Statistics and Its Interface Volume 4 (2011) 489–498

Propensity score stratification for observational comparison of repeated binary outcomes

Andrew C. Leon* and Donald Hedeker

A two-stage longitudinal propensity adjustment is described for bias reduction in treatment effectiveness estimates in observational studies. The initial stage characterizes those who receive various ordinal doses of treatment in a model of time-varying propensity for treatment intensity. The second stage incorporates the propensity adjustment in longitudinal treatment effectiveness analyses of a binary outcome that are stratified by propensity quantile. Mantel-Haenszel pooled parameter estimates are then calculated as weighted means of quantile-specific estimates. A simulation study compares four approaches to quantile stratification and shows that the quintile stratification reduces more bias than when fewer strata are used with longitudinal data. Statistical power, type I error and coverage are also evaluated. A longitudinal, observational study of antidepressant effectiveness illustrates the approach.

Keywords and phrases: Bias reduction, Propensity adjustment, Stratification, Treatment effectiveness.

1. INTRODUCTION

The randomized controlled clinical trial is the gold standard for evaluation of biomedical interventions because groups tend to be balanced at baseline in well-conducted trials. Nonetheless, there are clinical populations that do not lend themselves to a standard randomized experimental design. For instance, the randomization of acutely suicidal patients to a placebo condition would raise serious ethical concerns. It is in such a setting that observational designs can prove to be quite informative. The intervention evaluation, however, must accommodate the non-randomized treatment assignment. This is because selection bias, be it driven by the clinician or patient, will render non-equivalent comparison groups. That is, when treatment choice is determined by clinical need, and not by a randomization strategy, group comparisons can be influenced by confounding variables. For instance in psychiatry, those who are more severely ill tend to be treated with more aggressive interventions than the mildly ill. As a consequence, unadjusted comparisons could demonstrate that the less aggressive intervention is more effective. In contrast, surgical interventions might be offered to the less severely ill or the less fragile patients. These examples illustrate the difficulty in interpreting treatment effects in the presence of a confounding relationship between baseline illness severity with treatment assignment.

A simple, yet effective strategy for removing a confounding variable is stratification. This is because if a confounding variable is transformed into a constant through stratification, it can have no influence on the stratum-specific outcome. However, stratification is unwieldy with multiple confounding variables and, as the number of strata increases, the stratum-specific sample sizes become sparse. The propensity adjustment is a strategy that can be used for multivariable stratification. Rosenbaum and Rubin (1983) defined the propensity score as the conditional probability of group assignment, given a set of covariates. They describe how the confounding influence of those covariates can be ameliorated with a propensity score strategy implemented through stratification, matching, or covariate adjustment. Here we focus on the former in which observations are stratified into propensity score quantiles and separate analyses are conducted on each stratum.

Two other features of longitudinal observational studies are incorporated in the data analytic strategy described here. One is the repeated assessments of outcome over the course of the study. In such a case, an adjustment for confounding variables, such as the propensity stratification, must accommodate the longitudinal design. Second, not only is treatment assignment beyond the control of the investigator, but an observational protocol typically will not proscribe any interventions that are used in clinical practice. Thus, the statistical adjustment must also accommodate various doses, or treatment intensities. The theoretical properties of a propensity adjustment for dose-response functions have been described elsewhere (Joffe, 1999; Imbens, 2000). Furthermore, longitudinal propensity adjustments for ordinal doses have been evaluated for repeated measures of survival outcomes (Leon and Hedeker, 2005) and continuous variables (Leon and Hedeker, 2007). The implementation of the longitudinal propensity adjustment is extended here for repeated binary outcomes and the impact of various forms of quantile stratification is examined with regard to bias reduction in conditional odds ratios estimated in random intercept models.

Initially, a longitudinal model of propensity for treatment intensity is described that accommodates time-varying

*Corresponding author.

treatments, which are characteristic of the longitudinal course of a chronic illness, and incorporates time-varying predictors of treatment intensity (Section 2). As the predictors change over the course of an illness, with symptom exacerbation for example, the strategy that we describe allows for corresponding changes in the propensity score. The implementation of the longitudinal propensity adjustment in treatment effectiveness evaluations is then considered (Section 3). An observational examination of antidepressant effectiveness is used for illustration (Section 4). Finally, a simulation study compares the performance of four approaches to stratification: median-split, terciles, quartiles, and quintiles (Section 5).

2. A LONGITUDINAL MODEL OF PROPENSITY FOR TREATMENT INTENSITY

2.1

Stage one of the analyses involves the propensity model. The notation of Rosenbaum and Rubin (1983) is adapted for a longitudinal *propensity for treatment intensity score*. The propensity for the *kth* ordinal dose, denoted by the variable T, is:

$$e_k(x_{ij}, \upsilon_i) = P(T_{ij} > k \mid \upsilon_i, x_{ij}) \tag{2.1}$$

for subject i ($i = 1, \ldots, N$), at time j ($j = 1, \ldots, J_i$), for dose k ($k = 1, \ldots, K-1$). In the ideal setting, treatment assignment is ignorable given x_{ij} and υ_i, which is the subject-specific random effect that is normally distributed in the population with mean 0 and variance σ^2_υ. This can be specified using a mixed-effects ordinal logistic regression model (Hedeker and Gibbons, 1994) that includes covariates and the subject-specific random effect:

$$\ln\left[\frac{P(T_{ij} > k)}{1 - P(T_{ij} > k)}\right] = \gamma_k + x'_{ij}\beta + \upsilon_i \tag{2.2}$$

