LIKELIHOOD REWEIGHTING METHODS TO REDUCE POTENTIAL BIAS IN NONINFERIORITY TRIALS WHICH RELY ON HISTORICAL DATA TO MAKE INFERENCE[1]

BY LEI NIE, ZHIWEI ZHANG, DANIEL RUBIN AND JIANXIONG CHU

The U.S. Food and Drug Administration

It is generally believed that bias is minimized in well-controlled randomized clinical trials. However, bias can arise in active controlled noninferiority trials because the inference relies on a previously estimated effect size obtained from a historical trial that may have been conducted for a different population. By implementing a likelihood reweighting method through propensity scoring, a study designed to estimate a treatment effect in one trial population can be used to estimate the treatment effect size in a different target population. We illustrate this method in active controlled noninferiority trials, although it can also be used in other types of studies, such as historically controlled trials, meta-analyses, and comparative effectiveness analyses.

1. Introduction. The code of federal regulations (CRF) Chapter 21, Section 314.126, states that "The purpose of conducting clinical investigations of a drug is to distinguish the effect of a drug from other influences..." and the purpose is achieved through "adequate and well-controlled clinical investigation." According to the CRF, an adequate and well-controlled trial has a number of characteristics, including: (1) "The method of assigning patients to treatment and control groups minimizes bias and is intended to assure comparability of the groups with respect to pertinent variables such as age, sex, severity of disease..." and (2) "Adequate measures are taken to minimize bias on the part of the subjects, observers, ..." Characteristic (1) and part of (2) aim to minimize bias through balancing the population between the two treatment arms.

By conducting well-controlled clinical trials, we generally anticipate that systematic bias is minimized in superiority trials. However, this belief may be more tenuous in noninferiority trials. Note that noninferiority trials are the major vehicle to evaluate new treatments in many disease areas, after the pioneering consideration of ethical issues in Placebo-controlled trials by Rothman and Michels (1994).

Consider a Palivizumab-controlled noninferiority trial of Motavizumab for prophylaxis of serious respiratory syncytial virus (RSV) disease in high risk children [Carbonell-Estrany et al. (2010)]. This trial will be called MOTA throughout the paper. The goal of the trial was to evaluate whether Motavizumab was

noninferior to Palivizumab in the rate of hospitalization attributed to RSV. Let $\hat{\mu}_{TC}$ be the estimated log-odds ratio of Palivizumab vs. Motavizumab, and let $\hat{\mu}_{CP}$ be the estimated log-odds ratio of Placebo vs. Palivizumab. Because the log-odds ratio of Placebo vs. Motavizumab cannot be estimated directly (the noninferiority trial does not have a placebo arm), $\hat{\mu}_{TC} + \hat{\mu}_{CP}$ is often used as an indirect estimate, with a standard error of $\sqrt{\sigma_{TC}^2 + \sigma_{CP}^2}$. We may consider the noninferiority of Motavizumab to Palivizumab to be met at level α if $\hat{\mu} = (\hat{\mu}_{TC} + \hat{\mu}_{CP})/\sqrt{\sigma_{TC}^2 + \sigma_{CP}^2} > Z_\alpha$, where Z_α is the $(100 - \alpha)$th percentile of a standard normal distribution. In this example, $\hat{\mu}_{CP}$ and its standard error σ_{CP} were obtained from an earlier Placebo-controlled trial of Palivizumab, Impact-RSV Study Group (1998), in which $\hat{\mu}_{CP} = 0.86$ (corresponding to odds ratio of 2.4) with a standard error of 0.21. This trial will be called IMPACT throughout the paper. Now, keeping in mind the fact that the statistics $\hat{\mu}$ synthesizes $\hat{\mu}_{CP}$ and $\hat{\mu}_{TC}$, with the former estimated from the IMPACT population and the latter estimated from the MOTA population, we illustrate the following issues. First, both MOTA and IMPACT enrolled subjects exclusively from two disjoint subgroups: (1) children ≤ 24 months with a clinical diagnosis of Bronchopulmonary dysplasia (BPD); and (2) children with ≤ 35 weeks gestation and ≤ 6 months, who did not have BPD. The proportion of subjects with BPD was 51% in IMPACT and only 22% in MOTA. Second, treatment heterogeneity of Palivizumab was observed in these two subgroups in IMPACT. For subjects enrolled with BPD, the odds ratio was 4.88 with a 95% C.I. of (2.17, 10.96), and for subjects enrolled without BPD, the odds ratio was 1.72 a 95% C.I. of (1.06, 2.79) (see Section 4). The Wald Chi-square test of treatment by BPD interaction through a logistic regression was significant with a p-value of 0.03. Because of the population difference and treatment heterogeneity, an appropriate odds ratio $\hat{\mu}_{CP}$ used in $\hat{\mu}$ should reflect the population of MOTA, while the value of 0.86 instead reflects the population of IMPACT. Using data provided in Section 4, we obtain the adjusted incidence rate of Placebo in the MOTA population of $34/266 \times 22\% + 19/234 \times 78\% = 9.2\%$, and the adjusted incidence rate of Palivizumab of $39/496 \times 22\% + 9/506 \times 78\% = 3.1\%$. Therefore, the adjusted log-odds ratio is 1.14 and the adjusted odds ratio is 3.1. Consequently, the adjusted log-odds ratio of Placebo vs. Palivizumab in the MOTA population should be better quantified as 1.14 rather than 0.86, the unadjusted log-odds ratio $\hat{\mu}_{CP}$. The difference between 1.14 and 0.86 is a bias associated with this inference.

In the previous example, it was easy to adjust for the population difference, which only involved heterogeneity in BPD status. In some other examples, the situation could be more complicated. For example, in the development of Elvitegrevir [Molina et al. (2012)], the trial population was different from the historical trial population in several characteristics, for which treatment heterogeneity has been reported [Cooper et al. (2008)].

These examples show that analysis of a noninferiority trial may rely on a combination of information from the trial itself and one or more historical trials. The main issue is that the populations of the noninferiority and historical trials may be different. If treatment heterogeneity is present, an inference that does not adjust for the population difference can be biased.

Covariate adjustment approaches [Zhang (2009) and Nie and Soon (2010)] have been proposed to address the problem. Both approaches involve a regression model relating the clinical outcome to treatment and relevant covariates. They cannot be directly applied to obtain the marginal (crude) odds ratio, which is the prespecified primary endpoint in the aforementioned examples.

