Communications for Statistical Applications and Methods
2020, Vol. 27, No. 6, 675–688

https://doi.org/10.29220/CSAM.2020.27.6.675
Print ISSN 2287-7843 / Online ISSN 2383-4757

Estimating causal effect of multi-valued treatment from observational survival data

Bongseong Kim[a], Ji-Hyun Kim[1,a]

[a]Department of Statistics and Actuarial Science, Soongsil University, Korea

Abstract

In survival analysis of observational data, the inverse probability weighting method and the Cox proportional hazards model are widely used when estimating the causal effects of multiple-valued treatment. In this paper, the two kinds of weights have been examined in the inverse probability weighting method. We explain the reason why the stabilized weight is more appropriate when an inverse probability weighting method using the generalized propensity score is applied. We also emphasize that a marginal hazard ratio and the conditional hazard ratio should be distinguished when defining the hazard ratio as a treatment effect under the Cox proportional hazards model. A simulation study based on real data is conducted to provide concrete numerical evidence.

Keywords: generalized propensity score, inverse probability weighting, Cox's proportional hazards model, marginal hazard ratio

1. Introduction

Randomized experiments identify the causal effects of new treatments or new drugs in medical or pharmacological studies. However, randomly assigning treatments to experimental units may not be possible due to ethical or operational difficulties. To estimate causal treatment effects from the data obtained from observational studies, the covariate distributions in treatment groups should be adjusted to be balanced, so that the treatment effects are not confounded with the effects of covariates, or pre-treatment variables. It is not easy to adjust covariate distributions when the number of covariates is high; however, the propensity scores proposed by Rosenbaum and Rubin (1983, RR83) can make the required adjustment effectively. RR83 and many succeeding studies covered a two-treatment-group case. However, there may be more than two treatment groups or levels, such as multiple doses of a drug. For a multi-valued treatment, it is not simple to extend the analysis procedure used for a binary treatment. Imbens (2000) extended the propensity scores of RR83 to a multi-valued treatment called generalized propensity scores. Methods for removing the confounding effect by matching or subclassification using generalized propensity scores were dealt with in Yang *et al.* (2016), and by inverse probability weighting (IPW) in Linden *et al.* (2015). However, these studies handled complete data, not right-censored data that is commonly observed in survival analyses.

This paper deals with the issue of estimating multi-valued treatment effect from the right-censored data obtained from observational study. We focus on IPW method. Austin (2013) discussed on the binary treatment effect using propensity scores. He compared the performance of matching, IPW and other methods through simulation. IPW methods are often used instead of matching when right

[1] Corresponding author: Department of Statistics and Actuarial Science, Soongsil University, 369 Sangdo-ro, Dongjak-gu, Seoul 06978, Korea. E-mail: jxk61@ssu.ac.kr

Published 30 November 2020 / journal homepage: http://csam.or.kr
© 2020 The Korean Statistical Society, and Korean International Statistical Society. All rights reserved.

censored data are given. Matching methods require a large pool of data to find pairs of units with similar characteristics. In survival analyses that require long-term follow-ups, it is not easy to get adequate data. The IPW method has relatively fewer problems with data size. The IPW method has strengths when the values of treatment and covariates change over time (Robins *et al.*, 2000); however, this study assumes they do not depend on time. The first topic of this paper is about the confusion on the weights used for the IPW method. The literature has different explanations for the rationale of stabilized weights (SW) and non-stabilized weights (NSW). Second, in survival analyses the marginal hazard ratio and the conditional hazard ratio are not properly distinguished when the treatment effect is defined as the hazard ratio (HR). These two topics are covered in this paper. When using the IPW method to estimate the causal effect from observational survival data, it is shown that it is more appropriate to apply the stabilized weights. The marginal and the conditional HR are different in meaning and values; therefore, it is emphasized that the distinction should be made.

The paper proceeds as follows. Section 2 describes the IPW method using generalized propensity scores. It also addresses the issue of marginal and conditional HRs. In Section 3, a simulation study based on real data clarifies the topics covered in Section 2. Section 4 summarizes the important points and also mentions limitations when analyzing observational survival data for causal reasoning.

2. Inverse probability weighting and causal hazard ratio

To be able to estimate causal treatment effects from observational data, we need some crucial assumptions. To describe those assumptions some definitions come first. The observed data for each unit i are (Y_i, A_i, X_i), which are assumed to be independent and identically distributed for $i = 1, \ldots, n$. We denote the response by Y, the treatment by A which can take one of K possible levels, the pre-treatment characteristics of the unit by X which can be a random vector of multiple variables (The subscript i will be omitted when we state the general property of the variable). Before dealing with censored data, let us first consider the complete response Y. To define the treatment effects, we follow the counterfactual approach in Rubin (1974, 1990), who set up the framework of causal reasoning in observational studies. We define the counterfactual of Y, denoted by $Y_{A=a}$, as the value of Y when $A = a$ is 'assigned' to a unit. Here, 'assigned' means that the unit did not choose a treatment level for itself, but rather the experimenter randomly determined the treatment level for the unit. The treatment effect of two treatment levels a and b can be defined as

$$\tau_{a,b} = E\left[Y_{A=a} - Y_{A=b}\right].$$

The mean of the difference in responses obtained when two different treatment levels are assigned to the same unit can be defined as the treatment effect. In an observational study, if the observed value of A of a unit is a, then Y is assumed to be equal to $Y_{A=a}$, which is referred to as the stable-unit-treatment-value assumption (Rubin, 1980).

