2-10-2017

Adaptive Non-Inferiority Margins under Observable Non-Constancy

Brett S. Hanscom
Fred Hutchinson Cancer Research Center, bhanscom@fredhutch.org

Deborah J. Donnell
Fred Hutchinson Cancer Research Center, deborah@scharp.org

Brian D. Williamson
University of Washington - Seattle Campus, brianw26@uw.edu

Jim Hughes
University of Washington - Seattle Campus, jphughes@u.washington.edu

Adaptive Non-Inferiority Margins under Observable Non-Constancy

Brett Hanscom[†], Deborah Donnell[†], Brian Williamson[⋆], and James P Hughes[⋆†]

[†]Fred Hutch Cancer Research Center, Seattle, WA 98109
[⋆]Department of Biostatistics, University of Washington, Seattle, WA, 98195

February 10, 2017

Abstract

A central assumption in the design and conduct of non-inferiority trials is that the active-control therapy will have the same degree of effectiveness in the planned non-inferiority trial as it had in the prior placebo-controlled trials used to define the non-inferiority margin. This is referred to as the 'constancy' assumption. If the constancy assumption fails, the chosen non-inferiority margin is not valid and the study runs the risk of approving an inferior product or failing to approve a beneficial product. The constancy assumption cannot be validated in a trial without a placebo arm, and it is unlikely ever to be met completely. However, it is often the case that there exist strong, measurable predictors of constancy, such as dosing and adherence, and such predictors can be used to identify situations where the constancy assumption will likely fail. Here we propose a method for using measurable predictors of active-control effectiveness to specify non-inferiority margins targeted to the planned study population, and further use these predictors to adapt the non-inferiority margin at the end of the study. Population-specific margins can help avoid violations of the constancy assumption, and adaptive margins can help adjust for violations that will inevitably occur in real clinical trials, while at the same time maintain pre-specified levels of type I error and power.

1   Introduction

Non-inferiority (NI) trials are designed to determine whether a new therapy is as effective as, or at least not meaningfully worse than, an existing therapy that was previously approved based on placebo-controlled trials. By comparing the new therapy to the approved "active-control" therapy, an NI trial can be used to infer whether the new therapy is effective without having to incorporate a placebo-control arm. For this inference to be valid it must be assumed that the active-control therapy has a similar degree of effectiveness in the planned NI trial as it had in prior placebo-controlled trials. This is referred to as the "constancy" assumption. The constancy assumption cannot be validated in a trial without a placebo arm, and it is unlikely to ever be met completely; however, various strategies can

1

be used to minimize the risk of non-constancy.

A first approach is simply to specify entry criteria designed to enroll participants who match the participants from the prior trials, and specify in the protocol a dosing amount and schedule identical to the schedule and dosing used in prior trials. Unfortunately, identical entry criteria can yield very different participant populations, particularly if the new trial is conducted in a difference setting, city or country. Also, although it may be simple to prescribe a specific dosing schedule, actual dosing may vary dramatically depending on adherence.

Prior articles have suggested modifications to the non-inferiority margin when there exists evidence of constancy failure. Everson-Stewart [1] proposed a model for assessing constancy based on heterogeneous effectiveness in population subgroups, and recommended tightening the margin, or moving to a superiority design, if constancy is violated. This approach has an important limitation in that it only allows for tightening the margin, never relaxing the margin, which fails to account for situations where active-control effectiveness is higher than planned. Nie and Soon [2] developed a regression-model approach for identifying potential non-constancy, and use population characteristics to define a non-inferiority margin appropriate for the enrolled study participants. They did not, however, demonstrate whether modifying the margin could improve, or harm, the trial's operating characteristics.

Although our modeling approach is somewhat different from the one proposed by Nie and Soon, our objective is essentially to extend Nie and Soon's idea of an adjusted non-inferiority margin to the planning, monitoring, and analysis phases of a non-inferiority trial. In addition to modeling population characteristics, the proposed method incorporates dynamic features of the trial, such as dosing and adherence, which can only be measured during the trial or after the trial is complete. We propose several methods for adapting the non-inferiority margin, and evaluate how each method can influence operating characteristics such as Type-I error, power, and expected sample size. An example trial of HIV pre-exposure prophylaxis (PrEP) is used for illustration.

2 Sources of Non-Constancy and Defining the NI Margin

Non-constancy can occur for a variety of reasons, including differences in participant characteristics, differences in dosing or adherence, changes in background supportive care, or actual declines in biological benefit, for example due to increased resistance to an antibiotic. Many potential sources of non-constancy may be unobservable, but the associations between certain observable participant characteristics and the effectiveness of a therapy are sometimes known. These factors can be divided into two categories: (1) fixed population attributes such as race, gender, and age that do not change appreciably during a trial and can be measured at baseline, and (2) dynamic features, such as dosing and drug adherence, that could change during the course of a trial and cannot be assessed until the trial is complete.

An important aspect of dynamic features is that not only can they differ over time, they

2

can also differ between treatment arms. For example, in an HIV-prevention trial of long-acting injectable pre-exposure prophylaxis (PrEP) compared to daily oral pill-based PrEP, participants in the injectable arm are likely to have high adherence, assuming they receive directly observed injections, while participants in the oral arm may have mixed adherence, depending on how well they take the pills. This is true even in a double-blind, double-dummy trial (participants receiving active injections and placebo pills, versus participants receiving active pills and placebo injections) because adherence to the oral pills will only influence drug concentrations in participants receiving the active, oral pills.

To protect against falsely declaring non-inferiority (inflated type I error) if the active control is not as effective as in prior trials, NI margins are often chosen to be conservative (i.e. less likely to produce a significant result). Defining the non-inferiority margin typically involves estimating the effect of an active-contol therapy based on prior placebo-controlled trials, and choosing a non-inferiority margin that conserves some proportion ρ of the active-control effect [3]. For example, consider a series of two arm trials comparing the active-control therapy C to placebo P, and let $\hat{R}R_{C/P}$ be the estimated relative risk based on a meta analysis. The non-inferiority margin δ^* can be defined, conservatively, to preserve at least $(100 \times \rho)\%$ of the active-control benefit by setting

$$\delta^* = (LCL_{95}(\widehat{RR}_{P/C}))^{1-\rho} \tag{1}$$

where LCL_{95} is the lower limit of the 95% confidence interval and $\hat{R}R_{P/C} = \hat{R}R_{C/P}^{-1}$. The lower confidence limit for the active-control effect can be thought of as the 'assured effect', i.e. evidence from prior trials rules out a smaller effect, but does not assure anything larger. Using the LCL_{95} as the estimated active-control effect acknowledges the uncertainty associated with the meta-analytic estimate and therefore provides a degree of protection against non-constancy. Values for ρ can range between 0 and 1, where choosing $\rho = 0$ gives a margin that does not preserve any benefit of C but assures that the experimental therapy E is at least better than placebo, and choosing $\rho = 1$ gives a margin of 1.0 which is equivalent to requiring superiority of E over C. The value for ρ is most often taken to be 0.50, yielding a margin that preserves at least 50% of the benefit provided by C (on the log relative risk scale). If the results of a non-inferiority trial comparing an E to the active control satisfy

$$UCL_{95}(\widehat{RR}_{E/C}) < \delta^* \tag{2}$$

where UCL_{95} is the upper limit on the 95% confidence interval, then the trial results are consistent with non-inferiority.

