approval (FDA 2010; GAO 2010). The latter report identified 43 new drugs, among 175 total, that were submitted to FDA for review between 2002 and 2009 based at least in part on NI trial results. One concern discussed in the GAO report was the risk of bio-creep, which was defined in the report as "a concern that successive generations of drugs approved based on non-inferiority

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

trials, with the active control changing in each new generation, could lead to the adoption of decreasingly effective drugs”.

D’Agostino et al. (2003) suggest that bio-creep may be reduced if the standard is always the “best” comparator. Unfortunately, investigators are not required to use this criterion for selecting the active comparator, and it may be perceived not to be in their interest to do so. Furthermore, data may not be available for investigators to directly compare therapies to each other in order to select the best among them in the setting of the planned NI trial.

Everson-Stewart and Emerson (2010) assessed bio-creep by evaluating the rate of approving therapies that are no better than, or worse than, placebo, i.e. “ineffective” or “harmful” treatments (Everson-Stewart & Emerson 2010). However, when NI trials are designed to preserve a percentage of effect of the standard treatment, the rate of approval of “insufficiently effective” treatments not meeting this requirement should also be considered when evaluating bio-creep risk. Peterson et al. (2010) argued against preservation of effect, calling standards based on percent preservation “inherently arbitrary and lacking in objective clinical or scientific justification.” By contrast, Fleming et al. (2011) maintained that a stricter standard is clinically justified when randomization to placebo is considered unethical due to availability of effective therapies, especially when the active comparator has demonstrated efficacy with respect to irreversible morbidity or mortality. Although the debate about preservation of effect for regulatory approval remains active, we will address the scientifically well motivated setting where investigators wish to control the risk of approval of insufficiently effective therapies at a level of 0.025. In that case, the risk of approval of ineffective therapies would be much, much lower.

The margin Δ may be formulated based on an estimated effect of Standard compared to placebo incorporating statistical uncertainty about this estimate, or Δ may be a fixed value. Common methods for formulating Δ are the Synthesis(p) margin and the 95–95(p) margin (Fleming 2008; Rothmann et al. 2003). The former is based on imputing the effect of Experimental relative to placebo by “synthesis” of the historical effect estimate of Standard relative to Placebo and the NI trial estimate of Experimental to Standard effect, and testing the hypothesis to rule out that Experimental preserves less than a proportion p of the effect of Standard relative to placebo (Rothmann et al. 2003). The latter compares the upper limit of the 95% confidence interval (CI) for the estimate of a parameter for the effect of Experimental relative to Standard to a fraction p of the lower limit of the 95% CI for placebo to Standard effect (FDA 2010; Fleming 2008). Though other combinations of confidence intervals may be used, the two 95% confidence intervals approach is most common, leading to the “95–95” margin name.

One uncertainty related to use of the historical estimate of Standard effect is validity of the “constancy assumption”, that the efficacy of Standard relative to placebo in the setting of historical trial(s) is equal to the true efficacy of Standard relative to placebo in the setting of the NI trial. When this assumption is violated, use of the historical estimate introduces bias. Non-constancy due to the influence of effect modifiers, changes in supportive care, changes in disease etiology, changes in endpoint definitions, and changes in endpoint ascertainment procedures, as well as other sources of bias in the estimate of the effect of Standard relative

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

to placebo lead to inflated false positive rates and therefore contribute to bio-creep (Everson-Stewart & Emerson 2010; Fleming 2008; Fleming et al. 2011). Margins which account for these likely sources of bias have been proposed (Fleming 2008; Rothmann et al. 2003; Davis 2010; Odem-Davis & Fleming 2013). The 95–95(p) margin provides indirect adjustment for such bias and results in a reduced probability of conclusion of NI as compared to the Synthesis(p) approach (Fleming 2008), and has been advocated by FDA to account for uncertainties related to use of the estimate of Standard effect relative to placebo from prior studies in the setting of the NI trial (FDA 2010).

An adjusted Synthesis margin was proposed by Rothmann (2003) to simultaneously address the bias in the estimated effect of Standard in the NI trial and to preserve a fraction p of this effect. It is based on multiplying the historical estimate, and the corresponding standard error, by an attenuation factor in development of the test statistic for the hypothesis of preservation of effect (Rothmann et al. 2003). We will call this margin the Synthesis(λ, p) margin, where p is the proportion of effect to be preserved and $1 - \lambda$ is the corresponding attenuation factor. Suppose the bias in the historical estimate of the effect of Standard relative to placebo is known and equal to λ times the expectation of the estimate. Then, the Synthesis(λ, p) method appropriately adjusts the margin for bias (Rothmann et al. 2003; Davis 2010; Odem-Davis & Fleming 2013). Another adjusted Synthesis margin was recently proposed to address uncertainty about λ in development of the Synthesis(λ, p) margin (Davis 2010; Odem-Davis & Fleming 2013). In the absence of specific information about the degree of this uncertainty, the margin is based on multiplying the historical estimate by the attenuation factor, $1 - \lambda$, without also attenuating the corresponding standard error (Davis 2010; Odem-Davis & Fleming 2013). We will call this margin the Bias-adjusted(λ, p) margin. Comparison to the 95–95 margin for examples obtained from a systematic review of trials discussed at FDA advisory committee meetings provided a context for the choice of λ in the absence of more direct estimates; Using this reference, we take λ equal to 0.3 (Davis 2010; Odem-Davis & Fleming 2013).

Many factors may influence the risk of bio-creep, including:

- The method for choosing the NI margin
- The method for choosing Standard, including influence of publication bias and random high bias
- The efficacy of Standard relative to placebo in the historical setting
- The method for estimation of the effect of the Standard relative to placebo
- Bias in the estimate of Standard relative to placebo in the NI trial setting, including violation of the constancy assumption
- Distribution of effects of Experimental treatments
- NI trial sample size
- Outcome scale

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Among the factors listed above, investigators have control over the method for choosing the NI margin, method of estimation of the effect of Standard relative to placebo, method for choosing Standard, the NI trial sample size, and the outcome scale.

2 Methods: Simulation Studies

We designed simulation studies to assess the risk of bio-creep after serial use of NI trials designed for preservation of effect, comparing different NI margin methods, Standard selection methods, and distributions of Experimental treatments in three studies. Simulations for all studies were motivated by the setting of anti-infective drugs for treatment of Community Acquired Bacterial Pneumonia (CABP), with a first Standard being sulfonamides or penicillin.

All superiority and NI tests were conducted using the difference of proportions outcome scale, commonly used in regulatory studies of anti-infective drugs. Each simulation began with a Fixed Effects meta-analysis of three superiority trials with sample sizes of 150 subjects per arm comparing the first Standard to placebo, where only trials with statistically significant results were included. Clinical success rates were approximately based upon data regarding efficacy of sulfonamides or penicillin with respect to mortality as compared to no specific treatment from non-randomized studies, with true placebo success rate of 0.5 and the first Standard success rate of 0.8 in the setting of the “historical” trials for approval of the first Standard (Fleming & Powers 2008).

For the first NI trial, true success rates for subjects on new Experimental treatments were drawn from a distribution defined by s_E = .5*Z + .45 where Z was distributed Beta(a,b), with shape parameters a and b, the resulting distribution will be referred to as the Scaled Beta(a,b) distribution. This scaling resulted in success rates in the range of 0.45 to 0.95, allowing for a small number of harmful treatments compared to the event rate of 0.5 on placebo.

