Special Issue Paper

Statistics in Medicine

Received 12 January 2010, Accepted 12 November 2010 Published online 24 February 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4168

Generalized propensity score for estimating the average treatment effect of multiple treatments

Ping Feng,[a] Xiao-Hua Zhou,[b,c,d] Qing-Ming Zou,[e] Ming-Yu Fan[f] and Xiao-Song Li[g*†]

The propensity score method is widely used in clinical studies to estimate the effect of a treatment with two levels on patient's outcomes. However, due to the complexity of many diseases, an effective treatment often involves multiple components. For example, in the practice of Traditional Chinese Medicine (TCM), an effective treatment may include multiple components, e.g. Chinese herbs, acupuncture, and massage therapy. In clinical trials involving TCM, patients could be randomly assigned to either the treatment or control group, but they or their doctors may make different choices about which treatment component to use. As a result, treatment components are not randomly assigned. Rosenbaum and Rubin proposed the propensity score method for binary treatments, and Imbens extended their work to multiple treatments. These authors defined the generalized propensity score as the conditional probability of receiving a particular level of the treatment given the pre-treatment variables. In the present work, we adopted this approach and developed a statistical methodology based on the generalized propensity score in order to estimate treatment effects in the case of multiple treatments. Two methods were discussed and compared: propensity score regression adjustment and propensity score weighting. We used these methods to assess the relative effectiveness of individual treatments in the multiple-treatment IMPACT clinical trial. The results reveal that both methods perform well when the sample size is moderate or large. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: causal inference; generalized propensity score; treatment effect; Traditional Chinese Medicine (TCM); multiple treatment components

1. Introduction

The randomized controlled trial, theoretically, is the ideal experimental design for estimating the causal effect of a treatment versus a placebo. The randomized controlled trial is usually conducted under tightly controlled 'best-case' conditions. However, in practice, implementation of this kind of an experiment becomes difficult for many reasons, such as compliance, ethics, and patient preference. Furthermore, due to the complexity of many diseases, such as cancer, depression, asthma, chronic pain, and diabetes, the effective intervention plan usually contains multiple treatment components. For example, the treatment plan recommended by Traditional Chinese Medicine (TCM) for rehabilitation after acute stroke may include several components, e.g. herbs, acupuncture, and massage therapy.

This has led to the emergence of a new type of randomized controlled experimental design in research dealing with chronic diseases. The experimental design has the following features: the intervention

[a]*Institute of Clinical Trials, West China Hospital, Sichuan University, Sichuan, People's Republic of China*
[b]*Harbin Medical University, Harbin, People's Republic of China*
[c]*Department of Biostatistics, School of Public Health, University of Washington, Seattle, WA, U.S.A.*
[d]*Beijing International Center for Mathematical Research, Peking University, Beijing, People's Republic of China*
[e]*School of Economics and Management, Nanhua University, Hunan, People's Republic of China*
[f]*Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, U.S.A.*
[g]*West China School of Public Health, Sichuan University, Sichuan, People's Republic of China*
[*]*Correspondence to: Xiao-song Li, West China School of Public Health, Sichuan University, Sichuan, People's Republic of China.*
[†]*E-mail: lixiaosong1101@126.com*

consists of multiple components, subjects are randomly assigned either to the treatment or control groups, and subjects or their doctors make choices about which treatment component to use. This kind of a trial is also known as a pragmatic clinical trial, proposed in 1967 [1, 2]. With such a design, although it is relatively easy to assess the treatment effect of the entire intervention package versus the control, it is difficult to assess the relative effects of individual treatment components because patient assignment to different treatment components is no longer random. Measuring the relative effectiveness of individual components is important for two major reasons: one is that the institutions and the doctors who conduct the research need information about the relative contribution of particular components; the other is that the evaluation of a new treatment usually includes cost-effectiveness analysis, and information about the individual components is essential for such analysis.

The 'Improving Mood-Promoting Access to Collaborative Treatment' (IMPACT) program was a multi-center, randomized controlled trial that was designed to study the effectiveness of collaborative care versus ordinary primary care for late-life depression [3]. Dr Jürgen Unützer and his colleagues conducted the study from 1999 to 2001. Patients were randomly assigned either to an intervention group that received collaborative care or to a control group that received conventional primary care. Regardless of their assigned group, all patients still made their own choices about which treatment components to use: antidepressant medications, mental health therapy, both, or neither. IMPACT is an example of pragmatic clinical trial.

When the treatment groups are not randomly assigned, the covariate distribution that is associated with the outcome variable and the grouping variable in the comparison groups is often unbalanced. The method of assigning treatment matters when deciding how to estimate the treatment effect. If the assignment method is not taken into account, and instead conventional approaches to estimating the treatment effect are used, the analysis usually cannot draw the correct conclusions.

The methods commonly used to adjust treatment effects for covariate imbalance in non-randomized studies include matching, stratification, and standard regression analysis. The main limitation of matching and stratification is that it can be very difficult to implement if the set of covariates is large. Multiple regression analysis, although it can easily handle a large number of covariates, may have other limitations [4, 5]. Analysis of covariance (ANCOVA) is a specific application of linear regression analysis. One problem with this standard approach is that 'it imposes a linearity constraint. Although nonlinear terms can be added to the ANCOVA model, it is often difficult to know how the nonlinearity should be specified.' [4] When the specified nonlinear forms in the regression model are wrong, the results from the regression analysis may be biased. Furthermore, standard regression analysis cannot make good adjustments when there are substantial differences between the groups being compared [6].

A number of alternative methods have been proposed to complement these conventional methods, including instrumental variable (IV) analysis [7–10], propensity score analysis [11] and the combination of the two [12]. IV analysis involves, first, identifying one IV that predicts the probability that a subject will receive a particular treatment, but that has no independent effect on the outcomes except through the effect of the treatment received. A valid IV is not related to the outcome through any other pathway. This approach helps to separate the effect due to treatment from the effect due to observed and unobserved factors. The advantage of this approach is that it addresses both overt and hidden biases in estimating an average treatment effect (ATE). Some limitations of this approach are that it may be difficult to identify a valid IV among the covariates, that the IV may not be unique and that it is difficult to verify or confirm the assumptions of the method.

Another new method to adjust treatment effects for covariate imbalance in non-randomized studies is the propensity score method. The propensity score method was introduced by Rosenbaum and Rubin under the weakly ignorable treatment assignment assumption [11]. They define the propensity score as the conditional probability of being assigned to treatment given a vector X of observed covariates. The propensity score method is widely used in medical research to estimate the ATE in binary treatment scenarios. Three methods of using the propensity score are commonly applied: matching on the propensity score, stratification on the propensity score, and covariate adjustment [12–20]. A propensity score weighting method has also been proposed in some studies [21, 22].

Imbens extends Rosenbaum and Rubin's work to multiple treatments [23]. He defines the generalized propensity score as the conditional probability of receiving a particular level of the treatment given the pre-treatment variables (baseline covariates). This theoretical method has not yet been applied to making causal inferences in multi-component treatments.

Currently, very few appropriate methods are available to assess the relative effectiveness of the individual components in multi-component treatments involving non-randomized treatment assignment.

To address this gap, the current study attempts to assess the validity of the generalized propensity score methods in estimating the causal effects of multi-component treatments. We adopt Imbens' approach and consider two methods: propensity score regression and propensity score weighting. We then compare how well these methods work, and conclude with the implications for analyzing the clinical effectiveness of TCM treatments.

The paper is structured as follows. In Section 2 we describe the motivational example, the IMPACT trial and the goals of our analysis. In Section 3 we describe the two methods, one using propensity score regression and the other using propensity score weighting. In Section 4 we apply the two methods to analyze IMPACT data. In Section 5 we perform a simulation study and summarize our findings. Finally, we discuss the implications of our results for TCM in Section 6.