The meta-analysis-based margin δ^* in (1) effectively assumes that the active-control effect in the planned trial will correspond to the average effect observed in the prior placebo-controlled trials, and ignores attributes and dynamic features of the planned study population that predict effectiveness (i.e. effect modifiers). *Post hoc* subgroup analyses are regularly performed on clinical-trial data to identify correlates of effectivess. Although such analyses can be informative in setting a better margin, subgroup analyses are generally underpowered and exploratory in nature. In addition, analyses of effectiveness based on post-randomization factors such as drug adherence do not benefit from the protection of randomization and therefore may not yield valid results. In the next section we suggest a

3

strategy than could more reliably identify effect modifiers.

3 Trial-level variables, meta-regression, and population-specific NI margins

Rather than relying on subgroup analyses from individual trials, effect modifiers may be identified by using meta-analysis regression to combine results across multiple studies. The meta-analysis approach can improve power and precision for estimating effect modification in important demographic and clinical subgroups [4]. Improved power results from the increased sample sizes achieved by combining data from multiple trials. Second, by comparing post-randomization effects across studies, rather than within studies, the meta-analysis approach can reduce the potential for confounding. While it is well recognized that cross-trial comparisons may be influenced by ecological bias [5], the potential for confounding by post-randomization variables is a more serious concern.

For example, within-study estimation of the association between drug adherence and effectiveness could easily be confounded by medication side effects. That is, if adherence is influenced by the treatment, then the estimation of treatment effects among adherers and non-adherers would no longer rely on randomized comparisons. On the other hand, if multiple randomized, blinded, controlled trials with different levels of adherence show an association between adherence and effectiveness, we know that estimation of within-study treatment effects is protected by randomization and unconfounded by medication side effects. By treating dynamic features such as drug adherence as trial-level variables in a meta-regression model, dynamic effect modifiers can be identified and estimated without the usual hazards associated with analysis of post-randomization subgroups.

To estimate the size and importance of potential effect modifiers, mixed-effects meta analysis can be extended to a regression model that includes study-level, fixed-effects covariates [6] as follows:

$$\log(RR_j) = \beta_0 + \boldsymbol{\beta_1}\boldsymbol{x}_j + \boldsymbol{\beta_2}\boldsymbol{z}_j + b_j + \varepsilon_j \tag{3}$$

where $\boldsymbol{x}_j$ are population attributes, $\boldsymbol{z}_j$ are dynamic features, b_j are the study-specific random effects with $b_j \sim N(0, \tau^2)$ and τ^2 the between study variance, and ε_j are error terms corresponding to study-specific effect-size estimates with $\varepsilon_j \sim N(0, \sigma_j^2)$. Fixed-effect parameter estimates $\hat{\beta}_0$, $\hat{\boldsymbol{\beta_1}}$, and $\hat{\boldsymbol{\beta_2}}$ can then be used to estimate effectiveness in a new study with anticipated attributes $\boldsymbol{x}^A$ and features $\boldsymbol{z}^A$, as

$$\widehat{RR}_{P/C}(\boldsymbol{x}^A, \boldsymbol{z}^A) = \exp(\hat{\beta}_0 + \hat{\boldsymbol{\beta_1}}\boldsymbol{x}^A + \hat{\boldsymbol{\beta_2}}\boldsymbol{z}^A). \tag{4}$$

A population-specific NI margin can then be computed based on the lower confidence limit of the 95% confidenc limit of the regression model estimate, preserving at least the fraction ρ of the active control benefit, by setting

$$\delta(\boldsymbol{x}^A, \boldsymbol{z}^A, \rho) = (LCL_{95}(\widehat{RR}_{P/C}(\boldsymbol{x}^A, \boldsymbol{z}^A)))^{1-\rho}. \tag{5}$$

For illustration we consider a meta-analysis regression using results from randomized, placebo-controlled trials of oral PrEP. Adherence to oral PrEP is considered to be a strong

dynamic effect modifier, and there is some evidence that oral PrEP may be less effective in women, so both a dynamic and baseline factor are included in the model. Results from seven major randomized trials are included in the analysis [7; 8; 9; 10; 11; 12; 13], giving the estimated regression equation

$$\widehat{RR}_{P/C}(sex, adherence) = \exp(0.7520 - 0.1058 \times sex - 2.276 \times adherence) \qquad (6)$$

where *sex* is an indicator variable of sex at birth with $1 = male$, and *adherence* is a measure between zero and one of the proportion of participants with detectable plasma levels of PrEP. Figure 1 shows a scatterplot of the trial results as a function of adherence, divided by sex. The fitted regression line for men, along with confidence bounds, are shown. For a planned study in men, the lower confidence limits in Figure 1 would be used as the basis for computing the NI margin. The lower confidence limit is used as a conservative estimate for effectiveness in the regression setting because even though observable factors may be accounted for, unobservable or unrecognized sources of non-constancy cannot be ruled out.

For a hypothetical trial enrolling men, a range of possible margins derived from this regression equation are shown in Table 1. An essential result is that when projected adherence falls below approximately 40%, the lower confidence limit on projected effectiveness falls below 1.0, indicating that there is not strong evidence for effectiveness, and that a superiority trial is recommended. (A non-inferiority trial with a margin of 1.0 is equivalent to a superiority trial.) This suggests that in general the meta-regression model may be used not only to set the NI margin, but also to determine under what circumstances an NI trial is appropriate.

Adherence	Effectiveness	NI Margin
0.40	<1.0	1.0
0.50	1.17	1.08
0.60	1.50	1.23
0.70	1.89	1.37
0.80	2.30	1.52

Table 1: Predicted oral-PrEP effectiveness in men as a function of adherence, and suggested NI margins that preserve at least 50% of the benefit ($\rho = 0.5$).

4 Non-constancy, Type I Error and Power.

The study population in a clinical trial often is not identical to the populations in the prior placebo-controlled trials used for identifying the NI margin. For an intervention with substantial effect modifiers, even if a population-specific approach is used to select the margin, population attributes and dynamic features in the new trial may not match the values used in the planning phase. If these values are substantially different from the planning phase, there is a risk that the constancy assumption will fail and the selected NI margin will not be valid, in which case the trial runs the risk of declaring support for a product that doesn't work (type I error), or failing to support a product that does work (type II error). In

the context of a non-inferiority trial, a product that 'works' is one that provides a similar amount of benefit to the active control, where the active-control benefit is defined in relation to placebo, and a product that does not 'work' is one that provides signifantly less benefit than the active control, or provides no benefit at all (i.e. is equivalent to placebo, or causes harm).

