STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:317–332
Published online 5 March 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2855

Current issues in non-inferiority trials[‡]

Thomas R. Fleming[*,†]

University of Washington, Seattle, WA, U.S.A.

SUMMARY

Non-inferiority (NI) trials enable a direct comparison of the relative benefit-to-risk profiles of an experimental intervention and a standard-of-care regimen. When the standard has clinical efficacy of substantial magnitude that is precisely estimated ideally using data from multiple adequate and well-controlled trials, with such estimates being relevant to the setting of the NI trial, then the NI trial can provide the scientific and regulatory evidence required to reliably assess the efficacy of the new intervention. In clinical practice, considerable uncertainty remains regarding when such trials should be conducted, how they should be designed, what standards for quality of trial conduct must be achieved, and how results should be interpreted. Recent examples will be considered to provide important insights and to highlight some of the challenges that remain to be adequately addressed regarding the use of the NI approach for the evaluation of new interventions. 'Imputed placebo' and 'margin'-based approaches to NI trial design will be considered, as well as the risk of 'bio-creep' with repeated NI trials, use of NI trials when determining whether excess safety risks can be ruled out, higher standards regarding quality of study conduct required with NI trials, and the myth that NI trials always require huge sample sizes. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: margin; imputed placebo; active control trial; equivalence trial; bio-creep

1. INTRODUCTION

In many areas of clinical practice, interventions have been shown through well-controlled clinical trials to provide a favourable benefit-to-risk profile and have become standard-of-care options. Frequently, experimental therapeutic or prevention strategies (EXP) are proposed where the intention is that the EXP will reduce the side effects or improve the cost or convenience of this standard

[*] Correspondence to: Thomas R. Fleming, Department of Biostatistics, University of Washington, Box 357232, Seattle, WA 98195, U.S.A.

[†] E-mail: tfleming@u.washington.edu

[‡] Most of the information presented in this article was also presented in an invited lecture at the U.S. Food and Drug Administration, on 15 March 2006 in Washington, DC.

Contract/grant sponsor: NIH/NIAID; contract/grant number: R37 AI 29168

Received 13 July 2006
Accepted 18 January 2007

regimen (STD). While the EXP would provide useful advantages if these intentions can be achieved, it will be necessary to establish that the therapeutic efficacy of the EXP is not unacceptably worse than that of the STD.

Clinical studies designed to rule out that the efficacy of the EXP is unacceptably worse than that of STD have been called equivalence trials or, more recently, non-inferiority (NI) trials, and involve a direct randomization between these two regimens. Such trials not only provide this direct comparison, but also can provide evidence of interest to the scientific and regulatory community regarding whether the EXP is truly effective (i.e. superior to placebo). While studying the relative benefit-to-risk profiles may be sufficient when comparing two known effective agents, the additional objective of assessing efficacy is also essential when evaluating an unproven EXP regimen. The principal interest in this article will be in this setting where efficacy of EXP is unproven. While a placebo controlled trial would provide a more direct assessment of efficacy, randomization to placebo may not be ethical in settings where the STD has been established to provide clinically and statistically significant effects on measures of serious morbidity or mortality.

In active control NI trials in which participants are randomized to either EXP or STD, there are three conditions on STD that would allow the trial to provide a reliable assessment of efficacy of EXP. The STD should have clinical efficacy: (i) that is of substantial magnitude; (ii) that is precisely estimated; and (iii) with estimates that are relevant to the setting in which the NI trial is being conducted. Preferably, this evidence on the efficacy of STD is provided by multiple well-controlled trials. As stated in the International Conference on Harmonization (ICH-E9) [1]:

> A suitable active comparator could be a widely used therapy whose efficacy in the relevant indication has been clearly established and quantified in well-designed and well documented superiority trials and which can be reliably expected to have similar efficacy in the contemplated active control trial.

The most important challenge in the design of NI trials is reliably justifying this third criterion that the estimate of efficacy of STD obtained from previous clinical trials provides a reliable estimate of the true efficacy of STD in the setting of the NI trial [2–6]. This criterion is often called the ‘constancy assumption’. To illustrate its importance, suppose the antibiotic agent vancomycin, the STD, has been established to provide a large improvement, compared to no treatment, in cure rate in patients with various infections. However, suppose that a high prevalence of vancomycin-resistant enterococci (VRE) has developed following many years use. If the NI trial comparing an experimental antibiotic (EXP) with vancomycin is conducted in a patient population where VRE is prevalent, a biased estimate of the efficacy of EXP would be obtained if the constancy assumption is falsely presumed to hold in the NI study. In such settings, one might falsely judge that EXP and STD are equally effective when truly they are equally ineffective regimens in participants having VRE. If equally ineffective, neither intervention should be used in such patients.

The validity of the constancy assumption usually cannot be fully established, and is impacted by whether the design of the NI trial is similar to those upon which the effect of STD is estimated. Even in a disease where the effect of the STD has remained constant over time, the effect of STD may vary across trials due to between-trial differences in patient characteristics, in use of supportive care, in dose, schedule and level of adherence, and in the definition of trial endpoints. NI trials have a reliance on historical evidence that renders results from these trials to be inherently at much higher risk for bias relative to data from adequate and well-controlled superiority trials.

For diseases inducing serious morbidity or mortality, if there are standard interventions that have large and reproducible effects, NI trial designs can be a useful methodology. On the other hand, for diseases that are likely self-resolving and where treatment effects are small or not reproducible, such as the acute otitis media and acute bacterial sinusitis disease settings, use of NI trials is more difficult to justify. It is also in these latter situations where ethical concerns with conduct of placebo controlled trials are the least since the risk of serious morbidity or mortality in the placebo group are the least.

While much has been written about NI trials [3–13], considerable uncertainty remains in clinical practice regarding when such trials should be conducted, how they should be designed, what standards for quality of trial conduct must be achieved, and how results should be interpreted. Recent examples will be considered to provide important insights and to highlight some of the challenges that remain to be adequately addressed regarding the use of the NI approach to the evaluation of new interventions.

2. SOME NI TRIAL DESIGN CONSIDERATIONS

The 'imputed placebo' approach to NI trial design will be considered initially, with greater attention then given to the approach based on 'setting a margin'.

2.1. Imputed placebo approach

For the imputed placebo approach, it is assumed that one would like to use evidence from an NI trial to estimate the efficacy of EXP relative to placebo. For illustration, consider patients with acute coronary syndrome where hirudin was studied as an alternative agent to heparin. There is interest in estimating the odds ratio (OR) of cardiovascular (CV) death or new myocardial infarction (MI) by day 7 for patients with acute coronary syndrome receiving hirudin *versus* placebo. A conceptually straightforward approach would be to estimate this OR by the product of two ORs, that for hirudin (EXP) relative to heparin (STD), estimated using data in Figure 1 from the 'OASIS II' NI trial [14], and that for STD relative to placebo, where estimates of the latter OR would be obtained from trials in Figure 1 evaluating efficacy of STD [11]. This has been called the 'imputed placebo' approach and is discussed in more detail in the Appendix.

