Published in final edited form as:

Biom J. 2019 May ; 61(3): 698–713. doi:10.1002/bimj.201800049.

K-Sample Comparisons using Propensity Analysis

Sin-Ho Jung¹, Sang Ah Chi², and Hyun Joo Ahn³

¹Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, U.S.A.

²Department of Health Sciences and Technology, SAIHST, Sungkyunkwan University, Seoul, South Korea

³Department of Anesthesiology and Pain Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, South Korea

Abstract

In this paper, we investigate K -group comparisons on survival endpoints for observational studies. In clinical databases for observational studies, treatment for patients are chosen with probabilities varying depending on their baseline characteristics. This often results in non-comparable treatment groups because of imbalance in baseline characteristics of patients among treatment groups. In order to overcome this issue, we conduct propensity analysis and match the subjects with similar propensity scores across treatment groups or compare weighted group means (or weighted survival curves for censored outcome variables) using the inverse probability weighting (IPW). To this end, multinomial logistic regression has been a popular propensity analysis method to estimate the weights. We propose to use decision tree method as an alternative propensity analysis due to its simplicity and robustness. We also propose IPW rank statistics, called Dunnett-type test and ANOVA-type test, to compare 3 or more treatment groups on survival endpoints. Using simulations, we evaluate the finite sample performance of the weighted rank statistics combined with these propensity analysis methods. We demonstrate these methods with a real data example. The IPW method also allows us for unbiased estimation of population parameters of each treatment group. In this paper, we limit our discussions to survival outcomes, but all the methods can be easily modified for any type of outcomes, such as binary or continuous variables.

Keywords

ANOVA; Decision tree; Dunnett test; Inverse probability weighting; Multinomial logistic regression

1 Introduction

In a prospective study, such as a phase III clinical trial, patients are randomly assigned to different treatment groups independently of the baseline characteristics, also called predictors, so that the distribution of the predictors are well balanced among treatment groups and the statistical testing to compare the efficacy of the treatments controls the type I

error rate accurately even without adjusting for the predictors. We often use a stratified randomization method for a perfect balance of the predictor distributions among treatment groups by selecting some important predictors as the stratification factors.

In the data set of an observation study, however, treatment group is usually confounded with predictors. In this case, we may conduct a multivariable regression analysis including the group identifier and predictors as covariates to compare the groups adjusting for the potential bias due to confounding. In the presence of confounding, clinical investigators do not feel so comfortable with a comparison among treatment groups, so that they want to generate a subset of data with a balanced predictor distributions among treatment groups as if they were from a randomized trial.

In non-randomized (observational) studies, the use of propensity analysis has received much attention in clinical research to compare clinical endpoints between groups. A propensity score is a measure of how the treatment of each patient is selected depending on the values of predictors. If we can find equal number of patients from each group with the same propensity score, then we can make up a perfectly balanced data set, called matched data, by selecting only those patients for an efficacy analysis. With matched data, we can conduct a K -sample comparison using a standard univariate analysis method, but we often have to discard a large part of data set during the matching procedure. An alternative approach is to use a weighted test statistic for comparing groups with inverse probability weighting (IPW) by keeping all the original data, e.g. Curtis et al. (2007) and Breslow et al. (2009).

There have been plenty of publications about propensity method since the original works by Rosenbaum and Rubin (1983, 1984). Most of these publications, however, are focused on comparing two patient groups from non-randomized studies and its extension to $K(=3)$ has not been fully investigated yet, especially for survival analysis. Logistic regression method has been popularly used in propensity analysis for two treatment groups. Use of multinomial logistic regression method was proposed by Rubin (1998) in the estimation of propensity scores for K -sample comparison problems, and has been studied by many investigators including Imbens (2000) and McCaffrey et al. (2013). In a K -sample comparison, accurate estimation of propensity scores is a key component. If treatment selection probabilities have a non-monotone trend on a predictor, it is not easy to accurately estimate the propensity scores using a model-based regression method. To overcome this issue, some investigators (Stone et al. 1995; Pruzek & Cen 2002; Westreich et al. 2010) have applied classification tree method for propensity analysis as an alternative to the logistic regression method.

McCaffrey et al. (2013) propose to repeatedly apply the 2-sample tree method between each experimental treatment group and a common control, and adjust the $K-1$ dimensional propensity scores for K sample comparisons. But, they report that the K sample comparison result can be different depending on which group to be chosen as the control. We propose a decision tree method for a propensity analysis of K sample comparison through a direct classification of K treatment groups not requiring specification of a control group. We use a pruning method based on χ^2 -tests for testing if two propensity strata have different probabilities of treatment selection or not.

We also propose two types of weighted rank statistics to compare K groups on survival outcomes, using ANOVA-type and Dunnett-type (1955) testing methods combined with IPW. We evaluate the performance of the weighted testing methods combined with two propensity analysis methods to estimate the weights, multinomial logistic regression method and decision tree method, through simulations. We apply the proposed IPW method and a matched data method using the propensity scores estimated by the two propensity analysis methods to a real data example. We limit our discussions to survival outcome, but the results can be easily modified for other types of variables, such as continuous or binary.

2 Propensity Analysis

Suppose that there are n_k patients for treatment group $k (= 1, \dots, K)$ and $n = \sum_{k=1}^K n_k$. From each patient, we have data on group identity, m covariates $(z_{1i}, \dots, z_{mi})^T$, and a survival outcome. In the first step of propensity analysis, we estimate propensity scores using the predictors and groups identity. Outcome data are not used in this step.

2.1 Multinomial Logistic Regression Method

At first, we review the multinomial logistic regression method that was proposed by Rubin (1998). Let $(P_{1i}, \dots, P_{Ki})$ be the propensities of K groups for patient $i (= 1, \dots, n)$ with predictors $z_i = (1, z_{1i}, \dots, z_{mi})^T$. We consider logistic models

$$\log \frac{P_{ki}}{P_{Ki}} = \beta_k^T z_i = \beta_{0k} + \beta_{1k} z_{1i} + \dots + \beta_{mk} z_{mi} \quad (1)$$

for $k = 1, \dots, K-1$, where $\beta_k = (\beta_{0k}, \beta_{1k}, \dots, \beta_{mk})$ is a vector of regression coefficients. For each $k (= 1, \dots, K-1)$, β_k is estimated by fitting the logistic regression (1) using data from groups k and K as in a propensity analysis for a two-sample comparison. For $i = 1, \dots, n_k + n_K$, the binary response variable is $y_i = 1 (= 0)$ if patient i belongs to group k (group K).

Since $\sum_{k=1}^K P_{ki} = 1$ and $P_{ki} = P_{Ki} \exp(\beta_k^T z_i)$ for $k = 1, \dots, K-1$ from (1), we have

$$P_{Ki} = \frac{1}{1 + \sum_{k=1}^{K-1} \exp(\beta_k^T z_i)}$$

and

$$P_{ki} = \frac{\exp(\beta_k^T z_i)}{1 + \sum_{k'=1}^{K-1} \exp(\beta_{k'}^T z_i)} \quad \text{for } k = 1, \dots, K-1.$$

For $k = 1, \dots, K-1$, let $\hat{\beta}_k$ denote the estimator of β_k from the regression analysis between treatment groups k and K .

