STATISTICS IN MEDICINE
*Statist. Med.* 2007; **26**:2650–2665
Published online 8 November 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2732

# A comparison of mixed-effects quantile stratification propensity adjustment strategies for longitudinal treatment effectiveness analyses of continuous outcomes

Andrew C. Leon[1,*,†] and Donald Hedeker[2]

[1]*Departments of Psychiatry and Public Health, Cornell University, New York, NY, U.S.A.*
[2]*Division of Epidemiology and Biostatistics, School of Public Health, University of Illinois at Chicago, Chicago, IL, U.S.A.*

SUMMARY

The propensity adjustment is used to reduce bias in treatment effectiveness estimates from observational data. We show here that a mixed-effects implementation of the propensity adjustment can reduce bias in longitudinal studies of non-equivalent comparison groups. The strategy examined here involves two stages. Initially, a mixed-effects ordinal logistic regression model of propensity for treatment intensity includes variables that differentiate subjects who receive various doses of time-varying treatments. Second, a mixed-effects linear regression model compares the effectiveness of those ordinal doses on a continuous outcome over time. Here, a simulation study compares bias reduction that is achieved by implementing this propensity adjustment through various forms of stratification. The simulations demonstrate that bias decreased monotonically as the number of quantiles used for stratification increased from two to five. This was particularly pronounced with stronger effects of the confounding variables. The quartile and quintile strategies typically removed in excess of 80–90 per cent of the bias detected in unadjusted models; whereas a median-split approach removed from 20 to 45 per cent of bias. The approach is illustrated in an evaluation of the effectiveness of somatic treatments for major depression in a longitudinal, observational study of affective disorders. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: antidepressant; effectiveness; longitudinal; mixed-effects model; propensity adjustment

## 1. INTRODUCTION

Well-conducted randomized controlled clinical trials (RCT) are the primary source of information regarding the efficacy of therapeutic interventions. Through randomized group assignment and

*Correspondence to: Andrew C. Leon, Department of Psychiatry, Weill Medical College of Cornell University, PO Box 140, 525 East 68th Street, New York, NY 10021, U.S.A.
†E-mail: acleon@med.cornell.edu

Contract/grant sponsor: National Institute Health; contract/grant numbers: MH060447, MH068638

*Received 13 April 2006*
*Accepted 19 September 2006*

double-blinding, the RCT design reduces the risk of bias. However, there are clinical populations, such as the acutely suicidal, that cannot readily participate in an RCT and an observational study can provide a useful format for examining treatment effectiveness. However, if treatment is not randomly assigned, but instead determined by clinical needs, it is unrealistic to expect balance across groups. For instance, the more severely ill will likely receive more intensive treatment than the moderately ill. In such a case, illness severity is a confounding variable and, as a result, the estimate of treatment effectiveness will likely be biased. For that reason, a statistical adjustment must be used to reduce bias when examining treatment effectiveness in an observational study.

Several adjustment strategies have been proposed including those that incorporate the propensity adjustment. Rosenbaum and Rubin [1] described three standard adjustment procedures in which the propensity score can be applied: matching, subclassification, and covariate adjustment. Cases where covariate adjustment performed poorly have been documented [1]. Matching procedures have been examined extensively (e.g. [2–6]). Below we evaluate various forms of subclassification or stratification. The objective of this manuscript is to compare the bias reduction from various approaches to quantile stratification when a propensity adjustment is used for longitudinal studies in which the treatment effectiveness outcome is a continuous variable. The data analytic strategy as applied here involves two stages. Initially, characteristics that differentiate those who receive various doses of time-varying treatments are examined in a model of propensity for treatment intensity using mixed-effects ordinal logistic regression. The effectiveness of those ordered categorical doses is then compared in a mixed-effects linear regression model of the longitudinal continuous outcome that incorporates the propensity adjustment. We apply these models to longitudinal data in order to illustrate the effectiveness evaluations involving the repeated assessments of continuous outcomes that are often seen in observational studies. Our previous applications and evaluations of the propensity adjustment in longitudinal studies have involved survival intervals as the treatment effectiveness outcome [7–9]. Because it is probably the case that repeated continuous effectiveness outcomes are more common than repeated survival outcomes, this current evaluation may be of more interest to readers involved in longitudinal data analysis.

This paper is organized in the following manner. To provide the background, we initially describe the propensity model (Section 2) and then present the effectiveness model (Section 3). A simulation study (Section 4) then compares the performance of four forms of stratification: median-split, terciles, quartiles, and quintiles. Finally, the approach is used to examine the antidepressant effectiveness in a longitudinal, observational study of affective disorders (Section 5).

## 2. PROPENSITY FOR TREATMENT MODEL

The propensity score is defined as the conditional probability of assignment to treatment $k$, given demographic and clinical variables that are hypothesized to be related to receiving treatment. The propensity score, $e(x)$, using the Rosenbaum and Rubin notation,

$$e(x) = P(T_i = 1|x)$$

can be estimated for each subject $i$ $(i = 1, \ldots, N)$ from the logistic model

$$\ln\left[\frac{P(T_i = 1)}{1 - P(T_i = 1)}\right] = x_i'\beta$$



DOI: 10.1002/sim

where the vector of coefficients, $\beta$, are parameters to be estimated, and $x_i$ is a vector of demographic and clinical covariates hypothesized to be related to receiving treatment ($T_i = 1$) *versus* not receiving treatment ($T_i = 0$). Additionally, $x_i$ includes a constant for the intercept. The propensity score, $e(x)$, is obtained as follows:

$$e(x_i) = \frac{\exp(x_i'\beta)}{1 + \exp(x_i'\beta)}$$

### 2.1. Propensity model for longitudinal study of ordinal doses

Elsewhere we described an extension of the propensity model to examine the longitudinal study of $k$ time-varying ordinal doses denoted by the variable $T_{ij}$ [8]. Adapting the Rosenbaum and Rubin [1] notation, the *propensity score* for the $k$th ordinal dose is as follows:

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

for subject $i$ $(i = 1, \ldots, N)$, at time $j$ $(j = 1, \ldots, J_i)$ and dose $k$ $(k = 0, \ldots, k-2)$. Here, to be consistent with the notation of the dichotomous logistic regression model, the index starts at 0 and goes to $k-2$. Also, the subject-specific random effect, $\upsilon_i$, is normally distributed in the population with mean 0 and variance $\sigma_\upsilon^2$. The propensity score is estimated from a mixed-effects ordinal logistic regression model [10]

$$\ln\left[\frac{P(T_{ij} > k)}{1 - P(T_{ij} > k)}\right] = \gamma_k + \beta_0 + x_{ij}'\beta + \upsilon_i$$

where $\gamma_k$ represents the threshold for dose $k$, $\beta_0$ is the intercept, $x_{ij}$ is the $p \times 1$ vector of covariates and $\beta$ represents the corresponding regression coefficients. Vector $x$ can include both time-invariant and time-varying covariates, which must be assessed prior to the start of a particular course of treatment. As specified, the model allows for changes in an individual's propensity score and dose over time. For identification, the first threshold $\gamma_1$ is typically set to zero (or the intercept is set to zero). In this parameterization, the threshold values $\gamma_k$ are decreasing and reflect the marginal cumulative logits. Specifically, there are $k-1$ cumulative logits (indexed as $k = 0, 1, \ldots, k-2$) and, based on the proportional odds assumption [11], the covariates are assumed to have the same effects across these logits.