In Figure 1, the EXP-to-STD odds ratio (i.e. $\mathrm{OR}=0.84$) is favourable, although EXP provided increased risk of major bleeds and stroke. (Such risks lead to a need for increased rigour in ensuring that the EXP provides favourable effects on the risk of CV death and MI.) For estimation of the OR for STD to placebo, the sponsor provided six trials that were reviewed on 2 May 2000 at the 90th meeting of the FDA Cardio-renal Advisory Committee [15]. Some important issues arise. First, even though trial nos. 2–6 were very small, curiously all five provided OR estimates that were less than unity, raising concerns about potential publication bias and illustrating the importance of ensuring that the search for evidence about the efficacy of STD should not be limited to the published literature. Second, trial no. 1 providing $\frac{2}{3}$ of all events from the six historical trials has a surprisingly high event rate (i.e. $\frac{82}{285}=29$ per cent). One faces a dilemma regarding whether data from trial no. 1 should be included in the estimation of the effect of heparin, the STD. Excluding that trial compromises the principle that the efficacy of STD needs to be precisely estimated. Including it violates the principle that the efficacy of the STD needs to be assessed in settings similar to that in the OASIS II NI trial in order to achieve validity of the constancy assumption.

Such dilemmas can create important risk to the integrity of a NI trial. Given that there is not a clear answer as to whether trial no. 1 should be included, it allows the decision about inclusion

DOI: 10.1002/sim

Non-Inferiority Trial Illustration

Hirudin (EXP) *vs* Heparin (STD)
in Acute Coronary Syndrome
(Primary endpoint: "CV death / new MI", by day 7)

OASIS II	CV death/MI	Major Bleeds	Stroke
Hirudin	178/5045 (3.5%)	60 1.2%	35 0.7%
Heparin	211/5033 (4.2%)	37 0.7%	15 0.3%
	OR= .84		

Six Heparin Trials	#1	#2	#3	#4	#5	#6
Aspirin+Heparin	42/154	2/122	3/210	0/37	4/105	4/70
Aspirin	40/131	4/121	7/189	1/32	9/109	7/73
OR	**.85**	**.49**	**.38**	–	**.44**	**.58**

- **All six heparin trials are small**
- **All OR < 1, in small samples, (literature search bias?)**
- **Most events are from trial #1; #1 has a 29% event rate!**

Figure 1. Estimating the efficacy of hirudin using evidence from the NI 'OASIS II' trial and from six historical trials evaluating efficacy of heparin. 'OR' refers to the odds ratio.

to be influenced by whether the trial provides a favourable estimate of the effect of heparin. Such data-driven selection processes induce the risk of serious bias.

2.2. *Need for objective choice of evidence in estimating the efficacy of the STD regimen*

The importance of need for objectivity in the choice of evidence to be used to estimate the efficacy of the STD regimen is illustrated further in Figure 2. In previously treated non-small cell lung cancer patients, the hazard ratio, often called the relative risk (RR), for death rate for patients receiving pemetrexed (EXP) *versus* placebo can be estimated by a product of two terms: the RR for pemetrexed relative to docetaxel (STD) estimated to be 0.992 using the 'JMEI' NI trial [16], and the RR for STD relative to placebo using data from the TAX 317 and TAX 320 trials [17, 18]. Given that the pemetrexed *versus* docetaxel RR is near unity, whether pemetrexed is estimated to have a favourable effect depends entirely on whether one can justify docetaxel to be strongly beneficial in this setting. Together, the TAX 317 and 320 trials evaluated docetaxel in nearly 600 patients through four pairwise comparisons with control regimens. The designer/sponsor of JMEI chose to exclude three of these pairwise comparisons, excluding $\frac{5}{6}$ of the patients, basing their estimate of docetaxel's efficacy solely on the 75 mg *versus* control results from TAX 317. These exclusions were done presumably because only the 75 mg dose of docetaxel was used in the NI trial and because use of vinorelbine and ifosfamide in the control arm of the TAX 320 trial may have resulted in an underestimate of the effect of docetaxel (although data are not available to establish positive effects of these drugs in this setting). An alternative view point regarding the exclusion of $\frac{5}{6}$ of the docetaxel data is that the designer/sponsor of JMEI knew that inclusion of those patients would result in substantially less impressive estimates for the effect of pemetrexed.

Suppose the choice of the control regimen in an NI trial and, more importantly, the corresponding choice of the data sets to estimate the effect of the STD are made by the designer or sponsor of an NI trial who knows the results of the historical trials evaluating STD (such as the TAX 317

DOI: 10.1002/sim

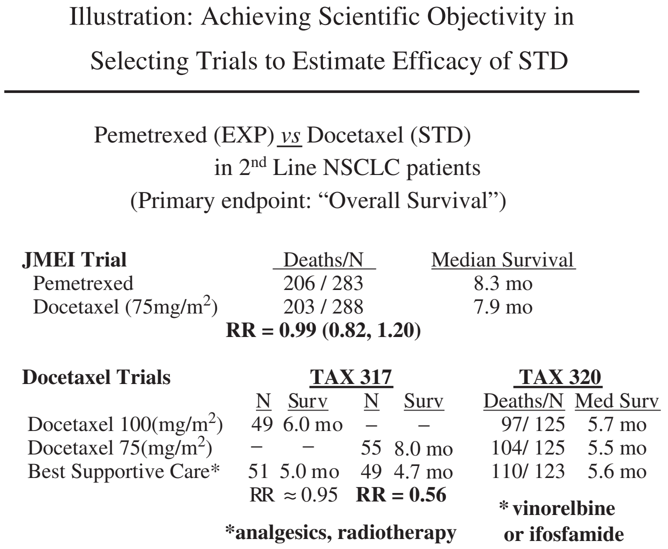

Illustration: Achieving Scientific Objectivity in Selecting Trials to Estimate Efficacy of STD

Pemetrexed (EXP) vs Docetaxel (STD)
in 2nd Line NSCLC patients
(Primary endpoint: “Overall Survival”)

JMEI Trial	Deaths/N	Median Survival
Pemetrexed	206 / 283	8.3 mo
Docetaxel (75mg/m^2)	203 / 288	7.9 mo
	RR = 0.99 (0.82, 1.20)	

Docetaxel Trials	**TAX 317**				**TAX 320**	
	N	Surv	N	Surv	Deaths/N	Med Surv
Docetaxel 100(mg/m^2)	49	6.0 mo	–	–	97/ 125	5.7 mo
Docetaxel 75(mg/m^2)	–	–	55	8.0 mo	104/ 125	5.5 mo
Best Supportive Care*	51	5.0 mo	49	4.7 mo	110/ 123	5.6 mo
	RR ≈ 0.95		**RR = 0.56**			

***analgesics, radiotherapy**

*** vinorelbine or ifosfamide**

Figure 2. Estimating the efficacy of pemetrexed using evidence from the NI ‘JMEI’ trial and from the TAX 317 and 320 trials evaluating docetaxel. ‘RR’ refers to the relative risk, which is also called the hazard ratio.

and 320 trials). When this knowledge can be used to guide these choices, it is likely substantial bias will be introduced in the estimate of the effect of STD, and hence of EXP.

