Published in final edited form as:

Biom J. 2016 January ; 58(1): 133–153. doi:10.1002/bimj.201400102.

Treatment effect heterogeneity for univariate subgroups in clinical trials: Shrinkage, standardization, or else

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

¹Division of Biostatistics and Bioinformatics, Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, MD, USA

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

³Office of Biostatistics, OTS/CDER, FDA, Silver Spring, MD 20993, USA

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

Abstract

Treatment effect heterogeneity is a well-recognized phenomenon in randomized controlled clinical trials. In this paper, we discuss subgroup analyses with prespecified subgroups of clinical or biological importance. We explore various alternatives to the naïve (the traditional univariate) subgroup analyses to address the issues of multiplicity and confounding. Specifically, we consider a model-based Bayesian shrinkage (Bayes-DS) and a nonparametric, empirical Bayes shrinkage approach (Emp-Bayes) to temper the optimism of traditional univariate subgroup analyses; a standardization approach (standardization) that accounts for correlation between baseline covariates; and a model-based maximum likelihood estimation (MLE) approach. The Bayes-DS and Emp-Bayes methods model the variation in subgroup-specific treatment effect rather than testing the null hypothesis of no difference between subgroups. The standardization approach addresses the issue of confounding in subgroup analyses. The MLE approach is considered only for comparison in simulation studies as the “truth” since the data were generated from the same model. Using the characteristics of a hypothetical large outcome trial, we perform simulation studies and articulate the utilities and potential limitations of these estimators. Simulation results indicate that Bayes-DS and Emp-Bayes can protect against optimism present in the naïve approach. Due to its simplicity, the naïve approach should be the reference for reporting univariate subgroup-specific treatment effect estimates from exploratory subgroup analyses. Standardization, although it tends to have a larger variance, is suggested when it is important to address the confounding of univariate subgroup effects due to correlation between baseline covariates. The Bayes-DS approach is available as an R package (DSBayes).

*Corresponding author: suejane.wang@fda.hhs.gov, Phone: +1-301-796-0831.
Both authors contributed equally for the preparation of the manuscript

Additional supporting information including source code to reproduce the results may be found in the online version of this article at the publisher's web-site

Keywords

Bayesian shrinkage estimate; Confounding; Empirical Bayes; Marginal subgroup analysis; Maximum likelihood estimate; Naïve estimate; Standardization; Subgroup analysis

1 Introduction

Traditionally, in a randomized controlled trial, the estimate of an overall treatment effect based on the primary efficacy endpoint, along with its estimated variability, is the basis for concluding whether a new treatment is effective relative to the standard of care or placebo. Treatment effect estimates are often reported for those subgroups within which sizable treatment differences are found. Such selectively reported treatment effect estimates will be biased because one may select extreme estimates. This is an instance of “random high” bias, a phenomenon closely related to regression to the mean (Fleming, 2010). It is obvious that this practice can lead to overestimation of the difference between treatments within the subgroups reported. Results of subgroup analyses are, rightfully, treated with skepticism due to their post hoc nature and to concerns about making Type I and Type II errors (Pocock et al., 2002; Brookes et al., 2004). Commonly used methods to account for these issues, such as multiplicity correction for Type I error, address only a part of the problem, but may not be completely satisfactory from a scientific perspective. Multiplicity adjustments further increase the Type II error, which already tends to be high in subgroup analyses due to the reduced sample size within subgroups. More importantly, the assumption that observed subgroup differences are purely due to randomness, which is generally the basis of multiplicity adjustment procedures, may not always be justifiable, especially in the setting of potential effect modifiers that are prespecified, clinically important subgroup indicators. A reasoned approach that recognizes the possibility that observed differences may be the result of random variability as well as the possibility that they are real may be warranted for exploratory subgroup analyses (e.g., Berry, 1990).

Another important issue is the possibility of confounding while comparing treatment effects across subgroups, which can affect the interpretation of univariate subgroup analyses. Suppose, for example, we find that the treatment effect is significantly different between men and women. This difference may actually be attributable to another characteristic that is different in men and women, for example, to age when the men in the trial are, in general, older than the women. Therefore, the observed differences in subgroup-specific treatment effects could be spurious due to correlations between baseline covariates (Varadhan and Wang, 2014).

In this paper, we examine treatment effect heterogeneity in prespecified subgroups defined by potentially important baseline factors. These baseline factors are generally known to be prognostic and some of them may also be effect modifiers or predictive factors. We explore various alternatives to the traditional, univariate subgroup analyses (naïve) to address the issues of multiplicity and confounding. Specifically, we consider a model-based Bayesian shrinkage (Bayes-DS) and a nonparametric, empirical Bayes shrinkage approach (Emp-Bayes) to temper the optimism of traditional univariate subgroup analyses; a standardization

approach (standardization) that accounts for the correlation between baseline covariates; and, a model-based maximum likelihood estimation (MLE) approach. The Bayesian and empirical Bayes methods are reasoned approaches to modeling the differences in subgroup-specific treatment effect, which recognize the possibility that observed differences may be the result of random variability as well as the possibility that they are real. The standardization approach addresses the issue of confounding in subgroup analyses. The MLE approach is a model-based, frequentist approach, which we use for comparison in simulation studies as the “truth” since the data were generated from the same model.

In Section 2, we introduce the terms univariate and multivariate subgroups and the corresponding subgroup-specific treatment effects. Our estimation approaches for the multivariate subgroup-specific treatment effects include Bayesian, empirical Bayes, and maximum likelihood estimates. These are introduced in Section 3. Even though our focus is on univariate subgroup analysis, we present multivariate subgroup analysis for modeling the underlying structure of treatment effect heterogeneity. Such a higher dimensional treatment effect heterogeneity structure automatically implies a corresponding univariate subgroup-specific treatment effect structure. We show this correspondence in Section 3.4. In Section 4, we briefly summarize the naïve approach and the standardization approach described in Varadhan and Wang (2014). We perform simulation studies in Section 5 for the scenarios considering a combination of sample size (large vs. moderate), correlation structure (strong vs. weak), and univariate prevalence of the prespecified binary baseline covariates (extreme vs. nonextreme). The results of the simulation are summarized in Section 6. In Section 7, we present a challenging setting described by an example study that is in the range of a typical large outcome trial to shed some light on the impact of shrinkage and standardization. In the example, two baseline covariates are strongly correlated with each other and one of which interacts with the new treatment. In Section 8, we articulate what type of estimator would be appropriate to report for subgroup-specific treatment effects. A short discussion is given in Section 9.

2 Problem setup

There are two types of reporting of subgroup-specific treatment effect estimates. The first focuses on the estimation of treatment effects in mutually exclusive subgroups defined on the basis of two or more baseline factors. We denote this as *multivariate subgroups*. The other way is to report the treatment effect estimates in subgroups defined on the basis of each baseline factor, separately. We denote this as *univariate subgroups*. Suppose, for example, that we are interested in treatment effect heterogeneity by race (Caucasian/non-Caucasian) and gender (men/women). In the multivariate subgroup analysis, we have four mutually exclusive subgroups: Caucasian men, Caucasian women, non-Caucasian men, and non-Caucasian women; in the univariate subgroup analysis, we have four subgroups, Caucasian, non-Caucasian, men, and women, which are not mutually exclusive (e.g., the groups Caucasian and men are not mutually exclusive).

The multivariate subgroup-specific treatment effects, Q_j in Table 1, are mutually exclusive and independent. However, the subgroup-specific effects typically vary substantially due to small sample sizes. Thus, the chance of false positives (and false negatives) of hypothesis

tests pertaining to subgroup-specific treatment effects can be high. Shrinkage of treatment effect estimates toward a common effect may reduce the variance of the estimate. Univariate subgroup effects, for example, $\hat{\pi}(x_j = 0)$ in Table 1, on the other hand, are less variable. However, they may be confounded, as stated in Section 1. We note that confounding is also possible in multivariate subgroups, but is less common than in univariate subgroups. Adjustment for confounding may help remove the bias of the subgroup-specific treatment effect estimator.