The mixed-effects propensity score for subject $i$ at time $j$ can be expressed using the logistic response function as follows:

$$e(x_{ij}, \upsilon_i) = \frac{\exp(\beta_0 + x_{ij}'\beta + \upsilon_i)}{1 - \exp(\beta_0 + x_{ij}'\beta + \upsilon_i)}$$

Under the proportional odds assumption, it is not necessary to include the thresholds, $\gamma_k$, for dose $k$ $(k = 0, \ldots, k-2)$ in this expression because the thresholds do not vary by subject or time. Likewise, the intercept, which is a constant, is not needed for the propensity-based ranking that is used for stratification. The propensity score, which ranges from 0 to 1, represents the probability of receiving a higher dose $(T)$ of treatment based on the contribution of covariates, $x$, and subject-specific effects, $\upsilon_i$. The model could also include time effects if a temporal trend in repeated dosing is hypothesized. A low propensity score indicates that the observation has the characteristics of someone not likely to receive a higher dose at a particular point in time, whereas a high propensity score indicates that the observation has characteristics associated with more intensive doses.



DOI: 10.1002/sim

### 2.2. Propensity quantile classification

Each observation for subject $i$ at time $j$ is classified into a propensity quantile, based on the propensity score $e(x_{ij}, \upsilon_i)$. Treatment effectiveness analyses are then stratified by these quantiles. Cochran [12] showed that quintiles are sufficient for about a 90 per cent reduction in bias for normally distributed independent variables; and that further subclassification is not generally needed in linear regression analysis. Thus, quintile stratification, $q_{(1)} \ldots q_{(5)}$, is often the subclassification strategy that has been used to incorporate the time-invariant propensity adjustment. Below we describe a simulation study that compares bias reduction from various forms of quantile stratification for longitudinal data. The primary focus is to evaluate the applicability of a propensity adjustment that is implemented using a quintile stratification and based on time-varying propensity scores.

Quantile-specific effectiveness analyses cannot be conducted until one determines whether each treatment is well-represented in each quantile. This is because a dose that is not represented in a particular quantile, of course, cannot be evaluated in treatment effectiveness analyses of that quantile. Examination of a dose by propensity quantile contingency table determines the extent of representation in each quantile. Although there are no precise guidelines regarding minimum representation, our analyses proceeded if at least 5–10 observations were present in each cell of the contingency table.

## 3. LONGITUDINAL TREATMENT EFFECTIVENESS ANALYSES

### 3.1. Propensity quantile-stratified effectiveness analyses

A mixed-effects linear regression model compares $k$ treatment doses on a longitudinal continuous-dependent variable, $y_{ij}$, such as the Hamilton Rating Scale for Depression (HRSD), a measure of illness severity. The model is specified as follows:

$$y_{ij} = \alpha_0 + \alpha_1 T_{ij1} \cdots + \alpha_{k-1} T_{ij,k-1} + \theta_i + \varepsilon_{ij}$$

where $\alpha_0$ is the intercept term, $\alpha_1$ is the coefficient for dummy-coded treatment dose ($T_{ij1}$), $\alpha_{k-1}$ is the coefficient for dummy-coded treatment dose ($T_{ij,k-1}$), $\theta_i$ is a subject-specific random intercept, distributed as $N(0, \sigma_\theta^2)$, and $\varepsilon_{ij}$ is the error term for subject $i$ at time $j$, distributed independently as $N(0, \sigma_\varepsilon^2)$. If a temporal influence on treatment effectiveness, such as development of a medication resistance, is hypothesized, the model could also include a term representing the slope over time and possibly an interaction of treatment by time.

The propensity adjustment can be incorporated through quantile stratification; that is, separate quantile-specific treatment effectiveness analyses are conducted. Those quantile-specific results are then pooled provided there is not a significant propensity by treatment interaction. At this point, we do not specify the particular form of quantile to apply because it is the purpose of the simulation study reported below to compare four quantile strategies: median-split, terciles, quartiles, and quintiles.

### 3.2. Pooling of quantile-specific effectiveness results

The quantile-specific results are pooled using the Mantel–Haenszel procedure (as described by Fleiss [13]) in which each quantile-specific parameter estimate is weighted by the inverse of its



DOI: 10.1002/sim

squared standard error and the weighted mean of each of those estimates is the pooled estimate of the corresponding treatment effect. In pooling the results, it is assumed that there is not a treatment by propensity interaction. This assumption is evaluated with a likelihood ratio test that compares two effectiveness models, each of which includes observations from all quantiles: (1) the main effects only model (treatment and propensity quantile vectors) and (2) the main effects model plus propensity by treatment interaction. (The test of this assumption involves the only effectiveness analyses that examine observations from all quantiles in one model.) If the interaction is statistically significant, the assumption is violated. An interaction signifies that the treatment effect varies across quantiles, in which case pooling of quantile-specific results is inappropriate and, instead, quantile-specific treatment effectiveness results are reported.

## 4. SIMULATION STUDY

### *4.1. Propensity score simulation specifications*

The performance of four forms of quantile stratification that incorporate the longitudinal propensity adjustment in a mixed-effects linear regression model was examined in a Monte Carlo simulation study. The following approach was used to generate simulated data. The simulation was designed to generate six observations over time for each subject. Initially, six time-varying propensity scores were calculated for each subject from a mixed-effects ordinal logistic model using four randomly generated predictor variables:

$$e(x_{ij}, \upsilon_i) = P(T_{ij} > k | \upsilon_i, x_{1i}, x_{2i}, x_{3ij}, x_{4ij})$$

$x_1$ and $x_2$ are each dichotomous variables with a 50:50 split of zeros and ones and $x_3$ and $x_4$ are each time-varying continuous variables. All four $x$'s were generated based on an underlying standard normal distribution. The odds ratios (ORs) for the four predictors of dose in the propensity model varied (1.0, 1.2, 1.4, 1.6, 1.8, 2.0) and were constrained such that the ORs for $x_1$ and $x_3$ were equivalent and those for $x_2$ and $x_4$ were equivalent. The correlations among the propensity model predictor variables were set to 0.20. The correlations involving $x_1$ and $x_2$ were specified based on the relationship of the latent variables (that were dichotomized to form $x_1$ and $x_2$) with the respective predictors of dose, $x_1$–$x_4$. Two samples sizes were examined ($N = 100$ and 250) and each subject had six repeated observations over the course of time. The intraclass correlation coefficient (ICC) $\rho$ among those repeated observations was set at 0.40 and calculated based on the logistic link, where $\rho = \sigma_\upsilon^2/(\sigma_\upsilon^2 + \pi^2/3)$ [14].