With any NI design approach, risk of bias in the estimate of the effect of STD must be addressed, whether the bias would be due to lack of objectivity in the choice of evidence to estimate the efficacy of STD or to other issues that would compromise the validity of the constancy assumption.

A criticism that is specific to the imputed placebo approach discussed in Section 2.1 is that it might not be answering the clinically most relevant question in those settings where an NI trial is necessary. Certainly, a placebo controlled trial would provide the most direct, efficient and reliable approach for estimating the efficacy of EXP relative to placebo. An NI trial is only necessary when the effect of STD is sufficiently substantial on an endpoint of serious morbidity or mortality that one cannot expose patients to a placebo. In that setting, to be of interest, typically EXP not only would need to be better than placebo but also would need to have an acceptable benefit-to-risk profile relative to that of STD. This can be addressed by obtaining evidence to rule out that efficacy of EXP is unacceptably worse than that on STD, motivating the ‘setting the margin’ approach to the design of NI trials.

2.3. *Setting the margin approach*

In design of NI trials, the approach involving ‘setting a margin’ is probably the one most frequently implemented [2]. With this approach, a conclusion of NI is achieved by ruling out that efficacy on EXP is worse than that on STD by an amount δ, called the margin. In general, δ should be chosen to allow the conclusion that EXP preserves a meaningful amount (say, for example, at least one-half) of the effect of STD. Achieving this would allow two important conclusions: first, the use of EXP rather than STD would not result in a clinically unacceptable loss of efficacy

DOI: 10.1002/sim

when taking into account the totality of information about these two regimens and, second, EXP is effective (i.e. superior to placebo). Information from previous trials regarding the effect of STD is incorporated into this NI design approach through the formulation of δ.

Again, it will be informative to consider recent experiences for insights into the method and challenges that must be addressed. The ‘REPLACE 2’ NI trial [19] was conducted in nearly 6000 patients to estimate the OR of death, MI or urgent revascularization (UR) by day 30 in the setting of percutaneous coronary intervention (PCI) for patients receiving bivalirudin (EXP) *versus* glycoprotein IIb/IIIa inhibitor (STD). While an NI assessment was motivated by bivalirudin’s ability to induce a lower rate of major bleeds, the REPLACE 2 trial (see Figure 3) revealed that EXP has a 9 per cent relative increase in the odds of death/MI/UR when compared to STD, largely due to allowing a higher rate of MIs. The upper limit of the 95 per cent confidence interval indicated that it could not be ruled out that EXP allows a 32 per cent relative increase in the odds of death/MI/UR.

In REPLACE 2, if δ is chosen to ensure that EXP preserves at least $\frac{1}{2}$ the effect of STD, one needs reliable information regarding the STD effect. The EPISTENT and ESPIRIT trials can be used to estimate the effect of the STD, glycoprotein IIb/IIIa inhibitor [20, 21], in the PCI setting. As seen in Figure 3, STD provides an estimated 45 per cent relative reduction in the odds of death/MI/UR, with 240 patients having the primary endpoint collectively in the two trials, in turn yielding moderate precision in the estimate of efficacy of STD. Figure 4 reflects the estimated ORs for death/MI/UR on bivalirudin and placebo, relative to the STD, glycoprotein IIb/IIIa inhibitor. Hence, by the results in Figure 3, we see that placebo (i.e. heparin alone) has an estimated OR 1.82 relative to STD, with 95 per cent confidence interval (1.40, 2.32).

When specifying the margin, δ, adjustment must be made for the variability in the estimate of STD effect. If one were to adjust for this variability as well as the need to ensure that EXP preserves

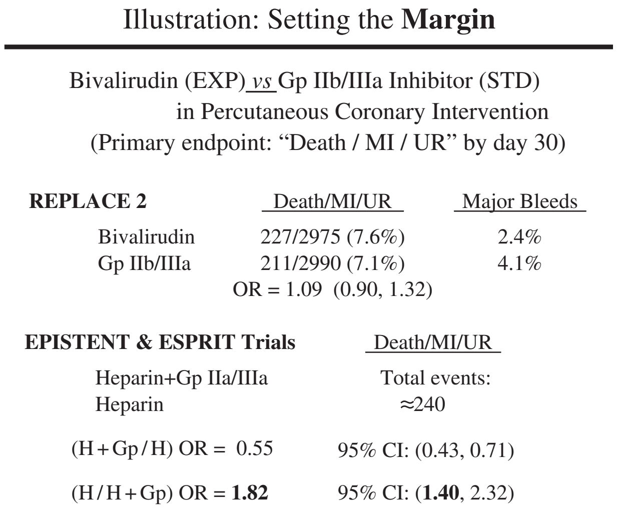

Illustration: Setting the **Margin**

Bivalirudin (EXP) *vs* Gp IIb/IIIa Inhibitor (STD)
in Percutaneous Coronary Intervention
(Primary endpoint: “Death / MI / UR” by day 30)

REPLACE 2	Death/MI/UR	Major Bleeds
Bivalirudin	227/2975 (7.6%)	2.4%
Gp IIb/IIIa	211/2990 (7.1%)	4.1%
	OR = 1.09 (0.90, 1.32)	

EPISTENT & ESPRIT Trials	Death/MI/UR
Heparin+Gp IIa/IIIa Heparin	Total events: ≈240

(H + Gp / H) OR = 0.55 95% CI: (0.43, 0.71)

(H / H + Gp) OR = **1.82** 95% CI: (**1.40**, 2.32)

Figure 3. Estimating the efficacy of bivalirudin using evidence from the NI ‘REPLACE 2’ trial and from the ‘EPISTENT’ and ‘ESPRIT’ trials evaluating the effect of the STD intervention, glycoprotein IIb/IIIa inhibitor. ‘MI’ denotes myocardial infarction; ‘UR’ denotes urgent revascularization.