Let X_i be a baseline binary subgrouping variable (baseline covariate), $i = 1, \dots, K$, where K is the number of baseline subgrouping variables. Using gender as an example, a binary variable has two values, for example, $X_i = 0$ represents male patients and $X_i = 1$ represents female patients for the i -th baseline covariate in a two-arm placebo controlled clinical trial with $T = 1$ receiving treatment and $T = 0$ receiving placebo. Let Y be the clinical outcome of primary interest. Our formulation is also applicable in a two-arm active controlled trial setting, which aims to show superiority of a new treatment.

We will use $K = 3$ throughout this paper, although the methods discussed here apply to any K . The three prespecified clinically or biologically important baseline covariates are likely to be influential, prognostic of disease state, or potentially predictive of treatment response. Table 1 presents the treatment effect estimates based on the type of reporting. There are eight (2^K) mutually exclusive subgroup-specific treatment effect estimates and six ($2K$) univariate subgroup-specific treatment effect estimates.

In this paper, we consider shrinkage, nonshrinkage, naïve, and standardization estimates. The key descriptions of each estimation approach will be briefly summarized in Section 3 for approaches that naturally consider multivariate subgroup-specific treatment effect estimates, and in Section 4 for approaches that use univariate subgroup-specific treatment effect estimates.

3 Multivariate subgroup-specific treatment effect estimates

We consider three estimation approaches that address multivariate subgroup-specific treatment effect estimates. These approaches are described below.

3.1 Maximum likelihood estimate

Consider a two-arm placebo controlled clinical trial where we observe a response Y with three prespecified baseline binary covariates (X_1, X_2, X_3). The response Y can be continuous, binary, counts, or time-to-event. Note that there is nothing special about the number of covariates being three in the subgroup treatment effect estimation problem limiting to those prespecified important baseline factors in confirmatory clinical trials. The approach can be applied to as many subgrouping variables that are prespecified with clinical or biological importance in confirmatory clinical trials. However, we also note that such subgroup analysis is not for ad hoc or data-dredging purpose. It is relatively rare to consider a higher number of covariates as potential effect modifiers. We can write down the following generalized linear model of Y versus K binary baseline covariates:

$$g(E[Y]) = \mu + \tau T + \sum_{i=1}^K (\beta_i X_i + \gamma_i T X_i), \quad (1)$$

where $g(\cdot)$ is a link function for the generalized linear model (e.g., identity link for continuous response, logit link for binary response, log link for counts). The response Y is assumed to be from the exponential family of distributions (e.g., normal for continuous response, binomial for binary response, Poisson for counts). We do not consider higher order interactions between T and X_i and between X_i s, since reliable estimation of such terms would require large sample sizes typically not available in clinical trials. In the MLE approach, the subgroup-specific treatment effects are estimated from the parameter estimates obtained by fitting model (1) using the maximum likelihood approach.

Model (1) can also include regression models for time-to-event data including Cox proportional hazards model and the accelerated failure time model. Here, the estimation may be based on partial likelihood or on estimating equations. Our discussion applies to these models as well, although we will not make any distinction and simply lump these together as the MLE approach. The μ is the mean response (on the link function scale) for placebo patients at the low covariate level ($X=0$, $T=0$), τ the mean response increment for treated patients at the low covariate level ($X=0$, $T=1$), β_i the mean response increment for placebo patients having the covariate characteristics ($X_i=1$, $T=0$) for covariate i , and γ_i is the treatment by covariate i interaction for $i=1, \dots, K$. For example, for $K=3$, the treatment effect (on the link function scale) in the subgroup, $(X_1=x_1, X_2=x_2, X_3=x_3)$, is equal to $\tau + x_1 \gamma_1 + x_2 \gamma_2 + x_3 \gamma_3$, for all $x_1, x_2, x_3 = \{0, 1\}$.

These are the treatment effects in multivariate subgroups. We use the technique described in Section 3.4 to convert the 2^K multivariate subgroup treatment effect estimates to the $2K$ univariate subgroup-specific treatment effect estimates.

3.2 Bayesian shrinkage estimate with a modified Jeffery prior (Bayes-DS)

For the Bayesian shrinkage estimates and their 95% confidence intervals, we adopt Dixon and Simon (1991). We refer to the approach as Bayes-DS and denoted $\Theta = (\mu, \tau, \beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2, \gamma_3)$, the parameter vector in the model for the mean of response $Y(1)$. Note that we do not specify the full data model. This provides a lot of flexibility for fitting various types of response (e.g., continuous, binary, time-to-event) and estimation approaches (e.g., maximum likelihood, quasi-likelihood, partial likelihood). However, we require that the mean response follow (1) and also that the covariates be binary or categorical factors.

The posterior distribution for the multivariate subgroup-specific treatment effects of the Bayes-DS was derived based on the following assumptions:

- (B1)** An initial analysis, based on the mean response structure (1), yields a set of estimates, $\hat{\Theta}$, having approximately a normal distribution,
- $$\hat{\Theta} | \Theta \sim N(\Theta, C), \text{ with } C \text{ a known positive-definite matrix.}$$

- (B2) A prior, $\mu \sim N(\mu_0, \sigma_\mu^2)$; $\tau \sim N(0, \sigma_\tau^2)$; $\beta_i \sim N(0, \sigma_{\beta_i}^2)$ for $i = 1, \dots, K$.
- (B3) The distribution, $\gamma_i | \xi^2 \sim N(0, \xi^2)$, with $p(\xi^2) = [\max(\xi^2, \epsilon)]^{-1}$, where ϵ is a small, positive quantity.

A value of $\epsilon = 0.005$ is used in Bayes-DS to modify the standard Jeffreys indifference prior $(\xi^2)^{-1}$ for the interaction terms so as to achieve a proper posterior density for Θ . The analysis results of Bayes-DS are insensitive to choices of ϵ as long as it is small. The

formulation of Bayes-DS allows the variances of the prior distributions, $\sigma_\mu^2, \sigma_\tau^2, \sigma_{\beta_i}^2$, to grow without limit, since we are only interested in locally uniform priors for the corresponding parameters. This model is deliberately neutral on the question of whether τ , β s, and γ s are positive or negative.

The distribution assumed for γ s in (B3) is an example of the exchangeability, which is a form of symmetry. In other words, none of the interaction parameters is considered, a priori, more likely to have a large magnitude than any of the others. Therefore, the situation of a single interaction of particular a priori interest is excluded. If such an interaction is of interest, it should be modeled separately. It is worth noting that the assumption that all of the covariates are binary is needed to make the regression coefficients commensurate. However, it may still be appropriate to apply this method when some covariates are factors with three or four levels, provided we believe that the exchangeability assumption is reasonable. With the above formulation and assumptions, the posterior distribution for Θ , given $\hat{\Theta}$ in Bayes-DS, is given as:

$$q(\Theta | \hat{\Theta}) = \int g(\hat{\Theta} | \xi^2) q(\Theta | \hat{\Theta}, \xi^2) p(\xi^2) d\xi^2 / \int g(\hat{\Theta} | \xi^2) p(\xi^2) d\xi^2 \quad (2)$$

Dixon and Simon (1991) show that the use of (2) results in proper density function, despite the use of improper priors for μ , τ , and β . In their approach, exchangeability among the interactions is a key assumption, which leads to a shrinkage of the posterior distribution of the interaction terms toward the natural origin of 0, counteracting the tendency of the estimated effects to disperse. Here, no subjective prior distributions are required. Vague priors are used for all the regression coefficients except the treatment-by-covariate interactions, which are assumed to follow a normal distribution.

Note that Bayes-DS also considers 2^K multivariate subgroups. Since we are interested in univariate subgroup-specific treatment effects defined by univariate baseline covariates, we convert treatment effects in 2^K multivariate subgroups to arrive at the $2K$ univariate subgroups. We show how to do this transformation in Section 3.4.

3.3 Empirical Bayes estimate (Emp-Bayes)

We consider the empirical Bayes estimation of multivariate subgroup-specific treatment effects in clinical trials proposed by David and Leffingwell (1990). The Emp-Bayes is a direct application of the methods in a series of articles written by Efron and Morris that are

summarized in Morris (1983). The empirical Bayes method shrinks the extreme estimates toward the overall measure of treatment difference. Both point and interval estimates can be obtained. The Emp-Bayes procedure described in David and Leffingwell (1990) is briefly described below.