2. IMPACT program

Dr Jürgen Unützer and his colleagues conducted the IMPACT clinical trial from 1999 to 2001. A total of 1801 patients aged 60 years or older with major depression from 18 primary care clinics across the United States were enrolled in the study. The patients' baseline covariates were collected before the randomization, and these included age, gender, and preference for different depression treatments. Patients were randomly assigned either to an intervention group that received collaborative care from a depression care manager or to a control group that received conventional primary care. As the major outcome, depression level was assessed at baseline and at 3, 6, and 12 months. At each assessment, the severity of depressive symptoms was measured by a validated instrument (SCL-20).

With this randomized design, the effectiveness of the treatment package as a whole is easily assessed. Unützer *et al.* [3] reported that at 12 months, 45 per cent of intervention patients had a 50 per cent or greater reduction in depressive symptoms from baseline, compared with 19 per cent of conventional care participants. Intervention patients also experienced greater rates of depression treatment, more satisfaction with depression care, lower depression severity, less functional impairment, and greater quality of life than participants assigned to the conventional care group.

In contrast to the ease of assessing the effectiveness of the overall treatment package, it is more difficult to evaluate the individual contributions of antidepressant drugs and psychotherapy to these overall results. This is because the IMPACT program is an example of a new type of randomized controlled trial design that has become frequent in chronic disease research. This clinical trial design belongs to the family of pragmatic clinical research. Although patients were randomly assigned to the intervention or control groups, all patients still made their own choices about which treatment components to use: antidepressant medications, mental health therapy, both, or neither. This design is more ethical and it is closer to clinical reality. But this type of design brings a new challenge to statistics: How do we evaluate the relative effectiveness of the individual components in each treatment arm? The main difficulty of estimating the relative effectiveness of individual or combined antidepressant medications and psychotherapy treatments is selection bias, since the treatment assignment is no longer random.

There were four major groups across the intervention and control arms in the IMPACT study. Let 'Drug' denote antidepressant medications and let 'Mental' denote mental health therapy. This leads to the following group descriptions: Group 1, take nothing (None); Group 2, take antidepressant medications only (Drug); Group 3, mental health therapy only (Mental); and Group 4, take both antidepressant medications and mental health therapy (Drug+Mental).

3. Methods

Developing causal models is a relatively well-accepted method for making causal inferences in statistics [24–27]. In these models, potential outcomes are used to define causal effects in order to analyze observational or experimental studies. In randomized studies, subjects are randomly assigned to different treatments and on average, there are no systematic differences in observed or unobserved covariates between subjects in different groups. However, in non-randomized studies, differences among subjects assigned to different treatment groups are not controlled. As a result, the groups that we wish to compare may show significant differences in their observed covariates, and this may lead to biased estimates of treatment effects.

3.1. Potential outcome framework for multiple treatments

In this study, we adopt the Neyman–Rubin potential outcome framework and consider a multiple treatment trial with non-random allocation. Neyman [28, 29] proposed multiple possible outcomes for each experimental subject. Rubin [30–32] developed a framework commonly referred to as Rubin's causal model and which is widely used for causal inference in statistics, epidemiology, and economics [31, 33–35].

Suppose we have m multiple treatments, $\mathcal{T}=\{1,2,\ldots,m\}$, and let $Y_i(t)$ denote the potential response of subject i if the subject has been assigned treatment t. Moreover, let Y_{it} denote the observed response for subject i, and let $I(t)$ be the indicator of receiving treatment t:

$$I(t)=\begin{cases}1 & \text{if } T=t,\\ 0 & \text{otherwise.}\end{cases}$$

Here T is the random variable indicating the treatment a subject receives. Therefore, each subject has m potential outcomes. However, in practice, we can observe only one outcome under the assigned treatment. Let T_i denote the treatment that subject i actually received. For subject i, the observed outcome Y_i can be expressed as

$$Yi=\sum_{t=1}^{m} Y_i(t)I(T_i=t). \tag{1}$$

The missing potential outcomes are called counterfactuals [25]. The treatment effect of treatment j versus treatment k $(j\neq k)$ for subject i is defined as

$$TE=Y_i(j)-Y_i(k). \tag{2}$$

Since either $Y_i(j)$ or $Y_i(k)$ can be observed, but not both, TE cannot be identified. In other words, the effect of treatments j and k on subject i cannot be evaluated. Instead, we consider the ATE of treatment j versus treatment k $(j\neq k)$ in the population, which is defined as

$$ATE_{jk}=E\{Y_i(j)\}-E\{Y_i(k)\}. \tag{3}$$

If we can observe all the potential outcomes, the ATE can be consistently estimated by

$$\widehat{ATE}_{jk}=\frac{1}{n}\sum_{i=1}^{n} Y_i(j)-\frac{1}{n}\sum_{i=1}^{n} Y_i(k). \tag{4}$$

In reality we can use only observed data to estimate ATE. In a completely randomized experiment, since the treatment assignment is independent of the outcomes, the groups being compared can be considered as representative samples of the corresponding populations. Let $\bar{y}j$ and $\bar{y}k$ denote the sample averages of the observed outcomes for the treatment groups j and k, respectively. The ATE of treatment j versus treatment k $(j\neq k)$ can be estimated by

$$\widehat{ATE}_{jk}=\bar{y}_j-\bar{y}_k, \tag{5}$$

which is an unbiased estimator of the ATE under the ignorable treatment assignment.

However, in non-randomized studies, the treatment status is likely to be dependent on outcomes. The baseline covariates among the treatment groups may have large differences and the difference of observed sample means is a biased estimator of ATE. Both overt bias and hidden bias exist. Nevertheless, the baseline covariates can provide some useful information for estimating the treatment effect, and we should use the covariates to obtain less biased estimators for ATE. In this paper, we employ two approaches to use the covariates in order to reduce the bias of the resulting estimators. One is to use propensity score regression adjustment, and the other is to use propensity score weighting.

3.2. Propensity scores in the case of multiple treatments

Imbens extends Rosenbaum and Rubin's work to multiple treatments and defines the generalized propensity score $r(t,X)$ as the conditional probability of receiving treatment t given the pre-treatment variables X,

$$r(t,X)=pr(T=t|X=X)=E\{I(t)|X=X\}, \tag{6}$$

where T denotes the treatment received and takes on a value in a set T and $I(t)$ is the indicator of receiving treatment t.

The key assumption is that the treatment assignment T is weakly unconfounded given the observed covariates X:

$$I(t)\perp Y(t)|X.$$

Here, the weak unconfoundedness is different from the strong unconfoundedness sometimes made for the binary treatments. The strong unconfoundedness means that the treatment T is independent of all potential outcomes, $Y(1),\ldots,Y(m)$, given X, and the weak unconfoundedness means that the treatment indicator at t, $I(t)$, is independent of the potential outcome at t, $Y(t)$, given X. Imbens shows that if the treatment assignment is weakly unconfounded given the observed covariates, then the treatment assignment is weakly unconfounded given generalized propensity score $r(t,X)$:

$$I(t)\perp Y(t)|r(t,X).$$

Let $\beta(t,r)$ denote the expected outcome of a subject under treatment t given generalized propensity score $r(t,X)=r$. If the treatment assignment is weakly unconfounded given covariates X, for all $t\in$T, we have the following results:

$$\beta(t,r)=E\{Y(t)|r(t,X)=r\}=E\{Y|T=t,r(T,X)=r\}. \tag{7}$$

The expected value of the potential outcome of a subject under treatment t is

$$E\{Y(t)\}=E\{\beta(t,r(t,X))\}, \tag{8}$$

where the expectation on the right side is taken with respect to the distribution of $r(t,X)$. The generalized propensity score, $r(t,X)$, can be estimated by multinomial, nested logit model, or ordinal regression model according to the characteristics of the values of the treatment.

3.3. Estimation of the ATE

Here, we summarize the two methods for estimating *ATE*, as proposed by Imbens. Let T_i be the treatment of the ith subject received; let Y_{it} denote the observed outcome of the ith subject with $T_i=t$; let $r(t,X_i)$ be the estimated generalized propensity score of the ith subject with covariate X_i and $T_i=t$; let n be the total sample size for all groups; and n_t, the observed sample size for treatment group t.