To illustrate, consider a trial that is designed under the assumed values $\boldsymbol{x} = \boldsymbol{x}^A$ and $\boldsymbol{z} = \boldsymbol{z}^A$, and using corresponding NI margin $\delta(\boldsymbol{x}^A, \boldsymbol{z}^A, \rho)$ as in (5). Unlike a superiority trial where the goal is to rule out $H_0 : RR = 1.0$, in a non-inferiority trial the goal is to rule out the NI margin, leading to the hypotheses:

$$
\begin{aligned}
H_0 &: RR_{E/C} = \delta(\boldsymbol{x}^A, \boldsymbol{z}^A, \rho) \\
H_1 &: RR_{E/C} < \delta(\boldsymbol{x}^A, \boldsymbol{z}^A, \rho),
\end{aligned}
\tag{7}
$$

where $\delta()$ will typically be larger than 1.0. Re-expressing these hypotheses in terms of incidence rates helps isolate the impact of non-constancy. The relative effectiveness $RR_{E/C}$ can be written as λ_E/λ_C^A, where λ_E is the incidence rate in E, and λ_C^A is the incidence rate in C when $\boldsymbol{x} = \boldsymbol{x}^A$ and $\boldsymbol{z} = \boldsymbol{z}^A$. The null hypothesis in (7) can then be written as

$$
H_0 : \lambda_E = \lambda_E^0 = \lambda_C^A \delta(\boldsymbol{x}^A, \boldsymbol{z}^A, \rho),
\tag{8}
$$

where λ_E^0 is defined to be the highest allowable incidence under E that would be considered clinically acceptable.

The specific alternative hypothesis used to compute sample size and power is often based on the desire to reject H_0 if E has an effect similar to C, that is, if $H_1 : RR_{E/C} = 1.0$. If there is reason to expect that E may outperform C then the study may be powered to establish a somewhat stronger effect, say $H_1 : RR_{E/C} = \xi$, where $\xi \in [0, 1]$. The alternative hypothesis is then expressed in terms of incidence rates as

$$
H_1 : \lambda_E = \lambda_E^1 = \lambda_C^A \xi.
\tag{9}
$$

Note that since an NI trial is designed to rule out the margin, not equality, the trial is powered to detect an effect size (ratio of the alternative to the null) equal to $\xi/\delta(\boldsymbol{x}^A, \boldsymbol{z}^A, \rho)$.

If the true study population characteristics turn out be $\boldsymbol{x}^T \neq \boldsymbol{x}^A$ and $\boldsymbol{z}^T \neq \boldsymbol{z}^A$ such that $RR_{C/P}(\boldsymbol{x}^T, \boldsymbol{z}^T) \neq RR_{C/P}(\boldsymbol{x}^A, \boldsymbol{z}^A)$ then the planned margin $\delta(\boldsymbol{x}^A, \boldsymbol{z}^A, \rho)$ will be incorrect and the study will not retain desired levels of type I error and power. By expressing $RR_{E/C}$ in terms of incidence rates, and using the fact that $\lambda_C^T = \lambda_P RR_{C/P}(\boldsymbol{x}^T, \boldsymbol{z}^T)$, where λ_P is incidence in a hypothetical placebo arm, type I error and power can be expressed as a function of $RR_{C/P}(\boldsymbol{x}^T, \boldsymbol{z}^T)$:

$$
\begin{aligned}
P[\text{Type I error}] &= P\Big[UCL_{95}\Big(\frac{\lambda_E}{\lambda_P RR_{C/P}(\boldsymbol{x}^T, \boldsymbol{z}^T)}\Big) < \delta(\boldsymbol{x}^A, \boldsymbol{z}^A, \rho) | \lambda_E = \lambda_E^0\Big] \\
\text{Power} &= P\Big[UCL_{95}\Big(\frac{\lambda_E}{\lambda_P RR_{C/P}(\boldsymbol{x}^T, \boldsymbol{z}^T)}\Big) < \delta(\boldsymbol{x}^A, \boldsymbol{z}^A, \rho) | \lambda_E = \lambda_E^1\Big].
\end{aligned}
\tag{10}
$$

6

Error rates for an example trial are plotted in Figure (2) as a function of the (hypothetical) percent risk reduction provided by C (versus placebo) in the true study population, or $(1 - RR_{C/P}(\boldsymbol{x}^T, \boldsymbol{z}^T)) \times 100$. If C is more effective than planned, E will not look as good in comparison, and the type II error rate will be high (i.e. the trial will have low power.) If C is less effective than planned, E will appear better in comparison to C, and the chances of a false-positive result will increase. It is clear that fairly small deviations from constancy can have a substantial impact on the probability of making a false conclusion. For instance if C is 60% effective instead of the planned 50%, the type II error probability increases from 10% to 48% (i.e. power drops from 90% to 52%). Likewise if C is only 40% effective, the type I error probability elevates from 2.5% to 16%. To reduce these error probabilities we propose an adaptive NI margin approach that makes use of observable study population characteristics.

5   Adaptive NI Margins

Although an NI margin must be specified in order to plan an NI trial, the margin used for planning may not always be the appropriate gauge to judge whether the new product is truly effective. As demonstrated in the previous section, using the wrong NI margin at the analysis stage can be detrimental to the operating characteristics of a trial, and the trial may fail to ensure that the experimental therapy provides sufficient benefit over placebo to warrant approval. In order to make certain that an NI trial will only support a therapy that meets a pre-specified level of effectiveness, we propose a method for adapting the planned margin that is designed to accomodate observed non-constancy.

In the presence of signifcant effect modifiers, we suggest an adaptive margin, where we pre-specify an equation for updating the non-inferiority margin at the end of the study based on meta-analysis regression, such as in (4). The method for updating incorportates pre-specified, observable characteristics of the trial population, including both baseline characteristics and dynamic features that can be measured during or after the trial. We use the terms $\boldsymbol{x}^T$ and $\boldsymbol{z}^T$ to refer to the observed, or 'true', values of the baseline characteristics and dynamic features, respectively, of the study population. A pre-specified level of benefit over placebo, which we refer to as 'absolute effectiveneses', (or an equation for absolute effectiveness,) is also required and is discussed below.

There are a number of possible ways to define an adaptive NI margin, and the rationale for each choice is slightly different. Determining the most appropriate strategy for adapting the margin will depend on the goals of the trial, the factors that influence non-constancy, and the investigators' perspective. As we describe the various methods for adjusting the margin below, we also discuss the rationale for each choice.

To facilitate the calculation of adaptive margins, it is useful to define the NI margin in relation to the absolute effectiveness the experimental therapy is required to provide with respect to a (hypothetical, unobserved) placebo. This is precisely the way NI margins are typically calculated, but here we explicitly define the margin in terms of the required experimental-therapy-versus-placebo relative risk, $RR_{E/P}$, which we call Δ. Specifically, the adaptive margin, δ^a, is expressed as a function of both Δ and the estimated active-control effect:

7

$$\delta^a = \Delta \times LCL_{95}(\widehat{RR}_{P/C}(\boldsymbol{x}^T, \boldsymbol{z}^T)), \tag{11}$$

Recall that the active-control effect takes the somewhat counterintuitive form of the placebo-versus-control relative risk (i.e. a measure of how much worse the placebo performs as compared to the active control therapy.) This results in a margin which defines how much worse the experimental therapy can be as compared to the control therapy. And again, the lower confidence limit of the estimated active-control effect (the 'assured effect') is used in order to account for unmeasured factors that may influence constancy.