Assumption 1. (Stable Unit Treatment Value Assumption; SUTVA) *If a unit is observed to receive treatment a, the observed Y of that unit will be equal to $Y_{A=a}$, that is $Y = \sum_{j=1}^{K} I(A = j)Y_{A=j}$. This assumption holds if there is no interference between units, so that the counterfactuals of each unit do not depend on the treatment status of other units.*

However, because Y for only one treatment level is observed for each unit, $Y_{A=a} - Y_{A=b}$ is not observed for each unit. Since $\tau_{a,b} = E[Y_{A=a}] - E[Y_{A=b}]$, the treatment effect can be estimated if the average of each possible outcome, or counterfactual, can be estimated, rather than the average of the differences between the two possible outcomes. Here the problem would be simple if the mean of the individuals

with $A = a$, i.e., $E[Y|A = a]$ equals $E[Y_{A=a}]$; however, for observational studies, the treatment level A is not randomly determined, so $E[Y|A = a]$ is different from $E[Y_{A=a}]$. That is

$$E[Y|A = a] = E[Y_{A=a}|A = a] \neq E[Y_{A=a}].$$

Therefore, $E[Y|A = a] - E[Y|A = b] \neq \tau_{a,b}$. This fact means that we cannot estimate the treatment effect with a simple estimate of the mean of observed Y by treatment level.

For causal reasoning to be possible in observational studies lacking randomization, a key assumption is required for X as follows.

Assumption 2. (Weak Unconfoundedness Assumption) *$Y_{A=a} \coprod I(A = a)|X$, that is when X is given, counterfactual $Y_{A=a}$ and indicator of the observed treatment level $I(A = a)$ are conditionally independent.*

A more strong assumption $(Y_{A=1}, \ldots, Y_{A=K}) \coprod A|X$ is not necessary here. Assumption 2, also called a weak version of 'no-unmeasured-confounders assumption', means that the data with the same value of X can be considered as randomized experimental data. To reveal causation rather than mere association between the treatment A and the response Y, the effect of every confounder, the common cause of A and Y, should be eliminated. The covariate vector X should contain all the confounders for the Assumption 2 to be satisfied. When the dimension of X is high, the propensity score plays an important role. For the treatment A with two levels, expressed as 0 and 1, RR83 defines the conditional probability $P(A = 1|X = x)$ as propensity score. Because this conditional probability is a function of x, it is sometimes denoted as $e(x)$. Imbens (2000) extended the propensity score into the case of two or more treatment levels, defining $P(A = a|X = x)$, $a = 1, 2, \ldots, K$ as the generalized propensity score. The generalized propensity score is marked as $e(a, x)$ because it is a function of the treatment level a and covariate x. The propensity score can be seen as a special case of the generalized propensity score; therefore, only the generalized propensity score is considered.

Assumption 3. (Overlap Assumption) *For all a and x, $0 < e(a, x) < 1$.*

This assumption indicates that every unit has a positive chance to be in any treatment group called an 'overlap-of-covariate-distributions assumption'. Combining this assumption with the strong version of Assumption 2, RR83 named it the 'strong-ignorability assumption'. Including as many observable pre-treatment variables as possible in the covariate vector X, such as demographic variables and pre-treatment health conditions, increases the possibility of satisfying the unconfoundedness assumption. However, a higher number of covariates increases the possibility that the overlap assumption will be violated. An important characteristic of the generalized propensity score is that if $Y_{A=a}$ and $I(A = a)$ are independent when vector X is given, they are also independent when only the scalar $e(a, X)$ is given (Imbens, 2000).

2.1. Non-stabilized weights and stabilized weights

We examine two kinds of weights in the IPW method to estimate the causal treatment effect. The inverse probability weights assigned to each unit are determined by the treatment level applied to the unit and the covariate vector representing the unit characteristics. For example, the inverse probability weight applied to a unit with $A = a$ and $X = x$ is $1/e(a, x)$. The inverse probability weight is the reciprocal of the generalized propensity score. However, we cannot say that it is the reciprocal of the propensity score, because if there are only two levels of treatment, the inverse probability weight for a unit with $a = 0$ is $1/e(0, x) = 1/(1 - e(x))$, which is not the reciprocal of the propensity score

$e(x)$. Since $e(a,x) = P(A = a|X = x)$ is a conditional probability for a treatment level, the inverse probability weights are sometimes called Inverse-Probability-of-Treatment-Weights (IPTW).

The inverse probability weight $w(a,x) = 1/e(a,x)$ become very large as $e(a,x)$ approaches zero, which also increases the variation of the statistics that reflect this weight. Robins *et al.* (2000) proposed to use stabilized weights $\text{sw}(a,x) = P(A = a)/e(a,x)$ to reduce the variation. However, there are other ways to reduce the variation and this explanation is less convincing as to why we use stabilized weights. We use a different perspective to explain the appropriate reason for the use of stabilized weights when censored data are given.

Hirano and Imbens (2001) explained why we need inverse probability weights for binary treatment case, which can be extended to multi-valued treatment case as follows.

$$\begin{aligned} E\left[\frac{YI(A=a)}{e(a,X)}\right] &= E\left[E\left[\frac{YI(A=a)}{e(a,X)}\middle|X\right]\right] \\ &= E\left[E\left[\frac{Y_{A=a}I(A=a)}{e(a,X)}\middle|X\right]\right] \\ &= E\left[\frac{E[Y_{A=a}|X]}{e(a,X)}E[I(A=a)|X]\right] \\ &= E[Y_{A=a}]. \end{aligned} \tag{2.1}$$

The second equality is established by Assumption 1, the third equality by Assumption 2, and the fourth equality by the definition of the generalized propensity score. The above expression indicates that if the parameter to be estimated can be expressed by the expected values of the counterfactuals, the observed data multiplied by the inverse probability weights may be considered and analyzed as counterfactuals.

