

A mixed-effects quintile-stratified propensity adjustment for effectiveness analyses of ordered categorical doses[‡]

Andrew C. Leon^{1,*},[†] and Donald Hedeker²

¹*Weill Medical College of Cornell University, U.S.A.*

²*University of Illinois at Chicago, U.S.A.*

SUMMARY

Observational studies can be used to evaluate treatment effectiveness among patients with a broader range of illness severity than typically seen in randomized controlled clinical trials. However, there are several difficulties with observational evaluations including non-equivalent comparison groups, treatment doses and durations that vary widely, and, in longitudinal studies, multiple courses of treatment per subject. A mixed-effects approach to the propensity adjustment for non-equivalent comparison groups is described that can account for each of these perturbations. The strategy involves two stages. First, characteristics that distinguish among subjects who receive various levels of treatment are examined in a model of propensity for treatment intensity using mixed-effects ordinal logistic regression. Second, the propensity-stratified effectiveness of ordered categorical doses is compared in a mixed-effects grouped time survival model of time until recovery. The model is applied in a longitudinal, observational study of antidepressant effectiveness. Then a Monte Carlo simulation study indicates that the strategy has acceptable type I error rates and minimal bias in the estimates of treatment effectiveness. Statistical power exceeds 0.90 for an odds ratio of 1.5 with $N = 250$ and 500, and is acceptable for an odds ratio of 2.0 with $N = 100$. Nevertheless, with $N = 100$, the models that had high intraclass correlation coefficients had greater tendency towards non-convergence. This approach is a useful strategy for observational studies of treatment effectiveness. It is capable of adjusting for selection bias, incorporating multiple observations per subject, and comparing effectiveness of ordinal doses. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: treatment effectiveness; propensity analyses; propensity adjustment; mixed-effects models; observational study

*Correspondence to: Andrew C. Leon, Weill Medical College of Cornell University, Department of Psychiatry, Box 140, 525 East 68th Street, New York, NY 10021, U.S.A.

[†] E-mail: acleon@med.cornell.edu

[‡] Presented, in part, at the *9th Biennial CDC/ATSDR Symposium on Statistical Methods*, 28–29 January, 2003, Atlanta, GA.

Contract/grant number: MH60447

1. INTRODUCTION

Randomized controlled clinical trials (RCT) are the scientific standard for treatment evaluation. The RCT design places a strong emphasis on internal validity with randomized group assignment, double-blinding, and control or comparison groups. However, the generalizability of RCT results is limited by the inclusion and exclusion criteria which tend to exclude those with comorbid disorders, those with a mild form of the illness, and, for both ethical and legal reasons, the most severely ill. Thus, RCT results provide valuable information about the efficacy of treatments in tightly controlled settings, but often do not apply to a substantial minority of those afflicted with the illness that the treatment is designed to help.

Observational studies, on the other hand, can provide supplementary information on treatment effectiveness among patients with a broader range of illness severity. However, observational investigators of treatment effectiveness are faced with several methodological difficulties including non-equivalent comparison groups, treatment doses and durations that vary widely, and, in longitudinal studies, multiple courses of treatment per subject. By design, an investigator observes, but does not manipulate treatment in an observational study. Treatment assignment is not randomized, but is typically influenced by characteristics of the patient. For instance, those with more severe symptoms might be likely to receive the most aggressive treatment; whereas those with mild symptoms might be likely to receive no treatment at all. In which case, symptom severity is a confounding variable because it is associated with both the independent and dependent variables. As a result, unadjusted comparisons of the untreated subjects with those who received aggressive treatment would likely draw the conclusion that more intensive treatment is associated with worse outcome. We describe an adjustment strategy for treatment effectiveness analyses in longitudinal, observational studies.

The approach involves two stages. First, characteristics that distinguish among subjects who receive various levels of treatment are examined in a model of *propensity for treatment intensity* using mixed-effects ordinal logistic regression. Second, the *effectiveness of ordered categorical* doses is compared in a mixed-effects grouped time survival model of time until recovery. Initially stratified effectiveness analyses are conducted, separately for each propensity quintile. Then, if there is not a propensity by treatment interaction, those quintile-specific results are pooled. It is through this subclassification, or stratification, that the bias of observed confounding variables is reduced [1]. Cochran showed that stratification of a continuous confounding variable into quintiles is sufficient to remove most associated bias for certain distributions [1].

