STATISTICS IN MEDICINE
*Statist. Med.* 2007; **26**:110–123
Published online 12 December 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2458

# Bias reduction in effectiveness analyses of longitudinal ordinal doses with a mixed-effects propensity adjustment¶

Andrew C. Leon[1,*,†], Donald Hedeker[2,‡] and Jedediah J. Teres[1,§]

[1]*Department of Psychiatry, Weill Medical College of Cornell University, New York, NY 10021, U.S.A.*
[2]*Division of Epidemiology & Biostatistics, University of Illinois at Chicago, Chicago, IL, U.S.A.*

SUMMARY

A mixed-effects propensity adjustment is described that can reduce bias in longitudinal studies involving non-equivalent comparison groups. There are two stages in this data analytic strategy. First, a model of propensity for treatment intensity examines variables that distinguish among subjects who receive various ordered doses of treatment across time using mixed-effects ordinal logistic regression. Second, the effectiveness model examines multiple times until recurrence to compare the ordered doses using a mixed-effects grouped-time survival model. Effectiveness analyses are initially stratified by propensity quintile. Then the quintile-specific results are pooled, assuming that there is not a propensity × treatment interaction. A Monte Carlo simulation study compares bias reduction in fully specified propensity model relative to misspecified models. In addition, type I error rate and statistical power are examined. The approach is illustrated by applying it to a longitudinal, observational study of maintenance treatment of major depression. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: treatment effectiveness; propensity adjustment; longitudinal study; observational study; misspecification

## 1. INTRODUCTION

An observational study of treatment effectiveness is typically challenged by non-equivalent comparison groups. That is, when subjects are not randomized to a treatment, there are selection biases that contribute to outcome. In clinical settings, for instance, patients who receive

*Correspondence to: Andrew C. Leon, Department of Psychiatry, Weill Medical College of Cornell University, Box 140, 525 East 68th Street, New York, NY 10021, U.S.A.
†E-mail: acleon@med.cornell.edu
‡E-mail: hedeker@uic.edu
§E-mail: jjt2002@med.cornell.edu
¶Presented, in part, at the Meeting of the Eastern North American Region (ENAR) of the International Biometric Society, 28–31 March 2004, Pittsburgh, PA, U.S.A.

Contract/grant sponsor: NIH; contract/grant numbers: MH60447, MH68638

*Received 23 December 2004*
*Accepted 17 October 2005*

the most intensive dose of an antidepressant are likely to be the most severely ill; whereas those receiving lower doses tend to be less severely ill. An unadjusted analysis would likely conclude that higher doses are less effective. This is because the factors that contribute to choice of treatment are confounds in that they are also associated with outcome. In addition, there are other fundamental aspects of observational data that complicate the treatment effectiveness analyses. For example, neither the dose nor the duration of treatment is determined by the investigator. Moreover, a longitudinal observational study will likely have repeated measures on each subject.

The propensity adjustment is a general approach to reduce the effects of confounding variables on effectiveness analyses [1] that can be implemented through matching, subclassification, or covariance adjustment. In an effort to apply the propensity strategy to longitudinal, observational studies, we proposed a dynamic adaptation of the propensity adjustment for ordinal doses [2, 3]. It is a two-stage, mixed-effects data analytic strategy that includes a model of propensity for treatment intensity and a model of treatment effectiveness. In stage one, the propensity model examines repeated measures of ordinal treatment doses over time. The model can incorporate multiple treatment intervals per subject and variations in both within-subject treatment intensity and within-subject propensity for treatment intensity during the course of the study. In stage two, the treatment effectiveness model examines time from the start of each treatment until relapse. The mixed-effects framework accounts for the correlated within-subject relapse times, which correspond to the successive within-subject treatment intervals.

Our initial implementation of the propensity score with longitudinal data involved covariate adjustment that incorporated four vectors to represent the propensity quintiles [2]. We have since applied the approach with analyses that were stratified by propensity quintile [3] and subsequently examined type I error, statistical power and bias of the quintile-stratified approach [4]. We now focus on bias reduction from the mixed-effects propensity adjustment. Initially, this article describes the longitudinal model of propensity for treatment intensity. We then discuss a mixed-effects treatment effectiveness model that is stratified by propensity quintiles. The approach is illustrated with an application to a longitudinal study of antidepressant treatment for relapse prevention. Finally, a simulation study examines the performance of this data analytic strategy.

## 2. MIXED-EFFECTS PROPENSITY FOR TREATMENT INTENSITY MODEL

The model of *propensity for treatment intensity* characterizes the longitudinal ordinal doses of treatment based on demographic and clinical features. Specifically, a mixed-effects ordinal logistic regression examines the repeated ordinal doses as a function of these subject characteristics. This approach assumes that an ordered categorical scale of therapeutic equivalents has been developed, such as that used for treatments of mood disorders [5, 6].

Rosenbaum and Rubin [1] defined the propensity score as 'the conditional probability of assignment to a particular treatment given a vector of observed covariates'. They have shown that propensity score adjustments can be used to reduce the bias in estimates of treatment effectiveness in an observational study [1, 7]. Adapting the Rosenbaum and Rubin notation, the mixed-effects *propensity for treatment intensity score*, for the $k$th ordinal dose denoted

 

by the variable $T$, is

$$e_k(x_{ij}, \upsilon) = P(T_{ij} > k | \upsilon, x)$$

for subject $i\,(i = 1, \ldots, N)$, at time $j\,(j = 1, \ldots, J)$, for dose $k\,(k = 1, \ldots, K-1)$. This is derived from a mixed-effects ordinal logistic regression model that includes covariates and random subject-specific effects in terms of the $k-1$ cumulative logits [8]:

$$\ln\left[\frac{P(T_{ij} > k)}{1 - P(T_{ij} > k)}\right] = \gamma_k + \alpha + x'_{ij}\beta + \upsilon_i$$

where $\gamma_k$ represents the threshold for dose $k$, $\alpha$ is the intercept, $x_{ij}$ is the $p \times 1$ vector of covariates, $\beta$ includes the corresponding regression coefficients, and $\upsilon_i$ is the subject-specific random effect that is normally distributed in the population with mean 0 and variance $\sigma^2_\upsilon$. This subject-specific random effect is included in the model to account for the within-subject clustering (i.e. repeated observations within subjects). Both time-varying and time-invariant covariates can be included in the vector $x$. The $k-1$ thresholds are strictly increasing parameters, and with the intercept in the model it is common to set the first threshold equal to zero for identification purposes.