Investigators evaluating treatments for CABP had historically used a fixed margin of 0.2 on the difference in success proportion scale, while a margin less than or equal to 0.1 was advocated in 2008 (Fleming & Powers 2008). Alternatively, though rescinded in 2001 (FDA 2001), a step-function defined in a 1992 Points to Consider (PTC) FDA briefing document (FDA 1992) has been used to develop margins for NI trials in the area of Anti-Infective Drugs in trials discussed as recently as 2009. The PTC margin depends on the maximum of the observed Experimental and Standard success rates in the NI trial (Rothmann et al. 2011). If this maximum is greater than or equal to 90%, then the margin equals 10%. If the maximum is between 80 and 89%, then the margin is 15%. Finally, for a maximum less than 80%, the margin is 20%. We also considered the 95–95 and Synthesis margins, which have been advocated in the regulatory setting as previously described (Fleming 2008; Rothmann et al. 2003; Davis 2010; Odem-Davis & Fleming 2013). The number of successes in each arm was sampled from the Binomial distribution, and the resulting estimates of success rates and corresponding standard errors were used to compute Δ according to each of the methods listed below:

- Points to Consider (PTC)

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

- Synthesis($\lambda = 0$, $p = 0.5$)
- 95–95($p = 0.5$)
- Synthesis($\lambda = 0.3$, $p = 0.5$)
- Bias-adjusted($\lambda = 0.3$, $p = 0.5$)
- Fixed Margins of 0.1, 0.15, 0.2

The same margin method was used for each iteration in sequence of ten NI trials. Standard for the next trial was chosen according to one of the rules: “Worst”, “Recent”, and “Best”. The “Worst” treatment was defined as the treatment with the smallest estimated effect among treatments approved in previous iterations of trials. The Recent treatment was the most recently approved. The Best treatment was defined as the treatment with the greatest estimated effect among treatments approved in previous iterations of trials. This process was completed for 10,000 iterations.

Margins that depend on the estimated effect of Standard, such as Synthesis and 95–95 margins, were updated each time a new Standard was selected. For example, suppose that the success rate of the Experimental treatment from the first NI trial is $\hat{s}_{E,1}$, the Standard success rate from the NI trial is $\hat{s}_{S,1}$, and the Standard and placebo success rates from the prior trial(s) are $\hat{s}_{S,0}$ and $\hat{s}_{P,0}$ respectively. Then, if the Experimental from NI Trial 1 becomes Standard for NI Trial 2, the Synthesis and 95–95 margins for Trial 2 are based on the estimate $\hat{s}_{E,1} - \hat{s}_{S,1} + \hat{s}_{S,0} - \hat{s}_{P,0}$, and the corresponding variance, equal to the sum of the variances for the success rate estimates.

Experimental treatment success rates were sampled according to Scaled Beta distributions a=3 and with different values for b. Plots of the densities for these distributions with b ranging between 1.5 and 10 are shown in Figure 1. When b=2, the corresponding scaled distribution has a mean of 0.75 and mode of 0.78. This setting, with an “Effective” pool of Experimental treatments, was intended to model a scenario where new treatments are from the same class as the initial Standard and are only slightly less effective on average. When b=9, the corresponding scaled distribution has a mean of 0.575 and mode of 0.55. This set, with an “Ineffective” pool of Experimental treatments, was intended to model a scenario where new treatments are from a different class than the initial Standard and are only marginally effective on average.

We also considered a scenario where effects of new Experimental treatments were decreasing over time, i.e., a “Decreasingly Effective” pool of Experimental treatments. In these simulations, a scaled Beta(a=3, b=b[i]) was used, where i is the iteration of NI trial and the vector b is (1.5, 2, 3, 4, 5, 6, 7, 8, 9, 10). One example of decreasing efficacy may be consecutive comparisons of treatments using the same or similar drug(s) administered at decreasing doses. Note that this scenario is not appropriate for modeling development of resistance, since resistance involves violation of the constancy assumption whereby previously approved therapies used as Standard also have reduced efficacy in later NI trials.

Per arm sample sizes in NI trials were computed based on the formula $[\Phi^{-1}(\alpha) + \Phi^{-1}(\beta)]^2[2p(1-p)](-\Delta)^2$, where Φ^{-1} is the inverse of the Standard Normal cumulative

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

distribution function, α was set to 0.025 and β was set to 0.1 for controlling the NI trial false positive and false negative error rates respectively (relative the margin specified at design stage). Then, a minimum of 150 per arm and a maximum of 400 per arm were imposed to avoid simulating unrealistic samples sizes relative to those typically seen in the motivating setting. This choice of range is approximately based upon a study by Higgins et al. (2008), in which authors reviewed all comparative studies that were submitted to the Office of Antimicrobial Products to support indications for CAP in the 8 years prior to 2008. They found that among the 7 studies submitted, sample sizes ranged in from about 300 to 500 subjects for oral antibiotics and from about 300 to 700 for intravenous antibacterial drugs.

Under each combination of settings, results were obtained regarding the average success rate on approved treatments, the risk of approving a treatment with success rate equal to Placebo, “False Positive Rate (Ineffective)”, and the risk of approving a treatment preserving half of the effect of the first Standard in the historical setting, “False Positive Rate (Insufficiently Effective)”.

In Study 1, simulations were conducted under the constancy assumption. In Study 2, constancy was violated due to greater success rates on both (unobserved) placebo and (observed) Standard arms, in a manner reducing the mean effect of Standard relative to placebo in the setting of the NI trial as compared to in the “historical” setting under which prior estimates of Standard effect were obtained. This model was intended to represent differences between the characteristics of subjects in the “historical” superiority trials and the NI trials, with healthier patients (or patients with better supportive care) for whom the treatment was less effective enrolled in the NI trials. The higher unobserved placebo success rate in the NI trial compared to the prior trials is not a placebo effect; rather, the healthier patients in the NI trial are more likely to recover from infection (e.g., due to a better immune responses), even when untreated. In Study 3, constancy was violated due to a reduction over time in the success rate among subjects on the (observed) Standard arm in the NI trial, while the (unobserved) Placebo arm remained constant. This model was intended to represent increasing the proportion of enrolled patients infected with pathogens resistant to the Standard treatment. Additional details for each study are provided in the subsequent sections.

2.1 Study 1: Under Constancy

In the first study, success rates on placebo and for each therapy satisfied constancy across non-inferiority trials. Bio-creep risk was compared across standard selection methods by assessment of the “False Positive” risk of approval of Insufficiently Effective or Ineffective treatments, under each of the margin methods, using models for Effective, Ineffective, and Decreasingly Effective pools of Experimental treatments.

2.2 Study 2: Imbalances in Patient Characteristics

The second study simulated a setting where rates on placebo and Standard may both differ between historical and NI trials due to imbalances in patient characteristics that are effect modifiers. Here, placebo and Standard mean success rates were scaled proportionally assuming that the NI trial means, s_{NI}, were equal to $\Lambda(1) + (1 - \Lambda)s_H$, where s_H represents

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the corresponding historical mean and Λ represents the non-constancy bias proportion. In the CABP setting where young non-bacteremic patients on both placebo and Standard have probability approximately equal to 1 of surviving the observed time and where older bacteremic patients have probability of s_H of surviving the same observation time, this model is equivalent to a trial that includes (1−Λ) of older bacteremic patients and Λ of younger patients. Simulations were based on value of Λ equal 0.5. In the CABP example, true values for Λ may actually be much greater than 0.5 when trials are conducted primarily in younger, non-bacteremic populations.

Success rates on Experimental treatments drawn from the Scaled Beta distributions described above were also scaled to the NI trial setting according to the formula $s_{E,NI} = s_{E,H} + \Lambda(1 - s_{E,H})$. Then, observed rates were based on draws from the corresponding Binomial(n, p = $s_{E,NI}$) distribution.