This paper proposes a calibration method through likelihood reweighting so that a study designed to estimate a marginal treatment effect size for one trial population (e.g., IMPACT) may be used to calibrate the effect size in a different but closely related study population (e.g., MOTA). We prove that the maximum likelihood estimator for this reweighted likelihood is a consistent estimator of the treatment effect size in the targeted population. In addition, we also propose a nonparametric approach based on the calibration method.

The proposed calibration approach using the likelihood reweighting method is (asymptotically) equivalent to the covariate adjustment approach in some cases such as linear regression, however, they are different in other cases. The choice between the two approaches can be subtle and subjective. An important consideration is to make sure that $\hat{\mu}_{TC}$ and $\hat{\mu}_{CP}$ in the statistics $\hat{\mu}_{ks} = (\hat{\mu}_{TC} + \hat{\mu}_{CP})/\sqrt{\sigma_{TC}^2 + \sigma_{CP}^2}$ have similar interpretations. Specifically, if $\hat{\mu}_{TC}$ is a marginal (i.e., overall) treatment effect as in the previous two examples and in most randomized clinical trials, then $\hat{\mu}_{CP}$ should probably be calibrated using the method presented in this paper so as to maintain the marginal interpretation. In Section 3.3 we also make some observations on the likelihood reweighing method as an alternative to the covariate adjustment approach used in randomized clinical trials, along with the differences noted in the literature.

Although this paper mainly targets noninferiority trials, the results are also applicable to historically controlled trials, which have similar issues [Friedman, Furberg and Demets (1998)]. A comparison of the likelihood reweighting method to related methods in historically controlled trials, for example, Zhang (2007), Signorovitch et al. (2010) and Signorovitch et al. (2011), is provided in the supplement [Nie et al. (2013)]. This paper focuses on calibrating the treatment effect size from one population to another population which is different but overlapping. It is related to but different from studies generalizing results from a subpopulation to a strictly larger population (whole population); see Cole and Stuart (2010), Greenhouse et al. (2008), Weisberg, Hayden and Pontes (2009) and Frangakis (2009), among others. These references are restricted to the clinical trial literature, although other areas, such as observational studies, involve similar problems.

2. Motivation, assumptions, and notation.

2.1. *Motivation.* The idea behind our method is simple. Recall the Introduction where we obtained the expected incidence rate of Placebo in the MOTA population as

$$12.8\% \times 22\% + 8.1\% \times 78\% = \frac{\sum_{i=1}^{500} y_i \, p_{x_i,\text{MOTA}} / p_{x_i,\text{IMPACT}}}{500},$$

where $p_{x_i,\text{MOTA}}$ and $p_{x_i,\text{IMPACT}}$ are the percentage of Placebo subjects with characteristics x_i in MOTA and IMPACT trials, respectively. When $x_i = 1$ (a diagnostic of BPD), the $p_{x_i,\text{MOTA}}$ and $p_{x_i,\text{IMPACT}}$ are 22% and 53%, respectively; when $x_i = 0$, they are 78% and 47%. By defining $p_{x_i,\text{MOTA}} / p_{x_i,\text{IMPACT}}$ as r_i, the expected incidence rate of Placebo in the MOTA population is simply the mean of reweighted response from all subjects and the weight reflects the change of population difference with respect to BPD status.

In many other situations, the parameters cannot be directly calibrated as shown in this example. They can, however, be estimated using a likelihood approach to be described shortly.

Robins and colleagues gave an intuitive explanation of how the inverse probability weighting approach reduces bias in the context of estimating marginal structural models (MSMs) in epidemiology [Robins, Hernán and Brumback (2000)]. Heuristically, weighting each subject by the inverse of the propensity score for the treatment actually received creates a confounding-free pseudo-population, where treatment assignment is independent of the potential outcomes. Typically, the inverse probability weighting approach is used to estimate marginal means of potential outcomes in an estimating equation framework. However, the insights of the work by Robins and colleagues certainly extend to likelihood-based inference and allow us to calibrate the treatment effect. Specifically, upon appropriately reweighting the likelihood function contributed by each subject, a calibrated treatment effect can be obtained. Before illustrating the reweighted likelihood approach, we introduce some notation and assumptions.

2.2. *Assumptions and notations.* Consider a trial conducted in a population P (e.g., the IMPACT population) to compare treatment 1 (e.g., Palivizumab) to treatment 2 (e.g., Placebo). We assume that a random sample from P is randomly assigned into these two treatment groups. The objective of the trial is to quantify the treatment effect size of treatment 1 relative to treatment 2 in population P. The objective of this paper is to calibrate the effect size of treatment 1 relative to treatment 2 from the original population to a different but closely related population P^* (e.g., MOTA population).

We assume that the populations P and P^* are different. In our first example, P refers to a population comprised of subjects with BPD (51%) and without BPD

(49%) and P^* refers to a population with a different composition (22% with BPD, 78% without BPD).

We also assume that the populations P and P^* are closely related and that the differences between P and P^* are entirely captured by the value of a predictive covariate (vector) X representing subjects' baseline disease characteristics. In addition, we assume that all subjects with covariate value $X = x$ are expected to have the same treatment effect, regardless of their origin (population P or population P^*). That is, subjects with the same covariate value X are exchangeable in P and P^*. In our first example, this means subjects with the same BPD diagnostic status, whether in the IMPACT population or the MOTA population, are exchangeable in terms of response to treatments.

The difference and close relationship between population P and P^* is further illustrated in mathematical form below after we clearly state the objective of the paper. Let Y be the response variable. We write $\mu_t(X) = E(Y|X, T = t)$ for the conditional mean response of subjects with covariate X who were assigned into treatment $T = t$, and $\mu_{tP} = \nu[E_{X \in P}\{\mu_t(X)\}]$ for the transformed marginal mean response with respect to population P. When $\nu(\cdot)$ is the identity function, μ_{tP} is the marginal mean; when Y is a binary variable and $\nu(\cdot)$ is the logit function, μ_{tP} is the log odds in the population P. μ_{tP} may be used to quantify the response of treatment $T = t$ from a historical trial, although this is not the focus of this paper but a by-product. Instead this paper focuses on noninferiority trials, in which we are interested in the treatment effect of treatment 1 vs. treatment 2. We thus consider $\mu_P = \pi[E_{X \in P}\{\mu_1(X)\}, E_{X \in P}\{\mu_2(X)\}]$ as a metric to measure treatment effect of treatment 1 vs. treatment 2.