### *4.2. Treatment effectiveness simulation*

Based on the propensity model, an ordinal dose ($k = 0, 1, 2, 3$) was obtained for each of the six observations per subject as a function of the propensity score and threshold values used to categorize the latent, or underlying, continuous propensity scores. The effects of the three doses (relative to a control, dose 0) on the continuous outcome were then specified in a linear mixed-effects regression model, with effect sizes of 0, 0.20 and 0.50. The effects of the covariates in the propensity model ($x_1$–$x_4$) on the continuous effectiveness outcome were incorporated and the covariate effects on dose and covariate effects on outcome were specified to be approximately equivalent. For each combination of simulation specifications, 1000 data sets were generated and analysed. Those results were evaluated, as described below.



DOI: 10.1002/sim

### 4.3. Misspecified propensity score

Bias reduction was evaluated relative to the bias in a misspecified propensity model and compared across four different propensity quantile stratification methods (median-split, terciles, quartiles, and quintiles). In order to do so, two different propensity scores were calculated for each observation. The *true propensity score* $e(x_{ij}, \upsilon_i)$ comprises all four of the propensity model predictors ($x_1$–$x_4$); whereas, the *misspecified propensity score*, $e^*(x_{ij}, \upsilon_i)$, included only two predictors:

$$e^*(x_{ij}, \upsilon_i) = P(T_{ij} > k | \upsilon_i, x_{1i}, x_{3ij})$$

Although the true propensity score was used to generate values for dose and treatment effectiveness outcome, the misspecified propensity score was used for quantile stratification in the treatment effectiveness analyses. Thus, the stratification process ignored two confounds ($x_2$ and $x_4$), even though they were components of the true propensity score. In this way the effect of a stratification process that ignores confounds was examined and used as a benchmark in evaluating bias reduction. (Note that when the omitted variables, $x_2$ and $x_4$, were specified to each have an OR of 1.0, the true and misspecified propensity scores were equivalent.)

### 4.4. Evaluation of model performance

Several criteria were used to evaluate the data analytic strategies. The primary criterion used was bias reduction. Bias was defined as the difference between the parameter estimate and the specified treatment effect. Bias reduction represents the proportion decrease in bias of the model that used the true propensity score relative to the bias in a model that failed to adjust for the true propensity score (i.e. that used only the misspecified propensity score). In addition, type I error, statistical power and coverage were also examined. Coverage is defined as the proportion of simulations in which the 95 per cent confidence interval for the parameter estimate included the specified value. Based on the work of Collins *et al*. [15], coverage that is less than 90 per cent is highlighted in the tables. All evaluation criteria were derived from the Mantel–Haenszel pooled results of quantile-specific estimates. The MIXOR program [16] was used for the propensity model simulations and the MIXREG program [17] was used for the effectiveness model simulations.

### 4.5. Simulation results

*4.5.1. Bias.* The magnitude of bias in the parameter estimates increases with the strength of the association of the propensity covariates (i.e. the confounds) with dose and outcome. This relation holds for each of the three doses and both $N = 100$ and 250 (Tables I and II, respectively). Bias decreases monotonically as the number of quantiles used for stratification on the propensity score increases from two to five.

*4.5.2. Bias reduction.* The bias reduction, expressed as a proportion relative to an unadjusted model, tends to increase with the number of quantiles, with quintiles typically having the greatest reduction (Figure 1). The exceptions to this appear with either small treatment effects or null effects of the propensity covariates (i.e. OR = 1.0). In cases where the bias in the unadjusted model approaches zero, for example with propensity OR of 1.0 or 1.2 (Tables I and II), there is minimal bias to remove. (Thus, in order to clearly convey the pattern of reduction across quantile strategies, in the presence of potential bias, propensity ORs of 1.0 are not presented in Figure 1.)



DOI: 10.1002/sim

Table I. Propensity adjustments applied with various quantile stratification strategies: comparison of type I error, statistical power, and bias for $N = 100$.