Assuming this mixed-effects ordinal logistic model, the propensity score, which ranges in value from 0 to 1, can be expressed using the logistic response function for subject $i$ at time $j$ as

$$e(x_{ij}, \upsilon_i) = \frac{\exp(\alpha + x'_{ij}\beta + \upsilon_i)}{1 + \exp(\alpha + x'_{ij}\beta + \upsilon_i)}$$

The above propensity is specifically for the first threshold, and one could calculate $k-1$ propensity scores, each indicative of the $k-1$ cumulative response probabilities. However, since the covariate effects and random subject effects are constant across the cumulative logits (i.e. proportional odds assumption) the results would not change for purposes of ranking propensity scores. Thus, here we define the propensity score for the ordinal dose in terms of the probability of responding in categories greater than the first category (i.e. in terms of the first threshold). As indicated above, the propensity score embodies the contribution of covariates $x$ and subject effects $\upsilon$ on the probability of receiving a more intensive dose of treatment (i.e. a higher value in terms of the ordinal dose $T$). An observation with a high propensity score presents as someone quite likely to receive intensive treatment at time point $j$, whereas an observation with a low propensity score has characteristics of someone less likely to receive intensive treatment.

*Propensity quintiles classification.* Each observation for subject $i$ at time $j$ is classified into a propensity quintile, $q_{(1)}, \ldots, q_{(5)}$, based on the corresponding propensity score. In an effort to remove confounding effects of the variables comprising the propensity score, treatment effectiveness is conducted separately for each quintile. Prior to conducting quintile-specific analyses, however, one must determine whether all treatments are well-represented in each quintile. This is because a treatment that is not represented in a particular quintile, of course, cannot be evaluated in treatment effectiveness analyses of that quintile. Quintile representation is evaluated by examining the frequencies in a treatment by propensity quintile contingency table. The amount of uncertainty around the treatment effect will naturally increase as the $N$'s

 

decrease, yet there are no formal guidelines regarding this issue. We proceed with analyses if there are at least 5 to 10 observations per cell.

### 2.1. Mixed-effects treatment analyses

*Effectiveness analyses stratified by propensity quintiles*. Once the quintiles are formed, separate quintile-specific treatment effectiveness analyses are conducted. The rationale underlying this strategy is that stratification on a confounding variable will remove the associated bias [9]. Based on the quintile-specific results, pooled estimates of treatment effectiveness are obtained under the condition that there is not a significant quintile $\times$ treatment interaction. This condition can be assessed, as described below, in a combined analysis that includes model terms for treatment, quintile, and quintile $\times$ treatment interactions.

In terms of treatment effectiveness, the time until recurrence of disease or event is the dependent variable, modelled using survival analysis methods. Here, because the measurement of survival is ascertained in time intervals (e.g. did the event occur since the last follow-up?) grouped-time survival analysis is appropriate. Additionally, since subjects can have multiple observations (i.e. multiple recurrences of the event) across the time period of the study, a mixed-effects grouped-time survival model [10] is used. This model examines the probability of *recurrence* up to, and including, time interval $t$ for subject $i$ and observation $j$ as

$$P_{ijt} = P(t_{ij} \leqslant t)$$

Using the complementary log–log link function yields a proportional hazards regression model describing the cumulative probability of recurrence as a function of treatment:

$$P_{ijt} = 1 - \exp(-\exp(\alpha_t + x'_{ij}\beta + \upsilon_i))$$

where $\alpha_t$ represents the intercept terms (i.e. the baseline hazard), $x$ is a vector of covariates (dummy coded to represent the treatment groups), $\beta$ is a vector of coefficients, and $\upsilon_i$ represents a random subject effect accounting for the clustering of repeated occurrences within subjects.

*Pooling the quintile-specific results*. The results from the quintile-specific models are pooled using the Mantel–Haenszel procedure as described by Fleiss [11]. Using this approach, each quintile-specific parameter estimate is weighted by the inverse of its squared standard error and those weighted estimates are summed for each treatment. As noted, this approach assumes that there is not a treatment $\times$ propensity interaction. The assumption is tested by comparing the fit of two mixed-effects grouped-time survival models, each of which includes observations from all quintiles: (1) the main effects only model (treatment and propensity quintile) and (2) the main effects model plus propensity $\times$ treatment interaction. The incremental contribution of the propensity $\times$ treatment interaction is tested with a likelihood ratio test and, if the interaction is statistically significant, the assumption is violated. Such an interaction would indicate that the treatment effect varies across propensity quintiles, and thus pooling quintile-specific results is not viable. Instead, conclusions regarding treatment effectiveness must be quintile-specific in this case.