In the historical trial (e.g., IMPACT), μ_{tP} or μ_P is estimated. However, the objective in this paper is to estimate $\mu_{tP^*} = \nu[E_{X \in P^*}\{\mu_t(X)\}]$ or $\mu_{P^*} = \pi[E_{X \in P^*}\{\mu_1(X)\}, E_{X \in P^*}\{\mu_2(X)\}]$ through calibration, without conducting a different trial in population P^* (MOTA population).

Let $F(x)$ and $F^*(x)$ denote the cumulative distribution functions of X in P and P^*, respectively, and let $f(x)$ and $f^*(x)$ be the corresponding probability density functions. We first assume that $f^*(x)/f(x) \neq 1$ for some $X = x$, which illustrates the differences between populations P and P^*. We also assume that $\infty > r(x) = f^*(x)/f(x)$ is well defined. Because the populations are fully described by X, this assumption means that any subject included in P^* should have representatives with the same measurements in population P. This highlights the close relationship between P and P^*. When a value of x does not present in P^*, then $r(x) = 0$. In this case, we shall not use the subjects in the historical trials with value x.

3. Calibration of treatment effect size through likelihood reweighting.
In our first example only the BPD status is considered and the weight is easy to define. However, in our second example many predictive covariates may need to be considered. In the latter case, the definition of the weight is straightforward using the concept of the propensity score [Rosenbaum and Rubin (1983)].

3.1. *Parametric approach.* Assume two random samples of size n_1, n_2 from population P are assigned into treatment 1 and treatment 2, respectively. We assume y_{it}, the ith subject's response from treatment group t, follows a generalized linear model (GLM) with canonical link,

$$(1) \qquad l_t(y, \theta_{tx}) = \exp\left\{\frac{y\theta_{tx} - b(\theta_{tx})}{a_{tx}(\varphi_{tx})} + c_{tx}(y, \varphi_{tx})\right\}.$$

Let $g(\cdot)$ be the canonical link function; then $\mu_t(X) = g^{-1}(\theta_t)$. We assume that $g(\cdot)$ is a monotone function with continuous second derivative functions. One possible metric to measure the treatment effect is $E_X(\theta_{tx})$ and another possible metric is $\mu_{tP} = g[E_{X\in P}\{\mu_t(X)\}]$. In the binomial-logistic regression case, we implicitly assume that the log odds is additive for the metric $E_X(\theta_{tx})$ and assume the proportion is additive for the metric $\mu_{tP} = g[E_{X\in P}\{\mu_t(X)\}]$. The former metric was used in the covariate adjustment approach of Nie and Soon (2010). The latter metric shall be used in the likelihood reweighting method, as introduced in this paper. These two metrics are related but usually are not identical in nonlinear models.

To estimate μ_{tP}, we construct the likelihood function

$$(2) \qquad \prod_{i=1}^{n_t} l_t(y_{it}, \alpha_t).$$

The maximum likelihood estimate (MLE) $\hat{\alpha}_t$ of α_t is a consistent estimate of the treatment effect size $\mu_{tP} = g[E_{X\in P}\{\mu_t(X)\}]$. The proof for this is standard and similar to that of Theorem 1 below, and is therefore omitted. However, in this paper, our goal is to provide a consistent estimate of $\mu_{tP*} = g[E_{X\in P*}\{\mu_t(X)\}]$. Our strategy is to "tilt" the population P so that it matches the population P^* and our matching tool is the propensity score.

In the likelihood (2) we reweight the contribution of the likelihood function from the ith subject from the historical trial with the weight $r(x)$, and form a new likelihood function (2*)

$$(2^*) \qquad \prod_{i=1}^{n_t}\{l_t(y_{it}, \alpha_t)\}^{r(x_i)}.$$

THEOREM 1. $\hat{\alpha}_t^*$, *the MLE which maximize* (2*), *is a consistent estimate of* $\mu_{tP*} = g[E_{X\in P*}\{\mu_t(X)\}]$. *In addition,* $\hat{\alpha}_t^* \sim N(\mu_{iP*}, A^{-1}(\mu_{iP*})B(\mu_{iP*}) \times A^{-1}(\mu_{iP*}))$, *where*

$$A(\alpha_t) = E\left\{r(x)\frac{d^2 \log l_t(y_{it}, \alpha_t)}{d\alpha_t^2}\right\},$$

$$B(\alpha_t) = E\left\{r^2(x)\frac{d \log l_t(y_{it}, \alpha_t)}{d\alpha_t}\frac{d \log_t(y_{it}, \alpha_t)}{d\alpha_t}\right\}.$$

The proof of Theorem 1 is given in the Appendix. Theorem 1 indicates that the calibrated treatment effect converges to the treatment effect that would be presented in population P^*. In other words, the likelihood function (2*) is reweighted in such a way that the units can be treated as randomly sampled from a target population, not the population of the study. Note that, if the two trials have the same population, then $r(x) = 1$, so that likelihood function (2*) reduces to (2).

Briefly, we note that the parametric approach easily extends to include some key covariates, including a treatment indicator as typically used in noninferiority trials, in the likelihood (2*) as follows:

$$\prod_{i=1}^{n} \{l(y_i, \alpha + \beta z)\}^{r(x_i)},$$

where $l(y_i, \alpha + \beta z)$ is the likelihood function contributed by the ith subject and z is a vector of treatment and/or covariate of interest. The MLE converges to the parameters in the target population P^*.

