Published in final edited form as:

J Biopharm Stat. 2015 ; 25(1): 206–225. doi:10.1080/10543406.2014.923738.

On Robustness of Non-Inferiority Clinical Trial Designs against Bias, Variability, and Non-Constancy

Qing Liu¹, Yulan Li², and Katherine Odem-Davis³

¹AbacusCloud, LLC., Long Valley, NJ, U.S.A

²Novartis Pharmaceutical Corporation, East Hanover, NJ, U.S.A

³University of Washington, Seattle, WA, U.S.A

Abstract

The regulatory guidelines on non-inferiority trials emphasize constancy not only in the treatment effect over time but also in the trial design, clinical practice, and quality of the trial conduct and execution. In practice, the constancy assumption is generally impossible to justify; often there are clear reasons to expect a loss of efficacy over time. There are also concerns about the inherent and publication bias in the historical data, and various sources of selection bias in the non-inferiority trial design. Thus, a conservative non-inferiority margin is often considered. However, different non-inferiority margin approaches are largely evaluated under the assumption of constancy and absence of bias, and therefore, controversies arise and are unresolved on the necessary degree of conservativeness. We develop a framework to quantify the robustness of any non-inferiority margin approach against inherent and publication bias in historical data, selection bias in trial design, non-constancy in reference effects. We introduce a *consistency principle* to address variability in the historical data. We control across-trial *conditional* error rates given a final non-inferiority trial design over a design specific robust range for reference effects. Following a *conditionality principle*, we provide a theoretical justification of the framework and the conditions for controlling across-trial *unconditional* type 1 error rates. We raise the issue of inherent bias in historical data with an illustrative example.

1 Regulatory Background

There are various statistical approaches for determining the non-inferiority (NI) margin for a non-inferiority trial. Depending on the tightness of the margins, some appear more “conservative” than others. There are no objective criteria to decide the degree of conservativeness necessary to ensure that valid conclusions can be reached following the completion of the trial.

Many existing statistical approaches make an explicit constancy assumption. The CHMP guideline (2006) on non-inferiority trials, which is referred to as the CHMP Guidance herein, requires constancy in the trial design and clinical practice and constancy in the treatment effect over time. The draft guidance issued by the U. S. Food and Drug

Administration (FDA) on non-inferiority guideline (2010), which we refer to as the FDA Guidance, gives a more specific definition of the constancy assumption: “*The conclusion that [historical evidence of sensitivity to drug effects] can be used to choose M_1 for the new NI study can be reached only when it is possible to conclude that the NI study is sufficiently similar to the past studies with respect to all important study design and conduct features that might influence the effect size of the active control.*” Furthermore, the FDA Guidance provides design and clinical details which may lead to the constancy assumption being violated. These include “*the characteristics of the patient population, important concomitant treatments, definitions and ascertainment of study endpoints, dose of active control, entry criteria, and analytic approaches*”. The FDA Guidance emphasizes that “*As a general matter, the historical and new NI studies should be as close to identical as possible in all important aspects.*”

In practice, the constancy assumption is generally very difficult to justify; often there are clear reasons to expect a loss of efficacy over time due to biological, clinical, or manufacturing changes. For example, the FDA Guidance raises the issue of possible lack of constancy in cardiovascular outcome trials. For biosimilarity clinical developments, as explained by Li et al. (2012), the constancy assumption is not expected to hold due to differences in raw materials, periodical changes in the manufacturing processes, development of neutralizing anti-bodies to the biologic product, and medical advances in general (e.g., diagnosis or standard care). A recent experience demonstrates that a regulatory agency (i.e., FDA) could grant a drug approval, notwithstanding non-constancy, by which differences in the control group were observed from historical trials (Fleming and Emerson, 2011).

In addition to non-constancy, there are also various sources of inherent bias in the historical data. A common source is due to missing data. Currently, emphasis on missing data has been on its impacts on type 1 error rates. Missing data bias on estimates and confidence intervals of the reference effects from historical trial poses difficulty in determining margins for future non-inferiority trials. Bias may also arise when the evaluation of subjective clinical response is left to local investigators, rather than a centralized adjudication committee. As shown later in Section 4, local tumor assessment in oncology trials has lead to substantial bias in the historical data.

The CHMP Guidance states that “*It may be that studies with a ‘positive’ outcome are more likely to be published than those with disappointing results. If it seems possible that such publication bias exists, a conservative approach should be taken in producing the historical confidence interval for reference versus placebo.*” Rothmann et al. (pp. 62–65, 2012) use “*statistical significance bias*” to reflect that publication bias is the result of decision to publish upon positive statistical significance test. A regulatory approval process further exacerbates publication bias. A reference drug would not have been approved if it has failed to demonstrate efficacy over placebo; therefore, either the reference drug would not be an appropriate comparator or the new drug would not even be developed. The latter case is best illustrated in the highly active field of biosimilar drug development following the Biologics Price Competition and Innovation (BPCI) Act of 2009 by the U. S. Congress. Even if the reference is approved, a small reference effect would also prevent non-inferiority trial with a

new drug from being conducted. For the non-inferiority trial to be practical, a statistically significant difference must also be clinically meaningful. For simplicity of description, we use the term “*minimum reference effect bias*” to account for publication bias, statistical significance bias, and selection of a clinically meaningful difference, say $\delta_{\min}$, in the reference effect.

Selection bias could be introduced during the design of non-inferiority trial. The CHMP Guidance recommends a systematic literature review to identify relevant historical trials and that “*The criteria used for selecting which of the available studies to include should be thoroughly documented so that it is clear that, as far as is possible, an unbiased selection of studies was made.*” Even with proper selection of the historical trials, bias could also be introduced due to choice of control and the ad-hoc nature of the statistical analysis of the historical data (Davis, 2010).

Thus, with the presence of inherent and publication bias in the historical data, selection bias in the non-inferiority trial design, and non-constancy in reference effects over time, there is no guarantee that for a non-inferiority trial, all important design or conduct features will be similar to historical trials. In recognition of these concerns that often arise in non-inferiority trials, it is important to examine if, or under what conditions, the non-inferiority test is a valid statistical inference that would lead to the clinical conclusion that the new drug is efficacious.

2 Motivation

In an effort to address the validity of the statistical inference via the 95% confidence interval, we take the first clue from the following intriguing statement by the CHMP Guidance: “*it is important to remember that, although 95% confidence and 5% significance have become commonly accepted, there is still a possibility of false positive results even using this significance level.*” With large samples, the 95% confidence intervals are known to have the correct probability of coverage, i.e., 95%. If the type 1 error rate for the underlying null hypothesis can be evaluated, is it possible to identify the conditions under which this type 1 error rate is not inflated? It is the pursuit of answers to this fundamental question that motivated our research. Specifically, we are interested in the robustness of the non-inferiority test against inherent and publication bias and variability in the historical data, and selection bias in the non-inferiority trial design, and the violation of the constancy assumption.

For simplicity, we consider a future non-inferiority trial where patients are randomized to receive a new test drug (T) or a reference drug (R) in a one-to-one randomization ratio with the sample size n per treatment group. It is assumed that the primary endpoint is binary and equal to 1 if treatment is successful and 0 otherwise (e.g., a treatment success is defined as a patient completing the trial and not deteriorating in the disease). Let the response rates of the test and reference drug be μ_T and μ_R , respectively. Notice that the trial does not have a placebo control. Following both the CHMP and the FDA Guidance, one statistical objective of the trial is to test against the null hypothesis $\mathcal{H}_0$ in favor of the alternative hypothesis $\mathcal{H}_A$ where

$$\mathcal{H}_0: \mu_T - \mu_P \leq 0 \text{ and } \mathcal{H}_A: \mu_T - \mu_P > 0, \quad (1)$$

for which the response rate μ_T of the new test drug (T) is measured against the response rate μ_P of a “putative” placebo control (P). Another important hypothesis testing framework is designed to rule out loss of a pre-specified fraction of the effect of standard. In this paper we introduce concepts under the minimal criteria of superiority to placebo, and we will expand to the more general framework of preservation of effect in future work.

