Published in final edited form as:

J Biopharm Stat. 2014 ; 24(1): 154–167. doi:10.1080/10543406.2013.856023.

STANDARDIZATION FOR SUBGROUP ANALYSIS IN RANDOMIZED CONTROLLED TRIALS

Ravi Varadhan^{1,2} and Sue-Jane Wang^{3,4}

¹Center on Aging and Health, Division of Geriatric Medicine and Gerontology, School of Medicine, Johns Hopkins University, Baltimore, Maryland, USA

²Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA

³Office of Biostatistics, Office of Translational Service, Center for Drug Evaluation and Research, US Food and Drug Administration, Washington, DC, USA

⁴Engineering and Applied Science Programs for Professionals, Johns Hopkins University, Baltimore, Maryland, USA

Abstract

Randomized controlled trials (RCTs) emphasize the average or overall effect of a treatment (ATE) on the primary endpoint. Even though the ATE provides the best summary of treatment efficacy, it is of critical importance to know whether the treatment is similarly efficacious in important, predefined subgroups. This is why the RCTs, in addition to the ATE, also present the results of subgroup analysis for preestablished subgroups. Typically, these are marginal subgroup analysis in the sense that treatment effects are estimated in mutually exclusive subgroups defined by only one baseline characteristic at a time (e.g., men versus women, young versus old). Forest plot is a popular graphical approach for displaying the results of subgroup analysis. These plots were originally used in meta-analysis for displaying the treatment effects from independent studies. Treatment effect estimates of different marginal subgroups are, however, not independent. Correlation between the subgrouping variables should be addressed for proper interpretation of forest plots, especially in large effectiveness trials where one of the goals is to address concerns about the generalizability of findings to various populations. Failure to account for the correlation between the subgrouping variables can result in misleading (confounded) interpretations of subgroup effects. Here we present an approach called standardization, a commonly used technique

Copyright © Taylor & Francis Group, LLC

Address correspondence to Ravi Varadhan, Suite 2-700, 2024 E. Monument Street, Baltimore, MD 21205, USA; rvaradhan@jhmi.edu.

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at <http://www.tandfonline.com/page/terms-and-conditions>

in epidemiology, that allows for valid comparison of subgroup effects depicted in a forest plot. We present simulations results and a subgroup analysis from parallel-group, placebo-controlled randomized trials of antibiotics for acute otitis media.

Keywords

Confounding; Consistency of treatment effect; Forest plot; Heterogeneity of treatment effects; Interaction; Marginal structural model

1. INTRODUCTION

Randomized controlled trials (RCTs) emphasize the overall or average effect of a treatment (ATE) on the primary endpoint. Even though the ATE provides the best summary of treatment efficacy, it is of critical importance to know whether the treatment is similarly efficacious in important, predefined subgroups. For example, if a treatment is efficacious on average, we may want to know whether it has a similar degree of efficacy among men and women, or young and old. The Food and Drug Administration (FDA) Guidance for Industry (FDA, 1998) states:

Frequently, large trials have relatively broad entry criteria and the study populations may be diverse with regard to important covariates such as concomitant or prior therapy, disease stage, age, gender or race. Analysis of the results of such trials for *consistency across key patient subsets* addresses concerns about generalizability of findings to various populations in a manner that may not be possible with smaller trials or trials with more narrow entry criteria.

The FDA guidance does not define what is meant by consistency of findings in key patient subsets. We provide a working definition of *consistency of treatment effect* as the absence of any statistically significant interaction between the treatment and a subgrouping variable using an appropriate model for interaction (at a significance level $\alpha = .05/K$, where K is the number of preestablished subgrouping variables considered in the analysis). In order to examine consistency of treatment effects across key patient subgroups, the RCTs, in addition to the ATE, also present the results of subgroup analysis for pre-established, baseline subgrouping factors. Typically, these are *marginal* subgroup analysis in the sense that treatment effects are estimated in subgroups defined by only one baseline characteristic at a time (e.g., men versus women, young versus old). This should be contrasted with a *joint* subgroup analysis where treatment effects would be estimated in mutually exclusive subgroups defined by multiple baseline characteristics (e.g., young–men, young–women, old–men, old–women). Forest plot is a popular graphical approach for displaying the results of marginal subgroup analysis in clinical trials. These plots were originally used in meta-analysis for displaying the treatment effects from independent studies. Treatment effect estimates of different subgroups from marginal analysis are, however, not independent. Correlation between the subgrouping variables should be addressed for proper interpretation of forest plots. This is especially important in large effectiveness trials where one of the goals of subgroup analyses is to address concerns about the generalizability of findings to various populations. Failure to account for the correlation between the subgrouping

variables can result in misleading interpretations of subgroup-specific treatment effects, and consequently treatment may be targeted to the wrong subgroups. For example, suppose that gender and age are correlated such that men are likely to be younger than women in the trial. Suppose the truth is that the effectiveness of the treatment declines with age independently of gender. A naive comparison of treatment effect in men and women would reveal that the treatment is only effective in men, which could result in wrongfully withholding the treatment from women. We should compare the effects of the treatment in men and in women in such a manner that the distribution of age is the same in men and women. This naturally leads to a consideration of the technique of standardization routinely employed by epidemiologists to control for confounding (Sato and Matsuyama, 2003). Epidemiologists calculate, for example, quantities such as standardized mortality/morbidity ratio by stratifying on a set of confounding factors. This technique is also known as inverse probability weighting (Robins et al., 2000). While the concept of standardization (or inverse probability weighting) is well known in epidemiology, our proposal for using it in the context of subgroup analysis in randomized controlled trials is novel.