3.2. *Nonparametric approach.* Section 3.1 is based on the model assumption (1). In this subsection we take a nonparametric approach similar to the reweighting method of Zhang (2007) [see also Signorovitch et al. (2010) and Signorovitch et al. (2011)] for a historical control problem, and estimate $E_{X \in P^*}\{\mu_t(X)\}$ by $\hat{\delta}_t = \sum_{i=1}^{n_t} y_{it} r(x_i) / \sum_{i=1}^{n_t} r(x_i)$. When $n_t \to \infty$,

$$\frac{\sum_{i=1}^{n_t} y_{it} r(x_i)}{n_t} \to E_{X \in P}\left\{\mu_t(X) \frac{f^*(X)}{f(X)}\right\} = E_{X \in P^*}\{\mu_t(X)\} = \mu_{iP^*}.$$

Here we used the fact that $r(x) = f^*(x)/f(x)$, shown in the proof of Theorem 1 in the Appendix. Similarly, $\sum_{i=1}^{n_t} r(x_i)/n_t \to E_X\{f^*(x)/f(x)\} = 1$. Therefore, $\sum_{i=1}^{n_t} y_{it} r(x_i)/n_t \to E_{X \in P^*}\{\mu_t(X)\}$ and, thus,

$$\mu_{tP} = \nu[E_{X \in P}\{\mu_t(X)\}]$$

can be estimated by $\nu(\hat{\delta}_t) = \nu\{\sum_{i=1}^{n_t} y_{it} r(x_i) / \sum_{i=1}^{n_t} r(x_i)\}$. The variance of the estimator and therefore the confidence interval for the desired parameter can be obtained, for example, through the bootstrap method proposed in Efron (1981).

3.3. *Likelihood reweighting method vs. the previous covariate adjustment approach.* Aside from the differences between two approaches previously mentioned in the Introduction, we have the following observations on the likelihood reweighing method as an alternative to the covariate adjustment approach used in randomized clinical trails, along with the differences noted in the literature.

In the covariate adjustment approach, only the covariates interacting with treatment are considered influential and relevant to the adjustment. However, there are other types of "influential" covariates. One type of "influential" covariates relates to noncollapsibility, as illustrated in Table 1 from Greenland, Pearl and Robins

(1999). Assuming that Table 1 represents the population P and that X and Z represent the treatments and status of a disease, 50% of enrolled subjects have a certain disease and the other 50% of them do not have it. The event rates are 40% and 20% for treatment 1 ($X = 1$) and 2 ($X = 0$), respectively, in subjects with the disease and are 80% and 60% in subjects without the disease. The treatment 1 vs. treatment 2 odds ratio is 2.67 whether subjects have the disease or not, hence no treatment heterogeneity (i.e., no treatment by covariate interaction when measured in odds ratio). Consider a population P^*, in which 86% of enrolled subjects have the disease and the other 14% of them do not have it. The overall odds ratio in population P^* is thus 2.44. While the covariate adjustment approach would find that 2.67 is the odds ratio of treatment 1 vs. treatment 2 in P and P^*, the likelihood reweighting method would find that 2.25 and 2.44 are the odds ratio in P and P^*, respectively. In other words, the covariate adjustment approach estimates the conditional odds ratio and the likelihood reweighting method estimates the marginal odds ratio.

In this example, because the conditional odds ratio is not the same as the marginal odds ratio, the issue of noncollapsibility arises, leading to some questions on using the odds ratio as the measure of treatment effect. The likelihood reweighting method provides a flexible way to circumvent this difficulty. With the percentage difference as an alternative metric (measure of treatment effect), we could work with $\tau_{P*} = \mu_{1P*} - \mu_{2P*}$ and $\mu_{tP*} = \nu[E_{X \in P*}\{\mu_t(X)\}]$, where $\nu(\cdot)$ is the identity function.

The covariate adjustment approach relies on the ability of the data to detect treatment effect heterogeneity, that is, the treatment by covariate interaction. However, in some situations, trials may not be large enough to detect moderate interactions because they are not designed for that purpose. Even if some trials are large, the rarity of events could hamper the ability to detect all heterogeneity. Therefore, it is likely that some important interactions are not going to be detected, leading to partial adjustment with a residual bias. In these scenarios, the likelihood reweighting method can be a good alternative, as it estimates the marginal effect by maximizing the "unadjusted" likelihood (2*). Although modeling was also used to estimate propensity scores, the dependent variable is the trial indicator, not the outcome of interest.

The covariate adjustment approach may face co-linearity problems when there are many covariates. When that happens, the model-based adjustment requires a difficult decision on how to omit some covariates and select a good model. When this problem occurs, the likelihood reweighting method can be a good alternative, as it is less susceptible to this issue: while co-linearity may also cause problems with parameter estimation in the propensity score models, it does not adversely affect prediction of the propensity score itself.

The covariate adjustment approach utilizes the same outcome data for both model selection and formal inference based on the chosen model, and the results

could be too optimistic. In case this becomes a concern, the likelihood reweighting method can be considered as an alternative, as propensity score modeling and model selection do not involve outcome data. The techniques presented here are expected to be useful in uncontrolled observational studies as well. Although uncontrolled observational studies inherit many more issues than these controlled trials, the difference or change in the patient populations associated to different comparing groups remains one of the key issues.

As pointed out by the reviewers, a possible weakness of both approaches is that they require subject-level data from the historical control trial. Readers are referred to Nie and Soon (2010) for some discussions. One possible solution could be defining the weight based on summary statistics, an idea as illustrated in Signorovitch et al. (2010) and Signorovitch et al. (2011).

4. Applications. In noninferiority trials, we aim to calibrate the effect size of the active control (e.g., Palivizumab) relative to Placebo from the historical trial population P (e.g., IMPACT) to the noninferiority trial population P^* (e.g., MOTA). Using Bayes' rule, we obtain $r(x) \propto \Pr(P^*|X = x)/\Pr(P|X = x)$. As the population P^* is associated with the experimental treatment (e.g., Motavizumab) and the population P with Placebo, $r(x) \propto \Pr(T = \text{MOTA}|X = x)/\Pr(T = \text{Placebo}|X = x)$, that is, $r(x)$ is proportional to the odds through the propensity score.

4.1. *Development of Motavizumab, a second generation of Palivizumab.* Palivizumab is a humanized monoclonal antibody, approved and marketed for passive immunoprophylaxis of respiratory syncytial virus (RSV) in infants at risk for serious RSV disease. It was studied in the IMPACT trial, a phase III randomized, double-blind, Placebo-controlled clinical trial that was conducted to evaluate the ability of prophylaxis with Palivizumab to reduce respiratory syncytial virus infection in high-risk infants. A total of 1502 children with prematurity or bronchopulmonary dysplasia (BPD), also called chronic lung disease (CLD) in infancy, were randomized to receive either Palivizumab or Placebo intramuscularly. The primary endpoint was RSV related hospitalization within 150 days since administration of the first dose of treatment. For more information of this trial please refer to the Impact-RSV Study Group (1998). This trial enrolled subjects exclusively from two disjoint subgroups: (1) children 24 months old or younger with a clinical diagnosis of BPD requiring ongoing medical treatment; and (2) children with 35 weeks gestation or less and 6 months old or younger, who did not have a clinical diagnosis of BPD.