 

## 3. APPLICATION

The National Institute of Mental Health Collaborative Depression Study (CDS) enrolled 955 subjects with affective disorders from 1978–1981 who sought treatment for one of the major affective disorders (major depressive disorder, mania, or schizoaffective disorder) at one of five academic medical centres in the United States (Boston, Chicago, Iowa City, New York, and St. Louis). Each subject provided written informed consent. The objectives and design of this longitudinal, observational study have been described previously [12]. Among those there were 431 subjects who had major depressive disorder at intake [13]. Eighty-six of those subjects developed bipolar disorder and another 43 of them did not recover from their intake episode. Here, we examine the effectiveness of somatic antidepressant treatment for relapse prevention among those who recovered from unipolar major depression as defined by the Research Diagnostic Criteria (RDC: [14]). These subjects have been followed for up to 20 years.

The propensity-adjusted analyses proceeded in two stages, as described above. Initially, analysis of propensity for treatment intensity examined characteristics that distinguished among those receiving various levels of somatic antidepressant treatment. In this propensity analysis, treatment intensity was the ordinal-dependent variable (ranging from 0 to 3, described in detail in References [3, 5] and covariates included several demographic and clinical variables that were hypothesized to be associated with treatment intensity. As mentioned, since subjects had multiple observations of treatment intensity across time, a mixed-effects ordinal logistic regression model [8] was used.

In the second stage, treatment effectiveness analyses examined time from the start of the course of a particular intensity of treatment until recurrence of a major depressive episode using the aforementioned mixed-effects grouped-time survival model. Survival time represented the 'time until recurrence', defined as the number of consecutive weeks during which treatment remained at one level of intensity during a 'well' period. A survival interval terminated in one of three ways: (1) recurrence of depression (2) change in antidepressant treatment intensity or (3) end of follow-up. The latter two were classified as censored. Censoring due to the end of follow-up was assumed to be unrelated to time until recurrence. Each subject accumulated additional survival intervals, which we refer to as 'treatment intervals', with each new episode and each change in treatment intensity while in episode. The unit of analysis for both the propensity and effectiveness models was treatment interval and a separate propensity score was calculated for each treatment interval. All analyses were performed with MIXOR [15] software.

### *3.1. Results*

*Propensity for treatment intensity*. The analyses included 1782 observations on 296 subjects. (Of the 302 subjects who otherwise met criteria for these analyses, six were excluded due to missing data necessary for the propensity model.) The propensity model included 6 independent variables (Table I). Those with more severe symptoms, those with primary major depression at intake into the CDS, and those with more prior episodes tended to get more intensive maintenance treatment, as did those who got more intensive treatment in their more recent episode. The younger subjects were significantly less likely to get aggressive treatment. Likewise, subjects from three of the study sites (New York, Iowa, and Chicago) tended to get

 

Table I. Mixed-effects ordinal logistic regression model of propensity for treatment intensity.

| Variable | *b* | *SE* | Odds ratio | CI low | CI high | $\chi^2$ | df | *Z* | *p* |
|---|---|---|---|---|---|---|---|---|---|
| Symptom severity (range: 1–6) | 0.16 | 0.04 | 1.17 | 1.09 | 1.26 | | | 4.31 | <0.001 |
| Primary depression at intake | 0.22 | 0.16 | 1.25 | 0.91 | 1.73 | | | 1.36 | 0.17 |
| Number prior episodes | | | | | | 16.88 | 4 | | 0.002 |
| 1 | | | 1.00 | | | | | | |
| 2 | −0.32 | 0.16 | 0.73 | 0.53 | 0.99 | | | −2.04 | 0.04 |
| 3 | −0.32 | 0.16 | 0.72 | 0.52 | 1.00 | | | −1.98 | 0.05 |
| 4 | −0.28 | 0.21 | 0.76 | 0.50 | 1.14 | | | −1.35 | 0.18 |
| 5+ | 0.26 | 0.19 | 1.30 | 0.90 | 1.87 | | | 1.42 | 0.16 |
| Age (years) | | | | | | 26.05 | 4 | | <0.001 |
| <30 | −0.77 | 0.17 | 0.46 | 0.33 | 0.65 | | | −4.49 | <0.001 |
| 30–39 | | | 1.00 | | | | | | |
| 40–49 | 0.09 | 0.16 | 1.09 | 0.80 | 1.48 | | | 0.55 | 0.58 |
| 50–59 | 0.15 | 0.19 | 1.17 | 0.80 | 1.71 | | | 0.79 | 0.43 |
| 60+ | 0.26 | 0.24 | 1.29 | 0.80 | 2.08 | | | 1.06 | 0.29 |
| Education | | | | | | 10.31 | 3 | | 0.016 |
| <12 Years | | | 1.00 | | | | | | |
| High school grad | 0.57 | 0.25 | 1.77 | 1.09 | 2.87 | | | 2.33 | 0.02 |
| Some college | 0.19 | 0.23 | 1.21 | 0.77 | 1.90 | | | 0.82 | 0.41 |
| College & higher | 0.64 | 0.26 | 1.89 | 1.14 | 3.14 | | | 2.47 | 0.01 |
| Study site | | | | | | 24.96 | 4 | | <0.001 |
| Boston | | | 1.00 | | | | | | |
| New York | 0.92 | 0.36 | 2.51 | 1.24 | 5.08 | | | 2.57 | 0.01 |
| St. Louis | 0.14 | 0.26 | 1.15 | 0.69 | 1.92 | | | 0.53 | 0.60 |
| Iowa | 0.96 | 0.25 | 2.61 | 1.60 | 4.26 | | | 3.86 | <0.001 |
| Chicago | 0.63 | 0.30 | 1.88 | 1.04 | 3.39 | | | 2.10 | 0.04 |

more intensive treatment than those from other sites. The intensity of treatment was not highly consistent within subjects across treatment intervals (intraclass correlation coefficient = 0.19).