Now, we rearrange the data so that $\tilde{z}_{ki}$ denotes the covariate vector for patient $i (= 1, \dots, n_k)$ belonging to treatment group $k (= 1, \dots, K)$. Using this notation, the allocation probability for this patient is estimated by

$$\hat{P}_{ki} = \frac{\exp(\hat{\beta}_k^T \tilde{z}_{ki})}{1 + \sum_{k'=1}^{K-1} \exp(\hat{\beta}_{k'}^T \tilde{z}_{ki})} \quad \text{if } 1 \leq k \leq K-1$$

and

$$\hat{P}_{Ki} = \frac{1}{1 + \sum_{k'=1}^{K-1} \exp(\hat{\beta}_{k'}^T \tilde{z}_{Ki})}.$$

We use $\hat{P}_{ki}$ as a propensity score. For a propensity matching, we may define strata by grouping patients with similar values for $(\hat{P}_{1i}, \dots, \hat{P}_{K-1,i})$ using a clustering method. A simple clustering method may be, for example with $K=3$, to partition the ranges of $\hat{P}_{1i}$ and $\hat{P}_{2i}$ from n patients into J_1 and J_2 intervals, respectively, and construct $J_1 \times J_2$ strata, each of which consists of a stratum of patients with similar propensity scores $(\hat{P}_1, \hat{P}_2, \hat{P}_3)$. Within each stratum, we can randomly select a certain number of patients from each of the three groups so that the set of allocation proportions $(\gamma_1, \gamma_2, \gamma_3)$ with $\sum_{k=1}^3 \gamma_k = 1$ is identical across the $J_1 \times J_2$ strata. This procedure is called vector matching by Abadie A & Imbens (2006). Lopez and Gutman (2017) provide a review of various matching method for multiple treatment comparisons.

If we want an inference based on IPW, then we use $w_{ki} = 1/\hat{P}_{ki}$ as the weights for patient i in group k .

2.2 Decision Tree Method

The multinomial logistic regression is a parametric approach, so that the final K group comparison results as well as the propensity analysis results can be sensitive to the assumed regression model. Decision tree is a robust alternative to avoid such issues of regression methods.

We define a stratum based on propensity scores that are associated with relative frequencies of the K groups. Ideally, each stratum should be so homogeneous that the propensity scores of the patients within each stratum should be similar. On the other hand, the relative frequencies of the K groups should be very different between any two different strata. The further away two propensity strata are, the more different the relative frequencies between the two strata will be. We derive a propensity score estimation method based on this concept.

A decision tree is a tool to find the best decision for an object based on the values of its predictors. While regression trees usually have a continuous or binomial outcome variable, decision trees have a nominal categorical outcome variable. If there are two possible

decisions (or groups to be compared), then a decision tree is identical to the binary regression tree. We consider decision trees with $K(=3)$ possible decisions that are called treatment groups in this paper. At a node, we go through all possible cutoff values for every predictor and identify the cutoff value of a predictor that gives the most significant partitioning in terms of the proportions among K groups. As the level of nodes goes down, the possible cutoffs of a predictor will be limited to a smaller range. If the cutoff value is too extreme, then one of two strata partitioned by the cutoff will be too small, especially with a continuous predictor. Since we do not want to consider too small strata, we may not consider extreme cutoff values.

In order to measure the significance of a classification, we propose to use the p-value from the chi-squared test for a $2 \times K$ frequency table (see Table 1) comprising two strata identified by a cutoff value of a predictor and K groups. We continue the classification procedure until there exist no more significant cutoff values for any predictors for all nodes. When this procedure is completed, each terminal node consists of a subset of patients with similar propensity scores. In this sense, we call each terminal node a propensity stratum, and the frequencies of the K groups within each stratum can be used for propensity matching.

For the purpose of statistical trimming of regression trees, Jung et al. (2014) propose to control the familywise error rate (FWER) in determining the significance of K tables accounting for the multiplicity of the predictors and the cutoff values of each predictor. This rule, however, will produce a very short decision tree and each of the resulting branches (i.e. strata) will consist of patients with very wide range of propensity scores. So, in this paper, we propose to control the marginal type I error rate based on the p-values of χ^2 -tests with $K - 1$ degrees of freedom. If the total sample size is large and we want each stratum to have very homogeneous propensity scores, then we use a large type I error rate, called α_1 .

When there is no more significant classification, each of the leaves, also called final nodes, can be counted as a stratum consisting of subjects with similar propensity scores and the frequencies among the K groups are summarized by a $K \times 1$ table. A decision tree may possibly result in over-classification. If there are terminal nodes with similar relative frequencies, then we may combine them as a single propensity stratum for efficient estimation of propensity scores. We make this decision if the p-value of the χ^2 -test to compare the relative frequencies between two strata is larger than a prespecified type I error rate, α_2 . We repeat this procedure until there are no more pairs of strata with p-value of χ^2 test larger than the specified α_2 level.

We can consider matching and inverse probability weight (IPW) approaches for an unbiased comparison of the outcome among K groups based on the final decision tree. For matching, we randomly select patients from each stratum using certain proportions among K groups so that the set of proportions is identical across the strata. A simple example will be to select equal number of patients from each group within each stratum, corresponding to a 1-to-1 matching in a two group propensity matching case. Oftentimes, control groups have much more subjects than case groups. If some groups have a lot more subjects than the others for all strata, then we may consider an unbalanced matching among K groups, corresponding to 1-to- m matching in a two group propensity matching. More specifically, suppose that we

want allocation proportions of $(\gamma_1, \dots, \gamma_K)$ among K groups with $\gamma_k > 0$ and $\sum_{k=1}^K \gamma_k = 1$.

Then, in a propensity matching, we randomly select subjects from large groups to satisfy the allocation proportions $(\gamma_1, \dots, \gamma_K)$ within each stratum.

If stratum j has n_{jk} patients in treatment group k , then the propensity score for these patients is estimated by $\hat{P}_{jk} = n_{jk} / \sum_{k'=1}^K n_{jk'}$ and the weights for IPW are given as

$w_{ki} = 1/\hat{P}_{jk} = n_{jk}^{-1} \sum_{k'=1}^K n_{jk'}$. In summary, a decision tree for propensity analysis may proceed as follows. The first node is the whole data.

[I] Classification procedure

[Ia] For each predictor, partition a node into two nodes using a value of the predictor as a cutoff value, and calculate the p-value of the χ^2_{K-1} test.

[Ib] Repeat [Ia] for all possible cutoff values with respect to all predictors. If the smallest p-value is smaller than α_1 , then split the node into two nodes using the corresponding cutoff value.

[Ic] If there exist no p-values smaller than α_1 , then the classification procedure ends.

[II] Pooling procedure: Suppose that a classification procedure resulted in D_1 nodes (or strata).

[IIa] For each of $D_1(D_1 - 1)/2$ pairs of nodes, calculate the p-value of the χ^2_{K-1} test comparing the allocation proportions between two nodes.

[IIb] In [IIa], if the largest p-value is larger than α_2 , combine the corresponding pair of nodes into one.

[IIc] Repeat [Ia] and [Ib] until there exists no pair of nodes with p-value larger than α_2 , to result in D_1 strata.

[III] Among D_2 strata, discard the strata with 0-frequency for any treatment groups, to result in D_3 strata.

Note that, while steps [I] and [II] do not change the sample size, step [III] can lower the sample size for the final analysis. Step [II] corresponds to variable deletion, while step [I] corresponds to variable addition, in stepwise regression.

3 Weighted Test Statistics

Propensity analysis discussed in the previous section uses group identity and predictors, but not outcome data. Once a propensity analysis is completed, we want to conduct a statistical testing to compare the groups on an outcome variable incorporating the propensity analysis result. Cole and Hernan (2004) and Xie and Chaofeng (2005) propose a weighted 2-sample log-rank test for survival outcomes.

In a K -sample test, the null hypothesis is that all K groups have the same survival distributions, i.e. $H_0 : S_1(t) = \cdots = S_K(t)$ for $t \geq 0$, where $S_k(t)$ denotes the survivor function of the population when all patients received treatment k . The statistical testing method for this null hypothesis is different depending on the study design and the study objective. For example, if all K groups are cases, then we may compare their survival distributions using a one-way ANOVA-type test. On the other hand, if a group is a control and the remaining $K - 1$ groups are cases, then we may use a Dunnett-type test to compare each case group with the control. Jung and Hui (2002) and Jung et al. (2008) propose sample size formulas for ANOVA-type and Dunnett-type rank tests, respectively, for comparing the survival distributions among K groups when the distributions of predictors are balanced among treatment groups. In this section, we propose weighted version of these rank tests and derive their asymptotic distributions for large n assuming that $n_k/n \rightarrow \gamma_k (> 0)$ for $k = 1, \dots, K$.