However, if Y is censored, as in the survival analysis, the treatment effect may not be expressed by $E[Y_{A=a}]$. For the Cox proportional hazards model, which is widely used in the analysis of censored data, the treatment effect is defined by hazard ratio. Therefore, in the case of censored data, the validity of the inverse probability weights cannot be explained by the above results. Another rationale for the validity of the IPW method is the balancing of covariate distributions, which can be found in Ridgeway *et al.* (2014) and Li *et al.* (2018). These descriptions are for the binary treatment case, and can be extended to the multi-valued treatment case. Let $f_{X|A}(x|a)$ represent the distribution of X when $A = a$, and $f_X(x)$ represent the distribution of X in the entire population that covers all the treatment levels (If all variables in the covariate vector X are continuous, $f_X(x)$ will represent the joint density function). Suppose we want the observed distribution $f_{X|A}(x|a)$ to be unrelated to a by placing a weight of $w^*(a,x)$ in a unit of $A = a$ and $X = x$. We want the distribution of the confounders to be irrelevant to the level of treatment and to be the same as the distribution of data obtained from the randomized experiment. This can be expressed as

$$w^*(a,x)f_{X|A}(x|a) = f_X(x), \quad \text{for each } a.$$

Here, the weighting means that a new pseudo dataset is composed, in which the number of observations increases or decreases by the weights. However, this is different from multiplying each observation by weights as in (2.1). We can now derive from the above expression that the weights

$w^*(a,x)$ should be the stabilized weights $\mathrm{sw}(a,x)$.

$$\begin{aligned} w^*(a,x) &= \frac{f_X(x)}{f_{X|A}(x|a)} \\ &= \frac{f_X(x)}{f_{A,X}(a,x)/f_A(a)} \\ &= \frac{f_X(x)f_A(a)}{f_X(x)f_{A|X}(a|x)} \\ &= \frac{f_A(a)}{f_{A|X}(a|x)} \\ &= \frac{P(A=a)}{e(a,x)}. \end{aligned}$$

If non-stabilized weights $w(a,x) = 1/e(a,x)$ is used instead of $\mathrm{sw}(a,x)$, the covariate distribution of the pseudo dataset becomes

$$w(a,x)f_{X|A}(x|a) = \frac{f_X(x)}{P(A=a)} \propto f_X(x),$$

at treatment a. This fact means that knowing the treatment level does not change the covariate distribution, so the treatment and covariates become independent. The balancing of the covariate distributions in treatment groups is achieved even if non-stabilized weights are used.

Now we examine the size of pseudo dataset formed by two types of weights, the stabilized and the non-stabilized weights. The non-stabilized weights $1/e(a,x)$ is always greater than 1 because $e(a,x)$ is greater than 0 and less than 1. However, the stabilized weights may be less than 1. The use of NSW increases the size of pseudo dataset by K times over a given sample size on average, which can be seen as follows.

$$\begin{aligned} E[w(A,X)] = E\left[\frac{1}{f(A|X)}\right] &= \sum_{a=1}^{K}\int \frac{1}{f_{A,X}(a,x)/f_X(x)} f_{A,X}(a,x)dx \\ &= \sum_a \int f_X(x)dx = \sum_a 1 \\ &= K. \end{aligned}$$

A slight modification of the above derivation can show that $E[\mathrm{sw}(A,X)] = 1$. Therefore, $\sum_{i=1}^n \mathrm{sw}_i(a_i,x_i)$ is n on average. If we use SW, the size of the pseudo dataset is equal to the size of the original sample size on average. Xu *et al.* (2010) showed that $\sum_{i=1}^n w(a_i,x_i) = 2n$ and $\sum_{i=1}^n \mathrm{sw}(a_i,x_i) = n$ for a binary treatment case when X is a single categorical variable. Xu *et al.* (2010) demonstrated by simulations that the sum of the weights varies from data to data when X contains a continuous variable, but did not prove the average relationship.

Pseudo data sets formed by the two weights are not only different in size, but also different in proportions of treatment levels. The proportion of $A = a$ in the pseudo dataset obtained by the SW is, because the distribution of X in the area of $A = a$ is $f_X(x)$,

$$\int_{A=a} f_X(x)dx = P(A=a).$$

So, in pseudo dataset obtained by SW, the proportion of each treatment level is maintained as the same as in the observed data. However, for NSW, the distribution of X in the area of $A = a$ is $f_X(x)/P(A = a)$, therefore the proportion of $A = a$ in the pseudo dataset is,

$$\frac{\int_{A=a} f_X(x)/P(A = a)dx}{\sum_{j=1}^{K} \int_{A=j} f_X(x)/P(A = j)dx} = \frac{1}{K}.$$

For NSW, the proportions of all the treatment levels are the same at $1/K$. To summarize the discussion, the covariate distribution balance is achieved for either weight, while the stabilized weight generates a pseudo dataset in which the size and the proportion of each level of treatment remain the same as the original sample, the non-stabilized weight generates pseudo dataset in which the size of data for each level of treatment is as large as the size of the whole original sample.