A propensity score was computed for each observation based on the results of the model in Table I. Observations were then classified into one of five propensity score quintiles. The propensity for treatment tended not to vary widely within subject, 47.6 per cent of the subjects remained in just one quintile (across their repeated observations) and another 36.5 per cent were in two quintiles. The representativeness of treatments in each quintile was evaluated by examining the quintile × treatment contingency table (Table II). As expected, observations in the lower quintiles tend to receive less intensive doses and, conversely, those in higher quintiles tend to receive more intensive doses. Nevertheless, each dose is represented in each quintile; therefore, quintile-specific treatment effectiveness analyses compared the four doses.

*Balance.* The purpose of the propensity adjustment is to adjust for imbalance in demographic and clinical characteristics across treatment groups (i.e. doses). The extent of balance was examined by comparing unadjusted and propensity-adjusted mixed-effects ordinal logistic regression models of ordinal dose. Separate models evaluated the strength of the association of each of the six variables that form the propensity score with dose. The results are presented in Table III, which compares the adjusted and unadjusted odds ratios (95 per cent CI) and

 

Table II. Cross-classification of ordinal treatment dose by propensity quintile.

| | Propensity quintile | | | | | |
|---|---|---|---|---|---|---|
| Dose | 1 | 2 | 3 | 4 | 5 | Total |
| 0 | 256 | 175 | 126 | 69 | 31 | 657 |
| 1 | 59 | 106 | 85 | 80 | 34 | 364 |
| 2 | 27 | 53 | 98 | 111 | 116 | 405 |
| 3 | 14 | 21 | 51 | 94 | 176 | 356 |
| Total | 356 | 355 | 360 | 354 | 357 | 1782 |

Cell entries represent frequencies of observations.
Data are based on 1782 treatment intervals (i.e. observations) from 296 subjects.

Table III. A comparison of unadjusted and propensity-adjusted mixed-effects ordinal logistic regression models of ordinal doses.

| | Unadjusted results | | | | Propensity-adjusted results | | | |
|---|---|---|---|---|---|---|---|---|
| Variable* | Odds ratio | CI low | CI high | *p*-value | Odds ratio | CI low | CI high | *p*-value |
| Symptom severity (range: 1–6) | 1.15 | 1.08 | 1.24 | <0.0001 | 1.08 | 1.02 | 1.14 | 0.009 |
| Primary depression at intake | 1.30 | 0.95 | 1.78 | 0.103 | 1.12 | 0.89 | 1.42 | 0.346 |
| Number prior episodes | | | | <0.001 | | | | 0.028 |
| 1 | 1.00 | | | | | | | |
| 2 | 0.88 | 0.65 | 1.19 | 0.390 | 1.02 | 0.78 | 1.33 | 0.882 |
| 3 | 0.87 | 0.64 | 1.17 | 0.350 | 1.01 | 0.77 | 1.32 | 0.938 |
| 4 | 1.04 | 0.71 | 1.51 | 0.850 | 1.04 | 0.74 | 1.46 | 0.843 |
| 5+ | 1.83 | 1.34 | 2.51 | <0.001 | 1.59 | 1.22 | 2.07 | 0.001 |
| Age (years) | | | | <0.0001 | | | | 0.954 |
| <30 | 0.51 | 0.37 | 0.71 | <0.001 | 0.99 | 0.73 | 1.34 | 0.949 |
| 30–39 | 1.00 | | | | 1.00 | | | |
| 40–49 | 1.11 | 0.86 | 1.42 | 0.435 | 1.01 | 0.80 | 1.29 | 0.913 |
| 50–59 | 1.31 | 0.90 | 1.90 | 0.156 | 1.13 | 0.83 | 1.54 | 0.450 |
| 60+ | 1.34 | 0.87 | 2.07 | 0.188 | 1.01 | 0.74 | 1.36 | 0.971 |
| Education | | | | 0.025 | | | | 0.837 |
| <12 Years | 1.00 | | | | 1.00 | | | |
| High school grad | 1.70 | 1.06 | 2.71 | 0.027 | 1.16 | 0.83 | 1.63 | 0.388 |
| Some college | 1.20 | 0.78 | 1.85 | 0.409 | 1.11 | 0.81 | 1.52 | 0.518 |
| College & higher | 1.83 | 1.10 | 3.06 | 0.020 | 1.10 | 0.78 | 1.56 | 0.584 |
| Study site | | | | <0.0001 | | | | 0.106 |
| Boston | 1.00 | | | | | | | |
| New York | 2.89 | 1.45 | 5.74 | 0.002 | 1.20 | 0.72 | 1.98 | 0.490 |
| St. Louis | 1.30 | 0.79 | 2.13 | 0.302 | 0.93 | 0.62 | 1.40 | 0.717 |
| Iowa | 2.61 | 1.61 | 4.24 | <0.001 | 1.35 | 0.91 | 1.99 | 0.138 |
| Chicago | 2.49 | 1.41 | 4.41 | 0.002 | 1.16 | 0.76 | 1.77 | 0.484 |

*Separate models evaluated each of the six variables.

 

$p$-values. Balance was achieved with four of the variables; whereas for symptom severity and number of prior episodes, the imbalance was greatly reduced, but not eliminated.

