STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:2197–2212
Published online 10 May 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2095

On the use of discrete choice models for causal inference

Rusty Tchernis[1,*,†], Marcela Horvitz-Lennon[2,3] and Sharon-Lise T. Normand[2,4]

[1] *Department of Economics, Indiana University, Bloomington, IN, U.S.A.*
[2] *Department of Health Care Policy, Harvard Medical School, Boston, MA, U.S.A.*
[3] *Department of Psychiatry, Harvard Medical School, Boston, MA, U.S.A.*
[4] *Department of Biostatistics, Harvard School of Public Health, Boston, MA, U.S.A.*

SUMMARY

Methodology for causal inference based on propensity scores has been developed and popularized in the last two decades. However, the majority of the methodology has concentrated on binary treatments. Only recently have these methods been extended to settings with multi-valued treatments. We propose a number of discrete choice models for estimating the propensity scores. The models differ in terms of flexibility with respect to potential correlation between treatments, and, in turn, the accuracy of the estimated propensity scores. We present the effects of discrete choice models used on performance of the causal estimators through a Monte Carlo study. We also illustrate the use of discrete choice models to estimate the effect of antipsychotic drug use on the risk of diabetes in a cohort of adults with schizophrenia. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: causal inference; discrete choice models; matching estimator

1. INTRODUCTION

The majority of clinical assessments of treatment effects rely on the results of randomized trials. Randomization allows researchers to create comparable groups of patients, such that, on average, differences in outcomes between two groups of patients are attributed solely to treatment. Unfortunately this is not the case with observational studies. Often a set of confounders, variables that affect both treatment assignment and outcome, is present. Methodology based on propensity scores, conditional probabilities of assignment to a particular treatment given a vector of observed covariates, has been developed and widely used for causal inference in the observational setting with binary treatments. Rosenbaum and Rubin [1] demonstrated that

*Correspondence to: Rusty Tchernis, Department of Economics, Indiana University, Wylie Hall Room 105, 100 S.Woodlawn, Bloomington, IN 47405, U.S.A.
†E-mail: rtcherni@indiana.edu

Contract/grant sponsor: National Institute of Mental Health; contract/grant number: RO1-MH61434
Contract/grant sponsor: Agency for Healthcare Research and Quality; contract/grant number: PO1-HS10803

Received May 2004
Accepted September 2004

Table I. Utilization of antipsychotic medications and 12-month incidence of diabetes.

	Diabetes			
	Episodes		Risk	
Treatment group	*N*	per cent	*N*	per cent
Therapeutic class: *conventionals*				
1. Low potency	608	9.01	25	4.11
2. Medium potency	937	13.88	39	4.16
3. High potency	1822	26.99	47	2.58
Therapeutic class: *atypicals*				
4. Dibenzapines	1819	26.94	64	3.52
5. Non-dibenzapines	1565	23.18	50	3.19
Total	6751	100.00	225	3.33

Adult Medicaid beneficiaries with schizophrenia.

balancing on propensity scores removes bias due to all observed confounders. Only recently have propensity score methods been extended to multiple treatment settings [2–6]. While these extensions are promising, the choice of analytic model for the treatment assignment mechanism as well as its subsequent impact on the causal estimator remains unclear. In particular, unlike in the case of binary treatments, it might be important to account for potential similarity between sets of available treatments.

Pharmacologic treatment of schizophrenia is a case in point. Table I describes utilization of antipsychotic medications in a cohort of adults with schizophrenia and the corresponding 12-month incidences of diabetes. The medications can be combined into five groups based on both their effects on clusters of schizophrenic symptoms and their adverse effects. The five groups span two therapeutic classes—the older conventional antipsychotics and the more recently available atypical antipsychotics. In the last decade, atypical antipsychotics have become the *de facto* treatment choice for schizophrenia. Atypicals may be superior to conventionals in terms of controlling certain manifestations of the illness, also posing a much lower risk for neurological side effects. However, studies have suggested that atypicals are associated with a higher risk of adverse outcomes such as adult onset diabetes [7, 8] increased levels of lipids (hyperlipidemia) [9, 10] and obesity [11, 12]. Diabetes risks range from 2.6 to 4.2 per cent in the schizophrenia cohort across the groups. Diabetes, like hyperlipidemia and obesity, is a significant risk factor for serious cardiovascular events such as heart attacks and strokes, is associated with blindness and kidney failure, and is the seventh leading cause of death in the U.S.

Although not all persons with schizophrenia are in treatment or take prescribed antipsychotic medications, broad clinical consensus exists that these medications are an essential component of the treatment for schizophrenia. For clinicians, the decision-making process is not whether to prescribe an antipsychotic, but rather *which* antipsychotic to prescribe. This is a discrete choice problem—how to choose from one of several distinct alternatives. Atypicals and conventionals combine more and less similar medications, and it might be important to account for the potential similarity of the medications. For example, high-potency conventional antipsychotics have similar potency as measured by the number of milligrams needed to achieve an effect, and a similar adverse effect profile which includes abnormal movements.

Dibenzapine atypical antipsychotics are similar in terms of chemical structure and adverse effect profile, which includes risk of sedation.

How can we estimate the effect of each medication on the risk of diabetes? Is it important to account for the potential similarity of treatments? In this article, we provide answers to these questions. We examine the impact of choice of probability model used for estimation of the treatment assignment mechanism on the resulting causal estimator. This is accomplished using a Monte Carlo study in which we simulate a variety of assignment mechanisms and evaluate the performance of causal estimators using different probability models for the propensity scores. We find that models that do not account for potential similarity between treatments lead to causal estimates that perform poorly when treatments are indeed correlated.

We review the framework for causal inference with multiple treatments in Section 2 and characterize the family of discrete choice models for inferring the treatment assignment mechanism in such settings in Section 3. Section 4 describes a set of Monte Carlo experiments examining the impact of treatment similarity on causal inferences. In Section 5, we illustrate methods through examination of the effect of antipsychotic medication use on the 12-month incidence of adult onset diabetes, henceforth referred to as risk or incidence of diabetes.