A large literature, like Cole and Hernan (2004), first describes the non-stabilized weight, and then explains that the stabilized weight can be used to reduce the variance of the treatment effect estimated by the non-stabilized weight. However, you can reduce the variance by multiplying by an arbitrary number less than 1 or by truncating when you have a value greater than a specified value, for example, 10. Therefore, this explanation is not accurate as the rationale of the stabilized weight. We have shown that the stabilized weight is more appropriate than the non-stabilized weight in that both the size and the proportion of each treatment level remain the same while achieving the objective of covariate distribution balance.

2.2. Marginal hazard ratio and conditional hazard ratio

For censored data we usually do not represent the treatment effect as an average because the complete value may not be observed. The treatment effect on survival time can be expressed by using a hazard function in the Cox proportional hazards model:

$$\lambda_{T_{A=a}}(t) = \lambda_0(t) \exp(\beta_a), \tag{2.2}$$

where $\lambda_0(t)$ is a baseline hazard function that does not need to be specified. β_a, $a = 1, \ldots, K$ are parameters that represent the treatment effects, and since there are K categories, the number of parameters required is $K-1$. The parameter corresponding to the last baseline category is assumed to be 0, i.e., $\beta_K = 0$. $\lambda_{T_{A=a}}(t)$ is a hazard function of the counterfactual survival time $T_{A=a}$, which is obtained when the treatment a is randomly assigned to a unit. Therefore

$$\frac{\lambda_{T_{A=a}}(t)}{\lambda_{T_{A=b}}(t)} = \exp(\beta_a - \beta_b) \tag{2.3}$$

is the hazard ratio obtained when treatment a is assigned to one of the two randomly selected units, with treatment b assigned to the other, and can be interpreted as a causal treatment effect. $\exp(\beta_a)$ is the effect of the treatment a over the baseline treatment K because $\beta_b = 0$ when $b = K$. If observed data are given by simple observation, not by randomized experiment, then the model applied to this data is not $\lambda_{T_{A=a}}(t)$, but

$$\lambda_T(t|A = a) = \lambda'_0(t) \exp(\beta'_a). \tag{2.4}$$

We cannot interpret β'_a or $\exp(\beta'_a - \beta'_b)$ causally. However, if covariate vector X, which characterizes the units, satisfies Assumption 2, the pseudo dataset obtained by applying the inverse probability

weights becomes the data for which the confounding effect has been eliminated. Thus, the parameters obtained by applying the Cox proportional hazards model to the pseudo dataset can be interpreted causally.

However, assuming a Cox model for pseudo dataset, covariates may be included as:

$$\lambda_{T_{A=a}}(t|X = x) = \lambda_0(t) \exp(\theta_a + \gamma x). \tag{2.5}$$

The covariate vector x included in the above model uses the same symbol as the covariate vector x used to generate the pseudo dataset, but it may be a part of that covariate vector. In the above model, it is assumed that there is no interaction effect between the covariate X and treatment A on the counterfactual survival time $T_{A=a}$. The inclusion of interaction term in the model complicates interpretation because the treatment effect can be modified by the covariate value. The hazard ratio under the model (2.5)

$$\frac{\lambda_{T_{A=a}}(t|X = x)}{\lambda_{T_{A=b}}(t|X = x)} = \exp(\theta_a - \theta_b) \tag{2.6}$$

is the conditional hazard ratio. If the Cox model is fitted to the observed data, not the pseudo dataset, then the analysis model can be expressed as

$$\lambda_T(t|A = a, X = x) = \lambda'_0(t) \exp\left(\theta'_a + \gamma' x\right). \tag{2.7}$$

Generally $\theta_a \neq \theta'_a$, but if the covariate vector X satisfies Assumption 2 and the model (2.5) truly holds, then $\theta_a = \theta'_a$. This is because all confounders have been included in the model and controlled.

As Austin (2013) noted, the conditional causal hazard ratio (2.6) should be distinguished from the marginal causal hazard ratio (2.3). The conditional hazard ratio is obtained when different treatments are assigned to the same unit or units with the same attributes, while the marginal hazard ratio is obtained when different treatments are assigned to units with different attributes. The answer to which of the two hazard ratios is more appropriate as a measure of treatment effect depends on the situation. Randomized experiments assign treatments to units with different attributes; therefore, the marginal hazard ratio can be seen as a more appropriate measure if a randomized experiment is a proper procedure to estimate the causal effect. This may be the case if you want to know the effect of a public policy or the effect on all patients. However, a conditional hazard ratio may be a more appropriate measure if one wants to know which of the selectable treatments is more effective for individual, or if one wants to know the effect on units with a specific attribute. Therefore, the choice of the type of measure depends on the situation, such as the purpose of the study or the subject of interest.

2.3. Data analysis procedure

When censored data (A_i, X_i, Y_i, d_i), $i = 1, \ldots, n$ are given, the procedure for estimating the treatment effect by applying the IPW method is summarized. We observe $Y_i = \min(T_i, C_i)$, C_i represents the censoring time, and $d_i = I(T_i \leq C_i)$ is the indicator of complete data. We assume the survival time T and the censoring time C are independent, and the first two assumptions in Section 2 hold for T, not Y.