3.3.1. Method 1: *propensity score regression adjustment.* This method involves estimating the conditional expectation $\beta(t,r)$ and the average of the potential response at each treatment level t as follows:

$$\hat{\beta}(t,r(t,X))=\hat{E}\{Y|T=t,r(t,X)\}, \tag{9}$$

$$\widehat{E}\{Y(t)\}=\frac{1}{n}\sum_{i=1}^{n}\hat{\beta}(t,r(t,X_i)). \tag{10}$$

The following are the steps for estimating *ATE* using propensity score regression adjustment:

Step 1: Estimation of the generalized propensity score $r(t,X)$.

We estimate the generalized propensity score $r(t,X)$ for each subject based on the values of the observed covariates, using all the observations:

$$r(t,X)=Pr(T=t|X).$$

We can use a multinomial logistic regression model to obtain $r(t,X)$. If we let $t=1$ be the reference category, the multinomial logistic regression model with logit link can be written as the following $m-1$ basic models, to which the interaction terms or higher order terms of X can be added:

$$\log it\{r(t,X)\}=\alpha_{t0}+\alpha_{t1}'X, \tag{11}$$

where

$$\log it\{r(t,X)\}=\ln\left(\frac{P(T=t|X)}{P(T=1|X)}\right).$$

For m treatments, there are m sets of generalized propensity scores for subject i:

$$r(1, Xi) = Pr(T=1|X_i), \tag{12}$$

$$r(2, Xi) = Pr(T=2|X_i), \tag{13}$$

$$\cdots\cdots$$

$$r(m, X_i) = Pr(T=m|X_i), \tag{14}$$

where

$$r(1, X_i)+r(2, X_i)+\cdots+r(m, X_i)=1. \tag{15}$$

Step 2: Obtaining $\hat{\beta}(t, r(t, X))$.

According to the previous explanations, we know that the estimator of $\beta(t, r(t, X))$ plays an important role in the evaluation of the *ATE*. To obtain the estimation of $\beta(t, r(t, X))$ based on the generalized propensity score, we use a generalized linear model (GLM) to model the relationship between outcomes and the generalized propensity scores. The following regression model is fit to the sample data for each observed treatment group separately:

$$E(Y_i|T_i=t, r(t, X_i))=\alpha_t+\gamma_t g(r(t, X_i)), \tag{16}$$

where

$$g(r(t, X))=\ln\left\{\frac{r(t, X)}{1-r(t, X)}\right\}.$$

After we obtain m sets of $\hat{\alpha}_t$ and $\hat{\gamma}_t$, we then use equation (16) to estimate $\beta(r, r(t, X_i))$, where $t=1, \ldots, m$, for each subject, and obtain that

$$\hat{\beta}(t, r(t, X_i))=\hat{\alpha}_t+\hat{\gamma}_t g(r(t, X_i)). \tag{17}$$

Step 3: Estimation of $E\{Y(t)\}$.

We then estimate $E\{Y(t)\}$ using equation (10).

Step 4: Estimation of *ATE*.

We estimate the *ATE* of two treatment groups j and k, $j \neq k$

$$\widehat{ATE}_{jk}=\hat{E}\{Y(j)\}-\hat{E}\{Y(k)\}. \tag{18}$$

3.3.2. Method 2: propensity score weighting. Another approach to estimate *ATE* is to use generalized propensity scores to weight observations in order to make them representative of the population of interest. This method is similar to the Horvitz–Thompson estimation [36]. The following equality shows the validation of this method:

$$E\{Y(t)\}=E\{Y(t)I(T=t)/r(T=t, X)\}. \tag{19}$$

The rationale behind weighting by the generalized propensity score is to create a sample in which covariates are balanced across all treatment groups, and then to calculate the average outcome for those treatments in subjects with $T=t$ in the sample to estimate $E\{Y(t)\}$. We normalize the weights so that they add up to one in each treatment group in expectation [37]. Thus, the weighted outcome for treatment t is given by

$$\widehat{E}\{Y(t)\}=\left[\sum_{i=1}^{n}\frac{Y_i \cdot I(T_i=t)}{r(t, X_i)}\right]\left[\sum_{i=1}^{n}\frac{I(T_i=t)}{r(t, X_i)}\right]^{-1}. \tag{20}$$

Next we outline the steps in this method.

Step 1: Estimation of the generalized propensity score $r(t, X)$.

Estimate the generalized propensity score $r(t, X)$ for each subject based on the values of the observed covariates. This step is the same as the first method in the propensity score regression adjustment method.

$$r(t, X)=Pr(T=t|X).$$

Step 2: Estimation of $E\{Y(t)\}$.

We apply equation (20) to calculate the sample average of all the weighted outcomes.
Step 3: Estimation of *ATE*.
Estimate the *ATE* of two treatment groups j and k, $j \neq k$.

$$\widehat{ATE}_{jk} = \widehat{E}\{Y(j)\} - \widehat{E}\{Y(k)\}.$$

4. Analysis of data in the IMPACT study

4.1. Data source and analysis software

IMPACT was a randomized controlled trial. The patients were randomly allocated to the intervention arm that received collaborative care or to the control arm that received conventional primary care. The major outcome depression level was assessed at baseline and at 3, 6, and 12 months using the SCL-20 instrument, in which a higher score corresponds to more severe depression. The patients' characteristics, including age, gender, and education status, were collected at baseline. Unützer and colleagues reported a total of 22 baseline characteristics of the patients in their JAMA paper [3].

We used one of the five imputed datasets and observations with missing data which were dropped from the analysis. This yielded a total of 1783 patients, with 897 in the intervention arm and 886 in the control arm. Table I shows that the selected baseline covariates were balanced between the intervention and control arms.

For simplicity, we evaluated the effect only at the first follow-up, which was three months after the baseline. There were different components in each arm; for example, there was a depression care manager in the intervention arm. Thus we looked at the treatment groups in the two arms separately in this paper. We defined the outcome variable SCL score in our study to be the improvement (reduction) in SCL-20 depression scores at three months, i.e. the baseline SCL-20 minus the SCL-20 at three months, in which a higher score corresponds to greater effect. In the intervention and control arms, patients made their own choices about which treatment component to use: antidepressant medications (Drug), mental health therapy (Mental), both (Drug+Mental), or neither (None). Our study focused on

Table I. Baseline patient characteristics for all patients in the IMPACT study [3]. The dataset is extracted from one of the imputed datasets, which come directly from the study authors. Some observations have been deleted because there are missing values.

	Intervention arm $N=897$	Control arm $N=886$	Comparisons	
Variable*	Mean(SD)	Mean(SD)	Statistics†	*P*-value
ORG	5.026(2.091)	5.019(2.098)	1.395	0.986
AGE	70.946(7.331)	71.308(7.567)	0.899	0.369
GENDER	0.357(0.479)	0.342(0.475)	0.427	0.513
WHITE	0.781(0.413)	0.760(0.428)	1.209	0.272
EDUCAT	2.637(1.054)	2.598(1.020)	3.564	0.312
MARRIED	0.443(0.497)	0.485(0.500)	3.274	0.070
NUMDIS1	3.767(1.949)	3.784(1.925)	0.402	0.688
GHLTH00	3.278(1.061)	3.322(1.083)	3.449	0.486
PREF00	1.798(0.805)	1.798(0.806)	0.307	0.959
WORK00	0.122(0.327)	0.142(0.349)	1.668	0.197
INC400	38 236(66 275)	36 259(60 427)	−0.618	0.537
TOTTH00	0.273(1.398)	0.142(0.719)	−1.123	0.262
ANTID00	0.431(0.496)	0.424(0.495)	0.091	0.763
SCL00	1.683(0.604)	1.674(0.609)	−0.562	0.574

Note: *Variable: ORG (Health care organization, 8 possible values); AGE (age in years); GENDER(0=female; 1=male); WHITE(0=ethnic minority; 1=white); EDUCAT (Education status, 4 possible values); MARRIED (Married or living with partner, 0=no; 1=yes); NUMDIS1 (Number of chronic diseases, a range of 1–10); GHLTH00 (Self-reported general health at baseline, 5 possible values); PREF00 (Preference for depression treatment at baseline, 4 possible values); WORK00 (working at baseline, 0=yes;1=no); INC400 (Total household income, in US Dollars, last year at baseline); TOTTH00 (Total number of psychotherapy sessions in the past 3 months at baseline, a range of 0–24); ANTID00 (Taking any anti-depressants at baseline, 0=no;1=yes); SCL00 (Baseline score SCL-20).
†Wilcoxon two-sample test for quantitative variables, χ^2 test for qualitative variables.