Among subjects enrolled with a diagnosis of BPD, the incidence rate of RSV-related hospitalization was 12.8% (34/266) in the Placebo arm and 7.9% (39/496) in the Palivizumab arm. Among subjects enrolled without a diagnosis of BPD, the incidence rate of RSV-related hospitalization was 8.1% (19/234) in the Placebo

TABLE 1
Subject distribution: numbers of subjects
(*numbers of RSV-hospitalizations*)

IMPACT trial	BPD	Non-BPD
Placebo	266 (34)	234 (19)
Palivizumab	496 (39)	506 (9)
MOTA trial		
Palivizumab	723 (28)	2607 (34)
Motavizumab	722 (22)	2583 (24)

arm and 1.8% (9/506) in the Palivizumab arm. Among the 500 subjects who received Placebo, 53 (10.6%) had an RSV-related hospitalization; among the 1002 subjects who received Palivizumab, 48 (4.8%) had an RSV-related hospitalization (see Table 1 for details). It is clear that the treatment effect of Palivizumab vs. Placebo was better in subjects enrolled without a diagnosis of BPD than in subjects enrolled with a diagnosis of BPD. The overall treatment effect size of Palivizumab, as measured in the odds ratio of the Placebo vs. Palivizumab, was 2.4 with a 95% C.I. of (1.6, 3.5).

To evaluate Motavizumab, a second generation version of Palivizumab, a phase 3, randomized, double-blind, Palivizumab-controlled, multi-center, multi-national noninferiority trial (MOTA) was conducted to assess whether Motavizumab was noninferior to Palivizumab. More precisely, the question was whether Motavizumab is at least not too much worse than Palivizumab in the sense that the difference of Motavizumab vs. Palivizumab is greater than the difference of Placebo vs. Palivizumab. With the risk difference metric this means that the rate difference of RSV hospitalization between Motavizumab and Palivizumab is smaller than the rate difference of RSV hospitalization between Placebo and Palivizumab. In the metric of odds ratio this means that the odds ratio between Motavizumab and Palivizumab is smaller than the odds ratio between Placebo and Palivizumab. In order to evaluate noninferiority, one possible test statistic is

$$\hat{\mu}_{ks} = \frac{\hat{\mu}_{TC} + \hat{\mu}_{CP}}{\sqrt{\sigma_{TC}^2 + \sigma_{CP}^2}},$$

where $\hat{\mu}_{TC}$ is the overall log-odds ratio of Palivizumab vs. Motavizumab and $\hat{\mu}_{CP}$ is the overall log-odds ratio of Palivizumab vs. Placebo.

4.2. *Calibrated effect size of Palivizumab vs. Placebo in the new MOTA study population.* Assume that Y_{ixt}, the incidence of RSV hospitalization of the ith subject, follows a logistic regression model, $y_{ixt} \sim$ Binomail$(1, p_{xt})$; logit$(p_{xt}) = \theta_{xt}$ with $x = 0, 1$, representing subjects enrolled without and with a diagnosis of BPD and $t = 0, 1$ representing Placebo and Palivizumab.

Whether to make inference on p_{xt} (the incidence rate) or θ_{xt} (the log odds of an event) is generally subjective. Both are used extensively in noninferiority trials. In IMPACT and MOTA, the log-odds ratio was the primary metric, but the risk difference is the primary metric in current HIV trials. Therefore, we shall illustrate both metrics in the Motavizumab example.

Let us first consider quantifying the treatment effect using the risk difference. Let p_n denote the proportion of subgroup with $x = 1$ in the target population (e.g., MOTA population) and p_h denote the proportion of subgroup with $x = 1$ in the historical population (e.g., IMPACT population). It is easy to show that the MLEs of likelihood in (2) and (2*) are

$$\hat{\alpha}_t = \frac{n_{1t}\bar{y}_{\cdot 1t} + n_{0t}\bar{y}_{\cdot 0t}}{n_{1t} + n_{0t}}; \qquad \hat{\alpha}_t^* = \frac{n_{1t}(p_n/p_h)\bar{y}_{\cdot 1t} + n_{1t}((1 - p_n)/(1 - p_h))\bar{y}_{\cdot 0t}}{n_{1t}(p_n/p_h) + n_{1t}(1 - p_n)/(1 - p_h)}.$$

In our example, these are

$$\hat{\alpha}_0 = \frac{266 \times 12.8\% + 234 \times 8.1\%}{500} = 10.6\%;$$

$$\hat{\alpha}_0^* = \left(266\frac{0.22}{266/500}12.8\% + 234\frac{0.78}{234/500}8.1\%\right)$$

$$\Big/\left(266\frac{0.22}{266/500} + 234\frac{0.78}{234/500}\right) = 9.1\%.$$

Similarly, $\hat{\alpha}_1 = 4.8\%$; $\hat{\alpha}_1^* = 3.1\%$. For the nonparametric approach presented in Section 3.2, the same results are obtained. Indeed, whether using the risk difference or the log-odds ratio as metrics, the parametric approach presented in Section 3.1 and the nonparametric approach presented in Section 3.2 lead to the same results for this example.

The standard error of the calibrated effect size $\hat{\alpha}_j^*$ can be calculated directly here as

$$\text{std}(\hat{\alpha}_0^*) = \sqrt{0.22^2 \times \frac{12.8\% \times (1 - 12.8\%)}{266} + 0.78^2 \times \frac{8.1\% \times (1 - 8.1\%)}{234}}$$

$$= 0.015.$$

Similarly, $\text{std}(\hat{\alpha}_1^*) = 0.005$. We could also use statistical software to obtain the standard error. In this paper we used the SAS procedure PROC GEMMOD with the generalized estimating equation (GEE) option to compute the standard error described in Theorem 1. The resulting standard errors for $\hat{\alpha}_0^*$ and $\hat{\alpha}_1^*$ are 0.015 and 0.005, which are the same as obtained in our direct computation.