We observe censored survival time X_{ki} , denoting the minimum of the survival and the censoring times, and event indicator δ_{ki} , taking 1 if an event is observed and 0 otherwise, for patient $i (= 1, \dots, n_k)$ in group $k (= 1, \dots, K)$ who has a weight w_{ki} from a propensity analysis. For each patient, the censoring time is independent of the survival time. Let $N_{ki}(t) = \delta_{ki} I(X_{ki} \leq t)$ and $Y_{ki}(t) = I(X_{ki} > t)$ denote the event and the at-risk processes for patient i in group k , and

$N_k(t) = \sum_{i=1}^{n_k} w_{ki} N_{ki}(t)$, $Y_k(t) = \sum_{i=1}^{n_k} w_{ki} Y_{ki}(t)$, $\bar{Y}_k(t) = \sum_{i=1}^{n_k} w_{ki}^2 Y_{ki}(t)$, $N(t) = \sum_{k=1}^K N_k(t)$, and $Y(t) = \sum_{k=1}^K Y_k(t)$. Further, let

$M_{ki}(t) = w_{ki} \int_0^t \{dN_{ki}(s) - Y_{ki}(s)d\Lambda(s)\}$, $M_k(t) = \sum_{i=1}^{n_k} M_{ki}(t)$, and $M(t) = \sum_{k=1}^K M_k(t)$. Let $w_k = \sum_{i=1}^{n_k} w_{ki}$, $w = \sum_{k=1}^K w_k = \sum_{k=1}^K \sum_{i=1}^{n_k} w_{ki}$, and $\bar{w}_k = \sum_{i=1}^{n_k} w_{ki}^2$.

Before discussing K -sample comparison problems, we briefly investigate a weighted Kaplan-Meier estimate as a 1-sample problem.

3.1 Weighted Kaplan-Meier estimate

A weighted estimator for the survival function $S_k(t)$, called weighted Kaplan-Meier (1958) estimator by Galimberti et al. (2002), of group k is obtained by

$$\hat{S}_k(t) = \prod_{s \leq t} \left\{ 1 - \frac{\Delta N_k(s)}{Y_k(s)} \right\} = \prod_{\{1 \leq i \leq n_k : X_{ki} \leq t\}} \left\{ 1 - \frac{\sum_{i'=1}^{n_k} w_{ki'} \delta_{ki'} I(X_{ki'} = X_{ki})}{\sum_{i'=1}^{n_k} w_{ki'} I(X_{ki'} \geq X_{ki})} \right\},$$

where $\Delta N_k(t) = N_k(t) - N_k(t-)$. Note that $M_{ki}(t)$ is a 0-mean martingale with variance $w_{ki}^2 \int_0^t Y_{ki}(s) d\Lambda_k(s)$. Using similar arguments of Fleming and Harrington (1991) for the Kaplan-Meier estimator, we can show that $\hat{S}_k(t)$ is approximately normal with mean $S_k(t)$ and variance $\sigma_k^2(t)$ that can be consistently estimated by

$$\hat{\sigma}_k(t)^2 = \hat{S}_k^2(t) \int_0^t \frac{\bar{Y}_k(s)}{Y_k(s)^2} d\hat{\Lambda}_k(s) = \hat{S}_k(t)^2 \sum_{i=1}^{n_k} \delta_{ki} w_{ki} \frac{\sum_{i'=1}^{n_k} w_{ki'}^2 I(X_{ki'} \geq X_{ki})}{\left\{ \sum_{i'=1}^{n_k} w_{ki'} I(X_{ki'} \geq X_{ki}) \right\}^3},$$

where $\hat{\Lambda}_k(t) = \int_0^t Y_k(s)^{-1} dN_k(s)$ is a weighted version of Aalen-Nelson estimator (Aalen 1978; Nelson 1969) for the cumulative hazard function $\Lambda_k(t) = -\log S_k(t)$ for group k .

3.2 One-Way ANOVA-Type Tests

In an ANOVA-type test, we want to test $H_0 : \Lambda_1(t) = \cdots = \Lambda_K(t) (= \Lambda(t))$ for $t \geq 0$ against $H_1 : \Lambda_k(t) \neq \Lambda_{k'}(t)$ for some $k \neq k'$.

A weighted version of Nelson-Aalen estimator for the common cumulative hazard function $\Lambda(t)$ under H_0 is given as $\hat{\Lambda}(t) = \int_0^t Y^{-1}(s) dN(s)$. For testing H_0 , we consider a weighted log-rank test $W = (W_1, \dots, W_{K-1})^T$, where

$$\begin{aligned} W_k &= \frac{\sqrt{n}}{w_k} \int_0^\infty Y_k(t) \{d\hat{\Lambda}_k(t) - d\hat{\Lambda}(t)\} \\ &= \frac{\sqrt{n}}{w_k} \left\{ \int_0^\infty dN_k(t) - \int_0^\infty \frac{Y_k(t)}{Y(t)} dN(t) \right\} \\ &= \frac{\sqrt{n}}{w_k} \left\{ \sum_{i=1}^{n_k} \delta_{ki} w_{ki} - \sum_{l=1}^K \sum_{i=1}^{n_l} \frac{\delta_{li} w_{li} \sum_{i'=1}^{n_k} w_{ki'} I(X_{ki'} \geq X_{li})}{\sum_{l'=1}^K \sum_{i'=1}^{n_{l'}} w_{l'i'} I(X_{l'i'} \geq X_{li})} \right\}. \end{aligned}$$

Appendix A shows that, under H_0 , W is asymptotically normal with mean 0 and variance V that can be consistently estimated by $\hat{V} = (\hat{v}_{k,k'})_{(K-1) \times (K-1)}$ with

$$\begin{aligned} \hat{v}_{k,k'} &= \frac{n}{w_k w_{k'}} \sum_{l=1}^K \int_0^\infty \left\{ \xi_{kl} - \frac{Y_k(t)}{Y(t)} \right\} \left\{ \xi_{k'l} - \frac{Y_{k'}(t)}{Y(t)} \right\} \frac{\bar{Y}_l(t)}{Y(t)} dN(t) \\ &= \frac{n}{w_k w_{k'}} \sum_{l=1}^K \sum_{i'=1}^{n_{l'}} \sum_{i=1}^{n_l} \delta_{l'i'} w_{l'i'} \left\{ \xi_{kl} - \frac{\sum_{i''=1}^{n_k} w_{ki''} I(X_{ki''} \geq X_{l'i'})}{\sum_{l''=1}^K \sum_{i''=1}^{n_{l''}} w_{l''i''} I(X_{l''i''} \geq X_{l'i'})} \right\} \\ &\quad \times \left\{ \xi_{k'l} - \frac{\sum_{i''=1}^{n_{k'}} w_{k'i''} I(X_{k'i''} \geq X_{l'i'})}{\sum_{l''=1}^K \sum_{i''=1}^{n_{l''}} w_{l''i''} I(X_{l''i''} \geq X_{l'i'})} \right\} \frac{\sum_{i''=1}^{n_l} w_{li''}^2 I(X_{li''} \geq X_{l'i'})}{\sum_{l''=1}^K \sum_{i''=1}^{n_{l''}} w_{l''i''} I(X_{l''i''} \geq X_{l'i'})}, \end{aligned}$$

where $\xi_{k,k'} = I(k = k')$. Hence, with a specified type I error rate α , we can reject H_0 if $Q = W^T \hat{V}^{-1} W$ is larger than $\chi_{K-1, 1-\alpha}^2$.