Table II. Baseline patient characteristics in the intervention arm in the IMPACT study [3]. The dataset is extracted from one of the imputed datasets, which come directly from the study authors. Some observations have been deleted because there are missing values.

	Group 1 None $N=188$	Group 2 Drug $N=312$	Group 3 Mental $N=133$	Group 4 Drug+Mental $N=264$	Comparisons	
Variable*	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)	Statistics†	*P*-value
ORG	4.739(2.155)	4.955(2.061)	4.932(2.203)	5.360(1.989)	36.618	0.019
AGE	72.144(7.613)	70.625(7.365)	71.677(7.087)	70.106(7.099)	10.104	0.018
GENDER	0.415(0.494)	0.311(0.464)	0.429(0.497)	0.333(0.472)	9.249	0.026
WHITE	0.729(0.446)	0.798(0.402)	0.752(0.434)	0.814(0.390)	5.925	0.115
EDUCAT	2.404(1.126)	2.654(1.031)	2.684(1.076)	2.758(0.995)	23.516	0.005
MARRIED	0.452(0.499)	0.458(0.499)	0.414(0.494)	0.432(0.496)	0.962	0.810
NUMDIS1	3.729(2.031)	3.824(1.847)	3.925(1.877)	3.648(2.043)	3.475	0.324
GHLTH00	3.191(1.063)	3.343(1.037)	3.331(1.013)	3.235(1.109)	6.596	0.883
PREF00	1.910(0.765)	1.744(0.836)	1.940(0.736)	1.712(0.814)	34.116	0.000
WORK00	0.096(0.295)	0.106(0.308)	0.135(0.343)	0.152(0.359)	4.358	0.225
INC400	34 681(75 332)	40 660(56 241)	34 745(78 165)	39 662(64 017)	9.754	0.021
TOTTH00	0.080(0.584)	0.173(0.728)	0.150(0.754)	0.591(2.315)	17.665	0.001
ANTID00	0.128(0.335)	0.599(0.491)	0.105(0.308)	0.614(0.488)	200.001	0.000
SCL00	1.523(0.635)	1.685(0.573)	1.687(0.629)	1.793(0.585)	19.144	0.000

Note: *Variable: ORG (Health care organization, 8 possible values); AGE (age in years); GENDER(0=female; 1=male); WHITE(0=ethnic minority; 1=white); EDUCAT (Education status, 4 possible values); MARRIED (Married or living with partner, 0=no; 1=yes); NUMDIS1 (Number of chronic diseases, a range of 1–10); GHLTH00 (Self-reported general health at baseline, 5 possible values); PREF00 (Preference for depression treatment at baseline, 4 possible values); WORK00 (working at baseline, 0=yes; 1=no); INC400 (Total household income, in US Dollars, last year at baseline); TOTTH00 (Total number of psychotherapy sessions in the past 3 months at baseline, a range of 0–24); ANTID00 (Taking any anti-depressants at baseline, 0=no; 1=yes); SCL00 (Baseline score SCL-20).

†Kruskal–Wallis test for quantitative variables, χ^2 for qualitative variables.

the two major components (Drug, Mental) and made comparisons among the four treatment groups within each arm.

Tables II and III display the selected baseline characteristics in the two arms. Approximately half of the baseline covariates were unbalanced among the four treatment groups in each arm.

All statistical analyses were conducted using SAS (Statistical Analysis System) software, version 9.1.

4.2. Estimate ATE using ANCOVA

We used 'proc GLM' procedure in SAS to analyze IMPACT data. Table IV displays the estimated ATE using ANCOVA.

4.3. Estimate ATE using propensity score regression

Step 1: Estimation of the generalized propensity score.

We developed a propensity score model to predict the probability that the patients would be given a particular treatment component. Unützer *et al.* have collected a rich set of variables that could affect the outcome and the treatment assignment. This serves as the foundation for the propensity score analysis. We used a multinomial logistic regression model with logit as the link function in order to obtain generalized propensity scores because the treatment groups were not ordered. The model selection rules we followed were those suggested by Rosenbaum and Rubin [19]. This is an iterative process. In theory, covariates that are known to be related to the treatment assignment but not to the outcome should not be included, because they may reduce the effectiveness of the estimation process. In practice, however, it is hard to identify the variables that are related to the treatment assignment but not to the outcome. Unimportant covariates will add noise to the model and inflate variance estimates; on the other hand, omitting an important covariate can result in serious bias. Rubin and Thomas [19, 20] indicate that it is better to include an unimportant covariate and lose some efficiency than increase the bias by omitting an important covariate.

When applying multinomial logistic regression, researchers typically do not include predictors whose coefficients are not significantly different from zero. However, the aim of the propensity score model

Table III. Baseline patient characteristics in the control arm of the IMPACT study [3]. The dataset is extracted from one of the imputed datasets, which come directly from the study authors. Some observations have been deleted because there are missing values.

	Group 1 None $N=410$	Group 2 Drug $N=320$	Group 3 Mental $N=53$	Group 4 Drug+Mental $N=103$	Comparisons	
Variable*	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)	Statistics†	*P*-value
ORG	4.888(2.184)	5.041(2.054)	5.075(1.741)	5.447(2.018)	60.241	0.000
AGE	72.051(7.668)	70.603(7.474)	72.075(7.673)	70.146(7.134)	9.253	0.026
GENDER	0.400(0.490)	0.284(0.452)	0.302(0.463)	0.311(0.465)	11.679	0.009
WHITE	0.756(0.430)	0.753(0.432)	0.755(0.434)	0.796(0.405)	0.860	0.835
EDUCAT	2.600(1.038)	2.509(1.017)	2.736(0.923)	2.796(0.984)	9.495	0.393
MARRIED	0.468(0.500)	0.500(0.501)	0.491(0.505)	0.505(0.502)	0.915	0.822
NUMDIS1	3.707(1.881)	3.888(1.979)	3.698(1.937)	3.816(1.934)	1.492	0.684
GHLTH00	3.361(1.075)	3.319(1.082)	3.340(1.108)	3.165(1.103)	12.103	0.438
PREF00	1.878(0.763)	1.678(0.849)	1.906(0.597)	1.796(0.890)	51.014	0.000
WORK00	0.154(0.361)	0.113(0.316)	0.264(0.445)	0.126(0.334)	9.432	0.024
INC400	36 241(62 555)	33 940(31 903)	27 491(32 069)	48 046(10 9934)	5.205	0.157
TOTTH00	0.046(0.271)	0.063(0.290)	0.472(1.395)	0.602(1.617)	81.753	0.000
ANTID00	0.122(0.328)	0.797(0.403)	0.132(0.342)	0.621(0.487)	370.170	0.000
SCL00	1.588(0.575)	1.682(0.642)	1.760(0.571)	1.951(0.573)	32.878	0.000

Note: *Variable: ORG (Healthcare organization, 8 possible values); AGE (age in years); GENDER(0=female; 1=male); WHITE(0=ethnic minority; 1=white); EDUCAT (Education status, 4 possible values); MARRIED (Married or living with partner, 0=no; 1=yes); NUMDIS1 (Number of chronic diseases, a range of 1–10); GHLTH00 (Self-reported general health at baseline, 5 possible values); PREF00 (Preference for depression treatment at baseline, 4 possible values); WORK00 (working at baseline, 0=yes;1=no); INC400 (Total household income, in US Dollars, last year at baseline); TOTTH00 (Total number of psychotherapy sessions in the past 3 months at baseline, a range of 0–24); ANTID00 (Taking any anti-depressants at baseline, 0=no;1=yes); SCL00 (Baseline score SCL-20).

†Kruskal–Wallis test for quantitative variables, χ^2 for qualitative variables.