*Treatment effectiveness analyses.* Results from the quintile-specific treatment effectiveness analyses were pooled as described above because the treatment × propensity interaction was not statistically significant ($\chi^2 = 6.146$, df $= 12$, $p = 0.909$). The effectiveness of each of the doses was then contrasted with dose 0. These results indicate that those treated with higher doses of somatic antidepressant therapy (dose 3) were half as likely to have a recurrence than those who received no somatic treatment (dose 0) (odds ratio: 0.50; 95 per cent confidence interval: 0.30–0.84; $Z = -2.60$; $p = 0.009$), after controlling for propensity for treatment intensity. In contrast, moderate doses (dose 2) were associated with marginal protection (OR: 0.65; 95 per cent CI: 0.41–1.01; $Z = -1.92$; $p = 0.055$); whereas lower doses (dose 1) were not associated with significant protection from recurrence (OR: 0.98; 95 per cent CI: 0.65–1.48; $Z = -0.09$; $p = 0.929$). In an observational study treatment effectiveness such as this, it is prudent to control for hypothesized confounding variables. Nevertheless, for comparative purposes, an unadjusted antidepressant effectiveness analysis was also conducted with these data. In this case, the unadjusted results are remarkably similar to the propensity-adjusted results. That is, higher doses of somatic antidepressant therapy offered significant protection from recurrence (OR: 0.63; 95 per cent CI: 0.44–0.89; $Z = -2.61$; $p = 0.009$), moderate doses were associated with marginal protection (OR: 0.70; 95 per cent CI: 0.49–1.00; $Z = -1.95$; $p = 0.051$), and lower doses did not offer significant protection (OR: 1.11; 95 per cent CI: 0.77–1.60; $Z = -0.56$; $p = 0.574$). This one example should not be seen as an indication that confounding variables can generally be ignored. The stimulation studies that follow highlight the benefit of the propensity adjustment.

## 4. SIMULATION STUDY

*Propensity score simulation specifications.* A Monte Carlo simulation study was conducted to evaluate the performance of the mixed-effects quintile-stratified propensity strategy that was just described. Simulated data were generated in the following way. First, propensity scores were calculated for each of 8 observations (i.e. repeated measures over time) per subject from a mixed-effects ordinal logistic regression model including four randomly generated predictor variables: (1) two time-invariant dichotomous variables ($x_1$, $x_2$), each with a 50:50 split of zeros and ones (2) two time-varying continuous variables ($x_3$, $x_4$) based on a uniform distribution. Odds ratios for each of the four predictors in the propensity model varied (1.25, 1.50, 1.75, 2.0) and the intercept was set to unity. The correlation among the propensity model predictor variables was set at 0.10. The intraclass correlation coefficient among the (time-varying) dependent variable within subjects varied (0.20, 0.40) in both the propensity and treatment effectiveness models. An ordinal dose (0, 1, 2, 3) was calculated for each observation as a function of the full propensity score (described below) and specified model threshold values that indicate the response probabilities of the ordinal outcome.

*Modelling a misspecified propensity score*. In order to evaluate the effect of misspecification of the propensity model, two different propensity scores were calculated for each observation. One, which we call the *full propensity score*, included all four of the propensity

 

model predictors. The other, called the *misspecified propensity score*, only included two predictors: one time-invariant dichotomous variable ($x_1$) and one time-varying continuous variable ($x_3$). The *full propensity score* was used to generate values for dose and for survival times for each observation. In contrast, the misspecified propensity score was used strictly for quintile stratification in the treatment effectiveness analyses. In this way, two confounds, $x_2$ and $x_4$, each components of the full propensity score, were ignored in the stratification process. Of course, when the propensity model was defined such that the odds ratios for $x_2$ and $x_4$ were both 1.0, the *full* and *misspecified propensity scores* were identical. This property was used to examine the effect of a misspecified propensity score. More specifically, the full propensity score was specified such that odds ratios corresponding to $x_1$ and $x_3$ varied, whereas those for $x_2$ and $x_4$ were constrained to unity. Hence, the terms to be omitted from the misspecified propensity score would not effect stratification. Conversely, in the misspecified models, the full propensity score was specified such that odds ratios corresponding to $x_1$ and $x_3$ were constrained to unity, whereas those for $x_2$ and $x_4$ varied. In this way, the effect of stratification that ignores confounds, whether purposefully or naïvely, was examined.

*Treatment effectiveness simulation specifications*. Latent survival times were then generated using a proportional hazards model as a function of both the full propensity score and treatment effectiveness. The effects of doses 1–3, relative to dose 0, were specified as odds ratios of 1.0, 1.25, and 1.75, respectively. *Time until recurrence*, which is the survival time outcome in the treatment effectiveness analyses, was modelled as a function of the dose for each subject at each time point and the confounds that were incorporated in the propensity score. In order to represent ten grouped-time survival times, the continuous latent survival times were transformed into deciles. The censoring rate was specified as 25 per cent in all effectiveness models. One thousand data sets were generated, each with 8 within-subject observations, for each of 32 combinations of the following data specifications: fully specified *versus* misspecified, $N$ (150, 300 subjects), odds ratios (1.25, 1.5, 1.75, 2.0) for the propensity model, and intraclass correlations (ICC: 0.20, 0.40), which represent the within-subject correlation of both the repeated propensity scores and survival times.

For each data set, several treatment effectiveness models were analysed. There were five quintile-specific models and one model in which the quintile-specific results were pooled using the Mantel–Haenszel procedure. In addition, 2 models were used to test the assumption of no propensity $\times$ treatment interaction: a main effects only model and a main effects plus interaction model. These last two models included observations from all quintiles. Thus, 8 models were analysed for each of 1000 data sets for each combination of specifications. An augmented version of MIXOR software [15] was used for the simulations.

*Evaluation of model performance*. Five criteria were used to evaluate the performance of the data analytic strategy: Type I error, statistical power, coverage, bias, and bias reduction. These were each derived from the Mantel–Haenszel pooled results. Bias was defined as the difference between each parameter estimate and the specified treatment effect. Bias reduction was defined as proportion decrease in bias of the model that used the full propensity score relative to the bias that failed to adjust for the propensity score (i.e. that only used the misspecified propensity score). Coverage represents the proportion of simulations in which the 95 per cent CI for the parameter estimate include the specified value. Feasibility was also examined and defined as the proportion of models in which convergence was achieved.