3.3 Dunnett-Type Tests

In K -sample problems, oftentimes one of the groups is a control and the remaining groups are cases. We assume that group K is the control. In this case, we usually want to test if each of the $K-1$ case groups is more efficacious than the control group. For $k = 1, \dots, K-1$, we want to test the null hypothesis $H_k: \Lambda_k(t) = \Lambda_K(t)$ against the alternative hypothesis $\bar{H}_k: \Lambda_k(t) \neq \Lambda_K(t)$. Denote $H_0 = \cap_{k=1}^{K-1} H_k$ and $H_a = \cup_{k=1}^{K-1} \bar{H}_k$. For $k = 1, \dots, K-1$, the weighted log-rank statistic (Peto R, Peto 1972; Jung & Hui 2002) for testing H_k is given by

$$U_k = \frac{\sqrt{n}(w_k + w_K)}{w_k w_K} \int_0^\infty \frac{Y_k(t) Y_K(t)}{Y_k(t) + Y_K(t)} \{d\hat{\Lambda}_K(t) - d\hat{\Lambda}_k(t)\} \\ = \frac{\sqrt{n}(w_k + w_K)}{w_k w_K} \left\{ \sum_{i=1}^{n_K} \frac{\delta_{Ki} w_{Ki} Y_k(X_{Ki})}{Y_k(X_{Ki}) + Y_K(X_{Ki})} - \sum_{i=1}^{n_k} \frac{\delta_{ki} w_{ki} Y_K(X_{ki})}{Y_k(X_{ki}) + Y_K(X_{ki})} \right\} \\ = \frac{\sqrt{n}(w_k + w_K)}{w_k w_K} \left\{ \sum_{i=1}^{n_K} \frac{\delta_{Ki} w_{Ki} \sum_{i'=1}^{n_k} w_{ki'} I(X_{ki'} \geq X_{Ki})}{\sum_{i'=1}^{n_k} w_{ki'} I(X_{ki'} \geq X_{Ki}) + \sum_{i'=1}^{n_K} w_{Ki'} I(X_{Ki'} \geq X_{Ki})} - \sum_{i=1}^{n_k} \frac{\delta_{ki} w_{ki} \sum_{i'=1}^{n_K} w_{Ki'} I(X_{Ki'} \geq X_{ki})}{\sum_{i'=1}^{n_k} w_{ki'} I(X_{ki'} \geq X_{ki}) + \sum_{i'=1}^{n_K} w_{Ki'} I(X_{Ki'} \geq X_{ki})} \right\}.$$

Note that a positive U_k value implies that group k has a longer survival time than group K . By Appendix B, under H_k , U_k is approximately normal with mean 0 and variance σ_k^2 that can be consistently estimated by

$$\hat{\sigma}_k^2 = \frac{n(w_k + w_K)^2}{w_k^2 w_K^2} \int_0^\infty \frac{Y_k^2(t) \bar{Y}_K(t) + Y_K^2(t) \bar{Y}_k(t)}{\{Y_k(t) + Y_K(t)\}^2} d\hat{\Lambda}_{(k)}(t) \\ = \frac{n(w_k + w_K)^2}{w_k^2 w_K^2} \left[\sum_{i=1}^{n_k} \delta_{ki} w_{ki} \frac{Y_k(X_{ki})^2 \bar{Y}_K(X_{ki}) + Y_K(X_{ki})^2 \bar{Y}_k(X_{ki})}{\{Y_k(X_{ki}) + Y_K(X_{ki})\}^3} + \sum_{i=1}^{n_K} \delta_{Ki} w_{Ki} \frac{Y_k(X_{Ki})^2 \bar{Y}_K(X_{Ki}) + Y_K(X_{Ki})^2 \bar{Y}_k(X_{Ki})}{\{Y_k(X_{Ki}) + Y_K(X_{Ki})\}^3} \right],$$

where $\hat{\Lambda}_{(k)}(t) = \int_0^t \{Y_k(s) + Y_K(s)\}^{-1} \{dN_k(s) + dN_K(s)\}$ denotes the weighted Nelson-Aalen estimator under H_k .

If we reject H_k with a type I error probability of α , the probability of false rejections for the global null hypothesis H_0 will be larger than the nominal α level due to the multiple testing issue. In order to avoid this issue, statisticians control the FWER. For a chosen critical value c , we reject H_k if $|U_k/\hat{\sigma}_k| > c$. Dunnett-type tests control the FWER at α by choosing $c = c_\alpha$ satisfying

$$\alpha = P\left\{ \cup_{k=1}^{K-1} (|U_k/\hat{\sigma}_k| > c) \mid H_0 \right\}. \quad (2)$$

The Bonferroni test, rejecting H_k with a type I error probability of $\alpha/(K-1)$, results in too a conservative testing result.

From Appendix B, for large n under H_0 , $(U_1/\hat{\sigma}_1, \dots, U_{K-1}/\hat{\sigma}_{K-1})$ is approximately normal with means 0, variances 1, and correlation coefficients

$$\hat{\rho}_{kk'} = \frac{w_k w_{k'} \bar{w}_K}{\sqrt{(w_K^2 \bar{w}_k + w_k^2 \bar{w}_K)(w_K^2 \bar{w}_{k'} + w_{k'}^2 \bar{w}_K)}}.$$

for $1 \leq k < k' \leq K-1$. Hence, we can obtain $c = c_\alpha$ from (2) by using a numerical method (Genz & Bretz 2000; Gassmann et al. 2002) or a simulation method (Bang et al. 2005). We focus on the former method in this paper. Let $\phi_{K-1}(u_1, \dots, u_{K-1})$ denote the joint probability density function (PDF) of the $(K-1)$ -variate normal distribution with marginal means 0, variances 1, and correlation coefficients $\hat{\rho}_{kk'}$. Then, from (2), we obtain $c = c_\alpha$ by solving

$$\alpha = 1 - \int_{-c}^c \dots \int_{-c}^c \phi_{K-1}(u_1, \dots, u_{K-1}) du_1 \dots du_{K-1} \quad (3)$$

with respect to c .

If (Z_1, Z_2) is bivariate normal with means (μ_1, μ_2) , variances (σ_1^2, σ_2^2) , correlation coefficient ρ , then it is well known that Z_1 conditioning on $Z_2 = z_2$ is normal with mean $\mu_1 + (\rho\sigma_1/\sigma_2)(z_2 - \mu_2)$ and variance $\sigma_1^2(1 - \rho^2)$. Hence, for $K=3$, (3) can be expressed using a one dimensional integration

$$\alpha = 1 - \int_{-c}^c \phi(u) \left[\Phi\left(\frac{c - \hat{\rho}_{12}u}{\sqrt{1 - \hat{\rho}_{12}^2}}\right) - \Phi\left(\frac{-c - \hat{\rho}_{12}u}{\sqrt{1 - \hat{\rho}_{12}^2}}\right) \right] du,$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density function and the cumulative density function of $N(0,1)$.

Our Dunnett-type test is conducted as follows.

1. Calculate the weights $(w_{ki}, i = 1, \dots, n_k, k = 1, \dots, K)$ using the multinomial logistic regression or decision tree method.
2. By solving (3), obtain the critical value c_α .
3. For $k = 1, \dots, K-1$, we reject H_k if $|U_k/\hat{\sigma}_k| > c_\alpha$.

Given an observed test statistic $z_k = U_k/\hat{\sigma}_k$ for group $k (= 1, \dots, K-1)$, the FWER-adjusted p-value (Jung et al. 2005), p_k , is given as

$$p_k = 1 - \int_{-|z_k|}^{|z_k|} \dots \int_{-|z_k|}^{|z_k|} \phi_n(u_1, \dots, u_{K-1}) du_1 \dots du_{K-1},$$

which is simplified to

$$p_k = 1 - \int_{-|z_k|}^{|z_k|} \phi(u) \left[\Phi \left(\frac{|z_k| - \hat{\rho}_{12}u}{\sqrt{1 - \hat{\rho}_{12}^2}} \right) - \Phi \left(\frac{-|z_k| - \hat{\rho}_{12}u}{\sqrt{1 - \hat{\rho}_{12}^2}} \right) \right] du,$$

for $K = 3$.