Table IV. Adjusted average treatment effect (ATE) using ANCOVA in the IMPACT study.

	Treatment arm		Control arm	
Groups compared*	Unadjusted ATE†	Adjusted ATE	Unadjusted ATE	Adjusted ATE
Drug versus None	0.099	0.005	−0.003	−0.119
Mental versus None	0.184	0.088	0.041	−0.056
(Drug+Mental) versus None	0.105	−0.065	0.247	−0.054
Mental versus Drug	0.085	0.083	0.044	0.063

Note: *None, patients received no intervention; Drug, patients took antidepressant medications only; Mental, patients received mental health therapy only; Drug+Mental, patients received both antidepressant medications and mental health therapy.

†ATE, average treatment effect.

in the present study was to obtain the best estimate for the probability of treatment assignment, so we were not concerned with over-parameterization. The final propensity model contained the main effect of 14 variables. We also tested the effects of including the interaction terms in the model, but this did not improve the model fit statistics. The final list of variables included in the model were organization, age, gender, race, education status, marriage status, number of chronic diseases, self-reported general health at baseline, preference for depression treatment at baseline, employment status at baseline, total household income last year at baseline, total number of psychotherapy sessions during the three months before baseline, taking any anti-depressants at baseline, and baseline SCL score.

We had four sets of generalized propensity scores, $r(0, X_i)$, $r(1, X_i)$, $r(2, X_i)$, and $r(3, X_i)$) for each subject, since there were four treatment groups in each arm:

$$T = \begin{cases} 0 & \text{None}, \\ 1 & \text{Drug}, \\ 2 & \text{Mental}, \\ 3 & \text{Drug+Mental}. \end{cases}$$

Step 2: Estimation of $\beta(t, r(t, X_i))$.

Once we obtained the estimated generalized propensity scores, a GLM was fit to the sample data:

$$Y_{it} = \alpha_t + \gamma_t g(r(t, X_i)) + \varepsilon_{it},$$

where Y_{it} denoted the observed SCL score for subject i under assigned treatment t, and the error term ε_{it} was assumed to be independent and identically distributed with mean 0 and common variance σ^2.

We obtained four sets of $\hat{\alpha}$ and $\hat{\gamma}$. We then estimated $\beta(t, r(t, X_i))$ for all subjects in the sample.

$$\hat{\beta}(t, r(t, X_i)) = \hat{\alpha}_t + \hat{\gamma}_t g(r(t, X_i)).$$

Step 3: Estimation of $E\{Y(t)\}$ and *Step* 4: Estimation of *ATE*.

We followed the steps described in Section 3.3.1.

Step 5: Calculation of the 95 per cent CI for *ATE*.

Finally, we calculated the 95 per cent CI for *ATE* by bootstrap [38, 39]. We re-sampled independently 10000 times for each group and then appended the observations together to form the bootstrap sample. Each bootstrap sample had the same size as the original sample. We used propensity score regression adjustment, as described above, to estimate the *ATE* for each bootstrap sample. We calculated a 95 per cent confidence interval of the *ATE* by selecting the bootstrap estimates that lay on the 2.5th and 97.5th percentiles.

4.4. Estimate ATE using propensity score weighting

This method used the inverse of the generalized propensity score $r(t, X_i)$ to adjust for the observed outcome. The following are the steps:

Step 1: Estimation of the generalized propensity score.

We estimated the generalized propensity score $r(t, X_i)$ for each subject based on the values of the observed covariates. This step was the same as described for method 1.

Step 2: Obtain $\hat{E}\{Y(t)\}$.

We used the inverse of the generalized propensity score to weight the observations and estimate $E\{Y(t)\}$ using equation (20).

Steps 3–5 were the same as described for method 1.

4.5. IMPACT analysis results

4.5.1. Baseline comparisons and statistics for the propensity model fit. IMPACT is a randomized controlled trial. As intended, almost all the baseline covariates are balanced between the intervention and control arms (Table I). However, within each arm the treatment assignment is not random, and approximately half of the baseline covariates are unbalanced among the four treatment groups (Tables II and III).

We used standard model fit statistics to assess and develop the propensity score model. The model fit statistics, including deviance, Pearson goodness-of-fit statistics, Akaike's information criterion (AIC), Schwarts criterion (SC) and $-2\log$ L and max-rescaled analogous R^2, are presented for each arm in Table V. These diagnostics suggested that we had a reasonable model. Austin [40] proposes goodness-of-fit diagnostics for the propensity score model when estimating treatment effect using covariate adjustment in binary treatments. Research on such diagnostics in the case of multiple treatments is needed.

4.5.2. The intervention arm. The unadjusted *ATE*s and bootstrap results in the intervention arm are shown in Table VI. Recall that the higher SCL score means the greater effect. The unadjusted scores

Table V. Model fit statistics for the multinomial logistic regression models used to estimate generalized propensity scores in both arms of the IMPACT study.

Criterion	Treatment arm	Control arm
AIC	2187.954	1588.060
SC	2403.911	1803.462
$-2\log$ L	2097.954	1498.060
Max-rescaled analogous R^2	0.307	0.499

Table VI. Unadjusted and adjusted average treatment effects (ATEs) in the intervention arm of the IMPACT study.

		Propensity score regression adjustment				Propensity score weighting			
Groups compared*	Unadjusted ATE†	Adjusted ATE	SD	CI_lo‡	CI_hi§	Adjusted ATE	SD	CI_lo	CI_hi
Drug versus None	0.099	0.017	0.007	0.004	0.026	0.138	0.100	−0.062	0.333
Mental versus None	0.184	0.156	0.007	0.143	0.164	0.263	0.126	0.008	0.508
(Drug+Mental) versus None	0.105	−0.077	0.010	−0.097	−0.062	0.089	0.109	−0.133	0.298
Mental versus Drug	0.085	0.140	0.004	0.130	0.147	0.125	0.090	−0.058	0.294

Note: *None, patients received no intervention; Drug, patients took antidepressant medications only; Mental, patients received mental health therapy only; Drug+Mental, patients received both antidepressant medications and mental health therapy.

†ATE, average treatment effect.

‡CI_lo, the lower boundary of the 95 per cent confidence interval.

§CI_hi, the upper boundary of the 95 per cent confidence interval.

show that the patients in the mental health group had the highest mean SCL score, and the patients who received neither drugs nor mental health therapy had the lowest mean SCL score. The unadjusted results show: (i) that mean SCL score of the patients in the antidepressant group is higher than that in the None group; (ii) that the mean SCL score of the patients in the mental health group is higher than that in the None group; (iii) that the mean SCL score of the patients who received both antidepressants and mental health is higher than that in the None group; and (iv) that the mean SCL score of the patients in mental health group is higher than that in the antidepressant group.

The results adjusted by propensity score regression reveal: (i) that the mean SCL score of the patients in the antidepressant group is significantly higher than that in the None group; (ii) that the mean SCL score of the patients in the mental health group is significantly higher than that in the None group; (iii) that the mean SCL score of the patients who received both antidepressants and mental health is significantly lower than that in the None group; and (iv) that the mean SCL score of the patients in mental health group is significantly higher than that in the antidepressant group. The 95 per cent CI for all the *ATE*s obtained by this method did not include 0.

The results adjusted by propensity score weighting indicate similar results as the unadjusted score. However, the magnitudes of *ATE*s are different from the unadjusted score. In contrast to propensity score regression, the results adjusted by weighting showed that the 95 per cent CI for three of the *ATE*s obtained by weighting included 0.

4.5.3. The control arm. The unadjusted ATEs and bootstrap results in the control arm are displayed in Table VII. The unadjusted scores show that the patients in the group receiving antidepressants and mental health therapy had the highest mean SCL score, while the patients who received only antidepressants had the lowest mean SCL scores. The unadjusted results show: (i) that mean SCL score of the patients in the antidepressant group is lower than that in the None group; (ii) that the mean SCL score of the patients in the mental health group is higher than that in the None group, (iii) that the mean SCL score of the patients who received both antidepressants and mental health is higher than that in the None group; and (iv) that the mean SCL score of the patients in mental health group is higher than that in the antidepressant group.