2. MARGINAL SUBGROUP EFFECTS

Consider a randomized experiment where a binary treatment T is randomly assigned to subjects. Let Y be the outcome and let $\mathbf{X} = \{X_1, \dots, X_K\}$ be a vector of categorical baseline covariates. Our interest lies in understanding how the effect of T on Y varies according to baseline characteristics. We can define the treatment effect within any given stratum $\mathbf{x} = \{x_1, \dots, x_K\}$ as follows:

$$\theta(\mathbf{x}) = g(E[Y|T=1, \mathbf{X}=\mathbf{x}]) - g(E[Y|T=0, \mathbf{X}=\mathbf{x}]), \quad (1)$$

where $g(\cdot)$ is a function denoting the scale in which the treatment effect is quantified. Commonly used treatment effect scales are identity ($g(x) = x$), log ($g(x) = \log(x)$), and logit ($g(x) = \log(x) - \log(1 - x)$). The identity scale is used when the outcome is continuous. For binary outcomes, all three scales may be used with the identity scale denoting risk difference, log scale denoting the risk ratio, and the logit scale denoting the odds ratio. The log scale is also used when the outcome is the number of occurrences (counts) or the time duration for an event to occur.

Suppose, for example, that we have two binary baseline covariates: sex (men/women) and age (young/old) defined with men ($X_1 = 1$), women ($X_1 = 2$), young ($X_2 = 1$), and old ($X_2 = 2$). There will be four treatment effects, possibly different, defined in the four strata: young men, young women, old men, and old women. The number of strata can become large when there are several baseline covariates. Treatment effect estimates in such high-dimensional problems will tend to be highly variable due to small sample size. Therefore, it is more common to focus on treatment effects in low-dimensional subgroups defined by fewer baseline covariates, the simplest of which is subgroups defined by a single baseline characteristic, for example, men ($X_1 = 1$) and women ($X_1 = 2$), or young ($X_2 = 1$) and old ($X_2 = 2$). We call these “marginal” subgroup-specific treatment effects, which are defined for the subgroups defined by the k th baseline variable:

$$\eta(x_k) = g(E[Y|T=1, X_k=x_k]) - g(E[Y|T=0, X_k=x_k]), \quad (2)$$

Using the fact that

$$f(y|t, x_k) = \int f(y|t, x_k, \mathbf{x}_{-k}) f(\mathbf{x}_{-k}|t, x_k) d\mathbf{x}_{-k},$$

and that T is independent of $\mathbf{X}$, we obtain:

$$E[Y|T=t, X_k=x_k] = \int_{\mathbf{x}_{-k}} \int_y y f(y|t, x_k, \mathbf{x}_{-k}) f(\mathbf{x}_{-k}|x_k) dy d\mathbf{x}_{-k} = \int_{\mathbf{x}_{-k}} E[Y|T=t, X_k=x_k, \mathbf{X}_{-k}=\mathbf{x}_{-k}] dF(\mathbf{x}_{-k}|x_k). \quad (3)$$

Therefore,

$$E[Y|T=1, x_k] - E[Y|T=0, x_k] = \int_{\mathbf{x}_{-k}} \{E[Y|T=1, x_k, \mathbf{x}_{-k}] - E[Y|T=0, x_k, \mathbf{x}_{-k}]\} dF(\mathbf{x}_{-k}|x_k). \quad (4)$$

Thus, for linear models we can obtain the marginal effects for the k -th subgrouping variable X_k , equation (2), by marginalizing the fully stratified treatment effect $\theta(\mathbf{x})$ over the conditional distribution of the other baseline covariates, $f(\mathbf{X}_{-k} | X_k = x_k)$:

$$\eta(x_k) = \int_{\mathbf{x}_{-k}} \theta(X_k = x_k, \mathbf{X}_{-k}) f(\mathbf{X}_{-k} | X_k = x_k) d\mathbf{X}_{-k}, \quad (5)$$

Equation (4) shows that when $g(x) = x$ in equation (2), the treatment effect is *collapsible* in the sense that the marginal effects can be obtained by integrating the fully stratified effects. Such collapsibility is not generally present for other, nonlinear link functions:

$$g(E[Y|T=1, x_k]) - g(E[Y|T=0, x_k]) \neq \int_{\mathbf{x}_{-k}} \{g(E[Y|T=1, x_k, \mathbf{x}_{-k}]) - g(E[Y|T=0, x_k, \mathbf{x}_{-k}])\} dF(\mathbf{x}_{-k}|x_k) \quad (6)$$

Therefore, when collapsibility is not present, we cannot exactly represent the marginal subgroup effects by integrating the fully stratified subgroup effects over appropriate conditional distributions. Noncollapsibility notwithstanding, such a representation is reasonably accurate even for subgroups of moderately large size.

The theoretical development presented is applicable to both continuous and categorical covariates $\mathbf{X}$. For categorical variables, all the integrals in the preceding equations reduce to weighted summations across strata given by $\{\mathbf{X}_{-k}, X_k = x_k\}$, where the weights are the conditional probabilities $P(\mathbf{X}_{-k} = \mathbf{x}_{-k} | X_k = x_k)$.