2. THE PROPENSITY AND TREATMENT EFFECTIVENESS MODELS

2.1. Propensity adjustment

Rosenbaum and Rubin [2], who defined the propensity score as ‘the conditional probability of assignment to a particular treatment given a vector of observed covariates’, have shown that adjustments on the propensity score can be used to reduce the bias in estimates of treatment effectiveness in an observational study [2, 3]. A propensity score, $e(x)$, using the Rosenbaum

and Rubin notation,

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

is derived for each subject i ($i = 1, \dots, N$) from the logistic model

$$\ln \left[\frac{P(T_i = 1)}{1 - P(T_i = 1)} \right] = \alpha + x'_i \beta$$

where the intercept, α , and the vector of coefficients, β , 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 ($T_i = 0$). The propensity score can be estimated as

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

Rosenbaum and Rubin [2] describe the application of the propensity score with three standard adjustment procedures: matching, subclassification, and covariance adjustment.

2.2. Mixed-effects propensity adjustment for effectiveness analyses of ordinal treatments

The strategy evaluated here is a variation of what we initially described as a *dynamic adaptation of the propensity adjustment* for the study of *ordinal* doses [4]. It applies the propensity methodology in a two-staged mixed-model framework that includes a mixed-effects model of propensity for treatment intensity and a mixed-effects model of treatment effectiveness. In the first stage, the propensity model examines repeated measures of ordinal doses of treatment over time. The model allows for multiple treatment intervals per subject and variations in both within-subject treatment intensity and within-subject propensity for treatment intensity during the course of the study. In the second stage, the treatment effectiveness model examines time from the start of each course of treatment until recovery. The mixed-effects framework can account for multiple within-subject recovery times that correspond to each of the treatment intervals endured by a subject.

2.2.1. Stage 1: the mixed-effects propensity adjustment. Considering the case of K ordinal dosages denoted by the variable T , the mixed-effects *ordinal propensity score*, adapting the Rosenbaum and Rubin notation, is

$$e(x, v) = P(T_{ij} > k | v, x)$$

for subject i ($i = 1, \dots, N$), at time j ($j = 1, \dots, J_i$), for dose k ($k = 1, \dots, K - 1$). The subject-varying random effect, v_i , is normally distributed in the population with mean 0 and variance σ_v^2 . The propensity score is derived from a mixed-effects ordinal logistic regression model [5] that evaluates the effect of covariates and random subject effects on $K - 1$ cumulative logits:

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

where γ_k represents the threshold for dosage k , x_{ij} is the $p \times 1$ vector of covariates, and β are the corresponding regression coefficients. Both time-varying and time-invariant covariates

can be included in the vector x . Also, the threshold values are strictly increasing and for identification the first threshold γ_1 is typically set to zero.[§] The number of cumulative logits in this model equals $K - 1$, and because the covariate effects do not carry the subscript k their effects are assumed to be the same across these cumulative logits. McCullagh [6] calls this assumption of identical odds ratios across the $K - 1$ cut-offs the proportional odds assumption.

Assuming this mixed-effects ordinal logistic model, the propensity score for subject i at time j can be expressed using the logistic response function as

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

which ranges in value from 0 to 1. Strictly speaking, the above formula yields the propensity for a response greater than the first dosage category. The propensity for responses greater than the remaining dosages would incorporate the thresholds $\gamma_2, \dots, \gamma_{k-1}$ into the formula. However, since the thresholds do not vary by subjects or time, their inclusion into the propensity score is unnecessary. The above propensity captures the essential contribution of covariates x and subject effects v on the probability of receiving more intensive dosage treatment (i.e. a higher value in terms of the ordinal dose T). Specifically, observations with low propensity scores have characteristics of someone unlikely to receive intensive treatment at a particular time point, whereas those with higher propensity scores have characteristics of someone more likely to receive intensive treatment at a particular time point.