| | | Type I error | Statistical power | | Bias | | | Coverage* | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Treatment effect† | | | Treatment effect† | | | Treatment effect† | |
| Quantile strategy | Propensity covariate odds ratio | 0.0 | 0.2 | 0.5 | 0.0 | 0.2 | 0.5 | 0.0 | 0.2 | 0.5 |
| Unadjusted model | 1.0 | 0.05 | 0.44 | 0.98 | 0.009 | 0.008 | 0.004 | 0.949 | 0.951 | 0.951 |
| | 1.2 | 0.04 | 0.48 | 0.99 | 0.015 | 0.007 | 0.027 | 0.957 | 0.959 | 0.941 |
| | 1.4 | 0.06 | 0.58 | 0.99 | 0.038 | 0.048 | 0.081 | 0.940 | 0.932 | **0.880** |
| | 1.6 | 0.07 | 0.68 | 1.00 | 0.051 | 0.088 | 0.157 | 0.927 | 0.914 | **0.767** |
| | 1.8 | 0.11 | 0.78 | 1.00 | 0.094 | 0.143 | 0.236 | **0.886** | **0.845** | **0.561** |
| | 2.0 | 0.13 | 0.85 | 1.00 | 0.123 | 0.192 | 0.330 | **0.869** | **0.754** | **0.289** |
| Median split | 1.0 | 0.05 | 0.39 | 0.92 | 0.009 | 0.008 | 0.003 | 0.945 | 0.953 | 0.948 |
| | 1.2 | 0.05 | 0.40 | 0.93 | 0.014 | 0.000 | 0.014 | 0.946 | 0.953 | 0.948 |
| | 1.4 | 0.06 | 0.49 | 0.93 | 0.023 | 0.038 | 0.040 | 0.943 | 0.941 | 0.932 |
| | 1.6 | 0.06 | 0.49 | 0.96 | 0.039 | 0.049 | 0.072 | 0.937 | 0.937 | 0.938 |
| | 1.8 | 0.10 | 0.56 | 0.97 | 0.068 | 0.072 | 0.104 | 0.905 | 0.920 | 0.914 |
| | 2.0 | 0.09 | 0.59 | 0.97 | 0.081 | 0.106 | 0.144 | 0.914 | **0.874** | **0.870** |
| Terciles | 1.0 | 0.06 | 0.38 | 0.92 | 0.009 | 0.008 | 0.004 | 0.945 | 0.961 | 0.934 |
| | 1.2 | 0.04 | 0.37 | 0.92 | 0.010 | 0.008 | 0.012 | 0.957 | 0.953 | 0.944 |
| | 1.4 | 0.05 | 0.44 | 0.91 | 0.020 | 0.028 | 0.023 | 0.945 | 0.938 | 0.943 |
| | 1.6 | 0.05 | 0.40 | 0.93 | 0.023 | 0.025 | 0.035 | 0.951 | 0.946 | 0.955 |
| | 1.8 | 0.07 | 0.43 | 0.90 | 0.045 | 0.035 | 0.039 | 0.927 | 0.935 | 0.952 |
| | 2.0 | 0.06 | 0.41 | 0.93 | 0.044 | 0.049 | 0.060 | 0.942 | 0.923 | 0.940 |
| Quartiles | 1.0 | 0.05 | 0.39 | 0.91 | 0.010 | 0.010 | 0.005 | 0.947 | 0.946 | 0.943 |
| | 1.2 | 0.05 | 0.36 | 0.91 | 0.009 | 0.006 | 0.010 | 0.950 | 0.940 | 0.942 |
| | 1.4 | 0.05 | 0.42 | 0.88 | 0.012 | 0.018 | 0.007 | 0.952 | 0.945 | 0.939 |
| | 1.6 | 0.05 | 0.38 | 0.88 | 0.017 | 0.012 | 0.011 | 0.953 | 0.940 | 0.955 |
| | 1.8 | 0.06 | 0.38 | 0.86 | 0.023 | 0.009 | 0.003 | 0.936 | 0.939 | 0.944 |
| | 2.0 | 0.05 | 0.38 | 0.86 | 0.021 | 0.021 | 0.012 | 0.945 | 0.926 | 0.945 |
| Quintiles | 1.0 | 0.05 | 0.38 | 0.90 | 0.009 | 0.008 | 0.005 | 0.947 | 0.949 | 0.936 |
| | 1.2 | 0.06 | 0.36 | 0.90 | 0.009 | 0.006 | 0.008 | 0.942 | 0.936 | 0.947 |
| | 1.4 | 0.05 | 0.40 | 0.87 | 0.009 | 0.012 | 0.001 | 0.945 | 0.935 | 0.936 |
| | 1.6 | 0.06 | 0.33 | 0.86 | 0.014 | 0.001 | 0.005 | 0.942 | 0.931 | 0.941 |
| | 1.8 | 0.08 | 0.34 | 0.84 | 0.021 | 0.008 | 0.009 | 0.923 | 0.930 | 0.951 |
| | 2.0 | 0.06 | 0.33 | 0.83 | 0.008 | 0.002 | 0.004 | 0.942 | 0.920 | 0.946 |

*Coverage represents the proportion of models in which the 95 per cent confidence interval includes the specified parameter. Coverage values lower than 0.90 are presented in bold font.

†Treatment effect is expressed as an effect size.

It is noteworthy that with quartile and quintile stratification, bias reduction for a non-null treatment effect typically exceeds 80 per cent when the OR for the propensity covariate is 1.6 or higher.



DOI: 10.1002/sim

Table II. Propensity adjustments applied with various quantile stratification strategies: comparison of type I error, statistical power and bias for $N = 250$.

| | | Type I error | Statistical power | | Bias | | | Coverage* | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Treatment effect† | | | Treatment effect† | | | Treatment effect† | | |
| Quantile strategy | Propensity covariate odds ratio | 0.0 | 0.2 | 0.5 | 0.0 | 0.2 | 0.5 | 0.0 | 0.2 | 0.5 |
| Unadjusted model | 1.0 | 0.05 | 0.83 | 1.00 | 0.005 | 0.004 | 0.000 | 0.949 | 0.946 | 0.938 |
| | 1.2 | 0.06 | 0.86 | 1.00 | 0.010 | 0.010 | 0.023 | 0.945 | 0.950 | 0.939 |
| | 1.4 | 0.07 | 0.92 | 1.00 | 0.030 | 0.049 | 0.083 | 0.935 | 0.916 | **0.819** |
| | 1.6 | 0.10 | 0.97 | 1.00 | 0.055 | 0.090 | 0.156 | **0.898** | **0.827** | **0.524** |
| | 1.8 | 0.17 | 0.99 | 1.00 | 0.089 | 0.142 | 0.238 | **0.829** | **0.661** | **0.187** |
| | 2.0 | 0.24 | 1.00 | 1.00 | 0.121 | 0.188 | 0.326 | **0.760** | **0.482** | **0.023** |
| Median split | 1.0 | 0.04 | 0.77 | 1.00 | 0.005 | 0.004 | 0.001 | 0.958 | 0.952 | 0.951 |
| | 1.2 | 0.06 | 0.78 | 1.00 | 0.009 | 0.003 | 0.012 | 0.943 | 0.955 | 0.943 |
| | 1.4 | 0.04 | 0.84 | 1.00 | 0.019 | 0.033 | 0.041 | 0.957 | 0.935 | 0.938 |
| | 1.6 | 0.07 | 0.89 | 1.00 | 0.043 | 0.052 | 0.073 | 0.928 | 0.919 | 0.914 |
| | 1.8 | 0.12 | 0.90 | 1.00 | 0.066 | 0.070 | 0.104 | **0.885** | **0.895** | **0.834** |
| | 2.0 | 0.15 | 0.94 | 1.00 | 0.087 | 0.101 | 0.137 | **0.853** | **0.841** | **0.758** |
| Terciles | 1.0 | 0.05 | 0.76 | 1.00 | 0.005 | 0.004 | 0.003 | 0.955 | 0.949 | 0.939 |
| | 1.2 | 0.06 | 0.77 | 1.00 | 0.006 | 0.000 | 0.006 | 0.938 | 0.956 | 0.941 |
| | 1.4 | 0.06 | 0.78 | 1.00 | 0.015 | 0.021 | 0.024 | 0.945 | 0.938 | 0.945 |
| | 1.6 | 0.06 | 0.77 | 1.00 | 0.028 | 0.026 | 0.036 | 0.941 | 0.947 | 0.952 |
| | 1.8 | 0.07 | 0.79 | 1.00 | 0.043 | 0.032 | 0.044 | 0.929 | 0.928 | 0.932 |
| | 2.0 | 0.08 | 0.80 | 1.00 | 0.051 | 0.045 | 0.053 | 0.924 | 0.927 | 0.933 |
| Quartiles | 1.0 | 0.04 | 0.74 | 1.00 | 0.004 | 0.004 | 0.002 | 0.956 | 0.943 | 0.941 |
| | 1.2 | 0.06 | 0.74 | 1.00 | 0.004 | 0.002 | 0.004 | 0.939 | 0.946 | 0.934 |
| | 1.4 | 0.05 | 0.74 | 1.00 | 0.009 | 0.014 | 0.012 | 0.955 | 0.936 | 0.943 |
| | 1.6 | 0.06 | 0.72 | 1.00 | 0.016 | 0.010 | 0.013 | 0.943 | 0.959 | 0.960 |
| | 1.8 | 0.07 | 0.70 | 1.00 | 0.026 | 0.009 | 0.008 | 0.932 | 0.940 | 0.950 |
| | 2.0 | 0.06 | 0.70 | 1.00 | 0.027 | 0.017 | 0.009 | 0.944 | 0.951 | 0.956 |
| Quintiles | 1.0 | 0.05 | 0.74 | 1.00 | 0.006 | 0.005 | 0.003 | 0.954 | 0.941 | 0.943 |
| | 1.2 | 0.06 | 0.73 | 1.00 | 0.003 | 0.004 | 0.002 | 0.942 | 0.950 | 0.937 |
| | 1.4 | 0.05 | 0.73 | 1.00 | 0.008 | 0.010 | 0.006 | 0.948 | 0.945 | 0.953 |
| | 1.6 | 0.05 | 0.68 | 1.00 | 0.010 | 0.001 | 0.001 | 0.953 | 0.955 | 0.947 |
| | 1.8 | 0.06 | 0.64 | 1.00 | 0.014 | 0.006 | 0.011 | 0.937 | 0.934 | 0.945 |
| | 2.0 | 0.05 | 0.63 | 0.99 | 0.012 | 0.004 | 0.015 | 0.951 | 0.950 | 0.953 |

