Estimating Subgroup Effects Using the Propensity Score Method

A Practical Application in Outcomes Research

Hester V. Eeren, MSc,† Marieke D. Spreeuwenberg, PhD,‡ Anna Bartak, PhD,§‖ Mark de Rooij, PhD,¶ and Jan J.V. Busschbach, PhD*†*

Objective: Our aim was to demonstrate the feasibility of the univariate and generalized propensity score (PS) method in subgroup analysis of outcomes research.

Methods: First, to estimate subgroup effects, we tested the performance of 2 different PS methods, using Monte Carlo simulations: (1) the univariate PS with additional adjustment on the subgroup; and (2) the generalized PS, estimated by crossing the treatment options with a subgroup variable. The subgroup effects were estimated in a linear regression model using the 2 PS adjustments. We further explored whether the subgroup variable should be included in the univariate PS. Second, the 2 methods were compared using data from a large effectiveness study on psychotherapy in personality disorders. Using these data we tested the differences between short-term and long-term treatment, with the severity of patients' problems defining the subgroups of interest.

Results: The Monte Carlo simulations showed minor differences between both PS methods, with the bias and mean squared error overall marginally lower for the generalized PS. When considering the univariate PS, the subgroup variable can be excluded from the PS estimation and only adjusted for in the outcome equation. When applied to the psychotherapy data, the univariate and generalized PS estimations gave similar results.

Conclusion: The results support the use of the generalized PS as a feasible method, compared with the univariate PS, to find certain subgroup effects in nonrandomized outcomes research.

From the *Viersprong Institute for Studies on Personality Disorders (VISPD), Halsteren; †Department of Psychiatry, Section Medical Psychology and Psychotherapy, Erasmus MC, Rotterdam; ‡Department of Health Services Research, Maastricht University, Maastricht; §Department of Clinical Psychology, University of Amsterdam (UvA); ‖Bos en Lommer Private Practice, Amsterdam; and ¶Department of Methodology and Statistics, Institute of Psychology, Leiden University, Leiden, The Netherlands.

The authors declare no conflict of interest.

Reprints: Hester V. Eeren, MSc, Department of Psychiatry, Section Medical Psychology and Psychotherapy, Erasmus MC, P.O. Box 2040, Room Na 20-07, Rotterdam 3000 CA, The Netherlands. E-mail: h.vaneeren.1@erasmusmc.nl.

Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Website, www.lww-medicalcare.com.

ISSN: 0025-7079/15/5304-0366

Key Words: propensity score, Monte Carlo method, mental health, outcomes research

(*Med Care* 2015;53: 366–373)

In nonrandomized studies, the propensity score (further denoted as PS) method has gained popularity as a statistical method to overcome selection bias due to differences in observed pretreatment variables of patient groups[1] and the "dimensionality" problem of alternative methods such as stratification and matching.[2] The univariate PS[3] is a valid solution to compare 2 treatment categories,[4,5] whereas the generalized PS can be used if >2 treatment categories are compared.[6–8] An equal distribution on the covariates is assumed after adjustment on the PS.[8–10] Although the PS can control for overt bias due to (many) observed pretreatment variables,[10,11] hidden bias could still be present.[11]

The PS is predominantly used to estimate treatment effects.[12,13] However, it may also be important to define which treatment is specifically effective for a (sub)group of patients.[14] Treatment options can then be applied more efficiently by directly allocating a group of patients to a relevant treatment. For instance, one could argue that long-term psychotherapy is more effective than short-term psychotherapy for patients having severe problems. Because the patients are not randomly assigned to the treatment options, patients having either mild or severe problems can differ on the observed pretreatment variables. Therefore, there is a need to apply PS modeling methods when studying subgroup effects.

Several authors have described methods to estimate treatment effects for particular subgroups while using the univariate PS. Rosenbaum and Rubin already recommended subclassifying or matching on additional covariates to identify differences in treatment effect between subgroups. To reduce bias in the estimated treatment effect, Rubin and Thomas[15,16] advised that the covariate together with the PS be included when estimating the treatment effect. Such an additional covariate could define subgroups. The treatment effect can also vary according to quantiles of the PS estimations. Effect modification (eg, interaction effects) can result in different estimated treatment effects for different PS quantiles.[17–21] However, it is not possible to relate a specific

subgroup to the PS quantiles. More recently, Liem et al[22] determined the treatment effect for subgroups by adding interaction terms in a multivariable-adjusted model in which the PS was also included. Another method was described by Radice et al[23] and Kreif et al,[24] who estimated the univariate PS within each subgroup separately. Yet, only the univariate PS was used in subgroup analyses and the PS was not made multiple, as a generalized PS, by crossing the treatment options with a subgroup variable.

Because the univariate PS is mainly used in subgroup analyses, this study investigated whether a generalized PS could be used to estimate subgroup effects in outcomes research, and compared it with using a univariate PS. We first used Monte Carlo simulations to investigate whether and how the generalized or univariate PS could be used to estimate subgroup effects. These 2 PS estimations were subsequently compared using data from a Dutch research project on psychotherapy effectiveness: SCEPTRE (*S*tudy on *C*ost-*E*ffectiveness of *P*ersonality *D*isorder *TRE*atment).[25]

METHODS

First, we describe the univariate PS, the generalized PS, and the simulation study in which we tested these methods. Then, we describe the case study where we compare the 2 PS methods. To estimate the treatment effects for subgroups of patients, we used the 2 PS estimations in covariate adjustment, as this method is the most frequently used PS method in the medical literature[9,16,26,27] because it leaves the sample size intact.