2. CAUSAL INFERENCE WITH MULTI-VALUED TREATMENTS

Suppose we have a cohort of N patients, each of whom is assigned to one of J treatments $t=1,\ldots,J$. The average causal effect of treatment t relative to treatment k on outcome Y is defined as

$$\tau = E[Y_i(t) - Y_i(k)] \tag{1}$$

where $Y_i(t)$ is a potential outcome if patient i was assigned treatment t and the expectation is taken over some population. Outcomes under treatments not assigned, the *counterfactual outcomes*, are missing. If patient i is assigned to treatment t, we denote the observed outcome as Y_{it}. In the absence of randomization, the key concern is whether the potential outcomes are independent of treatment assignment. Propensity scores methodology has been used extensively in empirical research to estimate the causal effect of binary treatments from observational data when outcomes are independent of treatment assignment conditional on the propensity score (for a review, see Reference [13]). The central role of the propensity score is to reduce the bias due to all the observed covariates when comparing two treatments [1].

When there are multiple treatment options, a number of different causal questions may be of interest however. Clinicians may want to estimate the average effect of: (1) assigning all patients to treatment t versus treatment k; (2) treatment t versus treatment k for those assigned to treatment t; or (3) treatment t versus any other treatment, $\bar{t}$, for those assigned to t. To answer the first question, we must assume that all patients can potentially be assigned to each treatment, and be willing to estimate potential outcomes under all treatments for each patient. Imbens [2] presents a weighted estimator similar to Horvitz–Thompson estimator [14] to answer the first question. The second question can be answered by comparing only the data from treatments t and k. This approach drastically reduces the sample size and does not make use of all the potential information. Lechner [3] presents a matching estimator to answer the second question. We consider the last question, which allows us to use all the available data and, more importantly, does not require out-of-sample predictions of counterfactuals.

To fix ideas, assume we observe a vector of treatment and patient-specific covariates, $\mathbf{X}_i = \{X_{it}\}$ and an outcome under the treatment received Y_{it}. A $J \times 1$ treatment indicator vector, $\mathbf{D}_i$, corresponding to the assigned treatment, with components

$$D_{it} = \begin{cases} 1 & \text{if } i \text{ assigned to } t \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The average effect of treatment t compared to any other treatment, $\bar{t}$, among *treated* subjects is

$$\tau_t = E[Y_i(t) - Y_i(\bar{t}) \mid D_{it} = 1] \tag{3}$$

where $E[Y_i(t) \mid D_{it}]$ is the expected outcome under treatment t over the population of individuals assigned treatment t.

Throughout we assume treatment assignment is strongly ignorable given the pre-treatment information [15] and the causal process for each patient is independent of other patients. The latter implies there is no competition for resources or externalities [16]. The former implies that given $\mathbf{X}_i$, appropriate adjustment for pre-treatment covariates is sufficient to directly estimate τ_t,

$$\begin{aligned} \tau_t \quad &= \quad E[Y_i(t) \mid D_{it} = 1, \mathbf{X}_i] - E[Y_i(\bar{t}) \mid D_{it} = 1, \mathbf{X}_i] \\ &\overset{\text{Ignorability}}{=} E[Y_{it} \mid \mathbf{X}_i] - E[Y_{i\bar{t}} \mid \mathbf{X}_i] \\ &\overset{\text{Balancing score}}{=} E[Y_{it} \mid p(\mathbf{X}_i)] - E[Y_{i\bar{t}} \mid p(\mathbf{X}_i)] \end{aligned} \tag{4}$$

where $p(\mathbf{X}_i) = p(D_{it} = 1 \mid \mathbf{X}_i)$ is the propensity score, the probability patient i receives treatment t. Strongly ignorable treatment assignment also assumes that in an infinite population there will be subjects treated with each treatment at each value of $\mathbf{X}$.

Under these assumptions, several methods can be used to adjust for $\mathbf{X}$, including matched sampling [17]. In a matched analysis, propensity scores are estimated for each subject and then matches for those assigned to treatment t, the *treated*, among those assigned to $\bar{t}$, the *controls* are found. The estimator of the average treatment effect on the treated is computed as

$$\hat{\tau}_t = \frac{1}{N_t^{\mathrm{m}}} \sum_s (Y_{s,t} - Y_{s,\bar{t}} \mid s, p(\mathbf{X}_i)) \tag{5}$$

where $Y_{s,t}$ is the observed outcome of the treated patient in pair s and N_t^{m} is the total number of matched pairs. Inference proceeds using inference for paired data such as paired t-tests or McNemar's test. It is important to note that the resulting estimators of $\hat{\tau}_t$ and $\hat{\tau}_{t'}$ may correspond to different subpopulations, namely those assigned to t and t', respectively. The control groups are comprised of patients assigned to different treatments. Our primary goal is to examine the impact of the choice of probability model to estimate $p(\mathbf{X}_i)$ on estimates of τ_t in the presence of multi-valued treatments. The main difference from the binary treatment setting is that some of the treatments may be perceived as similar and not accounting for the potential similarity might affect the performance of the estimator in equation (5).

3. DISCRETE CHOICE MODELS FOR TREATMENT ASSIGNMENT

Discrete choice models [18] are a general family of models that characterize the relationship between covariates, X_{it}, and the probability of treatment assignment, $p(X_{it})$. Three common models include the multinomial logit, nested logit, and multinomial probit, each differing with regard to the amount of correlation between treatment choices.

3.1. Multinomial logit

The multinomial logit model specifies the probability of treatment assignment as

$$P_{it}^{\mathrm{m}} = \Pr(D_{it} = 1 \mid X_{it}) = \frac{\mathrm{e}^{X_{it}\alpha}}{\sum_{t=1}^{J} \mathrm{e}^{X_{it}\alpha}} \tag{6}$$

The interpretation of the parameter α can be seen from the log probability ratio (LPR) of two treatments,

$$\log\left[\frac{P_{it}^{\mathrm{m}}}{P_{i\bar{t}}^{\mathrm{m}}}\right] = \alpha(X_{it} - X_{i\bar{t}})$$

implying that α measures the effect of change in one unit of X_{it} on the LPR.

