Balancing Treatment Comparisons in Longitudinal Studies

Sue M. Marcus, PhD; Juned Siddique, DrPH; Thomas R. Ten Have, PhD; Robert D. Gibbons, PhD; Elizabeth Stuart, PhD; and Sharon-Lise T. Normand, PhD

Sue M. Marcus, PhD, is with the Department of Psychiatry, Mount Sinai School of Medicine, New York. Juned Siddique, DrPH, is with the Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago. Thomas R. Ten Have, PhD, is with the Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia. Robert D. Gibbons, PhD, is Professor of Biostatistics and Psychiatry, and is Director of the Center for Health Statistics, University of Illinois at Chicago. Elizabeth Stuart, PhD, is with the Johns Hopkins Bloomberg School of Public Health, Baltimore. Sharon-Lise T. Normand, PhD, is with the Department of Health Care Policy, Harvard Medical School, and the Department of Biostatistics, Harvard School of Public Health, Boston.

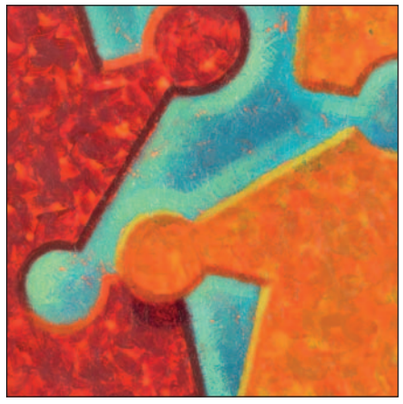
Address correspondence to Sue Marcus, PhD, Department of Psychiatry, Mount Sinai School of Medicine, One Gustave L. Levy Place, Box 1230, New York, NY 10029; fax 212-860-4630; or e-mail sue.marcus@mssm.edu.

Dr. Marcus, Dr. Siddique, Dr. Ten Have, Dr. Gibbons, Dr. Stuart, and Dr. Normand, have disclosed no relevant financial relationships.

Evaluation of treatment efficacy in psychiatric trials involves a comparison of outcomes for those who receive a treatment versus those who receive a control or comparison treatment. However, if the treatment and comparison groups are not comparable or "balanced," observed differences in outcomes between treated and comparison groups may be due, in part, to these imbalances. In such settings, estimates of treatment efficacy may be biased unless some adjustments are made to make the groups more comparable.

The preferred method of treatment assignment is randomization. Randomization ensures, on average, a balance of observed and unobserved baseline characteristics between those assigned to treatment and those assigned to the comparison group. In the absence of randomization, however, treatment groups could differ on the basis of both observed and non-observed characteristics. Longitudinal observational studies, studies that repeatedly measure outcomes on participants, are subject to additional analytic challenges. However, a) treatment groups may differ at baseline,[1] and b) treatment groups may quickly become less comparable over the course of the study due to subject dropout, treatment switching, noncompliance, and missing data.[2] Thus, estimates of treatment efficacy in longitudinal studies may result in over- or underestimates, unless comparisons can be balanced.

In this article, we show how techniques used in observational cross-sectional studies can be used to balance comparisons in longitudinal studies. We build upon the methods presented in Stuart et al (to be published in an upcoming issue of *Psychiatric Annals*), which will describe methods for balancing comparisons in cross-sectional trials.[3] We discuss when to adjust during the course of the study (eg, baseline, posttreatment, follow-up assessments, etc.). We also present methods for creating propensity scores (scalar summaries that measure how likely a subject is to receive the treatment rather than the control) for longitudinal studies such that the comparisons are balanced and inference is not compromised by adjusting for variables that have been affected by treatment.[4] Throughout this article, we assume that the treatment is assigned once and it is not time-varying.



A major goal of research in developmental psychopathology is to measure the effect of an event on a subsequent course of development.

Three longitudinal studies in which propensity scores are used to balance treatment comparisons illustrate our methods (see Table 1, page 807). These studies include the Runaway Youth Study, an intervention study aimed at preventing HIV transmission among runaway youths housed at shelters in New York City;[1] the Gang-joining Study, a Montreal-based study that evaluated the effect of gang joining at age 14 on subsequent violence; and the MTA Follow-up Study, an observational follow-up to the randomized Multimodal Treatment Study of Children with Attention Deficit Hyperactivity Disorder (MTA).[2] Each study has a different design and requires balancing at different points along the longitudinal course.