2.3 Study 3: Development of Resistance

The third study simulated a setting of development of resistance to anti-infective drugs, where Standard effect differs between historical and NI trials but where success rates on placebo remain constant. For this scenario, simulations were conducted using the Scaled Beta(3,2) and Scaled Beta(3,9) distributions for Experimental treatment effects in the setting of the historical trial, assuming no resistance until the first NI trial. Resistance development on both Standard and Experimental treatments was modeled by a formula for “gradually developing resistance”.

The formula for resistance was $s_{NI} = s_H - ((k - 0.5)/10)(s_H - s_{PLA})$ where k is the NI trial iteration, s_{NI} is the success rate in the NI trial setting, s_H would be the mean success rate for Standard or Experimental in the historical setting, and $s_{PLA} = 0.5$ is the mean success rate on placebo in both the historical and NI settings. One example where this model may apply is in antibiotics for the treatment of staph infection with increasing prevalence of MRSA infections. Then, the percentage of the study populations with MRSA infections increased from 5% to 95% from the first to the tenth NI trial.

Success rates on Experimental treatments considered for Standard drawn from the Scaled Beta distributions described above were also scaled to the NI trial setting according to the formulas for gradually developing resistance, assuming that no new treatments introduced were effective for treatment of infections with resistant strains.

3 Results

For each of the studies, Insufficiently Effective and Ineffective False Positive Error rates across simulations were compared between study characteristics and margin methods. Mean success rates on newly approved treatments at each trial iteration were also compared.

3.1 Study 1: Under Constancy

Results from Study 1, with simulations under the constancy assumption, are summarized in Figures 2 through 4. The first two plots respectively show the False Positive Rates for risk of approving an Insufficiently Effective or Ineffective therapies at each NI trial iteration under

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the corresponding margin and Standard selection method (Best, Worst, or Recently Approved) when the pool of Experimental treatments are Effective, Ineffective, or Decreasingly Effective.

In Figure 2 we see that Fixed 0.2 and PTC margins have high Insufficiently Effective treatment false positive error rates of approximately 0.17 and 0.09 respectively at the first iteration of NI trials under all scenarios. In subsequent trials for these margins, the Standard selection method is very influential. Under the Effective Pool, rejection rates increase significantly when the worst approved Standard is iteratively chosen, but approach the one-sided 0.025 rate after several generations of trials when the best approved Standard is repeatedly selected. Use of the most recently approved Standard results in an approximately level rate of Insufficiently Effective False Positive Error rates under an Effective pool of Experimental treatments, but these rates increase substantially under Ineffective or Decreasingly effective pools of treatments. The greatest Insufficiently Effective False Positive Error risk was under a Fixed 0.2 margin and decreasingly effective pool of treatments, with greater than 70% risk of approval of treatments half as effective as the first Standard.

As expected, since the true effect of the first Standard is 0.3 (equal to success rate of 0.8 on Standard minus 0.5 on placebo), the Fixed 0.15 margin has an approximate 0.025 Insufficiently Effective False Positive rate at the first iteration of NI trials. The Synthesis margin also has a 0.025 Insufficiently Effective False Positive Rate at the first iteration, suggesting that the method for estimating the first Standard using only trials achieving statistical significance does not lead to substantial random-high bias in this setting where Standard is highly effective and historical trials are sufficiently large. Then, even for these margins which are appropriate since the constancy assumption holds, “worst” Standard selection method results in a meaningful increase in the likelihood for approval of therapies which preserve less than half the effect of the first Standard after several generations of NI trials.

The “Recently Approved” Standard selection method is neutral with respect to bio-creep in the top panel of Figure 2 where new Experimental treatments considered for Standard are only slightly less effective than the first Standard on average. However, this selection method is subject to increases in false positive error rates in later iterations of NI trials when new Experimental treatments to be considered for Standard are only marginally effective on average or are decreasing in efficacy over successive trials as shown in the mid- and bottom-right panels.

In these simulations, the unadjusted Synthesis($\lambda = 0$, $p = 0.5$) margin resulted in greater than the desired 0.025 one-sided false positive error rate under the worst Standard selection model regardless of the pool of treatments, and under the recently approved Standard selection model when the pool of treatments was Ineffective or Decreasingly Effective. Use of the 95–95(0.5), Synthesis($\lambda = 0.3$, $p = 0.5$), Bias-adjusted($\lambda = 0.3$, $p = 0.5$), and Fixed 0.1 margins resulted in low risk of inflated Insufficiently Effective false positive error even after several generations of trials regardless of the Standard selection method or pool of Experimental treatments.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

The fact that the Fixed 0.1 margin is among the more “conservative” margins depends entirely on the magnitude of the true effect of Standard relative to placebo. The Fixed 0.1 margin would not be comparable to the 95–95(0.5) or adjusted Synthesis margins in a study with a smaller Standard effect size. Also, studies with a greater coefficient of variation, either due to a smaller Standard effect size or due to smaller studies, would result in a greater difference between the Bias-adjusted($\lambda = 0.3$, $p = 0.5$) and Synthesis($\lambda = 0.3$, $p = 0.5$) margins, with Bias-adjusted($\lambda = 0.3$, $p = 0.5$) margin leading to lower false positive error rates than Synthesis($\lambda = 0.3$, $p = 0.5$) than delineated here. Finally, note that Synthesis(0,0.5) margin remains closer to 0.025 error compared to the Fixed 0.15 margin when experimental treatments are drawn from an effective pool, due to the fact that when the best treatment is used down-stream Synthesis margins are greater than 0.15 (due to increasingly effective Standard treatments) and when the worst treatment is used down-stream Synthesis margins are less than 0.15 (due to decreasingly effective Standard treatments).

Figure 3 shows the Ineffective False Positive Error rates, corresponding to risk of approving Experimental treatments with a true success rate of 0.5, which is equivalent to placebo under constancy. When the hypothesis test of interest requires preservation of some proportion of the effect of Standard relative to Placebo, false approvals of therapies no better than Placebo should be controlled at a rate much less than the usual one-sided α of 0.025. A tolerance of three standard errors would require a false approval of approximately 0.001. The dashed and solid horizontal lines indicate one-sided 0.025 and 0.001 False Positive thresholds respectively.

When the “best” therapy was used as Standard, the risk for approval of ineffective therapies was negligible, even after many generations of NI trials and regardless of the distribution of Experimental therapies to be considered for use as Standard. However, when other methods for choosing Standard were used, the risk of bio-creep was great when the Fixed 0.2 or PTC margins were used. For example, when a fixed margin of 0.2 was used along with selection of the “worst” approved Standard, the approval rate for ineffective therapies exceeded 0.001 at the second iteration of NI trials and reached a high of 0.3 when the Fixed 0.2 margin was used under the decreasingly effective pool of treatments.

Next, success proportion means on approved treatments are shown in Figure 4. When treatments are on average as effective as, but not substantially better than, the first Standard, the mean success rate remains approximately constant over serial NI trials except under fixed 0.2 or PTC margins when choosing the worst standard as seen in the top panel of Figure 4.

When the pool of treatments considered for standard are only marginally effective (“Ineffective Pool”), approved treatment means under the use of the worst treatment as Standard or under the most recently approved standard decrease over successive NI trials under the Synthesis($\lambda = 0$, $p = 0.5$), Fixed 0.15, PTC, and Fixed 0.2 margins, with the greatest decrease under the Fixed 0.2 margin. Here, use of the best standard is protective from increasing rates of approvals of new treatments under the ineffective pool, and the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

newly approved success means are constant, though less than the efficacy of the first standard with success rate of 0.8.