Now, let us quantify the treatment effect using the log-odds ratio metric. The estimate can be obtained through PROC NLMIXED with the replicate statement. However, the standard error obtained from this procedure is not the standard error stated in Theorem 1. In order to obtain the correct standard errors, we could obtain

a bootstrap standard error [Efron (1981)], which is 0.25. Alternatively, we can also use PROC GEMMOD to obtain the same point estimate and the standard error. It results in the same estimate and standard error. Note that the unadjusted log-odds ratio is $\hat{\mu}_{CP} = 0.86$ with standard error 0.21.

The estimated log-odds ratio of Palivizumab vs. Motavizumab is 0.31 with a standard error of 0.20. Using the unadjusted or adjusted effect size, we calculate the unadjusted and adjusted statistics,

$$\hat{\mu} = \frac{\hat{\mu}_{TC} + \hat{\mu}_{CP}}{\sqrt{\sigma_{TC}^2 + \sigma_{CP}^2}} = \frac{0.31 + 0.86}{\sqrt{0.20^2 + 0.21^2}} = 4.0;$$

$$\hat{\mu}_{\text{adj}} = \frac{\hat{\mu}_{TC} + \hat{\mu}_{CP}}{\sqrt{\sigma_{TC}^2 + \sigma_{CP}^2}} = \frac{0.31 + 1.14}{\sqrt{0.20^2 + 0.25^2}} = 4.5.$$

The significance levels associated with the unadjusted and adjusted inference are 0.00003 and 0.000003, respectively. Other than this approach, one can also use a more conservative approach, the fixed margin approach [see FDA (2010) for details], to make the following inference:

$$\hat{\mu}_f = \frac{\hat{\mu}_{TC} + \hat{\mu}_{CP}}{\sigma_{TC} + \sigma_{CP}} = \frac{0.31 + 0.86}{0.20 + 0.21} = 2.9;$$

$$\hat{\mu}_{\text{adj},f} = \frac{\hat{\mu}_{TC} + \hat{\mu}_{CP}}{\sigma_{TC} + \sigma_{CP}} = \frac{0.31 + 1.14}{0.20 + 0.25} = 3.2.$$

The significance level associated with the unadjusted inference is 0.002 and it is 0.0006 for the adjusted inference. Although both are less than 0.05, the latter one is approximately 0.025^2, which means the significance level is as low as that of two independent clinical trials, each significant at a level of 0.025, fulfilling the regulatory requirement on the quantity of the evidence [see Soon et al. (2013)]. The quantity requirement has been interpreted in the FDA guidance of drug effectiveness [FDA (1998)] as follows: "With regard to quantity, it has been FDAs position that Congress generally intended to require at least two adequate and well-controlled studies, each convincing on its own, to establish effectiveness." Therefore, the adjusted approach could make a difference because Motavizumab was evaluated in a single noninferiority trial. However, we emphasize that this analysis only takes the published data into consideration and assumes that there are no other potential issues associated with the trial design and conduct.

4.3. *Calibrating the effect size using subject-level data.* Section 4.2 presented a simple example to illustrate calibration through likelihood reweighting. In general, subject-level data from a clinical trial is much more complex than the data presented in Table 1. Although the FDA typically has access to all subject-level data for regulatory purposes, we have no authority to use them for purposes other

than regulatory decision-making. Therefore, we cannot share our experiences analyzing real data with readers. For the purpose of illustrating methodology, we will use a simulated data set based on the IMPACT data set and the MOTA data set.

We randomly generate variables x_1, x_2, and x_3 so the generated data set, say, IMPACT_0 and MOTA_0, will have 5 variables: BPD status, treatment, x_1, x_2, and x_3. In IMPACT_0, we randomly generate a data set for 1502 subjects, with x_1, x_2, x_3 following three independent Bernoulli distributions with success rates of 0.4, 0.6, and 0.5. Similarly, in MOTA_0, we randomly generate another data set for 6635 subjects, with x_1, x_2, and x_3 following three independent Bernoulli distributions with success rates of 0.6, 0.5, and 0.4. We then pool the two data sets together and define a trial indicator to distinguish IMPACT_0 and MOTA_0.

To obtain the weight for each subject, we use logistic regression to model the logit of the trial probability as a linear function of BPD status, x_1, x_2, and x_3. Using the fitted model, we predict the probability of each subject in IMPACT_0 being located in MOTA_0. We then define the weight $r(x)$ by the odds of the predicted probability times $1502/6635$. Note here vector x means all variables: BPD status, treatment, x_1, x_2, and x_3. Now, the estimated propensity score ratio $r(x)$ is defined for all 1502 subjects.

The MLE of the reweighted likelihood (2*) can be obtained through implementation of SAS procedure PROC GEMMOD (see the attached programming code). With the repeat and weight statement [in the repeat statement, subject is the ID number and the weight is $r(x)$], the GEMMOD procedure provides the GEE [Zeger and Liang (1986)] type sandwich estimates for standard error, corresponding to the variance formula given in Theorem 1. All the programs, including the simulated data, are available upon request.

The point estimate of the adjusted log-odds ratio is 1.23 with a standard error of 0.28. The final inference of the noninferiority trial may use the following adjusted statistics:

$$\hat{\mu}_{\text{adj}} = \frac{\hat{\mu}_{TC} + \hat{\mu}_{CP}}{\sqrt{\sigma_{TC}^2 + \sigma_{CP}^2}} = \frac{0.31 + 1.23}{\sqrt{0.20^2 + 0.28^2}} = 4.5;$$

$$\hat{\mu}_{\text{adj}, f} = \frac{\hat{\mu}_{TC} + \hat{\mu}_{CP}}{\sigma_{TC} + \sigma_{CP}} = \frac{0.31 + 1.23}{0.20 + 0.28} = 3.2.$$

Although the problem did not occur in our example, we note that the weighted likelihood approach may result in an estimate with large variance if we have very small or very large values of $r(x)$. The problem has been described in the propensity score literature and is not unique to our setting. Some promising methods for dealing with extreme propensity score weights include using generalized boosted regression (GBR) [McCaffrey, Ridgeway and Morral (2004), Ridgeway and McCaffrey (2007) and Lee, Lessler and Stuart (2009)]. Based on our experience, we recommend that the propensity score model should be limited to effect modifiers,

that is, baseline variables that are associated with the treatment difference, echoing what was recommended by Cole and Hernán (2008). Including many variables that are not effect modifiers in the propensity score model generally increases the chance of extreme weight due to the larger population heterogeneity, with no clear benefits in reducing bias. The supplement [Nie et al. (2013)] provides more details and a simulation study to illustrate this recommendation. At the end of this section, we also describe some additional alternatives approaches. Before doing that, we would like to discuss some special issues when this problem occurs in historically controlled trials and noninferiority trials.