THREE LONGITUDINAL STUDIES

The Runaway Youth Study

The Longitudinal Study of HIV Risk among Runaway Youths involved assessing an intervention to prevent HIV transmission among runaway youths 11 to 18 years who were housed at shelters in New York City during 1988-1991.[1] The design involved randomizing shelters rather than youths: two shelters were randomized to receive an intervention involving staff training and a series of interactive group sessions (n = 167 youths); the other two shelters did not receive any intervention (n = 144). The primary outcome was the number of unprotected sexual acts measured at five follow-up time-points. The same measure was also available for the 3 months before baseline. Many participants had missing assessments, leading to unbalanced longitudinal outcome measures. Youths in the intervention and comparison groups differed on nine sociodemographic and substance-use characteristics: youths at control shelters were older, had a higher school dropout rate, and had more severe drug and alcohol use. Unadjusted analyses would result in an overestimate of the intervention efficacy because these sociodemographic and substance-use characteristics were positively associated with the outcome.

The Gang-joining Study

The gang-joining illustration uses data from the Montreal Longitudinal-Experimental Study of Boys.[5] A total of 1,037 boys from 53 schools in the lowest socioeconomic area of Montreal, Canada, were studied longitudinally from kindergarten until 17 years. We use information from 580 participants who reported no involvement with gangs from 11 through 13 years and also had

TABLE 1.

Three Longitudinal Studies

Study	Treatment	Comparison	Outcome	Repeated Measures	Confounders
Runaway Youth Study	Intervention to prevent HIV transmission	No intervention	Number of unprotected sexual acts	3-month period before baseline and 5 follow-up assessments	Age, dropout status, drug and alcohol use
Gang-joining Study	Joining a gang at age 14	Not joining a gang at age 14	Level of violence	11 to 17 years	Prior violence, popularity, aggression, hyperactivity, oppositionality, number of sexual partners
The MTA Follow-up Study	High medication use	Low medication use	ADHD symptoms (hyperactivity/impulsivity)	0, 14, 24, 36 months	Prior ADHD symptoms, sex, grade, ethnicity, prior medication use, family intactness, family income, emotional and behavioral problems

one or no missing measurements of their violence, delinquency, and gang involvement during this time period. Haviland et al[5] studied the effect of joining a gang at age 14 on those adolescent males who had not joined a gang prior to 14 years. Because none of the boys had been gang-joiners before 14 years, their levels of violence before 14 years were not impacted by gang membership. The primary outcome is level of violence.

Like the first study, there were imbalances in the youths who joined gangs at 14 years and those who did not: age 14 joiners were different from non-joiners in that they experienced greater violence, aggression, and hyperactivity at earlier ages. Analyses that do not consider these imbalances could lead to an overestimate of the effect of gang-joining on subsequent violence rates because, like the runaway youth study, these covariates are predictive of the outcome (in this case, violence).

The MTA Follow-up Study

The MTA is a large, randomized, federally-funded [by both the National Institute of Mental Health (NMIH) and the Department of Education] treatment study comparing the efficacy of behavioral and/or medication treatment for children with attention-deficit hyperactivity disorder (ADHD).[6] A group of 579 children with ADHD 7 to 9.9 years, were randomly assigned to 14 months of medication management (titration followed by monthly visits); intensive behavioral treatment (parent, school, and child components, with therapist involvement gradually reduced over time); the two combined; or standard community care (treatments by community providers). The study also conducted a passive 10-month follow-up evaluation to examine the continuing impact of the randomized treatment after 24 months. After the 24-month period, the sample was followed in a naturalistic longitudinal design for an additional 8 years (to 15 to 18 years). The primary outcome is level of hyperactivity/impulsivity.

We consider whether imbalances with respect to the group of children with ADHD who had low medication use compared with the group who had high medication use at the 36-month assessment contributed to the lack of medication advantage.[2] It was hypothesized that children with relativity more severe psychopathology at baseline or during the follow-up would be more likely to receive a higher level of medication after the initial randomized 14-month period. It was also hypothesized that the association of severity and long-term medication use would result in selective long-term treatment of the most severe cases, which could mask the potentially beneficial long-term effects of the medication.