Under the “Decreasingly Effective” pool of treatments, “Ineffective” and “Insufficiently Effective” False Positive Error rates increase to a level above that under the “Ineffective” pool when the “worst” standard selection method is used. Referring to Figure 1, note that the distribution of treatment success rates begins with a greater mode at the first trial as compared to the “Effective” pool and is drawn from the same distribution of success rates at the ninth trial as compared to the “Ineffective” pool. In Figure 4, we see that approved treatments success rate means under the “Decreasingly Effective” pool of treatments and using the “worst” standard selection method fall below levels under the “Ineffective” pool of treatments by the ninth iteration of trials, even though approved treatment success rate means are greater at the first NI trial iteration. This phenomenon is explained by the greater rate of approvals in later trials under the “Decreasingly Effective” pool as compared to the “Ineffective” pool when the “worst” treatment is used as standard (data not shown).

3.2 Study 2: Non-Constancy due to Imbalances in Patient Characteristics

Results from simulations for scenarios based on non-constancy bias of $\lambda = .5$ are provided in Figure 5. In these scenarios, the risk of approval of insufficiently effective and ineffective treatments is substantial, even at the first NI trial. Margins designed to adjust for the non-constancy bias and smaller margins such as the Fixed 0.1 are still somewhat protective, as is the selection of the best approved treatment as Standard.

Use of the best approved treatment as Standard helps decrease bio-creep risk under the model assuming that the pool of experimental treatments is effective on average. However, when the treatment pool is ineffective, both “Insufficiently Effective” and “Ineffective” False Positive Error rates remain constant, or even increase under some margin methods (e.g., Fixed 10%), under use of the best Standard.

Under the “Worst” standard selection method, bio-creep risk is greatest, in terms of risk of approval of “Insufficiently Effective” and “Ineffective” treatments. In this setting, use of smaller margins such as the Fixed 0.1, Synthesis($\lambda = 0.3$, $p = 0.5$), or Bias-adjusted($\lambda = 0.3$, $p = 0.5$) offer some protection, though even under these margins, ‘Insufficiently Effective” false positive error rates reach above 10%.

Next, the pool of experimental treatments is especially influential if the strategy of most recently approved Standard is used. The risk of approving ineffective treatments is much greater when treatment means are drawn according to the “Ineffective” pool distribution. This difference is shown in the bottom right panels of Figure 5. However, even under a pool of effective treatments and using the Best standard, use of a Fixed 20% margin leads to a 28% risk of approving an ineffective therapy.

3.3 Study 3: Non-Constancy due to Development of Resistance

False positive rates for NI test approval of ineffective treatments under the model for “gradually developing resistance” are shown in Figure 6. In these settings, large margins such as the PTC and Fixed 0.2 margins, led to substantial risk for approval of insufficiently

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

effective and ineffective therapies after a few iterations of NI trials. For example, under the “worst” selection method, the Fixed 0.2 margin led to an ineffective treatment approval rate of 0.17 and 0.19 at the fourth NI trial iteration under models for “Effective” and “Ineffective Pools” of experimental treatments respectively.

After a series of five trials in this setting, when 45% of subjects have resistant strains, Ineffective False Positive Error rates were inflated for 95–95 and larger margins. In contrast to the setting of imbalances in patient characteristics, risk for approval of ineffective therapies under use of smaller margins such as the Fixed 0.1 and Bias-adjusted(0.3,0.5) was inflated after six trial iterations and reached rates of 30 to 40% at the tenth generation when 95% of patients were infected with resistant strains. Insufficiently effective error rates were inflated for all margins in all scenarios for after four trials. Selection of the “best” treatment was not protective against bio-creep in this study.

4 Discussion

When large margins are used, risk of approval of “Insufficiently Effective” and “Ineffective” treatments may be great even when the best approved treatment is used as Standard. These simulations confirm that bio-creep risk is minimal under certain conditions: when NI margins are based on estimates of Standard effect relative to Placebo, incorporating uncertainty and preservation of effect, under the constancy assumption, and when the best approved treatment is used as Standard (D’Agostino et al. 2003; Everson-Stewart & Emerson 2010). However, under fixed margins that are large or under Synthesis margins that do not incorporate both uncertainty and preservation of effect, risk of approval of insufficiently effective or ineffective treatments may be substantial even under constancy, especially when the standard is not the best approved treatment.

As under any simulation study, interpretation of results depends upon how well the model reflects reality. Standard selection methods were assumed constant across all NI trials in a round of simulations, whereas in truth investigators evaluating new treatments may use different methods for choosing Standards across a series of NI trials. Furthermore, we simulated a “single trial” per iteration, whereas two trials are often required for regulatory approval of experimental therapies not already approved for another indication. However, the application of a minimum sample size of 150 per arm is just below the 340 total subjects that would be required for a combination of two trials each with fixed margins of 0.2, 90% power, one-sided 0.025 alpha, and control success probability of 0.8 (85 subjects per arm, per trial). Additional simulations evaluating the effect of a regulatory two-trial requirement on bio-creep risk is an area for future study.

When the NI trial design (e.g., sample size based on NI-trial power calculations) depends on results from the prior trials estimating the Standard effect, or when two or more trials are evaluated relative to the same Standard, then the distribution of estimates for active comparator efficacy from prior trials and for the experimental treatment from the NI trial will be correlated. Methods for addressing correlation in NI trial evaluations have been proposed but were not used in this study (Kang & Tsong 2005; Rothmann 2005).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

As previously discussed, bio-creep is of particular concern when the constancy assumption is violated and when new experimental therapies decrease in efficacy over time (Everson-Stewart & Emerson 2010). Validity of the constancy assumption cannot directly be assessed without a Placebo arm in the NI trial. Therefore, margins should account for uncertainty in its validity as well as for variability in the estimate of the effect of Standard. Unlike in previous studies, we note here that bio-creep with respect to increasing risk of insufficiently effective therapies may also be a problem under constancy, depending on the strategy for Standard selection and on the margin method.

The simulations in this study were designed to facilitate understanding of bio-creep risk when approved treatments are truly highly effective, such as sulfonamides or penicillin for treatment of CABP in older patients or in bacteremic patients. Here, use of large, fixed margins under violation of constancy (e.g., due to differences in population characteristics between prior and NI trials) could be disastrous, with treatments approved that may be ineffective or harmfully less effective than previously approved treatments in the target populations.

In the setting of developing resistance, the use of non-inferiority is not appropriate if a large proportion of subjects are likely to be infected by resistant strains. Use of a placebo or comparison to active comparators using superiority designs are better suited to evaluating new treatments when previously approved therapies are no longer effective for the majority of enrolled subjects.

Whenever possible, superiority tests should be used when comparing new therapies to active controls for evaluation of efficacy. When testing for non-inferiority is deemed appropriate, use of an active control which has been directly compared to placebo is ideal. Margins such as 95–95($p > 0$), Synthesis($\lambda > 0$, $p > 0$) or Bias-adjusted($\lambda > 0$, $p > 0$) are recommended. These are based upon the estimated effect of Standard, account for likely sources of bias in this estimate, and address the importance of preservation of some proportion of the effect of Standard. The recent GAO report on FDA evaluation of NI trial results in regulatory approval of new drugs suggests that this agency has taken steps to minimize bio-creep (GAO 2010). Simulated results presented here further justify the FDA's advocacy of smaller NI margins and their guidelines for active control choice in this setting.