Let $\Omega = (\lambda_1, \lambda_2, \dots, \lambda_g)$ be the eight multivariate empirical Bayes subgroup-specific treatment effects to be estimated from the observed subgroup-specific treatment effects, $Q_1, Q_2, \dots, Q_g$. As shown in Table 1, $Q_j = \hat{\pi}_{x_1 x_2 x_3}$, $j = 1, \dots, 8$, are the multivariate subgroup-specific treatment effect estimates assumed to be available. Conditional on λ_j , assume $\hat{\lambda}_j | \lambda_j \sim N(\lambda_j, \sigma_j^2)$, and $\lambda_j \sim N(\theta, \Gamma)$. Following Morris (1983), the empirical Bayes estimate for the multivariate subgroup-specific treatment effect estimates for λ_j is

$$\hat{\lambda}_j = (1 - \hat{B}_j) Q_j + \hat{B}_j \hat{\theta} \quad (3)$$

with $\hat{B}_j = [(g-3)/(g-1)] \times \sigma_j^2 / (\sigma_j^2 + \hat{\Gamma})$, $\hat{\theta} = \sum_{j=1}^g w_j Q_j / \sum_{j=1}^g w_j$, and, $w_j = (\sigma_j^2 + \Gamma)^{-1}$, $g > 3$,

$$\hat{\Gamma} = \max \left\{ 0, \left[\sum_{j=1}^g w_j \left\{ [g/(g-1)] (Q_j + \hat{\lambda}_j)^2 \right\} \right] / \sum_{j=1}^g w_j \right\}.$$

To obtain $\hat{\Gamma}$, one assumes a value for Γ initially that gives the corresponding estimate $\hat{\lambda}_j$, $j = 1, \dots, 8$ (i.e., $g = 2^K$ where $K = 3$), which allows iteration until the estimate of Γ converges. Generally, we get convergence with five to 10 iterations. Note that $0 < \hat{B}_j < 1$. Hence, $\hat{\lambda}_j$ is a weighted average of the $\hat{\theta}$ and the multivariate subgroup-specific treatment effect estimate Q_j , where $\hat{\theta}$, which may be viewed as the overall treatment effect estimate, is itself a weighted average of Q_j . The variance estimate for $\hat{\lambda}_j$ is based on a Taylor expansion derived by Morris (1983):

$$s_j^2 = \sigma_j^2 \left\{ 1 - [(g - r_j)/g] \hat{B}_j \right\} + v_j (Q_j - \hat{\theta})^2 \quad (4)$$

where $r_j = g w_j / \sum_{j=1}^g w_j$, $v_j = [2/(g-3)] \hat{B}_j^2 (\sigma^2 + \hat{\Gamma}) / (\sigma_j^2 + \hat{\Gamma})$, and $\sigma^2 = \sum_{j=1}^g w_j \sigma_j^2 / \sum_{j=1}^g w_j$.

An approximate $(1 - \alpha)100\%$ confidence interval for λ_j using the empirical Bayes estimate in (3) and variance estimate in (4) is

$$\hat{\lambda}_j \pm z_{\alpha/2} s_j \quad (5)$$

where $z_{\alpha/2}$ is the appropriate $(1 - \alpha/2)$ -th percentile value of the standard normal deviate.

For all practical purposes, one substitutes a consistent estimate of σ_j^2 in (3) and (4). Morris

(1983) pointed out that the claim of a $(1 - \alpha)100\%$ coverage in (5) is based on the following three observations: (i) if $\sigma_j^2 = \sigma^2$ for all j , these intervals have the claimed coverage probability, (ii) as g increases, $\hat{\Gamma}$ converges to Γ in probability and Bayes theory guarantees the correct coverage probability, and (iii) the coverage probabilities have held up in various computer simulations.

The empirical Bayes approach that we consider assumes full exchangeability of the treatment effects among the multivariate subgroups. The impact of the exchangeability assumption is that the estimated subgroup-specific treatment effects are shrunk toward a common mean.

3.4 Transformation of multivariate (joint) subgroup-specific treatment effect estimates to univariate (marginal) subgroup-specific treatment effect estimates

Hitherto, in Sections 3.1–3.3, we modeled the estimation of multivariate subgroup-specific treatment effects, where the subgroups are defined by multiple baseline characteristics. Typically, however, in most clinical trial reports, subgroup analyses are presented as univariate estimates, where the subgroups are defined by a single baseline characteristic. Therefore, we show how to transform the multivariate subgroup-specific treatment effects to univariate subgroup-specific treatment effects. From now on, only the transformed univariate subgroup-specific treatment estimates will be presented for these three methods.

To obtain the univariate subgroup-specific treatment effects, $\hat{\pi}(x_1)$, $\hat{\pi}(x_2)$, $\hat{\pi}(x_3)$ from $\hat{\pi}_{x_1, x_2, x_3}$, we integrate (or sum) the multivariate subgroup-specific treatment effect estimates over appropriate conditional distributions of the covariates as follows:

$$\hat{\pi}(x_1) = \sum_{x_3} \sum_{x_2} \hat{\pi}_{x_1, x_2, x_3} P(X_2 = x_2, X_3 = x_3 | X_1 = x_1) \quad (6)$$

For example, to estimate the treatment effect in the univariate subgroup $x_1 = 0$, one computes

$$\begin{aligned} \hat{\pi}(x_1=0) &= \hat{\pi}_{(0,0,0)} P(x_2=0, x_3=0 | x_1=0) + \hat{\theta}(0, 1, 0) P(x_2=1, x_3=0 | x_1=0) \\ &\quad + \hat{\theta}(0, 0, 1) P(x_2=0, x_3=1 | x_1=0) + \hat{\pi}_{(0,1,1)} P(x_2=1, x_3=1 | x_1=0). \end{aligned}$$

Similarly, we obtain $\hat{\pi}(x_2)$ and $\hat{\pi}(x_3)$. Since the multivariate subgroup-specific treatment effects, $\hat{\pi}_{x_1, x_2, x_3}$, are independent, we can easily compute the variances of the univariate subgroup-specific treatment effect estimates as follows:

$$\text{Var}(\hat{\pi}(x_1)) = \sum_{x_3} \sum_{x_2} p_{23,1}^2 \text{Var}(\hat{\pi}_{x_1, x_2, x_3}) \quad (7)$$

where $p_{23,1} = P(X_2 = x_2, X_3 = x_3 | X_1 = x_1)$. We can plug in the sample estimates of the conditional probability to get a consistent estimator of the variance for the univariate subgroup effects. Asymptotic normality is then used to obtain the confidence intervals. This

approach is used to obtain the interval estimates for univariate subgroup-specific treatment effects from MLE (MLE) and empirical Bayes (Emp-Bayes) approaches described in Sections 3.1 and 3.3. The credible intervals for univariate Bayesian (Bayes-DS) subgroup-specific treatment effects are directly obtained from the marginal posterior distribution.

4 Univariate subgroup-specific treatment effect estimates

Now we consider two direct approaches for estimating the univariate subgroup-specific treatment effects. These are “direct” in the sense that they are estimated from univariate models (i.e., models with only a single baseline subgrouping variable and the treatment indicator), whereas the approaches proposed in Sections 3.1–3.3 involve models with multiple subgrouping variables. First, we introduce the naïve estimates. Then, we consider the standardization approach for univariate subgroup analysis previously proposed by Varadhan and Wang (2014).

4.1 Naïve estimate (naïve)

A univariate subgroup analysis is based on the marginal interaction model (MIM)

$$g(E[Y_i]) = \mu_i + \tau_i T + \beta_i X_i + \gamma_i T X_i \quad (8)$$

where $g(\cdot)$ is a link function for the generalized linear model (e.g., identity link for continuous response, logit link for binary response, log link for counts) for baseline covariate i . The response Y is assumed to be from the exponential family of distributions (e.g., normal for continuous response, binomial for binary response, Poisson for counts).

Model (8) is the univariate version of model (1). The treatment effect in the subgroups $X_i = 0$ and $X_i = 1$ are τ_i and $\tau_i + \gamma_i$, respectively. These effects, which ignore other baseline covariates, are referred to as the naïve univariate subgroup-specific treatment effect estimates.