ADJUSTING FOR BIAS IN LONGITUDINAL STUDIES: WHEN TO BALANCE

The design and context of the study determine which comparisons over time need to be balanced. In longitudinal observational studies, it is often necessary to adjust for baseline imbalances between those belonging to the treatment group and those belonging to the comparison group. For example, the analysis of the longitudinal intervention study for runaway youths at shelters in New York City sought to balance important baseline characteristics such as high school dropout and substance use.[1] This baseline was the same for all individuals, making it easy to define the groups and match on variables known to be unaffected by the program.

A major goal of research in developmental psychopathology is to measure the effect of an event on a subsequent course of development. For example, Haviland et al[5] estimated the effects of

TABLE 2.

MTA 36-month Medication Use (low/high) by Propensity Score Quintile

Treatment Group	Quintile 1 (lowest probability of high medication)	Quintile 2	Quintile 3	Quintile 4	Quintile 5 (highest probability of high medication)	All
Low med	71 (75%)	47 (49%)	37 (39%)	23 (24%)	5 (5%)	183 (39%)
High med	24 (25%)	48 (51%)	57 (61%)	72 (76%)	89 (95%)	290 (61%)
All	95 (100%)	95 (100%)	94 (100%)	95 (100%)	94 (100%)	473

Entries are number of participants and column percent.

gang-joining on subsequent violence. It is both unethical and infeasible to randomize adolescents to joining a gang versus not, so a quasi-experimental study was designed. Comparison subjects who had not joined a gang by 14 years were matched to first-time gang joiners at 14 years. The boys in the study were divided into trajectory groups based upon earlier violence from 11 to 13 years and within these groups, joiners were matched to a variable number of controls. Within these balanced groups, gang membership status was evaluated at 14 years and also from 15 to 17 years. A more complex design could have used risk set matching, in which joiners at any age (not just 14 years) are matched with controls who have not yet joined gangs at the same age.[7,8] Risk set matching matches people who are at risk of joining a gang at any age and follows them so that the matched sets can start at different ages and will be followed during different observation periods. The advantage to using risk set matching rather than matching only at 14 years is that the risk set matching provides additional information as matched sets start at different ages and are followed for different periods of time.

The MTA study assessed the effect of the randomized treatment at 14 months and 24 months, then followed children in a naturalistic observational study for a total of 8 years. A naturalistic follow-up study of a randomized controlled trial is a type of hybrid design combining aspects of both efficacy and effectiveness, as recommended by a NIMH report.[9] The MTA Follow-up Study planned to assess the efficacy of medication at each follow-up time (36 months, 72 months, and 84 months) and to disseminate the results of analyses from each time-point sequentially. Thus, it was desired to balance the low and high medication use groups at each of these follow-up assessments. In this article on balancing longitudinal treatment comparisons, we use the results from the 36-month assessment.[1,10]

VALID LONGITUDINAL INFERENCE

When specifying a longitudinal propensity score for balancing comparisons, it is essential that inference is not compromised by adjusting for variables that may have been affected by treatment.[4] Generally, estimates of treatment effects will be biased when there is adjustment for a covariate that is an outcome of treatment rather than a pre-treatment covariate. For example, adjusting for intensity of treatment can produce biased treatment effects because more intensive treatment is often provided to more treatment resistant patients.

When propensity score matching for longitudinal studies occurs at baseline only, as in the longitudinal study of HIV risk among runaway youths, adjusting for posttreatment covariates is not an issue. However, in the other two studies, the propensity score is specified so that covariates are established before treatment initiation and outcomes are subsequent to treatment initiation.

Haviland et al[5] assess the effects of gang joining at 14 years on violence 14 to 17 years, while controlling for characteristics that occur from 11 to 13 years, including the trajectory of violence from 11 to 13 years. Because none of the boys in the study were in gangs before 14 years, there is no matching on characteristics that may have been impacted by gang membership at 14 years.

Propensity score analyses of the MTA 36-month follow-up study matched low and high medication use groups on baseline characteristics and severity of ADHD symptoms at 14 months and 24 month and evaluated 36-month ADHD symptom severity within matched quintiles.[2] We note that this strategy yields valid longitudinal inference for the estimated effect of medication at 36 months, rather than the effect of medication from baseline to 36 months. Some of those receiving medication at 36 months may have been receiving it for years, while others may have been receiving it for just a few weeks or months.