Acknowledgment

For both TR Fleming and K Odem-Davis, the sources of financial support for research described in this article are NIH/NIAID grants entitled " Statistical Issues in AIDS Research" (R37 AI 29168) and Clinical Research on AIDS Training Grant (T32 AI07450). Additional support for K Odem-Davis was on the NIH grant entitled Mechanisms and Cofactors of HIV Transmission to Women (P01-HD064915) and on the National Center for Advancing Translational Sciences of the National Institutes of Health NIH Award Number UL1TR000423. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

D'Agostino RB, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues - the encounters of academic consultants in statistics. Statistics in Medicine. 2003 Jan; 22(2):169–186. [PubMed: 12520555]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Davis KS. Non-constancy, estimation bias, biocreep, and an alternative to current methods used in non-inferiority trials. University of Washington, Department of Biostatistics. 2010

Everson-Stewart S, Emerson SS. Bio-creep in non-inferiority clinical trials. Statistics in Medicine. 2010; 29(27):769–802.

Fleming TR. Current issues in non-inferiority trials. Statistics in Medicine. 2008; 27(3):317–332. [PubMed: 17340597]

Fleming TR, Odem-Davis K, Rothmann M, Shen YL. Some essential considerations in the design and conduct of non-inferiority trials. Clinical Trials. 2011; 8(4):432–439. [PubMed: 21835862]

Fleming TR, Powers JH. Issues in noninferiority trials: the evidence in community-acquired pneumonia. Clin Infect Dis. 2008 Dec.47:S108–S120. Review. [PubMed: 18986275]

United States Food and Drug Administration. points to consider, rescinded February 2001, 1992. 1992. http://www.fda.gov

United States Food and Drug Administration. points to consider, 2001. 2001 Feb. http://www.fda.gov

United States Food and Drug Administration. Draft Food and Drug Administration guidance document: guidance for industry non-inferiority clinical trials. 2010 Feb. http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation

United States Government Accountability Office. New drug approval: Fda's consideration of evidence from certain clinical trials, gao-10-798. 2010 Jul. http://www.gao.gov/products/GAO-10-798

Higgins K, Singer M, Valappil T, Nambiar S, Lin D, Cox E. Overview of recent studies of community-acquired pneumonia. Clinical Infectious Diseases. 2008; 47(Supplement 3):S150–S156. PMID: 18986282. [PubMed: 18986282]

Kang S-H, Tsong Y. Strength of evidence of non-inferiority trials-The adjustment of the type I error rate in non-inferiority trials with the synthesis method. Statistics in Medicine. 2010; 29(14)

Odem-Davis K, Fleming TR. Adjusting for unknown bias in non-inferiority clinical trials. Statistics in Biopharmaceutical Research. 2013; 5(3)

Peterson P, Carroll K, Chuang-Stein C, Ho YY, Jiang Q, Gang L, Sanchez M, Sax R, Wang YC, Snapinn S. Pisc expert team white paper: Toward a consistent standard of evidence when evaluating the efficacy of an experimental treatment from a randomized, active-controlled trial. Statistics in Biopharmaceutical Research. 2010; 2(4)

Rothmann M. Type I error probabilities based on design-stage strategies with applications to noninferiority trials. Journal of Biopharmaceutical Statistics. 2005; 15(1)

Rothmann M, Li N, Chen G, Chi GY, Temple R, Tsou HH. Design and analysis of non-inferiority mortality trials in oncology. Statistics in Medicine. 2003; 22(2):239–264. [PubMed: 12520560]

Rothmann, MD.; Wiens, BL.; Chan, ISF. Design and analysis of non-inferiority trials. CRC Press; 2011.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

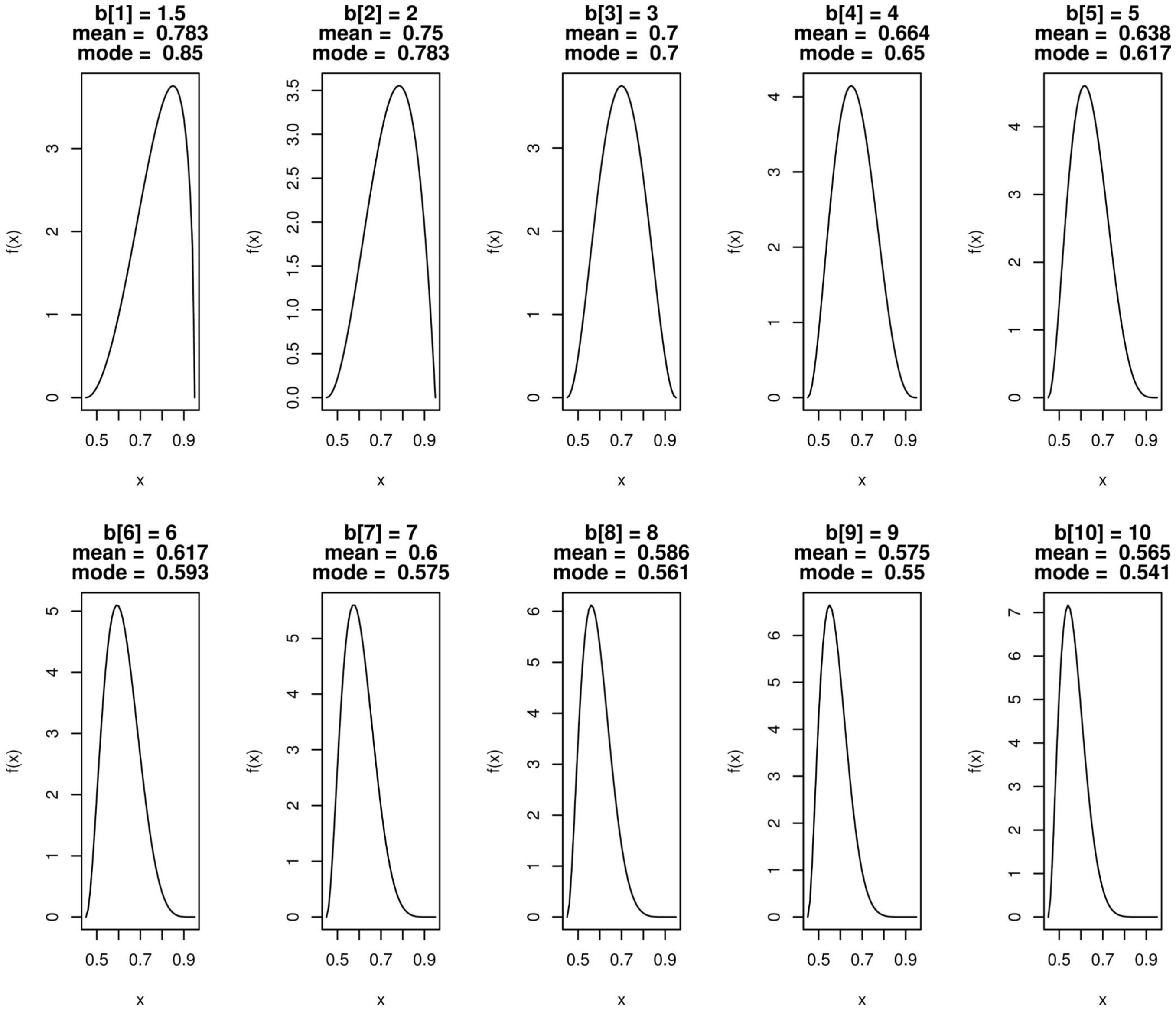

Figure 1.