where γ_k represents the threshold for dose k, x_{ij} represents a $p \times 1$ vector with an intercept and the covariates, and β contains the regression coefficients. Both time-varying (e.g., illness severity) and time-invariant (e.g., demographic characteristics such as gender or ethnicity) covariates can be included in the vector x. The subject-specific random effect is included in the model to account for the within-subject clustering that is expected in a longitudinal design with repeated within-subject observations. In the current parameterization, the $k-1$ thresholds are strictly decreasing parameters, which reflect the marginal cumulative logits. For identification purposes, either the first threshold or the intercept is set equal to zero. (Here we do the former.) Model parameters can be estimated using a maximum (marginal) likelihood solution that uses Gauss-Hermite quadrature to numerically integrate over the random effect distribution (Hedeker & Gibbons, 1994, 1996).

This time-varying propensity score, which can range from 0 to 1, can be expressed using the logistic response function for subject i at time j:

$$e(x_{ij}, \upsilon_i) = \frac{\exp(x'_{ij}\beta + \upsilon_i)}{1 + \exp(x'_{ij}\beta + \upsilon_i)}. \tag{2.3}$$

As expressed above, the propensity score is specifically for the first threshold, that is, dose $k = 1$ vs. doses $k > 1$. Separate propensity scores could be calculated for each of the $k-1$ cumulative response probabilities. Based on the proportional odds assumption (McCullagh, 1980), however, the random subject effects and covariate effects are constant across the cumulative logits. As a result, the ranking of propensity scores will not change with separate propensity scores per cumulative logit. Here we assume proportional odds and, for that reason, define the propensity score for the ordinal dose in terms of the probability of receiving a dose greater than the first dose (i.e., the first threshold). This assumption could be relaxed, however that would increase the complexity in propensity score calculation in a non-trivial manner, in that it would be necessary to calculate a separate score for each observation for each cumulative logit (Hedeker and Mermelstein, 1998). Also, as noted by McCullagh and Nelder (1989), this extension of the ordinal model can lead to negative estimated probabilities in some cases. The propensity score (2.3) includes the contribution of covariates x and subject effects υ on the probability of receiving more intensive treatment (i.e., a higher value of the ordinal dose T). An observation with a high propensity score has characteristics of someone more likely to receive intensive treatment at time point j. Conversely, an observation with a low propensity score has characteristics of someone less likely to receive intensive treatment. The propensity score can vary within-subjects over time. Thus, as illness severity increases over time, for example, the propensity for treatment intensity can change, presumably increasing. However, time-varying variables included in the propensity score must be assessed prior to the commencement of the predicted treatment.

2.2 Propensity quantile classification

As stated earlier, stratification on propensity score, $e(x_{ij}, \upsilon_i)$, will ameliorate confounding effects of the variables included in the propensity score. When stratification is used, Rosenbaum and Rubin recommend using quintiles. This is based on results from Cochran's (1968) evaluations of various subclassification strategies. Here we examine the applicability of quintile stratification to longitudinal studies by comparing bias reduction from various forms of quantile stratification. Each observation for subject i at time j is classified into a propensity quantile based on the propensity score, and this classification can vary within subject, over time, reflecting the waxing and waning of treatment needs during the course of a chronic illness. Treatment ef-

fectiveness analyses are then conducted separately for each quantile. An implicit assumption of quantile-specific effectiveness analyses is that all treatments are well-represented in each quantile. This can be verified by examining a contingency table of treatment by propensity quantile. We proceed with analyses if there are at least 5 observations per cell. However, there is not a consensus regarding the minimum stratum size.

3. LONGITUDINAL TREATMENT EFFECTIVENESS ANALYSES

3.1 Propensity quantile-stratified effectiveness analyses

Stage two of the procedure involves a treatment effectiveness model. Let Y_{ij} represent health status for subject i at time j, such that $Y_{ij} = 1$ indicates that the subject is sick and $Y_{ij} = 0$ indicates that the subject is well. Let $P(Y_{ij} = 1)$ represent the probability that subject i is sick at time j . The within quantile treatment effectiveness analyses of repeated binary outcomes can be evaluated with a mixed-effects logistic regression model, expressed as:

$$\ln\left[\frac{P(Y_{ij}=1)}{1-P(Y_{ij}=1)}\right] = \alpha_0 + T'_{ij}\alpha + \theta_i, \tag{3.1}$$

where Y_{ij} is the binary outcome for subject i ($i = 1, \ldots, N$), at time j ($j = 1, \ldots, J_i$), α_0 is the intercept, T_{ij} is the $(k-1) \times 1$ vector of treatment doses, α contains the corresponding regression coefficients, and θ_i is the subject-specific random effect that is normally distributed in the population with mean 0 and variance σ^2_θ. This vector T comprises time-varying treatment effects, allowing for changes in doses over time, which can be specified as conditional odds ratios (i.e., conditional on the random subject effect θ_i). (Note that unlike the mixed-effects ordinal logistic model of the propensity for treatment intensity which involves cumulative logits for multiple ordinal doses, there is only one logit for the binary outcome.)

As stated earlier, the treatment effectiveness analyses include observations from all times j ($j = 1, \ldots, J_i$) and are conducted separately for each quantile. (We do not specify the form of the quantile at this point because one objective of this manuscript is to compare the performance of four stratification strategies: median-split, terciles, quartiles, and quintiles.) The model includes a random subject effect because a given subject can have multiple observations within a quantile-specific analysis.