3. STANDARDIZED MARGINAL INTERACTION MODELS

The marginal effect in equation (2) can be estimated nonparametrically by replacing the expectations with sample means of response Y in the appropriate subgroups. For binary (or categorical) subgrouping variables, the estimand in equation (2) can, equivalently, be estimated from the following saturated regression model:

$$g(E[Y|T, X_k]) = \beta_{0,k} + \beta_{t,k}T + \beta_k X_k + \gamma_k T X_k, \forall k=1, 2, \dots, K. \quad (7)$$

We call this the *marginal interaction model* (MIM), since it only pertains to a single covariate X_k . When X_k is binary, this model is equivalent to a stratified subgroup analysis for $X_k = 0$ and $X_k = 1$, since it is a saturated model in T and X_k . The treatment effects in the subgroups $X_k = 0$ and $X_k = 1$ denoted by $\beta_{t,k}$ and $\beta_{t,k} + \gamma_k$, respectively, are the same as $\eta(X_k = 0)$ and $\eta(X_k = 1)$, respectively, from equation (2). An advantage of the marginal interaction model over marginal subgroup analysis is that we can directly assess the *consistency* of subgroup treatment effects, that is, we can test the presence of an interaction between T and X_k , since it is equivalent to testing whether the treatment effects differ between $X_k = 0$ and $X_k = 1$. We denote the subgroup-specific treatment effects estimated using this model as *naive* subgroup-specific treatment effects.

When X_k is not independent of $\mathbf{X}_{-k}$, the inferences from the marginal model, equation (7), regarding consistency of subgroup-specific treatment effects can be misleading. Suppose, for example, that X_1 and X_2 are positively correlated in the sample, but that only X_1 is a true predictor of treatment response. The marginal subgroup analysis based on equation (7) could lead us to conclude that both X_1 and X_2 are predictors of treatment response. When X_1 and X_2 are negatively correlated and are both treatment effect modifiers, we could conclude from the marginal subgroup analysis that neither modifies the effect of the treatment. In other words, the interaction between treatment and X_2 is confounded by X_1 . Conversely, the interaction between treatment and X_1 can also be confounded by X_2 . Groenwold (2009) discusses an example of this phenomenon. We need to account for the correlation between baseline subgrouping variables $\mathbf{X}$ in order to draw proper inferences.

This leads to the consideration of the following marginal treatment effect that is not affected by correlation between X_k and $\mathbf{X}_{-k}$:

$$\eta^\dagger(x_k) = \int_{\mathbf{x}_{-k}} \theta(X_k = x_k, \mathbf{X}_{-k}) f(\mathbf{X}_{-k}) d\mathbf{X}_{-k}, \quad (8)$$

where $f(\mathbf{X}_{-k})$ is the joint distribution of $K - 1$ baseline covariates excluding X_k . Using the fact that $f(A) = f(A | B)f(B)/f(B | A)$, we can rewrite the preceding integral as:

$$\eta^\dagger(x_k) = \int_{\mathbf{x}_{-k}} \theta(X_k = x_k, \mathbf{X}_{-k}) f(\mathbf{X}_{-k} | X_k = x_k) \frac{f(X_k = x_k)}{f(X_k = x_k | \mathbf{X}_{-k})} d\mathbf{X}_{-k}, \quad (9)$$

Now, a comparison of equation (5) with equation (9) suggests that we can recover the

unconfounded effects from the naive effects by using the weights, $\frac{f(X_k = x_k)}{f(X_k = x_k | \mathbf{X}_{-k})}$. This is the basis for our standardization approach to removing the confounding in the naive marginal treatment effects. When $\mathbf{X}_{-k}$ is independent of X_k we have $\eta^\dagger(x_k) = \eta(x_k)$, and the weights equal 1. The obvious and natural choice for standardizing distribution, $f(\mathbf{X}_{-k})$, is the actual sample distribution of baseline covariates in the trial.

We illustrate standardization with a simple example. Consider a 2×2 table of men/women versus young/old; see Table 1. Men are six times as likely to be young than women; hence, a naive comparison of treatment effect between men and women is likely to be misleading if

age is a predictor of treatment response. The probability of being young among men and women is $\frac{2}{3}$ and $\frac{1}{9}$, respectively, whereas it is $\frac{1}{3}$ in the overall sample.

We can *standardize* the age distributions of men and women to match that of the overall sample, by weighting them differently. We assign a weight of $\frac{1}{2}$ to each young man, which is equal to $\frac{\Pr(\text{Men})}{\Pr(\text{Men}|\text{Young})}$, and a weight of 2 to each old man, which is equal to $\frac{\Pr(\text{Men})}{\Pr(\text{Men}|\text{Old})}$. Similarly, we also assign a weight of 3 to each young woman and a weight of $\frac{3}{4}$ to each old woman. After applying the standardization weights, we get the results shown in Table 2.

In the standardized Table 2, the age distributions in men and women are identical; the gender distributions among young and old are identical, as well.