Densities, f(x), for random variables s_E = 0.5*Z + 0.45 with Z distributed Beta(a=3, b=b[i]) and with b = (1.5, 2, 3, 4, 5, 6, 7, 8, 9, 10). True success rates for simulations were drawn from these “Scaled Beta” distributions and then estimated from Binomial samples.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

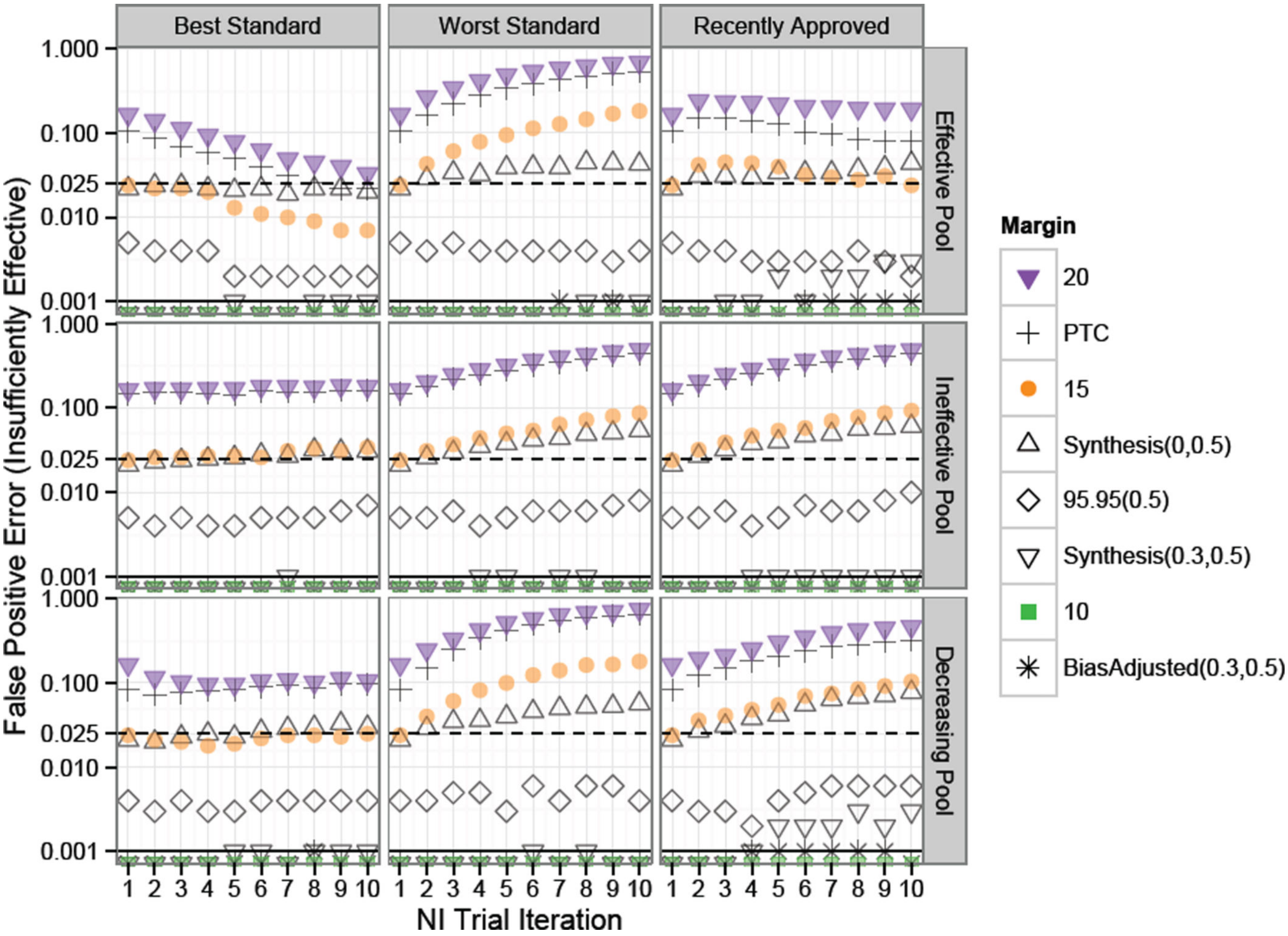

Figure 2.

Study 1 risk of approval of “Insufficiently Effective” treatments under constancy. Standard for each NI trial in a series of ten NI trial iterations chosen according to one of: “Best”, “Worst”, or “Recently Approved”. Success rates on experimental treatments were drawn from Scaled Beta distributions representing an “Effective Pool”, “Ineffective Pool” or “Decreasing Pool” of therapies as indicated by the row label. The dashed and solid horizontal lines indicate one-sided 0.025 and 0.001 False Positive thresholds respectively. Fixed 0.2 and Points to Consider (PTC) margins had high false positive rates in all scenarios, though with decreasing or constant rates under the Best Standard. Scenarios with Fixed 0.15 and Synthesis(0,0.5) margins were subject to increasing rates of False Positive Error in all settings except when the Best Standard was used on an Effective Pool of treatments.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

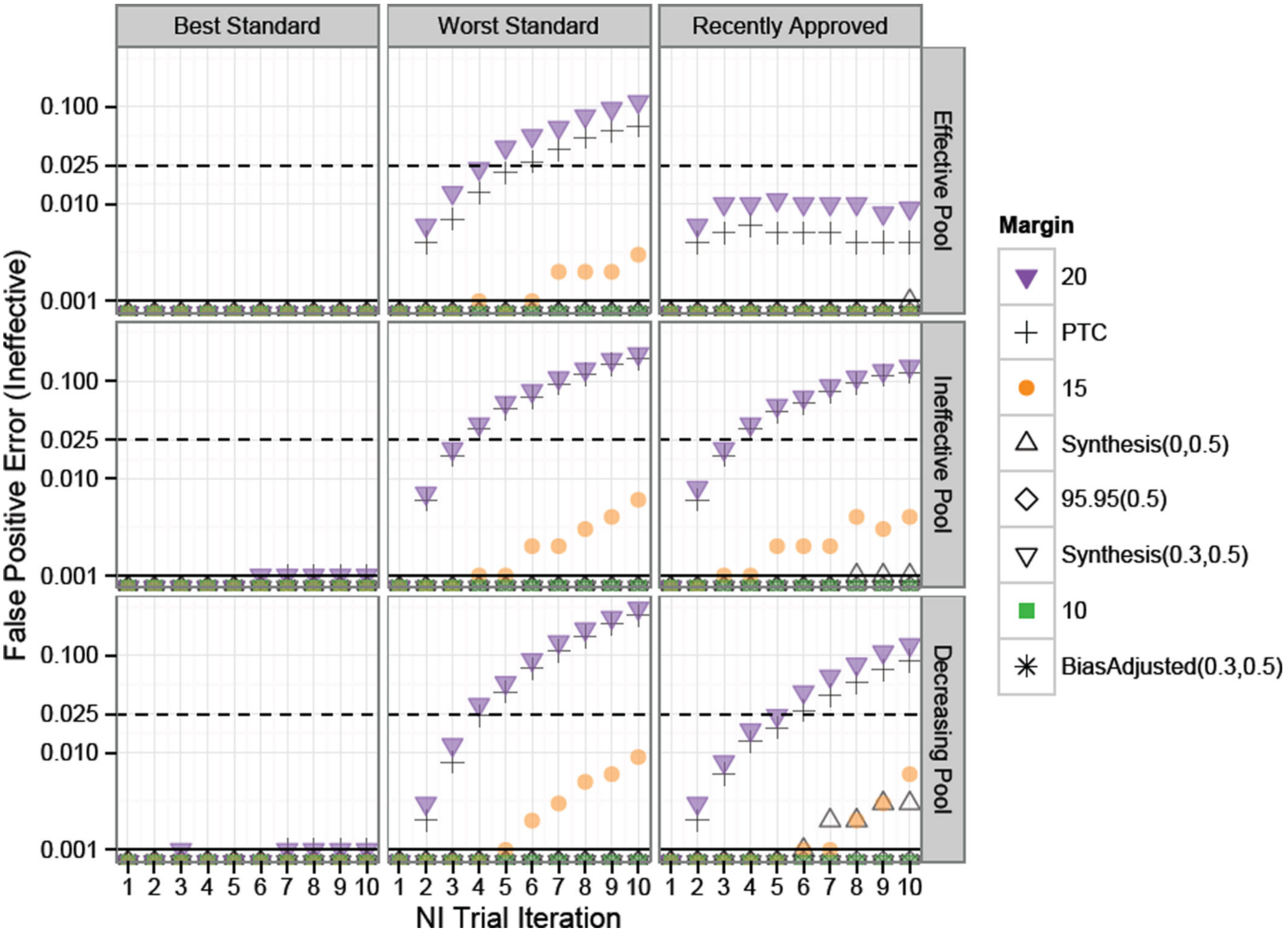

Figure 3.