The results adjusted by propensity score regression reveal: (i) that mean SCL score of the patients in the antidepressant group is significantly lower than that in the None group; (ii) that the mean SCL score of the patients in the mental health group is significantly higher than that in the None group, (iii) that the mean SCL score of the patients who received both antidepressants and mental health is significantly higher than that in the None group; and (iv) that the mean SCL score of the patients in mental health group is significantly higher than that in the antidepressant group. This method gives results similar to the unadjusted scores. However, the magnitudes of *ATE*s are different from the unadjusted value. It is worth noting that the 95 per cent CI for all the *ATE*s obtained by this method did not include 0.

The results adjusted by propensity score weighting show: (i) that mean SCL score of the patients in the antidepressant group is significantly lower than that in the None group; (ii) that the mean SCL score of the patients in the mental health group is lower than that in the None group; (iii) that the mean

Table VII. Unadjusted and adjusted average treatment effects (ATEs) in the control arm of the IMPACT study.

		Propensity score regression adjustment				Propensity score weighting			
Groups compared*	Unadjusted ATE†	Adjusted ATE	SD	CI_lo‡	CI_hi§	Adjusted ATE	SD	CI_lo	CI_hi
Drug versus None	−0.003	−0.118	0.006	−0.132	−0.108	−0.164	0.080	−0.331	−0.019
Mental versus None	0.041	0.180	0.028	0.141	0.237	−0.150	0.139	−0.456	0.098
(Drug+Mental) versus None	0.247	0.034	0.011	0.010	0.053	−0.139	0.146	−0.439	0.129
Mental versus Drug	0.044	0.298	0.027	0.262	0.353	0.014	0.131	−0.271	0.248

Note: *None, patients received no intervention; Drug, patients took antidepressant medications only; Mental, patients received mental health therapy only; Drug+Mental, patients received both antidepressant medications and mental health therapy.

†ATE, average treatment effect.

‡CI_lo, the lower boundary of the 95 per cent confidence interval.

§CI_hi, the upper boundary of the 95 per cent confidence interval.

SCL score of the patients who received both antidepressants and mental health is lower than that in the None group; and (iv) that the mean SCL score of the patients in mental health group is higher than that in the antidepressant group. The *ATE*s in (2)–(4) are not significant.

4.5.4. Comparison of the two methods. The ATEs estimated by the propensity score regression adjustment and by propensity score weighting showed significant differences in magnitude and sign from the unadjusted values. This indicates that both methods were capable of providing information about the treatment effects due to individual treatment components. The results of the propensity score regression adjustment and weighting methods agreed on some of the directions of the *ATE*s (positive or negative sign). However, the propensity score regression adjustment method produced much tighter confidence intervals than the propensity score weighting method did. Since we did not know the true ATEs in this example, we could not tell which method is better. To determine the performance of the two methods, we presented simulation study in Section 5.

5. Simulation study

5.1. Simulation setup

We conducted a simulation study to compare the performance of the two propensity score methods and ANCOVA. The simulation considered three treatment levels, $\mathscr{T}=\{1,2,3\}$. The simulation study used $s=1000$ replications, three covariates, and various sample sizes. All statistical analyses were conducted using SAS software, version 9.1.

First, we generated the three covariates X_1, X_2, and X_3. The covariates X_1 and X_2 were drawn from the normal distribution with $\mu=0$ and $\sigma^2=1$. The covariate X_3 was drawn from a Bernoulli distribution with $p=0.3$. Second, we generated potential outcomes $\mathbf{y}(t)$ for all the units in the sample with the following model:

$$\mathbf{y}_j(t)=\boldsymbol{\alpha}(t)+\boldsymbol{\beta}_1(t)X_{1i}+\boldsymbol{\beta}_2(t)X_{2i}+\boldsymbol{\beta}_3(t)X_{3i}+\varepsilon_i(t),$$

where

$$[\alpha_{(1)},\alpha_{(2)},\alpha_{(3)}]=[0,1,2],$$

$$[\beta_{1(1)},\beta_{1(2)},\beta_{1(3)}]=[1,0.5,1],$$

$$[\beta_{2(1)},\beta_{2(2)},\beta_{2(3)}]=[2,1,2],$$

$$[\beta_{3(1)},\beta_{3(2)},\beta_{3(3)}]=[3,2,1],$$

and $\varepsilon_{i(t)}$ was drawn from the normal distribution with $\mu=0$ and $\sigma^2=1$. Third, the treatment assignment T_i was generated by the multinomial distribution with parameter $(C_{1,i}, C_{2,i}, C_{3,i})$:

$$C_{1,i}=\frac{\exp(A_1 D_{1,i})}{\sum_{t=1}^{3}\exp(A_i D_{t,i})},\quad C_{2,i}=\frac{\exp(A_2 D_{2,i})}{\sum_{t=1}^{3}\exp(A_i D_t, i)},\quad C_{3,i}=\frac{\exp(A_3 D_{3,i})}{\sum_{t=1}^{3}\exp(A_i D_t, i)},$$

where

$$A_i=(1, X_{1,i}, X_{2,i}, X_{3,i}),\quad D_{1,i}=(0.3, 1, 1, 1)',\quad D_{2,i}=(0.8, 2, 2, 2)',\quad D_{3,i}=(0.1, 3, 3, 3)'.$$

Based on the above chosen models, we obtained the observed outcome for the ith subject, Y_i, as $y_i=\sum_{t=1}^{m} I_{[T_i=t]}y_i(t)$, which satisfied the following linear model:

$$\mathbf{y}_j=\sum_{t=1}^{m}\boldsymbol{\alpha}_{(t)}I_{[T_i=t]}+\sum_{t=1}^{m}I_{[T_i=t]}\boldsymbol{\beta}_{1(t)}X_{1i}+\sum_{t=1}^{m}\boldsymbol{\beta}_{2(t)}I_{[T_i=t]}X_{2i}+\sum_{t=1}^{m}I_{[T_i=t]}\boldsymbol{\beta}_{3(t)}X_{3i}+\sum_{t=1}^{m}I_{[T_i=t]}\varepsilon_{i(t)}.$$

Fourth, the algorithm of our approach was applied to each one of the 1000 generated datasets. We used the main effects of the three covariates in the propensity model for all the analyses. We estimated the ATEs using the two propensity score methods we considered. Finally, we calculated the bias and the MSE of each method:

$$\text{Bias}=\frac{1}{1000}\sum_{s=1}^{1000}(\widehat{ATE}_{jks}-ATE_{jks}),$$

$$MSE=\frac{1}{1000}\sum_{s=1}^{1000}(\widehat{ATE}_{jks}-ATE_{jks})^2,$$

where $\widehat{ATE}_{jks}$ was the estimated ATE_{jk}, based on the sth simulated dataset. Here

$$ATE_{21}=ATE_{32}=0.7\quad\text{and}\quad ATE_{31}=1.4.$$

To investigate the impact of the sample size, we used $n=100, 300, 1000, 3000$, and $10\,000$, where n is the total sample size of the three treatment groups in each replication.

5.2. *Simulation results*

The results were summarized by the estimated *ATE*, the sample variance, the bias, and the *MSE*. The estimated *ATE*s and sample variances for different sample sizes are displayed in Table VIII. The biases for different sample sizes are shown in Table IX. And the *MSE*s for different sample sizes are shown in Table X.

5.3. *Comparison of the three methods*

The results of the simulation study showed that the propensity score regression adjustment produced much smaller variance than the propensity score weighting method. However, the bias of the propensity score regression adjustment was larger than that of the propensity score weighting. The bias of ANCOVA stayed about the same as the sample size increased. The variance of ANCOVA was very small.

The bias for the propensity score weighting method was monotonically diminishing when the sample size increased. However, the bias for regression method was not monotonically decreasing when the sample size increased. We think there are two main reasons that the bias for propensity score regression method does not go away although we can see a slight trend of decreasing bias with the increasing sample size. The first reason is that we have used a nonlinear form of the propensity score to fit the observations to estimate ATE using propensity score regression adjustment whereas the true potential outcome is a linear function of the covariates. With the random generation of the simulated data, the stability of the nonlinear model is affected, which is reflected in the influence of the bias in regression adjustment. The second reason is that we consider 'local' ATE in this paper as we consider the ATE pairwise, which yields a different performance for a different pair of ATEs.