 

Table IV. Monte Carlo simulation results of the quintile-stratified, mixed-effects propensity adjustment for effectiveness analyses of ordinal treatment groups: coverage, bias, and bias reduction.

| Number of subjects | Propensity odds ratio | ICC | Coverage* Treatment odds ratio | | | Bias Treatment odds ratio | | | Bias reduction† Treatment odds ratio | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1.00 | 1.25 | 1.75 | 1.00 | 1.25 | 1.75 | 1.00 | 1.25 | 1.75 |
| 150 | 1.25 | 0.20 | 0.971 | 0.964 | 0.961 | 0.005 | 0.011 | 0.023 | 0.869 | 0.804 | 0.725 |
| | 1.25 | 0.40 | 0.976 | 0.975 | 0.981 | 0.006 | 0.012 | 0.012 | 0.848 | 0.776 | 0.847 |
| | 1.50 | 0.20 | 0.961 | 0.946 | 0.942 | 0.005 | 0.030 | 0.054 | 0.944 | 0.759 | 0.684 |
| | 1.50 | 0.40 | 0.978 | 0.970 | 0.969 | 0.013 | 0.016 | 0.004 | 0.833 | 0.859 | 0.975 |
| | 1.75 | 0.20 | 0.965 | 0.936 | 0.906 | 0.013 | 0.054 | 0.087 | 0.909 | 0.751 | 0.729 |
| | 1.75 | 0.40 | 0.961 | 0.967 | 0.964 | 0.015 | 0.008 | 0.013 | 0.899 | 0.962 | 0.959 |
| | 2.00 | 0.20 | 0.972 | 0.923 | **0.883** | 0.006 | 0.056 | 0.106 | 0.966 | 0.794 | 0.745 |
| | 2.00 | 0.40 | 0.967 | 0.963 | 0.968 | 0.018 | 0.007 | 0.021 | 0.909 | 0.975 | 0.951 |
| 300 | 1.25 | 0.20 | 0.963 | 0.921 | 0.924 | 0.007 | 0.014 | 0.024 | 0.799 | 0.745 | 0.678 |
| | 1.25 | 0.40 | 0.959 | 0.948 | 0.965 | 0.014 | 0.015 | 0.015 | 0.649 | 0.742 | 0.805 |
| | 1.50 | 0.20 | 0.980 | **0.899** | **0.882** | 0.011 | 0.032 | 0.058 | 0.857 | 0.725 | 0.648 |
| | 1.50 | 0.40 | 0.961 | 0.957 | 0.952 | 0.019 | 0.017 | 0.009 | 0.768 | 0.856 | 0.946 |
| | 1.75 | 0.20 | 0.957 | **0.867** | **0.809** | 0.008 | 0.050 | 0.090 | 0.940 | 0.764 | 0.713 |
| | 1.75 | 0.40 | 0.938 | 0.950 | 0.951 | 0.022 | 0.013 | 0.007 | 0.854 | 0.942 | 0.979 |
| | 2.00 | 0.20 | 0.954 | **0.867** | **0.755** | 0.006 | 0.050 | 0.104 | 0.969 | 0.815 | 0.745 |
| | 2.00 | 0.40 | 0.953 | 0.953 | 0.946 | 0.026 | 0.010 | 0.012 | 0.863 | 0.962 | 0.971 |

*Coverage represents the proportion of models in which the 95 per cent confidence interval includes the specified parameter.

†Bias reduction is the proportion reduced relative to bias in a null propensity model.

### 4.1. Simulation results

*Bias*. The simulations were designed such that each treatment effectiveness model estimated a treatment effect for each of three doses, relative to the lowest dose. The respective specified values for those treatment effects were 0, 0.223, and 0.560; or, expressed as odds ratios, 1.0, 1.25, and 1.75. Bias, the difference between the parameter estimate and the specified value, tended to be quite small (Table IV). The median bias was 0.009, 0.014, and 0.022 for each of the three doses, respectively, with $N = 150$ and 0.012, 0.016, and 0.020 for each of the three doses with $N = 300$.

*Bias reduction*. The magnitude of bias reduction was examined relative to a *misspecified* model (which was described in detail above). Reduction in bias is defined as the difference between the bias in the misspecified and fully specified quintile-stratified models as a proportion of bias in the misspecified model. The quintile-stratified propensity adjustment reduced substantial bias in the estimates of the treatment effect (Table IV), ranging from 68.4 per cent to 97.5 per cent (median = 85.4 per cent) for $N = 150$ and from 64.8 per cent to 97.9 per cent (median = 81 per cent) for $N = 300$.

*Coverage*. The coverage of the 95 per cent CI for each estimate was examined and is presented in Table IV as the percent of confidence intervals (among the 1000 simulated data sets) that included the specified value of the treatment effect. Following the format used by Collins *et al.* [16], coverage that is less than 90 per cent is highlighted in the tables. For

 

Table V. Monte Carlo simulation results of the quintile-stratified, mixed-effects propensity adjustment for effectiveness analyses of ordinal treatment groups: type I error, power and non-convergence.