3.2 Pooling the quantile-specific effectiveness results

The quantile-specific treatment effectiveness estimates are pooled by implementing the Mantel-Haenszel (1959) strategy as described by Fleiss (1981). The pooled parameter estimates are weighted means of quantile-specific parameter estimates, where each weight is the inverse of the respective squared standard error. The use of pooled estimates assumes that there are not differential treatment effects across the quantiles. The effectiveness analyses that test this assumption are conducted with data pooled across quantiles. The model includes treatment effects, a quantile class variable, the treatment by quantile interaction terms and a subject-specific random effect. A likelihood ratio test examines the incremental contribution of the quantile by treatment interaction terms relative to the more parsimonious model that does not include the interaction. (All other effectiveness analyses involve quantile-specific data.) The analyses and interpretation of results will differ when the assumption of no quantile by treatment interaction is violated. For instance, quantile-specific results cannot be pooled. Consequently, quantile-specific results must be presented and discussed separately. The interpretation of quantile-specific results will include a description of the characteristics of subjects comprising the various quantiles (based on the composition of the propensity score) and which treatment worked best in each quantile.

4. APPLICATION

4.1 Antidepressant effectiveness

The National Institute of Mental Health Collaborative Depression Study (CDS) is a longitudinal, observational study (Katz and Klerman, 1979). Patients with mood disorders were enrolled into the CDS at academic medical centers in five sites (Boston, MA; Chicago, IL; Iowa City, IA; New York, NY; and St. Louis, MO) from 1978 through 1981. After receiving a complete description of the study, each subject provided written informed consent. The 20 year follow-up data are used to illustrate the application of the longitudinal propensity adjustment. Episodes were defined according to the Research Diagnostic Criteria (Spitzer, Endicott, Robins, 1978). The analyses include 2230 observations of 187 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, and had at least one subsequent depressive episode.

We refer to the unit of analysis as a *treatment interval*, which was defined as follows. The somatic antidepressant treatments examined here include pharmacotherapy and electroconvulsive therapy. They were coded in five ordinal categories of intensity ranging from 0 (no treatment) to 4 (most intensive dose), which have been described in detail elsewhere (Keller, 1988; Leon, 2003). For example, the following intensity ratings were used for the antidepressant citalopram: 1 (1–19 mg), 2 (20–39 mg), 3 (40–59 mg), and 4 ($\geq$60 mg). Each treatment interval commenced during a

Table 1. A model of propensity for treatment intensity during a major depressive episode

Variable[1]	Adjusted Odds Ratio	CI low	CI high	Z	p-value
Symptom severity* (range: 1–6)	1.14	1.03	1.26	2.64	0.008
Symptom trajectory*					
worsening	1.48	1.19	1.83	3.59	< .001
improving	0.99	0.77	1.27	−0.08	0.934
Duration of Episode (number of months prior to treatment interval)	0.996	0.992	1.001	−1.65	0.100
Age (years)					
< 30	1.00				
30–39	1.90	1.41	2.57	4.18	< .001
40–49	2.22	1.53	3.21	4.22	< .001
50–59	1.60	1.02	2.51	2.04	0.041
60+	1.31	0.87	1.97	1.28	0.199
Study Site					
St. Louis	1.00				
New York	2.81	1.29	6.11	2.60	0.009
Boston	1.18	0.71	1.95	0.62	0.532
Iowa	1.94	1.32	2.84	3.40	0.001
Chicago	1.52	0.99	2.33	1.90	0.057

187 subjects; 2230 number of observations; ICC = 0.100; * prior 8 weeks.

Table 2. Cross-classification of treatment intensity by propensity quintile: Number of observations and response rates

	Propensity Quintile					
	Q1	Q2	Q3	Q4	Q5	Total
Treatment Intensity						
Dose 0	189	112	82	61	24	468
(Response[1] rate)	(21.7%)	(23.2%)	(20.7%)	(18.0%)	(12.5%)	(19.2%)
Dose 1	148	124	95	85	44	496
	(16.9%)	(29.0%)	(16.8%)	(9.4%)	(6.8%)	(15.8%)
Dose 2	73	123	152	122	117	587
	(27.4%)	(35.8%)	(36.2%)	(33.6%)	(17.9%)	(30.2%)
Dose 3	26	52	76	101	137	392
	(26.9%)	(32.7%)	(39.5%)	(26.7%)	(25.5%)	(30.3%)
Dose 4	10	35	41	77	124	287
	(30.0%)	(31.4%)	(41.5%)	(37.7%)	(38.7%)	(35.9%)
Total	446	446	446	446	446	2230

[1] "Response" is defined as reduction in symptom severity.

major depressive episode. The initial week of a treatment interval corresponded with the first week of a new treatment intensity. An interval terminated with a dose change or, if no dose change occurred, at the end of follow-up. By virtue of the longitudinal design, subjects typically had multiple treatment intervals (mean = 11.9; median = 8; sd = 13.4), and for each interval, had a unique propensity score.

The binary dependent variable in the effectiveness analyses was reduction of symptom severity (yes/no) from the beginning to the end of the treatment interval. The Psychiatric Status Rating, a component of the Longitudinal Interval Follow-up Evaluation (Keller et al., 1987), was used to assess symptom severity. It was hypothesized that more intensive somatic antidepressant treatment would correspond with a greater likelihood of reduction of severity.