*Coverage represents the proportion of models in which the 95 per cent confidence interval includes the specified parameter. Coverage values lower than 0.90 are presented in bold font.

†Treatment effect is expressed as an effect size.

*4.5.3. Coverage.* The coverage of 95 per cent confidence intervals for parameter estimates was examined and found to be appropriate for analyses that used at least three strata (Tables I and II). Not surprisingly, the unadjusted models had inadequate coverage probability for higher propensity



DOI: 10.1002/sim

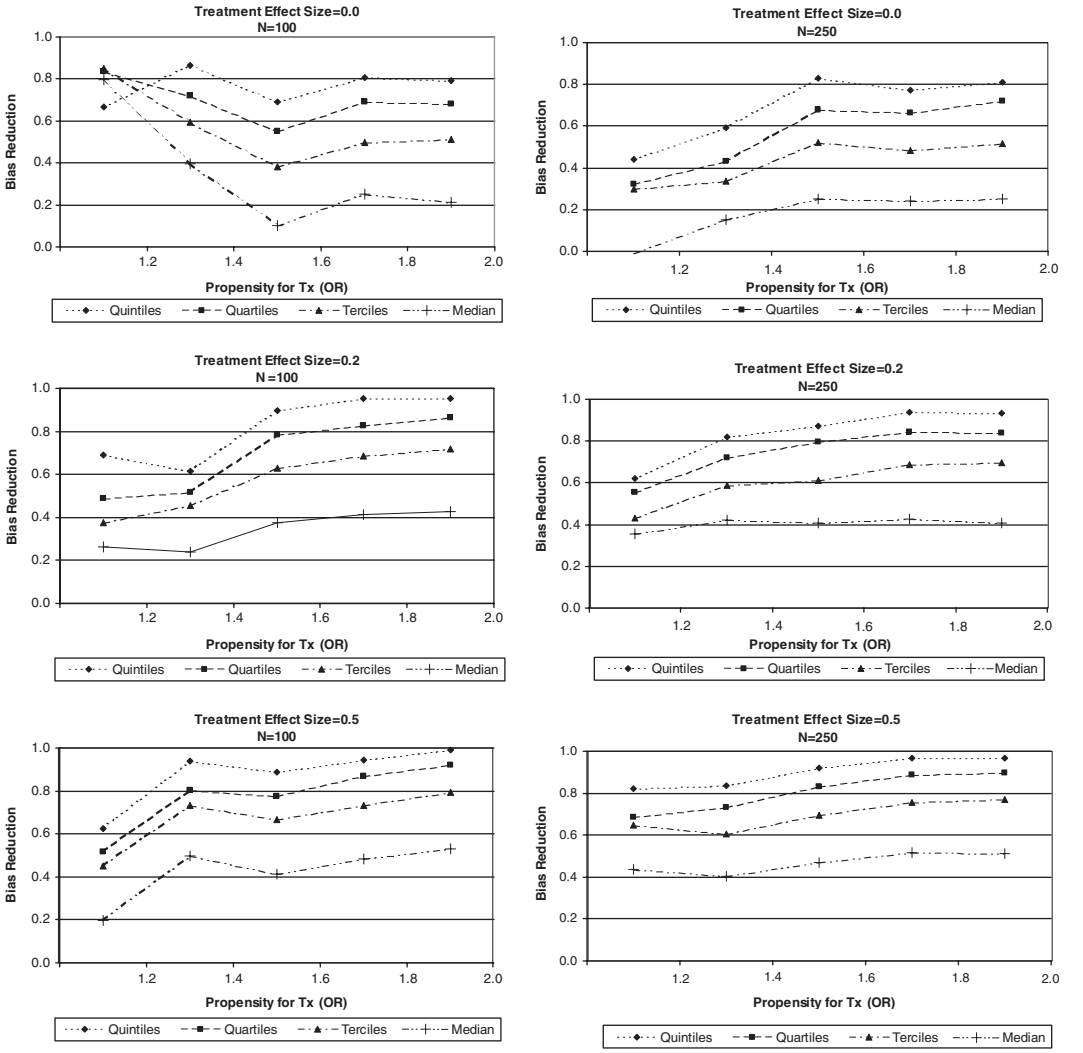


Figure 1. A comparison of mixed-effects quantile stratification propensity adjustment strategies: bias reduction.

covariate ORs ($\geqslant$1.4). In addition, the coverage was somewhat low for median-split stratification with covariate ORs of 1.8 and 2.0.