While the multinomial logit model can be easily estimated in a number of statistical and econometric software packages, it has an undesirable property known as *independence from irrelevant alternatives* (IIA). IIA stems from the fact that the ratio of predicted probabilities of two treatments does not depend on any information about the remaining $J-2$ treatments. The effects of IIA can be easily seen through an example.

Assume that patients and clinicians are indifferent between conventional and atypical antipsychotics, and are indifferent among types of atypicals. If two atypicals (clozapine and olanzapine) and one conventional antipsychotic (haloperidol) are available to clinicians, then the probability of antipsychotic assignment is 0.25 for clozapine, 0.25 for olanzapine, and 0.50 for haloperidol. Now assume that clozapine is taken off the market. The odds of receiving olanzapine relative to haloperidol should be 1 to 1, yet from equation (6) the multinomial logit predicts no change in the odds—2 to 1. The reason for the IIA property relates to the fact that the multinomial logit model does not permit similarity among treatments. That is, the ratio of probabilities of two treatments does not depend on information on other treatments.

3.2. Nested logit

The nested logit model relaxes the IIA property by specifying the probability of treatment assignment sequentially. First, a class or nest of drugs is chosen, and then a particular drug from the class is prescribed. The probability of drug assignment is thus

$$P_{it}^{\mathrm{n}} = \Pr(D_{it} = 1 \mid X_{it}) = P_{t|c} \times P_c = \frac{\mathrm{e}^{X_{it}\alpha/\rho_c}}{\sum_{t=1}^{J_c} \mathrm{e}^{X_{it}\alpha/\rho_c}} \times \frac{\mathrm{e}^{\rho_c I_{ic}}}{\sum_{c=1}^{C} \mathrm{e}^{\rho_c I_{ic}}} \tag{7}$$

where P_c is the probability of choosing class c, J_c is a number of treatments in class c, and $I_{ic} = \ln\{\sum_{t=1}^{J_c} \mathrm{e}^{X_{it}\alpha/\rho_c}\}$ is the *inclusive value*. The term ρ_c may be interpreted as a measure of dissimilarity between treatments within class and is restricted to the $(0,1]$ interval. If $\rho_c = 1$

the treatments are dissimilar and the model reduces to multinomial logit, and if $0<\rho_c<1$, the nested logit model is the correct model.

It is easy to show that this model is not subject to IIA because the ratio of probabilities of two treatments depends on values of covariates of *all* treatments. In equation (7), α measures the effect of one unit change in X_{it} on the LPR of two treatments of the same class.

The main disadvantage of the nested logit model is that the researcher must select *a priori* the set of treatments that are allowed to be potential substitutes or similar. In addition, because ρ_c is fixed within a class, it measures the average correlations between treatments in the class, which would combine different levels of pair-wise correlations.

3.3. Multinomial probit

The multinomial probit model explicitly specifies a correlation structure by assuming a latent variable, U_{it}, representing the utility patient i attributes to treatment t. Treatment t is chosen if it yields maximal utility, e.g. if $D_{it}=I(U_{it}=\max_j\{U_{ij}\})$. The multinomial probit assumes that $\mathbf{U}_i \sim N_J(X_i'\alpha, \Sigma)$ so that the probability of assignment to treatment 1

$$P_{i1}^{\mathrm{p}} = \Pr(D_{it}=1 \mid X_{it}) = \int_{A_{i12}} \cdots \int_{A_{i1J}} f(\eta_{i12},\ldots,\eta_{i1J})\,\mathrm{d}\eta_{i12},\ldots,\mathrm{d}\eta_{i1J} \tag{8}$$

where $A_{i12}=(-\infty,\alpha(X_{i2}-X_{i1})], \eta_{i1j}=\varepsilon_{ij}-\varepsilon_{i1},\ \varepsilon_i \sim N_J(0,\Sigma)$, and $f(\cdot)$ is a $J-1$ dimensional normal density function. The parameter Σ represents the marginal correlation matrix of utilities of all treatments, and each treatment can be correlated with all others. This model is the most difficult to estimate because it involves computing the integral in (8). In addition, further identifying restrictions must be imposed [19, 20] because only $J(J-1)/2-1$ parameters of the covariance matrix are identified. Typically the covariance matrix is restricted by setting one row and one column to a corresponding row and column of identity matrix, $\Sigma_j=I_j, \Sigma_j'=I_j'$.

The advantage of this model is that it allows for the most flexible correlation structure, estimating all pair-wise correlations subject to identifiability constraints.

4. MONTE CARLO STUDY

We conducted a Monte Carlo study to investigate the effect of the extent of similarity between treatments on the accuracy of the causal estimators. The main idea involved simulation of outcomes under a *true* model and comparison of the performance of the estimator in equation (5) based on propensity scores estimated using the discrete choice models described in Section 3. In addition to the three discrete choice models, we also estimated the propensity scores under a fourth, simpler, model—a logistic regression. The logistic regression may be viewed as a discrete choice model that collapses the J treatments into two groups: treatment t, the treatment of interest, and $\bar{t}$, all other treatments. The probability of receiving treatment t is

$$P_{it}^{\mathrm{l}} = \Pr(D_{it}=1 \mid X_{it}) = \frac{\mathrm{e}^{\tilde{X}_{it}\alpha}}{1+\mathrm{e}^{\tilde{X}_{it}\alpha}} \tag{9}$$

where $\tilde{X}_{it} = X_{it} - \bar{X}_i$, and $\bar{X}_i = 1/J \sum_t X_{it}$. The covariates are centered across treatments because, unlike under other discrete choice models, only one covariate is used to compute the probability of treatment assignment for each treatment.