4.2 Results

4.2.1 Propensity for treatment intensity

Mixed-effects ordinal logistic regression analyses were used to estimate the propensity score. The dependent variable in these models was the ordinal treatment intensity ranging from *dose 0* to *dose 4*. The components of the propensity score included two demographic and three clinical variables (Table 1). The clinical variables indicate that those with more severe depressive symptoms tended to get more aggressive treatment. For example, those whose symptoms were worsening during the eight weeks prior to the start of the treatment were nearly 50% more likely to get more intensive treatment than those whose symptom severity remained stable. The within-subject treatment intensity

Table 3. Comparison of unadjusted and propensity–adjusted mixed-effects ordinal logistic regression models of ordinal doses

	Unadjusted Results				Propensity-Adjusted Results			
Variable[1]	Odds Ratio	CI low	CI high	p-value	Odds Ratio	CI low	CI high	p-value
Symptom severity (range: 1–6)	1.22	1.12	1.32	<0.001	0.956	0.877	1.042	0.309
Symptom trajectory				0.007				0.046
Stable	1.00				1.00			
Worsening	1.37	1.10	1.69	0.004	0.79	0.64	0.97	0.023
Improving	1.10	0.87	1.40	0.406	0.99	0.78	1.25	0.918
Duration of Episode (number of months prior to treatment interval)	0.995	0.9920	0.9989	0.009	1.002	0.998	1.006	0.402
Age (years)				<0.0001				0.889
<30	1.00				1.00			
30–39	1.72	1.28	2.30	<0.001	0.98	0.78	1.24	0.888
40–49	1.84	1.30	2.59	0.001	0.94	0.70	1.27	0.704
50–59	1.42	0.93	2.17	0.101	0.96	0.66	1.39	0.825
60+	1.41	0.92	2.16	0.111	1.06	0.77	1.47	0.725
Study Site				0.0008				0.301
St. Louis	1.00							
New York	2.56	1.13	5.83	0.025	0.76	0.36	1.60	0.469
Boston	1.23	0.73	2.09	0.436	0.82	0.47	1.45	0.502
Iowa	1.86	1.34	2.59	<0.001	0.83	0.58	1.17	0.275
Chicago	1.52	1.04	2.23	0.030	0.81	0.54	1.22	0.317

[1]Separate models evaluated each of the variables.

over the longitudinal course of follow-up was erratic as quantified by the intraclass correlation coefficient (ICC) among observations of 0.10, perhaps a function of the 20 years of follow-up data.

A propensity score was calculated for each level one observation (i.e., the repeated within subject treatments over time), based on the results of the mixed-effects ordinal logistic regression model. Each observation was classified into the quintile that corresponded to the respective propensity score. The assumption that all treatments are represented in each quintile was evaluated by examining a contingency table that cross-classified the dose by propensity score quintile (Table 2). It is clear that, overall, observations in the lower quintiles tended to get less aggressive treatment; whereas observations in the higher quintiles tended to get more aggressive treatment. Nevertheless, each level of treatment was represented in each quintile and propensity quintile-stratified analyses were conducted.

4.2.2 Evaluation of propensity-adjusted balance

Bias reduction in the treatment effectiveness estimate is the motivation for using the propensity adjustment. This is referred to as the *balancing property* of the propensity adjustment (Rosenbaum and Rubin, 1983). We evaluated the degree to which balance was achieved by conducting separate mixed-effects ordinal logistic regression analyses for each component of the propensity score. In each model, ordinal treatment dose was the dependent variable (as in the propensity model) and the propensity score and one of its components were included as covariates. A comparison of the unadjusted and adjusted odds ratios (Table 3) reveals that the magnitude of the association between each component of the propensity score and dose is attenuated considerably with the propensity adjustment. The only propensity component that remains significantly associated with dose is symptom trajectory and the strength of that association was reduced substantially with the propensity adjustment.

4.2.3 Treatment effectiveness

The effectiveness of each of the four higher levels of treatment intensity was evaluated relative to no somatic treatment in quintile-stratified analyses. The quintile-specific response rates are presented in Table 2. The quintile-specific parameter estimates were pooled using the Mantel-Haenszel procedure (as described above) and those pooled results are discussed here. (Pooled estimates are used because the treatment by propensity-quintile interaction was not statistically significant: $-2LL = 19.518$; $df = 16$; $p = 0.243$.) There were statistically significant effects of the three higher doses (Table 4). Those who received the highest treatment intensity (dose 4) were more than twice as likely to have reduction in symptom severity (odds ratio = 2.27; 95% CI: 1.45–3.55; $p < .001$) than those who received no somatic treatment. Each of the two next highest treatment intensities also showed an increased chance of symptom reduction relative to who received no somatic treatment (dose 3: OR = 1.74;

Table 4. Treatment effectiveness analyses: Pooled results

Variable	Odds Ratio	95% Confidence Interval	
intercept	0.30	0.23	0.38
dose 1	0.87	0.63	1.21
dose 2	1.71	1.23	2.38
dose 3	1.74	1.17	2.57
dose 4	2.27	1.45	3.55

95% CI: 1.17–2.57; p = .006; dose 2: OR = 1.71; 95% CI: 1.23–2.38; p = .002). In contrast, the lower treatment intensity was not significantly beneficial (dose 1: OR = 0.87; 95% CI: 0.63–1.21; p = .401). It is also worth noting that, as with treatment intensity over time, subjects had inconsistent response to treatment over the course of follow-up (ICC = .08). In summary, longitudinal propensity-adjusted effectiveness analyses demonstrated that, despite a presentation of more severe depressive symptoms, those who received higher doses of treatment were more likely to have reduction in illness severity over the course of treatment.