When the weights are extremely small, such as when some subjects with certain characteristics in the historical trial have no or few counterparts (subjects with the same characteristics) in the noninferiority trials, $r(x)$ is 0 or near 0 for these subjects. For example, this may happen when more stringent inclusion criteria are implemented so that subjects with less severe disease conditions at baseline were included in the historical trials but are excluded from the noninferiority trial. It is understandable these subjects (e.g., with less severe disease condition) may not always be used to make inferences about the control vs. placebo (i.e., $\hat{\mu}_{CP}$) in the noninferiority trials subjects (e.g., with more severe disease condition) unless we make an assumption that the treatment effect $\hat{\mu}_{CP}$ in these subjects dose not depend on baseline disease status. Only with this assumption, we may multiply their weight $r(x)$ using a large number so that these subjects still represent the subjects with different characteristics. Without this assumption, the method expectedly leads to a relatively larger variance because we discard a portion of information from the historical trial.

The weights can be extremely large, such as when a subpopulation presented in a noninferiority trial is not well represented in the historical trial. For example, some subjects in noninferiority trials of HIV may use a newly approved potent background drug that was rarely used or never used in the historical trials. In this case, using historical data from another group (subjects who did not have the new background drug) to make inference about the relative effect of the control vs. placebo (i.e., $\hat{\mu}_{CP}$) may not be prudent without additional assumptions. In this case, we might have to consider alternative approaches. One possibility is to restrict the proposed analysis to the subpopulation of the current study that is also represented in the historical study. This is essentially equivalent to what propensity score matching would achieve, where unmatched subjects are automatically excluded. Another possibility is to consider the hybrid design idea presented in Soon et al. (2011).

Now we briefly describe a stratified approach based on stratified propensity scores. Suppose a control treatment is evaluated in historical trials but we would like to calibrate its effect size in a new population for which $r(x)$ is computed. We group $r(x)$ into a number of strata $g_1, \ldots, g_L$. The percentage of subjects falling into g_l is w_{lh} and w_{ln} in population P and P^*, respectively. Let $\hat{\beta}_l$ with variances s_l^2 be the treatment effect size in group l. Then a combined calibrated treatment

effect in the population P^* is $\sum_{l=1}^{L} \hat{\beta}_l w_{\text{ln}}$ with variance $\sum_{l=1}^{L} s_l^2 w_{ln}^2$. When a new treatment is evaluated by its comparison to the control in population P^*, a stratified analysis based on the subclasses of the propensity score can be implemented as follows. Let $\hat{\gamma}_l$ be an estimate of treatment effect of the new treatment, with variances s_{ln}^2, in the lth group of the propensity score. We evaluate the new treatment through a stratified analysis such as

$$\frac{\sum_{l=1}^{L} \{\hat{\gamma}_l - \hat{\beta}_l\} w_{l*}}{\sum_{l=1}^{L} \{s_{\text{ln}}^2 + s_{lh}^2\} w_{l*}^2}.$$

Depending on the objective, w_{l*} may be chosen differently. Alternatively, we may also use the stratification method as described in Section 4.3 or define a threshold $\Delta_2 > \Delta_1 > 0$ so that subjects with $r(x)$ beyond $[\Delta_1, \Delta_2]$ should redefine the weight of $[\Delta_1, \Delta_2]$, whichever is closer, similar to the method used in Cole and Hernán (2008). The actual determination of $[\Delta_1, \Delta_2]$ depends on the actual data and practical assumptions.

5. Concluding remarks. Motivated by a real example, we show that bias can arise in active controlled noninferiority trials when estimated treatment effect size for the control treatment is obtained from a historical trial that has been conducted for a different population. Covariate adjustment approaches [Zhang (2009) and Nie and Soon (2010)] have been proposed to address the problem. However, they may be directly applied to obtain the marginal treatment effects, which are often the pre-specified primary endpoints. This paper proposes a likelihood reweighting method through propensity scoring to estimate the marginal treatment effect size in the target population of a noninferiority trial based on data obtained from a historical trial that has been conducted for a different population.

APPENDIX

PROOF OF THEOREM 1. It is easy to verify that the generalized linear model satisfies Assumptions A1–A3 in White (1982), therefore, Theorem 2.2 is applicable. The log likelihood of (2*) is

$$\sum_{i=1}^{n} \frac{f^*(x_i)}{f(x_i)} \log l_t(y_{it}, \alpha_t).$$

According to Theorem 2.2 [White (1982)], the maximum likelihood estimate $\hat{\alpha}_t$ converges to the parameter, say, α_{t0}, which maximizes

$$\int E_{Y|X=x, T=t} \{\log f(x) + \log l(y_{it}|x, t) - r(x) \log l(y_{it}, \alpha_t)\} \, dF(x),$$

where $\log l(y_{it}|x, t)$ is the log-likelihood function obtained from model (1). Taking derivatives with respect to α_t, we know α_{t0} is the solution of the following equation:

$$\int E_{Y|X=x, T=t}\big[r(x)\{y - b'(\alpha_t)\}\big] dF(x) = 0.$$

Consequently, the estimating equation can be written as

$$\int E_{Y|X=x, T=t}\{y - b'(\alpha_t)\} dF^*(x) = 0.$$

Noting $b'(\cdot) = g^{-1}(\cdot)$, we have $\alpha_{t0} = \mu_{iP^*} = g[E_{X \in P^*}\{\mu_t(X)\}]$. As the assumptions A4–A6 are easily verifiable, the asymptotic properties of the MLE of (2*) are immediately obtained from Theorem 3.2 in White (1982). □

Acknowledgments. We would like to express our great appreciation to the three reviewers, an Associate Editor, and Dr. Paddock for all of their comments and suggestions, which substantially improved the quality of the paper.

