Application of inverse probability weights in survival analysis

Guoqiao Wang, PhD,[a] and Inmaculada Aban, PhD[a]

[a] Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL

Received Apr 23, 2015; accepted Apr 23, 2015
doi:10.1007/s12350-015-0157-9

See related article, pp. 600–607

Suppose that a researcher is interested in comparing two ''treatments''—A and B—and how the treatment affects an outcome of interest. The ideal study design would be to conduct a randomized trial where treatment assignment is randomly assigned. The random treatment assignment aims to make the subjects between the two treatments similar, i.e., it aims to balance any differences in baseline characteristics, so that any differences in outcome may be attributed to the treatment. The challenge with randomized trials is that they are typically costly and sometimes impractical. Observational studies are less costly and more practical for researchers. However, even with a well-designed observational study, subjects in different treatment groups are not likely to be comparable with respect to their baseline characteristics. For instance, consider the study discussed in the paper by Farzaneh-Far et al published in this issue comparing the prognostic value of using two different agents in myocardial perfusion imaging.[1] In this case, the two ''treatments'' of interest are the agents (regadenoson vs adenosine) used to determine the values of summed stress scores (SSS) and the summed difference scores (SDS). It is of interest to test whether the predictive ability of SSS and SDS on the composite outcome of time to cardiovascular death or myocardial infarction will depend on what agent is used. Some of the baseline characteristics in the adenosine group are not similar to those in the regadenoson group, e.g., there is significantly lower percentage of women and higher percentage of smokers and diabetics in the adenosine group. This issue of imbalance if not addressed in the analyses may result in misleading conclusions about the treatment effect due to potential selection bias and possible confounding variables.

How then can the data be used to enable a direct, meaningful, and valid comparison between the two treatments when subjects in the two groups are dissimilar? The more traditional method is to include the baseline variables showing significant differences between the treatment groups as covariates in the multivariable regression model that investigates the treatment effect. When a fitted univariate regression model with only treatment as the variable in the model shows significant treatment effect and, after adjusting for the baseline characteristics in a multivariate model, the treatment effect is still significant, then one has stronger evidence to conclude that there is a significant treatment effect. However, in studies where sample sizes may be small relative to the number of unbalanced variables, this method may not work, or worse, may not be appropriate.

An alternative method of addressing the issue of imbalance is the use of propensity scores which can overcome some of the shortcomings of the aforementioned method of adjusting using covariates in a regression model.[2,3] Propensity score is defined as the probability of an individual being assigned to one of two treatments given all information (e.g., baseline characteristics) available *before* assignment.[4] These scores are estimated based on the data collected such that individuals with similar baseline covariates would have similar scores, and vice versa. Thus, individuals with similar propensity scores are comparable except for the treatment assignment.

Propensity scores are used in the analyses in different ways: for matching to identify similar subjects between treatment groups; for defining strata based on the scores where one may perform stratified analysis; as a covariate in a regression model; and being used to obtain the inverse probability weights (IPW). The use of IPW in regression modeling, in particular to survival

Reprint requests: Guoqiao Wang, PhD, Department of Biostatistics, University of Alabama at Birmingham, 1720 Second Avenue South, Birmingham, AL 35294-0022; *guoqiao@uab.edu*.
J Nucl Cardiol 2015;22:611–3.
1071-3581/$34.00

model, to adjust for the imbalance in the demographic variables will be the main focus of the rest of this editorial. The idea of using IPW is to weigh individuals by the inverse of their propensity scores so that those with higher propensity scores(viewed as over-represented) will be assigned a lower weight and those with lower propensity scores(viewed as under-represented) will be assigned a higher weight.[5] Therefore, using IPW will generate a ''pseudo-sample'' in which the imbalanced set of covariates becomes balanced between treatment groups.[3] To illustrate using a simplistic hypothetical study, suppose it is of interest to compare two groups with identical baseline characteristics except for gender. Group A has 3 females while group B has 7. Therefore, the propensity of being in group A for a female is 0.3 (=3/10) and the propensity of a female being in group B is 0.7 (=7/10). Using IPW, the 3 females in group A and the 7 in group B will be assigned weights of 1/0.3 and 1/0.7, respectively. The resulting pseudo-sample has a total of 20 females equally distributed between the groups, i.e., $3*(1/0.3) = 10$ females in group A and $7*(1/0.7) = 10$ females in group B.