5. SIMULATION STUDY

A simulation study was conducted to examine the longitudinal propensity adjustment described above. The simulations compared the performance of quintile stratification to three alternative quantile-stratified approaches to propensity-based stratification: quartiles, terciles, median-split. These are successively more coarse approaches to stratification in that the within-stratum heterogeneity is increased with a smaller number of strata and, as a result, greater between treatment group imbalance on pretreatment covariates is expected. With each method of stratification, the subject-specific random effect is included in the models to account for the correlated observations seen in a longitudinal design with repeated within-subject observations. The following approach was used.

5.1 Specifications for the propensity score simulation

A logistic model was used to simulate propensity scores for each subject. Initially, four randomly generated predictor variables were generated. These include two time-invariant binary variables (X_1, X_2), each consisting of half zeros and half ones, based on dichotomization of randomly generated standard normal deviates, and two time-varying continuous variables (X_3, X_4) based on randomly generated standard normal deviates. The pairwise correlation among all propensity model predictor variables was set at .20. Also, for the time-varying predictors, the intraclass correlation coefficient (ICC) ρ was specified as 0.40. Odds ratios for each of the four predictors of dose in the propensity model varied (1.4, 1.6, 1.8, 2.0). Based on these specifications, a latent (continuous) propensity score was generated for each subject at each time point; these were categorized using the threshold concept (Agresti, 2002) to create the observed time-varying ordinal doses ($k = 0, 1, 2, 3$). In doing this, threshold values were specified to yield realistic distributions across the ordinal doses. In terms of sample size, we examined samples of 200 subjects, where each subject had ten repeated observations over the course of time.

5.2 Specifications for the treatment effectiveness simulation

A separate logistic model was used to simulate treatment outcomes across time. For this, we used the simulated doses, described above in 5.1, and specified effects of each of the three doses on the outcome (relative to the control, dose *0*) as odds ratios of 1.0, 1.5, and 2.0, respectively. The ICC among the repeated outcomes within subjects was set at 0.40. Again, a latent (continuous) outcome was generated for each subject at each timepoint, based on the effects of the doses and the ICC, and categorized using the threshold approach. One thousand data sets were generated and analyzed for each combination of simulation specifications.

5.3 Model performance criteria

The four quantile-stratification approaches were compared on several criteria, among which bias reduction was the focus. Bias is defined as the absolute value of the difference between the specified treatment effect and the parameter estimate and is reported on the scale of the parameter (not on the odds ratio scale). Bias reduction was defined relative to the bias in models that did not incorporate a propensity adjustment and expressed as percent reduction. Additional criteria included coverage for the effects each of the three higher doses ($k = 1, 2, 3$), type I error for tests of the second dose ($k = 1$), and statistical power of the tests for the higher two doses ($k = 2, 3$). Coverage represents the proportion of simulations in which the specified value was included in the 95% confidence interval for a particular parameter estimate. Coverage that is less than 90% is highlighted in the table (Collins, Schafer and Kam, 2001). All evaluation criteria are based on the Mantel-Haenszel pooled results of quantile-specific estimates. The MIXOR program (Hedeker and Gibbons, 1996) was used for these simulations.

5.4 Simulation results

5.4.1 Bias

As expected, the unadjusted models have a great deal more bias than the propensity-adjusted models. Bias decreases as the number of quantiles used for stratification increases from two to five (Table 5). Nevertheless, there was a small degree of bias with propensity score quintile stratification. Moreover, as the association of the propensity covariates (i.e., the confounding variables) with treatment dose and outcome increases, the magnitude of the bias of parameter estimates increases. Similar patterns are seen for the effects of each of the three treatment doses compared to dose 0.

Table 5. Results of simulation study: A comparison of type I error, statistical power, and bias for various quantile stratification methods

Quantile Strategy	Propensity Covariate Odds Ratio (OR)	Type I Error	Statistical Power		Bias[2]			Coverage[1]		
		Treatment Effect (OR)			Treatment Effect (OR)			Treatment Effect (OR)		
		1.0	1.5	2.0	1.0	1.5	2.0	1.0	1.5	2.0
Unadjusted	1.4	0.07	0.76	1.00	0.07	0.09	0.14	0.93	0.92	**0.88**
	1.6	0.09	0.85	1.00	0.10	0.15	0.26	0.91	**0.87**	**0.68**
	1.8	0.14	0.93	1.00	0.17	0.22	0.38	**0.86**	**0.78**	**0.38**
	2.0	0.21	0.97	1.00	0.22	0.31	0.51	**0.79**	**0.65**	**0.13**
Median Split	1.4	0.05	0.64	0.97	0.05	0.06	0.07	0.95	0.95	0.94
	1.6	0.06	0.69	0.98	0.08	0.08	0.14	0.94	0.93	0.91
	1.8	0.09	0.73	0.98	0.11	0.12	0.17	0.91	0.92	**0.88**
	2.0	0.11	0.80	1.00	0.14	0.17	0.23	**0.89**	**0.89**	**0.81**
Terciles	1.4	0.05	0.57	0.95	0.04	0.04	0.04	0.95	0.95	0.96
	1.6	0.06	0.57	0.95	0.06	0.05	0.07	0.94	0.95	0.95
	1.8	0.06	0.58	0.96	0.07	0.05	0.07	0.94	0.95	0.94
	2.0	0.07	0.64	0.97	0.09	0.09	0.11	0.93	0.93	0.93
Quartiles	1.4	0.05	0.53	0.94	0.03	0.03	0.03	0.95	0.95	0.95
	1.6	0.05	0.51	0.92	0.04	0.02	0.04	0.95	0.96	0.95
	1.8	0.04	0.50	0.92	0.04	0.02	0.02	0.96	0.95	0.95
	2.0	0.06	0.54	0.92	0.06	0.05	0.04	0.94	0.95	0.95
Quintiles	1.4	0.04	0.50	0.92	0.02	0.02	0.02	0.96	0.96	0.96
	1.6	0.06	0.47	0.90	0.03	0.01	0.02	0.94	0.96	0.96
	1.8	0.04	0.45	0.88	0.03	0.00	0.01	0.96	0.95	0.95
	2.0	0.05	0.47	0.89	0.03	0.02	0.00	0.95	0.95	0.96