SUPPLEMENTARY MATERIAL

Supplement to "Likelihood reweighting methods to reduce potential bias in noninferiority trials which rely on historical data to make inference" (DOI: 10.1214/13-AOAS655SUPP; .pdf). The supplement provides an assessment of the efficiency loss for the weighted likelihood method and a comparison between the likelihood reweighting method and related methods in historically controlled trials.

REFERENCES

CARBONELL-ESTRANY, X., SIMÕES, E. A. F., DAGAN, R. et al. (2010). Motavizumab for prophylaxis of respiratory syncytial virus in high-risk children: A noninferiority trial. *Pediatrics* **125** e35–e51.

COLE, S. R. and HERNÁN, M. A. (2008). Constructing inverse probability weights for marginal structural models. *Am. J. Epidemiol.* **168** 656–664.

COLE, S. R. and STUART, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *Am. J. Epidemiol.* **172** 107–115.

COOPER, D. A., STEIGBIGEL, R. T., GATELL, J. M. et al. (2008). Subgroup and resistance analyses of raltegravir for resistant HIV-1 infection. *N. Engl. J. Med.* **359** 355–365.

EFRON, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika* **68** 589–599. MR0637776

FDA (1998). Providing clinical evidence of effectiveness for human drug and biological products. FDA, 1998.

FDA (2010). Draft guidance for industry: Non-inferiority clinical trials. FDA, 2010.

FRANGAKIS, C. (2009). The calibration of treatment effects from clinical trials to target populations. *Clin Trials* **6** 136–140.

FRIEDMAN, L. M., FURBERG, C. D. and DEMETS, D. L. (1998). *Fundamentals of Clinical Trials*, 3rd ed. Springer, New York.

GREENHOUSE, J. B., KAIZAR, E. E., KELLEHER, K., SELTMAN, H. and GARDNER, W. (2008). Generalizing from clinical trial data: A case study. The risk of suicidality among pediatric antidepressant users. *Stat. Med.* **27** 1801–1813. MR2420346

GREENLAND, S., PEARL, J. and ROBINS, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology* **10** 37–48.

IMPACT-RSV STUDY GROUP (1998). Palivizumab, a humanized respiratory syncytial virus monoclonal antibody, reduces hospitalization from respiratory syncytial virus infection in high-risk infants. The IMpact-RSV Study Group. *Pediatrics* **102** 531–537.

LEE, B. K., LESSLER, J. T. and STUART, E. A. (2009). Using weight trimming to improve propensity score weighting. *Am. J. Epidemiol.* **169** S90–S90.

MCCAFFREY, D. F., RIDGEWAY, G. and MORRAL, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol Methods* **9** 403–425.

MOLINA, J.-M., LAMARCA, A., ANDRADE-VILLANUEVA, J. et al. (2012). Efficacy and safety of once daily elvitegravir versus twice daily raltegravir in treatment-experienced patients with HIV-1 receiving a ritonavir-boosted protease inhibitor: Randomised, double-blind, phase 3, non-inferiority study. *Lancet. Infect. Dis.* **12** 27–35.

NIE, L. and SOON, G. (2010). A covariate-adjustment regression model approach to noninferiority margin definition. *Stat. Med.* **29** 1107–1113. MR2756815

NIE, L., ZHANG, Z., RUBIN, D. and CHU, J. (2013). Supplement to "Likelihood reweighting methods to reduce potential bias in noninferiority trials which rely on historical data to make inference." DOI:10.1214/13-AOAS655SUPP.

RIDGEWAY, G. and MCCAFFREY, D. F. (2007). Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22** 540–543. MR2420459

ROBINS, J. M., HERNÁN, M. A. and BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11** 550–560.

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974

ROTHMAN, K. J. and MICHELS, K. B. (1994). The continuing unethical use of placebo controls. *N. Engl. J. Med.* **331** 394–398.

SIGNOROVITCH, J. E., WU, E. Q., YU, A. P. et al. (2010). Comparative effectiveness without head-to-head trials: A method for matching-adjusted indirect comparisons applied to psoriasis treatment with adalimumab or etanercept. *Pharmacoeconomics* **28** 935–945.

SIGNOROVITCH, J. E., WU, E. Q., BETTS, K. A. et al. (2011). Comparative efficacy of nilotinib and dasatinib in newly diagnosed chronic myeloid leukemia: A matching-adjusted indirect comparison of randomized trials. *Curr. Med. Res. Opin.* **27** 1263–1271.

SOON, G. G., NIE, L., HAMMERSTROM, T., ZENG, W. and CHU, H. (2011). Meeting the demand for more sophisticated study designs. A proposal for a new type of clinical trial: The hybrid design. *BMJ Open* **1** e000156.

SOON, G., ZHANG, Z., TSONG, Y. and NIE, L. (2013). Assessing overall evidence from noninferiority trials with shared historical data. *Stat. Med.* **32** 2349–2363.

WEISBERG, H. I., HAYDEN, V. C. and PONTES, V. P. (2009). Selection criteria and generalizability within the counterfactual framework: Explaining the paradox of antidepressant-induced suicidality? *Clin Trials* **6** 109–118.

WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50** 1–25. MR0640163

ZEGER, S. L. and LIANG, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42** 121–130.

ZHANG, Z. (2007). Estimating the current treatment effect with historical control data. *JP J. Biostat.* **1** 217–247. MR2522284

ZHANG, Z. (2009). Covariate-adjusted putative placebo analysis in active-controlled clinical trials. *Statistics in Biopharmaceutical Research* **1** 279–290.

L. NIE
D. RUBIN
DIVISION OF BIOMETRICS IV
OFFICE OF BIOSTATISTICS/CDER/FDA
10903 NEW HAMPSHIRE AVENUE
SILVER SPRING, MARYLAND 20993
USA

Z. ZHANG
J. CHU
DIVISION OF BIOSTATISTICS
OFFICE OF SURVEILLANCE
  AND BIOMETRICS/CDRH/FDA
10903 NEW HAMPSHIRE AVENUE
SILVER SPRING, MARYLAND 20993
USA
E-MAIL: zhiwei.zhang@fda.hhs.gov