Throughout, we simulated data on 1000 patients assigned to one of $J = 4$ treatments, and vary the level of confounding and degree of similarity among treatments. Outcomes were assumed to arise from either a normal distribution or a Bernoulli distribution. Treatments were assumed to have variable effects, the same effect, and no effect on the outcomes. For each true model (multinomial logit, nested logit, multinomial probit), the degree of confounding (large or small), the outcome type (normal or Bernouilli), and treatment effect (variable, same, no), we conducted 500 simulations.

4.1. Data simulation

4.1.1. Treatment assignment mechanism. We simulated covariates, X_{it} from a standard normal distribution, independently drawing J values for each individual. The parameters affecting treatment assignment included the level of confounding level, α, and treatment model-specific parameters denoted θ_{m}. For the nested logit model, $\theta_{\mathrm{m}} = \{J_C, \rho_C\}$, and $\theta_{\mathrm{m}} = \Sigma$ in the multinomial probit model.

We operationalized large and small covariates effects on treatment selection using the ratio of probabilities of two treatments with X_{it} one standard deviation apart. A large covariate effect was one in which $\alpha = \log 4$ and small with $\alpha = \log 1.2$. In the former case, this implied that increasing the covariate value by one standard deviation quadrupled the chances of receiving a different treatment whereas in the latter case, the chances of receiving a different treatment were increased by 20 per cent.

When assigning treatments under the nested logit model, we divided treatments into two therapeutic classes, with treatments 1, 2, and 3 in class 1, and treatment 4 in class 2. We simulated treatments with moderate amount of similarity, setting $\rho_1 = \rho_2 = 0.5$.

When treatment assignment is made under multinomial probit model we used

$$\Sigma = \begin{pmatrix} 1 & 0.9 & 0.5 & 0 \\ 0.9 & 1 & 0.5 & 0 \\ 0.5 & 0.5 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

which corresponds to a high correlation between treatments 1 and 2, a moderate correlation between treatments 1 and 3, and 2 and 3, while treatment 4 is uncorrelated with other treatments. For identifiability (Section 3.3) we placed restrictions on the correlation matrix for the multinomial probit model, which is consistent with the correlation structures used to simulate treatment assignment.

4.1.2. Outcomes. Continuous outcomes, Y_i^{c}, were simulated using

$$Y_i^{\mathrm{c}} \sim \mathrm{N}(\mathbf{D}_i' \boldsymbol{\beta} + X_i' \boldsymbol{\delta}, 1) \tag{10}$$

where $\mathbf{D}_i$ is the vector of indicators of assigned treatment from equation (2), $\boldsymbol{\beta}$ is a $J \times 1$ vector of treatment effects, and $\boldsymbol{\delta}$ is a $J \times 1$ vector of covariate effects. Throughout this paper $\boldsymbol{\delta} = \mathbf{1}\delta$, where $\mathbf{1}$ is a $J \times 1$ vector of ones. We simulated data under three competing assumptions. First, we assumed that all treatments affected the outcome but the size of the effect varied by treatment. In this case, we set $\boldsymbol{\beta} = [7, 1, 3, 4]$, corresponding to causal effects, τ_t, ranging between 1 and 4 standard deviations. Second, we assumed all treatments affected the outcome in exactly the same way, $\boldsymbol{\beta} = [1, 1, 1, 1]$. This implies that $\tau_t = 0$ for all treatments. Finally, we assumed that none of the treatments affected the outcomes, with $\boldsymbol{\beta} = [0, 0, 0, 0]$, again implying that $\tau_t = 0\ \forall t$. The values of δ corresponded to large ($\delta = 3$) or small ($\delta = 1$) covariate effects. Thus, $\delta = 3$ implied that for an increase of standard deviation in X_{it}, the outcome would increase by 3 standard deviations.

The expected outcome under treatment t is $E(Y_i \mid D_{it} = 1) = \beta_t$ from equation (10). The expected outcome under any other treatment, $\bar{t}$, is a weighted average of expected outcomes under treatments other than t,

$$E(Y_i \mid D_{it} \neq 1) = \sum_{j \neq t} E(Y_{ij}) \frac{p_{ij}}{\sum_{j \neq t} p_{ij}} \tag{11}$$

where p_{ij} is the *true* probability of patient i receiving treatment j, which depends on the values of X_{it}.

Binary outcomes were simulated using

$$Y_i^{\mathrm{b}} \sim \text{Bernoulli}\left(\frac{\exp(\mathbf{D}_i' \boldsymbol{\beta} + X_i' \boldsymbol{\delta})}{1 + \exp(\mathbf{D}_i' \boldsymbol{\beta} + X_i' \boldsymbol{\delta})} \right) \tag{12}$$

As in the continuous outcome case, we selected values of β to represent variable, constant, and null treatment effects. In the case of variable treatment effects, the odds of a patient event under a treatment relative to the odds under a different treatment ranged from about 4 to 1.2. We selected similar sizes to characterize large (odds ratio of 4) and small (odds ratio of 1.2) covariate effects. The covariate effects again corresponded to increasing the covariate value by one standard deviation.

The expected outcome under treatment t is equal to

$$\Pr(Y_i^{\mathrm{b}} = 1 \mid D_{it} = 1) = \frac{\exp(\beta_t D_{it} + X_i' \boldsymbol{\delta})}{1 + \exp(\beta_t D_{it} + X_i' \boldsymbol{\delta})} \tag{13}$$

and the expected outcome under $\bar{t}$ is computed using equation (11).

4.2. Estimation and evaluation

All discrete choice models were estimated in SAS, version 8.2 (SAS Institute Inc., Cary, NC) using PROC MDC. The correlation structures for the multinomial probit and the nesting structure for the nested logit were set to the same structures used to simulate the data. We also relaxed this assumption in a set of simulations.