| Number of subjects | Propensity odds ratio | ICC | Type I error / Power: Treatment odds ratio 1.00 | Power: Treatment odds ratio 1.25 | Power: Treatment odds ratio 1.75 | Non-convergence |
|---|---|---|---|---|---|---|
| 150 | 1.25 | 0.20 | 0.029 | 0.298 | 0.966 | 0.038 |
| | 1.25 | 0.40 | 0.024 | 0.257 | 0.945 | 0.066 |
| | 1.50 | 0.20 | 0.039 | 0.244 | 0.950 | 0.046 |
| | 1.50 | 0.40 | 0.022 | 0.281 | 0.937 | 0.073 |
| | 1.75 | 0.20 | 0.035 | 0.164 | 0.900 | 0.079 |
| | 1.75 | 0.40 | 0.039 | 0.256 | 0.902 | 0.117 |
| | 2.00 | 0.20 | 0.028 | 0.158 | 0.847 | 0.196 |
| | 2.00 | 0.40 | 0.033 | 0.242 | 0.892 | 0.197 |
| 300 | 1.25 | 0.20 | 0.037 | 0.606 | 1.000 | 0.003 |
| | 1.25 | 0.40 | 0.041 | 0.593 | 1.000 | 0.003 |
| | 1.50 | 0.20 | 0.020 | 0.505 | 1.000 | 0.000 |
| | 1.50 | 0.40 | 0.039 | 0.599 | 1.000 | 0.003 |
| | 1.75 | 0.20 | 0.043 | 0.401 | 0.999 | 0.000 |
| | 1.75 | 0.40 | 0.062 | 0.562 | 0.997 | 0.004 |
| | 2.00 | 0.20 | 0.046 | 0.393 | 0.997 | 0.018 |
| | 2.00 | 0.40 | 0.049 | 0.538 | 0.997 | 0.006 |

$N = 150$, the median coverage was 96.7 per cent (mean = 95.7 per cent, SD = 2.4 per cent), with just one value (of 24 parameter estimates) falling below 90 per cent. However, with $N = 300$, the median coverage was 95.1 per cent (mean = 92.5 per cent, SD = 5.4 per cent), with 6 values (of 24 parameter estimates) falling below 90 per cent, occurring only with an ICC of 0.20.

*Type I error*. Type I error rates (Table V) ranged from 0.022 to 0.039 for $N = 150$ (median = 0.031) and from 0.020 to 0.062 for $N = 300$ (median = 0.042). There was a tendency for somewhat higher type I error as the odds ratios in the propensity model increased.

*Statistical power*. Statistical power (Table V) typically exceeded 0.90 for treatment effectiveness odds ratios of 1.75 for models with sample size of 150 and 300 (median = 0.98). Nevertheless, there were two such models with lower power (0.85 and 0.89), each of which had $N = 150$ and odds ratios in the propensity model of 2.0. In contrast, statistical power for treatment effectiveness odds ratios of 1.25 ranged from 0.16 to 0.30 for $N = 150$ (median = 0.25) and 0.39 to 0.61 (median = 0.55) for $N = 300$.

*Feasibility*. There were some problems with convergence for the models with $N = 150$, in that, on average, about 10 per cent of the models did not converge (Table V). This problem typically stemmed from the quintile-specific analyses, particularly for the lowest or highest quintile and was more common as the odds ratio in the propensity model increased. In these cases, it sometimes happened that not all ordinal doses were present, or were present in very small numbers, for a given quintile, thus giving rise to computational difficulties. In contrast, less than 1 per cent of the models failed to converge for models with $N = 300$.

 

## 5. CONCLUSION

A two-staged mixed-effects propensity adjustment for longitudinal data was described and examined in a simulation study. A mixed-effects ordinal logistic regression model was used to estimate the propensity for treatment intensity and a mixed-effects grouped-time survival model was used to estimate the effect of ordinal doses. The primary objective of applying this method was to reduce bias in the estimate of the treatment effect. Bias tended to be quite small with the propensity adjustment. In comparison with models that incorporated a misspecified propensity score, the quintile-stratified propensity adjustment reduced substantial bias, typically in excess of 80 per cent.

Type I error rates were slightly conservative, particularly for $N = 150$ where it was less than 0.03 for half of the models. Statistical power readily exceeded 0.80 for treatment effectiveness odds ratios of 1.75, even with an $N$ of 150. However, for an effectiveness odds ratio of 1.25, even an $N$ of 300 provided inadequate power.

The simulations included 8 repeated observations per subject. Despite the resulting 1200 data points, there were problems with convergence about 10 per cent of the time for an $N$ of 150 when the propensity odds ratio was 1.75 or higher. In these situations, data sparseness contributed to collinearity or near-collinearity of model parameters in quintile-specific analyses. This situation could be avoided in actual data analysis, where an attentive analyst would realize that model simplification was necessary in these cases. In simulations, however, it is harder to program in this type of flexibility.

The extent of bias reduction was not dissimilar to that found by Cochran [9] for using quintile subclassification for non-normal continuous data. Drake [17] conducted a simulation study that, among other things, examined bias reduction with a logistic propensity score and a binary response variable. She found somewhat higher bias reduction than seen here; but unlike the mixed-effects approach examined in the current manuscript, Drake did not include repeated measurements per subject.

The findings of our simulation study are limited by the specifications used in that design. For instance, we did not examine the impact of hidden bias, which can result from a propensity model that includes some, but not all variables that are imbalanced across treatments. The effects of hidden bias can be quite insidious, in that it results from both observed and unobserved variables, and its impact can affect both the magnitude and the direction of bias. Instead, our simulations compared the performance of models that included all confounds with those that included none. Although our simulation study examined performance of the approach, we do not provide a proof that the balancing property extends to longitudinal study. Further work is needed to examine these issues.

Furthermore, all data generation in the simulations incorporated a 25 per cent censoring rate. We do not know how this strategy would perform with greater censoring or, for that matter, fewer observations per subject. Likewise, we did not evaluate the impact of ignoring a treatment $\times$ propensity interaction when pooling the parameter estimates nor did we examine the consequences of violations of the proportional odds assumptions. Furthermore, it is unclear how the approach would perform in a mixed-effects linear treatment effectiveness model. In the application, the propensity model showed that those with more severe symptoms tended to receive more intensive maintenance treatment for depression. The propensity adjustment removed or greatly reduced this imbalance. However, imbalance among unobserved measures could remain and contribute bias. Nonetheless, the effectiveness analyses showed that the

 

largest dose significantly reduced the risk of recurrence. Similar findings were seen in our earlier propensity-adjusted examination of acute treatment for major depression [3].