4.2 Standardized estimate (standardized)

The interpretation of univariate subgroup analyses can be problematic due to possible confounding, that is, when multiple baseline covariates can affect the interpretation of the subgroup-specific treatment effects. Suppose, for example, we find that the treatment effect is significantly different between men and women. This difference may actually be attributable to another characteristic that is different in men and women, for example, age, say, men tend to be older than women. Therefore, the differences in subgroup-specific treatment effects could be spurious when there is moderate to strong correlation, 0.3 or larger, between baseline covariates (Varadhan and Wang, 2014).

We will illustrate standardization with a simple example. Consider a 2×2 table of men/women versus Caucasian/non-Caucasian, see Table 2 under “before standardization”. Men are six times more likely than women to be Caucasian, hence a naïve comparison of treatment effect between men and women is likely to be misleading if race is a predictor of treatment response. The probability of being Caucasian among men and women is $2/3$ and

1/9, respectively, whereas it is 1/3 in the overall sample. Thus, the difference in treatment effect between men and women will be confounded by race.

Standardization (also known as inverse probability weighting) is a technique used to remove potential confounding effects. We can standardize the race distributions of men and women to match that of the overall sample, by weighting them differently. We assign a weight of 1/2 to each Caucasian man, which is equal to $\Pr(\text{men})/\Pr(\text{men}|\text{Caucasian})$, and a weight of 2 to each Non-Caucasian man, which is equal to $\Pr(\text{Men})/\Pr(\text{Men}|\text{non-Caucasian})$. Similarly, we also assign a weight of 3 to each Caucasian woman and a weight of 3/4 to each Non-Caucasian woman. After applying the standardization weights, we get the 2×2 table “after standardization” as shown in the second part of Table 2.

In the standardized table, the race distributions in men and women are identical; the gender distributions among Caucasian and non-Caucasian are identical, as well. Therefore, a weighted subgroup analysis will not be confounded. Conceptually, to standardize X_j such that the joint distribution of $\{X_{-j}\}$ excluding X_j is the same in $X_j = 0$ and $X_j = 1$, one can apply the following nonparametric weights as (Varadhan and Wang, 2014):

$$w_i = \frac{P(X_i = x_i)}{P(X_i = x_i | X_{-i} = x_{-i})} \quad (9)$$

When K is moderately large, the denominator of equation (9) can become very small, resulting in large weights causing the standardized subgroup-specific treatment effect estimates to be unstable with a large variance. In general, it is preferable to estimate the weights using parametric models to avoid unstable estimates. For instance, one can fit the logistic regression model for X_j in terms of the other covariates, X_{-j} to obtain the weights, which can be expressed in the form of (9) after model fitting. Then, we fit model (8), but with weights from (9) for covariate i for each subject. The resulting model is the standardized marginal interaction model (SMIM). The standardized univariate subgroup-specific treatment effect in the subgroups $X_j = 0$ and $X_j = 1$ are τ_i^* and $\tau_i^* + \gamma_i^*$, respectively, accounting for the correlation among prespecified baseline covariates. Varadhan and Wang (2014) note that the variances for the parameter estimates can be computed using Huber–White type estimation. When X_j is independent of $\{X_{-j}\}$, the weights w_j equal 1, and the SMIM model reduces to the naïve MIM model, namely, Eq. (8).

5 Simulation studies

We perform simulation studies to investigate the operational characteristics of the univariate subgroup-specific treatment effect estimators and to compare among the five approaches described in Sections 3 and 4. Before we discuss the simulation study, we would like to highlight the following crucial aspects of the five estimators that are compared here. It is important to recognize that these estimators are not all targeting the same truth. The target (or “truth”) of standardization estimator is different from that of all the other estimators. Standardization is aimed at providing an unbiased estimator for the unconfounded marginal effects, whereas the other four estimators are targeting the potentially confounded marginal

effects. The Bayes-DS and Emp-Bayes are shrinkage estimators and hence they attempt to minimize the mean squared error (MSE) with respect to the confounded marginal effects. The Emp-Bayes is nonparametric, whereas the Bayes-DS is model based. The MLE is the most efficient, unbiased estimator for the (potentially confounded) marginal effect, provided the model is correct. Finally, the naïve estimator is a nonparametric estimator unadjusted for any potential confounding. Given that each approach has its unique estimation feature, our presentation of the simulation results side-by-side is primarily to give a rough sense of the operational characteristics and of the trade-off between bias and variance of these estimators.

5.1 Simulation setup

We generated data from model (1), where the response Y was a binary variable and a logit link was assumed. We considered three prespecified clinically important baseline covariates (X_1 , X_2 , and X_3), two sample sizes (N_1 and N_2), two prevalence vectors for three baseline covariates (q_1 for nonextreme prevalence vs. q_2 for extreme prevalence) with two correlation structures (R_1 for weak correlation vs. R_2 for strong correlation), and two parameter vectors (γ^1_{123} for no treatment-by-covariate interactions vs. γ^2_{123} for one treatment-by-covariate interaction).

Suppose the clinical outcome is a composite endpoint from a large sample size of $N_1 = 15,750$. This resembles large outcome trials such as cardiovascular clinical outcome trials. The sample size $N_2 = 1575$ may be considered large in some disease indication, such as asthma trial, schizophrenia trial, but moderate or small depending on the treatment effect size a trial is designed for. We did not consider a sample size smaller than 1575, because it resulted in a small number of subjects in one of the subgroups (approximately 40 subjects per arm in the group with 5% prevalence for $N_2 = 1575$).

The nonextreme prevalence, $q_1 = (q_{11}, q_{12}, q_{13})$, refers to marginal prevalence of binary subgroups, for example, $q_{11} = 40\%$ for $X_1 = 1$, $q_{12} = 50\%$ for $X_2 = 1$, and $q_{13} = 60\%$ for $X_3 = 1$, that is, $q_1 = (0.4, 0.5, 0.6)$. The extreme prevalence q_2 we consider has low prevalence in the first two baseline covariates, that is, $q_{21} = 8\%$ ($X_1 = 1$) and $q_{22} = 5\%$ ($X_2 = 1$), although the prevalence for X_3 is not extreme, namely, $q_{23} = 50\%$. Thus, $q_2 = (0.08, 0.05, 0.50)$.

We will denote the correlation commonly seen in randomized controlled clinical trials as R_1 , which is weak-correlation among the three baseline covariates. The weak-correlation structure scenario we considered has $\rho_{12} = 0.10$ correlation between X_1 and X_2 , $\rho_{13} = 0.03$ between X_1 and X_3 , and $\rho_{23} = 0.09$ between X_2 and X_3 . That is, $R_1 = (\rho^1_{12}, \rho^1_{1,3}, \rho^1_{23}) = (0.10, 0.03, 0.09)$. For R_2 strong correlation structure that is not commonly seen in randomized controlled trials, we set $\rho_{12} = 0.55$ between X_1 and X_2 , leaving the other two pairwise correlations unchanged. Thus, $R_2 = (\rho^2_{12}, \rho^2_{1,3}, \rho^2_{23}) = (0.55, 0.03, 0.09)$.

We assumed the following values for the parameters of (1): $\mu = -1.0$, $\tau = 0.7$, and $\beta_1 = -0.5$, $\beta_2 = -0.5$, $\beta_3 = -0.5$. We will consider two scenarios for the interaction parameters: (i) no treatment-by-covariate interaction, which will be used for comparison, that is, $\gamma^1_{123} = (0, 0, 0)$, and (ii) one treatment-by-covariate interaction, $\gamma^2_{123} = (0.5, 0.0, 0.0)$.

Due to potentially large variability in subgroup-specific treatment effect estimates, it is important to choose a large number of simulations. Hence, we explored the number of simulations needed to reach stationary state of the parameter estimates. Ranging from 200 up to 5000 simulations, we observed that 500 to 1000 simulations appear to give reasonable convergence for a sample size of $N_1 = 15,750$. Thus, the results we report in Section 6 are based on 1000 simulations. We report four scenarios shown in Table 3. These scenarios apply to both the case of no treatment-by-covariate interaction and the case of treatment-by-covariate interaction.