4.2.1. Estimation of causal effect. The causal effect, $\hat{\tau}_t$, is estimated using the matching estimator described in equation (5). The structure of the matches were 1 to 1 and only considered matches if close, e.g. using a propensity caliper of 0.60 [17]. For each treatment, t, the treated subjects, $D_{it} = 1$, were randomly ordered. For the first selected treated subject we searched for the nearest available non-treated subject, $D_{it} \neq 1$, with estimated propensity score within the caliper. The matched pair was then removed from the sample. If no match was found for the treated subject, the subject was removed from the sample and not analysed further. We then selected the next treated subject and repeated the process. Differences in the observed outcomes were computed using the paired differences and inferences using paired t-tests.

4.2.2. Evaluation. We assessed the performance of the estimators by the absolute bias (multiplied by 100), mean squared error (MSE) relative to the true model, and per cent coverage of 95 per cent confidence intervals (CIs). We also tracked the width of the CIs. Bias is computed as $\frac{1}{500}\sum_{q=1}^{500}[\hat{\tau}_t^{(q)} - \tau_t^{*(q)}]$, where $\hat{\tau}_t^{(q)}$ is the estimated causal effect in data set q, and $\tau_t^{*(q)}$ is true treatment effect for data set q. MSE is computed as $\frac{1}{500}\sum_{q=1}^{500}[\hat{\tau}_t^{(q)} - \tau_t^{*(q)}]^2$. Per cent coverage is computed as the per cent of the 500 data sets in which the true causal effect fell within the 95 per cent CI, $\frac{1}{500}\sum_{q=1}^{500}[I(\tau_t^{*(q)} \in \mathrm{CI}_{95\,\text{per cent}}^{(q)})]$, where $\mathrm{CI}_{95\,\text{per cent}}^{(q)}$ is a 95 per cent CI.

We expected that models that accounted for similarity among treatments when treatments were correlated would have smaller absolute bias, smaller relative mean square, and nominal values of coverage of 95 per cent.

4.3. Results

4.3.1. Continuous outcomes. Tables II (variable treatment effect) and III (constant treatment effect) describe the performance summaries for normally distributed outcomes. Because of the

Table II. Continuous outcome and variable treatment effects.

	True treatment model					
	Small confounding ($\delta = 1, \alpha = \log 1.2$)			Large confounding ($\delta = 3, \alpha = \log 4$)		
Estimated treatment model	M.logit	N.logit	M.probit	M.logit	N.logit	M.probit
Absolute bias × 100						
Logistic regression	1.78	1.71	3.50	3.27	6.20	40.83
Multinomial logit	1.32	1.34	2.91	2.83	3.93	34.45
Nested logit	2.87	2.24	2.99	0.37	1.04	29.18
Multinomial probit	0.54	0.82	0.87	2.82	2.19	9.96
Relative MSE						
Logistic regression	1.03	1.09	0.84	0.94	0.96	1.34
Multinomial logit	1.00	1.15	0.91	1.00	1.12	1.42
Nested logit	1.01	1.00	0.88	0.89	1.00	1.30
Multinomial probit	0.95	1.09	1.00	0.96	0.98	1.00
Per cent coverage of 95 per cent intervals						
Logistic regression	94.40	94.00	95.80	94.80	96.40	91.20
Multinomial logit	93.40	95.00	96.00	93.80	93.80	92.40
Nested logit	93.40	94.60	95.80	94.60	95.40	92.80
Multinomial probit	95.40	94.20	94.60	95.00	95.80	95.80

Based on 500 simulated data sets, each of size 1000.

Table III. Continuous outcome and constant treatment effects.

	True treatment model					
	Small confounding ($\delta = 1, \alpha = \log 1.2$)			Large confounding ($\delta = 3, \alpha = \log 4$)		
Estimated treatment model	M.logit	N.logit	M.probit	M.logit	N.logit	M.probit
Absolute bias × 100						
Logistic regression	1.83	1.94	0.65	3.17	2.45	1.57
Multinomial logit	1.66	1.68	0.62	2.53	0.54	4.96
Nested logit	2.56	1.98	0.74	0.73	0.17	3.45
Multinomial probit	1.09	1.04	1.00	2.44	0.21	2.97
Relative MSE						
Logistic regression	1.06	1.12	0.89	0.93	0.98	0.97
Multinomial logit	1.00	1.16	0.93	1.00	1.14	1.15
Nested logit	1.00	1.00	0.86	0.89	1.00	1.09
Multinomial probit	0.99	1.02	1.00	0.95	1.01	1.00
Per cent coverage of 95 per cent intervals						
Logistic regression	94.20	94.40	95.80	95.40	96.20	96.00
Multinomial logit	95.00	93.40	95.00	94.20	93.80	93.40
Nested logit	95.00	95.20	98.00	95.20	96.00	95.20
Multinomial probit	97.00	93.80	94.80	94.80	94.80	95.40

Based on 500 simulated data sets, each of size 1000.

many treatment comparisons, table entries correspond to estimates of the effect of treatment 2 compared to any other treatment, $\hat{\tau}_2$ (equation (5)). We expected the performance characteristics of the causal estimator for treatment 2 to be more sensitive to the similarity between treatments than for other treatment effects because of the high correlation with treatment 1 under the multinomial probit model. In each table, we also compared performance characteristics when the data were simulated assuming different levels of confounding: small confounding is defined as $(\delta, \alpha) = (1, \log 1.2)$ while large confounding is $(3, \log 4)$.

When the data were simulated under the multinomial logit model, all models perform well—the degree of absolute bias is comparable across models, the relative MSE is not far from one, and coverage is close to 95 per cent. However, when treatments are correlated, such as when the true treatment assignment model is the nested logit or the multinomial probit model, and when the estimated treatment assignment model does not account for correlation, such as a logistic regression or multinomial logit model, then the estimator does not perform as well: the absolute bias and the relative MSE grow while the per cent coverage drops. For example, when the data were simulated under the multinomial probit with large levels of confounding the logistic regression produced four times more bias than the multinomial probit (40.83 versus 9.96), 34 per cent higher MSE, and covered 95 per cent CI only in 91 per cent of simulations.