[1]Coverage that is less than 90% is **bolded**. [2]Bias is presented on the scale of the parameter, not the odds ratio scale.

5.4.2 Bias reduction

Relative to an unadjusted model, bias reduction increases monotonically with the number of quantiles (Figure 1). Quintiles show the greatest bias reduction, removing from 66% to 100% (median 92%). Nonetheless, there is less bias reduction for the null treatment effect.

5.4.3 Coverage

The 95% confidence intervals coverage for parameter estimates is appropriate for the tercile, quartile, and quintile stratification procedures (Table 5). The unadjusted models had inadequate coverage probability, particularly for confounding variables with higher odds ratios. The acceptability of coverage with the median-split approach was erratic, dropping below 90% coverage for about one-third of the simulation specifications.

5.4.4 Type I error

Unadjusted models have inflated rates of type I error (Table 5). This is amplified with an increase in the propensity covariate odds ratios. In the adjusted models, Type I error rates decrease slightly with more quantiles. The median type I error rates are 7.5% (median-split), 6% (terciles), 5% (quartiles), and 4.5% (quintiles). Furthermore, when the number of quantiles is less than four, the type I error rates consistently increase as the association of the confounding variables with dose and outcome increases.

5.4.5 Statistical power

Statistical power for treatment effect odds ratios of 2.0 ranged from .88 to 1.00 (Table 5). There is a pattern of a decrease in power with a greater number of quantiles for treatment effects of 1.5. There appears to be no effect of the strength of the confounding variable on power.

6. DISCUSSION

A two-stage mixed-effects propensity adjustment for ordinal treatments with repeated binary outcomes was described and shown to reduce the bias in estimates of treatment effectiveness. The first stage involves a mixed-effects ordinal logistic regression model of propensity for treatment intensity. In the second stage, treatment effectiveness analyses are stratified by propensity for treatment quantiles. A simulation study evaluated the impact of various forms of quantile stratification on bias reduction. The longitudinal implementation of the propensity adjustment removed considerable bias, and that bias reduction increased with the number of quantiles. Relative to the unadjusted models, the 95% confidence interval coverage