For example, if we wish to ensure that the experimental therapy provides at least 25% risk reduction as compared to placebo, we would specify $\Delta = 0.75$. Suppose also that, based on observed study population characteristics, the estimated effect of the active control (placebo-versus-control relative risk) is 2.0 with lower confidence limit 1.8. Our adaptive margin would then be defined as $\delta = 0.75 \times 1.8 = 1.35$. Provided that the upper confidence limit of the experimental-versus-control relative risk falls below 1.35, we are assured that the experimental therapy is at least 25% more effective than placebo.

The general approach for computing δ^a is to compute $LCL_{95}(\widehat{RR}_{P/C}(\boldsymbol{x}^T, \boldsymbol{z}^T))$ based on measured values for $\boldsymbol{x}$ and $\boldsymbol{z}$, and then multply by Δ. The value of Δ may be fixed in advance, or may depend on $\boldsymbol{x}^T$ and $\boldsymbol{z}^T$, as described below. It is clear from (11) that even if Δ is fixed, δ^a is flexible and depends on the observed study population.

There are two general strategies for specifying Δ, one that requires a pre-specified level of absolute effectiveness, and the other that provides flexibility depending on how well the active control therapy works in the study population.

For method one, a fixed, pre-specified level of benefit Δ is chosen, which can be defined as either (a) the amount of benefit over placebo that was used to define the meta-analysis-based margin, or (b) an investigator-defined minimal clinically important difference (MCID). For example, if the pre-planned margin is the meta-regression-based margin $\delta(\boldsymbol{x}^A, \boldsymbol{z}^A, \rho)$ described in (5), we would define the planned value for Δ as

$$\begin{aligned}
\Delta_{Plan} &= \delta(\boldsymbol{x}^A, \boldsymbol{z}^A, \rho) \times LCL_{95}(\widehat{RR}_{P/C}(\boldsymbol{x}^A, \boldsymbol{z}^A))^{-1} \\
&= (LCL_{95}(\widehat{RR}_{P/C}(\boldsymbol{x}^A, \boldsymbol{z}^A)))^{1-\rho} \times LCL_{95}(\widehat{RR}_{P/C}(\boldsymbol{x}^A, \boldsymbol{z}^A))^{-1} \\
&= (LCL_{95}(\widehat{RR}_{P/C}(\boldsymbol{x}^A, \boldsymbol{z}^A)))^{-\rho},
\end{aligned} \tag{12}$$

which uses the planned values $\boldsymbol{x}^A$ and $\boldsymbol{z}^A$ in relation (11). If $\boldsymbol{x}^T = \boldsymbol{x}^A$ and $\boldsymbol{z}^T = \boldsymbol{z}^A$ then $\delta^a = \delta(\boldsymbol{x}^A, \boldsymbol{z}^A, \rho)$.

The rationale for this choice may be that investigators believe the original parameters used to plan the trial provide a margin that would, in a positive trial, establish a reasonable level of clinical effectiveness, and that deviations in participant characterisitics or behavior should have no bearing on the minimum required benefit. Note that the planned margin may be – and typically is – based on a simple, unconditional meta-analysis estimate of $RR_{P/C}$. In this situation $\Delta_{Plan} = (LCL_{95}(\widehat{RR}_{P/C})^{-\rho}$, where $\widehat{RR}_{P/C}$ is derived from a

8

standard (non-regression-based) meta analysis of prior placebo-controlled trials.

Similarly, an NI trial might have been planned based on a fixed MCID, which may be based on investigator consensus, expert opinion or both. For example, it might be determined that regardless of study population characteristics, it is essential that the experimental product provide a risk of reduction of at least 10% ($RR_{E/P} = 0.90$) over what would be expected with placebo. In this case Δ would be defined as

$$\Delta_{MCID} = RR_{MCID} = 0.90, \tag{13}$$

where RR_{MCID} is the risk reduction corresponding to the pre-specified MCID. Since relative risks increase in value as benefit decreases, RR_{MCID} is actually the maximum allowable relative risk. Note that this approach could also be used in the planning phase when defining the original margin.

The second strategy for specifying Δ allows the required level of effectiveness to vary depending on the estimated effectiveness of the active-control therapy in the current trial. This can be done by defining Δ as a proportion of the benefit provided by the active-control therapy in the *observed* study population. For instance, assume that study population characteristics $\boldsymbol{x}^T$ are observable at baseline, and let $\boldsymbol{z}^T$ be dynamic features assessed during followup. Then define $\Delta_{Estimated}$ based on the level of active-control effectiveness, as follows:

$$\Delta_{Estimated} = (LCL_{95}(\widehat{RR}_{P/C}(\boldsymbol{x}^T, \boldsymbol{z}^T)))^{-\rho}, \tag{14}$$

where $\hat{R}R_{P/C}(\boldsymbol{x}^T, \boldsymbol{z}^T)$ is computed using the meta-analysis equation (4), and ρ is the conserved proportion of the active-control treatment effect. This gives the adaptive margin:

$$\delta^a = LCL_{95}(\widehat{RR}_{P/C}(\boldsymbol{x}^T, \boldsymbol{z}^T))^{1-\rho}, \tag{15}$$

Alternatively, Δ may be defined by selecting the more stringent of the two choices Δ_{MCID} or $\Delta_{Estimated}$, which is accomplished by choosing the minimum:

$$\begin{aligned} \Delta_{Min} &= \text{Min}(\Delta_{MCID}, \Delta_{Estimated}) \\ &= \text{Min}(RR_{MCID}, (LCL_{95}(\widehat{RR}_{P/C}(\boldsymbol{x}^T, \boldsymbol{z}^T)))^{-\rho}) \end{aligned} \tag{16}$$

This will typically mean using $\Delta_{Estimated}$ when the active-control effect is large (i.e. substantially more than the MCID), and using Δ_{MCID} when the effect of the active control is estimated to be relatively small. Using this method prevents the level of required effectiveness from diminishing too far in a study population where the active-control therapy is thought to be not working well, for example due to low adherence.

Using a flexible value for Δ may be desirable when the control therapy has different levels of effectiveness in different populations. For example, assume that the control treatment has been shown to be more effective in adults that in adolescents, and that, despite efforts to recruit adults, the study population turned out to be mostly adolescents. Because the control treatment is expected to be less effective in the observed study population, it may be reasonable to reduce the requirements for the experimental therepy. By using the planned value for ρ, we can increase the value of Δ using equation (14), thus increasing the

9

NI margin δ to a level that preserves the desired proportion of the active-control benefit in a mostly adolescent population.