First, the multinomial logit model is fitted to estimate the general propensity score $P(A = a|X = x)$:

$$\log \frac{P(A = a|X = x)}{P(A = K|X = x)} = \alpha_a + \delta_a^T x, \quad a = 1, 2, \ldots, K - 1. \tag{2.8}$$

The above model imposes a constraint of $\alpha_K = \delta_K = 0$ for the baseline category parameters. The polynomial model may be expressed in another form as:

$$P(A = a|X = x) = \frac{\exp\left(\alpha_a + \delta_a^T x\right)}{\sum_{h=1}^{K} \exp\left(\alpha_h + \delta_h^T x\right)}, \quad a = 1, 2, \ldots, K.$$

If the programming language R is used as an analytical tool, the `multinom` function in the `nnet` package can be used to estimate $e(a, x) = P(A = a|X = x)$ (Venables and Ripley, 2002). For the estimation of $e(a, x)$ more flexible semi-parametric or non-parametric methods may be applied. The estimate of $P(A = a)$ is required to use stabilized weights. If a random sample has been taken, the sample proportion, $\sum_{i=1}^{n} I(A_i = a)/n$, can be used as the estimate of $P(A = a)$.

Fit the Cox model with only a treatment variable A

$$\lambda_{T_{A=a}}(t) = \lambda_0(t) \exp(\beta_a)$$

to the pseudo dataset generated using estimated stabilized weights. Then marginal hazard ratio $\lambda_{T_{A=a}}(t)/\lambda_{T_{A=K}}(t) = \exp(\beta_a)$, which represents the effect of the treatment a compared to the baseline treatment K can be estimated. If the model with covariate vector X in addition to treatment variable A

$$\lambda_{T_{A=a}}(t|X = x) = \lambda_0(t) \exp(\theta_a + \gamma x)$$

is fitted, a conditional hazard ratio $\lambda_{T_{A=a}}(t|X = x)/\lambda_{T_{A=K}}(t|X = x) = \exp(\theta_a)$ can be estimated. However, if a Cox model is fitted to the observed data rather than the pseudo dataset, the estimates of the treatment effects can be biased due to confounding. The weights are not integers, so that it is not clear how to generate a pseudo dataset. But the non-integer weight causes no problems if you use analysis methods that allow non-integer weighting, such as the `survival::coxph` function (Therneau, 2015). For a theoretical description of the analysis using non-integer weights, see Section 7.3 of Therneau and Grambsch (2000).

When you obtain estimates of the Cox model regression coefficient, β_a or θ_a, you obtain estimates of the hazard ratio, e^{β_a} or e^{θ_a}. Now let us consider the variance of the estimators. The authors of this paper recommend the use of stabilized weights for IPW method. If non-stabilized weights are used, the sample size is a problem. If there are K treatment levels, the size of pseudo dataset proves to be K times the size of the original data on average. Therefore, the standard error of the estimate will be underestimated by $1/\sqrt{K}$ times on average. The use of stabilized weights can eliminate sample size problems. However, the problem that observations in pseudo dataset are not independent still remains. Pseudo data constructed by applying the weights are not a random sample. In this case, robust standard error is recommended. The robust standard error was proposed by Lin and Wei (1989) as a less sensitive estimate when the model was incorrectly specified. Binder (1992) has extended this estimate and proposed a new robust standard error that can be used even if the sampling probabilities of elements are not equal or the samples are not independent. If you give the option `robust=TRUE` in the `survival:::coxph` function, you can obtain Binder's robust standard error. However, Binder's robust standard error assumes that the weights are known. The weights should be estimated in the actual analysis. However, as the following section shows, it seems that the impact of additional uncertainty from the estimation of weights does not matter much.

3. Simulation study

Simulations are conducted to confirm the contents of Section 2. The following is what we like to confirm through a simulation. First, when estimating causal effects from censored data obtained from

observational studies, we check how wrong the results can be if we simply estimate them without distinguishing between observational and causal studies. Next, we check how different the results will be depending on if the non-stabilized or stabilized weights are used when applying the IPW method. We also look at what is appropriate as the standard error of the estimate.

To ensure that the simulations are not too distant from the actual situation, the simulations have been done based on the real data. A model was fitted to the real data and then the fitted model was assumed to be the true model, and data for the simulation were generated from it. Matheson *et al.* (2012) conducted a study on the association between healthy lifestyle habits and mortality. Data in their study were used for simulation. National Health and Nutrition Examination Survey (NHNES) III data were used in the study of Matheson *et al.* (2012). However, the same data as in Matheson *et al.* (2012) could not be rebuilt because the NHNES III raw data were released (https://wwwn.cdc.gov/nchs/nhanes/nhanes3/default.aspx) and the refining process was not disclosed. In this simulation, it does not matter if data are exactly the same because the data compares the performance of the estimation methods rather than the exact magnitude of the causal effect of the healthy lifestyle habit on mortality. Also, the NHNES III data were not sampled with equal probability, but were considered random sample to simplify the simulation. From the date of registration between October 1988 and October 1994, 12,522 men and women over the age of 21 were followed up until December 31, 2006. The average follow-up time was 162 months and the total death toll was 3600. Healthy lifestyle habits (a treatment variable) is a categorical variable with five categories, which is a combination of diet, exercise, drinking and smoking habits of the observed at the time of registration. Values were given from 1 to 5, and level 5 is the category with the healthiest habit. The sample proportions of the five categories were 0.073, 0.253, 0.3474, 0.2230, and 0.0661, respectively. The response variable time is from the time of registration to the time of death or censoring; in addition, there are age, gender, race, marital status, and educational level as covariates, all of which are categorical variables.

