Kao-Tai Tsai* and Karl Peace

Analysis of Subgroup Data of Clinical Trials

Abstract: Large randomized controlled clinical trials are the gold standard to evaluate and compare the effects of treatments. It is common practice for investigators to explore and even attempt to compare treatments, beyond the first round of primary analyses, for various subsets of the study populations based on scientific or clinical interests to take advantage of the potentially rich information contained in the clinical database. Although subjects are randomized to treatment groups in clinical trials, this does not imply the same degree of randomization among sub-populations of the original trials. Therefore, comparisons of treatments in sub-populations may not produce fair and unbiased results without properly addressing this issue. Covariate adjustments in regression analysis and propensity score matching are commonly used to address the non-randomized nature of the sub-populations issue with various degrees of success. However, further improvements to these methods are still possible. In this article, we propose an analysis strategy that shows improvement to conventional methods. Treatment effects and their differences are estimated after adjustment for background imbalances. Treatment groups are then compared using confidence intervals whose limits are determined using the Robbins–Monro stochastic approximation. Data from a recent clinical trial are used to illustrate the methodology.

Keywords: propensity score, Genetic matching, Robbins–Monro confidence intervals

***Corresponding author: Kao-Tai Tsai,** Jiann-Ping Hsu College of Public Health, Georgia Southern University, Statesboro, GA, USA, E-mail: tsai0123@yahoo.com
Karl Peace, Jiann-Ping Hsu College of Public Health, Georgia Southern University, Statesboro, GA, USA, E-mail: kepeace@georgiasouthern.edu

1 Introduction

As stated by Peto et al. [1], "There is simply no serious scientific alternative to the generation of large-scale randomized evidence. If trials can be vastly simplified, ..., and thereby made vastly larger, then they have a central role to play in the development of rational criteria for the planning of health care throughout the world." Recruitment of a large number of eligible patients from a general population is both a major strength and weakness of large pragmatic trials. Deliberately broad and sometimes ill-defined entry criteria mean that the overall result can be difficult to apply to particular groups. In modern medical practice, individualized medicine is often advocated, and clinicians frequently need to make decisions about how best to use results of randomized clinical trials (RCT) and systematic reviews to maximize the wellbeing of their patients. Therefore, subgroup analyses have become increasingly necessary, if heterogeneity of treatment effects is likely to occur. However, various views seem to exist among scientific and clinical communities about the proper justifications and conducts for this kind of analysis.

Some statisticians and non-clinical epidemiologists have warned about the dangers of subgroup analyses due to the concerns of multiplicities, data dredging, false positive subgroup treatment effects, and the rarity of qualitative heterogeneity of relative treatment effect (e.g. see the discussions and references in Wang et al. [2] and the simulation study by Brookes et al. [3]). On the other hand, groups of practicing clinicians and statisticians have warned of the dangers of applying the overall results of large trials to individual patients without consideration of patho-physiology or other determinants of individual response [4–6]. Rothwell [7] put it in this way.

> The main potential of subgroup analysis is not in the identification of groups that differ in their response to treatment for reasons of patho-physiology, but is in answering practical questions about how treatments should be used most effectively. Subgroup analyses related to questions of the practical application of interventions can be vital to effective clinical practice.

In this article, without adding to the debates, we focus on the better methodological conduct of subgroup data analyses. In practice, sometimes even with stratified randomization specifically for subgroups of interest, the subgroup data may still have certain degrees of imbalance among some covariates, the accidental bias as stated by Efron [8] and Lachin [9], which may create bias in data analysis. In general, smaller trials are more likely to have covariate imbalance; however, imbalance can sometimes also occur in larger trials. For example, in a post hoc analysis, Wright et al. [10] compared the outcomes in hypertensive black and non-black patients treated with chlorthalidone, amlodipine, and lisinopril in the trial conducted by the ALLHAT Collaborative Research Group. Even with the sample size of several thousands in each stratum, they still showed significant imbalance in several covariates among the strata. Therefore, blindly analyzing subgroup data without proper adjustment of covariate imbalances may be problematic and even produce misleading results. Unfortunately, this is not uncommon in medical journals. Our intention in this article is to bring the matching methods, which had been used quite often in other scientific disciplines, into clinical trial subgroup data analysis as a tool to better adjust the covariate imbalance. Our simulation study had shown some remarkable merits of the matching methods for this purpose.

The organization of this manuscript is as follows. In Section 2, we briefly describe the statistical background of propensity score and the methods of matching to adjust the potential covariate imbalance, followed by comparisons of their performance in a simulation. In Section 3, we discuss the estimation of confidence interval of treatment effect using stochastic approximation. An example is provided in Section 4 to illustrate the procedures discussed herein using a data set from a recent clinical trial. Further discussions are presented in Section 5.

2 Statistical method and notations

The concept of propensity scores is thoroughly discussed by Rosenbaum and Rubin [11] as well as by other authors. In the following, we describe a few key points for analytical purposes. Let Y_{i1} denote the response to the experimental treatment of subject $i(1 \le i \le N)$ and Y_{i0} denote the response to the control treatment of subject i. Let $\boldsymbol{X}_i$ denote the vector of covariates associated with subject i and $T_i = 1(0)$, if subject i receives experimental (control) treatment. The observed outcome for subject i is then $Y_i = T_i Y_{i1} + (1 - T_i) Y_{i0}$.

If the subjects were appropriately randomized between experimental treatment and control groups, then

$$E(Y_{ij}|T_i = 1) = E(Y_{ij}|T_i = 0), \quad j = 0, 1,$$

even though $E(Y_{i0}|T_i = 1)$ of the experimental treatment group and $E(Y_{i1}|T_i = 0)$ of the control group cannot be estimated from the data since each subject can receive only either experimental or control treatment, but not both.

If the data were appropriately randomized, the estimand of the average treatment effect, which can be estimated empirically using the observed data, can be written as

$$\tau = E(Y_{i1}|T_i = 1) - E(Y_{i0}|T_i = 0),$$

which can be further re-expressed as

$$\begin{aligned}
\tau &= a_1 E(Y_{i1}|T_i = 1) - a_2 E(Y_{i0} \mid T_i = 0) + b_1 E(Y_{i1} \mid T_i = 1) - b_2 E(Y_{i0} \mid T_i = 0) \\
&= a_1 E(Y_{i1} \mid T_i = 1) - a_2 E(Y_{i0} \mid T_i = 1) + b_1 E(Y_{i1} \mid T_i = 1) - b_2 E(Y_{i0} \mid T_i = 1) \\
&= a_1 E(Y_{i1} \mid T_i = 1) - a_1 E(Y_{i0} \mid T_i = 1) + b_1 E(Y_{i1} \mid T_i = 0) - b_1 E(Y_{i0} \mid T_i = 0) \\
&= a_1 [E(Y_{i1} \mid T_i = 1) - E(Y_{i0} \mid T_i = 1)] + b_1 [E(Y_{i1} \mid T_i = 0) - E(Y_{i0} \mid T_i = 0)],
\end{aligned}$$

where a_1, b_1, a_2, b_2 are all positive with $a_1 + b_1 = 1$, $a_2 + b_2 = 1$, with

$$\tau_1 = [E(Y_{i1} \mid T_i = 1) - E(Y_{i0} \mid T_i = 1)] \text{ and } \tau_0 = [E(Y_{i1} \mid T_i = 0) - E(Y_{i0} \mid T_i = 0)]$$

being the (unobserved) treatment effects from the experimental treatment and control groups, respectively.