In conclusion, we have evaluated a mixed-effects quintile-stratified propensity adjustment for bias reduction in longitudinal, observational studies. Overall, the approach performed quite well, although some of the evaluation criteria were sensitive to smaller sample size.

ACKNOWLEDGEMENTS

Conducted with current participation of the following investigators: M.B. Keller, MD (Chairperson, Providence), W. Coryell (Co-Chair Person, Iowa City); D.A. Solomon, MD (Providence); W.A. Scheftner, MD (Chicago); W. Coryell, MD (Iowa City); J. Endicott, PhD, A.C. Leon, PhD, J. Loth, MSW (New York); J. Rice, PhD (St. Louis). Other current contributors include: H.S. Akiskal, MD, J. Fawcett, MD, L.L. Judd, MD, P.W. Lavori, PhD, J.D. Maser, PhD, T.I. Mueller, MD. This manuscript has been reviewed by the Publication Committee of the Collaborative Depression Study, and has its endorsement. The data for this manuscript came from the National Institute of Mental Health (NIMH) Collaborative Program on the Psychobiology of Depression-Clinical Studies [12]. The Collaborative Program was initiated in 1975 to investigate nostologic, genetic, family, prognostic and psychosocial issues of Mood Disorders, and is an ongoing, long-term multidisciplinary investigation of the course of Mood and related affective disorders. The original Principal and Co-principal investigators were from five academic centres and included Gerald Klerman, MD✠ (Co-Chairperson), Martin Keller, MD, Robert Shapiro, MD✠ (Massachusetts General Hospital, Harvard Medical School), Eli Robins, MD,✠ Paula Clayton, MD, Theodore Reich, MD,✠ Amos Wellner, MD✠ (Washington University Medical School), Jean Endicott, PhD, Robert Spitzer, MD (Columbia University), Nancy Andreasen, MD, PhD, William Coryell, MD, George Winokur, MD✠ (University of Iowa), Jan Fawcett, MD, William Scheftner, MD (Rush-Presbyterian-St. Luke's Medical Center). The NIMH Clinical Research Branch was an active collaborator in the origin and development of the Collaborative Program with Martin M. Katz, PhD, Branch Chief as the Co-Chairperson and Robert Hirschfeld, MD as the Program Coordinator. Other past contributors include: J. Croughan, MD, M.T. Shea, PhD, R. Gibbons, PhD, M.A. Young, PhD, D.C. Clark, PhD.

This research was supported, in part, by NIH grants MH60447 and MH49762.

REFERENCES

1. Rosenbaum P, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**:41–55.
2. Leon AC, Mueller TI, Solomon DA, Keller MB. A dynamic adaptation of the propensity score adjustment for effectiveness analyses of ordinal doses of treatment. *Statistics in Medicine* 2001; **20**:1487–1498.
3. Leon AC, Solomon DA, Mueller TI, Endicott J, Rice JP, Maser JD, Coryell W, Keller MB. A 20-year longitudinal, observational study of somatic antidepressant treatment effectiveness. *American Journal of Psychiatry* 2003; **160**:727–733.
4. Leon AC, Hedeker D. A mixed-effect propensity adjustment for effectiveness analyses of ordered categorical doses. *Statistics in Medicine* 2005; **24**:647–658.
5. Keller MB. Undertreatment of major depression. *Psychopharmacology Bulletin* 1988; **24**:75–80.
6. Sackeim HA. The definition and meaning of treatment-resistant depression. *Journal of Clinical Psychiatry* 2001; **62**(suppl):10–17.
7. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 1984; **79**:516–524.
8. Hedeker D, Gibbons RD. A random-effects ordinal regression model for multilevel analysis. *Biometrics* 1994; **50**:933–944.
9. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 1968; **24**:295–313.

✠Deceased.

 

10. Hedeker D, Siddiqui O, Hu FB. Random-effects regression analysis of correlated grouped-time survival data. *Statistical Methods in Medical Research* 2000; **9**:161–179.
11. Fleiss JL. *Statistical Methods for Rates and Proportions*. Wiley: New York, 1981.
12. Katz MM, Klerman GL. Introduction: overview of the clinical studies program of the NIMH clinical research branch collaborative study on psychobiology of depression. *American Journal of Psychiatry* 1979; **136**:49–51.
13. Keller MB, Lavori PW, Mueller TI, Endicott J, Coryell W, Hirschfeld RMA, Shea T. Time to recovery, chronicity and levels of psychopathology in major depression: A 5-year prospective follow-up of 431 subjects. *Archives General Psychiatry* 1992; **49**:809–816.
14. Spitzer R, Endicott J, Robins E. Research diagnostic criteria: rationale and reliability. *Archives of General Psychiatry* 1978; **35**:773–782.
15. Hedeker D, Gibbons RD. MIXOR: a computer program for mixed-effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine* 1996; **49**:157–176.
16. Collins LM, Schafer JL, Kam C-M. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods* 2001; **6**:330–351.
17. Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics* 1993; **49**:1231–1236.
18. Endicott J, Spitzer R. A diagnostic interview: schedule for affective disorders and schizophrenia. *Archives of General Psychiatry* 1978; **35**:837–844.
19. Keller MB, Lavori PW, Friedman B, Nielsen E, Endicott J, McDonald-Scott P, Andreason NC. The longitudinal interval follow-up evaluation: a comprehensive method for assessing outcome in prospective longitudinal studies. *Archives of General Psychiatry* 1987; **44**:540–548.

 