4 Numerical Studies

4.1 Simulations

We conduct simulation studies to evaluate the performance of the K -sample comparison methods combined with propensity analysis methods under various settings. We consider comparing the survival distribution of $K = 3$ groups using the ANOVA or Dunnett test combined with the IPW approach. The third group ($K = 3$) is the control group for Dunnett test. The weights are estimated by the multinomial logistic regression or decision tree method.

For patient i , the predictor z_i was generated from $U(0, 3)$. Given a predictor value z_i , the allocation proportion for each experimental group is determined by one of following models.

(A1) (Multinomial logistic model)

$$\gamma_1 = \gamma_2 = \frac{\exp(0.15z_i)}{1 + 2 \exp(0.15z_i)}, \quad \gamma_3 = 1 - 2\gamma_1$$

(A2) (Multinomial logistic model with an interval with controls only)

$$\gamma_1 = \gamma_2 = \frac{\exp(0.15z_i)}{1 + 2 \exp(0.15z_i)}, \quad \gamma_3 = 1 - 2\gamma_1$$

if $z_i \in [0, 2]$, and $\gamma_3 = 1$ if $z_i \in (2, 3]$.

The survival distribution of patient i in treatment group k with predictor z_i is generated from an exponential distribution with an annual hazard rate of

$$\lambda_i = 0.2 \exp(\beta_k + 0.5z_i).$$

Note that β_k denotes the effect of group k and $0.5z_i$ denotes the impact of the predictor on the survival outcome. We consider a null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, and alternative hypotheses $H_1 : (\beta_1, \beta_2, \beta_3) = (-0.35, -0.35, 0.05)$ and $H_2 : (\beta_1, \beta_2, \beta_3) = (-0.45, -0.25, 0.15)$.

We consider another model with allocation probabilities non-monotone in z_i .

(A3) (A non-multinomial logistic model)

$$\gamma_1 = \gamma_2 = \frac{1.25}{\sqrt{2\pi}} \exp\left[-\frac{(z_i - 1.5)^2}{2}\right], \quad \gamma_3 = 1 - 2\gamma_1$$

For (A3), the propensity score for group 3 is close to 1 for z_i close to 0 or 3, and close to 0 for $z_i = 1.5$. This propensity model is matched with an exponential survival distribution whose hazard rate is non-monotone in z_i ,

$$\lambda_i = 0.4 \exp\{\beta_k + 0.3 - 0.6I(1 < z_i < 2)\}.$$

The censoring variable is generated from $U(2, 7)$ mimicking a study with 5 years of patient registry and 2 years of additional follow-up to generate about 20% of censoring. The parameter values for propensity models and survival distributions are chosen for at least 80% of power by the decision tree method.

For the decision tree method, we use $\alpha_1 = 0.05$ for classification and $\alpha_2 = 0.3$ for pooling, and extreme cutoff values resulting in less than 5 patients in one stratum are not considered. We consider classification only (I), classification and pooling (I+II), classification and discarding the strata with 0-frequency groups (I+III), or all of the three steps (I+II+III), and report the average number of strata D for each of these procedures.

Under each simulation setting, we generate 10,000 simulation data sets of size $n = 1,000$, and apply the ANOVA or Dunnett test using the weights estimated by multinomial logistic regression or decision tree method to each sample. The empirical type I error rate and power are estimated by the proportion of simulation samples rejecting the null hypothesis with nominal $\alpha = 0.05$ among 10,000 simulation samples.

Table 2(a) summarizes the simulation results. If the treatment selection follows a multinomial logistic model (A1), then the weighted rank tests using the weights from both multinomial logistic and decision tree methods control the type I error rate closely under H_0 . Under allocation model (A1), the decision tree method does not seem to need steps (II) and (III) in addition to classification (I). If there exists an interval of predictor z_i with control patients only (A2) or if the allocation probabilities do not have monotone trends in z_i (A3), then the weighted rank tests using the propensity scores from multinomial logistic regression do not control the type I error accurately.

Under (A2) and (A3) allocation models, the decision tree method controls the type I error accurately when step (III) is incorporated. From the average number of strata D , as expected, step (III) much decreases the number of strata for model (A2), and step (II) much decreases the number of strata for model (A3).

Under H_1 and H_2 , the empirical powers of the cases where type I is accurately controlled are bold faced. The Dunnett-type test looks slightly more powerful than the ANOVA-type test in most of the simulation settings except for model (A3) under H_2 . If a multinomial logistic

allocation model is valid, then the weighted rank tests based on multinomial logistic regression method have a slightly higher power than those based on decision tree method.

A reviewer suggests to add some simulations with a more general survival distributions, such as Weibull, comparisons among $K = 4$ treatment groups, and a similar setting to that of the real data example presented below. We generate $m = 4$ correlated covariates as follows. At first, a random vector (u_1, u_2, u_3, u_4) is generated from the multivariate normal distribution with marginal means 0 and variances 1 and a exchangeable dependency with correlation coefficient 0.5. These random variables are transformed for $z_1 = u_1$, $z_2 = [z_2]$, $z_3 = \Phi^{-1}(u_3)$, and $z_4 = I(u_1 - 0.5)$, where $\langle a \rangle$ denotes the round down of a . Note that, marginally, z_3 is $U(0,1)$ and z_4 is Bernoulli(0.5) random variables. The treatment selection follows a multinomial logistic model with $\beta_{0k} = 0$, $\beta_{1k} = \beta_{2k} = 0.3$, and $\beta_{3k} = \beta_{4k} = 0.2$ for $k = 1, 2, 3$. By this multinomial logistic model, about 25.8% are allocated to each of groups 1, 2, and 3, and the remaining 22.5% are allocated to the control group 4. The survival time for a patient in treatment group k with covariate values (z_1, z_2, z_3, z_4) is generated from Weibull hazard function of given as

$$\lambda(t|z) = \nu \lambda_0 (\lambda_0 t)^{\nu-1} \exp(\beta_k + \theta_1 z_1 + \theta_2 z_2 + \theta_3 z_3 + \theta_4 z_4)$$

with $\theta_1 = \theta_2 = \theta_3 = \theta_4 = 0.25$ and $\nu = 0.5$. We consider a null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$, and alternative hypotheses $H_1 : (\beta_1, \beta_2, \beta_3, \beta_4) = (-0.65, -0.65, -0.65, 0)$ and $H_2 : (\beta_1, \beta_2, \beta_3, \beta_4) = (-0.75, -0.5, -0.25, 0)$. Under each hypothesis, λ_0 is selected for about 33% censoring by $U(2,7)$ censoring distribution. Other simulation parameters are identically set to those of Table 2(a) for $K = 3$. Table 2(b) reports simulation results with $n = 400$. We observe that, under H_0 , the weighted ANOVA and Dunnett tests control the type I error rate closely to the nominal level $\alpha = 0.05$ by multinomial logistic regression and decision tree with different options, while the unweighted rank tests are severely anticonservative. Using multinomial logistic regression, the weighted ANOVA test is slightly more powerful than the weighted Dunnett test under both H_1 and H_2 , while, using decision tree method, the weighted Dunnett test is slightly more (less) powerful than the weighted ANOVA test under H_1 (H_2). The decision tree method has similar power with different options, and it looks slightly less powerful than the multinomial logistic regression method.

4.2 A Real Data Example

The proposed methods are applied to a real clinical observational study. Postoperative analgesic methods are suggested to have an impact on long term prognosis after cancer surgery through opioid-induced immune suppression. Lee et al. (2017) report analysis results comparing the recurrence-free survival (RFS), defined as time to cancer recurrence or death, from surgery among three analgesic methods ($K = 3$): intravenous patient controlled analgesia (PCA, group 1), paravertebral block (PVB, group 2), and thoracic epidural analgesia (TEA, group 3), for lung cancer surgery. Excluding cases with missing data, our analysis includes a total of 363 patients (111 for PCA, 137 for PVB, and 115 for TEA) among whom 111 patients had disease recurrence. We consider four covariates of body mass

index (BMI), smoking, cancer stage and blood transfusion during surgery (BT) that are known to be important predictors for the outcome and study population.