How can we obtain estimates of the propensity scores in a general setting where there are more than one variable involved? For the reason that the outcome of interest in the propensity scores is binary (belonging to group A or B), the most commonly used method of estimating propensity scores is by fitting a multivariate logistic regression model with the treatment indicator as the dependent variable and other covariates measured *before* the treatment assignment as the independent variables. Although there is no consensus as to which covariates to include in the logistic model, in most circumstances, it is appropriate to include all measured pre-treatment baseline characteristics regardless if they are balanced or not between the two groups.[2] Including post-treatment characteristics in this logistic model should be considered with great caution and should be included only if they are not affected by the treatment.[2] In most settings, the IPW is obtained by simply taking the inverse of the estimated propensity scores (i.e., 1/score) (henceforth, referred to as original IPW). Unfortunately, using original IPW may artificially increase the total sample size as previously seen in the gender example. Furthermore, in cases where individuals have extremely small propensity scores, IPW will be large (e.g., when PS is 0.0001, the IPW would be 1000) that the estimation of the treatment effect is then dominated by these few observations with very large weights.[6] As a result, this can lead to a noticeable increase in the variances of estimated effects.[7] Fortunately, one may address this issue and achieve stabilization in the modeling by redefining the IPW as the ratio of the marginal probability of being treated to the propensity score (henceforth, referred to as stabilized IPW).[8] The marginal probability of being treated is the proportion of individuals in the treatment group. For example, if 10 of a total 40 actual samples were in the treatment group, and then the marginal probability of being treated is 0.25, which is the same for all 10 individuals regardless of their baseline characteristics. Additionally, using stabilized IPW will not artificially increase the sample size.[8]

The paper by Afshin et al is an excellent illustration of the application of stabilized IPW in survival analysis. Logistic model was used to obtain the propensity scores with 8 covariates measured before the agent administration, namely: age, gender, race, diabetes, hypertension, smoking, hyperlipidemia, and history of myocardial infarction. Both SSS (unbalanced at baseline) and SDS (balanced at baseline) are consequences of the agent used, and hence, should not be included in the estimation of propensity scores. Evidence that the use of stabilized IPW worked to balance all 8 baseline characteristics included in the logistic model is shown in Table 4 of supplementary section of the paper. Kaplan-Meier curves (Figures 1 and 2) presented to compare the estimated survival curves of the two agents are based on the stabilized IPW adjusted data so that any differences that may be observed are not confounded by the differences in the baseline characteristics of the subjects in each agent. Without the use of IPW and given the imbalance at baseline, constructing Kaplan-Meier curves to compare the curves for the two agents will not be appropriate. The authors were also interested in investigating how the prognostic ability of the SSS and SDS is affected by the type of agent used. This was achieved by fitting a proportional hazards survival model with SSS, agent, and SSS and agent interaction terms in the model using adjusted data based on stabilized IPW. The resulting model found no significant interaction between SSS and agent, and therefore the authors concluded that there is no evidence that the agent modifies the prognostic ability of SSS (or SDS) with regard to their outcomes of interest. Note that it was no longer necessary in this case to add any of the baseline characteristics as covariates in the survival model. Using IPW resulted to a model that is simpler, more parsimonious, and more efficient with regard to the number of variables in the model.

References

1. Farzaneh-Far A, Shaw LK, Dunning A, O'Connor CM, Borges-Neto S. Comparison of the prognostic value of Regadenoson and Adenosine maocardial perfusion imaging. J Nucl Cardiol (in press).
2. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivar Behav Res 2011;46:399–424.

3. Heinze G, Jüni P. An overview of the objectives of and the approaches to propensity score analyses. Eur Heart J 2011;32:1704–8.
4. Luellen JK, Shadish WR, Clark M. Propensity scores an introduction and experimental test. Eval Rev 2005;29:530–58.
5. Xie J, Liu C. Adjusted Kaplan-Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. Stat Med 2005;24:3089–110.
6. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res 2013;22:278–95.
7. Freedman DA, Berk RA. Weighting regressions by propensity scores. Eval Rev 2008;32:392–409.
8. Cole SR, Hernán MA. Adjusted survival curves with inverse probability weights. Comp Methods Programs Biomed 2004;75:45–9.