6   Adapting the Statistical Hypotheses

Once the adapted margin δ^a has been specified, this margin becomes the null hypothesis for statistical inference. Although the nominal value of the null hypothesis will have changed from the planning stage, the adapted null still corresponds to the pre-planned amount of benefit that the experimental therapy is required to provide. When control-therapy effectiveness is variable, and the null hypothesis is expressed in relation to the control therapy, the nominal value of the null hypothesis must change in order to determine whether the experimental treatment produces the required level Δ of absolute effectiveness over placebo. Note that if a trial is planned based on (5), the analysis margin (and hence the null hypothesis) will be the same as the planning margin if the constancy assumption is met, i.e. if $\boldsymbol{x}^T = \boldsymbol{x}^A$ and $\boldsymbol{z}^T = \boldsymbol{z}^A$.

Just as the nominal value of the null hypothesis can change under an adaptive NI margin strategy, so too can the nominal value of the alternative hypothesis. As discussed in section 4, in an NI trial the alternative hypothesis is expressed as a relative risk in relation to the active control, but the effect size that defines power and sample size depends on both the alternative hypothesis and the planned NI margin. We define the effect size Ω as the ratio ξ/δ, which represents the difference between the null and alternative hypotheses on the relative risk scale. For example, if the alternative hypothesis is $\xi = 0.90$ and the NI margin δ is 1.1, the effect size used to compute power and sample size is $\Omega = \xi/\delta = 0.82$.

Because active-control effectiveness – and hence the NI margin – is a moving target under non-constancy, it is useful to anchor the alternative hypothesis to the hypothetical placebo arm, just as we did for the null. Let Ω_{Plan} be defined as the target effect size of the experimental treatment over placebo, defined as

$$\Omega_{Plan} = \xi/LCL_{95}(\widehat{RR}_{P/C}(\boldsymbol{x}^A, \boldsymbol{z}^A)). \tag{17}$$

The effect size Ω_{Plan} can be thought of as the 'absolute' target benefit of the experimental product, as opposed to the relative benefit over the active-control therapy, and this target remains fixed even when the active-control effect changes as a result of non-constancy. The planned alternative hypothesis can be expressed as a function of Ω_{Plan} as

$$\xi = \Omega_{Plan} \times LCL_{95}(\widehat{RR}_{P/C}(\boldsymbol{x}^A, \boldsymbol{z}^A)), \tag{18}$$

and once the values for $\boldsymbol{x}^T$ and $\boldsymbol{z}^T$ have been observed we can adjust the nominal value of ξ to reflect the pre-planned target effect. Using the observed study population characteristics gives the following adaptive alternative hypothesis:

$$\xi^a = \Omega_{Plan} \times LCL_{95}(\widehat{RR}_{P/C}(\boldsymbol{x}^T, \boldsymbol{z}^T)). \tag{19}$$

Although the nominal value of ξ will have changed, by fixing Ω we preserve the initial, planned target effect of the experimental treatment over placebo.

10

There are circumstances in which an investigator may not want to change the alternative hypothesis, even when faced with non-constancy. For example, a common non-inferiority alternative hypothesis is $\xi = RR_{E/C} = 1.0$, meaning that trial is designed to establish non-inferiority in the situation where experimental and active-control treatments are equally effective. Regardless of whether the study population turns out as planned, investigators may want to ensure that the trial is powered to establish non-inferiority when the experimental treatment is equivalent to the control treatment (often the standard of care) in the study population.

7 Type-I Error and Power

With the adapted null and alternative hypotheses established, we now reassess the Type-I error probability and power. In contrast to the scenario described in section 4, after the NI margin and null hypothesis have been adjusted to reflect the observed study population, there is no longer a mismatch between the observed relative risk under the null (based on $\boldsymbol{x}^T$ and $\boldsymbol{z}^T$) and the unadjusted NI margin (based on $\boldsymbol{x}^A$ and $\boldsymbol{z}^A$), and hence no inflation or reduction in Type-I error.

Statistical power will depend on sample size and the ratio of the adjusted alternative and null hypotheses, i.e. the adjusted effect size. Provided that this ratio does not change, power will be not be affected. For example, if the NI margin is planned using the meta-analysis regression in (5), the planned effect size can be written as the ratio of (18) and (5) as follows,

$$
\begin{aligned}
\frac{\xi}{\delta} &= \frac{\Omega_{Plan} \times LCL_{95}(\widehat{RR}_{P/C}(\boldsymbol{x}^A, \boldsymbol{z}^A))}{(LCL_{95}(\widehat{RR}_{P/C}(\boldsymbol{x}^A, \boldsymbol{z}^A)))^{1-\rho}} \\
&= \Omega_{Plan} \times (LCL_{95}(\widehat{RR}_{P/C}(\boldsymbol{x}^A, \boldsymbol{z}^A)))^{\rho}.
\end{aligned}
\tag{20}
$$

If δ is computed using the pre-specified value Δ_{Plan}, the adjusted effect size does not change:

$$
\begin{aligned}
\frac{\xi^a}{\delta^a} &= \frac{\Omega_{Plan} \times LCL_{95}(\widehat{RR}_{P/C}(\boldsymbol{x}^T, \boldsymbol{z}^T))}{\Delta_{Plan} \times LCL_{95}(\widehat{RR}_{P/C}(\boldsymbol{x}^T, \boldsymbol{z}^T))} \\
&= \frac{\Omega_{Plan}}{(LCL_{95}(\widehat{RR}_{P/C}(\boldsymbol{x}^A, \boldsymbol{z}^A)))^{-\rho}} \\
&= \Omega_{Plan} \times (LCL_{95}(\widehat{RR}_{P/C}(\boldsymbol{x}^A, \boldsymbol{z}^A)))^{\rho}.
\end{aligned}
\tag{21}
$$

In other words, if the null and alternative hypotheses are adjusted by using the pre-specified values for Ω_{Plan} and Δ_{Plan}, the effect size ratio remains the same, and there is no loss or gain in power. However, if the minimum required benefit Δ is allowed to vary depending on observed population characteristics, the effect size will no longer remain constant. Using estimated-effectiveness to define Δ as in (14), the adjusted effect becomes

11

$$\frac{\xi^a}{\delta^a} = \frac{\Omega_{Plan} \times LCL_{95}(\widehat{RR}_{P/C}(\boldsymbol{x}^T, \boldsymbol{z}^T))}{\Delta_{Estimated} \times LCL_{95}(\widehat{RR}_{P/C}(\boldsymbol{x}^T, \boldsymbol{z}^T))} \tag{22}$$
$$= \Omega_{Plan} \times (LCL_{95}(\widehat{RR}_{P/C}(\boldsymbol{x}^T, \boldsymbol{z}^T)))^\rho,$$

which is a function of observed population characteristics. This means that if the control therapy is observed to be more effective than planned, the effect size, and consequently the power, will be higher than planned. On the other hand, if the control therapy is estimated to be worse than planned, the effect size will be smaller than planned and power will be less than planned.