DOI: 10.1002/sim

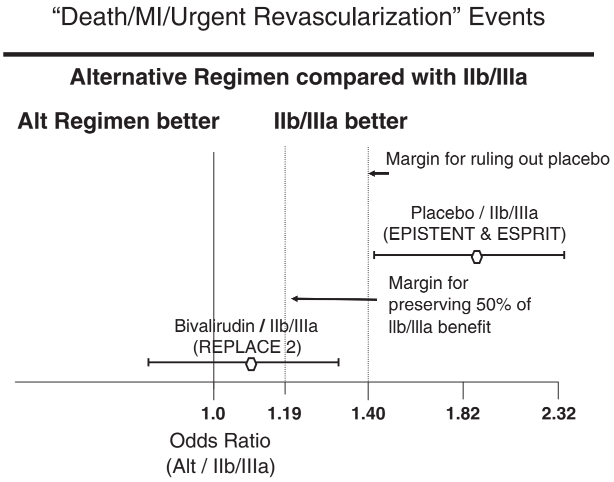

Figure 4. A graphical illustration of formulating the 'margin', δ, with implementation of this NI trial approach in evaluation of bivalirudin in PCI. 'IIb/IIIa' refers to glycoprotein IIb/IIIa Inhibitor.

at least $\frac{1}{2}$ of the effect of STD then, using inequality (A6) in the Appendix, the margin would be given by $\delta = 1.30$. However, choosing a margin that does not make an additional adjustment to take into account the uncertainty of the validity of the constancy assumption would result in an NI trial whose validity would be as suspect as an NI analysis based on 'the imputed placebo' approach. Such approaches allow substantially increased risk of false-positive conclusions as noted in ICH-E10 [2]:

> The determination of the margin in a non-inferiority trial is based on both statistical reasoning and clinical judgment, and should reflect uncertainties in the evidence on which the choice is based, and should be suitably conservative.

Confidence in the validity of the constancy assumption (i.e. that an unbiased estimate of the effect of glycoprotein IIb/IIIa inhibitor in REPLACE 2 is provided by estimates from EPISTENT and ESPRIT) is weakened by differences in these two settings: (i) in patient characteristics (with REPLACE 2 having more low-risk patients, a potential effect modifier); (ii) in use of supportive care (with higher doses of heparin in REPLACE 2, or such as use of clopidogrel); (iii) in dose, schedule and level of adherence; and (iv) in efficacy and safety endpoints (with REPLACE 2 requiring only a single CK-MB elevation in the definition of MI, with UR events counted for any rather than just target vessels, and with a 3 g/Dl rather than 5 g/Dl decrease in haemoglobin counted as a major bleed event, leading to an increase in the rates of such events).

To address the uncertainty regarding the validity of the constancy assumption, the CBER at FDA introduced the '95–95' method for formulating the margin [22]. As illustrated in Figure 4, the lower limit of the 95 per cent confidence interval of the OR for placebo to STD (i.e. 1.40), is used to identify the position of placebo relative to STD. This lower limit is used to simultaneously address the variability in the estimate of the effect of STD as well as to adjust for the uncertainty regarding the validity of the constancy assumption. Then, to ensure that the EXP preserves at least half the effect of STD, the margin would be set to be $1.19 = \sqrt{1.40}$ (i.e. the margin is the midpoint

DOI: 10.1002/sim

between 1.0 and 1.4 on a log scale rather than linear scale since the primary parameter estimated, as noted in the Appendix, is the logarithm of OR).

Before finalizing the selection of 1.19 as the margin in this setting, one must also be able to justify that up to a 19 per cent relative increase in the odds of death/MI/UR would be clinically acceptable when assessed in the context of overall properties of EXP and STD. These overall properties not only include the pre-specified primary endpoint that receives particularly strong weight, but also additional measures of efficacy, safety, convenience of administration and possibly even cost. We see from Figure 3, per 100 person years, that EXP would have approximately 1.7 fewer major bleeds, but the 19 per cent relative increase in the odds of the primary endpoint would represent approximately 1.25 additional MIs (since the difference between EXP and STD is predominantly in MI rather than death or UR components of the REPLACE 2 composite primary endpoint). Given that MI events represent irreversible morbidity, these considerations further motivate that the margin should not exceed approximately 1.19.

Now that the margin is formulated (done before the REPLACE 2 NI trial is initiated), the NI assessment can be conducted. Given that the upper 95 per cent confidence limit for the EXP-to-STD OR from REPLACE 2 is 1.32 (see Figure 4) this trial would not establish non-inferiority of bivalirudin relative to the glycoprotein IIb/IIIa inhibitor when using the NI margin of 1.19.

3. SOME SELECTED ISSUES

Several additional issues deserve consideration in design, conduct and analysis of NI trials. Among these are whether very large sample sizes are inevitable in NI trials having rigorous margins, the risk of 'bio-creep' with repeated NI trials, use of NI trials when determining whether excess safety risks can be ruled out, and higher standards that are required regarding quality of study conduct when conducting NI trials.

3.1. Sample sizes in NI trials

A widely held myth is that NI trials having scientifically rigorous margins always require very large sample sizes. This myth was stated by some industry representatives at the 19 February 2002 meeting of the FDA Anti-Infective Drugs Advisory Committee [23], held to consider the FDA's proposal to reduce the margins used in many NI trials conducted in this setting.

This myth is challenged by the illustration in Figure 5 relating to the evaluation of the effect of antibiotics on cure rate in settings such as community or hospital-acquired pneumonia. Traditionally, when randomizing between experimental and standard antibiotic regimens in such settings, trials have assessed endpoints such as cure rate and have had a total of 300–400 patients. In Figure 5, the difference in EXP and STD cure rates are plotted and four scenarios are considered. In each scenario, assume that the STD antibiotic regimen provides an 80 per cent cure rate. In the first, where the intention is to establish superiority of EXP relative to STD, if in truth EXP provides a 12 per cent higher cure rate, then 340 patients would provide 90 per cent power to rule out equality at the traditional one-sided 0.025 false-positive error rate. While such a result would be impressive, it would be uncommon to find such potent new antibiotics. In scenario no. 2, where the intention is to establish NI by ruling out that the cure rate on EXP is at least 15 per cent lower than that on STD, this could be achieved with 90 per cent power with only 300 patients when the true cure rates on EXP and STD are equal. This ability to have a high probability of

DOI: 10.1002/sim

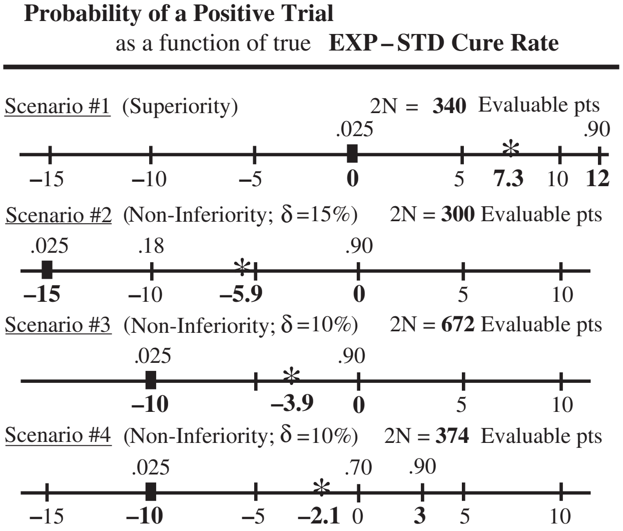

Figure 5. In the antibiotic setting, an illustration, use of rigorous NI margins does not require prohibitively large sample sizes. Each $*$ denotes the least favourable point estimate that yields a positive result in that setting.