2.2.2. Stage 2: mixed-effects treatment effectiveness analyses. Classification and examination of propensity quintiles: Once the propensity score has been estimated for subject i at time j , the observations are classified into propensity score quintiles: $q_{(1)}, \dots, q_{(5)}$. In fact, because the logistic response function is monotonic, the quintiles can be formed using: $\alpha + x'_{ij}\beta + v_i$. In the second stage of this approach, separate quintile-specific treatment effectiveness analyses are conducted. In this way, the putative confounding effects of the variables that make up the propensity score are removed from subsequent estimates of treatment effectiveness through stratification [1].

However, prior to conducting quintile-specific analyses, a contingency table of treatment by propensity quintile must be examined to determine whether each treatment is well-represented in each quintile. This, of course, is because a treatment that is not represented in a particular quintile cannot be evaluated in treatment effectiveness analyses of that quintile. Somewhat paradoxically, if the propensity score faultlessly classified most observations with regard to treatment intensity, there would be a poor representation among doses in each of the five quintiles. For instance, all those with a low propensity for treatment intensity would have received a low dose and none would have received high doses. Conversely, all those with a high propensity score would only have received a high dose, and none would have received low doses. However, in our experience, perfect concordance is not likely when attempting to explain human behaviour. Nevertheless, it is important to verify that each treatment is well-represented in each quintile.

[§]An alternative specification for identification is to set the model intercept α to zero and to estimate $K - 1$ thresholds.

2.2.3. *Mixed-effects grouped time survival models: stratified by propensity quintiles.* The treatment effectiveness analyses of ‘time until recovery’ are then conducted using mixed-effects grouped time survival models that are stratified by propensity quintile. That is, separate treatment effectiveness analyses are conducted for those least likely to get higher doses of treatment (q_1), those somewhat more likely (q_2), and so on. The mixed-effects grouped time survival model, as specified by Hedeker *et al.* [7], examines the probability of *recovery* up to, and including, time interval t for subject i , observation j : $P_{ijt} = P(t_{ij} \leq t)$. It is a proportional hazards model that uses a complementary log–log function to describe the cumulative probability of recovery as a function of treatment (as applied here)

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

where α represents the intercept term (i.e. the baseline hazard), x is a vector of covariates (i.e. dummy coded to represent the treatment groups), β is a vector of coefficients, and v_i represents a random intercept. Both the propensity and treatment effectiveness models can be analysed using MIXOR software [8].

2.2.4. *Pooling the quintile-specific results: assumption.* In order to draw one unified conclusion about treatment effectiveness, the results of the quintile-specific analyses are pooled. To do so, however, the data must meet the assumption of no treatment by propensity interaction. This assumption is tested in a set of analyses that includes observations from all quintiles. Initially, a mixed-effects grouped time survival analysis examines the main effects of treatment (three vectors for four doses) and propensity quintile (four vectors). The incremental contribution of the propensity by treatment interaction is then tested by comparing the -2 difference in log-likelihood of the models with and without the interaction terms. If the interaction is statistically significant, the assumption is violated. In view of the fact that an interaction indicates that treatment effects vary across propensity quintiles, the pooling of quintile-specific results cannot proceed. Conclusions regarding treatment effectiveness cannot be integrated if the assumption is not fulfilled, but instead must be quintile-specific.

2.2.5. *Pooling the quintile-specific results: method.* Assuming that the propensity by treatment interaction is not significant, the results of the quintile-specific models are pooled using the Mantel–Haenszel procedure as described by Fleiss [9]. This approach weights each quintile-specific parameter estimate by the inverse of its squared standard error and sums those weighted estimates for each treatment. Pooled standard errors are calculated in a similar fashion. Finally, the test statistic, which is the ratio of each pooled parameter estimate and the corresponding standard error, and the p -value are calculated.

3. APPLICATION

The two-stage model was recently applied to the study of somatic antidepressant effectiveness in a 20-year longitudinal, observational study, The National Institute of Mental Health Collaborative Depression Study [10]. The analyses included 3141 observations of 285 subjects who met criteria for major depressive disorder at intake into the Collaborative Depression Study. The propensity model shows that those with more severe depressive symptoms tended

Table I. Mixed-effects ordinal logistic regression analysis of propensity for somatic antidepressant treatment intensity^{*,†}.