When the covariate imbalance occurs, proper matchings of subjects to better balance covariates are usually recommended in order to obtain a more appropriate estimate of treatment effect. Given covariate $\boldsymbol{X}_i$ and following the results of Rubin [12, 13], one can show that

$$E(Y_{ij} \mid \boldsymbol{X}_i, T_i = 1) = E(Y_{ij} \mid \boldsymbol{X}_i, T_i = 0).$$

Therefore, the treatment effect of the experimental treatment group

$$\tau_1 = E_{\{\boldsymbol{X}_i \mid T_i = 1\}}\{E(Y_i \mid \boldsymbol{X}_i, T_i = 1) - E(Y_i \mid \boldsymbol{X}_i, T_i = 0)\},$$

where the expectation is taken over $\{\boldsymbol{X}_i \mid T_i = 1\}$, can be estimated.

Define the propensity score as

$$e(\boldsymbol{X}_i) = P(T_i = 1 | \boldsymbol{X}_i) = E(I\{T_i = 1\} \mid \boldsymbol{X}_i),$$

namely, the probability of patient i being assigned to experimental treatment given the covariate. Assume

$$(i)\ 0 < P(T_i = \delta_i \mid \boldsymbol{X}_i) < 1 \text{ and } (ii) P(T_1 = \delta_1, \ldots, T_N = \delta_N \mid X_1, \ldots, X_N) = \prod_{i=1}^{N} e(\boldsymbol{X}_i)^{\delta_i}(1 - e(\boldsymbol{X}_i))^{(1-\delta_i)},$$

where $\delta_i = 0$ or 1, Rosenbaum and Rubin [11] showed that

$$\tau_1 = E_{\{e(\boldsymbol{X}_i) | T_i = 1\}}\{E(Y_i \mid e(\boldsymbol{X}_i), T_i = 1) - E(Y_i \mid e(\boldsymbol{X}_i), T_i = 0) | T_i = 1\},$$

where the expectation is taken over $\{e(\boldsymbol{X}_i) | T_i = 1\}$, and τ_0 can be expressed similarly. Therefore, the average treatment effect can be derived from the estimates of τ_1 and τ_0. More details about the propensity score can be found in Rosenbaum [14] in addition to the articles mentioned herein.

Let $\boldsymbol{X}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})'$ and $m \leq k$ be the vector of covariates. A common method to estimate $e(\boldsymbol{X}_i)$ is via the logit function, i.e.

$$\text{logit}(e(\boldsymbol{X}_i)) = \beta_0 + h_1(\eta_{1i}) + h_2(\eta_{2i}), \tag{1}$$

where h_1 and h_2 are known functions and

$$\eta_{1i} = \sum_{r=1}^{m} f_r(x_{ir}) \text{ and } \eta_{2i} = \sum_{r,q=1}^{m} f_r(x_{ir}) f_q(x_{iq})$$

represent the main effects and interactions, respectively. The parameters in eq. (1) can be estimated using MLE-based methods. Goodness-of-fit can be checked graphically via, for example, Landwehr et al. [15] or Tsai [16].

According to Rosenbaum and Rubin [11], it is advantageous to sub-classify or match not only on $e(\boldsymbol{X})$ but also for other functions of $\boldsymbol{X}$. In particular, such a refined procedure may be used to obtain estimates of the average treatment effect in a subpopulation defined by the components of $\boldsymbol{X}$, for example, gender or different disease classifications.

In addition to matching by the propensity score defined above, other matching schemes exist. Two of the more commonly used methods are Mahalanobis and Genetic matching. Given two covariates, $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$, the distances between them used in Mahalanobis and Genetic matching are defined as

$$md(\boldsymbol{X}_i, \boldsymbol{X}_j) = \{(\boldsymbol{X}_i - \boldsymbol{X}_j)' S^{-1} (\boldsymbol{X}_i - \boldsymbol{X}_j)\}^{1/2},$$

and

$$gmd(\boldsymbol{X}_i, \boldsymbol{X}_j) = \{(\boldsymbol{X}_i - \boldsymbol{X}_j)' S^{-1/2} W S^{-1/2} (\boldsymbol{X}_i - \boldsymbol{X}_j)\}^{1/2},$$

respectively, where $S^{1/2}$ is the Cholesky decomposition of the covariance matrix of $\boldsymbol{X}$, and $\boldsymbol{W}$ is a diagonal positive definite weight matrix. The elements of $\boldsymbol{W}$ can be chosen objectively to simultaneously minimize the distributional difference and location difference of covariates between the experimental treatment and control groups based on the Kolmogorov–Smirnov test and *t*-test, respectively [17]. On the other hand, $\boldsymbol{W}$ can also be chosen somewhat subjectively depending on the relative importance among the matched variables. When certain variables are considered as more important and higher degree of balance for the selected variables is desired, one can assign higher weights for those variables during the matching processes.

The matching can be performed with either pair-matching or full-matching depends on the distributions of the data. The treatment effect can then be estimated between the matched data in the control and experimental groups. The overall treatment effect can be estimated by a weighted average of the individual matched groups. In the example below, we used the full-matching to estimate the treatment difference.

The conventional test of covariate balance between groups based on the *t*-test focuses only on location and can miss distributional differences between the covariates. On the other hand, the Kolmogorov–Smirnov test compares distributional differences and can miss differences in locations. By combining these two tests, matching can often be better assessed.

2.1 Comparison of matching methods via simulation

To further investigate the performance of various matching methods, a simulation with 5,000 iterations was conducted under various scenarios. Specifically, the simulation plan was designed as follows.

1. *Sample size:* two sets of sample size were used in simulation. The first set assumes equal sample size ($N = 50, 100, 250, 400, 450, 500$) for both experimental treatment and control groups. The second set also assumes these sample sizes for the experimental treatment group with the control group being about 20% smaller so that to mimic the different sample size allocations in many RCT and to study the effect of these methods with different sizes between samples.
2. Assume three covariates ($x_{i1}, x_{i2},$ and x_{i3}) will be matched between experimental treatment and control groups. The covariates were assumed to have somewhat different distributions between experimental treatment and control groups. Four different distributions are assumed. They consist of standard normal distributions with possibly different means and variances, or contaminated normal distributions with either symmetric or asymmetric contaminations from either tail. The list of distributions is shown in Table 1. In a separate simulation, we also include a binary covariate as suggested by a reviewer.
3. The response variable (Y) was assumed to follow two different models. The first model is

$$Y_i = \text{treatment effect} + x_{i1} + x_{i2} + x_{i3} + \text{error},$$

and the second model is

$$Y_i = \text{treatment effect} + x_{i1} + x_{i2} + x_{i3} + x_{i1} \times x_{i2} + x_{i1} \times x_{i2} + x_{i2} \times x_{i3} + \text{error}.$$

Table 1 Distributions of covariates simulated.

X_i	Group	F_1	F_2	F_3	F_4
x_{i1}	Treated	$N(0,1)$	$N(1,1)$	$0.9\,N(1,1) + 0.1N(1,3)$	$0.9\,N(1,1) + 0.1\,\lvert N(1,3)\rvert$
	Control	$N(0,1)$	$N(0,1)$	$0.9\,N(0,1) + 0.1N(0,3)$	$0.9\,N(0,1) + 0.1\,\lvert N(0,3)\rvert\,(-1)$
x_{i2}	Treated	$N(0,1)$	$N(0,2)$	$0.9\,N(0,2) + 0.1N(0,3)$	$0.9\,N(0,2) + 0.1N(0,3)$
	Control	$N(0,1)$	$N(1,2)$	$0.9\,N(1,2) + 0.1N(1,3)$	$0.9\,N(1,2) + 0.1N(1,3)$
x_{i3}	Treated	$N(0,1)$	$N(1,3)$	$0.9\,N(1,3) + 0.1N(1,4)$	$0.9\,N(1,3) + 0.1\,\lvert N(1,4)\rvert\,(-1)$
	Control	$N(0,1)$	$N(0,3)$	$0.9\,N(0,3) + 0.1N(0,4)$	$0.9\,N(0,3) + 0.1\,\lvert N(0,4)\rvert$

The treatment effect difference between experimental treatment and control groups is assumed to be a constant, e.g. 1. The purpose of assuming two different models is to compare these methods when the model is incorrectly specified. These two models were modified accordingly when the binary covariate was included.