We estimate the weights by multinomial logistic regression and decision tree methods. Table 3 reports the regression estimates, standard errors and p-values from multinomial logistic propensity analysis. Since the regression estimates of smoking from the two logistic regressions have negative sign, we find that more patients with smoking tend to belong to TEA. Figure 1 shows the strata identified by the decision tree method. In the decision tree analysis, we used $\alpha = 0.2$ for a little finer classification and $\alpha = 0.3$ for pooling. In Figure 1, each leaf represents a stratum of patients with similar propensity scores, and the frequencies for the three groups of each stratum are given within a box. The p-value at each branch is from χ^2 test with $2(= (3-1) \times (2-1))$ degrees of freedom for classification or pooling. The classification step has identified 8 strata, but two of them are combined during the pooling step to result in a total of 7 strata to be used for comparing RFS among the three treatment groups. Figure 2 reports the weighted Kaplan-Meier curves using these propensity analysis methods based on multinomial logistic regression (Fig 2.a) and decision tree (Fig 3.b). The IPW method using the multinomial logistic regression for testing the three groups gives us p-value = 0.040 by the ANOVA test, and FWER-adjusted p-value = 0.022 between PCA and TEA and 0.704 between PVB and TEA by the Dunnett test. On the other hand, the IPW method using the decision tree gives us p-value = 0.025 by the ANOVA test, and FWER-adjusted p-value = 0.011 between PCA and TEA and 0.469 between PVB and TEA by the Dunnett test. From these analyses, it seems that, compared to TEA, PCA has significantly longer RFS and PVB seems to have similar RFS. All p-values in this analysis are 2-sided.

We also analyze the data after balanced (1-to-1-to-1) matching. Using the multinomial logistic regressions, we partitioned the propensity scores of each group into $J_1 = J_2 = 3$ intervals to result in 234 matched observations. From Fig 1, the seven strata from decision tree method will result in 240 matched observations by keeping all patients of the smallest group within each stratum. From Table 4, we observe that the distributions of predictors are well balanced among the three groups using both multinomial regression and decision tree methods. Figures 3 displays the Kaplan-Meier curves of RFS among the three matched groups using the multinomial regression and decision tree methods. We observe that the Kaplan-Meier curves for matched data using multinomial logistic regression method (Figure 3a) are similar to the weighted Kaplan-Meier curves of Figure 2a, while those of PVB and TEA groups using decision tree method (Figure 3b) are more separated than the corresponding weighted Kaplan-Meier curves in Figure 2b. Using the propensity scores from multinomial logistic regression, the matched data have ANOVA p-value = 0.066 by the (unweighted) ANOVA test (Jung & Hui 2002), and FWER-adjusted p-value = 0.052 between PCA and TEA and 0.979 between PVB and TEA by the (unweighted) Dunnett test (Jung et al. 2008). Using the propensity analysis from decision tree, the matched data have ANOVA p-value = 0.057 by the ANOVA test, and FWER-adjusted p-value = 0.029 between PCA and TEA and 0.500 between PVB and TEA by the Dunnett test. With the decreased sample sizes, the unweighted statistical tests using matched data have a little lower significance level than, but similar conclusions as, the weighted tests using the full data.

The clinical study was designed to show that, compared to the local anesthetic based analgesic methods (TEA and PVB), the opioid based analgesic method (PCA) would decrease RFS. Due to a high failure rate and unstable hemodynamics frequently observed in the local anesthetic based analgesic methods, however, the local anesthetic based analgesic methods were shown to have shorter RFS than the opioid based analgesic method.

4.3 Discussions and Conclusions

We have investigated K -group comparisons for observational studies with a survival endpoint. The weighted rank tests proposed in this paper can be easily modified for other types of endpoints, such as binary or continuous types. In order to estimate the weights (propensity scores) of the rank tests, we reviewed the multinomial logistic regression and proposed a decision tree method as an alternative.

Through simulations, the proposed weighted rank tests were shown to perform well as far as the weights were accurately estimated by the propensity analysis methods. We found that the decision tree method provided very robust propensity scores overall if it incorporate a step to discard strata with zero frequency for any treatment group to ensure comparability among K treatment groups for all strata included in the analysis. The popular multinomial logistic model was found to provide powerful and robust propensity scores if the allocation proportion to each group is monotone and there are no ranges of predictors with zero frequency for any group. When there exists a range with no frequencies for some treatment groups, Zanutto et al. (2005) propose to find a common support of $(K - 1)$ -dimensional propensity scores and discard the observations not in the common support. In a real data analysis, however, it is not always easy to identify a common support since the estimated propensity scores are discrete in nature. We can do this if we partition the range of propensity scores and define strata as in our real data analysis.

Zhu & Lu (2015) propose a Dunnett-type test derived from a Cox regression model

$$\lambda_j(t) = \lambda_0 \exp(\beta_1 x_1 + \cdots + \beta_{K-1} x_{K-1})$$

for patients in propensity stratum j with covariates $(x_1, \dots, x_{K-1})$, where $x_k = 1$ if a patient belongs to treatment group k and $= 0$ otherwise. They claim that β_k measures the difference in survival distribution between group k and the control group K , but this is not true. Actually, β_k measures the difference between group k and the remaining $K - 1$ groups combined. Furthermore, they assume a specific covariance structure, called one-factor structure, among the regression estimators to derive the multiplicity adjusted critical value c using Hsu's (1992) approach. Our method does not require this assumption which may not hold for a given data set.

We have focused on IPW method in this paper. But one may want to use standard (unweighted) test statistics using matched data. In this case, the decision tree method automatically and optimally defines strata, which makes data matching easy, while the multinomial logistic regression method requires an additional step to define strata. In the analysis of our example, we simply partitioned each dimension of propensity scores so that

each interval has similar propensity scores, but we could use a more technical $(K-1)$ -dimensional clustering method, such as decision tree or machine learning, e.g. Westreich et al. (2010), Abadie & Imbens (2006), and Lee et al. (2009). The computer programs are composed in Fortran and available upon request from the first author.

Appendices: Asymptotic Distribution of Weighted Log-Rank Tests

In the following appendices, we use the notations of $\psi_k = \lim_{n \rightarrow \infty} w_k/n$, $\psi = \lim_{n \rightarrow \infty} w/n = \sum_{k=1}^K \psi_k$, and $\bar{\psi}_k = \lim_{n \rightarrow \infty} \bar{w}_k/n$.

Appendix A.: An ANOVA-type Test

Under H_0 , we have

$$\begin{aligned} W_k &= \frac{\sqrt{n}}{w_k} \int_0^\infty Y_k(t) \{d\hat{\Lambda}_k(t) - d\hat{\Lambda}(t)\} \\ &= \frac{\sqrt{n}}{w_k} \left\{ \int_0^\infty dM_k(t) - \int_0^\infty \frac{Y_k(t)}{Y(t)} dM(t) \right\} \\ &= \frac{\sqrt{n}}{w_k} \sum_{l=1}^K \int_0^\infty \left\{ \xi_{kl} - \frac{Y_k(t)}{Y(t)} \right\} \sum_{i=1}^{n_l} dM_{li}(t). \end{aligned}$$

Note that $\{M_{kl}(t), i=1, \dots, n_k, k=1, \dots, K\}$ are independent 0-mean martingales under H_0 , and $R_k(t) = Y_k(t)/Y(t)$ are predictable function that uniformly converges to $r_k(t) = \psi_k S_k(t) / \sum_{l=1}^K \psi_l S_l(t)$. Hence, by the martingale central limit theorem [21], $W = (W_1, \dots, W_{K-1})^T$ is asymptotically normal with mean 0 and variance $V = (v_{k,k'})_{(K-1) \times (K-1)}$, where

$$v_{k,k'} = \frac{1}{\psi_k \psi_{k'}} \sum_{l=1}^K \int_0^\infty \left\{ \xi_{kl} - r_k(t) \right\} \left\{ \xi_{k'l} - r_{k'}(t) \right\} \bar{y}_l(t) d\Lambda(t).$$

$\bar{y}_k(t) = \lim_{n \rightarrow \infty} \bar{Y}_k(t)/n = \bar{\psi}_k S_k(t) G(t)$, and $G(t)$ is the survivor function of the censoring distribution.