6 Results and interpretations

As illustrated in Sections 3 and 4, the basic quantity summarizing the treatment effect estimates differs among the five methods considered, that is, either a multivariate or a univariate subgroup-specific treatment effects. Bayes-DS, Emp-Bayes, and MLE approaches naturally provide multivariate subgroup-specific treatment effects, whereas the naïve and standardization methods provide univariate subgroup-specific treatment effects. Our primary interest in this paper is the univariate subgroup-specific treatment effect estimates, which we obtain by transforming the multivariate estimates to univariate estimates (Section 3.4). In what follows, we present the univariate subgroup-specific treatment effect (interval) estimates using a forest plot format, followed by a comparison on the MSE across the five methods. In all forest plot figures, the dashed vertical line represents the true univariate subgroup-specific treatment effects.

6.1 Univariate subgroup-specific treatment effect estimates

We present the univariate subgroup-specific treatment effect estimates in a forest plot format that can capture the magnitude and the direction of the bias on the point estimates and the variability of the interval estimates. We label the results by scenario for each subgroup, $X_1 = 0$, $X_1 = 1$, $X_2 = 0$, $X_2 = 1$, $X_3 = 0$, and $X_3 = 1$, respectively. Since strong correlation refers to a correlation between X_1 and X_2 of 0.55, we depict the point and interval estimates for X_1 and X_2 in one graph, see Fig. 1 for a positive treatment-by- X_1 interaction. We omit the figure for X_1 and X_2 in the setting of no treatment-by-covariate interaction and explain the results in Section 6.1.1.

6.1.1 No treatment-by-covariate interaction—When there is no treatment-by-covariate interaction, there is essentially no bias in the univariate subgroup-specific treatment effect estimate regardless of nonextreme or extreme prevalence, weak or strong correlation structure, or both. However, the interval estimates are wider for extreme prevalence, that is, $X_1 = 1$ (8%) and $X_2 = 1$ (5%). The Bayes-DS shrinkage interval estimates have the shortest length followed by naïve, MLE, and Emp-Bayes interval estimates. The standardized interval estimates yield the widest intervals in the scenario where the prevalence is extreme and the correlation is also strong (scenario (iv)). The interval width is greater for $X_2 = 1$ than for $X_1 = 1$ in scenarios (iii) and (iv). These results are not shown.

6.1.2 Treatment-by- X_1 interaction only—When there is only one baseline covariate (X_1) interaction with the treatment, the univariate subgroup-specific treatment effect estimates are unbiased for $X_1 = 0$ and $X_2 = 0$ shown in the first and the third rows of Fig. 1. However, a small positive bias is observed with the Emp-Bayes estimate for $X_1 = 0$. When compared to $X_1 = 0$, not only the magnitude of bias increases, the direction of bias is also reversed for $X_1 = 1$ with the Emp-Bayes estimates. A similar behavior is only observable to a lesser degree for $X_2 = 1$ as compared to $X_2 = 0$ in the “nonextreme prevalence, extreme correlation” of scenario (ii). The interval estimates are generally wide for $X_1 = 1$ and for $X_2 = 1$ with “extreme prevalence” scenarios (iii) and (iv), due to small sample sizes in these subgroups, which can be wider when the correlation structure is also “strong” (scenario (iv)).

It is worth noting that the MLE estimates are essentially unbiased in all scenarios considered, see Fig. 1. This is because the data-generating model in the simulation study is the same model used in the MLE approach. Bias is mostly seen in the Emp-Bayes estimates and to a much less extent with the Bayes-DS estimates. As discussed in Section 4.2, the standardized estimate falls under a different category since it estimates unconfounded subgroup-specific treatment effects. The treatment effects in the subgroups $X_1 = 1$ and $X_2 = 1$ need to be interpreted as the effects “after adjusting for its correlation with the other covariate”. In this context, the univariate subgroup-specific treatment effect estimate is sort of a two dimensional concept since it needs an adjustment for its correlated baseline covariate. Therefore, when discussing subgroup-specific treatment effect estimates in the presence of other potentially correlated baseline covariates, the naïve estimates may be biased without accounting for the correlation. But, the standardized estimates are unbiased, see Varadhan and Wang (Brookes et al., 2004). From Fig. 1, the naïve estimate and the standardized estimate differs for $X_2 = 1$ in scenario (ii) and scenario (iv), where X_1 and X_2 are strongly correlated with the naïve estimate being optimistic. The magnitude of the difference is dampened for $X_2 = 0$ that consists of 95% of the subjects evaluated.

For a sample size that is not large, $N_2 = 1575$, we pick scenario (iv), in the setting that X_1 interacts with treatment, as an example to illustrate the features of various estimators. From Table 4, the interval estimates for all five methods are wider than for $N_1 = 15,750$, which is obviously expected due to smaller sample size (the interval estimates for MLE, naïve and standardization approaches can be unmanageably wide in scenario (iv) due to extremely low prevalence for $X_2 = 1$ subgroup). The finding of wider interval estimates is similar for scenario (iii) (results not shown). Bayes-DS and Emp-Bayes provide greater shrinkage of univariate subgroup-specific treatment effects in $X_1 = 1$ and $X_2 = 1$ in this smaller sample size setting as compared to that with larger sample size N_1 . This is a desirable feature. The interval estimates with Emp-Bayes are wider than those for Bayes-DS.

6.1.3 Baseline covariate X_3 —An intuitive observation is that when one of the prespecified baseline covariates (X_3) does not interact with treatment, and which also has weak correlation with the other two covariates (0.03 with X_1 and 0.09 with X_2), there is no inherent bias in univariate subgroup-specific treatment effect estimates for $X_3 = 0$ and $X_3 = 1$ (Fig. 2) for N_1 . This finding is consistent with that observed for X_1 and X_2 when there is

no treatment-by-covariate interaction. These phenomena prevail, except with wider interval estimates when N_2 is the total sample size.

When there is T-by- X_1 interaction, there is a small negative bias with the naïve and Emp-Bayes estimates for the “nonextreme prevalence” scenarios (i) and (ii) for moderate sample size N_2 , results not shown. This is due to the small, but nonzero, correlation of X_3 with X_1 and with X_2 .

6.2 Mean squared error

We report the MSE of the different estimators under different scenarios from the simulation studies.

6.2.1 No treatment-by-covariate interaction—Under “no-interaction”, the MSEs are generally small for subgroup of sufficient sample sizes, that is, $X_1 = 0$, $X_2 = 0$, $X_3 = 0$, and $X_3 = 1$. The MSEs are also small when the correlation is weak and the univariate prevalence is nonextreme (scenario (i)). These results are not shown.

In the case where the univariate prevalence is extreme and the correlation structure is strong, the MSE is greatest for the standardization approach due to its large variance. As we have noted previously, this may not be a proper comparison with other methods, since the standardization provides unconfounded estimates (Brookes et al., 2004). The naïve estimates perform as well as the MLE in terms of the MSE measure.

6.2.2 Treatment-by- X_1 interaction only—In Table 5, we present simulation results for small $n = 1575$, when T-by- X_1 interaction is present and X_2 is correlated with X_1 . We focus on the comparison of the shrinkage estimators (Bayes-DS and Emp-Bayes) to the naïve estimators for univariate subgroups.

In general, when compared to the naïve estimator, the shrinkage estimators (Bayes-DS and Emp-Bayes) have smaller MSE, except for $X_1 = 0$ and $X_1 = 1$ in scenarios (i) and (ii) (nonextreme prevalence). Most dramatic difference is seen for scenario (iv) in subgroup $X_2 = 1$, where the MSE for shrinkage estimators is more than 25-fold smaller than that of the naïve estimator.

We reiterate that, strictly speaking, standardization cannot be compared to the other estimators, since the target (or “truth”) of standardization is different from that of the other estimators. Standardization is aimed at providing an unbiased estimator for the unconfounded marginal effects, whereas the naïve estimator is a crude estimator and the other three estimators are targeting the confounded marginal effects. The MSE for standardization is much larger than other estimators because it is computed with respect to a “different” target, for example, the MSE with the standardized estimates for $X_1 = 1$ and $X_2 = 1$ in scenarios that include a strong correlation between X_1 and X_2 , especially for small subgroup sample sizes. The existence of a positive treatment-by- X_1 interaction will further complicate the MSE calculation. We emphasize that the standardization is in the context inclusive of other baseline covariates, especially when these baseline covariates are

correlated. Therefore, comparison of MSEs for the standardized estimates with other estimates is not appropriate when it is important to estimate unconfounded marginal effects.