Similarly, if the null hypothesis is adjusted but the alternative hypothesis remains constant at, for example, $\xi = 1.0$, the effect size, and hence power, will vary depending on the adjusted null, following the equation

$$\frac{\xi}{\delta^a} = \frac{1}{\Delta \times LCL_{95}(\widehat{RR}_{P/C}(\boldsymbol{x}^T, \boldsymbol{z}^T))} \tag{23}$$

where Δ could be Δ_{Plan}, $\Delta_{Adjusted}$, Δ_{MCID}, or Δ_{Min}. In this case the effect size is just the inverse of the non-inferiority margin. Stronger observed effects in the control arm (i.e. larger values for placebo-versus-control relative risk) will imply larger effect sizes (smaller absolute value of the relative risk) and hence higher power. Conversely, weaker than expected effects in the control arm would yield smaller effect sizes and reduced power. This is intuitive, since if the two interventions are equally effective, it is more difficult to establish non-inferiority using a narrower margin.

8   Examples

In trials with a fixed, pre-planned number of endpoint events, statistical power will depend directly on the ratio of the adjusted alternative and null hypotheses, as determined at the end of the trial. As described in previous sections, there are a variety of choices for adjusting the statistical hypotheses, and hence a number of possible combinations of adjusted null and adjusted alternatives. In the following we consider three distinct scenarios that represent realistic options for evaluating a new HIV prevention intervention. The planned study population has 60% adherence to the active-control drug, and is 50% female.

1. *Preserve Planned Benefit* – In this scenario the investigators wish to ensure that the experimental intervention provides a pre-planned level of benefit over placebo, Δ_{Plan}, based on preserving 50% of the active control benefit ($\rho = 0.5$), and wish to power the trial based on the alternative hypothesis that the experimental intervention has a specific amount of benefit over placebo, Ω_{Plan}, corresponding to 20% risk reduction ($\xi = 0.80$) at the planned covariate levels. The study team believes that the primary source of non-constancy is likely to be variable adherence to the active-control drug. Since the primary objective of the trial is to assess biological efficacy, investigators wish to minimize the impact of unpredictable drug-adherence on the study outcome.

12

2. *Preserve Proportional Benefit* – In this scenario the investigators are not sure how well the active-control therapy will work in the study population, but wish to ensure that the experimental therapy preserves at least half of the benefit that the active-control therapy can provide in whatever population is recruited ($\rho = 0.5$). To this end they will use $\Delta_{Estimated}$, and with the hope that the new intervention will be about 10% better than the control intervention in the study population (whatever its composition might be), they will use the fixed alternative $\xi = 0.90$.

3. *Preserve Proportional Benefit and MCID* – This is is similar to scenario 2 except that the investigators wish to place a limit on how flexible the adaptive margin can be. There is some concern that the control therapy may perform so poorly in the study population as to be little better than placebo, and the study team wishes to assure that the experimental therapy provide at least 10% benefit over placebo, even if the active-control effectiveness is very low. For this purpose they use $\Delta_{Min} = Min(\Delta_{Estimated}, 0.90)$. The alternative hypothesis is again fixed at $\xi = 0.90$.

Figure 3 shows power as a function of observed effectiveness (vs placebo) for each of the three scenarios. Sample sizes for each scenario are computed based on the planned study population composition (60% adherence and 50% female), yielding approximately 49% effectiveness, and hence it is at that point that power is always exactly 90% for all scenarios.

Scenario 1 requires a sample size of 204 events to achieve 90% power and a type-I error rate of 2.5%. The adjusted null and alternative hypotheses are based on the pre-planned values for $\Delta_{Plan} = LCL_{95}(\widehat{RR}_{P/C}(x = 0.5, z = 0.6))^{-0.5} = 0.79$ and $\Omega_{Plan} = 0.8/LCL_{95}(\widehat{RR}_{P/C}(x = 0.5, z = 0.6)) = 0.50$. As per equation (21), regardless of the degree of non-constancy, the adjusted effect size does not change from the planned effect size,

$$\frac{\xi^a}{\delta^a} = \frac{\Omega_{Plan}}{\Delta_{Plan}} = \frac{\xi}{\delta} = 0.50/0.79 = 0.63, \tag{24}$$

and there is no loss of power (solid line, Figure 3).

Scenarios 2 and 3 both have a planned sample size of 371 events. In scenario 2 the adjusted null depends on the observed values x^T and z^T, while the alternative hypothesis remains fixed at 0.90, yielding an effect size that depends on the observed values,

$$\frac{\xi^a}{\delta^a} = \frac{\xi}{\delta^a} = 0.90/LCL_{95}(\widehat{RR}_{P/C}(x^T, z^T))^{0.5}, \tag{25}$$

and hence lower levels of active-control effectiveness lead to smaller effect sizes and loss of power (dashed line, Figure 3).

Senario 3 mimics scenario 2 until active-control effectiveness drops to the point where $\Delta_{Estimated}$ becomes larger than 0.90, at which point the effect size drops as a function of $LCL_{95}(\widehat{RR}_{P/C}(x^T, z^T))$, (as opposed to a function of the square root of the LCL_{95} as in (25)), as

13

$$
\begin{aligned}
\frac{\xi^a}{\delta^a} &= \frac{\xi}{\delta^a} \\
&= 0.90/(RR_{MCID} \times LCL_{95}(\widehat{RR}_{P/C}(x^T, z^T))) \\
&= 0.90/(0.90 \times LCL_{95}(\widehat{RR}_{P/C}(x^T, z^T))) \\
&= 1/LCL_{95}(\widehat{RR}_{P/C}(x^T, z^T))
\end{aligned}
\tag{26}
$$

and hence power drops more rapidly as effectiveness drops (dotted line, Figure 3) below approximately 33%. In both scenarios 2 and 3 power becomes inflated for higher-than-planned levels of active-control effectiveness due to larger effect sizes in combination with constant sample sizes.

9 Dynamic features and sampling variation

It will not always be possible to measure dynamic features completely and accurately. Lab-based drug adherence assessment, for example, requires costly, regular collection and testing of samples. A more efficient approach may be to sample a random subset of participants, at a random set of time points, and generate an estimate of adherence rates. To compute the adapted margin, a sample-based estimate $\hat{z}^T$ may be substituted into (11) in place of z^T. The degree to which sampling variation may influence the margin depends on both the estimated regression coefficients in (4) and the sampling distribution of z^T. Potential bias introduced by sampling can be quantified by

$$
\delta(\boldsymbol{x}^T, \boldsymbol{z}^T, \rho) - \int \delta(\boldsymbol{x}^T, \hat{\boldsymbol{z}}^T, \rho) f(\hat{\boldsymbol{z}}^T)
\tag{27}
$$

where $f()$ is the sampling distribution of $\hat{\boldsymbol{z}}^T$ and $\delta()$ is defined as in equation (5). The true value of $\boldsymbol{z}^T$ will generally not be known, but sample sizes for $\hat{\boldsymbol{z}}^T$ should be chosen to provide precise estimates of $\boldsymbol{z}^T$ and minimize potential bias.