Variable	Odds ratio	(95 per cent CI)
Number prior affective episodes		
1	1.00	
2	1.08	(0.88–1.33)
3+	1.39	(1.15–1.69)
Symptom severity [‡]	1.24	(1.20–1.29)
Trajectory of symptom severity [‡]		
Stable	1.00	
Increasing	1.62	(1.36–1.94)
Decreasing	1.11	(0.86–1.43)
Treatment in prior episode		
None	1.00	
Lower dose	1.53	(1.09–2.16)
Moderate dose	1.68	(1.28–2.20)
Higher dose	1.99	(1.55–2.57)
Treatment in well interval		
None	1.00	
Lower dose	1.43	(1.11–1.85)
Moderate dose	2.84	(2.22–3.62)
Higher dose	5.06	(3.92–6.55)

Notes:

*From Leon *et al.* [10]. Reproduced with permission of the American Psychiatric Association.

†Data are based on 3141 treatment intervals (i.e., observations) from 285 subjects.

‡In the 8 weeks prior to commencing treatment (severity is ranked from 1 (lowest) to 6 (highest)).

to receive higher doses of somatic antidepressant treatment than those with less severe symptoms (Table I). The intraclass correlation coefficient (ICC) for the propensity model is 0.059. This indicates that there is very little within-subject consistency among repeated doses over time. This suggests that within-subject variation in symptom severity may have played a role in the doses prescribed. The contingency table shows that each level of treatment was well-represented in each of the propensity quintiles (Table II).

Treatment effectiveness analyses of time until recovery were initially stratified by quintile of propensity for treatment intensity (mean ICC across the five quintiles: 0.26). Then, because the treatment by propensity interaction was not statistically significant ($-2LL = 5.817$, $df = 12$, $p = 0.925$), pooled estimates of the treatment effects were calculated. Higher doses of somatic antidepressant treatment were associated with nearly twice the chance of recovery (odds ratio: 1.86; 95 per cent CI: 1.27–2.72) than no somatic treatment, despite more severe presentation of depressive symptoms. In contrast, those treated with low (OR: 0.86; 95 per cent CI: 0.55–1.23; $Z = -0.93$; $p = 0.347$) or moderate doses (OR: 1.13; 95 per cent CI: 0.79–1.63; $Z = 0.67$; $p = 0.504$) were no more likely to recover than those who did not receive somatic treatment.

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

Treatment	Propensity quintile					Subtotal
	$Q1$	$Q2$	$Q3$	$Q4$	$Q5$	
None	457	172	118	95	104	946
Lower dose	82	198	141	112	105	638
Moderate dose	60	162	269	195	194	880
Higher dose	31	83	105	236	222	677
Subtotal	630	615	633	638	625	3141

Notes: From Leon *et al.* [10]. Reproduced with permission of the American Psychiatric Association. Cell entries represent frequencies of observations. $Q1$ represents the group least likely to receive intensive treatment and $Q5$ represents the group most likely to receive intensive treatment. Data are based on 3141 treatment intervals (i.e. observations) from 285 subjects.

4. SIMULATION STUDY

A Monte Carlo simulation study was conducted to evaluate the performance of this approach. The study varied the magnitude of the treatment effect, the sample size and other aspects of the simulated data to identify conditions in which the approach is and is not useful.

4.1. Simulation specifications

The data for the Monte Carlo simulation study were generated in the following manner. First, a propensity score was calculated for each of the eight observations (i.e. points in time) per subject from a logistic model. Two randomly generated predictor variables were included in the propensity score, one a time-varying continuous variable (based on a uniform distribution) and the other a time invariant dichotomous variable with a 50:50 split of zeros and ones. The odds ratio for each these two propensity model predictors varied (1.0, 1.5, 2.0) and the intercept was set to 1.0. An ordinal dose (0, 1, 2, 3) for each observation was then calculated based on the continuous propensity score and specified threshold values (to categorize the continuous propensity scores). Next, the latent survival times (i.e. time until recovery) were generated under a proportional hazards model based on treatment effectiveness odds ratios that were specified for doses 1–3 (each relative to dose 0). This treatment effectiveness model thus indicated the survival time as a function of the dose for each subject at each time point. The effects of the confounding variables were thus mediated through dose. The continuous latent survival times were then grouped into deciles in order to represent 10 grouped-time survival times. In this, the effectiveness model, the censoring rate was fixed at 25 per cent. One thousand data sets were generated, each with eight within-subject observations, for each of 243 combinations of the following data specifications: N (100, 250, 500), correlation (0.10, 0.40, 0.70) among propensity model variables, odds ratios (1.0, 1.5, 2.0) for the propensity model, odds ratios (1.0, 1.5, 2.0) for the effectiveness model, intraclass correlations (ICC: 0.20, 0.40, 0.60) to represent the within-subject correlation of both the repeated propensity scores and survival times.