4. The statistical methods to be compared are:
   (a) Empirical mean difference,
   (b) Least squares (LS) fit (assuming the first model is correct),
   (c) LS fit (assuming the second model is correct),
   (d) Matching using the propensity score based on $x_{i1}, x_{i2},$ and $x_{i3},$
   (e) Matching on $x_{i1}, x_{i2},$ and x_{i3} using all available data,
   (f) Matching on $x_{i1}, x_{i2},$ and $x_{i3},$ and the propensity score using all available data,
   (g) Matching on $x_{i1}, x_{i2},$ and x_{i3} but excluding data in either tail outside of two times MAD (MAD is defined as $1.483\ \text{med}_i\{\,|x_{iu} - \text{med}_j(x_{ju})|\}$) from the median for each covariate (to mimic Tukey's robust trimmed estimate),
   (h) Matching on $x_{i1}, x_{i2},$ and $x_{i3},$ and the propensity score but excluding data in either tail outside of two times MAD from the median for each covariate.
5. Two criteria for comparisons are examined:
   (a) The estimates of the true treatment effect and the variation of the estimates,
   (b) Balancing the covariates between experimental treatment and control groups. This will be assessed by examining the minimum p-value of the Kolmogorov–Smirnov test for the distributional equality of each covariate between experimental treatment and control groups before and after matching. Large p-values indicate greater comparability of the experimental treatment and control groups in terms of the covariates and hence reflect better covariate balance between groups.

2.2 Summary of simulation results

By examining the median, the inter-quartile distance, and the overall range of the box plots of the estimated treatment effects, we make the following conclusions:

1. The simple observed treatment difference can be a very poor estimate when the covariate distributions are different and deviate from standard normal distributions as shown in Panels 2 and 4 of Figures 1 and 2.
2. For the main effect model, the LS fit (when the model is correctly specified or even over-fitted with interaction terms) is generally better than other methods in estimating the treatment effect. But the LS fit with main effect only can perform poorly, if the true model includes interactions; however, the LS fit with interactions (correct model) outperforms other methods. This finding seems to be quite consistent among various sample sizes.
3. Matching purely based on propensity scores usually performs worse than Genetic matching either with all available data or with the trimmed dataset in estimating the true treatment effect. The trimmed estimate using Genetic matching to match both covariates and propensity scores performs almost uniformly better than any other method regardless of model specification and sample size, except for the LS fit when the model is correctly specified as discussed in eq. (2).
4. When the covariates of experimental treatment and control groups have identical normal distributions, the LS method outperforms all other methods since there is no need for matching. Additional effort to match seems to be redundant. The propensity score matching seems to make the covariate matching worse more often than not. However, the Genetic matching seems to perform reasonably well, especially when the outliers were trimmed away (Panel 1 of Figures 3 and 4).
5. However, when the covariate distributions are different between experimental treatment and control groups and deviate from the standard normal, the effect of matching from all methods becomes very

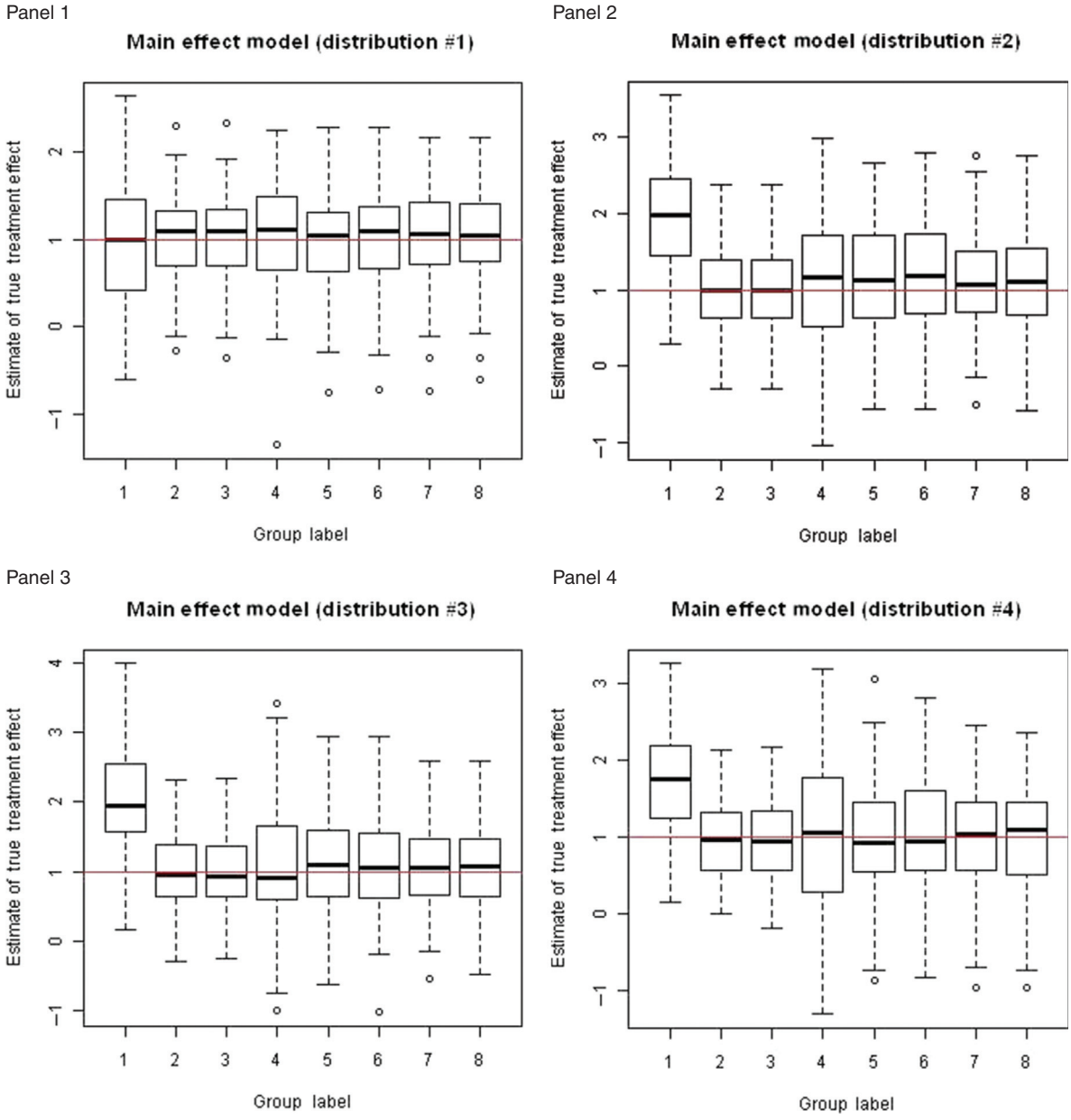

Figure 1 Estimation of true treatment effect (=1). Panel 1: main effect model with distribution #1 (group labels 1–8 correspond to methods a–h of Section 2.1). Panel 2: main effect model with distribution #2 (group labels 1–8 correspond to methods a–h of Section 2.1). Panel 3: main effect model with distribution #3 (group labels 1–8 correspond to methods a–h of Section 2.1). Panel 4: main effect model with distribution #4 (group labels 1–8 correspond to methods a–h of Section 2.1).

visible. This can be seen in Panels 2–4 of Figures 3 and 4. Genetic matching with trimmed outliers tends to outperform all other methods either matched only on all covariates or with propensity score included. This is true for all distributions and sample sizes tested here.

6. The sample size in Figures 1 and 2 was 450, and in Figures 3 and 4, it was 400. The patterns for other sample sizes are similar with somewhat higher variations for smaller sample sizes, therefore, are not shown here.