10 Discussion

Specifying non-inferiority margins can be a somewhat subjective process, and the validy of these margins relies heavily on the assumption of constancy. In clinical-trial settings with known effect modifiers, we have shown that meta-analysis regression can provide objective, population-specific margins for planned trials, and thus reduce the chances and degree of constancy violation. In addition, once a trial is complete, the same meta-analysis regression model can be used to adapt the end-of-study margin according to observed effect modifiers in the study population, thereby preserving planned levels of Type-I error and power.

Adaptive margins depend on adapted statistical hypotheses, and careful consideration must be given to how the null and alternative hypotheses should be updated. Certain choices for updating the hypotheses, however logical and reasonable, will result in low (or

14

high) power due to sample sizes that do not correspond to the adapted effect size. As illustrated in examples 2 and 3 above, if active-control effectiveness is less than planned, and the amount of benefit (Δ) required by the adapted NI margin is allowed to vary depending of estimated effectiveness, power will be less than planned. This does not imply that such choices for Δ are incorrect - such decisions should be pre-specified based on sound clinical judgement - but the potential for reduced power is something investigators should be aware of.

A key element of our approach is that NI-margin adaptation does not depend on the observed effect size. Meta-regression parameters are estimated prior to starting the new trial, and depend entirely on efficacy data which is external to the trial. The end-of-study margin depends only on observed effect modifiers in the new trial, and, to avoid the potential for bias, should be computed in the absence of any knowledge of the observed effect size. It may be reasonable to update the meta-regression with data from external trials that conclude during the conduct of the new trial, but the decision to update the model should also be pre-specified.

Our meta-analysis regression method is an important extension of Nie and Soon's approach [2] in that it allows for the inclusion of post-randomization dynamic features, which in some settings may be the most influential effect modifiers. In trials of HIV PrEP in high-risk populations, for example, the importance of medication adherence substantially outweighs any known effect modifier that could be measured at baseline. In fact, for any medication controlled by the participant, adherence may be the most important effect modifier.

Rohmel and Kieser [14] address the idea of "variable margins" which have been proposed as way to construct more reasonable NI margins for binary-endpoint trials when failure rates in the active-control arm are substantially different than expected. The variable-margin approach allows the end-of-study margin to depend on observed failure rates, and when these rates are lower than a preset threshold, the margin switches from a difference-in-proportions scale to the odds-ratio scale. Although changing the scale provides some flexibility in the face on non-constancy, unlike the meta-analysis approach it does not address the question of how to define an appropriately sized margin based on observed levels of active-control effectiveness.

The choice of endpoint-assessment scale is nevertheless important. Although our development of the adaptive NI margin approach is limited to the relative risk scale, the same approach can be used for risk differences by applying the log transformation to (5), so that the conserved proportion ρ becomes a multiplier on a risk difference instead of an exponent on a risk ratio. The required benefit Δ and target benefit Ω would then become additive differences instead of multipliers. In applications where event rates are fairly high (a threshold 20% is often used in the variable-margin approach) the risk-difference scale may be more intuitive and useful.

Meta-analysis based adaptive NI margins have important limitations. First, meta-analysis results are only as good as the trials upon which they are based, and it is not always the case that multiple, high-quality trials are available to analyze. Particularly for meta-analysis regression, multiple trials with accurate measures of important effect modi-

15

fiers are necessary to achieve a reliable regression model. While some fields of study are rich with existing, high-quality trial data, others are not. Also, critical effect modifiers such as adherence can be measured in very different ways, yielding different results, for example self-reported drug adherence has been shown to be consistently and significantly higher than lab-based adherence measures [11; 10]. Unless parallel measures are used across studies, the validity of meta-regression model parameters is questionable.

Similarly, assessment of effect modifiers in the new trial must be consistent with the prior studies used to construct the margin. If measurement is not consistent, regression-based estimates of active-control effectiveness are unlikely to be accurate, in which case the adapted NI margin will not result in the desired trial characteristics. In addition, if measurement methods are not clearly pre-specified, the door may be open for inappropriate manipulation of the NI margin. For example, if it is known that higher measured adherence corresponds to higher effectiveness, one need only choose an inflated adherence measure, such as self-report, to increase the estimated active-control effect and relax the NI margin, thus making drug approval more likely.

Relying on on between-trial effect-modifier estimates does not come without risks. Unmeasured differences in study populations can introduce ecological bias [5], and the evaluation of gender as an effect modifier in oral PrEP trials represents a classic example. If drug adherence were not included in the meta-regression model, it would likely appear that oral PrEP is much less effective in women than in men. This result, however, would be due to the fact the in several large PrEP trials in women, adherence was very low, whereas in most trials that include men, adherence was moderate or high (Figure 1). It is therefore important to use caution when using trial-level data to evaluate effect modifiers, and if post-randomization factors are not included in the model, consider using individial participant data as proposed by Hua et al. [5].

Sampling error may also introduce bias to an adapted NI margin. In situations where it is difficult to obtain sufficiently precise estimates of $\boldsymbol{z}^T$, the potential for bias may be higher. We would suggest assessing the potential for bias by evaluating (27) using a variety of reasonable values for $\boldsymbol{x}^T$ and $f()$. A sampling plan should select a random subset of participants, and also, if appropriate, consider and random subset of time points during the trial in order to account for temporal trends.

Even if effect modifiers could be measured perfectly, it is important to remember that it is always possible that there are unmeasured effect modifiers. If non-constancy results from factors that cannot be measured or have otherwise not been included in the meta-regression, the adaptive margin will not correct for these unmeasured factors. In the study of HIV prevention, for instance, measuring exposure to the HIV virus can be challenging. In a population where sexual contact is the primary form of transmission, it is often not possible to determine whether an individual has engaged in unprotected sex, and if so, whether their sex partner was HIV infected. If exposure to HIV is substantially less than expected, and hence HIV incidence is very low, active-control effectiveness may be significantly lower than is predicted by the model [15].

Meta-analysis regression methods offer a way to define NI margins appropriate to a specific study population, and to adapt end-of-study margins to observed characteristics of

16

study populations. In the presence of known, measurable effect modifiers, these methods can substantially reduce the undesirable effects of non-constancy. Our ongoing work includes investigating ways to adapt NI margins during interim analyses in group sequential trials, and appropriate ways to update samples sizes as result of adaptations to the margin and hypotheses.

References

[1] Everson-Stewart S. PhD Thesis, University of Washington 2010.

[2] Nie L, Soon G. A covariate-adjustment regression model approach to noninferiority margin definition. *Statistics in Medicine* 2010; .

[3] FDA. Guidance for industry: Non-inferiority trials. *Technical Report* 2010.

[4] Thompson S, Higgins J. How should meta-regression analysis be undertaken and interpreted? *Statistics in Medicine* 2002; .

[5] Hua H, Burke DL, Crowther MJ, Ensor J, Smith CT, Riley RD. One-stage individual participant data meta-analysis models: estimation of treatment-covariate interactions must avoid ecological bias by separating out within-trial and across-trual information. *Statistics in Medicine* 2016; .