The simulation study showed that both propensity score methods perform better when the sample size is moderate or large,which is consistent with that of the other reports [41, 42]. However, the performance of ANCOVA did not change when the sample size increased.

Table VIII. The estimated average treatment effects and sample variances in simulation study.

			n				
Method	Parameter	ATEjk	100	300	1000	3000	10 000
Propensity score regression adjustment	Mean	ATE21	0.952	0.125	0.413	0.424	0.529
		ATE31	0.618	1.375	1.083	0.929	1.029
		ATE32	−0.334	1.249	0.670	0.504	0.500
	Variance	ATE21	0.318	0.102	0.024	0.010	0.003
		ATE31	0.587	0.097	0.025	0.010	0.003
		ATE32	0.572	0.055	0.014	0.006	0.002
Propensity score weighting	Mean	ATE21	1.263	1.006	0.858	0.754	0.745
		ATE31	2.333	1.923	1.634	1.487	1.473
		ATE32	1.070	0.917	0.775	0.733	0.728
	Variance	ATE21	0.816	0.631	0.488	0.311	0.101
		ATE31	1.141	0.923	0.723	0.467	0.168
		ATE32	0.614	0.441	0.263	0.204	0.067
ANCOVA	Mean	ATE21	1.135	1.120	1.123	1.127	1.127
		ATE31	1.933	1.935	1.943	1.943	1.943
		ATE32	0.798	0.816	0.820	0.816	0.816
	Variance	ATE21	0.104	0.032	0.010	0.003	0.001
		ATE31	0.162	0.052	0.016	0.005	0.001
		ATE32	0.093	0.033	0.010	0.003	0.001

Note: *ATEjk is the average treatment effect of treatment j versus treatment k.

Table IX. The bias for different sample sizes in simulation study.

		n				
Method	ATEjk	100	300	1000	3000	10 000
Propensity score regression adjustment	ATE21	0.252	−0.575	−0.287	−0.276	−0.171
	ATE31	−0.782	−0.025	−0.317	−0.471	−0.371
	ATE32	−1.034	0.549	−0.030	−0.196	−0.200
Propensity score weighting	ATE21	0.563	0.306	0.158	0.054	0.045
	ATE31	0.933	0.523	0.234	0.087	0.073
	ATE32	0.370	0.217	0.075	0.033	0.028
ANCOVA	ATE21	0.435	0.420	0.423	0.427	0.427
	ATE31	0.533	0.535	0.543	0.543	0.543
	ATE32	0.098	0.116	0.120	0.116	0.116

Note: *ATEjk is the average treatment effect of treatment j versus treatment k.

Table X. The MSE for different sample sizes in simulation study.

		n				
Method	ATEjk	100	300	1000	3000	10 000
Propensity score regression adjustment	ATE21	0.381	0.432	0.106	0.086	0.032
	ATE31	1.197	0.097	0.125	0.232	0.140
	ATE32	1.640	0.357	0.015	0.045	0.042
Propensity score weighting	ATE21	1.132	0.724	0.513	0.313	0.103
	ATE31	2.010	1.196	0.776	0.474	0.173
	ATE32	0.750	0.487	0.268	0.205	0.068
ANCOVA	ATE21	0.293	0.208	0.188	0.186	0.184
	ATE31	0.446	0.339	0.310	0.299	0.297
	ATE32	0.102	0.046	0.024	0.017	0.014

Note: *ATEjk is the average treatment effect of treatment j versus treatment k.

6. Discussion

The propensity score method has been used extensively in binary treatments in the medical, social, and economic fields ever since Rosenbaum and Rubin first proposed it in 1983. Numerous studies have been published on methods that use the propensity score, including matching on the propensity score, stratification on the propensity score, covariate regression adjustment, propensity score weighting, or some combination of all four. In 2000, Imbens proposed the generalized propensity score. Since then, however, very few researchers have applied this theoretical method to the comparison of groups receiving multiple treatments. Lu [43] and Zanutto [44] reported the use of generalized propensity score matching and subclassification methods, respectively, in an antidrug media campaign study involving multiple doses. Imai [45] described the theoretical properties of the generalized propensity function. Cattaneo reported efficient semiparametric estimation of multi-valued treatment effects [46].

In the present study, we have chosen the propensity score methodology to make causal inferences in experimental data involving multiple treatments in cases where the treatment assignment is not random. We chose this methodology because it has several advantages. First, propensity scores have a simple and useful interpretation. Second, simulation studies have shown that the propensity score methodology is robust to model misspecification [47]. Third, propensity score methodology does not need to find an IV, which can be very difficult [48]. Finally, estimated propensity scores perform better than true propensity scores for removing bias because they also remove some chance imbalances in covariates [49].

Thus, the generalized propensity score methodology can be thought of as a useful tool for researchers when they need to estimate the treatment effect in multiple-treatment interventions where the treatment group is not randomly assigned. One type of clinical context that may benefit from this improved methodology is TCM.

TCM has been used in China for more than 5000 years. In 2006, the healthcare service provided by the TCM sector accounted for approximately 10–20 per cent of the health care in China [50]. TCM has also been recognized as a complementary and alternative medicine in Western countries. Although TCM has been in use for a long time, the effectiveness of the overall treatment and/or of the individual components is usually based solely on the physician's anecdotal experience.

Systematic scientific evaluation of the effectiveness of TCM is therefore needed. However, applying Western approaches to TCM can sometimes be problematic. TCM diagnoses are based on Zheng, which is similar to the concept of a syndrome, defined by symptoms and signs. The same disease defined in Western medicine can have different Zhengs, and the Zheng guides the TCM treatment. Multi-herb drugs have frequently been used to improve treatment outcomes and reduce side effects. Different patients with the same disease may be given different combinations of multiple Chinese herbs based on the patients' characteristics. For example, Chen [51] reported that the most common combination of Chinese herbal medicines used for treating chronic hepatitis is three to six kinds of Chinese herbal drugs. Indeed, most treatments in TCM research consist of multiple components.

This makes TCM clinical studies well suited to the pragmatic randomized clinical trial design. For example, Vickers *et al.* [52] reported the study of acupuncture for chronic headache, and Thomas *et al.* [53] reported a protocol for the study of the longer term clinical and economic benefits of acupuncture for chronic low back pain. In these studies, the patients were randomized into the treatment group receiving acupuncture or into the control group receiving conventional care. However, in the treatment arm, the patients received different doses of acupuncture treatment based on the patient's individual characteristics. The generalized propensity score methodology could be useful for obtaining more accurate estimates of the effectiveness of acupuncture compared to conventional therapy in these and similar clinical studies. There are also some similar ongoing TCM trials. For example, the treatment arm in a study of the TCM Synthesis Rehabilitation on early rehabilitation in patients with ischemic stroke contains Chinese herbs, Chinese Medicine instant washing, acupuncture, and massage therapy. The choices of different components is based on the characteristics of the patients.

There are two limitations of the propensity score methodology. One limitation is that when the sample size is limited, the propensity score method does not perform well. Another limitation is that it only controls for observed covariates [49]. This is always a limitation of non-randomized studies. To take advantage of this method, researchers should design the study well and try their best to collect data on the variables that could affect the outcome and the treatment assignment.

There is one limitation in our simulation study, which prevents us from assessing the efficiency of the propensity score regression method. Owing to the difficulty of generating potential outcome data, which satisfy a linear relationship between the response and the logit of the propensity score, we

were not able to assess the bias and MSE of the propensity score regression method when this linear relationship holds. A future research study is needed to assess the relative performance of the two propensity score methods when the response has a linear relationship with the logit of the estimated propensity scores. An alternative approach is to develop the non-parametric propensity score regression method, where the relationship between the response and the logit of the estimated propensity scores is non-parametrically estimated.