Standardization is readily extended to multiple binary variables. Suppose that we have K binary subgrouping variables: $\mathbf{X} = \{X_1, X_2, \dots, X_K\}$. To standardize X_1 such that the joint distribution of $\{X_2, \dots, X_K\}$ is the same in $X_1 = 0$ and $X_1 = 1$, we apply the following weights:

$$w_1 = \frac{P(X_1=x_1)}{P(X_1=x_1|X_2=x_2, \dots, X_K=x_K)}, \quad (10)$$

where $x_k \in \{0, 1\}$, $\forall k = 1, 2, \dots, K$. When K is even moderately large, the denominator of equation (10) can become very small, and consequently, these nonparametric weights can become extremely large. The resulting standardized subgroup-specific treatment effects would become unstable with a large variance. Therefore, it is preferable to estimate the weights using parametric models. Let the observed data be $\{X_{1i}, X_{2i}, \dots, X_{Ki}, T_i, Y_i\}$, $i = 1, \dots, n$, where T is the binary treatment indicator and Y is the outcome. To obtain the weights for standardizing the subgroups $X_1 = 0$ and $X_1 = 1$, we fit the logistic regression model:

$$\text{logit}[P(X_1=1)|X_2, \dots, X_K] = \alpha_0 + \sum_{k=2}^K \alpha_k X_k. \quad (11)$$

This is about the simplest model for calculating the weights. It is also possible to fit more complex models, such as, for example, adding pairwise or higher order interactions in equation (11). It is generally not advisable to fit more than pairwise interactions due to potential instability in weights. Weights for the individuals are calculated by plugging in the model predicted probabilities from equation (11) into the denominator of equation (10). Standardization can also be readily applied to categorical factors with more than two levels. Multinomial models would be used instead of binomial models for calculating standardization weights.

When we incorporate weights, w_{1i} , for individuals in equation (7), we obtain the *standardized marginal interaction model* (SMIM):

$$g^{(w_k)}(E[Y|T, X_k]) = \beta_{0,k}^\dagger + \beta_{t,k}^\dagger T + \beta_k^\dagger X_k + \gamma_k^\dagger T \cdot X_k, \forall k=1, 2, \dots, K, \quad (12)$$

where the standardization weights, w_k , are given by:

$$w_k = \frac{P(X_k = x_k)}{P(X_k = x_k | \mathbf{X}_{-k} = \mathbf{x}_{-k})}, \forall k=1, 2, \dots, K. \quad (13)$$

In this model, the treatment effects in the standardized subgroups $X_k = 0$ and $X_k = 1$ are denoted by $\beta_{t,k}^\dagger$ and $\beta_{t,k}^\dagger + \gamma_k^\dagger$ respectively. The variances for the parameter estimates can be computed using Huber–White type estimation. A Wald test of interaction can be performed by testing the significance of the interaction coefficient estimate $\hat{\gamma}_1^\dagger$ with the Huber–White standard error. It should be noted that when X_k is independent of $\mathbf{X}_{-k}$, the weights $w_{k,i}$ equal 1, and the SMIM model reduces to the naive MIM model, equation (7).

4. SIMULATION EXPERIMENTS

The goal of the simulation experiments was twofold: (1) to evaluate the bias, standard error, and mean-squared error of the two types of estimation procedures for the subgroup-specific treatment effects: naive or usual subgroup-specific treatment effect estimated using the model, equation (3) and the standardized subgroup-specific treatment effects estimated using equations (2) and (4); and (2) to compare the type I error of naive and standardization methods for hypothesis testing when there is no interaction between the treatment and the subgrouping variable. The outcome Y was assumed to be normally distributed with mean given by the following model:

$$E[Y|T, \mathbf{X}] = \beta_0 + \beta_t T + \sum_{k=1}^K \beta_k X_k + \sum_{k=1}^K \gamma_k T X_k, \forall k=1, 2, \dots, K, \quad (14)$$

and with $\sigma^2 = 1$.

Under this data generating model, we can obtain the true stratified (joint) and marginal subgroup-specific treatment effects, respectively, as follows:

$$\theta(\mathbf{X} = \mathbf{x}) = \beta_t + \sum_{i=1}^K \gamma_i x_i, \quad (15)$$

and

$$\eta^\dagger(X_k = x) = \beta_t + x \gamma_k + \sum_{i \neq k}^K \gamma_i P(X_i = 1), x = \{0, 1\}; k=1, 2, \dots, K. \quad (16)$$

The naive marginal subgroup-specific treatment effects for each scenario can also be calculated:

$$\eta(X_k=x) = \beta_t + x\gamma_k + \sum_{i \neq k}^K \gamma_i P(X_i=1|X_k=x), x=\{0, 1\}; k=1, 2, \dots, K. \quad (17)$$

We considered three binary subgrouping variables, X_1 , X_2 , and X_3 , which are correlated. The

correlation matrix is:
$$\begin{pmatrix} 1 & -0.4 & -0.4 \\ -0.4 & 1 & 0.5 \\ -0.4 & 0.5 & 1 \end{pmatrix}.$$

The marginal probabilities were: $P(X_1 = 1) = 0.4$; $P(X_2 = 1) = 0.5$; $P(X_3 = 1) = 0.6$. Let us denote the parameter vector as: $\theta = (\beta_0, \beta_t, \beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2, \gamma_3)$. In all the simulations, we consider $\theta = (-1.0, 0.7, -0.5, -0.5, -0.5, \gamma_1, \gamma_2, \gamma_3)$, where we only varied one or more of the interaction parameters, $\gamma_1, \gamma_2, \gamma_3$ between different scenarios. We consider four different scenarios: (1) The treatment effect is homogeneous, (2) only X_1 is an effect modifier, (3) X_2 and X_3 are positive effect modifiers, and (4) X_2 interaction is positive and X_3 interaction is negative. The parameter vectors for these 4 scenarios are:

Scenario 1 $\theta = (-1.0, 0.7, -0.5, -0.5, -0.5, 0.0, 0.0, 0.0)$.