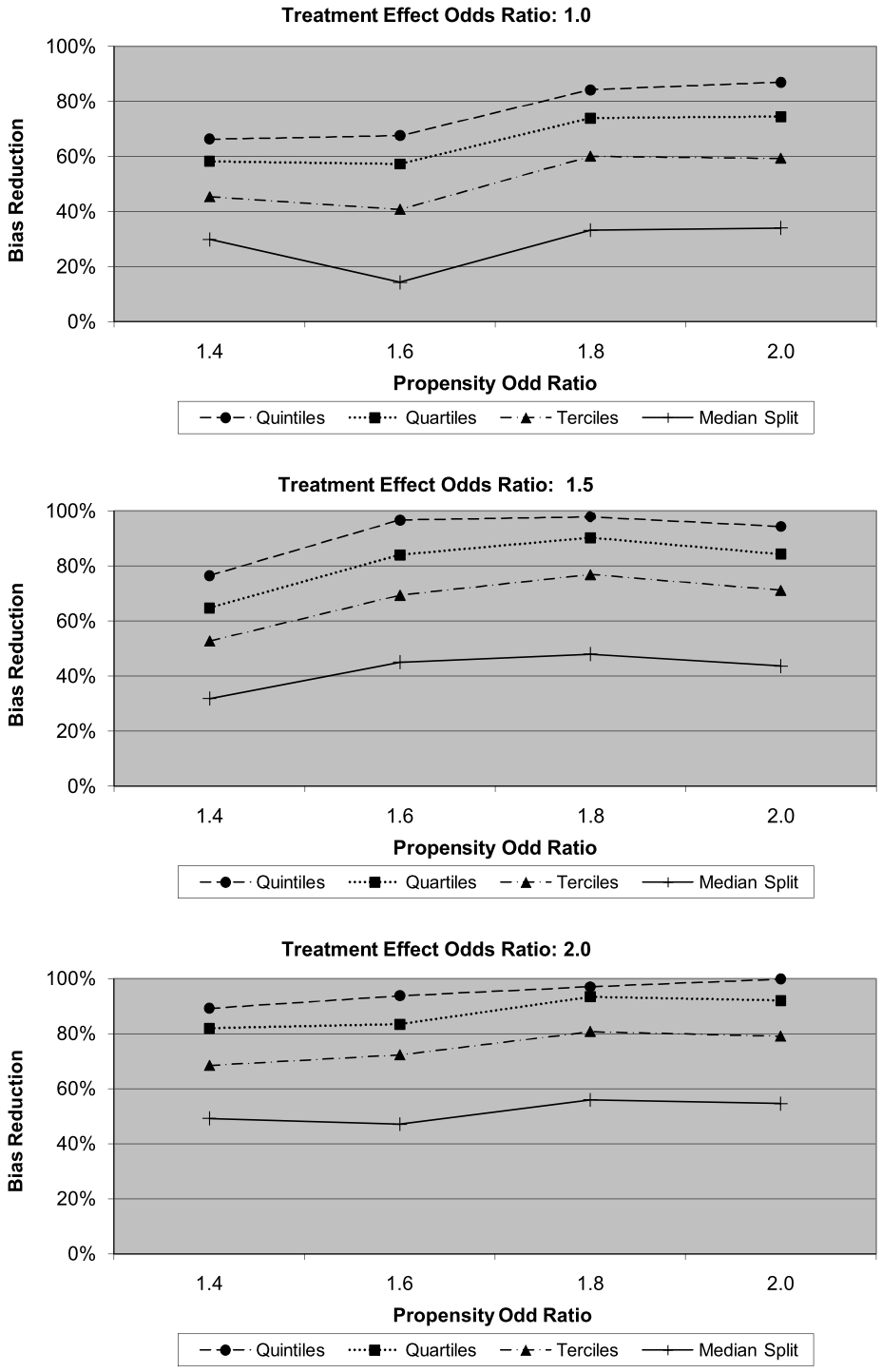

Figure 1. Percent bias reduction for various propensity-quantile stratification methods.

for three or more strata was much improved. Type I error was more acceptable for a greater number of strata, but statistical power decreased with more quantiles. Overall, if a stratification approach is adopted, the use of quartiles or quintiles is recommended. The results of the simulation study show that the mixed-effects approach to longitudinal data adequately accounts for non-independent units of analysis, whether within or between quantiles.

The model was illustrated with data from a longitudinal, observational study of antidepressant effectiveness. Although those who had more severe symptoms tended to receive more intensive treatment, this approach was able to

detect statistically significant beneficial effects of the more intensive treatments on the probability of symptom reduction. This is due to the success of the propensity adjustment in reducing imbalance across treatment groups.

As mentioned earlier, there was a small degree of bias remaining with the propensity score quintile stratification approach that we examined. This is not to be unexpected, based on the work of Gail et al. (1984) who, in work that that focused on the problem of omitted covariates, showed bias with odds ratios that are used to quantify the treatment effect in RCTs. It is conceivable that an alternative approach to longitudinal analyses would yield unbiased estimates as shown with cross sectional data (e.g., rate ratios examined in Austin et al., 2007).

There are several limitations to the evaluation of the approach that we have described. First, it is unclear to what extent the results of this simulation study apply to settings that differ from the specifications that were examined here. For instance, with considerably small sample sizes or a small number of repeated within-subject observations, the quintile stratification strategies could become infeasible because of the risk that some treatments will be inadequately represented in at least one quintile. In such a case, inverse probability weighting (IPW) or matching on the longitudinal propensity score might be useful. Such approaches have been examined with cross-sectional data (Lunceford and Davidian, 2004; Forbes and Shortreed, 2007). Marginal structural models (MSM) provide an alternative approach to examining time-varying treatments. MSM uses inverse probability weighting in an effort to produce between-group balance in covariates. Research is needed to compare the performance of stratification, IPW and matching with longitudinal data. For instance will the bias reduction from an alternative to stratification also come at the expense of power?

Furthermore, our simulations did not examine model performance with a larger number of propensity covariates that might be applied in practice. Likewise, the simulations did not examine performance when more than five strata are used. This was based on the practice of using quintiles that was established by the work of Cochran (1968) who showed that little bias reduction was gained with more strata for variables generated from various distributions. In addition, a limitation of the model is the simplifying assumption of proportional odds for the covariate effects. However, as discussed earlier, if the assumption were relaxed, it would be necessary to calculate a separate propensity score for each observation for each cumulative logit (Hedeker and Mermelstein, 1998) and that would substantially increase the complexity involving the use of propensity scores in this context. Finally, with the exception of unadjusted models, the simulations did not examine the issue of hidden bias. Rosenbaum (2002) has described sensitivity analyses that can examine the impact of hidden bias. The impact of misspecified propensity models has been considered for both cross-sectional (Drake, 1993) and longitudinal data (Leon and Hedeker, 2007).

In conclusion, the longitudinal propensity adjustment that has been described and evaluated here provides a data analytic strategy for treatment effectiveness estimates of ordinal doses in observational studies. A well-conducted RCT with minimal attrition is certainly preferable to an observational study of treatment. Yet there are post-marketing pharmaceutical studies where randomized treatment assignment is not feasible. It is in that context that the longitudinal propensity method can be applied to provide guidance for those seeking evidence-based treatment options.