achieving a 'positive' result in relatively small trials led many industry sponsors to advocate the use of a large 15 per cent margin. Unfortunately, using such a large margin results in having an 18 per cent probability of 'establishing NI' even when EXP truly has a 10 per cent lower cure rate than STD, where a 10 per cent lower cure rate would correspond to a clinically important relative 50 per cent increase in the non-cure rate on EXP *versus* STD. Rectifying this by using the smaller 10 per cent NI margin in scenario no. 3, as proposed by FDA, in turn would lead to doubling in sample size. However, if the true cure rate is only slightly (i.e. 3 per cent) better on EXP than on STD, as illustrated in scenario no. 4, then one can still use a rigorous 10 per cent NI margin without requiring large increases in sample size. To be specific, when the cure rate on EXP is in truth only 3 per cent better than that on STD, a trial with 374 patients would have 90 per cent power to rule out that the EXP cure rate is 10 per cent worse.

These scenarios point out a common flaw in reasoning that the EXP must either be truly much better than STD, such that superiority can be assessed as in scenario no. 1, or be truly the same as STD, as in scenario nos. 2 and 3. In fact, none of these might be the truth. The EXP might provide a small improvement relative to STD. In scenario no. 4, with a typical sample size, one can rule out a rigorous NI margin with 90 per cent power if EXP in truth provides a small 3 per cent increase in cure rate, and 70 per cent power even if EXP and STD in truth have the same cure rate. From a public health perspective, this means that there is a low risk of missing an antibiotic that is at least slightly better than the standard.

3.2. Bio-creep with repeated NI trials

In Figure 5, the asterisks represent the least favourable point estimates that would allow one to rule out the null hypothesis being tested in that scenario. When a large NI margin is used, as in

DOI: 10.1002/sim

Scenario no. 2, a ‘positive’ result would be obtained even with the substantially unfavourable point estimate that EXP has an absolute (not relative) 5.9 per cent lower cure rate. Such an approach provides a substantial risk that after two or three generations of NI trials have been conducted in a given clinical setting, ineffective interventions would be allowed into the market place. This outcome resulting from use of non-rigorous margins has been called ‘bio-creep’.

Consider, for illustration, the setting of empiric use of anti-fungal therapy of febrile neutropenic patients. The first generation therapy, amphotericin B deoxycholate, was approved for use in the 1970s based on trials by Pizzo and EORTC that had less than 100 patients [24, 25] and did not provide reliable evidence for benefit. The second generation anti-fungal, liposomal amphotericin B, was compared with amphotericin B deoxycholate in an NI trial that showed both regimens provided a 49 per cent success rate, where failure was occurrence of death, breakthrough fungal infection, or persistent fever [26]. The third generation anti-fungal, voriconazole, was compared with liposomal amphotericin B in an NI trial where voriconazole had a lower success rate (23.7 *versus* 30.1 per cent, 95 per cent confidence interval for the difference: −12 per cent, −0.1 per cent) [27]. The interpretation of these results was further complicated by the fact that the definition of persistent fever differed between the two NI trials, resulting in substantially different success rates on liposomal amphotericin B across the trials. On 4 October 2001, the FDA Anti-viral Advisory Committee concluded that these data failed to provide adequate evidence to establish favourable efficacy for voriconazole. However, if voriconazole had been approved based on the use of more lenient 15 or 20 per cent margins, an ineffective or meaningfully less effective fourth generation anti-fungal treatment readily could satisfy an efficacy criterion if compared with voriconazole in an NI trial using such lenient margins. Voriconazole and later generation anti-fungal treatments that could be much less effective than earlier generation agents could be approved for use if evaluated in NI trials using non-rigorous margins.

Recently, it has been proposed to further modify the definition of fever resolution for future trials conducted in this clinical area [28]. While these modifications may result in a more clinically meaningful endpoint, using a new endpoint to evaluate the next generation anti-fungal treatment would require the use of a superiority trial due to constancy assumption issues.

The hazards of bio-creep are broadly apparent in anti-infective settings such as acute otitis media, acute exacerbation of chronic bronchitis, and acute bacterial sinusitis, where generations of NI trials have been performed leading to FDA approval of more than a dozen antibiotics in each of these three settings [29, 30]. Such practice, unfortunately, has enabled widespread use of antibiotics that may not be providing true clinical benefit to patients in those settings, yet may be inducing safety risks and development of resistance such that these antibiotics will lose effectiveness in clinical settings where they truly would have been helpful.

3.3. Use of NI trials in evaluating safety risks

NI trials provide a useful approach when evaluating whether unacceptable safety risks can be ruled out for an intervention established to be effective in earlier trials, yet where those trials suggest a safety signal which could substantially alter the benefit-to-risk profile of the intervention. For illustration, in the setting of rheumatoid arthritis and osteoarthritis, clinical trials have established that celecoxib, a Cox-2 inhibitor, provides favourable analgesic effects. However, studies have raised concerns about adverse effects of Cox-2 inhibitors on the rate of CV death, MI and stroke. The recently designed Precision Trial will assess whether one can rule out the hypothesis that celecoxib (relative to naproxen) induces at least a 33 per cent increase in the rate of ‘CV death/MI/stroke’.

DOI: 10.1002/sim

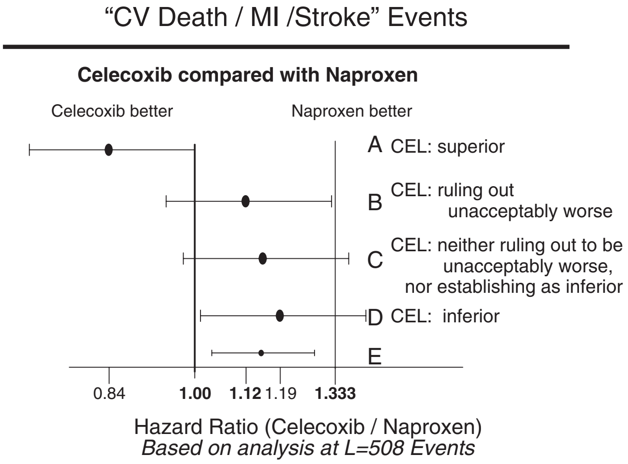

Figure 6. Interpreting outcomes in the 'Precision' NI trial, conducted to assess whether unacceptable safety risks from celecoxib can be ruled out.