As discussed above, when the model is correctly specified, the simple LS method outperforms other methods as expected. However, as in most of the data analysis, one rarely knows the correct model or

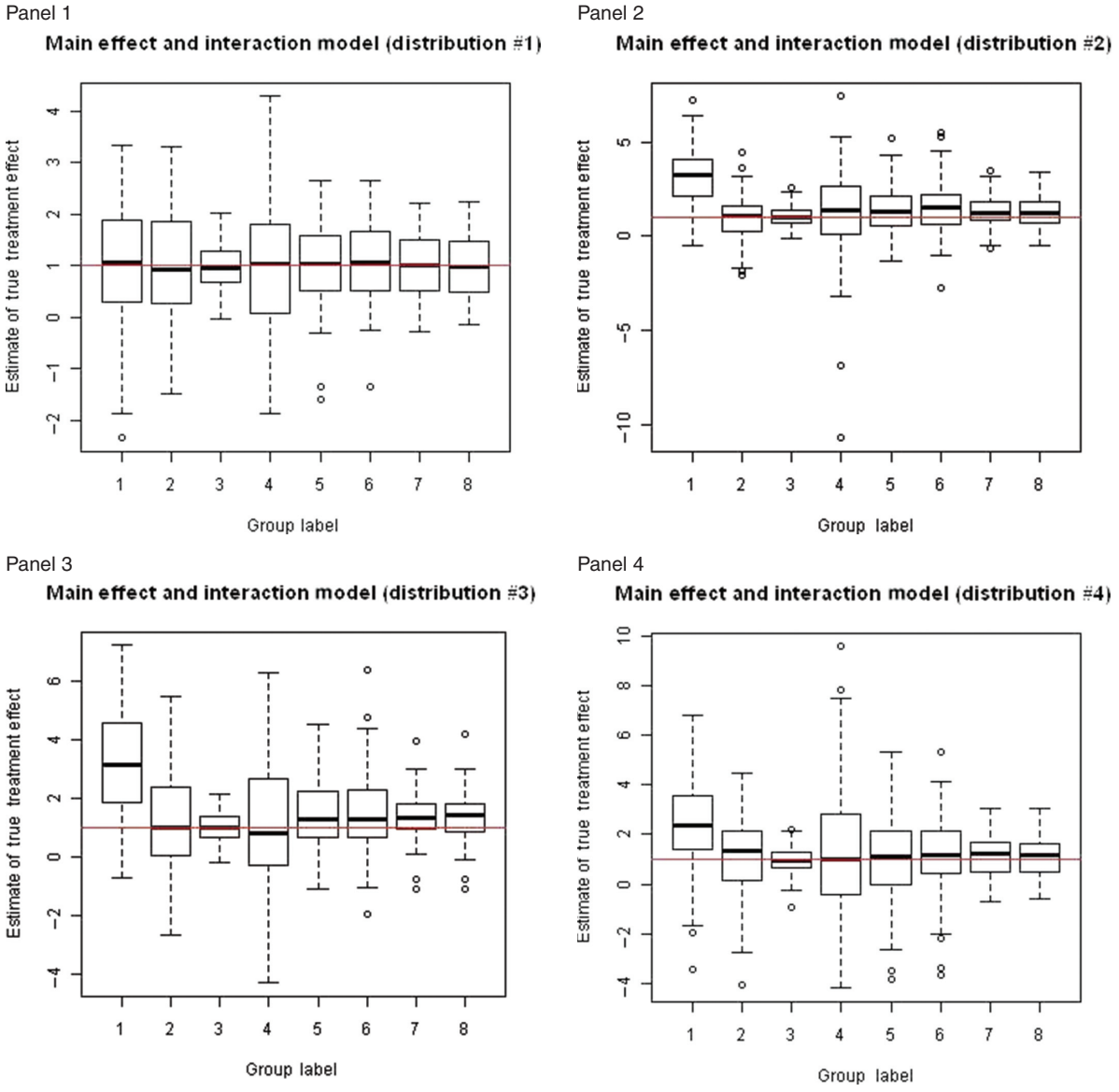

Figure 2 Estimation of true treatment effect (= 1). Panel 1: main effect and interaction model with distribution #1 (group labels 1–8 correspond to methods a–h of Section 2.1). Panel 2: main effect and interaction model with distribution #2 (group labels 1–8 correspond to methods a–h of Section 2.1). Panel 3: main effect and interaction model with distribution #3 (group labels 1–8 correspond to methods a–h of Section 2.1). Panel 4: main effect and interaction model with distribution #4 (group label 1–8 correspond to method a–h of Section 2.1).

the distribution from which the data was generated. Therefore, the performance of LS method can sometimes be expected to diminish in the analysis of real data. On the other hand, the performance of Genetic matching seems to be almost always comparable to the LS method when the model is correctly specified and performs much better when the model is mis-specified as shown in Panels 1, 2, and 4 of Figure 2. Therefore, the Genetic matching seems to serve as a "model mis-specification proof" tool for general data analysis. It is interesting to note that Diamond et al. [18] also similarly concluded that Genetic matching is preferred over other matching methods, because it is more efficient (smaller MSE) and is less biased.