SPECIFYING A LONGITUDINAL PROPENSITY SCORE

With careful selection of pretreatment covariates and posttreatment outcomes, the problem of specifying a propensity score in a longitudinal study can be translated into that of specifying

a propensity score in a cross-sectional study. For example, the propensity score analysis of the MTA 36-month follow-up examines the propensity to take a high level of medication as a function of baseline characteristics and ADHD symptoms at 14 and 24 months. Because the goal is to determine the efficacy of high-level medication use at 36 months (rather than baseline through 36 months), the longitudinal follow-up study can be viewed as a cross-sectional study taking place at 36 months. Consequently, the propensity score analyses described in Stuart et al can be applied.[3]

COMBINING LONGITUDINAL METHODS AND PROPENSITY SCORE MATCHING

There are various ways of combining longitudinal methods with propensity score matching to evaluate outcomes within balanced treatment comparison groups. Longitudinal observational studies that match treatment groups at baseline can use any longitudinal model (eg, mixed-effects regression or finite mixture modeling of trajectories) to compare matched sets.

For example, Song et al[1] used linear mixed-effects regressions fitted within each propensity score subclass. To get a combined estimator across subclasses, they used a stratified estimator with the results weighted to the number of cases in the subclasses.

Haviland et al[5] used finite mixture modeling to identify groups of boys with similar trajectories of violence from 11 to 13 years. They used propensity score matching to match on these trajectories as well as other characteristics. For example, the propensity to join a gang at 14 years was estimated conditional on violence trajectory prior to 14 years, peer-rated popularity, and other covariates. Within matched sets, this analysis evaluates violence at each age (14, 15, 16, and 17) separately; however, alternate analyses could use violence classes defined by trajectories as outcomes.

Other techniques for combining propensity scores with longitudinal models have been suggested (eg, the use of mixed-effects ordinal logistic regression to distinguish among participants who receive various ordered doses of treatment across time).[11,12] We note that these techniques differ from those that are described in this paper as they do not make it clear in a transparent way that a) mathematically, the groups will be balanced, on average and b) the covariates for adjustment are established prior to the start of treatment.

Marginal structural models (MSMs) and their associated estimation method (inverse probability weighting, or IPW) represent an extension of standard propensity score approaches to accommodate time-varying changes in treatment under observational longitudinal studies with a final endpoint outcome.[13] Such an approach will yield unbiased estimation of the effect of the time-varying treatment on the final endpoint outcome, if all potential time-varying confounders of this effect are measured and incorporated into the estimation approach. In standard cross-sectional propensity score contexts, the standard approach of correctly adjusting for the confounders as covariates in the outcome model will yield unbiased estimates of treatment effects on outcomes under appropriate conditions. However, this is not possible with standard methods for longitudinal data, because there is not a correct model specification of

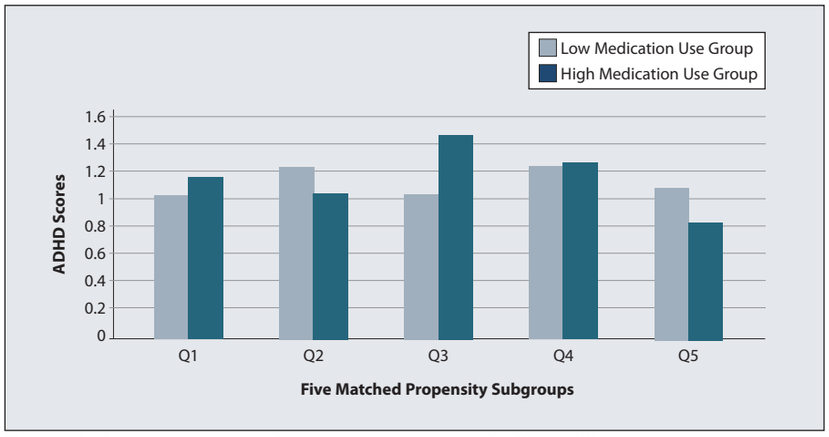


Figure 1. All five matched propensity subgroups showed no significant difference between the low and high medication use groups with respect to the 36-month ADHD assessment.

time-varying confounders and treatment when the treatment influences confounders.[14] The resulting bias of standard adjustment methods arises from treating the confounders only as confounders when in effect they are both confounders and mediators, given the time-varying nature of the confounders and treatment. That is, the confounders do influence treatment and outcome, but the confounders are also impacted by previous treatment. The interchanging of treatment and confounders across time and the resulting switching of roles for the confounder is accommodated by the IPW estimation scheme, which essentially recreates a sequentially randomized trial, where subjects are randomized at baseline and then re-randomized at subsequent times.[15] The recreation of such a sequentially randomized trial is achieved with a two-stage estimation process, where the second stage is the specification and IPW estimation of the MSM with the endpoint outcome as the dependent variable. The first stage is the construction of the weights for the IPW estimation.