The null hypothesis $\mathcal{H}_0$ is commonly tested via a non-inferiority test with a “fixed” margin, which is described below. Let $\hat{\mu}_T$ and $\hat{\mu}_R$ be the observed response rates of the test and reference drug, and $\hat{\sigma}_{TR}$ be the observed standard error of the treatment difference $\hat{\mu}_T - \hat{\mu}_R$. For a one-sided type 1 error rate α , usually taking the value of .025 for a single clinical trial, the critical value of the standard normal distribution at the cumulative probability $1 - \alpha$ is denoted by z_α . For the hypothesis testing problem in equation (1), let δ_{NI} be a pre-specified, and therefore a fixed, non-inferiority margin such that $\delta_{NI} > 0$. It is required that the lower $100(1 - \alpha)\%$ confidence bound for $\mu_T - \mu_R$ is at least as large as $-\delta_{NI}$. That is,

$$(\hat{\mu}_T - \hat{\mu}_R) - z_\alpha \hat{\sigma}_{TR} \geq -\delta_{NI}. \quad (2)$$

The problem is that this test is necessarily at most an indirect test against the null hypothesis $\mathcal{H}_0$. Depending on the method for δ_{NI} , issues regarding inherent and publication bias in the historical data, selection bias in the non-inferiority trial design, and non-constancy in reference effects over time may or may not have been adequately addressed. Therefore, by the CHMP Guidance, there is no apparent reason to believe that this indirect test controls the type 1 error rate for rejecting the underlying null hypothesis $\mathcal{H}_0$ at a specified significance level α .

The FDA Guidance uses the traditional terminology of type 1 error, which is also referred to as across-trial or unconditional type 1 error by Hung et al. (2003), Lawrence (2005), Rothmann (2005) and Hung et al. (2007). These authors investigate the impacts of the sampling variability in the historical trial data or the future non-inferiority trial on the type 1 error rates to compare different non-inferiority margin approaches under the assumptions of constancy and absence of bias. Hung et al. (2007) further suggest that the type 1 error rates “*should also be evaluated assuming that the constancy assumption is violated to a plausible range of extent.*”

In addition to allowing non-constancy in reference effects, assessing the type 1 error rates should also address issues of inherent and publication bias in historical data and selection bias in the non-inferiority trial design. Fleming et al. (2011) describe some sources of bias and suggest that “*the non-inferiority margin should be formulated using adjustments to account for bias ...*” Following this idea, Odem-Davis and Fleming (2013) describe a bias-adjustment introduced by Rothmann et al. (2003), and they propose similar non-inferiority margin which accounts for uncertainty in the bias-correction. Recently, Liu and Li (2014) propose a predictive bound approach by which the reliability of the non-inferiority test can

be controlled by the choice of predictive confidence level. However, the amount of the correction or the predictive confidence level is yet to be determined.

To improve upon these previous works, we present in Section 3 a general framework for controlling the across-trial unconditional type 1 error rate of the non-inferiority test in equation (2) for testing against the null hypothesis $\mathcal{H}_0$ given in equation (1). The framework encompasses issues of inherent and publication bias and variability in the historical data, selection bias in the non-inferiority trial design, and non-constancy in the reference effects. Different issues are addressed with different statistical tools.

First, the theoretical foundation rests on the conditionality principle (Cox and Hinkley, p. 38, 1974) and on the consistency principle (Sections 3.6 and 3.7):

- i. To account for a certain level of *minimum reference effect bias* (see Section 1), we introduce across-trial *conditional* type 1 error rates given the final decision made about the non-inferiority trial design (Section 3.3). Such decision can involve various aspects of the clinical and statistical considerations, including but not limited to the choice of ad-hoc analysis of the historical data, all of which may contribute to selection bias. We propose to control the across-trial conditional type 1 error rates at a pre-specified α -level. Then by the adaptive design theory of Liu, Proschan and Pledger (2002), we establish that the across-trial unconditional type 1 error rates are also controlled at the specified α level.
- ii. The uncertainty in the underlying statistical model due to sampling variability of the historical data is addressed via a *consistency principle* by which all statistical models that are consistent with the historical data are used to evaluate the across-trial conditional type 1 error rates.

Next, the degree of robustness, or *robust range*, of any non-inferiority margin approach for a given decision on trial design is then defined by setting the across-trial conditional type 1 error rates at the pre-specified α level. We define the robust range as the decrease in the difference of the reference effect size of the non-inferiority trial from that of the historical trial(s) for which the α level holds. For example, a robust range of 5% means that the non-inferiority test controls the unconditional across-trial type 1 error rate even if the reference effect size in the non-inferiority trial is at most 5% less than that of the historical trial(s). With this robust range, it is now possible to decide the degree of robustness of any non-inferiority margin approach, which in turn allows rational and objective comparison of different non-inferiority margin approaches.

As different decisions on the non-inferiority trial design may be subject to different non-constancy in the reference effects and degrees of selection bias, different robust ranges may be required. The robust range expands the concept of “*plausible range of extent*” suggested by Hung et al. (2007) as it accounts for not only the degree of non-constancy in reference effects but also plausible inherent bias in the historical data and selection bias in the trial design. Once a desired robust range is chosen, the predictive confidence level of Liu and Li (2014) or the amount of bias correction of Odem-Davis and Fleming (2013) can then be determined.

Then, using principles (i) and (ii), we account for bias in the following ways:

- A. *Minimum reference bias* is addressed by a worst case simulation model that is consistent with the historical data, accounting for the uncertainty (or “variability”) in the historical data and a *minimum reference effect* below which a non-inferiority trial would not be conducted due to cost considerations.
- B. Residual bias (i.e., inherent bias in historical data and selection bias in the trial design) and non-constancy are addressed by calculating the minimum robust range among all simulation models consistent with estimates from the worst case simulation model from A.

In Section 4, we raise the issue of inherent bias in historical data and illustrate its impact on non-inferiority trial design with an actual example following the proposed framework. The paper focuses on settings where the goal is to establish efficacy over a putative placebo control. For simplicity, we consider binary endpoints. In Section 5, we list critical areas of future research.

3 Robustness Against Uncertainty

3.1 A False Paradox

From Liu and Li (2014), it is seen that different non-inferiority margin approaches can result in designs with very different margins and sample sizes. Most approaches, including the lower predictive bound approach developed by Liu and Li (2014), make the explicit constancy assumption in reference effects over time as well as the *implicit* assumptions of absence of inherent and publication bias in the historical data and absence of selection bias in the trial design. If these assumptions were indeed true, then it would seem that the only characterization that can be attributed to any approach is its conservativeness, which is not a satisfactory characterization of properties of the approach.

Because of concerns about the constancy assumption and selection bias, all regulatory guidelines state that the non-inferiority margin chosen for a given trial must be properly tight. This motivates us to evaluate the robustness of the non-inferiority test against non-constancy and various sources of bias for which the margin is derived by different approaches. The idea of evaluating the robustness of a non-inferiority margin approach may seem paradoxical when the approach makes an explicit assumption of constancy in reference effects and implicit assumption of absence of bias. However, under these assumptions a lower confidence or predictive bound obviously underestimates the reference effect size. Consequently, choosing a lower confidence or predictive bound as the non-inferiority margin may lead to a non-inferiority test that is robust against a certain level of non-constancy and bias.