Scenario 2 $\theta = (-1.0, 0.7, -0.5, -0.5, -0.5, 0.5, 0.0, 0.0)$.

Scenario 3 $\theta = (-1.0, 0.7, -0.5, -0.5, -0.5, 0.0, 0.5, 0.5)$.

Scenario 4 $\theta = (-1.0, 0.7, -0.5, -0.5, -0.5, 0.0, 0.5, -0.5)$.

We considered two different sample sizes, $N = 200$ and $N = 2500$, representing small efficacy trials and large effectiveness trials, respectively. We conducted 100,000 simulations for $N = 200$ and 10,000 simulations for $N = 2,500$. Table 3 presents the simulation results for small sample size and Table 4 presents the results for large sample size.

Scenario 1

In this scenario, the treatment effect was constant across all subgroups. Both the naive and the standardized approaches yield unbiased estimates, which is 0.7 (from equations (16) and (17)), for small (Table 3) and large (Table 4) sample sizes. Standardization, while eliminating the bias, increases the variance of the estimate. Thus, the root mean squared

error (RMSE), defined as $\sqrt{\text{bias}^2 + \text{variance}}$, is larger for standardization. The rejection rates for the naive and standardization are consistent with a two-sided nominal level of 0.05.

Scenario 2

In this scenario, the treatment effect varied only across subgroups defined by X_1 . The marginal subgroup-specific treatment effects calculated using equation (16) for $X_1 = 0, X_1 = 1, X_2 = 0, X_2 = 1, X_3 = 0, X_3 = 1$ are, respectively, 0.7, 1.2, 0.9, 0.9, 0.9, 0.9. The naive estimates of the subgroup-specific treatment effects are unbiased only for $X_1 = 0$ and $X_1 = 1$ when X_1 is the only effect modifier. The subgroup-specific treatment effect estimates for the noneffect modifier, X_2 and X_3 , are biased, with an upward bias for $X_2 = 0$ and $X_3 = 0$, and a downward bias for $X_2 = 1$ and $X_3 = 1$. This results in a negatively biased estimate of the null

interaction between T and X_2 and X_3 . As expected, the RMSE is larger for standardization compared to naive approach in small samples, but it is smaller in large samples.

Also, the type I error probabilities of the naive approach for testing the null interaction between T and X_2 and X_3 are greater than the nominal rate of .05. The type I errors are 0.102 and 0.100 for the small sample size (Table 3) and 0.664 and 0.657 for large sample size (Table 4). Standardization has a type I error slightly larger than the nominal rate for small sample size.

Scenario 3

Here the treatment effect varied (in the same direction) across subgroups defined by X_2 and X_3 . The marginal subgroup-specific treatment effects calculated using equation (16) for $X_1 = 0$, $X_1 = 1$, $X_2 = 0$, $X_2 = 1$, $X_3 = 0$, and $X_3 = 1$ are, respectively, 1.25, 1.25, 1.00, 1.50, 0.95, and 1.45. The naive approach has a substantial bias for all subgroups. It has an upward bias for $X_1 = 0$, $X_2 = 1$, and $X_3 = 1$, and a downward bias for $X_1 = 1$, $X_2 = 0$, and $X_3 = 0$. Therefore, the magnitude of all three interaction effects is overestimated. The RMSE of the naive approach is much larger than that of the standardization approach for large samples, but is smaller for small samples. The type I error probabilities of the naive approach for testing the null interaction between T and X_1 were much larger than the nominal rate in both small sample (0.26) and large sample (0.997) simulations.

Scenario 4

As in Scenario 3, the treatment effect varied across subgroups defined by X_2 and X_3 , with the only difference being that the treatment effects varied in opposite directions. The marginal subgroup-specific treatment effects calculated using equation (16) for $X_1 = 0$, $X_1 = 1$, $X_2 = 0$, $X_2 = 1$, $X_3 = 0$, and $X_3 = 1$ are, respectively, 0.65, 0.65, 0.40, 0.90, 0.95, and 0.45. Interestingly, the naive approach is unbiased for X_1 subgroups, where it was biased in Scenario 3. This is due to the fact that the interactions between T and X_2 and T and X_3 have same magnitude and opposite signs, and X_2 and X_3 are negatively correlated. This has a counteracting effect. Also, with the naive approach there is an upward bias for $X_2 = 0$, $X_3 = 1$ and a downward bias for $X_2 = 1$, $X_3 = 0$. Therefore, the magnitude of the interaction effects $T * X_2$ and $T * X_3$ are underestimated. Once again, as expected, the RMSE is smaller for standardization compared to naive approach in large samples, but not always in small samples. The type I error probabilities for testing the null interaction between T and X_1 was consistent with the nominal rate for both approaches in both small (Table 3) and large (Table 4) sample simulations.