ACKNOWLEDGEMENTS

This research was supported, in part, by grants from the National Institute Health (MH060447, MH092606 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, M.D. (Chairperson, Providence), W. Coryell (Co-Chair Person, Iowa City); D.A. Solomon, M.D. (Providence); W.A. Scheftner, M.D. (Chicago); W. Coryell, M.D. (Iowa City); J. Endicott, Ph.D., A.C. Leon, Ph.D., J. Loth, M.S.W. (New York); J. Rice, Ph.D., (St. Louis). Other current contributors include: H.S. Akiskal, M.D., J. Fawcett, M.D., L.L. Judd, M.D., P.W. Lavori, Ph.D., J.D. Maser, Ph.D., T.I. Mueller, M.D. This manuscript has been reviewed by the Publication Committee of the Collaborative Depression Study, and has its endorsement. The data for the application in this manuscript came from the National Institute of Mental Health (NIMH) Collaborative Program on the Psychobiology of Depression-Clinical Studies (Katz and Klerman, 1979). 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 centers and included Gerald Klerman, M.D.† (Co-Chairperson), Martin Keller, M.D., Robert Shapiro, M.D.† (Massachusetts General Hospital, Harvard Medical School), Eli Robins, M.D.,† Paula Clayton, M.D., Theodore Reich, M.D.,† Amos Wellner, M.D.† (Washington University Medical School), Jean Endicott, Ph.D., Robert Spitzer, M.D. (Columbia University), Nancy Andreasen, M.D., Ph.D., William Coryell, M.D., George Winokur, M.D.† (University of Iowa), Jan Fawcett, M.D., William Scheftner, M.D. (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, Ph.D., Branch Chief as the Co-Chairperson and Robert Hirschfeld, M.D. as the Program Coordinator. Other past contributors include: J. Croughan, M.D., M.T.

†Deceased.

Shea, Ph.D., R. Gibbons, Ph.D., M.A. Young, Ph.D., D.C. Clark, Ph.D.

Received 21 October 2010

REFERENCES

Agresti, A. (2002). *Categorical Data Analysis*, 2nd ed. John Wiley and Sons, Hoboken, NJ. MR1914507

Austin, P. C., Grootwndorst, P., Normand, S.-L. T. and Anderson, G. M. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: A Monte Carlo study. *Statistics in Medicine* **26** 754–768. MR2339172

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* **24** 295–313. MR0228136

Collins, L. M., Schafer J. L. and Kam C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods* **6** 330–351.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of the treatment effect. *Biometrics* **49** 1231–1236.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions*. John Wiley and Sons, New York, 161–175. MR0622544

Forbes, A. and Shortreed, S. (2008). Inverse probability weighted estimation of the marginal odds ratio: Correspondence regarding "The performance of different propensity score methods for estimating marginal odds ratios". *Statistics in Medicine* **27** 5556–5559. MR2542369

Gail, M. H., Wieand, S. and Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* **7** 431–444. MR0775390

Hedeker, D. and Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics* **50** 933–944.

Hedeker, D. and Gibbons, R. D. (1996). MIXOR: A computer program for mixed-effects ordinal regression analysis. *Computational Methods and Programs for Biomedicine* **49** 157–176.

Hedeker, D. and Mermelstein, R. J. (1998). A multilevel thresholds of change model for analysis of stages of change data. *Multivariate Behavioral Research* **33** 427–455.

Imbens, G. W. (2000). The role of propensity scores in estimating dose-response functions. *Biometrika* **87** 706-710. MR1789821

Joffe, M. M. and Rosenbaum, P. R. (1999). Invited commentary: Propensity scores. *American Journal of Epidemiology* **150** 327–333.

Katz, M. M. and Klerman, G. L. (1979). Introduction: Overview of the clinical studies program of the NIMH clinical research branch collaborative study on psychobiology of depression. *American Journal of Psychiatry* **136** 49–51.

Keller, M. B. (1988). Undertreatment of major depression. *Psychopharmacology Bulletin* **24** 75–80.

Keller, M. B., Lavori, P. W., Friedman, B., Nielsen, E., Endicott, J., McDonald-Scott, P. and Andreason, N. C. (1987). The longitudinal interval follow-up evaluation: A comprehensive method for assessing outcome in prospective longitudinal studies. *Archives of General Psychiatry* **44** 540–548.

Leon, A. C. and Hedeker, D. (2005). A mixed-effects quintile-stratified propensity adjustment for effectiveness analyses of ordered categorical doses. *Statistics in Medicine* **24** 647–658. MR2134531

Leon, A. C. and Hedeker, D. (2007). A comparison of mixed-effects quantile stratification propensity adjustment strategies for longitudinal treatment effectiveness analyses of continuous outcomes. *Statistics in Medicine* **26** 2650–2665. MR2370830

Leon, A. C. and Hedeker, D. (2007). Quintile stratification based on a misspecified propensity score in longitudinal treatment effectiveness analyses of ordinal doses. *Computational Statistics and Data Analysis* **51** 6114–6122. MR2407702

Leon, A. C., Solomon, D. A., Mueller, T. I., Endicott, J., Rice, J. P., Maser, J. D., Coryell, W. and Keller, M. B. (2003). A 20-year longitudinal, observational study of somatic antidepressant treatment effectiveness. *American Journal Psychiatry* **160** 727–733

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine* **23** 2937–2960.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal National Cancer Institute* **22** 719–748.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *J. Roy. Statist. Soc. B.* **42** 109–142. MR0583347

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd ed. Chapman and Hall, New York. MR0727836

Rosenbaum, P. R. (2002). *Observational Studies*. 2nd ed. Springer-Verlag, New York. MR1899138

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974

Spitzer, R., Endicott, J. and Robins, E. (1978). Research diagnostic criteria: Rationale and reliability. *Archives of General Psychiatry* **35** 773–782.

Andrew C. Leon
Weill Cornell Medical College
Department of Psychiatry
Box 140
525 East 68th Street
New York, NY 10065
USA
Tel.: (212) 746-3872
E-mail address: acleon@med.cornell.edu

Donald Hedeker
University of Illinois at Chicago
USA