*4.5.4. Type I error.* The unadjusted models have excessive rates of type I error with propensity covariate ORs of 1.4 or higher and the problem is magnified as the strength of the association of the propensity covariates with dose and outcome increases. Type I error is reduced with the propensity adjustment, particularly as the number of quantiles increases. It is closest to the nominal level of 0.05 for quartile and quintile stratification. Even so, for propensity covariate ORs of 1.8, type I error is inflated for quartile (0.07 for $N = 250$) and quintile (0.08 for $N = 100$) stratification.



DOI: 10.1002/sim

*4.5.5. Statistical power.* Statistical power for a dose effect size of 0.20 tends to increase with the strength of the propensity confound for both $N = 100$ and 250. However, this is not the case for quartile and quintile stratification, where the propensity adjustment appears to reduce the variability in power across levels of the propensity covariate. Nonetheless, it appears that the additional strata involved in quartile and quintile stratification came at a cost of slightly reduced power. For $N = 100$, there is a trend of more power in results using fewer strata. The statistical power of the models is adequate for a dose effect of 0.50, for both $N = 100$ (typically $\geqslant 0.85$) and $N = 250$ ($\geqslant 0.99$), regardless of stratification strategy. In fact, the ceiling effect for power with $N = 250$ renders the stratification comparison uninformative for the larger dose effect.

## 5. APPLICATION

### *5.1. Effectiveness of somatic antidepressant treatment*

The mixed-effects propensity adjustment was applied to the study of effectiveness of somatic antidepressant treatment for reducing the severity of depressive symptoms. Data from the National Institute of Mental Health Collaborative Depression Study (CDS), a longitudinal, observational study with up to 20 years of follow-up, were used for these analyses. From 1978 through 1981, inpatients and outpatients with mood disorders were enrolled into the CDS at the academic medical centres in Boston, MA; Chicago, IL; Iowa City, IA; New York, NY; and St Louis, MO. Each subject provided written informed consent after receiving a complete description of the study [18]. The analyses presented here include 2185 observations of 182 subjects who met criteria for major depressive disorder at intake into the CDS, did not develop bipolar disorder during follow-up, recovered from the intake episode of major depression, had at least one subsequent depressive episode and had at least one change in treatment while in an episode subsequent to the intake episode. These data are used to illustrate the application of the mixed-effects propensity adjustment in which the treatment effectiveness outcome is a continuous measure. We hypothesized that more intensive somatic antidepressant treatment would correspond to a greater reduction in symptom severity.

Throughout follow-up, somatic antidepressant treatment intensity was coded in five ordinal categories ranging from 0 (no treatment) to 4 (intensive dose) as described in detail elsewhere [19, 20]. For example, the ordinal doses for sertraline are as follows: dose 1 (1–49 mg), dose 2 (50–100 mg), dose 3 (101–199 mg), and dose 4 ($\geqslant$200 mg). Somatic treatments include psychopharmacological and electroconvulsive therapies. The unit of analysis, which we refer to as a *treatment interval*, was defined in the following manner. A treatment interval commenced in the first week of a new dose of antidepressant treatment for a subject who was in a major depressive episode. The interval ended with a subsequent change in dose, or, if no dose change occurred, at the end of follow-up. Treatment intervals at intake into the CDS were excluded from these analyses because, by design, not all clinical characteristics at the start of the interval that were required to estimate the propensity for treatment intensity were assessed prior to intake. Given the longitudinal nature of this study, subjects tended to contribute multiple treatment intervals (mean = 12.0; median = 8.0; sd = 13.5).

The dependent variable for the evaluation of effectiveness was change from the beginning to the end of the treatment interval on the Psychiatric Status Rating (PSR), a component of the Longitudinal Interval Follow-up Evaluation (LIFE) [21], which quantifies level of psychopathology



DOI: 10.1002/sim

and ranges from 1 (no symptoms) to 6 (full criteria for major depression along with psychosis or extreme impairment in functioning).

### 5.2. Results

*5.2.1. Propensity for treatment intensity.* The propensity analyses involved mixed-effects ordinal logistic regression analyses, in which the dependent variable was the ordinal dose, ranging from dose 0 to 4 (as described above). The propensity model included several demographic and clinical variables (Table III). This model shows that those with more severe symptoms, worsening symptoms, and episodes of shorter duration tended to receive higher doses of somatic antidepressant treatment. Likewise, there were significant age and site differences in treatment intensity. Specifically, subjects from the New York and Iowa sites received more intensive treatments than those from St Louis. Those aged 30–49 years old tended to receive more intensive treatment than those younger than age 30. The ICC among these observations was estimated as 0.101, indicating that, within-subjects, there are highly variable levels of treatment while in episode over the longitudinal course of follow-up. (Recall that the propensity score is time-varying, resulting from the clinical characteristics that can change over the course of follow-up.)

A propensity score was calculated for each observation, based on the model just described, and the observations were then classified into propensity quintiles. The cross-classification of dose by propensity score quintile was then examined to evaluate the assumption that all treatments

Table III. A model of propensity for treatment intensity during a major depressive episode*.

| Variable | Adjusted odds ratio | CI low | CI high | $Z$ | $p$-value |
|---|---|---|---|---|---|
| Symptom severity† (range: 1–6) | 1.15 | 1.04 | 1.27 | 2.74 | 0.006 |
| Symptom trajectory† | | | | | |
| Worsening | 1.49 | 1.20 | 1.84 | 3.63 | <0.001 |
| Improving | 1.03 | 0.81 | 1.33 | 0.26 | 0.794 |
| Duration of episode (number of months prior to treatment interval) | 0.996 | 0.992 | 1.000 | −1.74 | 0.082 |
| Age (years) | | | | | |
| <30 | 1.00 | | | | |
| 30–39 | 1.84 | 1.35 | 2.50 | 3.87 | <0.001 |
| 40–49 | 2.07 | 1.42 | 3.03 | 3.78 | <0.001 |
| 50–59 | 1.56 | 0.99 | 2.46 | 1.91 | 0.056 |
| 60+ | 1.24 | 0.82 | 1.90 | 1.02 | 0.309 |
| Study site | | | | | |
| St Louis | 1.00 | | | | |
| New York | 2.71 | 1.22 | 6.02 | 2.44 | 0.015 |
| Boston | 1.18 | 0.71 | 1.96 | 0.62 | 0.533 |
| Iowa | 1.91 | 1.30 | 2.83 | 3.26 | 0.001 |
| Chicago | 1.49 | 0.96 | 2.31 | 1.76 | 0.078 |

*2185 observations from 182 subjects.

†Prior 8 weeks.