5. AN APPLICATION: ANTIBIOTICS FOR ACUTE OTITIS MEDIA IN CHILDREN

We demonstrate how standardization works in a typical study using the individual patient data (IPD) meta-analysis of six randomized controlled trials for evaluating the efficacy of antibiotics for acute otitis media (AOM) in children (Rovers, 2006). Table 5 depicts the joint distribution of the two binary subgrouping variables examined in the RCTs, age <2 /age ≥ 2 and unilateral AOM/bilateral AOM. The table shows that age at least 2 years and bilaterality

are negatively correlated; that is, children at least 2 years old are less likely to be bilateral than children less than 2 years old ($\hat{\rho} = -0.29$, p -value $< .0001$).

Naive and standardized marginal effects of antibiotics are shown in Table 6 for the 4 subgroups: age < 2 years, age ≥ 2 years, unilateral AOM, and bilateral AOM. There is significant treatment-by-laterality interaction. This interaction effect is exaggerated in the naive approach ($p = 0.0065$), but standardization provides a more conservative inference ($p = 0.014$). The treatment effect appears to be consistent between children less than 2 years old and children at least 2 years old. The nominal p -values for treatment-by-age interaction are 0.61 and 0.78, respectively, based on the naive and standardization approaches. As per our working definition of consistency of treatment effect, the treatment effect is not consistent across all subgroups since the treatment efficacy is significantly different by laterality.

From the simulation studies, we learnt that when only one subgrouping variable (e.g., laterality) interacts with treatment and the other (e.g., age) does not, the estimated laterality-specific treatment effects should be similar between the naive approach and the standardization approach, but the estimated age-specific treatment effects can be overestimated if the correlation between age and laterality is positive, whereas it can be underestimated with a negative correlation. Indeed, we observe this phenomenon in this case example. That is, the estimated treatment effects and their 95% confidence intervals expressed by odds ratio are similar for the two methods in the unilateral subgroup, 0.81 (0.58, 1.12) with the naive approach and 0.83 (0.60, 1.16) with the standardization approach, and also in the bilateral subgroup, 0.39 (0.26, 0.59) with the naive approach and 0.41 (0.26, 0.64) with the standardization approach. However, the treatment effects are underestimated in the less than 2 years old and the at least 2 years old subgroups based on the naive approach (odds ratios closer to 1 with the naive method than with standardization). The estimated odds ratio and their 95% confidence intervals in children less than 2 years old are 0.56 (0.39, 0.81) with the naive approach and 0.66 (0.45, 0.98) with the standardization approach, and are 0.65 (0.45, 0.93) with the naive approach and 0.61 (0.42, 0.90) with the standardization approach.

6. DISCUSSION

Treatment effects in subgroups defined by baseline covariates and the interaction effects (i.e., difference in treatment effect between levels of a baseline covariate) can be confounded even in randomized controlled trial due to the correlation between covariates. Randomization guarantees that the distribution of covariates will be the same in the two arms of a reasonably large trial, but it has no impact on the correlation between the baseline covariates in the sample. It is important to account for this correlation when interpreting the differences in treatment effect across subgroups. We have shown that standardization provides a proper (unconfounded) comparison of treatment effects among strata of a baseline subgrouping variable, when the distribution of other important baseline variables is different in those strata. We have demonstrated using simulations of linear regression models that standardization can remove the bias in the estimation of subgroup-specific treatment effects, as well as in the estimation of the interaction effect between subgrouping

variable and treatment. However, standardization increases the variance of the subgroup-specific treatment effects. Thus, the bias–variance trade-off between naive approach and standardization approach might depend on the specific problem setting. In general, standardization should be preferred. However, in smaller efficacy trials, the choice of the method would depend upon the relative importance placed on bias elimination versus variance reduction.

Standardization provides the correct type I error under the null hypothesis of no interaction, whereas the naive analysis can inflate the type I error. These inferential aspects were also reinforced by a real example involving the RCTs evaluating the efficacy of antibiotics for acute otitis media in children (section 5). In this example, the two binary subgrouping variables, age less than 2 years (yes/no) and unilateral otitis media (yes/no) were negatively correlated ($\hat{\rho} = -0.29$). Although standardization did not alter final conclusions regarding subgroup differences, the example did illustrate quantitative impact of standardization.

Calculation of standardization weights is straightforward. For a binary subgrouping variable, logistic regression can be used to calculate the weights; for a categorical variable with more than two levels, multinomial regression can be used. Standardization can also be applied to continuous baseline covariates such as age or serum albumin levels. We did not consider continuous covariates because that would not fit in the category of subgroup analysis. Standardization is applicable to all types of outcome regression models: linear, binary, count, time-to-event, or other types. Standardization has a negligible impact when the subgrouping variables are weakly correlated (e.g., correlation coefficient of magnitude less than 0.2), or when the magnitude of interaction is very small.

Which reference distribution of covariates should be used to standardize? The distribution of baseline subgrouping variables in the entire trial sample is a natural and obvious choice. However, it is also possible to use a different reference distribution for standardization, for example, the covariate distribution in a particular target population.

vanderWeele and Knol (2011) discussed when confounding is an issue in subgroup analyses of randomized controlled trials. They argued that control for confounding is not needed if the goal is to target subpopulations for intervention; however, if the goal is to intervene on factors defined by subgroups to increase treatment effect then control for confounding must be done. We believe that this distinction is not necessary and that control for confounding is useful for both targeting and intervening. Consider the example of correlation between sex and age in a randomized trial where men are on average younger than women. Suppose the truth is that the efficacy of the treatment wanes with age. We might incorrectly conclude that the treatment is less efficacious in women, and possibly withhold treatment for them—even the younger ones! Therefore, control for confounding is potentially useful regardless of the goal. In conclusion, we recommend that RCTs report the standardized subgroup-specific treatment effects and the corresponding forest plots when the established clinically important baseline covariates are strongly correlated.