For each data set, eight treatment effectiveness models were analysed. There was one model for each of the five propensity quintiles and one model in which the quintile-specific results

were pooled using the Mantel–Haenszel procedure. The final two models, a main effects only model and a main effect and propensity by treatment interaction model, included observations from all five quintiles and tested the assumption of no treatment by propensity interaction. An augmented version of MIXOR software [8] was used for the simulation.

4.2. Evaluation of model performance

The following information was used to evaluate the performance of the data analytic strategy: type I error, statistical power, the proportion of models in which there was a significant propensity by treatment interaction, and the proportion of models in which convergence was not achieved. In addition, for each combination of data specifications, the mean (among the 1000 simulated data sets) was calculated for the parameter estimate and standard error for each of three doses. Finally, the bias was calculated as the ratio of the difference between the mean parameter estimate and the specified treatment effect to the mean of the standard error. The treatment effect odds ratios for each of the three doses were pre-specified to be identical within each combination of simulation specifications. Hence, type I error, power, and bias are presented below as the mean across the three doses.

4.3. Simulation results

4.3.1. Bias. The results of the simulation study are presented in Tables III–V for $N = 500$, 250, 100, respectively. The estimated values of the treatment effects were very close to the specified values. The bias ranged from about 3–12 per cent of a standard error.

4.3.2. Type I error and statistical power. Type I error was well within the nominal rate of 0.05. There was substantial statistical power (i.e. >0.90) for $N = 500$ and 250 for each condition with odds ratios of both 1.5 and 2.0. In contrast, for $N = 100$ and an odds ratio of 2.0, the statistical power was in excess of 0.80 with $ICC = 0.20$ or $ICC = 0.40$, but at best marginal (i.e. >0.70) for $ICC = 0.60$.

4.3.3. Feasibility. The model assumes that there is no treatment by propensity interaction. This assumption was violated at a rate of approximately nominal type I error. Furthermore, the model must converge for it to be useful. For $N = 100$ there were problems with convergence and the problem was greater with an ICC of 0.60, which is a level that might be seen in a very short clinical trial. More specifically, with $N = 100$ non-convergence occurred in about 22 per cent of the models with $ICC = 0.20$ and 0.40, and over 40 per cent of the models with $ICC = 0.60$. Although not shown in the table, the source of the non-convergence stemmed from the quintile-specific model for either the lowest or highest quintile.

5. CONCLUDING REMARKS

A mixed-effects quintile-stratified propensity adjustment has been described for longitudinal, observational evaluation of ordinal doses of treatments. The performance of the model was evaluated in a simulation study and the model was applied in a longitudinal, observational study of antidepressant effectiveness.

Table III. Monte Carlo simulation results of the mixed-effects propensity adjustment for effectiveness analyses of ordinal treatment groups ($N = 500$)*.