DOI: 10.1002/sim

Table IV. Cross-classification of treatment intensity (dose) by propensity quintile.

| Treatment intensity (doses) | Propensity quintile | | | | | |
|---|---|---|---|---|---|---|
| | Q1 | Q2 | Q3 | Q4 | Q5 | Total |
| 0 | 191 | 104 | 83 | 58 | 17 | 453 |
| 1 | 138 | 124 | 99 | 77 | 45 | 483 |
| 2 | 70 | 125 | 133 | 133 | 114 | 575 |
| 3 | 24 | 54 | 73 | 95 | 141 | 387 |
| 4 | 13 | 31 | 49 | 74 | 120 | 287 |
| Total | 436 | 438 | 437 | 437 | 437 | 2185 |

are represented in each quintile (Table IV). Although, by nature of the propensity for treatment intensity model, observations in the lower quintiles tended to get less intensive treatment and conversely, observations in the higher quintiles tended to get more intensive treatment, each level of treatment was represented in each quintile.

*5.2.2. Examination of propensity-adjusted balance.* The primary goal of the propensity adjustment is to account for imbalance on hypothesized confounding variables across treatments. In an effort to examine the degree to which this was achieved, separate mixed-effects ordinal logistic regression analyses were conducted, in which the dependent variable was the ordinal dose, and the independent variables included the propensity score and the respective variables in the propensity model. Consistent with the belief that the propensity adjustment will remove or reduce the influence of the confounding variables, we hypothesized that the propensity-adjusted ORs for the respective variables will be of smaller magnitude than the unadjusted ORs and will not be statistically significant. In fact, this hypothesis was supported in the analyses of balance (Table V). Each variable was significantly associated with dose in unadjusted models, but none were in the adjusted models. (Note that the results in Table V examined balance in separate models for each variable in the propensity score. In contrast, Table III presents results from one model that includes all propensity variables.)

*5.2.3. Treatment effectiveness.* Treatment effectiveness analyses were then stratified by quintile of propensity for treatment intensity. The mean ICC for the effectiveness models, across the quintiles was 0.15, indicating substantial within-subject variability across treatment intervals in change in symptom severity. The treatment by propensity interaction was not statistically significant ($-2\text{LL} = 18.032$, $\text{df} = 16$, $p = 0.322$); therefore, pooled estimates of the treatment effects were calculated using the procedure described above. The three highest levels of somatic antidepressant treatment were associated with significantly greater reduction in symptom severity than no somatic treatment, despite more severe presentation of depressive symptoms. The effects for these three levels were quite similar: (dose 2: $b = 0.33$; 95 per cent CI: 0.17–0.49; $p < 0.001$; dose 3: $b = 0.33$; 95 per cent CI: 0.14–0.52; $p < 0.001$; dose 4: $b = 0.37$; 95 per cent CI: 0.15–0.59; $p < 0.001$). In contrast, those treated with the lower level did not have significantly greater reduction in severity than those who did not receive somatic treatment (dose 1: $b = 0.15$; 95 per cent CI: $-0.01$–0.30; $p = 0.068$). These results are reflected in the dose-specific change in symptom severity from the



DOI: 10.1002/sim

Table V. 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) | 1.23 | 1.13 | 1.34 | <0.0001 | 0.961 | 0.878 | 1.052 | 0.389 |
| Symptom trajectory | | | | | | | | |
| Worsening | 1.67 | 1.38 | 2.03 | <0.001 | 0.96 | 0.78 | 1.18 | 0.700 |
| Improving | 1.14 | 0.91 | 1.43 | 0.257 | 0.98 | 0.78 | 1.22 | 0.829 |
| Duration of episode (number of months prior to treatment interval) | 0.995 | 0.9918 | 0.9987 | 0.007 | 1.001 | 0.997 | 1.006 | 0.560 |
| Age (years) | | | | <0.0001 | | | | 0.859 |
| <30 | 1.00 | | | | 1.00 | | | |
| 30–39 | 1.67 | 1.24 | 2.26 | 0.001 | 0.98 | 0.77 | 1.25 | 0.884 |
| 40–49 | 1.72 | 1.20 | 2.46 | 0.003 | 0.92 | 0.68 | 1.25 | 0.604 |
| 50–59 | 1.37 | 0.89 | 2.09 | 0.151 | 0.93 | 0.64 | 1.35 | 0.692 |
| 60+ | 1.34 | 0.87 | 2.07 | 0.189 | 1.06 | 0.76 | 1.47 | 0.725 |
| Study site | | | | <0.0001 | | | | 0.443 |
| Boston | 1.00 | | | | | | | |
| New York | 2.52 | 1.10 | 5.79 | 0.030 | 0.72 | 0.31 | 1.70 | 0.457 |
| St Louis | 1.25 | 0.75 | 2.07 | 0.390 | 0.79 | 0.43 | 1.44 | 0.438 |
| Iowa | 1.88 | 1.34 | 2.64 | <0.001 | 0.87 | 0.60 | 1.27 | 0.463 |
| Chicago | 1.50 | 1.02 | 2.21 | 0.041 | 0.74 | 0.46 | 1.18 | 0.202 |

*Separate models evaluated each of the five variables.

beginning to the end of a treatment interval where the mean (sd) change increased across the treatment levels: dose 0: 0.06 (1.36); dose 1: 0.23 (sd = 1.08); dose 2: 0.46(1.26); dose 3: 0.48 (1.24); dose 4: 0.59 (1.29).

For comparison, the unadjusted parameter estimates of the treatment effects were derived from one mixed-effects linear regression model of the pooled data. Each was somewhat higher than the propensity–adjusted estimates reported above: (dose 1: $b = 0.17$; 95 per cent CI: 0.01–0.33; $p = 0.032$; dose 2: $b = 0.38$; 95 per cent CI: 0.23–0.53; $p < 0.001$; dose 3: $b = 0.44$; 95 per cent CI: 0.27–0.61; $p < 0.001$; dose 4: $b = 0.56$; 95 per cent CI: 0.38–0.75; $p < 0.001$). We caution that, based on the results from our simulation study, this one example is not meant to encourage the use of unadjusted analyses. Moreover, the larger parameter estimates from the unadjusted analyses are not necessarily the correct results.

## 6. CONCLUSIONS

A mixed-effects propensity adjustment for bias reduction in longitudinal, observational effectiveness analyses of continuous outcomes has been described and evaluated. The strategy examined here involves two stages: (1) a mixed-effects ordinal logistic regression model of propensity for treatment intensity and (2) a mixed-effects linear regression model comparing the effectiveness of



DOI: 10.1002/sim

the ordinal doses. The approach can accommodate both treatments and propensity scores that vary within-subjects over time, a scenario that is commonly seen over the course of a chronic illness. An example of SAS code that implements this procedure is available from the first author.