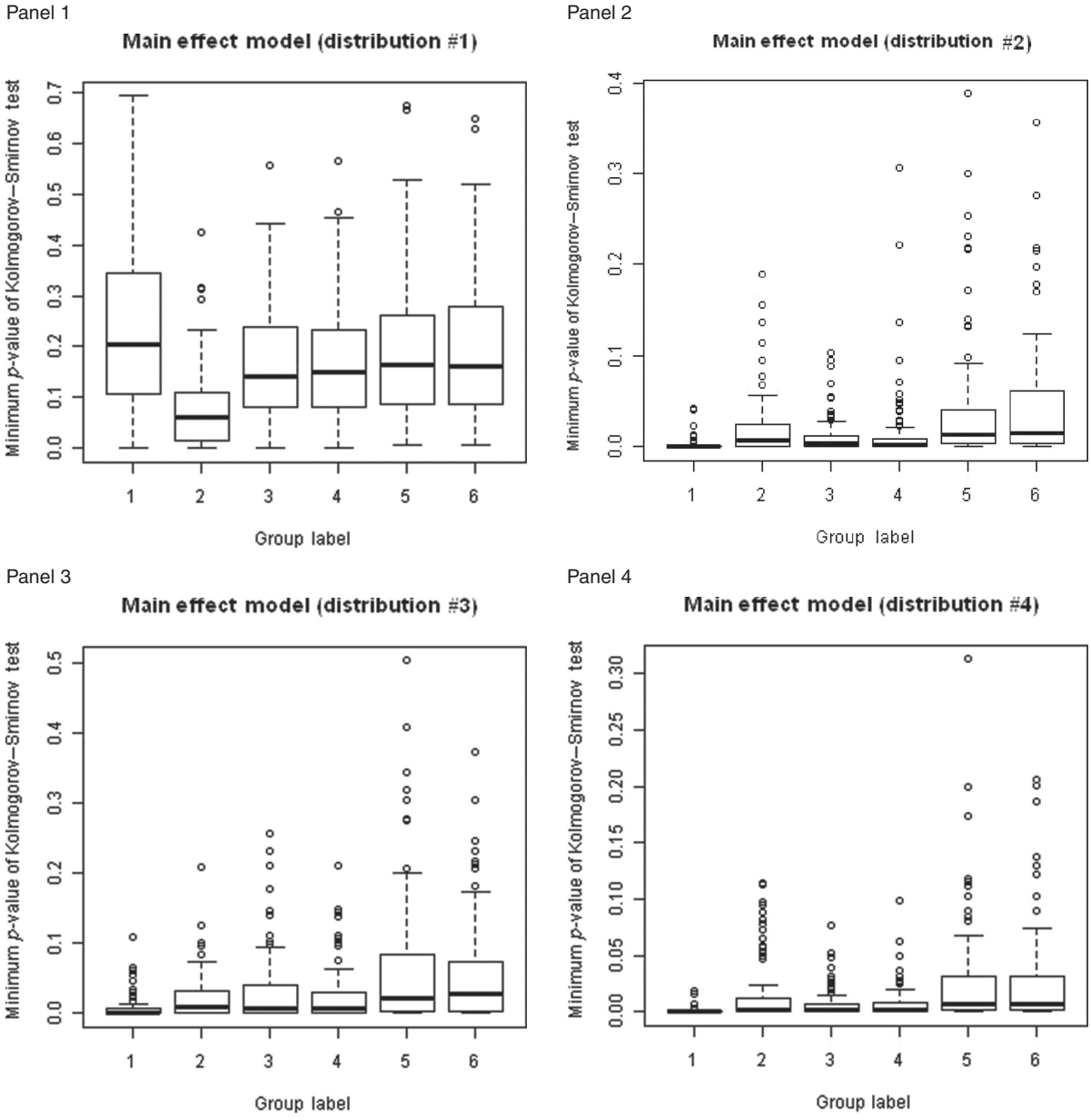

Figure 3 Minimum *p*-value of K–S test to compare the equality of covariates. Panel 1: main effect model with distribution #1 (group label 1 is before matching, 2–6 correspond to post-matching of methods d–h of Section 2.1). Panel 2: main effect model with distribution #2 (group label 1 is before matching, 2–6 correspond to post-matching of methods d–h of Section 2.1). Panel 3: main effect model with distribution #3 (group label 1 is before matching, 2–6 correspond to post-matching of methods d–h of Section 2.1). Panel 4: main effect model with distribution #4 (group label 1 is before matching, 2–6 correspond to post-matching of methods d–h of Section 2.1).

3 Estimation of confidence interval of treatment effect

As discussed by Lachin [9] and other researchers, in most clinical trial practices, participants are actually more of a convenient sample than a truly randomized sample from the intended population with a specific disease for treatment. After a group of study subjects has been recruited, trialists then give their best efforts to randomly assign subjects to treatments. That is one of the primary reasons the randomization model is

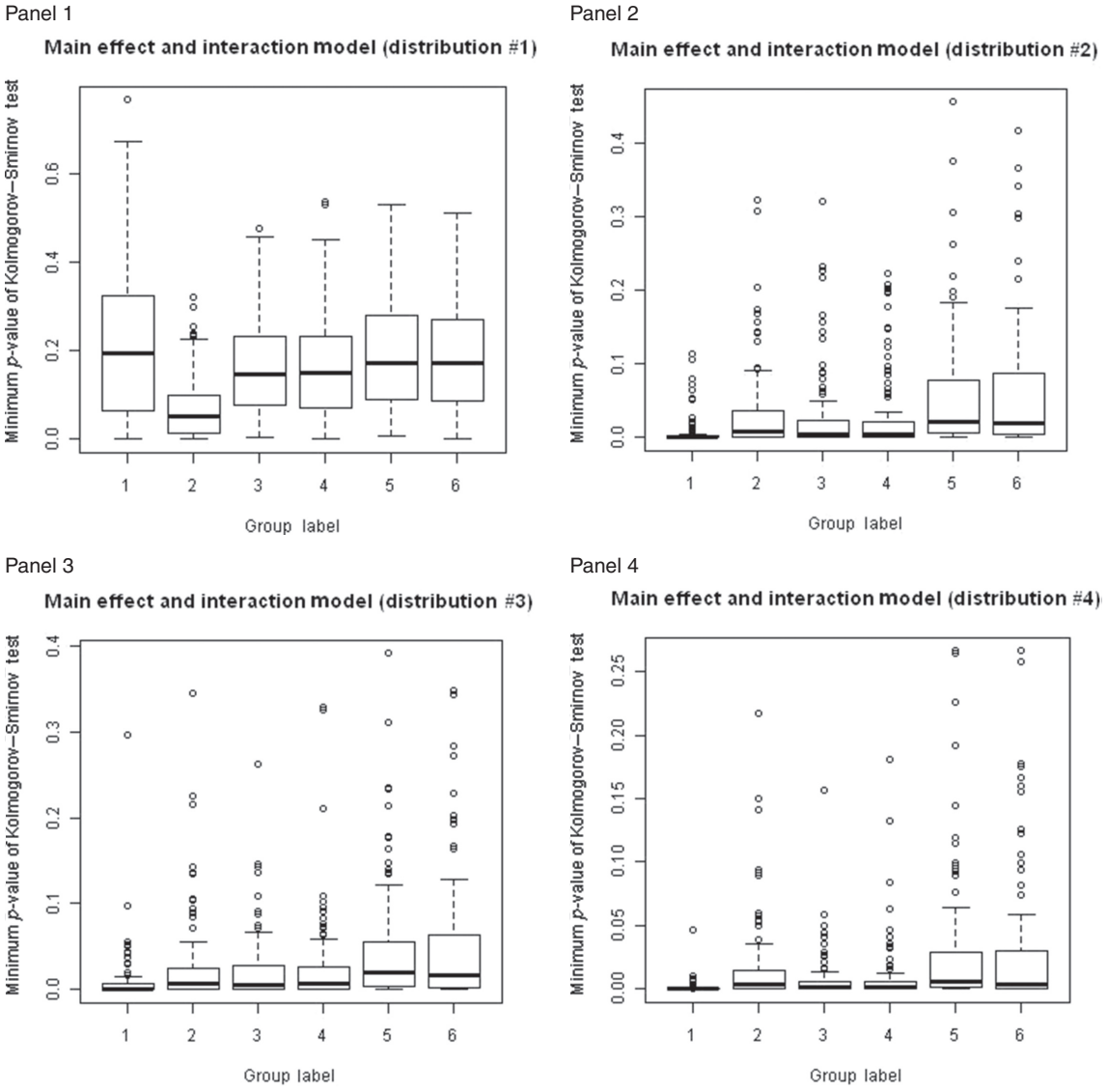

Figure 4 Minimum p-value of K–S test to compare the equality of covariates. Panel 1: main effect and interaction model with distribution #1 (group label 1 is before matching, 2–6 correspond to post-matching of methods d–h of Section 2.1). Panel 2: main effect and interaction model with distribution #2 (group label 1 is before matching, 2–6 correspond to post-matching of methods d–h of Section 2.1). Panel 3: main effect and interaction model with distribution #3 (group label 1 is before matching, 2–6 correspond to post-matching of method d–h of Section 2.1). Panel 4: main effect and interaction model with distribution #4 (group label 1 is before matching, 2–6 correspond to post-matching of method d–h of Section 2.1).

preferred to the population model by these researchers for statistical inferences. Since subgroups, defined either pre- or post-randomization, also inherit these properties, randomization model seems to be a natural choice for inferences.

It is a common statistical practice to accompany the point estimate of treatment effect with the corresponding confidence interval so that the magnitude of the effect can be better judged by clinical practitioners. Lachin suggested using randomization model to estimate the treatment effect and invoking the concept of population model to estimate the confidence interval. As an alternative, one can use the stochastic approximation as proposed by Robbins and Monro [19] and implemented by Garthwaite [20] to estimate the confidence interval of the treatment effect. Briefly, for treatment effect θ_0, a randomization test is performed to test the hypothesis $H_0 : \theta = \theta_0$ against both one-sided alternatives $H_1 : \theta < \theta_0$ and $H_1 : \theta > \theta_0$.

A separate search is performed for each endpoint of the corresponding confidence interval. The upper and lower endpoints of the confidence interval are updated according to an algorithm after every randomization test. The asymptotic property of the search process is discussed by Garthwaite and Buckland [21]. In addition, under weak regularity conditions, the estimates converge in probability to the correct confidence limits [22].