	Treatment odds ratio	Type I error	Statistical power	Bias [†]	Significant interaction	Non-converge	
	1.0	0.039		0.026	0.055	0.004	
	1.5		1.000	0.054	0.057	0.004	
	2.0		1.000	0.104	0.053	0.004	
<i>Treatment odds ratio</i>							
1.0	Propensity odds ratio	1	0.039	0.019	0.054	0.004	
		1.5	0.043	0.028	0.049	0.004	
		2	0.038	0.024	0.056	0.004	
	Correlation: propensity model variables	0.1	0.041	0.024	0.051	0.004	
		0.4	0.038	0.027	0.055	0.004	
		0.7	0.039	0.027	0.057	0.004	
	ICC	0.2	0.038	0.026	0.054	0.006	
		0.4	0.037	0.027	0.048	0.005	
		0.6	0.047	0.023	0.058	0.003	
	1.5	Propensity odds ratio	1	1.000	0.057	0.059	0.005
			1.5	1.000	0.054	0.055	0.004
			2	1.000	0.055	0.057	0.004
Correlation: propensity model variables		0.1	1.000	0.055	0.055	0.005	
		0.4	1.000	0.060	0.056	0.004	
		0.7	1.000	0.053	0.062	0.004	
ICC		0.2	1.000	0.057	0.054	0.004	
		0.4	1.000	0.060	0.054	0.005	
		0.6	1.000	0.043	0.069	0.005	
2.0		Propensity odds ratio	1	1.000	0.104	0.051	0.004
			1.5	1.000	0.081	0.055	0.005
			2	1.000	0.117	0.058	0.005
	Correlation: propensity model variables	0.1	1.000	0.094	0.053	0.004	
		0.4	1.000	0.088	0.050	0.003	
		0.7	1.000	0.117	0.058	0.006	
	ICC	0.2	1.000	0.108	0.051	0.005	
		0.4	1.000	0.080	0.050	0.005	
		0.6	1.000	0.117	0.062	0.003	

*Cell entries are medians of proportions from 4 factor simulation study in which 1000 runs for each of 81 combinations of parameter specifications for $N = 500$.

[†]Bias is defined as the ratio of the difference between the mean parameter estimate across runs and the specified treatment effect to the mean of the standard error.

In the simulation study, the type I error rate was acceptable for all three sample sizes. Statistical power exceeded 0.90 for $N = 250$ and 500 for odds ratio of both 1.5 and 2.0 and was acceptable for $N = 100$, but only for odds ratio of 2.0. Bias in the treatment effectiveness estimates was small, typically less than 10 per cent of the standard error.

Table IV. Monte Carlo simulation results of the mixed-effects propensity adjustment for effectiveness analyses of ordinal treatment groups ($N = 250$)*.

	Treatment odds ratio	Type I error	Statistical power	Bias [†]	Significant interaction	Non-converge	
	1.0	0.030		0.032	0.060	0.007	
	1.5		0.963	0.047	0.063	0.009	
	2.0		1.000	0.094	0.056	0.008	
<i>Treatment odds ratio</i>							
1.0	Propensity odds ratio	1	0.030		0.016	0.060	0.007
		1.5	0.033		0.036	0.062	0.008
		2	0.028		0.033	0.062	0.006
	Correlation: propensity model variables	0.1	0.030		0.030	0.060	0.006
		0.4	0.027		0.032	0.059	0.007
		0.7	0.030		0.034	0.061	0.007
	ICC	0.2	0.027		0.026	0.060	0.007
		0.4	0.025		0.022	0.056	0.006
		0.6	0.038		0.039	0.067	0.010
1.5	Propensity odds ratio	1		0.980	0.046	0.064	0.010
		1.5		0.967	0.053	0.061	0.006
		2		0.957	0.054	0.060	0.008
	Correlation: propensity model variables	0.1		0.972	0.046	0.058	0.007
		0.4		0.959	0.059	0.062	0.009
		0.7		0.950	0.046	0.063	0.009
	ICC	0.2		0.983	0.054	0.058	0.009
		0.4		0.967	0.047	0.060	0.008
		0.6		0.917	0.046	0.070	0.009
2.0	Propensity odds ratio	1		1.000	0.096	0.065	0.008
		1.5		1.000	0.097	0.050	0.007
		2		1.000	0.091	0.056	0.011
	Correlation: propensity model variables	0.1		1.000	0.097	0.061	0.009
		0.4		1.000	0.092	0.055	0.008
		0.7		1.000	0.103	0.062	0.009
	ICC	0.2		1.000	0.098	0.052	0.009
		0.4		1.000	0.091	0.055	0.007
		0.6		1.000	0.092	0.066	0.011

*Cell entries are medians of proportions from a 4 factor simulation study in which 1000 runs for each of 81 combinations of parameter specifications for $N = 250$.

[†]Bias is defined as the ratio of the difference between the mean parameter estimate across runs and the specified treatment effect to the mean of the standard error.

Table V. Monte Carlo simulation results of the mixed-effects propensity adjustment for effectiveness analyses of ordinal treatment groups ($N = 100$)*.