It should be, however, appropriate to compare the MSEs of standardized estimates across the four scenarios. From Table 5, the largest magnitude of MSE is observed mainly for $X_2 = 1$ in the extreme scenario (iv), followed by that either in scenario (iv) for $X_1 = 1$ or scenario (ii) for $X_2 = 1$, followed by that in scenario (iii) for $X_1 = 1$, then, in scenario (ii) for $X_1 = 1$. The remaining cases all have small MSEs. The large MSE for $X_2 = 1$ in the extreme scenario (iv) is likely due to the extreme prevalence of X_2 , which is not only strongly correlated with X_1 , X_1 also interacts with treatment and X_1 is also of extreme prevalence.

7 An example with strong correlation structure and extreme univariate prevalence

In the following, we describe a hypothetical large outcome trial that can occur in practice when some of the important baseline covariates have extreme prevalence. We are interested in learning the impacts due to strong correlation structure and extreme univariate prevalence. Suppose a total of 15,750 patients are randomized in a 1:1 ratio to receive either the placebo or the experimental treatment, where randomization is stratified on their background medications (A vs. B).

In a multiregional clinical trial, it is generally expected that treatment effect should be consistent between regions. However, there is also a general interest to check if a consistent treatment effect estimate between regions is indeed observed in an actual clinical trial. Suppose baseline covariate X_1 is prespecified as the region of specific interest (North America region " $X_1 = 1$ " versus non-North America region " $X_1 = 0$ "). At the planning stage, both genomic biomarker characteristics, X_2 ($X_2 = 1$ indicating "having the biomarker characteristics" and $X_2 = 0$ "without the biomarker characteristics"), and background medication, X_3 ($X_3 = 1$ representing "receiving background medication A" and $X_3 = 0$ "receiving background medication B") are also identified to be important baseline factors.

The primary endpoint is a MACE (major adverse cardiovascular event), which is a composite endpoint defined as a subject having any of the three events: stroke, myocardial infarction, or cardiovascular cause-specific death following treatment administration. The treatment duration is 12 months. Suppose the univariate prevalence distribution of the three prespecified important baseline factors from the data collected are 8% subjects from North America region, 5% subjects having the genomic biomarker characteristics, and 50% subjects received background medication A. The example dataset is simulated based on the parameter vector of $\mu = -2.06$, $\tau = -0.18$, and $\beta_1 = -0.76$, $\beta_2 = 1.11$, $\beta_3 = -0.31$ with $\gamma_1 = 1.04$, $\gamma_2 = -0.57$, $\gamma_3 = -0.10$. The results of the primary composite endpoint are shown in Table 6.

The study clearly demonstrates a superior treatment effect with an estimated 20% risk reduction ($p < 0.0001$) (Table 6). However, the differences were observed in treatment effects between regions ($X_1 = 1$ vs. $X_1 = 0$) and between presence and absence of genomic biomarker ($X_2 = 1$ vs. $X_2 = 0$) in exploratory analyses based on the naive and MLE

approaches (Fig. 3). More specifically, it appears that treatment may be inferior in the North America region ($X_1 = 1$) with odds ratio of new treatment relative to placebo being 1.58 (95% CI: 1.10, 2.28) based on the naïve approach and 1.72 (95% CI: 1.18, 2.52) based on the MLE method, but the treatment appears to be effective in the non-North America region ($X_1 = 0$) with odds ratio of 0.79 (95% CI: 0.71, 0.89) based on the naïve approach and 0.79 (95% CI: 0.70, 0.88) based on the MLE method.

The nominal p -value for the treatment-by- X_1 interaction test is less than 0.0001. It may be of interest to learn whether regional difference observed is likely to be real in a multiregional clinical trial. The pairwise correlation among the three prespecified important baseline factors are summarized. The genomic biomarker (X_2) is strongly correlated with region (X_1) ($\rho_{12} = 0.55$), followed by weak correlation between the genomic biomarker and background medication ($\rho_{23} = 0.09$) and between region and background medication ($\rho_{13} = 0.03$), respectively.

It is suspected that the regional difference may be due to strong correlation with the genomic biomarker characteristics, which are likely to be intrinsic biological factors potentially causing the discrepancy in the treatment effect heterogeneity. We study the impact of correlation on univariate subgroup-specific treatment effect estimates by comparing the results from the naïve approach ignoring the correlation and the results from the standardization approach that adjusts for the nonzero correlation. In this comparison, the interpretation of the standardized univariate subgroup-specific treatment effect estimate should account for its high correlation with baseline covariate X_2 . The univariate treatment effect estimate for $X_1 = 1$ (OR= 2.18, 95% CI: 1.25, 3.78) based on the standardization is highly statistically significant and clearly inferior after adjusting for its correlation of X_1 with X_2 , 0.55, which is large. Thus, we get an even stronger heterogeneity after standardization.

If there is no heterogeneity between regions, the univariate subgroup-specific treatment effects for $X_1 = 1$ and $X_1 = 0$ should be similar in truth. We apply Emp-Bayes and Bayes-DS to the hypothetical example dataset. It is interesting to note that by shrinking the variance estimate in the posterior distribution of treatment-by-covariate interaction assuming no preferential order in the interaction, the Bayes-DS shrinkage method yields a univariate treatment effect odds ratio estimate of 1.48 (95% CI: 0.99, 2.23) for $X_1 = 1$, a borderline significance showing treatment inferiority. However, the degree of shrinkage in the Emp-Bayes treatment effect estimate is higher, that is, odds ratio estimate of 1.15 (95% CI: 0.62, 2.13), results in an inconclusive finding. These shrinkage estimates on the odds ratio for $X_1 = 0$ is 0.80 (95% CI: 0.71, 0.89) with Bayes-DS and is 0.81 (95% CI: 0.68, 0.97) with Emp-Bayes, respectively, see Fig. 3.

The true (potentially confounded) univariate subgroup-specific treatment effects are 0.79 for $X_1 = 0$ and 1.71 for $X_1 = 1$, respectively. We present the true subgroup-specific treatment effects and the 95% interval estimates of the univariate subgroup-specific treatment effects of the example study across all five methods in Table 7. It appears that the MLE yields the univariate subgroup-specific treatment effect estimate closest to the true effects with no observable discrepancy, followed by the naïve estimates, then, the Bayes-DS shrinkage

estimates, with the Emp-Bayes shrinkage estimates being most conservative for $X_1 = 1$ and only slight bias if observable for $X_2 = 1$ (Table 7).

It is also of interest to investigate the univariate subgroup-specific treatment effects for $X_2 = 0$ and $X_2 = 1$. From Fig. 3, the univariate subgroup-specific treatment effect estimate for $X_2 = 1$ varies in the range of 2–7% risk reduction, but is consistent and around 17% risk reduction for $X_2 = 0$ with the shrinkage and nonshrinkage methods. It can be seen that the univariate subgroup-specific treatment effect estimates for $X_2 = 1$ has wide 95% interval estimates across all methods due to there being only 394 patients per treatment group, whereas the univariate subgroup-specific treatment effect estimate derived from 7481 patients per treatment group has a much narrower interval for $X_2 = 0$.

It is noticeable that the standardization estimate is very different from the other four estimates for both $X_1 = 1$ and $X_2 = 1$. This is because the standardization estimate makes adjustment for the strong correlation between X_1 and X_2 . In addition, an interaction exists between the treatment and covariate X_1 . The univariate treatment effect estimates are similar between $X_3 = 0$ and $X_3 = 1$ across all methods considered including the standardized estimates, see the bottom panel of Fig. 3. From the finding in the simulation studies, this is likely due to the following reasons. First, baseline covariate X_3 does not interact with treatment. Second, X_3 is at most weakly correlated with either X_1 (correlation = 0.03) or X_2 (correlation = 0.09). Third, the univariate prevalence of X_3 is nonextreme, that is, 50%.