[6] Sutton A, Abrams K, Jones D, Sheldon T, Song F. *Methods for meta-analysis in medical research*. John Wiley and Sons, 2000.

[7] Baeten JM, Donnell D, Ndase P. Antiretroviral prophylaxis for hiv prevention in heterosexual men and women. *N Engl J Med* 2012; .

[8] Thigpen M, Kebaabetswe P, Paxton L. Antiretroviral preexposure prophylaxis for heterosexual hiv transmission in botswana. *N Engl J Med* 2012; .

[9] Choopanya K, Martin M, Suntharasamai P. Antiretroviral prophylaxis for hiv infection in injecting drug users in bangkok, thailand (the bangkok tenofovir study): a randomised, double-blind, placebo-controlled phase 3 trial. *Lancet* 2013; .

[10] Marrazzo J, Ramjee G, Richardson B. Tenofovir-based preexposure prophylaxis for hiv infection among african women. *N End J Med* 2015; .

[11] Van Damme L, Corneli A, Ahmed K. Preexposure prophylaxis for hiv infection among african women. *N End J Med* 2012; .

[12] Molina L, Capitant C, Spire B, Pialoux G, Cotte L. On-depand preexposure prophylaxis in men at high risk for hiv-1 infection. *N End J Med* 2015; .

[13] Grant R, Lama J, Anderson P, McMahan V. Preexposure chemorophylaxis for hiv prevention in men who have sex with men. *N End J Med* 2010; .

[14] Rohmel J, Kieser M. Investigations on non-inferiority - the food and drug administration draft guidance on treatments for nosocomial pneumonia as a case for exact tests for binomial proportions. *Statistics in Medicine* 2012; .

[15] Coley RY, Brown ER. Estimating effectiveness in hiv preention trials with bayesian hierarchical compound poisson frailty mode. *Statistics in Medicine* 2016; .

Figure 1: PrEP effectiveness plotted against trial-level adherence, as measured by plasma TDF, for all randomized trials of oral PrEP versus placebo where an objective adherence measure was available. Circle sizes are proportional to the number of incident HIV infections during the trial The fitted regression line and 95% confidence bounds (dashed lines) are shown for men.
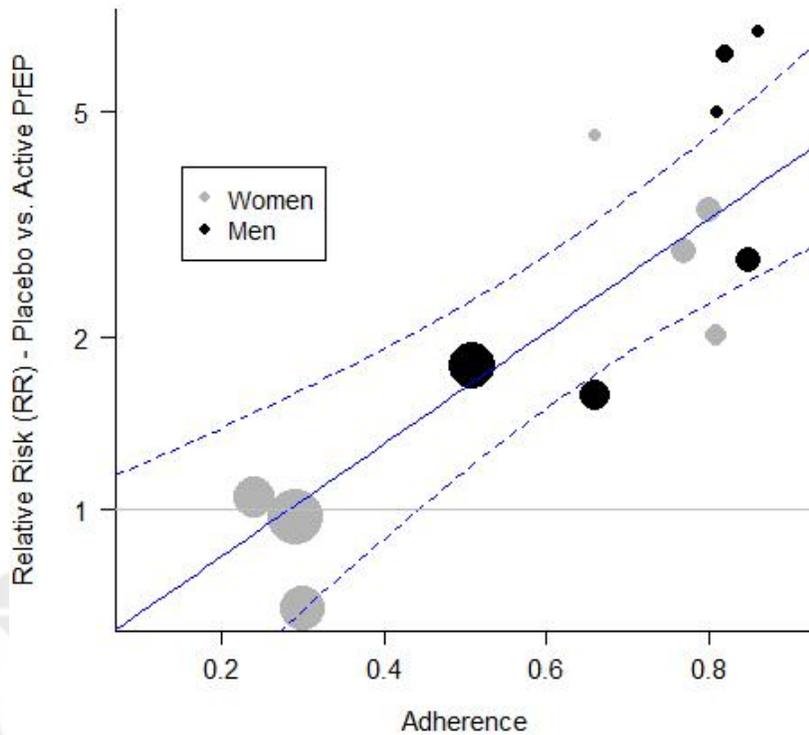
Figure 2: Type I and type II error probabilities according to the true level of effectiveness (% risk reduction vs. placebo) in the active-control arm. Rates are for a hypothetical NI trial with NI-margin=1.3, effect size versus active-control=0.7, sample size = 110 events, and planned active-control effectivess=50%.
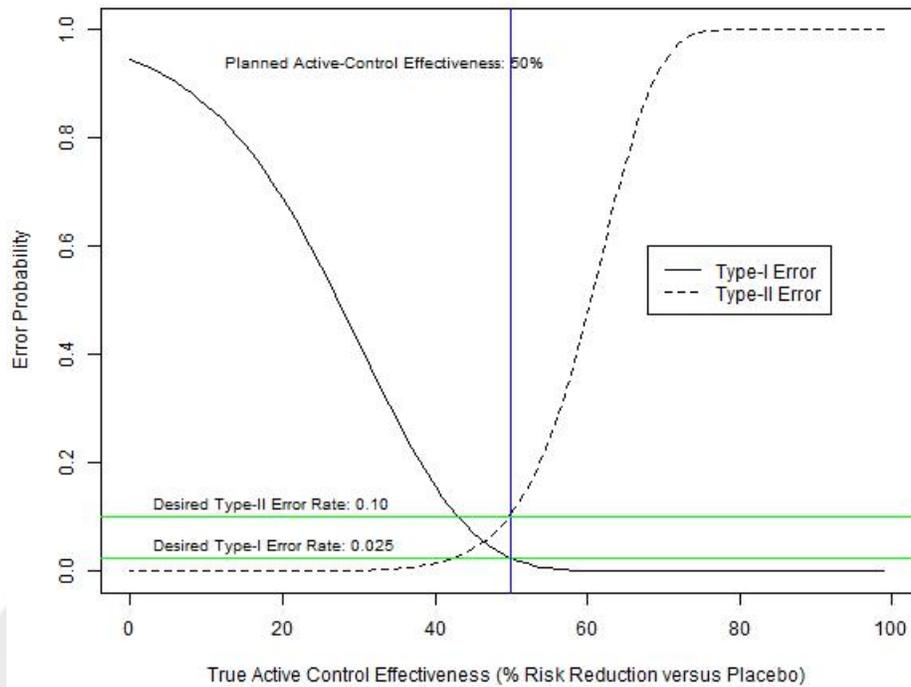
Figure 3: Power for three adaptive-margin scenarios, according to the true level of effectiveness (% risk reduction vs. placebo) in the active-control arm, for a HIV prevention trials. In scenario 1 both the null and alternative hypotheses are adapted accoring to pre-planned levels of required benefit. In scenario 2 and 3 the level of required benefit is allowed to vary according to observed study population characteristics. In scenario 3 the null hypothesis is also required to preserve at least a (pre-specified) MCID over a hypothetical placebo arm. Scenario 1 requires 204 observed events, while scenarios 2 and 3 require 371 observed events.