	Treatment odds ratio	Type I error	Statistical power	Bias [†]	Significant interaction	Non-converge	
	1.0	0.006		0.029	0.047	0.241	
	1.5		0.235	0.029	0.047	0.245	
	2.0		0.793	0.054	0.046	0.236	
<i>Treatment odds ratio</i>							
1.0	Propensity odds ratio	1	0.006		0.034	0.048	0.241
		1.5	0.004		0.019	0.045	0.243
		2	0.005		0.034	0.049	0.235
	Correlation: propensity model variables	0.1	0.004		0.021	0.045	0.243
		0.4	0.006		0.026	0.047	0.235
		0.7	0.006		0.034	0.048	0.241
	ICC	0.2	0.005		0.022	0.047	0.230
		0.4	0.004		0.026	0.044	0.228
		0.6	0.009		0.039	0.051	0.426
1.5	Propensity odds ratio	1		0.271	0.029	0.047	0.246
		1.5		0.246	0.029	0.046	0.243
		2		0.218	0.043	0.047	0.244
	Correlation: propensity model variables	0.1		0.260	0.029	0.050	0.239
		0.4		0.237	0.044	0.043	0.246
		0.7		0.223	0.036	0.044	0.253
	ICC	0.2		0.325	0.021	0.046	0.233
		0.4		0.235	0.025	0.043	0.243
		0.6		0.200	0.047	0.050	0.426
2.0	Propensity odds ratio	1		0.858	0.065	0.046	0.229
		1.5		0.810	0.054	0.045	0.249
		2		0.739	0.049	0.048	0.236
	Correlation: propensity model variables	0.1		0.806	0.058	0.046	0.218
		0.4		0.803	0.054	0.043	0.250
		0.7		0.755	0.048	0.046	0.236
	ICC	0.2		0.896	0.066	0.044	0.218
		0.4		0.810	0.035	0.042	0.227
		0.6		0.702	0.052	0.053	0.413

*Cell entries are medians of proportions from 4 factor simulation study in which 1000 runs for each of 81 combinations of parameter specifications for $N = 100$.

[†]Bias is defined as the ratio of the difference between the mean parameter estimate across runs and the specified treatment effect to the mean of the standard error.

The convergence problems with sample size of 100 limits the usefulness of the procedure with smaller N s. This could result too few observations in an ordinal dose category. A data analysis protocol could specify that this approach be attempted initially. If convergence problems were found to be an obstacle, the propensity score might be used for either matching or as a covariate. However, those approaches were not evaluated in this study.

The method that was evaluated here differs from the mixed-effects approach that we proposed in our earlier manuscript [4] in which the propensity quintiles were incorporated as covariates in the treatment effectiveness analyses. Limitations of using the propensity score for covariate adjustment have been described by Rosenbaum and Rubin [2]. In contrast in the current presentation, the treatment effectiveness analyses are conducted separately for each propensity quintile and then pooled. In both cases, the assumption of no propensity by treatment interaction is tested. The stratified approach used here is more clearly an extension of Cochran's method of removing effects of confounding variables through subclassification [1].

In conclusion, this approach is a useful strategy for longitudinal, observational studies of treatment effectiveness. It is capable of adjusting for selection bias, incorporating multiple observations per subject, and comparing effectiveness of ordinal doses.

ACKNOWLEDGEMENTS

This research was supported, in part, by Grant MH60447 awarded to Dr Leon.

REFERENCES

1. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 1968; **24**:295–313.
2. Rosenbaum P, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**:41–55.
3. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 1984; **79**:516–524.
4. 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.
5. Hedeker D, Gibbons RD. A random-effects ordinal regression model for multilevel analysis. *Biometrics* 1994; **50**:933–944.
6. McCullagh P. Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society B* 1980; **42**:109–142.
7. Hedeker D, Siddiqui O, Hu FB. Random-effects regression analysis of correlated grouped-time survival data. *Statistical Methods in Medical Research* 2000; **9**:161–179.
8. Hedeker D, Gibbons RD. MIXOR: a computer program for mixed-effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine* 1996; **49**:157–176.
9. Fleiss JL. *Statistical Methods for Rates and Proportions*. Wiley: New York, 1981.
10. 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.