4 Example

A phase III, multi-national randomized, double blind, placebo-controlled clinical trial was conducted by a pharmaceutical company to compare the treatment effect of drug A and drug B to placebo in controlling disease activity in subjects with rheumatoid arthritis having an inadequate clinical response to methotrexate. (Due to the restriction of the data provider, the names of drug A and drug B are not revealed.) The study was not originally designed to compare drug A and drug B directly. However, a post hoc analysis to compare these two drugs in a subgroup of countries of the original study is of clinical interest. A total of 156 and 165 patients were randomized to drugs A and B in these countries, respectively. The primary endpoint of the study was the disease activity score based on 28 joints (DAS28).

Comparisons of several baseline covariates using the *t*-test did not show particular imbalance between the two treatment groups. However, a more in-depth investigation of the baseline distributions by quantile–quantile (Q–Q) plots showed some deviations between the two populations. The objective in this analysis is to properly estimate the treatment difference under the situation of baseline imbalance.

The first step in this analysis is to match the patients from drugs A and B. Both the propensity score and the Genetic matching methods were applied with the covariates including age, baseline pain score, baseline CRP, and other components of DAS28, so that we can compare the relative performance of these two matching methods. As an example, the baseline pain scores between the treatment groups are compared and shown in Figure 5. The original Q–Q plot of pain scores between drug A and drug B is shown in Panel 1. The Q–Q plots of this covariate using propensity score matching and Genetic matching are shown in Panels 2 and 3, respectively. One can clearly see substantial improvement in covariate balance of Genetic matching over the propensity score matching.

Permutation distributions of the treatment effect before and after Genetic matching were also generated and are shown in Figure 6. The observed treatment difference prior to matching is about −0.19. However, the magnitude of the treatment difference was reduced substantially to −0.048 after matching. This indicates the importance of the proper matching of patients in the two treatment groups. Without this step, the treatment difference may potentially be over-estimated. Even though the permutation test did not show a significant treatment difference in either pre- or post-matching, the test prior to matching had a higher significance level than after matching.

The 95% confidence interval of the treatment effect difference was estimated using the procedure described previously. A total of 5,000 randomized samples were generated and analyzed. The estimates fluctuate substantially in the beginning of the approximation process. The process began to stabilize after about 2,500 randomizations. Figure 7 shows the stochastic approximation for the upper and lower limits of the confidence interval. The resulting 95% confidence interval is (−0.110, 0.4858).

5 Discussion

Subgroup data analysis is common practice in medical research in order to better understand or explore treatment effects in different groups of patients. This is an important step toward individualized medicine. However, how to do it properly to get a more or less unbiased (since no one knows what the truth is) treatment effect is a difficult task, especially when the data are not appropriately randomized or only

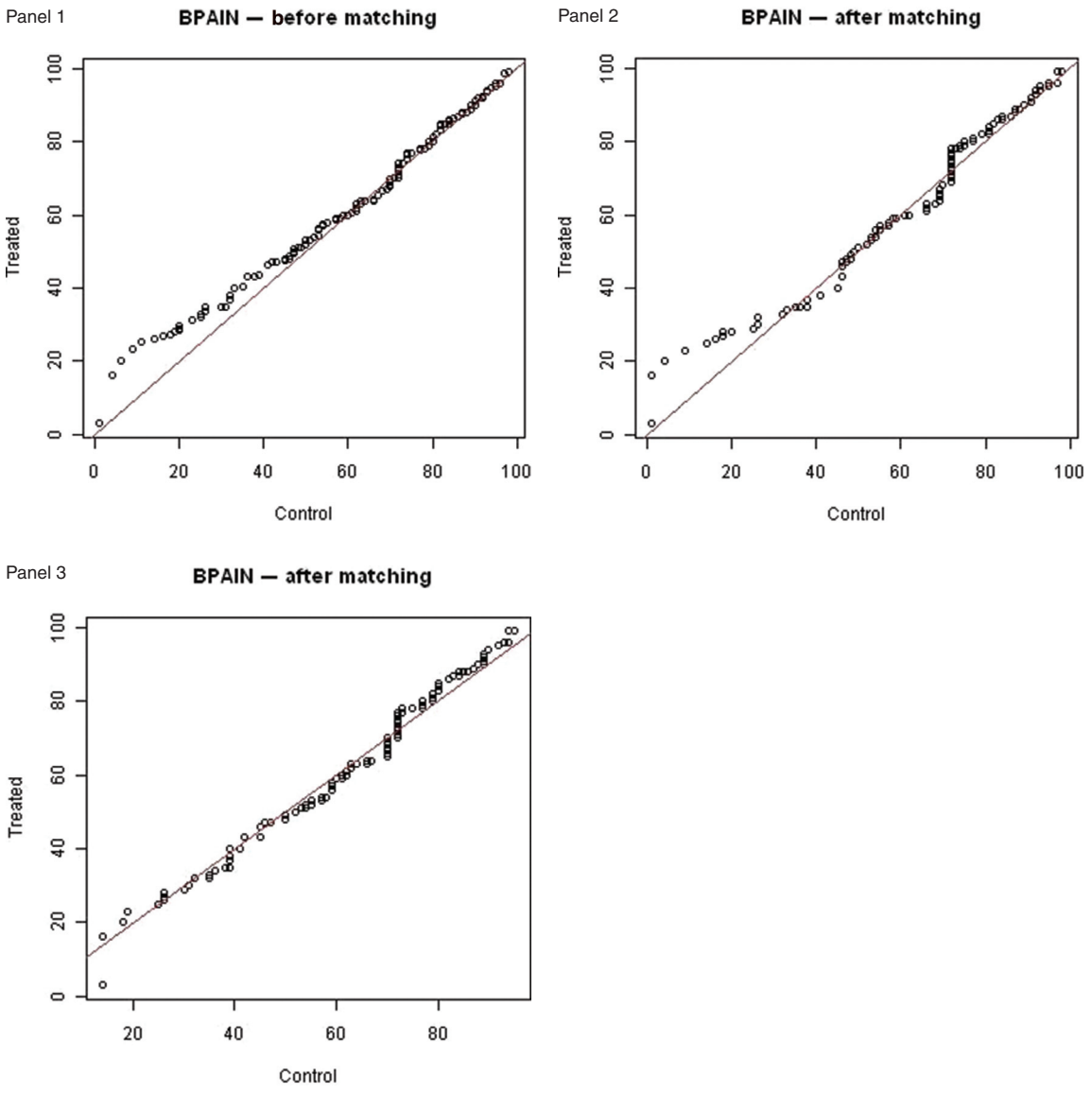

Figure 5 Balancing baseline pain score via propensity score and Genetic matching. Panel 1: original Q–Q plot without any matching. Panel 2: after propensity score matching. Panel 3: after Genetic matching.

observational. Researchers have proposed several classes of methods to analyze this kind of data (e.g. see [23] and the discussions therein) and matching methods, based on propensity score, Mahalanobis matching, Genetic matching, and their variants, are among the important tools for this purpose. Particularly, Genetic matching provides the extra flexibility of weighting the selected covariates for the desired matching, so that the treatment effect estimate can also reflect the preferences of the investigators. With a good matching between the experimental treatment and control subjects and a higher degree of association between the covariates and the response, the treatment effect can be more accurately estimated.

In this manuscript, we conducted a simulation to further look into the performance of these methods under various scenarios with respect to their ability to better balance the covariates between the experimental treatment and control groups and also to produce unbiased estimate of treatment effect. The methods we compared ranged from the usual linear regression, conventional matching techniques with all available data to more robust alternatives, which exclude possible outliers. In general, Genetic matching is preferred to other methods under various data distributions of the covariates and various sample sizes.