Matheson *et al.* (2012) considered the association between healthy habits and mortality, but in this simulation we consider the causal effects of healthy habits on mortality. In order to create a situation in which covariate distribution varies by treatment level, treatment variable A, healthy lifestyle habits, is created using the following multinomial logit model. The covariate vector x consists of a bunch of dummy variables corresponding to five categorical covariates.

$$P(A = a|X = x) = \frac{\exp\left(\alpha_a + \delta_a^T x\right)}{\sum_{h=1}^{K} \exp\left(\alpha_h + \delta_h^T x\right)}, \quad a = 1, 2, 3, 4, 5. \tag{3.1}$$

The values of α_a and δ_a, the parameters of the above model, are obtained from the fitted model to the actual data.

In order to generate a response variable corresponding to the generated treatment $A = a$ and a given vector of covariates x, the survival time T and the censoring time C are generated using the model (2.5). The parameter values of the model were determined by the estimates obtained by fitting the model (2.7) to the actual data. Since the last category is the base category of treatment variable A, $\theta_5 = 0$, and $\theta_1 = 0.8273$, $\theta_2 = 0.5155$, $\theta_3 = 0.3657$, $\theta_4 = 0.2267$ as a result of fitting the model (2.7). Refer to Bender *et al.* (2005) and Kim and Kim (2018) for information on how to generate survival times and censoring times when the Cox proportional hazards model is assumed. The former proposed parametric methods, the latter proposed nonparametric methods, and this simulation has applied a nonparametric method.

The causal effect of healthy lifestyle on mortality is defined as hazard ratio. The focus shall be

on the marginal hazard ratio. This is because the conditional hazard ratio can be estimated without methodology for causal studies, such as the IPW, if the covariate vector satisfies the strong ignorability as described in the previous section. However, the marginal hazard ratio (2.3) cannot be estimated simply by fitting a model (2.4) to the observed data.

The model (2.5) for generating data is a model with covariates, and (2.6) is a conditional hazard ratio. We need to know the value of the marginal hazard ratio to compare the performance of various estimators. A randomized experiment is simulated to find the true marginal hazard ratio corresponding to the model (2.5). For this, the treatment A is not determined in accordance with the multinomial model (3.1), but is randomly determined. The probability of $A = a$ may be set $1/K$ for all a, but the observed proportion of $A = a$ in the actual data was used. The survival times are generated in accordance with model (2.5). The model (2.2) with no covariates is fitted to the data generated, and the marginal hazard ratio (2.3) is estimated. The average of the estimates obtained by repeated implementation of this process would be close to the true marginal hazard ratio. The number of repetitions is 5,000.

The overall procedure of the simulation is as follows.

0. The parameters of the true model for generating simulated data are determined by fitting the multinomial model and the Cox model to the given actual data. In addition, the true marginal hazard ratios are obtained by the process described above.

1. Generate treatment A_i^*, $i = 1, 2, \ldots, n$, from the model (3.1). Estimate the inverse probability weights, both $\mathrm{sw}(a_i^*, x_i)$ and $w(a_i^*, x_i)$, from the model obtained by fitting the multinomial model (2.8) to the generated data.

2. Generating T_i^* and C_i^* from the model (2.5), we get the response variable $Y_i^* = \min(T_i^*, C_i^*)$ and $d_i^* = I(T_i^* \le C_i^*)$. Apply the IPW method to the generated data $(A_i^*, X_i, Y_i^*, d_i^*)$, $i = 1, \ldots, n$, and then fit the Cox model (2.4). We also apply Cox model (2.4) without applying IPW method for comparison with the IPW method.

3. Repeat the previous Step 1 and Step 2 1,000 times to compare the performance of three estimators of treatment effects, one without using the IPW method, one with the stabilized weights, and one with non-stabilized weights.

To compare the point estimation performance, three measures, bias, relative bias and mean squared error, are computed. Relative bias is calculated as the absolute value of the bias divided by the true marginal hazard ratio, then expressed as a percentage. Actual coverage probability and length of confidence interval are also obtained to compare the interval estimation performance.

The true marginal hazard ratios obtained in Step 0, $\exp(\beta_a - \beta_5) = \exp(\beta_a)$, $a = 1, 2, 3, 4$, are 1.8273, 1.4741, 1.3222, and 1.1920, respectively (The standard errors, standard deviation divided by $\sqrt{5000}$, of the marginal hazard ratios are 0.0019, 0.0015, 0.0010, and 0.0009, respectively, indicating that the values obtained by simulations are close enough to the true values). Also note that the true conditional hazard ratios, $\exp(\theta_a)$, $a = 1, 2, 3, 4$, are 2.2872, 1.6744, 1.4415, and 1.2544, which are different from the marginal hazard ratios.