We recommend to describe the robustness, rather than the conservativeness, of the non-inferiority test with a given approach for deriving the non-inferiority margin and the final decision on trial design. This is both conceptually and practically appealing as it can give rise to a proper margin once an appropriate degree of robustness is chosen, provided of course that the underlying hypotheses are specified at the onset.

3.2 Conditional Type 1 Error Rate

The key to developing a framework for evaluating robustness is to understand fully the correct null hypothesis of the non-inferiority test given in equation (2). For a given non-inferiority margin δ_{NI} , irrespective of how it is derived from the historical data, the non-inferiority test in equation (2) is used to test against the “working” conditional null hypothesis

$$\mathcal{H}_0(\delta_{NI}): \mu_T - \mu_R \leq -\delta_{NI}.$$

Assuming large samples, the type 1 error rate of the non-inferiority test against this conditional null hypothesis is α , which is universally known. However, as indicated in Section 2, the non-inferiority test in expression (2) is used as an indirect test against null hypothesis $\mathcal{H}_0: \mu_T - \mu_P = 0$ in favor of the alternative hypothesis $\mathcal{H}_A: \mu_T - \mu_P > 0$ in equation (1). Thus, the null hypothesis of relevance for the non-inferiority test in equation (2) is $\mathcal{H}_0: \mu_T - \mu_P = 0$. This was first suggested by Hung et al. (2003), and was later explicitly stated by the CHMP Guidance.

Let $\Delta = \mu_R - \mu_P$, which is the effect of the reference drug (R) against the putative placebo control (P) of the future non-inferiority trial. Then $\mathcal{H}_0: \mu_T - \mu_P = 0$ is equivalent to the null hypothesis

$$\mathcal{H}_0(\Delta): (\mu_T - \mu_R) + \Delta \leq 0. \quad (3)$$

Following the ideas of Lawrence (2005) or Hung et al. (2007), it is easy to derive that under the null hypothesis

$$H_0(\Delta): (\mu_T - \mu_R) + \Delta = 0$$

the conditional type 1 error rate of the fixed margin non-inferiority test in equation (2) given the non-inferiority margin δ_{NI} is approximately

$$CE(\Delta; \delta_{NI}, \hat{\sigma}_{TR}) = 1 - \Phi\{z_\alpha + (\Delta - \delta_{NI})/\hat{\sigma}_{TR}\} \quad (4)$$

for large samples.

It is immediately obvious that if the non-inferiority margin δ_{NI} is larger than Δ , the non-inferiority test in equation (2) inflates the conditional type 1 error rate, i.e.,

$$CE(\Delta; \delta_{NI}, \hat{\sigma}_{TR}) > \alpha.$$

Thus, the reliability of the non-inferiority test as an indirect test against the null hypothesis $\mathcal{H}_0$ in equation (3) depends heavily on how the non-inferiority margin δ_{NI} is chosen. For

example, if the historical point estimate, say, the maximum likelihood estimate (MLE), is used as the non-inferiority margin, then the probability that the conditional type 1 error rate $CE(\delta_{NI}, \hat{\sigma}_{TR})$ of the non-inferiority test is inflated is at least 50%, assuming that the historical point estimate is unbiased for the reference effect of the future non-inferiority trial. Another approach is to use the 97.5% lower confidence bound as the non-inferiority margin. Again under the constancy assumption, the probability of inflation is at least 2.5%, as the use of lower confidence bound does not take into consideration the selection bias of the historical data. On the other hand, the non-inferiority margin δ_{NI} by the 97.5% lower predictive bound approach (Liu and Li, 2014) may ensure that the probability of inflation is not more than 2.5%. Liu and Li (2014) also provide the synthesis approach for fixed non-inferiority margins. We show in Section 4.5 that the probability of inflation is around 30% with the synthesis approach.

It is also seen that increasing the standard error $\hat{\sigma}_{TR}$ can also increase the conditional type 1 error rate $CE(\delta_{NI}, \hat{\sigma}_{TR})$. It is of concern for analysis using the intent-to-treat (ITT) population if its observed standard error is larger than what is used for the trial design. This concern is often dealt with via a per-protocol (PP) population to reduce the standard error. However, the definition of a PP population usually involves factors that could introduce selection bias. Thus, using a PP population does not address the issue (FDA Guidance; Fleming et al., 2011; Fleming and Emerson, 2011). In Section 3.3, we show how to control the unconditional type 1 error rate of the non-inferiority test in equation (2) for the ITT population.

3.3 Across-trial Conditional Type 1 Error Rate

The conditional type 1 error rate $CE(\delta_{NI}, \hat{\sigma}_{TR})$ given in equation (4) depends on the actual value of the non-inferiority margin δ_{NI} and ignores altogether the process of selecting the historical data and the approach for deriving the non-inferiority margin and its distributional properties. Lawrence (2005) shows that with a certain choice of the reference effect in the non-inferiority trial and large variance ratio, the conditional error rate of the synthesis method can reach its upper limit of 50%. This leads to his conclusion that nearly half of the nine inferior drugs would be on the market following non-inferiority trials with the synthesis method. By equation (4), the conditional type 1 error rate $CE(\delta_{NI}, \hat{\sigma}_{TR})$ is an increasing function of the margin δ_{NI} , converging to 1 as δ_{NI} converges to $+\infty$. A similar calculation is also shown by Rothmann (2005). Thus, it is easy to construct numerical examples under a non-constancy assumption and large sample to show that the conditional error rate can be as high as 90% when the non-inferiority margin is set to be the lower 97.5% confidence bound of the historical data. Using the multiple drug example by Lawrence (2005), nearly eight of the nine non-inferior drugs would be on the market following the non-inferiority trials with the fixed margin non-inferiority test. This would suggest that the fixed margin test in equation (2), where the margin is derived from the 97.5% lower confidence bound, is more problematic than the synthesis approach.

To resolve this false paradox, we follow Lawrence (2005) and Hung et al. (2007) and adopt the concept of averaging the within-trial conditional type 1 error rate with respect to a distribution of the non-inferiority margin following a particular non-inferiority margin

approach and a given analysis of historical data. There are, however, two critical differences:

1. To account for a minimum reference effect bias (see Section 1), we use the conditional distribution of the non-inferiority margin given $\delta_{NI} \geq \delta_{\min}$ for some $\delta_{\min} > 0$.
2. Following the suggestion by Hung et al. (2007), we permit non-constancy in reference effect, as well as selection bias in the non-inferiority trial design.

Without a better terminology, we call this average of conditional type 1 error rates with respect to the conditional distribution of the non-inferiority margin as *across-trial conditional type 1 error rate*. Note that in Kang and Tsong (2010), the conditional across-trial error rate is simply the conditional type 1 error rate in this paper and literature. Our use of the term across-trial conditional type 1 error rate reflects that we integrate over a conditional distribution of the margin, so it is both across trial (over the distribution of the margin) and conditional (on positive historical trial).

Let $\sigma_{\max}$ be a maximum standard error for $\hat{\sigma}_{TR}$. For the difference in binomial proportions, we choose $\sigma_{\max} = (.5/n)^{1/2}$. Because $CE(\Delta; \delta_{NI}, \hat{\sigma}_{TR})$ is increasing in $\hat{\sigma}_{TR}$ for $\delta_{NI} < 0$, we have that

$$CE(\Delta; \delta_{NI}, \hat{\sigma}_{TR}) \leq CE(\Delta; \delta_{NI}, \sigma_{\max}).$$

It is easy to evaluate $CE(\Delta; \delta_{NI}, \sigma_{\max})$ when designing the future non-inferiority trial. In contrast, $CE(\Delta; \delta_{NI}, \hat{\sigma}_{TR})$ depends on the unknown $\hat{\sigma}_{TR}$ when the non-inferiority trial is being designed.