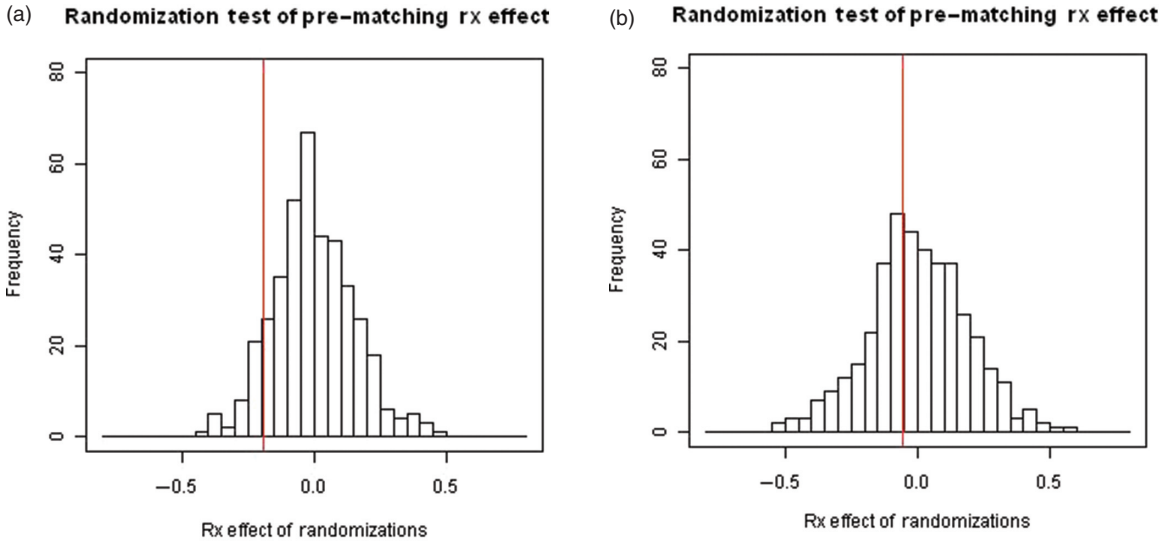

Figure 6 Comparison of pre- and post-Genetic matching estimates of treatment effects after 5,000 permutation simulations.

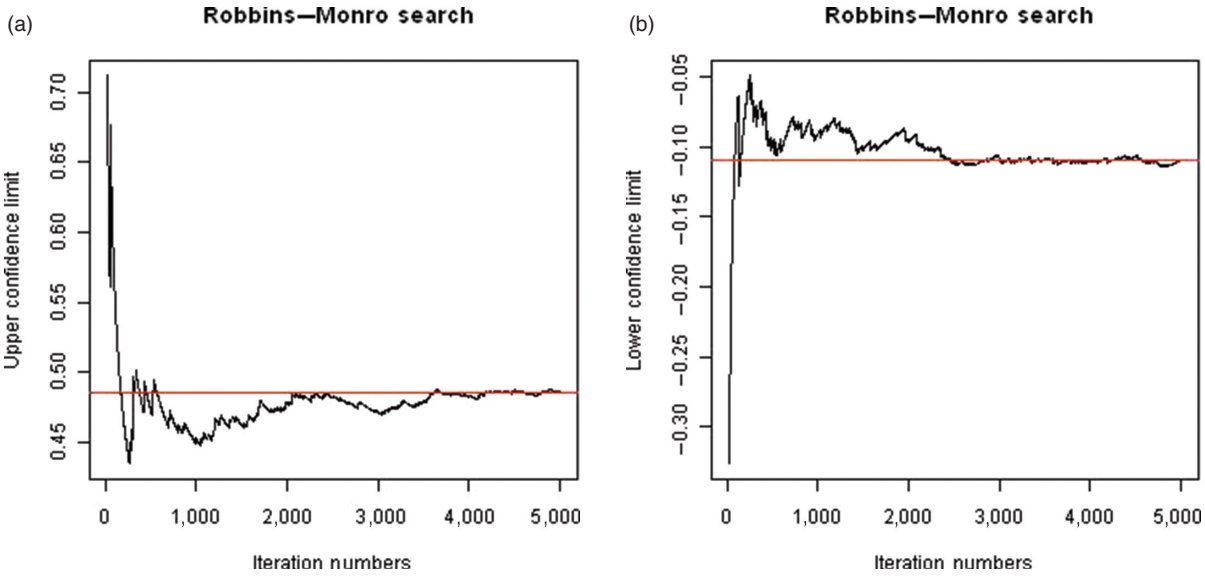

Figure 7 Stochastic approximation of the 95% confidence interval of treatment effect difference (based on 5,000 simulated randomizations).

Variable selection to be used in these procedures is an important point to consider. Several authors have proposed various approaches to incorporate covariates to estimate the propensity score [24–26]. The general findings are to incorporate variables which are thought to be related to outcomes, and variables thought to be confounded with both treatment assignment and outcomes. The model which incorporates as many covariates as possible and the model which includes obvious covariates such as age, gender, and race do not always seem to perform well. One should note that pre-randomization variables will not be confounded with treatment assignment in RCT and a successful randomization process is likely to correct for both the known and the unknown confounders. However, under the scenario of possible missing confounding variables, known or unknown, compounded with possible covariate imbalance, the performance of

matching methods relative to other approaches is still not well-understood, therefore, it will be further researched and reported in the future.

It is generally recommended that careful examination of covariate balance between treatment groups be conducted prior to statistical inferences. Besides the formal test procedures, graphical methods can quite often reveal the subtle data details which are not easily detected in test procedures. In addition to the treatment effect estimate, it is more informative to provide readers with a confidence interval that brackets the estimate, which can be quite useful for the clinicians to gauge the clinical significance of the treatment effect. Toward this purpose, among other approaches, one can use the Robbins–Monro stochastic approximation to estimate the confidence interval of the treatment effect difference (the R-program is available from the authors). One may notice that the confidence interval is asymmetric to the estimate in our data example. The stochastic approximation estimates the upper and lower bounds of the confidence interval separately by comparing the randomization test statistics using the original and re-randomized data, which is different from the population model approach. Hence, the asymmetry may be part of the intrinsic properties of the combination of randomization test and stochastic approximation; however, a more detailed investigation of this phenomenon seems to be worthwhile.

Toward the goal of individualized medicine, medical research, and practices often use the multi-stage therapeutic strategies, for example, the dynamic treatment regimes (DTRs), in which dose or treatment is modified at each stage according to a patient's current history, disease status, and response to the most recent treatment in the testing of experimental treatments for serious diseases such as cancer or psychiatric problems. Statistical research in design and analysis of studies aimed at evaluating the effects of these strategies also has an active history, for example, Zelen [27] and Wei and Durham [28] on the play-the-winner rule; Lavori and Dawson [29, 30], Thall et al. [31], Murphy [32], Oetting et al. [33], on the designs of randomized trials that aim at the evaluation of DTRs; Robins [34–37] on the g-estimation of structural nested models; and Murphy [38], Robins [39], and Moodie et al. [40] on the optimal treatment regime estimation. As explained by Moodie et al., the inferences in Robins and Murphy are based on the difference between the empirical and the counterfactual observations, which is also utilized in our proposed method. Given the number of treatments in DTRs and the usually moderate number of subjects, their methods utilized the parametric or semi-parametric method for more efficient estimation and modeling. Even though they could have employed subject matching as we have done here, the moderate trial size may post a severe limitation. Alternatively, Zhao et al. [41] proposed a non-parametric individualized treatment rule using outcome weighting learning which circumvents the need for conditional mean modeling, the counterfactual assumptions, and essentially turning the optimal treatment selection into a weighted classification problem using SVM techniques. Their individual treatment rule assigns treatments to each subject only based on subject's prognostic information. Presumably, they could have modified their weighting schemes to increase the flexibility, as provided by the Genetic matching, to adjust the weights preferred by the investigators beyond the prognostic factors.

The ultimate goal of these researches is to use the results of the trials as a basis for generating hypotheses and planning a future, larger scale confirmatory trial, and to tailor the specific treatments for patient subgroups with specific characteristics in diseases as well as genome and biomarker profiles, so that potentially better responses can be achieved. As more trials with more subjects have been conducted based on the best knowledge accumulated from these experimental findings, inevitably, additional questions will be raised to further identify and compare the patient subgroups with respect to treatment efficacy and adverse effects. The subgroup analysis method proposed in this article can become an important and handy tool to perform rigorous post hoc analysis to further understand the inner-workings of the new treatments. Therefore, the integration of the multi-state therapeutic strategies and careful post hoc subgroup analysis can become an important effort for medical advancement.