To have 90 per cent power to rule out this hypothesis when celecoxib truly provides no increase in CV risk, the trial will need to be of sufficient size so that $L = 508$ patients in the two arms have this outcome event. Hence, the trial will follow approximately 7000 randomized patients per arm for an average of 2+ years. Scenarios A, B, C and D in Figure 6 provide several important outcomes that might emerge in the trial, with the corresponding interpretation that may be made regarding this safety outcome.

Scenario E displays a possible outcome had the trial had been designed to provide $L = 1000$ events in these two arms. While NI would be established in this scenario, nevertheless, there is an apparent paradox because data also would establish celecoxib to be inferior to naproxen. This clearly illustrates that establishing NI does not allow one to conclude that EXP is 'at least as good as' STD. Rather, the conclusion established by NI trials is that EXP is not 'unacceptably worse' than STD. In turn, this further illustrates the importance of choosing a sufficiently rigorous margin that truly represents the smallest difference that would be clinically meaningful. In this setting, for example, one would need to justify that up to a 33 per cent relative increase in 'CV death/MI/stroke' could be accepted in view of the analgesic effects of celecoxib and its relatively low level of inducing gastro-intestinal ulceration.

3.4. Ensuring sensitivity to true differences in efficacy between STD and EXP

In any trial, irregularities in quality of trial conduct can lead to biased evaluations of efficacy and safety. This bias resulting from non-adherence, withdrawals from therapy, missing data, violations in entry criteria, or other deviations in the protocol, often tend to reduce sensitivity to differences between regimens. Hence, this can be of greater concern in an NI setting since it would increase the risk of inappropriately concluding that EXP is not unacceptably worse than STD. The occurrence of EXP-to-STD cross-ins (i.e. where EXP patients later receive STD) and STD-to-EXP cross-ins in NI efficacy and NI safety trials is very problematic, particularly if these cross-ins occur relatively soon after randomization.

DOI: 10.1002/sim

Even in NI trials, intention-to-treat (ITT) analyses have an advantage over 'per-protocol' analyses that are based on excluding from analysis the outcome information generated in patients with irregularities, since per-protocol analyses fail to maintain the integrity of randomization. Hence, to increase the sensitivity of ITT analyses to true efficacy differences between EXP and STD, it is especially important in the NI setting to design and conduct trials in a manner to reduce the occurrence of these irregularities.

4. SOME CONCLUSIONS AND RECOMMENDATIONS

Complexities inherent in evaluating the benefit-to-risk profile of EXP interventions through NI trials motivate this study design to be used only when ethics and science dictate it to be necessary. In order to justify a non-zero margin (i.e. justifying that efficacy can be established with results weaker than achieving superiority of EXP to STD), the STD must have clinical efficacy of substantial magnitude that is precisely estimated, with the estimates satisfying the constancy assumption.

Regulatory approval of a drug does not *de facto* mean that it can serve as an appropriate control in a future NI trial unless substantial benefit relative to placebo is established in a reliable and reproducible fashion. It follows that regulatory approval of interventions that have weak evidence regarding efficacy and safety not only places the clinical community at risk of using treatments having an unfavourable benefit-to-risk profile, but also makes it more difficult to evaluate and approve other new promising interventions in that setting since substantial margins cannot be justified if an NI trial approach is used.

It is especially problematic to obtain reliable evidence about the benefit-to-risk profile of EXP through an NI trial using surrogate endpoints. To provide reliable evidence of favourable efficacy, the trial must allow one to rule out that EXP is 'unacceptably worse' than STD with respect to a true clinical efficacy end point, i.e. a measure of tangible benefit to patients such as prolongation of survival, prevention or relief of disease symptoms, or improvement in functional status. While it is rare to have sufficient evidence to validate that a treatment effect on a surrogate endpoint reliably establishes a treatment effect on a clinical efficacy end point, it is even more rare to have sufficient evidence to determine the amount of loss of treatment effect on the surrogate (the NI margin when using the surrogate) that would correspond to the amount of loss of effect on the clinical end point that must be ruled out.

Further exploration of NI trial methodology is needed in several areas. First, whether one uses the 'imputed placebo' or 'setting the margin' approach, an improved process is needed that would achieve a more objective choice of evidence to be used to estimate the efficacy of the STD regimen, thereby reducing the substantial bias arising currently when study designers make choices to maximize the impression of efficacy of the STD (and in turn of EXP). Jorgensen *et al*. conducted an overview that provides additional evidence of the bias in industry-supported meta-analyses of drug effects and they reviewed some of methodological considerations used by the Cochrane reviews in conducting meta-analyses [31]. Second, when irregularities in quality of trial conduct cannot be prevented, methodological approaches are needed that preserve more effectively the integrity of randomization than is achieved by the popular 'per-protocol' analyses. Third, it would be valuable to have improved methods to define adjustments to margins to account for uncertainties about the validity of the constancy assumption. NI trial designs that do not address these uncertainties, such as using the 'imputed placebo approach' or defining a margin based on inequality (A6) in the Appendix, can result in substantially inflated risks of false-positive

DOI: 10.1002/sim

conclusions. While the '95–95' method in expression (A7) does provide adjustments to the margin to address the constancy assumption, it is apparent from expression (A8) that this adjustment is based on the variability of the estimates of treatment effects from the NI trial and the trials evaluating STD. However, it is bias more than variability that is caused by invalidity of the constancy assumption. Logically, methods are needed that can guide constancy assumption-related adjustments to margins based more on the bias in estimates of the effect that STD has in the NI trial than on the precision of the estimates.

The goal of a clinical development plan should be to determine whether a new intervention provides an important advance in clinical care. Using NI trials to evaluate interventions not expected to provide important efficacy or safety advantages over STD introduces substantial risks of eroding the progress made in benefits delivered by current therapies, especially when NI trials are proposed that do not have rigorously justified margins. It is tempting in such settings to propose the use of NI trials having large margins to enable reductions in sample size and increased likelihood of obtaining a positive efficacy conclusion. However, it is logically inconsistent to judge it to be clinically acceptable for EXP to have a level of efficacy that could approach being δ less than what is provided by STD if new and more toxic therapies providing increases of smaller magnitude than δ would be viewed to represent clinically important advances. NI trial designs must be implemented in a rigorous manner to ensure that reliable and interpretable evidence is provided to the clinical community, and to enhance the likelihood that newly available interventions truly represent a step ahead rather than a step back in quality of clinical care.