When treatments are correlated, treatment models that permitted correlation performed better than models that assumed independence. When data were simulated using the nested logit model with large levels of confounding, absolute bias under the multinomial probit model for treatment assignment was smaller than the multinomial logit (2.19 versus 3.93) and than logistic regression (absolute bias = 6.20).

Lower levels of confounding resulted in less pronounced differences among the models, yet similar results hold. We note that when confounding is large, the per cent of treated cases matched to 'control' cases is reduced relative to the situation when confounding is small. When the data were simulated using the multinomial probit the per cent of treated cases matched was around 70 per cent for large levels of confounding and almost 100 per cent for small.

When the treatment effects are constant (Table III), we observed very little difference in performance of models with no clear directionality, and very little difference in the magnitude of the results for small and large confounding. These results are similar to the case of no treatment effects (data not shown).

4.3.2. Binary outcomes. When outcomes are binary (Tables IV and V), the same pattern of results were observed as with continuous variables. Table IV demonstrates that the estimators that account for correlation among treatment choices when correlation exists have better performance characteristics than estimators that do not. The results are most apparent when data are simulated using large levels of confounding and results are less pronounced when confounding is small. Similar to the continuous outcomes, we observed very little difference in performance of discrete choice models when all treatments have the same effect (Table V).

Table IV. Binary outcome and variable treatment effects.

	True treatment model					
	Small confounding ($\delta = 1, \alpha = \log 1.2$)			Large confounding ($\delta = 3, \alpha = \log 4$)		
Estimated treatment model	M.logit	N.logit	M.probit	M.logit	N.logit	M.probit
Absolute bias × 100						
Logistic regression	0.14	0.60	0.26	0.08	0.16	2.23
Multinomial logit	0.36	0.41	0.34	0.14	0.35	1.52
Nested logit	0.45	0.52	0.42	0.16	0.01	1.29
Multinomial probit	0.18	0.34	0.29	0.03	0.11	0.47
Relative MSE						
Logistic regression	0.93	1.03	1.09	0.97	1.02	1.06
Multinomial logit	1.00	0.94	1.09	1.00	1.14	1.18
Nested logit	0.98	1.00	1.08	0.91	1.00	1.05
Multinomial probit	0.88	0.97	1.00	0.87	1.00	1.00
Per cent coverage of 95 per cent intervals						
Logistic regression	94.20	94.40	95.00	93.80	95.00	95.00
Multinomial logit	93.60	94.60	94.60	93.80	92.60	94.60
Nested logit	93.20	94.00	94.40	95.40	96.40	94.00
Multinomial probit	96.60	94.40	94.80	94.60	95.60	95.20

Based on 500 simulated data sets, each of size 1000.

Table V. Binary outcome and constant treatment effects.

	True treatment model					
	Small confounding ($\delta = 1, \alpha = \log 1.2$)			Large confounding ($\delta = 3, \alpha = \log 4$)		
Estimated treatment model	M.logit	N.logit	M.probit	M.logit	N.logit	M.probit
Absolute bias $\times$ 100						
Logistic regression	0.04	0.62	0.04	0.20	0.35	0.24
Multinomial logit	0.29	0.52	0.00	0.08	0.27	0.30
Nested logit	0.28	0.63	0.13	0.07	0.16	0.08
Multinomial probit	0.06	0.41	0.18	0.03	0.18	0.20
Relative MSE						
Logistic regression	0.94	1.05	1.07	0.95	1.02	1.00
Multinomial logit	1.00	1.02	0.96	1.00	1.10	1.14
Nested logit	0.95	1.00	1.08	0.93	1.00	1.07
Multinomial probit	0.91	1.00	1.00	0.93	0.98	1.00
Per cent coverage of 95 per cent intervals						
Logistic regression	94.20	95.00	93.80	95.00	95.00	96.20
Multinomial logit	94.20	95.00	96.60	95.20	94.80	95.40
Nested logit	93.60	94.80	95.40	96.00	94.40	94.60
Multinomial probit	96.40	95.60	94.00	94.40	96.20	95.80

Based on 500 simulated data sets, each of size 1000.

4.3.3. Summary. Across the simulation studies we examined, it appears important to account for the potential correlation between treatments, particularly when the covariates have a significant effect on treatment assignment and outcomes. This observation holds even when we have a smaller sample size—in a set of simulations with $N = 500$ (not reported here), the only substantive differences in conclusions related to wider CIs for the treatment effect. It is important to bear in mind that our findings are based on a specific set of simulations and not necessarily apply for all possible situations.

A natural concern relates to the risk of mis-specifying the treatment assignment model. We also conducted a small set of simulations where treatment assignment was simulated using a multinomial probit model but estimated using a multinomial probit model with incorrect correlation structure (not shown). In this case, the multinomial probit still performed best, suggesting that there is some benefit for using a more flexible discrete choice model to estimate the treatment assignment mechanism.

5. EFFECT OF ANTIPSYCHOTICS ON RISK OF DIABETES

In 2003, the U.S. Food and Drug Administration (FDA) issued a warning on the risk of diabetes associated with the use of atypical antipsychotics. The FDA warning was largely based on evidence generated by studies that compared one or more atypical antipsychotics with any conventional antipsychotic. We sought to provide a more nuanced assessment of risk

Table VI. Characteristics for Florida Medicaid beneficiaries with schizophrenia.