Even though the large-scale randomized controlled trials are generally considered as the gold standard to generate convincing clinical information, one should not underestimate the importance of subgroup analysis. One cannot ignore that concerns exist in conducting and reporting subgroup analysis, and problems persist even with the CONSORT [42] and ICH [43], guidance; however, when properly planned,

reported, and interpreted, subgroup analysis can provide valuable information. In some clinical trial settings, subgroup analysis can also be among the primary objectives. For example, the FDA had granted marketing approval for Pemetrexed plus Cisplatin [44, 45] to be used to treat non-small cell lung cancer patients with non-squamous even though the entire study did not show significance in overall survival.

Acknowledgments: The authors would like to thank the Editor and all the reviewers for their insightful comments. Their inputs substantially improve the quality of this manuscript.

References

1. Peto R, Collins R, Gray R. Large-scale randomized evidence: large, simple trials and overviews of trials. J Clin Epidemiol 1995;48:23–40.
2. Wang R, Lagakos S, Ware J, Hunter D, Drazen J. Statistics in medicine – reporting of subgroup analyses in clinical trials. New Engl J Med 2007;357:2189–94.
3. Brookes ST, Whitely 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. J Clin Epidemiol 2004;57:229–36.
4. Yusef S, Collins R, Peto R. Why do we need some large, simple randomized trials? Stat Med 1984;3:409–22.
5. Feinstein AR. The problem of cogent subgroups: a clinic statistical tragedy. J Clin Epidemiol 1998;51:297–9.
6. Tukey JW. The future of data analysis. Ann Math Stat 1962;33:13–14.
7. Rothwell PM. Subgroup analysis in randomized controlled trials: importance, indications, and interpretation. The Lancet 2005;365:176–86.
8. Efron B. Forcing a sequential experiment to be balanced. Biometrika 1971;58:403–17.
9. Lachin JM. Statistical properties of randomization in clinical trials. Control Clin Trials 1988;9:289–311.
10. Wright J, Dunn J, Cutler J, Davis B, Cushman W, Ford C, et al. Outcomes in hypertensive black and non-black patients treated with chlorthalidone, amlodipine, and lisinopril. New Engl J Med 2005;293:595–1607.
11. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;79:516–24.
12. Rubin DB. Estimating causal effects of treatments in randomized and non-randomized studies. J Educ Psychol 1974;66:688–701.
13. Rubin DB. Assignment to a treatment group on the basis of a covariate. J Educ Stat 1977;2:1–26.
14. Rosenbaum PR. Observational studies. New York: Springer-Verlag, 1995.
15. Landwehr JM, Pregibon D, Shoemaker AC. Graphical methods for assessing logistic regression models. J Am Stat Assoc 1984;79:61–71.
16. Tsai KT. Assessing regression modeling with ordinal responses. Presentation at the Joint Statistical Meetings of the American Statistical Association, 2008.
17. Sekhon JS. Alternative balance metrics for bias reduction in matching methods for causal inference. Working paper, 2006. Available at: http://sekhon.berkeley.edu/papers/SekhonBalanceMetrics.pdf
18. Diamond A, Sekhon JS. Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies. Berkeley, CA: Institute of Governmental Studies, University of California, 1996. Available at: http://escholarship.org/uc/item/8gx4v5qt
19. Robbins H, Monro S. A stochastic approximation method. Ann Math Stat 1951;22:400–07.
20. Gartwaite PH. Confidence intervals from randomization tests. Biometrics 1996;52:1387–93.
21. Gartwaite PH, Buckland ST. Generating Monte Carlo confidence intervals by the Robbins-Monro process. Appl Stat 1992;41:159–71.
22. Blum JR. Approximation methods that converge with probability one. Ann Math Stat 1954;25:390–4.
23. Shadish WR, Clark MH, Steiner PM. Can non-randomized experiments yield accurate answers? A randomized experiment comparing random and nonrandomized assignments. J Am Stat Assoc 2008;103:1334–56.
24. Rubin DB, Thomas N. Matching using estimated propensity score: relating theory to practice. Biometrics 1996;52:249–64.
25. Rubin DB. Estimating causal effects from large data sets using the propensity score. Ann Intern Med 1997;127:757–63.
26. Brookhart MA, et al. Variable selection for propensity score models. Am J Epidemiol 2006;163:1149–56.
27. Zelen M. Play the winner rule and the controlled clinical trial. J Am Stat Assoc 1969;64:131–46.
28. Wei LJ, Durham S. The randomized play-the-winner rule in medical trials. J Am Stat Assoc 1978;73:840–3.
29. Lavori P, Dawson RA. Design for testing clinical strategies: biased individually tailored within-subject randomization. J R Stat Soc Ser A 2000;163:29–38.

30. Lavori P, Dawson R. Dynamic treatment regimes: practical design considerations. Clin Trials 2004;1:9–20.
31. Thall P, Sung H, Estey E. Selecting therapeutic strategies based on efficacy and death in multi-course clinical trials. J Am Stat Assoc 2002;97:29–39.
32. Murphy S. An experimental design for the development of adaptive treatment strategies. Stat Med 2005;24:1455–81.
33. Oetting A, Levy R, Weiss S, Murphy S. Statistical methodology for a SMART design in the development of adaptive treatment strategies. In: Shrout PE, editor. Causality and psychopathology: finding the determinants of disorders and their cures. Arlington, VA: American Psychiatric Publishing, 2011:175–205.
34. Robins J. A new approach to causal inference in mortality studies with sustained exposure periods – application to control of the healthy survivor effect. Math Model 1986;7:1393–512.
35. Robins J. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In: Sechrest L, Freeman H, Mulley A, editors. Health service research methodology: a focus on AIDS. Rockville, MD: U.S. Public Health Service, 1989:113–59.
36. Robins J. Analytic methods for estimating HIV treatment and cofactor effects. In: Ostrow G, Kessler R, editors. Methodological issues of AIDS mental health research. New York: Plenum Publishing, 1993:213–90.
37. Robins J. Causal inference from complex longitudinal data. In: Berkane M, editor. Latent variable modeling and applications to causality. New York: Springer-Verlag, 1997:69–117.
38. Murphy S. Optimal dynamic treatment regimes (with discussion). J R Stat Soc Ser B 2003;65:331–66.
39. Robins J. Optimal structural nested models for optimal sequential decisions. In: Lin DY, Heagerty P, editors. Proceedings of the second Seattle symposium on biostatistics. New York: Springer, 2004:189–326.
40. Moodie EE, Richardson TS, Stephens DA. Demystifying optimal dynamic treatment regimes. Biometrics 2007;63:447–55.
41. Zhao Y, Zeng D, Rush A, Kosorok M. Estimating individualized treatment rules using outcome weighted learning. J Am Stat Assoc 2012;107:1106–18.
42. Moher D, Schulz KF, Altman DG, et al. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. Available at: http://www.consort-statement.org/. Accessed: 1 Nov 2007.
43. International Conference on Harmonisation (ICH). Guidance for industry: E9 statistical principles for clinical trials. Rockville, MD: Food and Drug Administration, September 1998. Available at: http://www.fda.gov/cder/guidance/ICH-E9-fnl.PDF. Accessed: 1 Nov 2007.
44. Scagliotti G, Parikh P, von Pawel J, Biesma B, Vansteenkiste J, Manegold C, et al. Phase III study comparing Cisplatin plus Gemcitabine with Cisplatin plus pemetrexed in chemotherapy-naive patients with advanced-stage non-small cell lung cancer. J Clin Oncol 2008;26:3543–51.
45. HemOncToday. FDA approved pemetrexed plus cisplatin for nonsquamous NSCLC. 2008.