Denote by $\mathcal{M}$ the underlying probability model incorporating the distribution and analysis of the historical data, and the selected statistical approach for calculating the non-inferiority margin δ_{NI} . Let $\psi_{\mathcal{M}}(\cdot)$ be the conditional probability density function of the resulting non-inferiority margin δ_{NI} given that a statistically significant and clinically meaningful difference $\delta_{\min}$ is achieved with the historical data (see Section 3.4). Then the across-trial conditional type 1 error rate is given by

$$C(\Delta; \mathcal{M}) = \int CE(\Delta; \delta_{NI}, \sigma_{\max}) \psi_{\mathcal{M}}(\delta_{NI}) d\delta_{NI}. \quad (5)$$

For a given probability model $\mathcal{M}$, we can determine the set of Δ values such that

$$C(\Delta; \mathcal{M}) \leq \alpha.$$

This set provides the robust range (see Section 3.5) for the reference effect Δ of the future non-inferiority trial against non-constancy in reference effects, inherent and publication bias in historical data, and selection bias in the final trial design, by which the non-inferiority test

in expression (2) with a particular approach for deriving the non-inferiority margin controls the across-trial conditional type 1 error rate.

Note that because the maximum standard error $\sigma_{\max}$ is used in evaluating $C(\delta; \mathcal{M})$, the across-trial conditional type 1 error rate for the non-inferiority test in expression (2) with the ITT population is controlled for any reference effect within the robust range. This provides a theoretical justification for the ITT non-inferiority suggested by Fleming and Emerson (2011) to avoid bias in per-protocol analysis.

3.4 Simulation Model

The challenge in calculating the across-trial conditional type 1 error rate is how to define the probability model $\mathcal{M}$ that incorporates the distribution and analysis of the historical data, and the statistical approach for calculating the non-inferiority margin $\delta_{\mathcal{M}}$. To simplify matters, we consider in this paper only the setting that the analysis of the historical data is based on the pre-specified statistical analysis plan (SAP) from a single historical trial.

We now develop a simulation model to incorporate the minimum reference effect bias. We use the simulation model as a substitute for the true probability model $\mathcal{M}$ to evaluate the across-trial conditional type 1 error rate $C(\delta; \mathcal{M})$ in equation (5). The simulation model matches the context of the historical data. For example, if the historical data arise from a single trial, then the distribution model with the same endpoint type (e.g., binary) as well as ITT sample sizes of the reference and placebo groups is used. We simulate data sets from the distribution model and toss out any data set if the 97.5% lower confidence bound is less than a minimum effect size $\delta_{\min}$. We then evaluate the across-trial conditional type 1 error rate $C(\delta; \mathcal{M})$ for different values of the reference effect δ of the future non-inferiority trial.

Note that the reference effect under the simulation model for the historical trial is not required to match the reference effect δ of the future non-inferiority trial. Therefore, the resulting across-trial conditional type 1 error rate $C(\delta; \mathcal{M})$ can incorporate various degrees of non-constancy and bias. Of course, the simulation model can be expanded to mimic more complex processes arising from other clinical settings (see discussion in Section 5).

3.5 Robust Range

Following the definition of $C(\delta; \mathcal{M})$ given in equation (5), it is seen that $C(\delta; \mathcal{M})$ is a decreasing function of δ . The larger the δ , the less the chance that the non-inferiority margin, following any approach, exceeds δ . An approach for deriving a non-inferiority margin is admissible if there exists a δ by which $C(\delta; \mathcal{M}) = \alpha$ where α is the pre-specified type 1 error rate. For an admissible approach, let $\delta_{\alpha, \mathcal{M}}$ be the effect size such that

$$C(\delta_{\alpha, \mathcal{M}}; \mathcal{M}) = \alpha. \quad (6)$$

Denote by $\Delta_{\mathcal{M}}$ the reference effect size for the historical trial(s) under the simulation model $\mathcal{M}$. Then the robust range is defined as

$$\lambda_{\mathcal{M}} = \Delta_{\mathcal{M}} - \delta_{\alpha, \mathcal{M}}. \quad (7)$$

Using a design in Section 4.4, we illustrate the calculation of the robust range $\lambda_{\mathcal{M}}$ with Figure 1 where the across-trial conditional type 1 error rate $C(\cdot; \mathcal{M})$ is plotted as a function of the reference effect δ for a simulation model with $\Delta_{\mathcal{M}} = .25$. It is seen that $C(\cdot; \mathcal{M})$ is a decreasing function of δ . As δ decreases to $\delta_{\alpha, \mathcal{M}} = .101$ (solid vertical line) for which $\alpha = .025$, the error equals to .025 (dotted horizontal line). Thus, for $\delta > \delta_{\alpha, \mathcal{M}} = .101$ with the given simulation model $\mathcal{M}$, the across-trial type 1 error rates are below $\alpha = .025$. This gives rise to the robust range $\lambda_{\mathcal{M}} = \Delta_{\mathcal{M}} - \delta_{\alpha, \mathcal{M}} = .25 - .101 = .149$.

The decrease of the reference effect δ compared to that of the historical trial(s) is $\Delta_{\mathcal{M}} - \delta$. As $\Delta_{\mathcal{M}} - \delta = \lambda_{\mathcal{M}}$ is equivalent to $\delta_{\alpha, \mathcal{M}}$, the robust range $\lambda_{\mathcal{M}}$ is interpreted as follows:

Provided that the simulation model $\mathcal{M}$ is true, the non-inferiority test given in expression (2) is robust against inherent and publication bias in historical data, selection bias in the non-inferiority trial design, and non-constancy in reference effects by a decrease of the reference effect δ not larger than $\lambda_{\mathcal{M}}$.

It is worth noting that the robust range does not distinguish the extent of non-constancy from that of various sources of bias. Following the CHMP Guidance, it is clear that non-constancy reflects “natural” change in the reference effects over time, independent of the presence or absence of the future non-inferiority trial; while selection bias, as seen below in details, aggregates deviations in reference effects in the decision making process for the final non-inferiority trial design. For a practical application, non-constancy and various sources of bias need to be evaluated separately, with subjective judgment, to assess if a certain robust range is adequate.

3.6 Conditionality Principle

Unlike some adaptive designs for which it is possible to specify the adaptation rules, the decision process for non-inferiority trials does not lend itself to a set of algorithms without over-simplification. The process may start with an unmet medical need to develop new drugs that have better safety profiles or public pressures through congressional legislation (e.g., BPCI Act, 2009) for developing alternatives to innovative drugs to lower the health-care costs or to expand public access. A drug company may or may not decide to develop a new drug based on its R&D portfolio or technical or commercial considerations. A clinical development team would face a huge range of clinical and statistical issues, including the study population, endpoint, choice of reference and putative control, selection or availability of historical trial and data, their method of statistical analysis for quantifying the reference effect, as well as the method for deriving the non-inferiority margins. The decision process is complex and iterative, often involving consultation with experts as well as multi-regional regulatory inputs. Each party, including a regulatory agency, has its own preference or bias. In addition, a lack of understanding of the statistical method or its misuse by both regulatory agencies and industry exacerbates the bias. Thus, the process inevitably introduces various sources of selection bias.