The simulation study demonstrated that greater bias reduction is achieved with a larger number of strata and this was particularly pronounced with stronger effects of the confounding variables (i.e. the propensity covariates). The quartile and quintile strategies typically removed in excess of 80–90 per cent of the bias detected in unadjusted models. The magnitude of bias reduction is similar to that reported by Cochran [12] for fixed-effects linear regression, likewise using quartiles or quintiles. The median-split approach did remove bias in our simulations, but the extent of reduction was substantially lower, by and large ranging from 20 to 45 per cent.

Nevertheless, in a study with a smaller sample size, median-split or tercile stratification might be necessary to assure that each treatment is represented in each stratum. Although this will yield less biased estimates than unadjusted analyses, considerable bias could remain. In such a case, further bias reduction might be achieved by including the propensity score as a covariate in the quantile-specific analyses, particularly if there is substantial within-quantile propensity score heterogeneity. However, this is a somewhat unusual application of the propensity score approach and perhaps needs further investigation.

In addition, the stratification procedure that was used affected type I error rate, statistical power and coverage. As the number of quantiles increased, the rates of type I error are closest to the nominal level of 0.05. However, there appears to be a small cost in statistical power associated with the additional strata involved in quartile and quintile stratification. The coverage was appropriate for analyses that used terciles, quartiles, and quintiles.

The approach was illustrated with an evaluation of the effectiveness of ordinal doses of somatic treatments for major depression in a longitudinal, observational study of affective disorders. The propensity adjustment removed the imbalance initially seen across doses. Despite the tendency that greater psychopathology was associated with receiving higher doses, the quintile-adjusted effectiveness analyses showed that the three higher doses were significantly more effective on outcome than no treatment.

Although our simulation study exclusively focused on methods of stratification, direct covariate adjustment and matching are two alternative strategies for propensity adjustment implementation. The former has been shown to perform poorly (with non-longitudinal data) if the relationship between the linear discriminant and the propensity score is not monotonic [1]. This was not examined here because the simulated data assumed a monotonic relationship between the propensity score and dose. Furthermore, a thorough examination of the representativeness of each treatment across the propensity score spectrum is needed with the covariate adjustment. Matching procedures have been examined in a comprehensive series of studies (e.g. [2–6]). Matching is likely to be more easily understood by the clinician and it is particularly useful if the sample size is insufficient for stratification. Yet, it is not readily apparent how analyses would be conducted with longitudinal data that are structured such that two level-two factors are crossed. That is, there would be correlated repeated measures within-subject and correlated observations within propensity score-matched dyads, presumably from distinct subjects, who each receive different treatments. Although the use of an average dose for a subject over time might simplify the structure, it would ignore the very dose-specific effects that we seek to evaluate.

The results presented here apply to settings where the design specifications are similar to those examined in the simulation study. Wider generalization is unwarranted. For instance, the simulated data sets included six observations per subject with an ICC of 0.40, sample sizes of 100 and



DOI: 10.1002/sim

250 subjects, and a correlation among the confounds of 0.20. It is unclear how the model would perform with smaller samples, less stable treatment intensity or outcome over time, or with much more highly correlated confounding variables.

This extends our earlier evaluation of a mixed-effects propensity adjustment with repeated survival intervals as outcomes [9], where we demonstrated that, in analyses of survival intervals, quintile stratification removed substantially more bias than when a smaller number of strata were used. In summary, the strategy described here provides a viable option for evaluating treatments in situations where RCT data, though preferred, are unavailable.

ACKNOWLEDGEMENTS

This research was supported, in part, by grants from the National Institute Health (MH060447 and MH068638). The authors thank Jed Teres for conducting data analyses. The Collaborative Depression Study was 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, M. S. W. (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 [18]. The Collaborative Program was initiated in 1975 to investigate nosologic, 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 (see footnote ‡) (Massachusetts General Hospital, Harvard Medical School), Eli Robins, MD (see footnote ‡) Paula Clayton, MD, Theodore Reich, MD (see footnote ‡) Amos Wellner, MD (see footnote ‡) (Washington University Medical School), Jean Endicott, PhD, Robert Spitzer, MD (Columbia University), Nancy Andreasen, MD, PhD, William Coryell, MD, George Winokur, MD (see footnote ‡) (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 Co-ordinator. Other past contributors include: J. Croughan, MD, M. T. Shea, PhD, R. Gibbons, PhD, M. A. Young, PhD, D. C. Clark, PhD.

REFERENCES

1. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**:41–55.
2. Rubin DB, Rosenbaum PR. The bias due to incomplete matching. *Biometrics* 1985; **41**:103–116.
3. Gastwirth JL, Krieger AM, Rosenbaum PR. Dual and simultaneous sensitivity analysis for matched pairs. *Biometrika* 1998; **85**:907–920.
4. Ming K, Rosenbaum PR. Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics* 2000; **56**:118–124.
5. Lu B, Zanutto E, Hornik R, Rosenbaum PR. Matching with doses in an observational study of a media campaign against drug abuse. *Journal of American Statistical Association* 2001; **96**:1245–1253.
6. Ming K, Rosenbaum PR. A note on matching with variable controls using the assignment algorithm. *Journal of Computational and Graphical Statistics* 2001; **10**:455–463.

‡Deceased.



DOI: 10.1002/sim

7. 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.
8. Leon AC, Hedeker D. A mixed-effects quintile-stratified propensity adjustment for effectiveness analyses of ordered categorical doses. *Statistics in Medicine* 2005; **24**:647–658.
9. Leon AC, Hedeker D, Teres JJ. Bias reduction in effectiveness analyses of longitudinal ordinal doses with a mixed-effects propensity adjustment. *Statistics in Medicine* 2005 [Epub ahead of print].
10. Hedeker D, Gibbons RD. A random-effects ordinal regression model for multilevel analysis. *Biometrics* 1994; **50**:933–944.
11. McCullagh P. Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society*, *Series* 1980; **42**:109–142.
12. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 1968; **24**:295–313.
13. Fleiss JL. *Statistical Methods for Rates and Proportions*. Wiley: New York, 1981.
14. Agresti A. *Categorical Data Analysis* (2nd edn). Wiley: New York, 2002.
15. 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.
16. Hedeker D, Gibbons RD. MIXOR: a computer program for mixed-effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine* 1996; **49**:157–176.
17. Hedeker D, Gibbons RD. MIXREG: a computer program for mixed-effects regression analysis with autocorrelated errors. *Computer Methods and Programs in Biomedicine* 1996; **49**:229–252.
18. 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.
19. Keller MB. Undertreatment of major depression. *Psychopharmacology Bulletin* 1988; **24**:75–80.
20. 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.
21. 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.



DOI: 10.1002/sim