Without the application of the IPW method, the marginal hazard ratio cannot be properly estimated. It can be expected that biased estimates will be obtained when model (2.4) is fitted to simulated data without applying the IPW method. The performance of the estimator by this method is summarized in the first row of Table 1. It can be seen that bias occurs at all treatment levels, the

Table 1: Point estimation performance: MSE, bias, relative bias in %

	MSE				Bias (Relative Bias in %)			
	trt 1	trt 2	trt 3	trt 4	trt 1	trt 2	trt 3	trt 4
No IPW	0.0875	0.0215	0.0207	0.0142	−0.2560 (14.0)	−0.0927 (6.3)	−0.1041 (7.9)	−0.0726 (6.1)
IPW: $w(a,x)$	0.0303	0.0164	0.0127	0.0122	0.0044 (0.2)	0.0049 (0.3)	0.0037 (0.3)	0.0050 (0.4)
IPW: $\mathrm{sw}(a,x)$	0.0304	0.0165	0.0128	0.0112	0.0051 (0.3)	0.0050 (0.3)	0.0037 (0.3)	0.0050 (0.4)

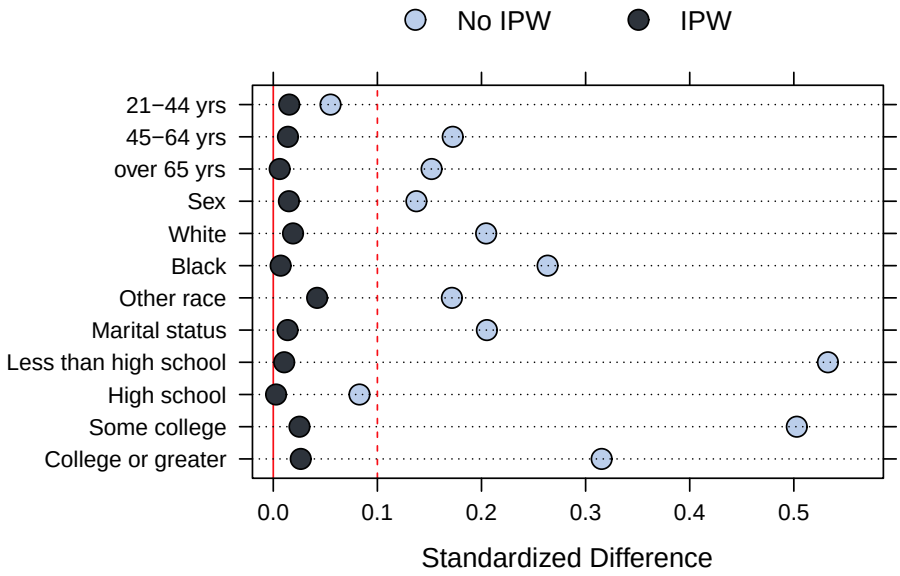

Figure 1: *Balance of covariate distribution before and after IPW: Each point represents a standardized difference between the maximum and minimum values of the proportions of each category of covariate at five treatment levels. The balance of covariate distribution has been achieved if the point lies below 0.1. We do not distinguish them because the two results of stabilized and non-stabilized weights coincide.*

relative bias reaching up to 14%, and the mean square error of this estimate gets large due to the large bias.

We now apply the IPW method, and report the performance of the two estimators obtained when SW and NSW are used. The performance of the two estimators are very similar in bias and mean square error. In the 1,000 simulations, we have found that 1,000 pairs of estimates at each of four levels were very similar. The pseudo dataset generated depends on the weight used (For example, if you use SW, the proportion of each treatment level is maintained as the same as in the observed data, but if you use NSW, the proportions are changed to be all equal). However, the covariate distribution balance in treatment levels is achieved no matter what weight is used (Figure 1). Consequently, there is little difference in the two point estimates of treatment effect by SW and NSW.

We now investigate the interval estimates of the treatment effects when using the IPW method. First, we compare the size of pseudo dataset. The average size of pseudo dataset is expected to be five times the sample size of 12,522, when non-stabilized weights are used. As expected, the average is 62610 with standard deviation of 7.2. When stabilized weights are used, the average is 12522 with standard deviation 0.5. Non-stabilized weights increase the size of pseudo dataset and therefore the

Table 2: Interval estimation performance: actual coverage probability of the 95% confidence interval

		trt 1	trt 2	trt 3	trt 4
IPW:	likelihood	0.420	0.456	0.454	0.473
$w(a,x)$	robust	0.978	0.974	0.972	0.973
IPW:	likelihood	0.957	0.946	0.946	0.942
$\text{sw}(a,x)$	robust	0.977	0.975	0.972	0.974

Table 3: Interval estimation performance: length of the 95% confidence interval (mean ± standard error)

		trt 1	trt 2	trt 3	trt 4
IPW:	likelihood	0.1796 ± 0.0006	0.1500 ± 0.0005	0.1371 ± 0.0005	0.1262 ± 0.0004
$w(a,x)$	robust	0.8149 ± 0.0029	0.5671 ± 0.0020	0.5021 ± 0.0018	0.4728 ± 0.0017
IPW:	likelihood	0.6821 ± 0.0025	0.4833 ± 0.0018	0.4300 ± 0.0016	0.4055 ± 0.0015
$\text{sw}(a,x)$	robust	0.8164 ± 0.0029	0.5678 ± 0.0020	0.5026 ± 0.0018	0.4732 ± 0.0017

standard error of the estimator is underestimated. An underestimated standard error would shorten the confidence interval and would not guarantee a nominal confidence level of 95%. The first row of Table 2 shows that the actual coverage probability dropped below 50%. From the second row, you can see that the problem is solved using the robust standard error. A data size problem does not occur if stabilized weights are used. The third row of Table 2 shows that, if you use stabilized weights, there is no problem with the actual coverage probability even if you do not use the robust standard error. However, the fourth rows of Table 2 and Table 3 show that the robust standard error slightly overestimates the standard error of the estimate, thereby increasing the actual coverage probability above the nominal confidence level and increasing the length. At least in this data there is no need to use the robust standard error if you use stabilized weights. However, when you use non-stabilized weights, you must use the robust standard error.