It is impossible without oversimplification to develop a model for the decision making process to account for various sources of selection bias. Without such a model, across-trial unconditional type 1 error rates cannot be calculated. An alternative approach is to control the across-trial conditional type 1 error rates developed in Section 3.3 over the robust range to account for a plausible extent of non-constancy and bias for a given decision made on the final non-inferiority trial design. Then, following the conditionality principle (Cox and Hinkley, p. 38, 1974), it is possible to control the across-trial unconditional type 1 error rates.

We now present a more formal measure-theoretic description. Let (Ω, F, P) be a σ -field. The sub- σ field G in F represents the totality of the historical data. Let g be a G -measurable adaptation rule (see Liu, Proschan and Pledger, 2002) representing the decision process with a countable range D . An element d in D represents a particular decision made on the final non-inferiority trial design. For a given d , let $\mathcal{M}_d$ be the underlying probability model for the historical data and $C_d(\cdot; \mathcal{M}_d)$ be a G -measurable across-trial conditional error function.

Theorem: For each $d \in D$, let $C_d(\cdot; \mathcal{M}_d) \leq \alpha$ for $\delta_{\alpha, \mathcal{M}_d}$. Then

$$E\{C_g(\Delta; \mathcal{M}_g)\} \leq \alpha \quad (8)$$

for $\max(D) = \max_{d \in D} \delta_{\alpha, \mathcal{M}_d}$.

The left side of expression (8) is the across-trial unconditional type 1 error rate. By construction, $C_g(\cdot; \mathcal{M}_g) \leq \alpha$ over the event $\{g = d\} \in G$ for $\delta_{\alpha, \mathcal{M}_d}$ any $d \in D$. Thus, $E\{C_g(\cdot; \mathcal{M}_g) | g\} \leq \alpha$ for $\max(D)$. By double expectation,

$$E\{C_g(\Delta; \mathcal{M}_g)\} = E[E\{C_g(\Delta; \mathcal{M}_g) | g\}] \leq \alpha.$$

The theorem requires that the decision process g is measurable with a countable range, which is a condition that is always met in practice. Through controlling the across-trial conditional type 1 error rate for any $d \in D$, the conditionality principle, by the theorem, ensures that the across-trial unconditional type 1 error rate is also controlled without the need of specifying the decision process g .

Following the measure-theoretical formulation, we note that non-inferiority trials are essentially adaptive trials. The fundamental difference between (adaptive) non-inferiority trials and adaptive superiority trials is the presence of the “global” requirement $\max(D)$ for governing the decision making process g . This is because non-inferiority trials are concerned with inference where the type 1 error rates are assessed at non-zero points of the reference effect size, while adaptive superiority trials are only concerned with null hypotheses of zero effect size. Thus, if $\max(D) = \max_{d \in D} \delta_{\alpha, \mathcal{M}_d}$ is larger than a plausible reference effect size, due to possibility of substantial selection bias of a single decision on the final non-inferiority design, then it is impossible in principle to control the across-trial unconditional type 1 error rates. Although selection bias has been recognized by CHMP

Guidance, Fleming et al. (2011) and Rothmann et al. (pp. 62–70, 2012), there is little *adequate* research to address various sources of selection bias, let alone to quantify the extent of selection bias.

3.7 Consistency Principle

In reality, it would be impossible to know whether a particular simulation model or its underlying effect size $\Delta_{\mathcal{M}}$ is true due to the variability of the historical data. Therefore, there is a need to evaluate the robust range $\lambda_{\mathcal{M}}$ for all simulation models with all possible values for the effect size $\Delta_{\mathcal{M}}$ which are consistent with the historical data while taking into consideration its variability. We consider a simulation model as consistent if there is a lack of statistical evidence against the model in light of the historical data. Two approaches are described in the following.

Let μ_P^* and μ_R^* be the success rates for the placebo control (P) and the reference drug (R) of the distribution model of the historical trial. Then the effect size of the simulation model $\mathcal{M}$ is $\Delta_{\mathcal{M}} = \mu_R^* - \mu_P^*$. The observed success rates are denoted by $\hat{\mu}_P^*$ and $\hat{\mu}_R^*$ for the placebo control (P) and the reference drug (R), respectively. Let $\sigma_P = \{\mu_P^*(1 - \mu_P^*)/n_P\}^{1/2}$ and $\sigma_R = \{\mu_R^*(1 - \mu_R^*)/n_R\}^{1/2}$. Then it is easy to show that $\text{Cov}(\hat{\mu}_R^* - \hat{\mu}_P^*, \hat{\mu}_P^*) = -\sigma_P^2$.

Let $B = \{(\hat{\mu}_R^* - \hat{\mu}_P^*) - \Delta_{\mathcal{M}}, \hat{\mu}_P^* - \mu_P^*\}$ and $V = \text{Var}(B)$. Then $V_{[1,1]} = \sigma_R^2 + \sigma_P^2$, $V_{[1,2]} = V_{[2,1]} = -\sigma_P^2$ and $V_{[2,2]} = \sigma_P^2$ where $V_{[i,j]}$ is the element of the matrix V at the i th row and j th column for $i = 1, 2$ and $j = 1, 2$. We use the “chi-square” statistic $\chi_{\mathcal{M}}^2 = BV^{-1}B'$ to test against the simulation model $\mathcal{M}$, for which μ_P^* and μ_R^* are the parameters of the underlying distribution model. Then through simulation, we can evaluate the p -value

$$p_{\mathcal{M}} = P_{\mathcal{M}}\{\chi_{\mathcal{M}}^2 \geq bV^{-1}b'\} \quad (9)$$

where b is the observed value of B based on the historical data. The simulation model $\mathcal{M}$ is not inconsistent with the historical data if and only if $p_{\mathcal{M}} \geq \alpha_{\mathcal{M}}$ for some significance level $\alpha_{\mathcal{M}}$, say $\alpha_{\mathcal{M}} = .05$ or $.1$.

An alternative and perhaps more easily generalizable approach is using the marginal confidence bounds for the observed placebo control rate $\hat{\mu}_P^*$ and reference effect $\hat{\mu}_R^* - \hat{\mu}_P^*$ to identify a “worst” case simulation model that is still consistent with the historical data. Then any other simulation models for which the control rate and reference effect are within the confidence bounds are consistent with the historical data. The worst case model can then be used to calculate the minimum robust range, $\lambda_{\min}$, among all consistent simulation models. This approach is illustrated later.

4 Inherent Bias in Historical Data

4.1 Issue

Existing regulatory guidelines and major research papers in the literature have so far been focusing on issues relating to various aspects of design and analysis of the non-inferiority trials. It has been assumed that although estimates of the reference effects of the historical data can be biased for the non-inferiority trial due to publication and selection bias, the historical data and analysis are themselves unbiased for the historical trials.

Sackett (1979) discusses various clinical bias in analytical research. Fleming (1997) points out several types of regulatory reviewers' bias. Liu and Chi (2010) provides a comprehensive list of sources of statistical related bias in the design and analysis of controlled clinical trials in drug development. For superiority trials, the main emphasis has been on controlling type 1 error rates of testing against the null hypotheses of absence of clinical effects; while less attention has been paid to bias in point estimates. A recent review of a non-inferiority trial design by the European Medicines Agency (EMA) identifies the difficulty of using non-inferiority margins derived from historical data with missing observations. There are many other types of bias in the historical data or analysis that can also impair the design and analysis of a non-inferiority trial. We illustrate the problem through an example involving evaluations of tumor response in oncology trials.

4.2 Non-Hodgkin's Lymphoma Trial

Blinded, independent review committees (BIRCs) have been used to determine the potential presence of bias of locally evaluated tumor response. In particular, full BIRC review has been recommended by the FDA guideline (2007) for unblinded cancer drugs and biologics trials. A recent report by Amid et al. (2010) only addresses potential bias of locally evaluated progression time, and provides a summary of the PhRMA PFS Working Group recommendations of establishing a central repository of all scans but *limiting* BIRC review of PFS events based on statistical calculation of sample size for differential discordance.