APPENDIX

Consider the setting where one has EXP, STD and placebo regimens. Let the term β_{ES} denote the true effect of EXP relative to STD, β_{SP} denote the true effect of STD relative to placebo, and β_{EP} denote the true effect of EXP relative to placebo. In the specific setting of a time-to-event outcome evaluated using the Cox regression proportional hazards model, let β_{EP} be the natural logarithm of the hazard ratio (also called the log relative risk) for EXP relative to placebo. In the setting of a dichotomous outcome, let β_{EP} be the natural logarithm of the odds ratio of failure for EXP relative to placebo. The terms β_{ES} and β_{SP} have analogous definitions in these two settings. Note $\beta_{\mathrm{ES}}, \beta_{\mathrm{SP}}$ and β_{EP} are negative when EXP is more efficacious than STD, when STD is more efficacious than placebo, and when EXP is more efficacious than placebo, respectively.

Assume $\hat{\beta}_{\mathrm{ES}}$ is the estimate of β_{ES} obtained using data from the NI trial, and $\hat{\beta}_{\mathrm{SP}}$ is the estimate of β_{SP} obtained using data from adequate and well-controlled trials evaluating the effect of STD relative to placebo. As a result, these two estimates are statistically independent. The imputed placebo approach is based on a natural estimate of the effect of EXP relative to placebo:

$$\hat{\beta}_{\mathrm{EP}} = \hat{\beta}_{\mathrm{ES}} + \hat{\beta}_{\mathrm{SP}} = \hat{\beta}_{\mathrm{ES}} - \hat{\beta}_{\mathrm{PS}} \tag{A1}$$

where $\hat{\beta}_{\mathrm{PS}}$ denotes the estimate of the effect of placebo relative to STD. Due to the statistical independence of the estimates, the variance of the sum is the sum of the variances; i.e.

$$\operatorname{var}\hat{\beta}_{\mathrm{EP}} = \operatorname{var}\hat{\beta}_{\mathrm{ES}} + \operatorname{var}\hat{\beta}_{\mathrm{PS}} \tag{A2}$$

DOI: 10.1002/sim

The one-sided $(1-\alpha)$ per cent confidence interval for β_{EP} using this imputed placebo approach has upper bound

$$(\hat{\beta}_{\mathrm{ES}} - \hat{\beta}_{\mathrm{PS}}) + Z_{1-\alpha}(\widehat{\mathrm{var}}\hat{\beta}_{\mathrm{ES}} + \widehat{\mathrm{var}}\hat{\beta}_{\mathrm{PS}})^{1/2} \tag{A3}$$

where $Z_{1-\alpha}$ denotes the $1-\alpha$ quantile of the standard normal distribution, so $Z_{1-\alpha}=1.96$ in the common setting when $\alpha=0.025$, and where $\widehat{\mathrm{var}}\hat{\beta}$ denotes the estimate of var $\hat{\beta}$.

When the hypothesis that $\beta_{\mathrm{EP}}=0$ holds, then asymptotically, with probability α,

$$\hat{\beta}_{\mathrm{ES}} + Z_{1-\alpha}(\widehat{\mathrm{var}}\,\hat{\beta}_{\mathrm{ES}})^{1/2} < \hat{\beta}_{\mathrm{PS}} - Z_{1-\alpha}[(\widehat{\mathrm{var}}\,\hat{\beta}_{\mathrm{ES}} + \widehat{\mathrm{var}}\,\hat{\beta}_{\mathrm{PS}})^{1/2} - (\widehat{\mathrm{var}}\,\hat{\beta}_{\mathrm{ES}})^{1/2}] \tag{A4}$$

When EXP preserves more than a proportion **p** of the effect of STD, then $\beta_{\mathrm{EP}}<\mathbf{p}\beta_{\mathrm{SP}}$ and, in turn,

$$\beta_{\mathrm{ES}} - (1-\mathbf{p})\beta_{\mathrm{PS}} < 0 \tag{A5}$$

Following a derivation similar to that above, it follows that when the hypothesis $\beta_{\mathrm{ES}}-(1-\mathbf{p})\beta_{\mathrm{PS}}=0$ holds, then asymptotically, with probability α,

$$\hat{\beta}_{\mathrm{ES}}+Z_{1-\alpha}(\widehat{\mathrm{var}}\,\hat{\beta}_{\mathrm{ES}})^{1/2} < (1-\mathbf{p})\hat{\beta}_{\mathrm{PS}}-Z_{1-\alpha}[(\widehat{\mathrm{var}}\,\hat{\beta}_{\mathrm{ES}}+(1-\mathbf{p})^2\widehat{\mathrm{var}}\,\hat{\beta}_{\mathrm{PS}})^{1/2}-(\widehat{\mathrm{var}}\,\hat{\beta}_{\mathrm{ES}})^{1/2}] \tag{A6}$$

If one could presume the validity of the constancy assumption, then the right-hand side of the inequality in (A6) could be used to formulate the NI margin when one wishes to address whether EXP preserves more than a proportion **p** of the effect of STD [32].

To add an adjustment for the uncertainty of the validity of the constancy assumption, the CBER at FDA introduced the '95–95' method (with its name based on the classical setting where $\alpha=0.025$) for the formulation of the margin. That margin is given by

$$(1-\mathbf{p})\hat{\beta}_{\mathrm{PS}} - Z_{1-\alpha}[((1-\mathbf{p})^2\widehat{\mathrm{var}}\,\hat{\beta}_{\mathrm{PS}})^{1/2}] \tag{A7}$$

By taking the difference between (A7) and the right-hand side of the inequality in (A6), it follows that the added adjustment in the margin that is made by the '95–95' method to address the uncertainty of the validity of the constancy assumption is the non-negative term

$$Z_{1-\alpha}\{[A+B+2\sqrt{A}\sqrt{B}]^{1/2} - [A+B]^{1/2}\} \tag{A8}$$

where $A=\widehat{\mathrm{var}}\,\hat{\beta}_{\mathrm{ES}}$ and where $B=(1-\mathbf{p})^2\,\widehat{\mathrm{var}}\,\hat{\beta}_{\mathrm{PS}}$.

Note that to obtain a margin for the hazard ratio or odds ratio, the expression in (A7) or on the right hand side of (A6) must be exponentiated.

ACKNOWLEDGEMENTS

The source of financial support for research described in this article is an NIH/NIAID grant entitled 'Statistical Issues in AIDS Research' (R37 AI 29168). The author is very appreciative of detailed comments on earlier drafts of this article provided by the referees, and by Susan Ellenberg and Katherine Odem-Davis.