Study 1 risk of approval of "Ineffective" treatments under constancy. Standard for each NI trial in a series of ten NI Trial Iterations was chosen according to one of the rules: "Best", "Worst", or "Recently Approved", as indicated by the column label. Success rates on experimental treatments were drawn from Scaled Beta distributions representing an "Effective Pool", "Ineffective Pool" or "Decreasing Pool" of therapies as indicated by the row label. The dashed and solid horizontal lines indicate one-sided 0.025 and 0.001 False Positive thresholds respectively. When the "Best" therapy was used as Standard, risk for ineffective therapy approval was negligible, even after many generations of NI trials and regardless of the distribution of Experimental therapies. However, under other methods for choosing Standard, the false positive risk was great after just two generations when the Fixed 0.2 or PTC margins were used and after five to seven generations when Synthesis or Fixed 0.15 margins were used.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

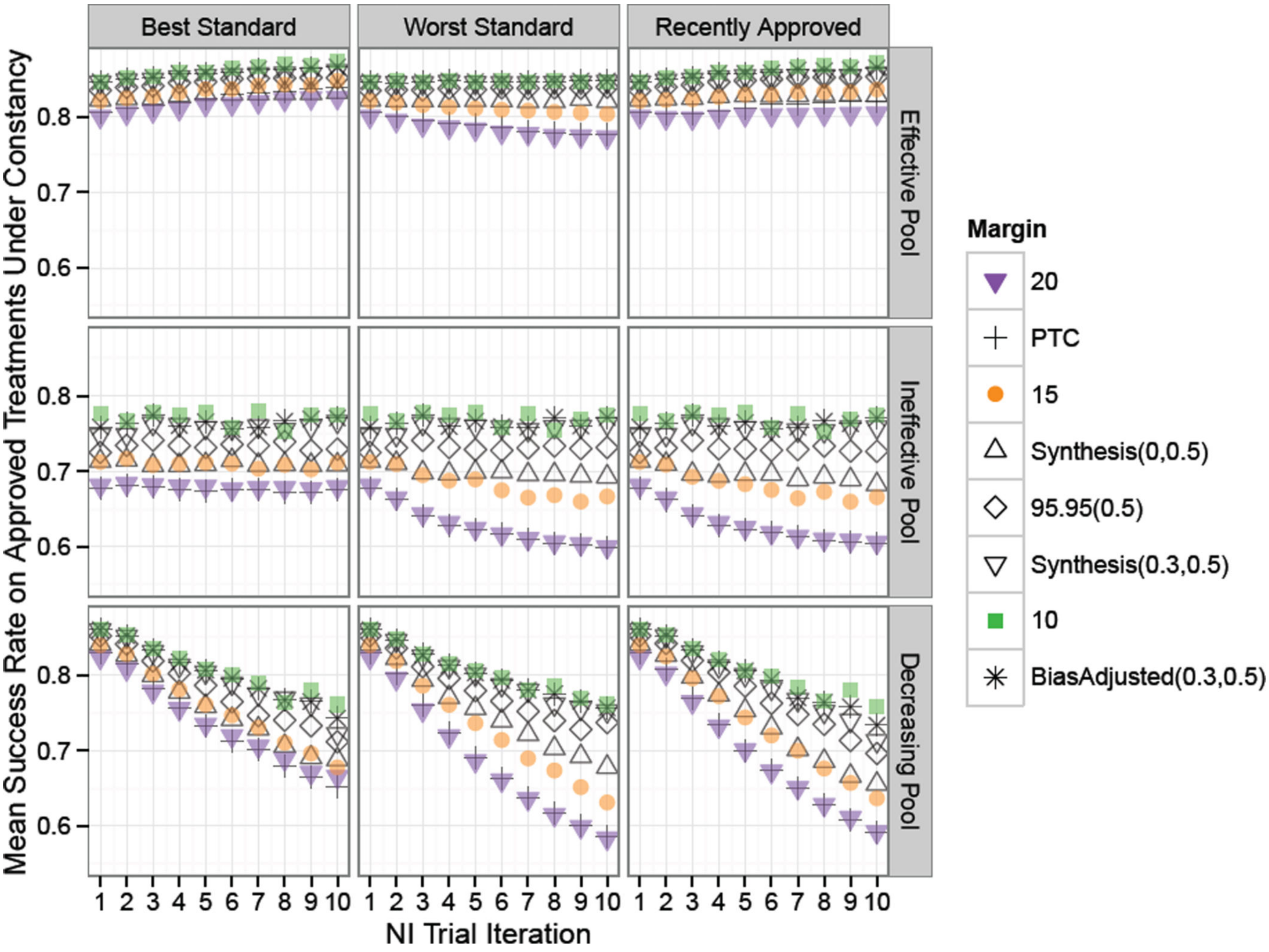

Figure 4.

Study 1 clinical success proportion means on newly approved treatments under constancy. Standard for each NI trial in a series of ten NI Trial Iterations was chosen according to one of the rules: “Best”, “Worst”, or “Recently Approved”, as indicated by the column label. Success rates on experimental treatments were drawn from Scaled Beta distributions representing an “Effective Pool”, “Ineffective Pool” or “Decreasing Pool” of therapies. For all but the Decreasing Pool of treatments, use of the Best approved Standard and 95–95 and smaller margins prevented the mean efficacy of newly approved therapies from declining over time.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