In a randomized and controlled trial to study the efficacy of an innovative biologic in patients with follicular lymphoma, PFS for the entire treatment period of the study, as well as tumor response following the induction therapy, were evaluated both locally by unblinded investigator sites and a blinded independent review committee. For the primary analysis, there is no differential discordance between treatment groups for the time-to-tumor progression (TTP) endpoint. However, a summary of the overall tumor response (OTR) data, which is extracted from a regulatory review, given in Table 1 shows systematic bias in the OTR rates. It is seen that the local OTR rate for the control is 14% higher than that of the central rate. An examination of their naive 95% confidence intervals, which are unadjusted for the correlation between local and central evaluations, reveals that the difference is due to bias in the local evaluation, in addition to the expected increased variability. By their 95% confidence intervals for the treatment difference, the presence of such bias does not affect the conclusion that the new biologic is more effective to induce OTR than the control.

4.3 Design of a Non-Inferiority Trial

Consider a double-blind biosimilarity clinical trial where patients with previously untreated advanced non-Hodgkin's lymphoma are randomized in a 1-to-1 ratio to receive either the regimen with a biosimilar (T) plus a chemotherapy (C) (T -plus- C) or the reference (R) of the innovative biologic plus the chemotherapy (C) (R -plus- C) regimen. The primary endpoint is the OTR by a BIRC review for all patients. Such BIRC review is justified as follows.

1. The FDA Guidance states that the non-inferiority margin is based on “*assessment of the quality of the NI trial, particularly looking for defects that could reduce a difference between the active control and the new drug (... , ‘bias toward the alternative’).*”
2. The CHMP Guidance assumes that “*the measurements make it possible to distinguish between desired (positive) and undesired (negative) effects.*”
3. Fleming et al. (2011) emphasize that “*Irregularities in quality of the conduct of the non-inferiority trial induce increased risk of both bias and variability.*” And the irregularities “*are even more problematic in a non-inferiority trial since they often dilute the sensitivity to true differences between the experimental intervention and Standard regimens, leading to an increased risk of falsely declaring non-inferiority in settings where the test treatment truly is clinically inferior to Standard.*”
4. The FDA guideline (2007) on clinical endpoints for cancer drugs and biologics specifically prefers blinding and recommends blinded review of objective response rate.

The historical trial was based on the comparison with the effective chemotherapy (C), as opposed to a placebo control. Following the recent FDA guideline (2010) on combination drugs, a primary objective of the biosimilarity trial is to establish superiority of T -plus- C over C . There is no recent trial with the chemotherapy C alone, and thus, an adequate robust range is necessary. With the type 1 error rate $\alpha = .025$, power 92.5% (see Li et al., 2012), and lower predictive confidence level 99.5%, the required sample size is 199 patients per treatment group with the non-inferiority margin $\delta_{NI} = .1437$.

To evaluate the minimum robust range, we consider the worst case simulation model, following the summary in Table 1 of the 95% confidence intervals by the central review, for which the control response rate 50% is taken near the upper bound while the reference effect size 25% is near the lower bound. We use the minimum reference effect $\delta_{\min} = 10\%$ when computing the across-trial conditional type 1 error rate, since we believe that a non-inferiority trial with delta less than 10% would not be funded due to cost in the motivating setting of biosimilars. The resulting minimum robust range is $\lambda_{\min} = 14.89\%$.

4.4 Impact of Bias

By Amid et al. (2010), the PhRMA PFS Working Group has not evaluated the consequence of limited BIRC review on subsequent drug or biologic development requiring a non-inferiority or biosimilarity design. Had the regulatory agency not provided the summary of

the OTR data, the non-inferiority margin for the biosimilar would be based on the local evaluations.

With the type 1 error rate $\alpha = .025$, power 92.5%, and default lower predictive confidence level 97.5%, the required sample size is 318 patients per treatment group. The non-inferiority margin is $\delta_{NI} = .106$ based on the lower predictive bound approach described in Li et al. (2012). The minimum robust range is evaluated under the worst case simulation model, following the summary in Table 1 of the 95% confidence intervals by the local review, for which the control response rate 65% is taken near the upper bound while the reference effect size 15% is near the lower bound. We use $\delta_{\min} = 10\%$ for the across-trial conditional type 1 error rate since studies with smaller margins would not be conducted in this setting. The resulting minimum robust range is $\lambda_{\min} = 4.8\%$.

As seen in Table 1, the local evaluations are negatively biased against the reference. This has direct impact on the design of subsequent non-inferiority or biosimilarity trials. The sample size has unnecessarily increased by nearly 60% with a much smaller non-inferiority margin .106. The minimum robust range has dropped from 14.9% to 4.8%. Had the central review data not been made available, it would not come to the conclusion that the 4.8% robust range would be adequate because after adjusting for the 10% local review bias, the correct robust range would be much larger. The harm would be that development of the biosimilar may be put on hold due to other commercial priorities, leaving the issue of health care cost and broader public access to effective medicines un-addressed.

4.5 Comparison of Margin Approaches

There are considerable debates on the appropriateness of the synthesis approach. It has been believed that the synthesis approach is adequate when the constancy assumption holds. However, existing literature ignores the issue of what we call in Section 3.3 “*minimum reference effect bias*”. Furthermore, there is no method to assess if the confidence bound approach for non-inferiority margins is adequate given a particular decision on the final non-inferiority design. In the following, we present comparisons of the non-inferiority designs and its minimum robust ranges for different non-inferiority margin approaches. For illustration, we use both the central and local review of the OTR of the follicular lymphoma trial. All designs use the type 1 error rate $\alpha = .025$ and power 92.5%.

The worst case simulation models in Sections 4.3 and 4.4 are used for central and local reviews, respectively. The resulting designs are summarized in Table 2 where LPB_{.995} and LPB_{.975} are used to denote designs with the lower predictive bound approach with predictive confidence level 99.5 and 97.5%. In particular, we only compare LPB_{.995} by the central review with LPB_{.975} by the local review. Designs with lower confidence bound approach at the confidence levels 99.5% and 97.5% are denoted by LCB_{.995} and LCB_{.975}. We also consider the design by the fixed margin synthesis-type approach (denoted by Syn) derived by Liu and Li (2014) from the synthesis test of Hasselblad and Kong (2001). Following Liu and Li (2014), we present the naive matching confidence levels $1 - \tau$ for the non-inferiority margins of the lower predictive bound and synthesis approach. Findings from these comparisons are discussed later in Section 5.2.

5 Discussion

5.1 Summary

The reliability of non-inferiority trials is adversely affected not only by the trial conduct, but also by external factors affecting the trial design, including inherent bias and variability in historical data, publication and selection bias, lack of constancy in reference effects, choice of the non-inferiority margin, and method of non-inferiority tests. Many crucial factors are not present in typical superiority trials. Therefore, standards and methods that may be acceptable to superiority trials are questionable for non-inferiority trials.

Past research has mainly focused on issues of sampling variability in the historical data, choice of non-inferiority margin and method of non-inferiority tests; and the study of across-trial unconditional type 1 error rates does not take into account publication and selection bias. In addition, the issue of inherent bias in historical data and analysis has been ignored.