Acknowledgments

FUNDING

The analyses upon which this publication is based were performed under contract HHSF2232010000072C, entitled “Partnership in Applied Comparative Effectiveness Science,” sponsored by the Food and Drug Administration, Department of Health and Human Services. This research was also supported by the RSR grant 08-48 awarded by the Center for Drug Evaluation and Research, US Food and Drug Administration. Ravi Varadhan is a Brookdale Leadership in Aging Fellow at the Johns Hopkins University.

REFERENCES

- Food and Drug Administration. Providing clinical evidence of effectiveness of human drug and biological products. Washington, DC: US Department of Health and Human Services; 1998. Guidance for Industry.
- Groenwold RHH, Donders ART, van der Heijden GJMG, Hoes AW, Rovers MM. Confounding of subgroup analyses in randomized data. *Archives of Internal Medicine*. 2009; 169:1532–1533. [PubMed: 19752412]
- Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000; 11:550–560. [PubMed: 10955408]
- Rovers MM, Glasziou P, Appelman CL, Burke P, McCormick DP, Damoiseaux RA, et al. Antibiotics for acute otitis media: A meta-analysis with individual patient data. *Lancet*. 2006; 368:1429–1435. [PubMed: 17055944]
- Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. *Epidemiology*. 2003; 14:680–686. [PubMed: 14569183]
- vanderWeele TJ, Knol MJ. Interpretation of subgroup analyses in randomized trials: Heterogeneity versus secondary interventions. *Annals of Internal Medicine*. 2011; 169:1532–1533.

Table 1A 2×2 table before standardization

	Young	Old	
Men	400	200	600
Women	100	800	900
	500	1000	

Table 2The 2×2 table after standardization

	Young	Old	
Men	200	400	600
Women	300	600	900
	500	1000	

Table 3

Results for small sample size: $N = 200$

Scenario	Subgroup-specific treatment effect (standard error) $(RMSE)_d$						Type-I error of interaction test ^c		
	$\hat{\theta}_{X_1=0}$	$\hat{\theta}_{X_1=1}$	$\hat{\theta}_{X_2=0}$	$\hat{\theta}_{X_2=1}$	$\hat{\theta}_{X_3=0}$	$\hat{\theta}_{X_3=1}$	$\gamma_1 = 0$	$\gamma_2 = 0$	$\gamma_3 = 0$
1. $\gamma_1 = 0, \gamma_2 = 0, \gamma_3 = 0$									
Naive ^a	0.70	0.70	0.70	0.70	0.70	0.70	0.049	0.051	0.051
	(0.20)	(0.24)	(0.21)	(0.21)	(0.23)	(0.19)			
	(0.20)	(0.24)	(0.21)	(0.21)	(0.23)	(0.19)			
Standardized ^b	0.70	0.70	0.70	0.70	0.70	0.70	0.050	0.048	0.050
	(0.23)	(0.30)	(0.25)	(0.28)	(0.33)	(0.22)			
	(0.23)	(0.30)	(0.25)	(0.28)	(0.33)	(0.22)			
2. $\gamma_1 = 0.5, \gamma_2 = 0, \gamma_3 = 0$									
Naive	0.70	1.20	1.00	0.80	1.02	0.82	—	0.102	0.100
	(0.20)	(0.24)	(0.21)	(0.21)	(0.23)	(0.19)			
	(0.20)	(0.24)	(0.23)	(0.23)	(0.26)	(0.20)			
Standardized	0.70	1.20	0.90	0.90	0.90	0.90	—	0.061	0.064
	(0.22)	(0.30)	(0.25)	(0.28)	(0.32)	(0.22)			
	(0.22)	(0.30)	(0.25)	(0.28)	(0.32)	(0.22)			
3. $\gamma_1 = 0.0, \gamma_2 = 0.5, \gamma_3 = 0.5$									
Naive	1.41	1.01	0.88	1.62	0.80	1.55	0.260	—	—
	(0.19)	(0.23)	(0.21)	(0.21)	(0.23)	(0.19)			
	(0.25)	(0.33)	(0.24)	(0.24)	(0.28)	(0.21)			
Standardized	1.25	1.25	1.00	1.50	0.94	1.45	0.052	—	—
	(0.21)	(0.29)	(0.25)	(0.28)	(0.32)	(0.22)			
	(0.21)	(0.29)	(0.25)	(0.28)	(0.32)	(0.22)			
4. $\gamma_1 = 0.0, \gamma_2 = 0.5, \gamma_3 = -0.5$									
Naive	0.65	0.65	0.52	0.78	0.80	0.55	0.049	—	—
	(0.20)	(0.25)	(0.22)	(0.21)	(0.23)	(0.19)			
	(0.20)	(0.25)	(0.24)	(0.24)	(0.28)	(0.22)			

Scenario	Subgroup-specific treatment effect (standard error) (RMSE) ^d						Type-I error of interaction test ^c		
	$\hat{\theta}_{X_1=0}$	$\hat{\theta}_{X_1=1}$	$\hat{\theta}_{X_2=0}$	$\hat{\theta}_{X_2=1}$	$\hat{\theta}_{X_3=0}$	$\hat{\theta}_{X_3=1}$	$\gamma_1 = 0$	$\gamma_2 = 0$	$\gamma_3 = 0$
Standardized	0.65 (0.23)	0.65 (0.30)	0.40 (0.25)	0.90 (0.29)	0.95 (0.32)	0.45 (0.22)	0.065	—	—
	(0.23)	(0.30)	(0.25)	(0.29)	(0.32)	(0.22)			

^aNaive approach using equation (7).