Treatment group	Confounders Episodes N (per cent)	 Female N (per cent)	 Age (yr) Mean (SD)	 White N (per cent)	 Comorbid. N (per cent)	Diabetes risk N (per cent)
Conventionals						
1. Low potency	608 (9)	323 (53)	40 (11)	291 (48)	97 (16)	25 (4)
2. Medium potency	937 (14)	550 (59)	42 (11)	379 (40)	188 (20)	39 (4)
3. High potency	1822 (27)	822 (45)	41 (11)	555 (30)	268 (15)	47 (3)
Atypicals						
4. Dibenzapines	1819 (27)	924 (51)	38 (11)	777 (43)	348 (19)	64 (4)
5. Non-Dibenzapines	1565 (23)	847 (54)	39 (12)	678 (43)	311 (20)	50 (3)
Total	6751 (100)	3466 (51)	40 (11)	2680 (40)	1212 (18)	225 (3)

Comorbidity is defined as presence of chronic medical conditions with the exception of hypertension.

by investigating incidence of diabetes in antipsychotics grouped in a manner that is consistent with clinicians' notions of similar and dissimilar medications.

We examined the risk of diabetes among adult Florida Medicaid beneficiaries who were diagnosed with schizophrenia and treated with antipsychotic medications during the period July 1, 1997–June 30, 2001. Only adults between ages of 18 and 64 who were continuously enrolled for at least 21 months were included in the study. We excluded subjects with dual insurance coverage by Medicare and Medicaid as we did not have complete Medicare data. We also excluded subjects who received any diabetes diagnoses or treatments during a 6-month pre-antipsychotic medication exposure period as we were interested in assessing the effect on diabetes incidence. We required a minimum exposure to the drug group of 3 months for each patient in the sample. This implied that we excluded patients who crossed over the medication groups during the first 3 months. Diabetes incidence was defined as new diagnosis of adult-onset diabetes or initiation of anti-diabetic treatment.

Table VI summarizes the distribution for selected confounders and risk of diabetes across treatment groups. The proportion of patients assigned to each group varies between 9 (low-potency conventionals) and 27 per cent (high-potency conventionals). Half of the patients are female and about 40 per cent are white, roughly a fifth of the sample had a chronic medical condition, while roughly 3 per cent developed diabetes. The summaries in Table VI suggest that healthier males were more likely to be prescribed high-potency conventionals.

In addition to the confounders listed in Table VI, the propensity score models included the total number of days hospitalized during the minimum exposure period, basis of Medicaid eligibility, comorbid psychiatric disease (bipolar diagnosis, substance abuse), and use of medications associated with diabetes, hyperlipidemia, or obesity. All the covariates were interacted with a treatment-specific covariate—the date of medication introduction into the market. The average (for the group) introduction dates varied from the mid-1970s for low-potency conventionals to the mid-1990s for dibenzapine antipsychotics. We included this variable because the amount of time the medication was available in the market affects prescribing practices. The date of introduction variable for each medication group was computed as a weighted average of introduction dates for each treatment within the group with weights equal to the proportion of patient receiving the medication.

The propensity scores were computed using logistic regression, multinomial logit, nested logit, and multinomial probit. For the nested logit, the five groups were divided into two classes—conventionals and atypicals. Four variables were included both in the probability of choosing a class of medications, as well as the probability of treatment conditional on therapeutic class—age, length of hospitalization, white, and female. When the propensity scores were estimated using the multinomial probit model, for identifiability, the diagonal elements of Σ were fixed at 1, the correlation between high-potency conventionals and dibenzapines, and all the correlations for non-dibenzapines were fixed at 0.

5.1. *Results*

For the nested logit model, the coefficients on the inclusive value ($\hat{\rho}=0.40$ [se $=0.16$]) for conventionals and $\hat{\rho}=0.65$ [se $=0.25$]) for atypicals) indicated a moderate amount of average correlation within each class. In terms of the probit models, the estimated correlations demonstrated a strong correlation between low- and medium-potency conventional antipsychotics ($\rho_{21}=0.94$, [se $=0.02$]), no significant correlation between medium- and high-potency

Table VII. Estimated causal effects of antipsychotic use on diabetes incidence.

Model	Treatment group	Per cent difference*	McNemar χ^2
Logistic regression	Low potency	0.49	0.09
	Medium potency	0.32	0.05
	High potency	−0.55	0.81
	Dibenzapines	0.27	0.13
	Non-dibenzapines	−0.06	0.00
Multinomial logit	Low potency	1.64	2.03
	Medium potency	0.85	0.72
	High potency	−1.15	3.60
	Dibenzapines	0.27	0.14
	Non-dibenzapines	0.06	0.00
Nested logit	Low potency	0.00	0.02
	Medium potency	0.43	0.12
	High potency	−1.70	7.44
	Dibenzapines	−0.06	0.00
	Non-dibenzapines	−0.19	0.04
Multinomial probit	Low potency	1.81	2.56
	Medium potency	0.75	0.52
	High potency	−1.48	5.98
	Dibenzapines	−0.06	0.00
	Non-dibenzapines	−0.06	0.00

*Difference is $\hat{\tau}_t$, diabetes risk caused by treatment t − diabetes risk caused by any other treatment. McNemar statistic should be compared to $\chi^2_{0.95,1\,\text{df}}=3.84$. Based on $N=6751$ episodes.

conventionals ($\rho_{32} = 0.11$, [se $= 0.07$]), and moderate correlation (between 0.25 and 0.4) between other groups.

Table VII summarizes the causal estimates of the difference in the risk of diabetes caused by treatment in group t versus any other group, for patients treated with t, along with McNemar χ^2 statistics. The critical value of χ^2 with one degree of freedom and level of significance of 0.05 is 3.84. For example, the value of 0.49 associated with low-potency antipsychotics for the logistic regression model indicates that the incidence of diabetes among schizophrenics treated with low-potency antipsychotic medications is 0.49 per cent higher than similar patients treated with any other antipsychotic medication group. The results indicate no significant difference in the risk of diabetes caused by any of the treatment groups when the propensity scores were estimated on the basis of models that do not account for similarities among treatments—none of the estimates of the causal effect are significant. However, when we account for this correlation, we find that high-potency conventionals significantly reduce the risk of diabetes by 1.7 per cent assuming the nested logit model or by 1.5 per cent assuming the multinomial probit model. The discrepancy between the results of the four models is not surprising given the results in Section 4.