The objective of this paper is to develop a theoretically sound and practically implementable framework to incorporate these various external factors. Within this framework, it is now possible to address the fundamental issues of variability in historical data and future non-inferiority trials, various sources of bias, and non-constancy in reference effects that affect the reliability of non-inferiority trials. We recognize that the decision making process leading to the final non-inferiority trial design can be subject to various sources of selection bias, and furthermore, no pre-specified algorithms exist prior to the design of a non-inferiority trial that allow quantifications of the degree of selection bias. Therefore, a direct calculation of across-trial unconditional type 1 error rates that incorporate the decision making process for adjusting selection bias is not possible.

Following the conditionality principle (Cox and Hinkley, p. 38, 1974), we propose to control the across-trial conditional type 1 error rates over a robust range for the reference effects given the decision made on the final non-inferiority trial design while adjusting for the minimum reference effect bias. With an appropriate choice of the non-inferiority margin and an adequate sample size, the robust range would account for inherent and publication bias in the historical data and analysis, selection bias in the final non-inferiority trial design, and non-constancy in reference effects. The robust range can be evaluated for the non-inferiority test with any non-inferiority margin approach.

We also touch on the issue of analysis population. As the maximum standard error of the observed treatment difference is used, the across-trial conditional type 1 error rates of the ITT analysis are controlled when the reference effect is within the robust range. Therefore, the ITT population can now be used as the primary analysis, while the per-protocol population should be used for supportive analysis.

The notion of accounting for the variability of historical data has been very confusing. In the future non-inferiority trial, the historical data that gave rise to the margin or design do not change, and therefore, there is no (conditional) variability. We propose to recognize that given the historical data the uncertainty is entirely in the underlying distribution model that gave rise to the historical data. To account for this uncertainty, we introduce the consistency

principle, by which all distribution or simulation models that are consistent with the historical data are used to evaluate the minimum robust range. We illustrate this with the design example.

5.2 Findings and Recommendations

5.2.1 Synthesis Approach—From the point of view of adaptive designs, the so-called synthesis test of Hasselblad and Kong (2001) relies on a form of combination test statistic that incorporates historical data in the analysis of non-inferiority trial. However, on a fundamental ground, the synthesis test ignores the process of adaption or decision making in the design of the non-inferiority trial along with its various sources of selection bias. The problem is first recognized by Lawrence (2005) who demonstrates that the within-trial type 1 error rates of the synthesis test can reach 50%. Subsequently, Rothmann et al. (pp. 62–65, 2012) point out the issues of publication and statistical significance bias, which has not been considered in evaluation of across-trial unconditional type 1 error rates. When such bias is taken into account, as shown through numerical evaluation of the minimum robust range of actual design examples, we find that the common perception that the synthesis test controls the across-trial unconditional type 1 error rates under the constancy assumption is incorrect. This is easily understood as the key difference or relationship between the across-trial conditional and unconditional type 1 error rates lies entirely on the choice of the minimum reference effect $\delta_{\min}$ used for adjusting the minimum reference effect bias (see Section 3.4). The across-trial conditional type 1 error rate converges to the across-trial unconditional type 1 error rate as $\delta_{\min}$ goes to $-\infty$. Only under $\delta_{\min} = -\infty$, or absence of minimum reference effect bias, that the synthesis test controls the across-trial unconditional type 1 error rate when the constancy assumption is true. As seen from Table 2, after adjusting for the minimum reference effect bias with $\delta_{\min} = .1$, the minimum robust range is $\lambda_{\min} = -.052$. This means that the reference effect over the putative control has to increase at least 5.2% for the synthesis approach to control the across-trial conditional type 1 error rates.

Another fundamental difficulty with the synthesis test is that its corresponding non-inferiority margin increases as the sample size decreases. This can lead to more negative robust range. However, the plausible range of extent for addressing pre-trial issues of non-constancy in reference effects and various sources of bias is unaffected by the size of the future non-inferiority trial. Thus, an increased margin reduces the reliability of the non-inferiority trial with the synthesis approach. Due to space limitation, the numerical results supporting this and other findings for the confidence and predictive bound approaches are not presented but are available from the authors.

In addition to the issue of the negative robust range, a small sample size also increases the chance for within-trial bias (e.g., imbalance of known or unknown risk factors). Given also that there are always concerns of inherent bias in historical data and reduction in reference effect over placebo over time, it is difficult to envision a practical setting in which the synthesis test can be reliably applied. These issues still persist even when a retention null hypothesis is used for the synthesis approach. Thus, it is necessary to investigate these properties of the synthesis-like tests with bias-correction proposed by Odem-Davis and Fleming (2013).

5.2.2 Confidence Bound Approach—Compared to the synthesis approach, the confidence bound approach for non-inferiority margins offer wider robust range. In addition, the confidence bound approach has the flexibility to further increase the robust range with the choice of higher confidence levels. However, once the confidence level is about 99%, the approach becomes ineffective to tighten the margins further. This can be an issue when a meta-analysis, especially when inappropriately applied, leads to large lower confidence bounds. Another drawback of the confidence bound approach is that the non-inferiority margin is sensitive to the variability of the historical data (Odem-Davis and Fleming, 2013).

An interesting and surprising finding for the confidence bound approach is that the robust range increases as the sample size decreases. A closer examination of the conditional error rate in equation (4) reveals that $(\hat{\mu} - \delta_{NI})/\sigma_{TR}$ is proportional to $n^{1/2}$. Thus, for a given δ_{NI} which does not depend on n , the conditional type 1 error rate for $\hat{\mu} < \delta_{NI}$ decreases as n decreases. This results in increasing robust range for decreasing sample size, although the magnitude of the max increase is negligible to raise concerns for practical applications. However, this does suggest if the non-inferiority margin δ_{NI} can be constructed to depend on n , then it is possible that the underlying approach could either decrease the robust range, as the synthesis approach, or meaningfully increase the robust range as the sample size decreases.

5.2.3 Predictive Bound Approach—The predictive bound approach of Liu and Li (2014) incorporates variability both in historical data and the future non-inferiority trial. Therefore, the non-inferiority margin is less sensitive to the variability of the historical data. There is also a wider choice of tightness of the non-inferiority margins by varying the predictive confidence level. Another attractive feature is that for a given predictive confidence level, the non-inferiority margin decreases as the trial sponsor reduces the sample size with a lower power. Per early discussion, it is unsurprising that this in turn leads to increases in the robust range. Thus, the reliability of non-inferiority trials with smaller sample size can still be maintained within a certain robust range. This property would also be the foundation for developing non-inferiority group sequential designs with reliable interim results. Finally, the predictive bound approach, by construction, permits putative confidence intervals for the effect of new drug over placebo. This feature is not shared by either the synthesis nor the confidence bound approach.

5.2.4 Bias in Historical Data—Last but not the least, we illustrate that inherent bias in historical data can greatly affect the design of future non-inferiority or biosimilarity trials. Although bias in the local evaluations of overall tumor response in the follicular lymphoma trial has negatively affected the size or potentially the feasibility of subsequent non-inferiority or biosimilarity trials, there are many other ways, such as informative differential dropouts, that inherent bias in historical data can inflate the type 1 error rates of the future non-inferiority or biosimilarity trials. For the issue of local versus central review, we are particularly concerned with the PhRMA PFS Working Group's limited research that has failed to evaluate the overall impact on clinical science, health-care costs, and public access to alternative effective medicines.

5.3 Future Work

Our work is limited to the mean difference scale for binomial distributions. Additional work is necessary to fulfill the need for other scales or endpoints with different distributions (see the FDA Guidance). For time-to-event endpoints, we are generally concerned with the use of the hazard ratio or its natural log transformation as its validity depends on the assumption of the proportional hazards model being correct.