V can be consistently estimated by replacing ψ_k , $\bar{y}_k(t)$, $r_k(t)$ and $\Lambda(t)$ with their consistent estimators w_k/n , $\bar{Y}_k(t)/n$, $R_k(t)$ and $\hat{\Lambda}(t)$, respectively, i.e. $\hat{V} = (\hat{v}_{k,k'})_{(K-1) \times (K-1)}$ with

$$\hat{v}_{k,k'} = \frac{n}{w_k w_{k'}} \sum_{l=1}^K \int_0^\infty \left\{ \xi_{kl} - \frac{Y_k(t)}{Y(t)} \right\} \left\{ \xi_{k'l} - \frac{Y_{k'}(t)}{Y(t)} \right\} \bar{Y}_l(t) d\hat{\Lambda}(t).$$

Appendix B.: A Dunnett-Type Test

Under $H_0 : \Lambda_1(t) = \dots = \Lambda_K(t) (= \Lambda(t))$, for $k=1, \dots, K-1$, we have

$$U_k = \frac{\sqrt{n}(w_k + w_K)}{w_k w_K} \int_0^\infty \frac{Y_k(t)Y_K(t)}{Y_k(t) + Y_K(t)} \left\{ \frac{dM_K(t)}{Y_K(t)} - \frac{dM_k(t)}{Y_k(t)} \right\},$$

which can be expressed as

$$U_k = \frac{\psi_k + \psi_K}{\sqrt{n}\psi_k\psi_K} \int_0^\infty \frac{y_k(t)y_K(t)}{y_k(t) + y_K(t)} \left\{ \frac{dM_K(t)}{y_K(t)} - \frac{dM_k(t)}{y_k(t)} \right\} + o_p(1). \quad (A1)$$

Using the martingale central limit theorem, we can show that, under H_0 , U_k is asymptotically normal with mean 0 and variance

$$\sigma_k^2 = \frac{(\psi_k + \psi_K)^2}{\psi_k^2 \psi_K^2} \int_0^\infty \frac{y_k^2(t)\bar{y}_K(t) + y_K^2(t)\bar{y}_k(t)}{\{y_k(t) + y_K(t)\}^2} d\Lambda(t)$$

which can be consistently estimated by

$$\hat{\sigma}_k^2 = \frac{n(w_k + w_K)^2}{w_k^2 w_K^2} \int_0^\infty \frac{Y_k^2(t)\bar{Y}_K(t) + Y_K^2(t)\bar{Y}_k(t)}{\{Y_k(t) + Y_K(t)\}^2} d\hat{\Lambda}_{(k)}(t)$$

by replacing the parameters with their estimators obtained using the data of groups k and K .

By the multivariate martingale central limit theorem applied to (A1), $(U_1, \dots, U_{K-1})$ is approximately normal with mean 0 and variance-covariance matrix $\Sigma = (\sigma_{kk'})_{(K-1) \times (K-1)}$ with $\sigma_{kk} = \sigma_k^2$ and

$$\sigma_{kk'} = \frac{(\psi_k + \psi_K)(\psi_{k'} + \psi_K)}{\psi_k \psi_{k'} \psi_K^2} \int_0^\infty \frac{y_k(t)y_{k'}(t)\bar{y}_K(t)}{\{y_k(t) + y_{k'}(t)\}\{y_{k'}(t) + y_K(t)\}} d\Lambda(t)$$

for $k \neq k'$. Under H_0 , we have $y_k(t) = \psi_k S(t)G(t)$ and $\bar{y}_k(t) = \bar{\psi}_k S(t)G(t)$, where $S(t) = \exp\{-\Lambda(t)\}$ denotes the common survivor function. Hence, under H_0 , we have

$$\sigma_k^2 = \left(\frac{\bar{\psi}_k}{\psi_k^2} + \frac{\bar{\psi}_K}{\psi_K^2} \right) \int_0^\infty G(t) dS(t)$$

and

$$\sigma_{kk'} = \frac{\bar{\psi}_K}{\psi_K^2} \int_0^\infty G(t) dS(t)$$

for $1 \leq k < k' \leq K-1$, so that the correlation coefficient between U_k and $U_{k'}$ is given as

$$\rho_{kk'} = \frac{\psi_k \psi_{k'} \bar{\psi}_K}{\sqrt{(\psi_K^2 \bar{\psi}_k + \psi_k^2 \bar{\psi}_K)(\psi_K^2 \bar{\psi}_{k'} + \psi_{k'}^2 \bar{\psi}_K)}}.$$

Note that the correlation coefficients depend only on the weights, but not on the censoring or survival distribution.

By replacing ψ_k and $\bar{\psi}_k$ with their consistent estimators w_k/n and $\bar{w}_k/n$, respectively, we obtain a consistent estimator of $\rho_{kk'}$

$$\hat{\rho}_{kk'} = \frac{w_k w_{k'} \bar{w}_K}{\sqrt{(w_K^2 \bar{w}_k + w_k^2 \bar{w}_K)(w_K^2 \bar{w}_{k'} + w_{k'}^2 \bar{w}_K)}}.$$

for $1 \leq k < k' \leq K-1$.

REFERENCES

- Aalen OO (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics*, 6, 701–726.
- Abadie A & Imbens GW (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74, 2352–2376.
- Bang HJ, Jung SH, & George SL (2005). A simulation-based multiple testing procedure and sample size calculation. *Journal of Biopharmaceutical Statistics*, 15, 957–967. [PubMed: 16279354]
- Breslow NE, Lumley T, Ballantyne CM, Chambless LE, & Kulich M (2009). Using the whole cohort in the analysis of case-cohort data. *American Journal of Epidemiology*, 169, 1398–1405.
- Cole SR & Hernan MA (2004). Adjusted survival curves with inverse probability weights. *Computer Methods and Programs in Biomedicine*, 75, 454–49.
- Curtis LH, Hammill BG, Eisenstein EL, Kramer JM, & Anstrom KJ (2007). Using inverse probability-weighted estimators in comparative effectiveness analyses with observational databases. *Medical Care*, 45, S103–S107. [PubMed: 17909367]
- Dunnett CW (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of American Statistical Association*, 50, 1096–1121.
- Fleming TR & Harrington DP (1991). *Counting processes and survival analysis*. Wiley: New York.
- Galimberti S, Sasieni P, & Valsecchi MG (2002). A weighted KaplanMeier estimator for matched data with application to the comparison of chemotherapy and bone-marrow transplant in leukaemia. *Statistics in Medicine*, 21, 3847–3864.
- Gassmann HI, Deak I, & Szantai T (2002). Computing multivariate normal probabilities: A new look. *Journal of Computational and Graphical Statistics*, 11, 920–949.
- Genz A & Bretz F (2000). Numerical computation of critical values for multiple comparison problems. *ASA Proceedings of the Sections on Statistical Computing and Statistical Graphics*, 84–87.
- Hsu JC (1992). The factor analytic approach to simultaneous inference in the general linear model. *Journal of Computational and Graphical Statistics*, 1, 151–168.
- Imbens GW (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87, 706–710.
- Jung SH, Bang H, & Young S (2005). Sample size calculation for multiple testing in microarray data analysis. *Biostatistics*, 6, 157–169. [PubMed: 15618534]
- Jung SH, Chen Y, & Ahn H (2014). Type I error control for tree classification. *Cancer Informatics*, 13, 11–18.