8 What is an appropriate estimator for reporting metric of a subgroup-specific treatment effect?

We explored the sampling properties of five different estimators of univariate subgroup-specific treatment effects via simulation studies. Bayes-DS and Emp-Bayes methods use the idea of shrinkage (engendered by the exchangeability assumption) to derive the univariate estimates from their corresponding multivariate estimates. Bayes-DS approach is model based in the sense that it imposes a simple mean response structure for treatment–covariate interactions. It also employs a modified Jeffery prior for the hyper-prior of the variance component of the interaction terms (locally uniform or flat priors are assumed for all the other parameters), and therefore, does not require any subjective prior distributions. Emp-Bayes is nonparametric and it uses an empirical normal prior, which is estimated based on available data. It is essentially equivalent to a frequentist random effects modeling approach.

8.1 Naïve estimates

The naïve univariate subgroup-specific treatment effect estimates routinely reported in clinical journals ignore correlation between the baseline covariates, and thus are susceptible to confounding. However, it is important to recognize that the naïve estimates are simple and provide crude marginal estimates. We believe the naïve univariate subgroup-specific treatment effect interval estimates should always be reported mainly to serve as reference estimates to which other estimators can be compared and contrasted. Furthermore, naïve estimates may also be useful for evidence synthesis (meta-analysis) of subgroup effects. The naïve estimates are similar to the MLE both in the point estimates and the interval estimates

from simulation studies. This is mainly because the MLE approach used the correct model, that is, the data-generating model.

8.2 MLE

In the simulation studies, the MLE estimator appears to be unbiased in all scenarios. Once again, this is because the MLE is estimated based on the correct model. However, it is unlikely the unbiasedness of the MLE univariate subgroup effects will be preserved when the model is misspecified. This needs further investigation. We plan to study the properties of MLE and Bayes-DS under model misspecification in future research.

8.3 Bayes-DS

Compared to Emp-Bayes, the Bayes-DS credible interval estimates of univariate subgroup-specific treatment effects not only shrink toward overall treatment effect estimate, the credible intervals of Bayes-DS are also generally narrower even for the small total sample size considered in the simulation studies, for example, $X_2 = 1$ consisted of approximately 40 patients per arm (Table 4). Furthermore, shrinkage in Bayes-DS does not really influence the univariate subgroup-specific treatment effect credible interval estimates when the subgroup of interest is weakly correlated with other baseline covariate or when the prevalence is nonextreme. In fact, the credible intervals for Bayes-DS become similar to the naïve estimates under these favorable settings. These features also hold for small studies. It is important to note that these observations regarding the Bayes-DS approach assume that the mean response model is correctly specified. Our conjecture is that when the model is incorrectly specified, Bayes-DS might have a larger bias than Emp-Bayes. However, even in that case the MSE might still be smaller for Bayes-DS than for Emp-Bayes. This would be important, if it were true. We plan to study the impact of model misspecification in a future research.

In the original paper on the Bayes-DS model (Dixon and Simon, 1991), it was suggested that the method was applicable primarily to binary subgroups. However, this does not appear to be essential. It seems to reason that the exchangeability assumption can also apply to categorical subgroups. Therefore, the method is also applicable to categorical subgroups with more than two levels (personal communication with Dr. Rich Simon of National Cancer Institute).

8.4 Empirical Bayes

As shown in all figures including the hypothetical example study, the Emp-Bayes interval estimate is always wider than that for the Bayes-DS across all scenarios investigated in the simulation studies. This is due to the nonparametric nature of Emp-Bayes, in the sense that it does not assume a simpler data structure for how the subgroup-specific treatment effects are related. On the contrary, Bayes-DS assumes a simpler structure (i.e., only first-order interactions exist between treatment and covariates) for the mean response, thus borrowing more information across the subgroups. In addition, the Emp-Bayes interval estimates can be biased either upward or downward, depending on the univariate prevalence distribution, the correlation structure, and whether there is treatment-by-covariate interaction. Among the prespecified covariates for reporting if a covariate X_1 interacts with treatment, the MSE for

$X_1 = 1$ is larger with the Emp-Bayes approach than with the Bayes-DS approach, if the prevalence of $X_1 = 1$ is nonextreme. However, an important advantage of Emp-Bayes approach is that it is not model dependent.

8.5 Standardization

Standardization addresses a different question from that of the other estimators, namely, estimating unconfounded univariate subgroup-specific treatment effects. We recommend that standardized subgroup-specific treatment effects be reported, along with the traditional naïve estimates, when the correlation between prespecified subgrouping variable is 0.3 or larger.

9 Discussion

Subgroup analyses are most often exploratory. When the subgroups are prespecified, preferably on the basis of biological or clinical rationale, it is sensible to characterize the sampling properties of subgroup-specific treatment effect estimators. On the other hand, when subgroups are chosen on a post hoc basis, that is, after looking at the treatment response, the sampling properties of estimators cannot be properly characterized. In this work, we studied the sampling properties of different subgroup-specific treatment effect estimation approaches via simulation studies, which closely mimicked clinical trials for large and moderate sample sizes. We suggest that the naïve estimates should be the reference metric for reporting univariate subgroup-specific treatment effects. Our investigation showed that the univariate Bayes-DS shrinkage approach can protect against optimism in traditional univariate reporting of a naïve estimate or an MLE estimate provided the model for the mean response is correctly specified (i.e., higher order treatment–covariate interactions are negligibly small). This protection applies to both large and moderate sample size scenarios, even when the smallest subgroup is quite small. This approach is implemented in an R package called DSBayes (Varadhan and Yao, 2014). The R code for all the estimation approaches considered here can be obtained from the authors.

It is an open problem whether the subgroup-specific treatment effect estimates based on univariate Bayes-DS approach will likely shrink further or become anticonservative when the underlying model for the mean response is misspecified (i.e., when the higher order treatment–covariate interactions are not negligible). Examination of the properties of Bayes-DS under model misspecification is a topic of our future research.

In reporting subgroup-specific treatment effect estimates, our simulation studies show that the empirical Bayes estimates can underestimate or overestimate the true effects when two baseline covariates are strongly correlated and one of them interacts with the treatment. When there is concern about the presence of confounding, we recommend that the standardization approach is used. However, it is likely that the standardization approach may introduce greater variability than the naïve approach when the sample size is not sufficiently large.

It is common practice to create subgroups by categorizing continuous baseline variables (e.g., continuous age is categorized into discrete age groups). This practice is not ideal, since categorization is often arbitrary and may not have plausible clinical/biological rationale. It

can be especially misleading to choose the cut-offs for categories in a post hoc manner (i.e., on the basis of treatment response).

Our work on the different estimators for subgroup-specific treatment effects may also be applicable in reporting of subgroup-specific treatment-induced safety effects in comparative effectiveness research. The standardization estimate and the naïve estimate may, in particular, be reported. However, the concept of shrinkage for subgroup-specific treatment-induced safety effect estimates is new and may not be appropriate since safety concerns within specific subgroups are generally signal detection oriented.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We wish to thank Prof. Dieter Hauschke and Dr. Claudia Schmoor for their kind invitation. The first author (R.V.) would like to acknowledge the funding from the contract provided by the Patient-Centered Outcomes Research Institute (PCORI) under contract ME-1303-5896.

Conflict of interest

This article reflects the views of the authors and should not be construed to represent the views or policies of the U.S. Food and Drug Administration nor of the Patient-Centered Outcomes Research Institute (PCORI).

References

- Berry DA. Subgroup analyses. *Biometrics*. 1990; 46:1227–1230. [PubMed: 2085637]
- Brookes ST, Whitley E, Egger M, Smith GD, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup specific analyses; power and sample size for the interaction test. *Journal of Clinical Epidemiology*. 2004; 57:229–236. [PubMed: 15066682]
- David CE, Leffingwell DP. Empirical Bayes estimates of subgroup effects in clinical trials. *Controlled Clinical Trials*. 1990; 11:37–42. [PubMed: 2157579]
- Dixon D, Simon R. Bayesian subset analysis. *Biometrics*. 1991; 47:471–881.
- Fleming TR. Clinical trials: discerning hype from substance. *Annals of Internal Medicine*. 2010; 153:400–406. [PubMed: 20855804]
- Jones HE, Ohlssen DI, Neuenschwander B, Racine A, Branson M. Bayesian models for subgroup analysis in clinical trials. *Clinical Trials*. 2011; 8:129–143. [PubMed: 21282293]
- Morris CN. Parametric empirical Bayes inference: theory and application. *Journal of American Statistical Association*. 1983; 78:47–55.
- Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine*. 2002; 21:2917–2930. [PubMed: 12325108]
- Varadhan R, Wang SJ. Standardization for subgroup analysis in randomized controlled trials. *Journal of Biopharmaceutical Statistics*. 2014; 1:154–167.
- Varadhan R, Yao W. DSBayes: Bayesian subgroup analysis in clinical trials. R package, Version 1.1. 2014 <http://cran.r-project.org/web/packages/DSBayes/index.html>.