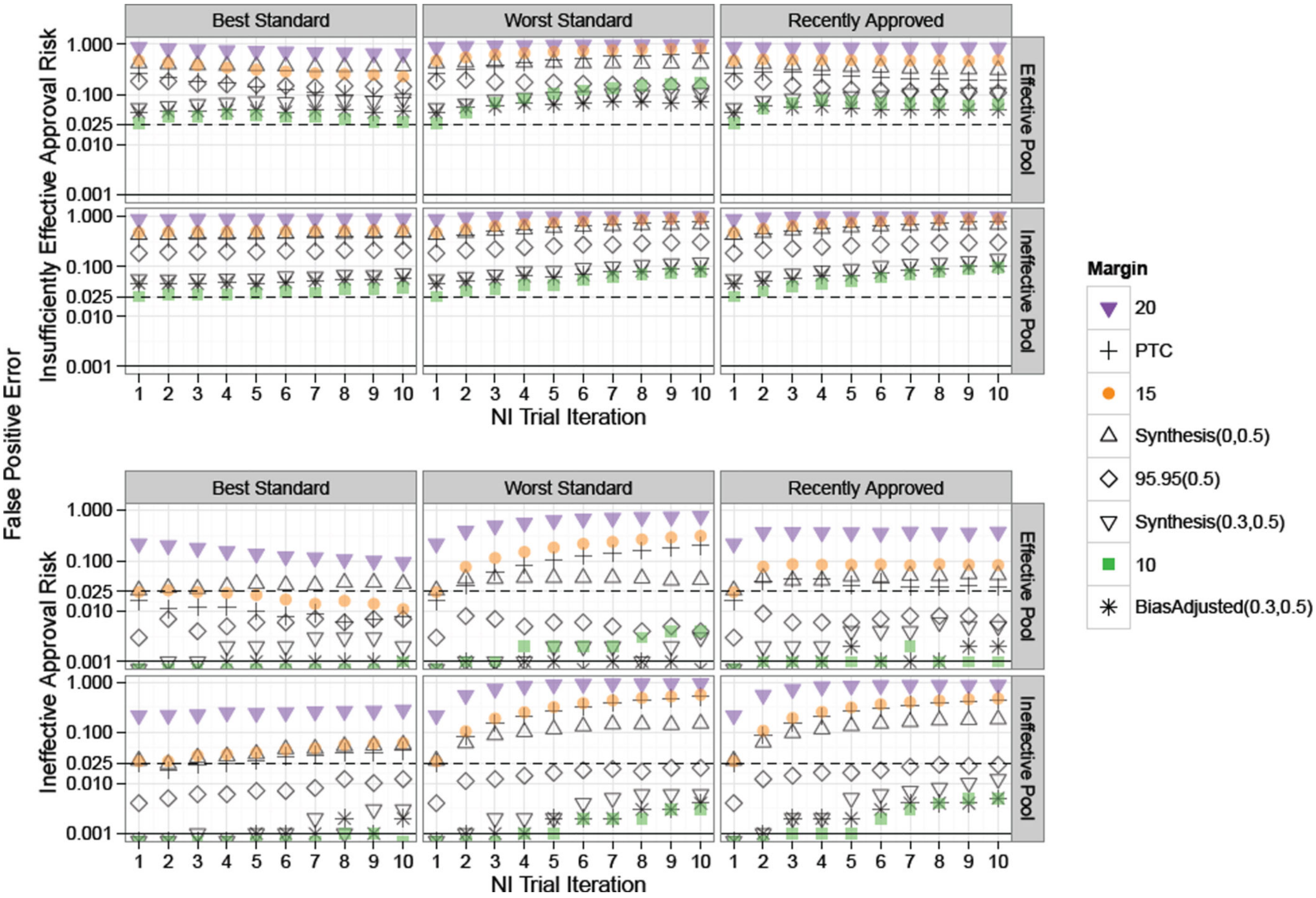

Figure 5.

Study 2 plot of risk of approval of “Insufficiently Effective” and “Ineffective” treatments under non-constancy due to imbalances in patient characteristics. Standard for each NI trial in a series of ten NI Trial Iterations was chosen according to one of the rules: “Best”, “Worst”, or “Recently Approved”. Success rates on experimental treatments were drawn from Scaled Beta distributions representing an “Effective Pool” or “Ineffective Pool” of therapies. Solid black and dashed horizontal lines indicate tolerance thresholds of 0.025 and 0.001 for “Insufficiently Effective” and “Ineffective” false positive error rates respectively. “Insufficiently Effective” False Positive Error risk is high across all scenarios, and “Ineffective” Error risk is inflated even using the Best treatment as Standard for 95–95 and larger margins.

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

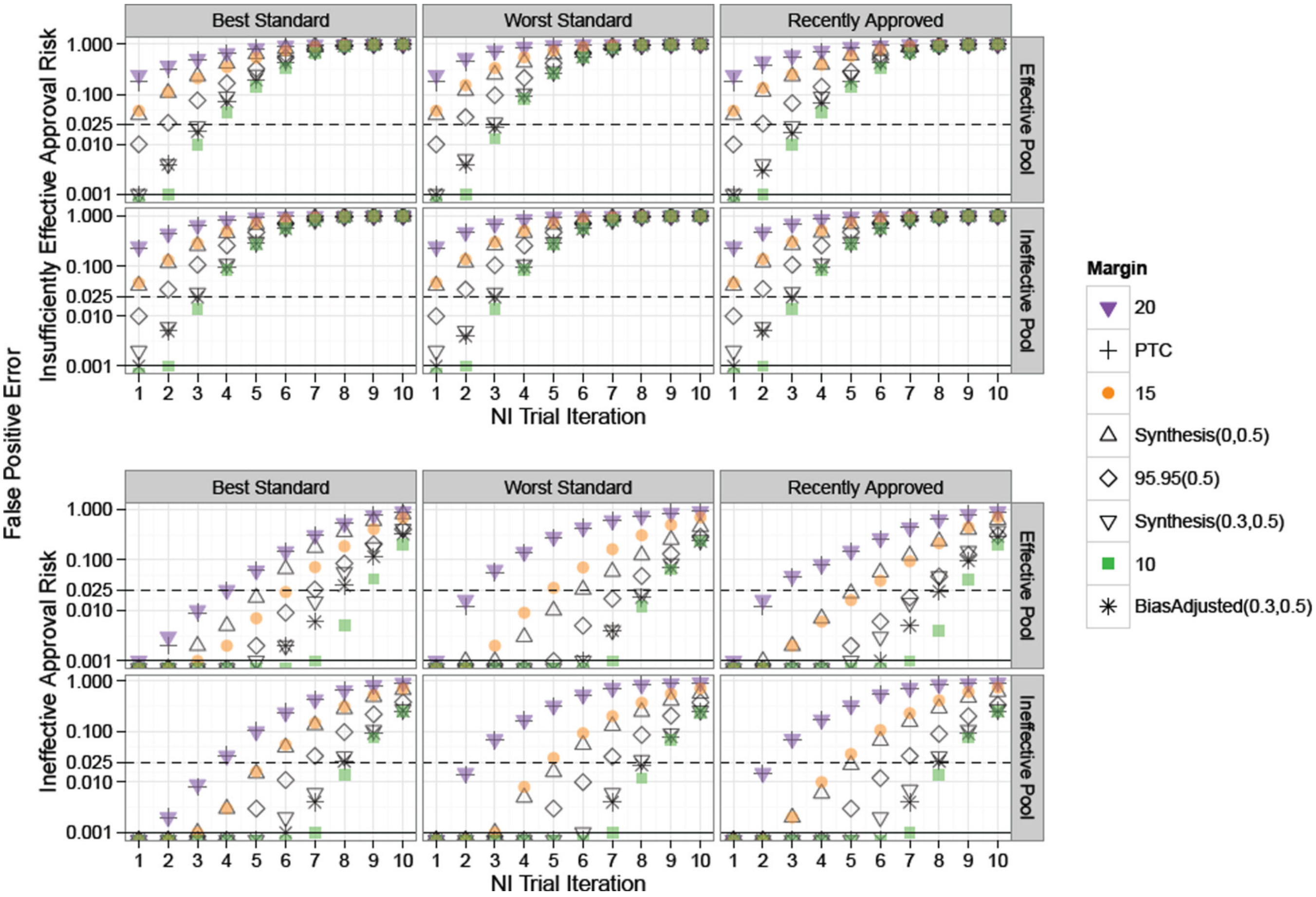

Figure 6.

Study 3 plot of risk of approval of “Insufficiently Effective” and “Ineffective” treatments under non-constancy due to development of resistance. Standard for each NI trial in a series of ten NI Trial Iterations was chosen according to one of the rules: “Best”, “Worst”, or “Recently Approved” . Success rates on experimental treatments were drawn from Scaled Beta distributions representing an “Effective Pool” or “Ineffective Pool” of therapies. The solid black horizontal lines indicate tolerance thresholds of 0.025 and 0.001 for “Insufficiently Effective” and “Ineffective” false positive error rates respectively. After a series of five trials in this setting, when 45% of subjects have resistant strains, Ineffective False Positive Error rates were inflated for 95–95 and larger margins. Insufficiently effective error rates were inflated for all margins in all scenarios for after four trials.