Acknowledgements

We are grateful to Dr Unützer and his IMPACT study team for making their data available. We thank Linling Zou and Min Feng for organizing the tables. Financial support for this research was generously provided through the Overseas Young Scholar Collaboration Research Grant of the National Science Foundation of China (NSFC grant 30728019), and the Basic Research for Application Grant from the Department of Science and Technology of Sichuan Province, People's Republic of China (grant 2009JY0020).

References

1. Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutic trials. *Journal of Chronic Diseases* 1967; **20**(8):637–648.
2. MacPherson H. Pragmatic clinical trials. *Complementary Therapies in Medicine* 2004; **12**:136–140.
3. Unützer J, Katon W, Callahan CM, Williams JW, Hunkeler E, Harpole L, Hoffing M, Della Penna RD, Noël PH, Lin EHB, Areán PA, Hegel MT, Tang L, Belin TR, Oishi S, Langston C. Collaborative care management of late-life depression in the primary care setting: a randomized controlled trial. *JAMA* 2002; **288**(22):2836–2845.
4. Winship C, Morgan SL. The estimation of causal effects from observational data. *Annual Review of Sociology* 1999; **25**:659–706.
5. Posner MA, Ash AS, Freund KM, Moskowitz MA, Shwartz M. Comparing standard regression, propensity score matching, and instrumental variables methods for determining the influence of mammography on stage of diagnosis. *Health Services and Outcomes Research Methodology* 2001; **2**:279–290.
6. Rubin DB. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology* 2001; **2**:169–188.
7. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 1996; **91**:444–455.
8. Rassen JA, Brookhart MA, Glynn RJ. Instrumental variables i: instrumental variables exploit natural variation in nonexperimental data to estimate causal relationships. *Journal of Clinical Epidemiology* 2009; **62**:1226–1232.
9. Rassen JA, Brookhart MA, Glynn RJ. Instrumental variables II: instrumental variable application-in 25 variations, the physician prescribing preference generally was strong and reduced covariate imbalance. *Journal of Clinical Epidemiology* 2009; **62**:1233–1241.
10. Basu A, Heckman JJ, Navarro-lozano S, Urzua S. Use of instrumental variables in the presence of heterogeneity and self-selection: an application to treatments of breast cancer patients. *Health Economics* 2007; **16**:1133–1157.
11. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**:41–55.
12. Ichimura H, Taber C. Propensity-score matching with instrumental variables. *The American Economic Review* 2001; **91**(2):119–124.
13. Normand SLT, Landrum MB, Guadagnoli E, Ayanian JZ, Ryan TJ, Cleary PD, McNeil BJ. Validating recommendations for coronary angiography following an acute myocardial infarction in the elderly: a matched analysis using propensity scores. *Journal of Clinical Epidemiology* 2001; **54**:387–398.
14. Perkins SM, Tu W, Underhill MG, Zhou XH, Murray MD. The use of propensity scores in pharmacoepidemiologic research. *Pharmacoepidemiology and Drug Safety* 2000; **9**:93–101.
15. Austin PC, Mamdani MM. A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine* 2006; **25**:2084–2106.
16. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine* 1997; **127**:757–763.
17. Austin PC, Mamdani MM, Stukel TA, Anderson GM, Tu JV. The use of the propensity score for estimating treatment effects: administrative versus clinical data. *Statistics in Medicine* 2005; **24**:1563–1578.
18. D'Agostino Jr RB. 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.
19. 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.
20. Rubin DB, Thomas N. Matching using estimated propensity scores: relating theory to practice. *Biometrics* 1996; **52**:249–264.
21. Hirano K, Imbens GW. Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. *Health Services and Outcomes Research Methodology* 2001; **2**:259–278.
22. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* 2004; **23**:2937–2960.
23. Imbens GW. The role of propensity score in estimating dose-response functions. *Biometrika* 2000; **87**(3):706–710.

24. Pearl J. Causal inference in the health sciences: a conceptual introduction. *Health Services and Outcomes Research Methodology* 2001; **2**:189–220.
25. Greenland S. An overview of methods for causal inference from observational studies. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. Wiley: New York, 2004; ISBN:0-470-09043-X.
26. Heckman JJ. Causal inference and nonrandom samples. *Journal of Educational Statistics* 1989; **14**:159–168.
27. Greenland S, Pearl J, Robins J. Causal diagrams for epidemiologic research. *Epidemiology* 1999; **10**:37–48.
28. Neyman J. On the application of probability theory to agricultural experiments: essay on principles, section 9. *Translated in Statistical Science* 1990; **5**:465–480.
29. Rubin DB. Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science* 1990; **5**:472–480.
30. Rubin DB. Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference* 1990; **25**:279–292.
31. Rubin DB. Teaching statistical inference for causal effects in experiments and observational studies. *Journal of Educational and Behavioral Statistics* 2004; **29**(3):343–367.
32. Rubin DB. Causal inference using potential outcomes: design, modeling, decisions. *Journal of the American Statistical Association* 2005; **100**:322–331.
33. Holland P. Statistics and causal inference. *Journal of the American Statistical Association* 1986; **81**:945–970.
34. Hofler M. Causal inference based on counterfactuals. *BMC Medical Research Methodology* 2005; **5**:28.
35. Pratt JW, Schlaifer R. On the interpretation and observation of laws. *Journal of Econometrics* 1988; **39**:23–52.
36. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite population. *Journal of the American Statistical Association* 1952; **47**:663–685.
37. Flores CA, Mitnik OA. Evaluating nonexperimental estimators for multiple treatments: evidence from experimental data. IZA Discussion Paper No. 4451, September 2009.
38. Barker N. A practical introduction to the Bootstrap using the SAS system. Available from: http://www.lexjansen.com/phuse/2005/pk/pk02.pdf [1 March 2008].
39. Efron, Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC: London, 1993. IBSN: 0-412-04231-2.
40. Austin PC. Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score. *Pharmacoepidemiology and Drug Safety* 2008; **17**:1202–1217.
41. Zhao Z. Using matching to estimate treatment effects: data requirements, matching metrics, and Monte Carlo evidence. *The Review of Economics and Statistics* 2004; **86**(1):91–107.
42. Anstrom KJ, Tsiatis AA. Utilizing propensity scores to estimate causal treatment effects with censored time-lagged data. *Biometrics* 2001; **57**(4):1207–1218.
43. Lu B, Zanutto E, Hornik R, Rosenbaum PR. Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association* 2001; **96**(456):1245–1253.
44. Zanutto E, Lu B, Hornik R. Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *Journal of Educational and Behavioral Statistics* 2005; **30**(1):59–73.
45. Imai K, Van Dyk DA. Causal inference with general treatment regimes: generalizing the propensity score. *Journal of the American Statistical Association, Theory and Methods* 2004; **99**(467):854–866.
46. Cattaneo MD. Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics* 2010; **155**:138–154.
47. Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics* 1993; **49**:1231–1236.
48. Landrum MB, Ayanian JZ. Causal effect of ambulatory specialty care on mortality following myocardial infarction: a comparison of propensity score and instrumental variable analyses. *Health Services and Outcomes Research Methodology* 2001; **2**:221–245.
49. Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. *American Journal of Epidemiology* 1999; **150**(4): 327–333.
50. Tang JL, Liu BY, Ma KW. Traditional Chinese Medicine. *The Lancet* 2008; **372**(9654):1938–1940.
51. Chen FP, Kung YY, Chen YC, Jong MS, Chen TJ, Chen FJ, Hwang SJ. Frequency and pattern of Chinese herbal medicine prescriptions for chronic hepatitis in Taiwan. *Journal of Ethnopharmacology* 2008; **117**:84–91.
52. Vickers AJ, Rees RW, Zollman CE, McCarney R, Smith C, Ellis N, Fisher P, Van Haselen R. Acupuncture for chronic headache in primary care: large, pragmatic, randomised trial. *BMJ* 2004; **328**:744–747.
53. Thomas KJ, Fitter M, Brazier J, MacPherson H, Campbell M, Nicholl JP, Roman M. Longer term clinical and economic benefits of offering acupuncture to patients with chronic low back pain assessed as suitable for primary care management. *Complementary Therapies in Medicine* 1999; **7**(2):91–100.