A full investigation of robustness of various non-inferiority margin approaches, including the synthesis type approaches in Odem-Davis and Fleming (2013), is necessary to solidify or perhaps expand the findings in this paper. There is also a need to examine various approaches under retention hypotheses, for which the index method proposed by Li and Chi (2011) may be used to develop the retention rates for different disease and therapeutic areas.

The paper relies on the simple simulation model that the analysis of the historical data is based on the pre-specified SAP of the historical trial protocol. Progress is necessary for settings where meta-analyses are used to define non-inferiority margins. Such meta-analyses need to incorporate not only random effects but also non-random systematic differences between trials and changing reference effects over time. Simulation models reflecting the meta-analyses need to account for deficiencies of the historical data for lacking information on non-random systematic trial to trial differences and reference effects over time. With multiple historical trials, the meta-analysis itself becomes a target of investigation as it is often possible to select the meta-analysis to maximize the difference and then provide a “justification” for the analysis.

The most challenging task is to determine the robust range that reflects the plausible range of bias and non-constancy (Hung et al., 2007). There is a need for more research to identify and then quantify various sources of inherent bias in historical data, selection bias in the trial design, and non-constancy in reference effects. Such research requires evidence modeling with sound clinical judgment.

In conclusion, it is clear that there are mountains of difficulties to overcome in order that results from non-inferiority trials be reliably interpreted and risks for incorrect clinical and statistical inference be minimized. There may be situations where these troubles can be avoided if the non-inferiority trial is designed using a minimally clinically meaningful difference or a superiority trial is employed. However, for drugs (e.g., biosimilars) truly expected to be similar to the reference, neither option is feasible. Then it is necessary to overcome the difficulties of the non-inferiority trials.

Acknowledgments

The authors would like to thank Dr. Paul Gallo, a FDA anonymous reviewer, and a referee for their careful review of the manuscript and suggestions for improving the content and presentation. The authors are grateful to Dr. Fleming for his guidance and insights which have substantially increased the scope and depth of the paper. Contributions by Odem-Davis were supported by the National Center for Advancing Translational Sciences of the National Institutes of Health (NIH) under Award Number UL1TR000423.

References

1. Amit O, Bushnell W, Dodd L, Roach N, Sargent D. Blinded independent central review of the progression-free survival endpoint. *The Oncologist*. 2010; 15:492495.
2. Biologics Price Competition and Innovation Act. 2009. www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/UCM216146.pdf
3. Committee for Medicinal Products for Human Use (CHMP). Guideline on the choice of the non-inferiority margin. 2006. www.emea.europa.eu/pdfs/human/ewp/215899en.pdf
4. Cox, DR.; Hinkley, DV. *Theoretical Statistics*. Chapman & Hall; London: 1974.
5. Davis, KS. Dissertation. University of Washington, Department of Biostatistics; 2010. Non-constancy, estimation bias, biocreep, and an alternative to current methods used in non-inferiority trials.
6. Fleming TR, Emerson SS. Evaluating rivaroxaban for nonvalvular atrial fibrillation regulatory considerations. *The New England Journal of Medicine*. 2011; 356:1557–1559. [PubMed: 21992495]
7. Fleming TR, Odem-Davis K, Rothmann MD, Shen YL. Some essential considerations in the design and conduct of non-inferiority trials. *Clinical Trials*. 2011; 8:432–439. [PubMed: 21835862]
8. Fleming, Z. *Regulatory Science Course*. FDA/CDER; Rockville: 1997. Efficacy Review.
9. Food and Drug Administration. Guidance for industry: codevelopment of two or more nonmarketed investigational drugs for use in combination. 2010. www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM236669.pdf
10. Food and Drug Administration. Guidance for industry: clinical trial endpoints for the approval of cancer drugs and biologics. 2007. www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM071590.pdf
11. Food and Drug Administration. Guidance for industry: non-inferiority clinical trials. 2010. www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM202140.pdf
12. Food and Drug Administration. Statement regarding the demonstrations of effectiveness of human drug products or devices. *Federal Register*. 2005; 60:39180–39181.
13. Hasselblad V, Kong DF. Statistical methods for comparison to placebo in active-control trials. *Drug Information Journal*. 2001; 35:435–449.
14. Hung HMJ, Wang SJ, Tsong Y, Lawrence J, O'Neil RT. Some fundamental issues with non-inferiority testing in active controlled trials. *Statistics in Medicine*. 2003; 22:213–225. [PubMed: 12520558]
15. Hung HMJ, Wang SJ, O'Neil RT. Issues with statistical risks for testing methods in noninferiority trial without a placebo arm. *Journal of Biopharmaceutical Statistics*. 2007; 17:201–213. [PubMed: 17365218]
16. Kang SH, Yi 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; 29:1477–1487. [PubMed: 20535762]
17. Lawrence J. Some remarks about the analysis of active control studies. *Biometrical Journal*. 2005; 47:616622.
18. Li G, Chi GYH. Inferiority index and margin in non-inferiority trials. *Statistics in Biopharmaceutical Research*. 2011; 3:288–301.
19. Li Y, Liu Q, Wood P, Johri A. Statistical considerations for biosimilar clinical efficacy trials with asymmetrical margins. To appear in *Statistics in Medicine*. 2012
20. Liu Q, Chi GYH. Understanding the FDA guidance on adaptive designs: historical, legal and statistical perspectives. Special issue of *Journal of Biopharmaceutical Statistics*. 2010; 20:1178–1219.
21. Liu Q, Li Y. On the lower predictive bound approach for non-inferiority clinical trials with binary endpoint. To appear in *Journal of Biopharmaceutical Statistics special issue on biosimilarity*. 2014
22. Liu Q, Proschan MA, Pledger GW. A unified theory of two-stage adaptive designs. *Journal of the American Statistical Association*. 2002; 97:1034–1041.

23. Odem-Davis K, Fleming TR. Adjusting for unknown bias in non-inferiority clinical trials. Submitted to Journal of Pharmaceutical Statistics. 2013
24. Rothmann M. Type 1 error probabilities based on design-stage strategies with applications to non-inferiority trials. Journal of Biopharmaceutical Statistics. 2005; 15:109–127. [PubMed: 15702608]
25. Rothmann M, Li N, Chen G, Chi GYH, Temple RT, Tsou HH. Non-inferiority methods for mortality trials in oncology. Statistics in Medicine. 2003; 22:239–264. [PubMed: 12520560]
26. Rothmann, M.; Wiens, B.; Chan, I. Design and Analysis of Non-Inferiority Trials. Chapman Hall/CRC Press; Boca Raton: 2011.
27. Sackett DL. Bias in analytic research (with discussion) (Com: p 64–68). Journal of Chronic Diseases. 1979; 32:51–63. [PubMed: 447779]

Figure 1.
Solution $\delta_{\alpha, M} = .101$ for simulation model with $\Delta_M = .25$ and $\alpha = .025$

Table 1

Summary of Overall Tumor Response Data

	Local			Central		
	<i>n</i>	Rate	95% CI	<i>n</i>	Rate	95% CI
Control	159	.57	(.493, .646)	150	.43	(.351, .509)
Reference	162	.81	(.740, .870)	150	.77	(.703, .837)
Difference		.24	(.142, .338)		.34	(.236, .444)

Table 2

Designs with central and local OTR reviews

	Central				Local			
	n	δ_{NI}	λ_{min}	$1 - \tau$	n	δ_{NI}	λ_{min}	$1 - \tau$
LPB _{.995}	199	.144	.149	.9999	318	.106	.048	.9964
LCB _{.995}	99	.203	.096	.9950	287	.111	.039	.9950
LCB _{.975}	74	.236	.067	.9750	176	.142	.013	.9750
Syn	41	.316	-.008	.6723	80	.212	-.052	.7146