REFERENCES

1. ICH E-9. *International Conference on Harmonisation*: *Statistical Principles for Clinical Trials*. Published in the Federal Register of 9 May 1997 (62 FR 25712).
2. ICH E-10. *International Conference on Harmonisation*: *Choice of Control Group in Clinical Trials*. Published U.S. May 2001. http://www.fda.gov/cder/guidance/index.htm

3. Temple R. Difficulties in evaluating positive control trials. *Proceedings of the American Statistical Association*, *Biopharmaceutical Section*, 1983; **Section I**:1–7.
4. Fleming TR. Treatment evaluation in active control studies. *Cancer Treatment Reports* 1987; **71**:1061–1065.
5. Hung HMJ, Wang SJ, Tsong Y, Lawrence J, O'Neill RT. Some fundamental issues with non-inferiority testing in active controlled trials. *Statistics in Medicine* 2003; **22**:213–225.
6. Hung HMJ, Wang SJ, O'Neill RT. A regulatory perspective on choice of margin and statistical inference issue in non-inferiority trials. *Biometrical Journal* 2005; **1**:28–36.
7. Fleming TR. Evaluation of active control trials in acquired immune deficiency syndrome. *Journal of AIDS* 1990; **3**(Suppl. 2):82–87.
8. Temple R, Ellenberg S. Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 1: Ethical and scientific issues. *Annals of Internal Medicine* 2000; **133**:455–463.
9. Ellenberg S, Temple R. Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 2: Practical issues and specific cases. *Annals of Internal Medicine* 2000; **133**:464–470.
10. Fleming TR. Design and interpretation of equivalence trials. *American Heart Journal* 2000; **139**:S171–S176.
11. Hasselblad V, Kong DF. Statistical methods for comparison to placebo in active-controlled trials. *Drug Information Journal* 2001; **35**:435–449.
12. Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJW, the CONSORT Group. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT Statement. *JAMA* 2006; **295**:1152–1160.
13. Kaul S, Diamond GA. Good enough: a primer on the analysis and interpretation of noninferiority trials. *Annals of Internal Medicine* 2006; **145**:62–69.
14. Organization to Assess Strategies for Ischemic Syndromes (OASIS-2) Investigators. Effects of recombinant hirudin (lepirudin) compared with heparin on death, myocardial infarction, refractory angina, and revascularization procedures in patients with acute myocardial ischaemia without ST elevation: a randomized trial. *Lancet* 1999; **353**:429–438.
15. FDA Cardio-renal Drugs Advisory Committee Briefing Document. Hirudin in acute coronary syndrome. *90th Meeting of the FDA Cardio-renal Advisory Committee*, 2 May 2000. www.fda.gov/ohrms/dockets/ac/00/transcripts/3612t2.rtf
16. FDA Oncology Drugs Advisory Committee Briefing Document. NDA 21-677. *Alimta as Single Agent is Indicated for Treatment of Patients with Locally Advanced or Metastatic Non-small Cell Lung Cancer* (*NSCLC*) *After Prior Chemotherapy*. Eli Lilly and Company, 27 July 2004. www.fda.gov/ohrms/dockets/ac/04/briefing/2004-4060b1.htm
17. TAX 317, Shepherd FA, Dancey J *et al*. Prospective randomized trial of docetaxel versus best supportive care in patients with non-small-cell lung cancer previously treated with platinum-based chemotherapy. *Journal of Clinical Oncology* 2000; **18**(10):2095–2103.
18. TAX 320, Fassella FV, DeVore R *et al*. Randomized phase III trial of docetaxel versus vinorelbine or ifosfamide in patients with advanced non-small-cell lung cancer previously treated with platinum containing chemotherapy regimens. *Journal of Clinical Oncology* 2000; **18**(12):2354–2362.
19. Lincoff AM, Bittl JA *et al*. Bivalirudin and provisional glycoprotein IIb/IIIa blockade compared with heparin and planned glycoprotein IIb/IIIa blockade during percutaneous coronary intervention: REPLACE 2 randomized trial. *JAMA* 2003; **289**(7):853–863.
20. EPISTENT Investigators. Randomized placebo-controlled and balloon-angioplasty-controlled trial to assess safety of coronary stenting with use of platelet glycoprotein IIb/IIIa blockade. *Lancet* 1998; **352**:87–92.
21. ESPRIT Investigators. Novel dosing regimen of eptifibatide in planned coronary stent implantation (ESPRIT): a randomized, placebo-controlled trial. *Lancet* 2000; **356**:2037–2044.
22. Snapinn SM. Alternatives for discounting in the analysis of noninferiority trials. *Journal of Biopharmaceutical Statistics* 2004; **14**(2):263–273.
23. FDA Anti-Infective Drugs Advisory Committee Briefing Information. *Approach for Selection of Delta in Non-inferiority* (*Equivalence*) *Clinical Trials*. 19 February 2002.
24. Pizzo PA, Robichaud KJ, Gill FA *et al*. Empiric antibiotic and antifungal therapy for cancer patients with prolonged fever and granulocytopenia. *American Journal of Medicine* 1982; **72**:101–111.
25. EORTC International Antimicrobial Therapy Cooperative Group. Empiric antifungal therapy in febrile granulocytopenic patients. *American Journal of Medicine* 1989; **86**:668–672.
26. Walsh TJ, Finberg RW, Arndt C *et al*. Liposomal amphotericin B for empirical therapy in patients with persistent fever and neutropenia. National Institute of Allergy and Infectious Diseases Mycoses Study Group. *New England Journal of Medicine* 1999; **340**:764–771.

DOI: 10.1002/sim

27. Walsh TJ, Pappas P, Winston DJ *et al*. Voriconazole compared with liposomal amphotericin B for empirical antifungal therapy in patients with neutropenia and persistent fever. *New England Journal of Medicine* 2002; **346**(4):225–234.
28. de Pauw BE, Sable CA, Walsh TJ, Lupinacci RJ, Bourque MR, Wise BA, Nguyen BY, DiNubile MJ, Teppler H. Impact of alternate definitions of fever resolution on the composite endpoint in clinical trials of empirical antifungal therapy for neutropenic patients with persistent fever: analysis of results from the Caspofungin Empirical Therapy Study. *Transplant Infectious Disease* 2006; **8**:31–37.
29. FDA Anti-Infective Drugs Advisory Committee. Factive (gemifloxacin mesylate) Tablets for treatment of acute bacterial sinusitis. *Transcript for 12 September 2006 Meeting*. http://www.fda.gov/ohrms/dockets/ac/cder06.html#AntiInfective
30. FDA Anti-Infective Drugs Advisory Committee. KETEK (telithromycin) for acute bacterial exacerbations of chronic bronchitis, acute bacterial sinusitis, and community acquired pneumonia. *Transcript for 14–15 December 2006 Meeting*. http://www.fda.gov/ohrms/dockets/ac/cder06.html#AntiInfective
31. Jorgensen AW, Hilden J, Gotzsche PC. Cochrane reviews compared with industry supported meta-analysis and other meta-analysis of the same drugs: systematic review. *British Medical Journal* 2006; **333**:782–786.
32. 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.

DOI: 10.1002/sim