^bStandardization approach using equation (12).

^cRejecting the hypothesis of no interaction at $\alpha = 0.05$.

^dRoot mean-squared error (RMSE) = $\sqrt{\text{bias}^2 + \text{variance}}$.

Table 4

Results for large sample size: $N = 2500$

Scenario	Subgroup-specific treatment effect (standard error) (RMSE) ^d						Type-I error of interaction test ^c		
	$\hat{\theta}_{X_1=0}$	$\hat{\theta}_{X_1=1}$	$\hat{\theta}_{X_2=0}$	$\hat{\theta}_{X_2=1}$	$\hat{\theta}_{X_3=0}$	$\hat{\theta}_{X_3=1}$	$\gamma_1 = 0$	$\gamma_2 = 0$	$\gamma_3 = 0$
1. $\gamma_1 = 0, \gamma_2 = 0, \gamma_3 = 0$									
Naive ^a	0.70 (0.056)	0.70 (0.068)	0.70 (0.060)	0.70 (0.059)	0.70 (0.066)	0.70 (0.054)	0.049	0.051	0.045
Standardized ^b	0.70 (0.063)	0.70 (0.084)	0.70 (0.069)	0.70 (0.076)	0.70 (0.088)	0.70 (0.061)	0.050	0.051	0.048
2. $\gamma_1 = 0.5, \gamma_2 = 0, \gamma_3 = 0$									
Naive	0.70 (0.055)	1.20 (0.068)	1.00 (0.058)	0.80 (0.059)	1.02 (0.065)	0.82 (0.054)	—	0.664	0.657
Standardized	0.70 (0.062)	1.20 (0.083)	0.90 (0.068)	0.90 (0.076)	0.90 (0.087)	0.90 (0.062)	—	0.051	0.048
3. $\gamma_1 = 0.0, \gamma_2 = 0.5, \gamma_3 = 0.5$									
Naive	1.41 (0.053)	1.01 (0.065)	0.88 (0.058)	1.62 (0.058)	0.80 (0.065)	1.55 (0.053)	0.997	—	—
Standardized	1.25 (0.058)	1.25 (0.077)	1.00 (0.068)	1.50 (0.075)	0.95 (0.087)	1.45 (0.061)	0.042	—	—
4. $\gamma_1 = 0.0, \gamma_2 = 0.5, \gamma_3 = -0.5$									
Naive	0.65 (0.055)	0.65 (0.070)	0.52 (0.061)	0.78 (0.060)	0.80 (0.066)	0.55 (0.053)	0.049	—	—
Standardized	0.65 (0.063)	0.65 (0.077)	0.52 (0.068)	0.78 (0.075)	0.80 (0.087)	0.55 (0.061)	0.049	—	—

Scenario	Subgroup-specific treatment effect (standard error) (RMSE) ^d						Type-I error of interaction test ^c		
	$\hat{\theta}_{X_1=0}$	$\hat{\theta}_{X_1=1}$	$\hat{\theta}_{X_2=0}$	$\hat{\theta}_{X_2=1}$	$\hat{\theta}_{X_3=0}$	$\hat{\theta}_{X_3=1}$	$\gamma_1 = 0$	$\gamma_2 = 0$	$\gamma_3 = 0$
Standardized	0.65 (0.062)	0.65 (0.084)	0.40 (0.069)	0.90 (0.078)	0.95 (0.088)	0.45 (0.061)	0.0532	—	—
	(0.06)	(0.08)	(0.07)	(0.08)	(0.09)	(0.06)			

^a Naive approach using equation (7).

^b Standardization approach using equation (12).

^c Rejecting the hypothesis of no interaction at $\alpha = .05$.

^d Root mean-squared error (RMSE) = $\sqrt{\text{bias}^2 + \text{variance}}$.

Table 5

A 2×2 table of subgrouping variables in the meta-analysis of antibiotics for acute otitis media

	Unilateral AOM	Bilateral AOM	
Age < 2 yrs	261	273	534
Age ≥ 2 yrs	611	183	794
	872	456	

Note. Correlation between row and column variables $\rho = -0.29$ (p -value < .0001).

Table 6
Results of acute otitis media (AOM) example. Interaction effects of antibiotics by age and by laterality

Scenario	Naive approach			Standardization		
	OR (95% CI)	$\hat{\gamma}$	Interaction <i>p</i> -value	OR (95% CI)	$\hat{\gamma}$	Interaction <i>p</i> -value
Age						
<2 years	0.56 (0.39, 0.81)			0.66 (0.45, 0.98)		
2 years	0.65 (0.45, 0.93)	0.14	0.61	0.61 (0.42, 0.90)	−0.08	0.78
Laterality of AOM						
Unilateral	0.81 (0.58, 1.12)			0.83 (0.60, 1.16)		
Bilateral	0.39 (0.26, 0.59)	−0.73	0.0065	0.41 (0.26, 0.64)	−0.71	0.014