4. Conclusion

In survival analysis of observational data, the inverse probability weighting (IPW) method and the Cox proportional hazards model are widely used when estimating the effects of multiple-valued treatment. In this paper, the two kinds of weights have been examined in the IPW method, and the cautionary points have been emphasized when defining the treatment effect as a hazard ratio.

When the IPW method is applied in the survival analysis, the explanation of the valid reasons for the weights is confusing because different literature has a different perspective. In this paper we explain the IPW method from the aspect of composing a pseudo dataset where the confounding effect of covariates is eliminated. The two sets of pseudo data formed by the stabilized and the non-stabilized weights differ in terms of the proportion of the treatment level and the size, there is no difference in that the covariate distribution becomes balanced in treatment levels. It is recommended to use stabilized weights when using IPW method in survival analysis since there seems to be no reason to use non-stabilized weights.

It is important to distinguish between causal and associational hazard ratios when using the Cox model for survival data; however, it is also important to distinguish between marginal and conditional hazard ratios. Austin (2013) distinguished the two hazard ratios, but in some cases, it goes without clearly distinguishing them, as in Sugihara (2010). We emphasized the difference between the two hazard ratios again. The marginal hazard ratio cannot be estimated by fitting a Cox model with only a treatment variable or by fitting a model with covariates to the observed data. Causal inference methods, such as the IPW method, should be applied.

The above facts were confirmed through simulations based on real data. The standard error of

the estimator must be estimated in order to perform hypothesis tests on the treatment effect or to obtain confidence intervals. You have to use the robust standard error when non-stabilized weights are applied. It was noted that the robust standard error may overestimate the true standard error when the stabilized weights are applied, but this cannot be generalized due to the limitations of the simulation.

All methods of causal reasoning from the observational study have limitations of relying on the assumption of strong ignorability. Sensitivity analysis for the assumption is sometimes performed because it is not possible to test from observed data whether the assumption holds or not, but sufficient research has not been done yet on the sensitivity analysis in survival analysis.

References

Austin PC (2013). The performance of different propensity score methods for estimating marginal hazard ratios, *Statistics in Medicine*, **32**, 2837–2849.

Bender R, Augustin T, and Blettner M (2005). Generating survival times to simulate Cox proportional hazards models, *Statistics in Medicine*, **24**, 1713–1723.

Binder DA (1992). Fitting Cox's proportional hazards models from survey data, *Biometrika*, **79**, 139–147.

Cole SR and Hernan MA (2004). Adjusted survival curves with inverse probability weights, *Computer Methods and Programs in Biomedicine*, **75**, 45–49.

Hirano K and Imbens GW (2001). Estimation of causal effects using propensity score weighting: an introduction to data on right heart catheterization, *Health Services & Outcomes Research Methodology*, **2**, 259–278.

Imbens GW (2000). The role of the propensity score in estimating dose-response functions, *Biometrika*, **87**, 706–710.

Kim J and Kim B (2018). Generating censored data from Cox proportional hazards models, *The Korean Journal of Applied Statistics*, **31**, 761–769.

Li F, Morgan KL, and Zaslavsky AM (2018). Balancing covariates via PS weighting, *Journal of the American Statistical Association*, **113**, 390–400.

Lin DY and Wei LJ (1989). The robust inference for the Cox proportional hazards model, *Journal of the American Statistical Association*, **84**, 1074–1078.

Linden A, Uysal SD, Ryan A, and Adams JL (2015). Estimating causal effects for multivalued treatments: a comparison of approaches, *Statistics in Medicine*, **35**, 534–542.

Matheson EM, King DE, and Everett CJ (2012). Healthy lifestyle habits and mortality in overweight and obese individuals, *Journal of the American Board of Family Medicine*, **25**, 9–15.

Ridgeway G, McCaffrey D, Morral AR, Burgette LF, and Grin BA (2014). Toolkit for weighting and analysis of nonequivalent groups: A tutorial for the twang package.

Robins JM, Hernan M, and Brumback B (2000). Marginal structural models and causal inference in epidemiology, *Epidemiology*, **11**, 550–560.

Rosenbaum PR and Rubin DB (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika*, **70**, 41–55.

Rubin DB (1974). Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology*, **66**, 688–701.

Rubin DB (1980). Comment on "Randomization Analysis of Experimental Data: The Fisher Randomization Test," by D. Basu, *Journal of the American Statistical Association*, **75**, 591–593.

Rubin DB (1990). Formal modes of statistical inference for causal effects, *Journal of Statistical Planning and Inference*, **25**, 279–292.

Sugihara M (2010). Survival analysis using inverse probability of treatment weighted methods based on the generalized propensity score, *Pharmaceutical Statistics*, **9**, 21–34.

Therneau TM (2015). A Package for Survival Analysis in S. version 2.38, Available from: https://CRAN.R-project.org/package=survival

Therneau TM and Grambsch PM (2000). *Modeling Survival Data*, Springer.

Venables WN and Ripley BD (2002). *Modern Applied Statistics with S* (4th Ed), Springer, New York.

Xu S, Ross C, Raebel MA, Shetterly S, Blanchette C, and Smith D (2010). Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals, *Value in Health*, **13**, 273–277.

Yang S, Imbens GW, Cui Z, Faries DE, and Kadziola Z (2016). Propensity score matching and subclassification in observational studies with multi-level treatments, *Biometrics*, **72**, 1055–1065.

Received August 11, 2020; Revised September 30, 2020; Accepted November 10, 2020