The MSM is a cross-sectional model with the final endpoint outcome as the dependent variable and a summary score of treatment history as the covariate. The summary score of treatment has taken a variety of forms. Standard MSM approaches use a summary score across time. Under certain assumptions, the structural aspect refers to the causal effect of the summary of treatment on outcome. The marginal aspect pertains to the absence of time-varying covariates from the model, although baseline covariates can be included in the model. Several assumptions are necessary for interpreting the effect of the treatment summary on outcome as causal. First, there are no unmeasured confounders (ie, all variables that are associated with both the outcome and treatment are measured). Secondly, the MSM is correctly specified. Third, we assume that the series of propensity score models across time are correctly specified. In comparison, analogous standard statistical models assume that time-varying confounders of current and future treatment and outcome are not impacted by previous treatment. This assumption is not required when using a MSM with the IPW estimation approach. The weights only adjust for the relationship between current treatment and current and past confounding but not future confounding unlike standard regression methods.

Estimation of the MSM is achieved with IPW estimation using standard software that allows weights (eg, Proc Logistic in SAS). The weights are obtained from propensity scores in the first stage by fitting a separate logistic model at each point in time for which the amount of treatment received is measured. For each logistic model, treatment at a given time point is modeled as a function of current and previous confounders, yielding predicted values or propensity scores for that particular time point. The weights of the IPW estimation approach are then constructed by multiplying these propensity scores across the time points at which treatment is measured with the resulting product inverted to produce the final weight for each subject. In this way, the weights create a pseudo-population in which previous and current covariates are balanced between the subjects of the different treatment groups at a particular time. However, this pseudo-randomization created by weighting does not balance covariates between the treatment groups in the future. In this way, the approach adjusts for confounding but not mediation of the time-varying covariates. Because of the possibility of small numbers of subjects within a treatment at a given time, the propensity scores at a particular time point may be small leading to unstable weights. As in survey sampling, weight stabilization methods have been used to accommodate such cases.

As an illustration of propensity-score matching in a longitudinal context, we consider the MTA 36-month follow-up analysis, which used propensity score matching to evaluate whether baseline characteristics and ADHD symptom severity at 14 and 24 months were associated with selection of drug treatment, thereby masking drug effects at the 36-month follow-up assessment. ADHD symptom severity at 36 months was evaluated within propensity score quintiles. Table 2 (see page 808) gives medication use at 36 months by propensity quintile (those in Quintile 1 have the lowest propensity to take medication, while those in Quintile 5 have the greatest propensity to take medication). We note that there is little overlap in Quintile 1 (and

also Quintile 5) between those who did and did not take medication. This means that children with the greatest risk factors are very likely to receive a high level of medication and very unlikely to be in the low medication group. On the other hand, children with the fewest risk factors are more likely to receive a low level of medication and are very unlikely to take a higher level of medication. Those in the high medication group tended to be male, younger, non-white, previous medication users before the start of the study and had a higher level of ADHD symptomatology at baseline.[2]

All five matched propensity subgroups showed no significant difference between the low and high medication use groups with respect to the 36-month ADHD assessment (see Figure 1, page 810). Although we did find selection bias, we failed to confirm our hypothesis that selection bias is masking the potentially beneficial effect of medication at 36 months.

DISCUSSION

We have shown how the problem of specifying propensity scores for longitudinal assessments can be thought of as similar to specifying propensity scores in cross-sectional studies, if we take caution not to adjust for variables that have been affected by treatment (see Sidebar for key recommendations). Thus, we must consider the same issues that apply to all propensity score analyses, such as the importance of identifying whether there are substantially overlapping covariate distributions and whether hidden bias might change the conclusions.