Figure 1.

Univariate subgroup-specific treatment effect estimates for X_1 and X_2 across scenarios by five methods when treatment-by- X_1 interaction exists; $N_1 = 15,750$.

Figure 2. Univariate subgroup-specific treatment effect estimates for X_3 across scenarios by five methods; X_3 is weakly correlated with X_1 and X_2 , and not interacts with treatment; $N_1 = 15,750$.

Figure 3.

Univariate subgroup-specific treatment effect estimates for a hypothetical example with X_1 a potential predictor that is strongly correlated with X_2 and interacts with treatment, whereas X_3 is at most weakly correlated with X_1 (corr = 0.03) and X_2 (corr = 0.09).

Table 1

Notation used for subgroup-specific treatment effect estimates by type of reporting.

Type of reporting	Multivariate subgroups	Univariate subgroups
Treatment effect estimates	$Q_j = \hat{\pi}_{x_1 x_2 x_3}, j = 1, \dots, 8$, for example, $Q_1 = \hat{\pi}_{000}$ is the treatment effect estimate for the subgroup $(x_1 = 0, x_2 = 0, x_3 = 0)$	$\hat{\pi}(x_i = 0), \hat{\pi}(x_i = 1), i = 1, 2, 3$, for example, $\hat{\pi}(x_1 = 0)$ is the treatment effect estimate for the subgroup $x_1 = 0$

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

A 2×2 table of gender by race illustrating standardization.

		Before standardization		Race	
		Caucasian		Non-Caucasian	
Gender	Men	400	200	600	
	Women	100	800	900	
		500	1000		
		After standardization		Race	
		Caucasian		Non-Caucasian	
Gender	Men	200	400	600	
	Women	300	600	900	
		500	1000		

Table 3

The scenarios of correlation and prevalence for investigation via simulation studies.

Correlation\prevalence	Nonextreme prevalence $q_1 = (0.4, 0.5, 0.6)$	Extreme prevalence $q_2 = (0.08, 0.05, 0.50)$
Weak correlation $R_1 = (\rho^1_{12}, \rho^1_{1,3}, \rho^1_{23}) = (0.10, 0.03, 0.09)$	Scenario (i)	Scenario (iii)
Strong correlation $R_2 = (\rho^2_{12}, \rho^2_{1,3}, \rho^2_{23}) = (0.55, 0.03, 0.09)$	Scenario (ii)	Scenario (iv)

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 4

Univariate subgroup-specific treatment effect interval estimates (95% confidence interval) for scenario (iv) by method when X_1 interacts with treatment for $N_2 = 1575$.

$N_2 = 1575$	Bayes-DS	Emp-Bayes	MLE	Naïve	Standardized
$X_1 = 0$	2.06 (1.64, 2.60)	2.05 (1.46, 2.89)	2.03 (1.40, 2.95)	2.00 (1.59, 2.52)	2.00 (1.57, 2.55)
$X_1 = 1$	2.16 (1.40, 5.36)	2.42 (0.74, 7.87)	3.91(0.02, *)	3.41 (1.30, 8.90)	3.58 (1.07, *)
$X_2 = 0$	2.08 (1.65, 2.61)	2.07 (1.45, 2.96)	2.08 (1.65, 2.62)	2.05 (1.63, 2.57)	2.08 (1.66, 2.62)
$X_2 = 1$	2.13 (1.20, 5.76)	2.26 (0.55, 9.25)	3.77 (0.00, *)	3.61 (0.00, *)	3.58 (0.38, *)
$X_3 = 0$	2.08 (1.59, 2.75)	2.08 (1.42, 3.04)	2.12 (1.15, 3.91)	2.08 (1.54, 2.82)	2.09 (1.54, 2.83)
$X_3 = 1$	2.09 (1.56, 2.83)	2.08 (1.34, 3.23)	2.16 (0.80, 5.84)	2.09 (1.49, 2.94)	2.09 (1.49, 2.94)

* A large number indicating inadequate sample size in the subgroup.

Table 5

The MSEs for univariate subgroups by the analysis method and by scenario when T-by- X_1 interaction is present ($N_1 = 1575$).

Scenario	Method	$X_1 = 0$	$X_1 = 1$	$X_2 = 0$	$X_2 = 1$	$X_3 = 0$	$X_3 = 1$
(i) Nonextreme prevalence, weak correlation	Bayes-DS	0.031	0.066	0.019	0.021	0.020	0.019
	Emp-Bayes	0.030	0.071	0.017	0.019	0.017	0.018
	MLE	0.025	0.050	0.029	0.039	0.034	0.032
	Naïve	0.024	0.048	0.028	0.037	0.034	0.031
	Standardized	0.024	0.049	0.028	0.040	0.034	0.031
(ii) Nonextreme prevalence, strong correlation	Bayes-DS	0.028	0.071	0.022	0.032	0.021	0.020
	Emp-Bayes	0.028	0.077	0.021	0.034	0.018	0.018
	MLE	0.024	0.055	0.028	0.041	0.035	0.032
	Naïve	0.023	0.053	0.027	0.040	0.035	0.032
	Standardized	0.029	0.097	0.057	0.079	0.035	0.032
(iii) Extreme prevalence, weak correlation	Bayes-DS	0.014	0.200	0.013	0.021	0.016	0.017
	Emp-Bayes	0.014	0.140	0.013	0.048	0.014	0.015
	MLE	0.017	0.250	0.014	1.900	0.025	0.038
	Naïve	0.014	0.220	0.014	1.800	0.024	0.030
	Standardized	0.014	0.240	0.014	2.000	0.024	0.030
(iv) Extreme prevalence, strong correlation	Bayes-DS	0.014	0.210	0.013	0.120	0.016	0.016
	Emp-Bayes	0.013	0.130	0.013	0.110	0.014	0.014
	MLE	0.015	0.960	0.014	3.400	0.027	0.044
	Naïve	0.014	0.330	0.014	3.300	0.024	0.030
	Standardized	0.015	0.550	0.014	6.400	0.024	0.030

Table 6

Primary efficacy analysis results of the MACE endpoint in the example study.

	Experimental treatment	Placebo
Event rate	8.1%	10.1%
Relative risk		0.80
<i>p</i> -Value		<0.0001

Univariate subgroup-specific treatment effects (truth and 95% interval estimates) of the example study.

Table 7

Example study	Univariate subgroup-specific treatment effects					
	$X_1 = 0$	$X_1 = 1$	$X_2 = 0$	$X_2 = 1$	$X_3 = 0$	$X_3 = 1$
True effect	0.79	1.71	0.83	0.96	0.88	0.80
Bayes-DS	0.80 (0.71, 0.89)	1.48 (0.99, 2.23)	0.83 (0.74, 0.93)	0.98 (0.68, 1.43)	0.88 (0.76, 1.01)	0.80 (0.68, 0.94)
Emp-Bayes	0.81 (0.68, 0.97)	1.15 (0.62, 2.13)	0.83 (0.69, 1.01)	0.93 (0.53, 1.62)	0.86 (0.72, 1.04)	0.81 (0.64, 1.02)
MLE	0.79 (0.70, 0.88)	1.72 (1.18, 2.52)	0.83 (0.74, 0.93)	0.96 (0.65, 1.43)	0.88 (0.77, 1.02)	0.80 (0.67, 0.94)
Naïve	0.79 (0.71, 0.89)	1.58 (1.10, 2.28)	0.83 (0.75, 0.93)	0.94 (0.64, 1.39)	0.88 (0.76, 1.02)	0.79 (0.67, 0.94)
Standardized	0.76 (0.67, 0.86)	2.18 (1.25, 3.78)	0.86 (0.77, 0.96)	0.47 (0.20, 1.13)	0.88 (0.76, 1.02)	0.80 (0.67, 0.94)