Our results are consistent with clinical experience (Reference [21] contains detailed results) and with previous research on antipsychotic-associated diabetes risk which has generally found no increased risk for high-potency conventionals and a greater risk for low-potency conventionals [22] and dibenzapine atypicals [7, 8].

6. DISCUSSION

In this article we investigated the importance of accounting for potential similarity between treatments for causal inference with multiple treatments. While there are a number of causal questions of interest, we concentrated on the effect of receiving a particular treatment versus any other treatment, for those assigned to this treatment—that is, which treatment to prescribe. We found that it is important to account for treatment similarity, and failure to account for similarity can lead to biased point estimates and under-coverage. Our Monte Carlo study indicated that not much is lost in terms of mean square error and bias when using the more flexible models even in the absence of correlation among treatments.

In our study we used a matching estimator, although a number of competing estimators, including weighted or stratified, are available. The doubly robust estimators proposed by Robins [23] have the attractive property of producing unbiased estimates if either the treatment or the outcome equation is specified correctly. It is possible that this robustness feature would be preserved in the multi-valued treatment setting.

The effect of model flexibility on the resulting estimates is best illustrated in our study of the effects of antipsychotic medications on risk of diabetes. When potential similarity between treatments is ignored, the causal estimates are different from those using more flexible models. We found that only when we account for the similarity between medication groups the high-potency conventionals are protective of risk of diabetes, contrary to the current belief that atypical antipsychotics increase the risk.

In summary, based on our simulations we recommend the use of more flexible discrete choice models for causal inference in multi-valued treatment settings. More research is needed to explore the performance of other causal estimators outside the binary treatment setting.

ACKNOWLEDGEMENTS

This research was funded by Grant R01-MH61434 from the National Institute of Mental Health. The first author was partially supported by Grant PO1-HS10803 from the Agency for Healthcare Research and Quality. The authors thank Larry Zaborski, Harvard Medical School, for data management, expert programming, and analyses.

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. Imbens GW. The role of the propensity score in estimating dose–response functions. *Biometrika* 2000; **87**: 706–710.
3. Lechner M. Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric Evaluation of Labour Market Policies*, Lechner M, Pfeiffer F (eds). Physica, Springer: Heidelberg, 2001; 43–58.
4. Joffe M, Rosenbaum P. Invited commentary: propensity scores. *American Journal of Epidemiology* 1999; **150**:327–333.
5. Lu B, Zanutto E, Hornik R, Rosenbaum P. Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statisitcal Association* 2001; **96**:1245–1253.
6. Yunfei P, Propert K, Rosenbaum P. Balanced risk set matching. *Journal of the American Statisitcal Association* 1968; **96**:870–882.
7. Lund BC, Perry PJ, Brooks JM, Arndt S. Clozapine use in patients with schizophrenia and the risk of diabetes, hyperlipidemia, and hypertension: a claims-based approach. *Archives of General Psychiatry* 2001; **58**: 1172–1176.
8. Sernyak MJ, Leslie DL, Alarcon RD, Losonczy MF, Rosenheck R. Association of diabetes mellitus with use of atypical neuroleptics in the treatment of schizophrenia. *American Journal of Psychiatry* 2002; **159**:561–566.
9. Sheitman BB, Bird PM, Binz W, Akinli L, Sanchez C. Olanzapine-induced elevation of plasma triglyceride levels. *American Journal of Psychiatry* 1999; **156**:1471–1472.
10. Koro CE, Fedder DO, L'Italien GJ, Weiss S, Magder LS, Kreyenbuhl J, Revicki D, Buchanan RW. An assessment of the independent effects of olanzapine and risperidone exposure on the risk of hyperlipidemia in schizophrenic patients. *Archives of General Psychiatry* 2002; **59**:1021–1026.
11. Wirshing DA, Wirshing WC, Kysar L, Berisford MA, Goldstein D, Pashdag J, Mintz J, Marder SR. Novel antipsychotics: comparison of weight gain liabilities. *Journal of Clinical Psychiatry* 1999; **60**:358–363.
12. Allison DB, Mentore JL, Heo M, Chandler LP, Cappelleri JC, Infante MC, Weiden PJ. Antipsychotic-induced weight gain: a comprehensive research synthesis. *American Journal of Psychiatry* 1999; **156**:1686–1696.
13. D'Agostino RBJ. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine* 1998; **17**:2265–2281.
14. Horvitz D, Thompson D. A generalization of sampling without replacement from a finite population. *Journal of the American Statistical Association* 1952; **47**:663–685.
15. Rosenbaum PR. The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A—General* 1984; **147**:656–666.
16. Rubin DB. Comments on randomization analysis of experimental data: the Fisher randomization test. *Journal of the American Statistical Association* 1980; **75**:591–593.
17. Gu XS, Rosenbaum PR. Comparison of multivariate matching methods: structures, distances, and algorithms. *Journal of Computational and Graphical Statistics* 1993; **2**:405–420.
18. Maddala GS. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University Press: Cambridge, 1983.
19. Geweke J, Keane M, Runkle D. Alternative computational approaches to inference in the multinomial probit model. *The Review of Economics and Statistics* 1994; **76**:609–632.
20. Weeks M. The multinomial probit model revisited: a discussion of parameter estimability, identification and specification testing. *Journal of Economic Surveys* 1997; **11**:297–320.
21. Horvitz-Lennon M, Tchernis R, Zaborski L, Normand S. Incidence of serious metabolic and nutritional adverse effects associated with antipsychotic use in a medicaid schizophrenic population. Under review, 2004.
22. Korenyi C, Lowenstein B. Chlorpromazine induced diabetes. *Diseases of the Nervous System* 1968; **29**: 827–828.
23. Robins JM. Robust estimation in sequentially ignorable missing data and causal inference models. *ASA Proceedings of the Section on Bayesian Statistical Science* 1999; 6–10.