Haviland et al[5] found three violence trajectory groups from 11 to 13 years: a low-violence group (46%), a group for which violence declines with increasing age (48%), and a chronic group (about 6% of the population) who experienced consistently high levels of violence. Their results suggested that it would be difficult to balance the gang joiners in the chronic group. Therefore, it was not possible to estimate the effect of gang membership for those with a chronic trajectory as there was no suitable control group.

On the other hand, propensity score analyses for the MTA Follow-up showed overlap across quintiles. All five propensity score quintiles showed initial advantage of medication that disappeared by 36 months and consequently did not support the hypothesis that self-selection masked a beneficial medication effect at 36 months.

Propensity score matching adjusts for observed factors related to treatment selection and balances on those observed characteristics that are used to calculate the propensity score. However, hidden bias due to unobserved characteristics (residual bias after accounting for overt bias) is always possible.[16] We recommend the use of a sensitivity analysis when the significance of a treatment effect is questioned due to possible hidden bias. For example, Haviland et al[5] showed that the effect of gang joining on violence at 14 years is not sensitive to small biases but is sensitive to moderate biases.

SIDEBAR.

Key Recommendations

1. Identify plausible confounders of treatment selection and outcomes.
2. Identify appropriate time-point for balancing given context of study.
3. Establish covariates prior to balancing time-point.
4. Create propensity scores such that points 2 and 3 hold.
5. Match on propensity score so that balance is "transparent."
6. Assess balance after matching.
7. Assess overlap of matches.
8. Assess impact of treatment adjusting for matches.
9. Consider sensitivity to bias from unobserved covariates.

REFERENCES

1. Song J, Belin TR, Lee MB, Gao X, Rotheram-Borus MJ. Handling baseline differences and missing items in a longitudinal study of HIV risk among runaway youths. *Health Services and Outcome Research Methodology*. 2001;2:317-329.
2. Swanson JM, Hinshaw SP, Arnold LE, et al. Secondary evaluations of MTA 36-month outcomes: propensity score and growth mixture model analyses. *J Am Acad Child Adolesc Psychiatry*. 2007;46(8):1002-1013.
3. Stuart EA, Marcus SM, Horvitz-Lennon MV, Gibbons RD, Normand S-LT. Using non-experimental data to estimate treatment effects. *Psychiatric Annals*. In press.
4. Rosenbaum PR. The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society*. Series A. 1984;147:656-666.
5. Haviland A, Nagin DS, Rosenbaum PR. Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychol Methods*. 2007;12(3):247-267.
6. MTA Cooperative Group (Multimodal Treatment Study of Children with ADHD). A 14-month clinical trial of treatment strategies for attention-deficit/hyperactivity disorder. *Arch Gen Psychiatry*. 1999;56(12):1073-1086.
7. Li YP, Propert KJ, Rosenbaum PR. Balanced risk set matching. *Journal of the American Statistical Association*. 2001;96:870-882.
8. Lu B. Propensity scores with time dependent covariates. Biometrics. 2005;61:721-728.
9. National Advisory Mental Health Council. Bridging Science and Service: a report of the National Advisory Mental Health Council's Clinical Treatment and Services Research Work Group. Rockville, MD; 1999.
10. Jensen PS, Swanson JM, Arnold LE, et al. Three-year follow-up of the NIMH MTA study. *J Am Acad Child Adolesc Psychiatry*. 2007;46(8):988-1001.
11. Leon AC, Hedeker D. A mixed-effects quintile-stratified propensity adjustment for effectiveness analyses of ordered categorical doses. *Stat Med*. 2005;24(4):647-658.
12. Leon AC, Hedeker D, Teres JJ. Bias reduction in effectiveness analyses of longitudinal ordinal doses with a mixed-effects propensity adjustment. *Stat Med*. 2007;26(1):110-123.
13. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11(5):550-560.
14. D'Agostino RB, Lee ML, Belanger AJ, Cupples LA, Anderson K, Kannel WB. Relation of pooled logistic regression to time dependent Cox regression analysis: The Framingham Heart Study. *Stat Med*. 1990;9(12):1501-1515.
15. Murphy SA. An experimental design for the development of adaptive treatment strategies. *Stat Med*. 2005;24(10):1455-1481.
16. Rosenbaum PR. *Observational Studies*. 2nd ed. New York, NY: Springer-Verlag; 2002.