- Jung SH & Hui S (2002). Sample size calculations to compare K different survival distributions. *Lifetime Data Analysis*, 8, 361–73. [PubMed: 12471945]
- Jung SH, Kim C, & Chow SC (2008). Sample size calculation for the log-rank tests for multi-arm trials with a common control. *Journal of Korean Statistical Society*, 37, 11–22.
- Kaplan EL & Meier P (1958). Nonparametric estimation from incomplete observations. *Journal of American Statistical Association*;53:457481.
- Lee BK, Lessler J, & Stuart EA (2009). Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29, 337–346.
- Lee EK, Ahn HJ, Zo JI, Kim K, Jung DM, & Park JH (2017). Paravertebral block does not reduce cancer recurrence, but is related to higher overall survival in lung cancer surgery: A retrospective cohort Study. *Anesthesia and Analgesia*, 125, 1322–1328. [PubMed: 28857802]
- Lopez MJ & Gutman R (2017). Estimation of causal effects with multiple treatments: a review and new ideas. arXiv:1701.05132 [stat.ME].
- McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, & Burgette LF (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine*, 32, 33883414.
- Nelson W (1969). Hazard plotting for incomplete failure data. *Journal of Quality Technology*, 1, 27–52.
- Peto R & Peto J (1972). Asymptotically efficient rank invariant test procedures (with discussion). *Journal of the Royal Statistical Society, Series A*, 135, 185–206.
- Pruzek RM & Cen L (2002). Propensity score analysis with graphics: A comparison of two kinds of gallbladder surgery Paper presented at the annual meeting of the Society for Multivariate Experimental Psychology, Charlottesville, VA.
- Rosenbaum PR & Rubin DB (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.
- Rosenbaum PR & Rubin DB (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of American Statistical Association*, 95, 749–759.
- Rubin DB (1998). Estimation from nonrandomized treatment comparisons using subclassification on propensity scores In *Nonrandomized comparative clinical studies*, ed. Abel U and Koch A, 85–100. Dusseldorf, Germany: Symposion.
- Stone RA, Obrosky DS, Singer DE, Kapoor WN, Fine MJ. (1995). Propensity score adjustment for pretreatment differences between hospitalized and ambulatory patients with community-acquired pneumonia. *Medical Care*, 33, AS56–66. [PubMed: 7723462]
- Westreich D, Lessler J, & Funk MJ (2010). Propensity estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternative to logistic regression. *Journal of Clinical Epidemiology*, 63, 826–833. [PubMed: 20630332]
- Xie J & Liu C (2005). Adjusted KaplanMeier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Statistics in Medicine*, 24, 30893110.
- Zanutto E, Lu B, & Hornik R (2005). Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *Journal of Educational and Behavioral Statistics*, 30, 59–73.
- Zhu H & Lu B (2015). Multiple comparisons for survival data with propensity score adjustment. *Computational Statistics and Data Analysis*, 86, 42–51. [PubMed: 25663729]

Figure 1.

Decision tree analysis of Lee et al. (2017). Each leaf denoting a stratum has frequencies of treatment groups of PCA, PVB, and TEA.

Figure 2.
Weighted Kaplan-Meier curves: (a) Using multinomial logistic regression

Figure 2.
Weighted Kaplan-Meier curves: (b) Using decision tree method

Figure 3. Kaplan-Meier curves of matched data: (a) Based on propensity analysis using multinomial logistic regression

Figure 3.
Kaplan-Meier curves of matched data: (b) Based on propensity analysis using decision tree method

Table 1.

$K \times 2$ frequency table defined by cutoff value c for predictor z

Propensity stratum	Treatment group			
	1	2	...	K
$z \leq c$	n_{11}	n_{12}	...	n_{1K}
$z > c$	n_{21}	n_{22}	...	n_{2K}

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2.

Empirical rejection probabilities for the weighted log-rank tests based on multinomial logistic regression and decision tree propensity methods. The decision tree method, we also report average number of strata (D) after the first classification (I), after pooling the strata with similar allocation proportions (I+II), after discarding the strata with no frequency in any of the three treatment groups (I+III), and pooling the strata with similar allocation proportions and discarding the strata with no frequency in any of the three treatment groups (I+II+III).(a) When $K = 3$

Allocation	Censoring	Multinomial			Tree (I)			Tree (I+II)			Tree (I+III)			Tree (I+II+III)		
		ANO	Dun		ANO	Dun	D	ANO	Dun	D	ANO	Dun	D	ANO	Dun	D
(i) Undet $H_0 : \beta_1=\beta_2=\beta_3$																
A1	19.6%	0.034	0.034	0.039	0.039	0.039	4.1	0.039	0.040	3.5	0.039	0.039	3.8	0.039	0.039	3.2
A2	19.6%	0.098	0.109	0.664	0.680	0.680	4.2	0.665	0.679	3.8	0.047	0.044	3.0	0.047	0.045	2.6
A3	17.4%	0.798	0.834	0.083	0.085	0.085	11.7	0.083	0.085	8.0	0.060	0.063	10.3	0.059	0.064	7.0
(ii) Undet $H_1 : \beta_1=\beta_2<\beta_3$																
A1	26.6%	0.962	0.971	0.942	0.955	0.955	4.1	0.940	0.955	3.5	0.942	0.955	3.8	0.941	0.955	3.2
A2	24.4%	0.998	0.998	1.000	1.000	1.000	4.2	1.000	1.000	3.8	0.921	0.936	2.9	0.920	0.935	2.6
A3	23.4%	1.000	1.000	0.854	0.866	0.866	11.7	0.853	0.864	8.0	0.847	0.871	10.3	0.846	0.870	7.0
(iii) Undet $H_2 : \beta_1<\beta_2<\beta_3$																
A1	24.6%	0.900	0.912	0.871	0.884	0.884	4.1	0.869	0.883	3.5	0.972	0.884	3.8	0.870	0.883	3.2
A2	21.8%	0.999	0.999	1.000	1.000	1.000	4.2	1.000	1.000	3.8	0.964	0.973	3.0	0.964	0.972	2.6
A3	21.2%	1.000	1.000	0.897	0.839	0.839	11.7	0.898	0.839	8.1	0.891	0.842	10.3	0.892	0.844	7.0

Table 2.(b)

When $K = 4$ (including results for unweighted ANOVA and Dunnett tests)

Hypothesis	Unweighted		Multinomial		Tree (I)		Tree (I+II)		Tree (I+II+III)	
	ANO	Dun	ANO	Dun	ANO	Dun	ANO	Dun	ANO	Dun
H_0	0.123	0.153	0.037	0.037	0.040	0.047	0.043	0.045	0.039	0.038
H_1	0.764	0.815	0.948	0.966	0.867	0.909	0.866	0.904	0.866	0.894
H_2	0.857	0.796	0.947	0.957	0.904	0.888	0.897	0.886	0.896	0.879

Table 3:

Regression estimates (EST), standard errors (SE) and p-values (PVAL) of logistic regression models for propensity analysis using TEA group as the reference group

Parameter	PCA vs. TEA			PVB vs. TEA		
	EST	SE	PVAL	EST	SE	PVAL
Intercept	−0.636	0.690	0.178	1.039	0.5766	0.036
BMI (<18.5 vs. 18.5–25 vs. 25)	0.235	0.275	0.197	−0.336	0.240	0.081
Smoking (yes/no)	−0.706	0.340	0.019	−0.631	0.303	0.019
Stage (<3 vs. 3)	−0.973	0.351	0.003	0.134	0.276	0.313
BT (yes/no)	1.353	0.316	0.000	−0.034	0.314	0.457

Table 4:

Matched data: distributions of covariates among three patient groups and p-values from χ^2 tests

		Multinomial Regression				Decision Tree			
Covariate		PCA	PVB	TEA	p-value	PCA	PVB	TEA	p-value
BMI	<18.5	4	3	4	0.889	3	3	1	0.861
	18.5–25	55	53	55		52	50	52	
	25	27	22	19		25	27	27	
Smoking	No	72	68	66	0.815	66	65	63	0.828
	Yes	14	10	12		14	15	17	
Stage	<3	66	58	59	0.939	62	62	62	1.000
	3	20	20	19		18	18	18	
BT	No	62	61	53	0.351	58	60	60	0.917
	Yes	24	17	25		22	20	20	