Univariate PS Method

The univariate PS is defined according to Rosenbaum and Rubin[3] as:

$$\mathrm{PS}(x) = \mathrm{pr}\langle D = 1 \mid X = x\rangle, \quad (1)$$

where if $D=1$ the PS defines the conditional probability ofassignment to the treatment of interest, given a set of observed covariates (X).[3] The ignorability assumption defines that the potential outcomes and the treatment assignment are independent given the observed covariates (X).[3,10] The PS was estimated in a univariate logistic regression function.[28] To estimate the treatment effect for subgroups of patients, this PS estimation was used as an extra predictor in a linear regression model with treatment outcome (Y) as the dependent variable:

$$\text{Outcome} = \beta_0+\beta_1 PS+\beta_2 D+\beta_3 Z+\beta_4 DZ. \quad (2)$$

The treatment groups (D), subgroups (Z), and the interaction between these (DZ) were the independent variables, and the effects of interest.[22]

Generalized PS Method

The generalized PS is an extension of the univariate PS defined by Rosenbaum and Rubin[3] and is further defined by Imbens.[7] To calculate the generalized PS used in this study, we estimated the joint conditional probability of the treatment assignment (D) and subgroup (Z) given all covariates (X):

$$\mathrm{PS}(d,z,x) = \mathrm{pr}\langle D = d, Z = z \mid X = x\rangle. \quad (3)$$

Using 2 treatment options and 2 subgroups, the PS was estimated for 4 groups. The assumption on the strongly ignorable treatment assignment is crucial.[3,7,10] When this assumption was adjusted to the combined categories on which the generalized PS was estimated, the joint distributions of the potential outcomes and the covariates X should be equal between the 4 groups.[29] The generalized PS was estimated in a multinomial regression model. To estimate the treatment effects, 3 estimated generalized PSs and 3 dummy variables indicating group membership (G) were adjusted for in a regression model with treatment outcome (Y) as the dependent variable:

$$\text{Outcome} = \beta_0+\beta_1 \mathrm{PS}_1+\beta_2 \mathrm{PS}_2+\beta_3 \mathrm{PS}_3+\beta_4 G_1+\beta_5 G_2+\beta_6 G_3. \quad (4)$$

The coefficients related to the 3 dummy variables were the effects of interest.

Monte Carlo Simulation Study

A Monte Carlo simulation study was designed to test the 2 PS estimations. Therefore, we simulated 2 treatment categories, a subgroup variable with 2 categories and 3 additional variables that served as covariates. These 3 covariates were continuous variables (such as age, length, or body weight) related to (a) only the treatment assignment; (b) both treatment assignment and outcome, such that it is a true confounder[30]; or (c) outcome alone (Table 1).

The covariates were multivariate normally distributed with a mean of zero and variance of 1, except for the subgroup, that followed a Bernoulli distribution with, for example, a probability of having severe problems of 0.4 (Table 1). The outcome was simulated from a linear regression model and its error term was multivariate normally distributed, just as for the error terms of the treatment assignment (Table 1). The correlation between the error terms was set to zero as only overt bias was simulated. We simulated 2 scenarios: in scenario 1 the subgroup was not related to the treatment assignment, whereas in scenario 2, this relationship was simulated. In both the scenarios, the subgroup was related to the outcome (Table 1). Within each scenario we varied the simulated data on 3 levels of the correlation between the covariates, on the presence or absence of a correlation with the subgroup, and on 3 different sample sizes (Table 1). Under each combination of characteristics of the simulated data, 1000 datasets were created, which resulted in 18,000 datasets per scenario.

In the literature, there is no consensus on how to select the variables in PS estimation.[30,31] Therefore, we varied the inclusion of variables in the PS estimations (Table 1). For the univariate PS, we investigated whether the subgroup variable should be included or excluded in the PS estimation. The subgroup variable cannot be selected for the generalized PS, as it is part of its definition (Table 1).

To evaluate the performance of the 2 PS methods in the simulation study, we estimated the bias, mean squared error (MSE) and SE of the relevant effects over the total number of simulations. In Eq. (2), the relevant effects were the treatment effect, the effect of the subgroups, and the interaction term. In Eq. (4), these were the coefficients

TABLE 1. Variables and Characteristics of Monte Carlo Simulation*

Variables	Type	Function
X_1	Covariate	Multivariate normal distribution (0, 1)
X_2	Covariate	Multivariate normal distribution (0, 1)
X_3	Covariate	Multivariate normal distribution (0, 1)
Z	Covariate—forms subgroups	Bernoulli distribution (1, 0.4)
D	Treatment assignment	Defined in treatment assignment, values 0 or 1
Y	Outcome	Defined in outcome
ϵ_1, ϵ_2	Error terms	Multivariate normal distribution (0, 1)
Treatment assignment: Scenario 1: $f \sim 0.5X_1+0.5X_2+\varepsilon_1$; if $f<0$, $D=0$; otherwise $D=1$ Scenario 2: $f \sim 0.5X_1+0.5X_2+0.3Z+\varepsilon_1$; if $f<0$, $D=0$; otherwise $D=1$		
Outcome: $Y \sim 0.5X_2+0.5X_3+\alpha_1 D+\alpha_2 Z+\alpha_3 DZ+\varepsilon_2$; linear regression model where $\alpha_1=0.7$, $\alpha_2=0.4$, $\alpha_3=0.2$		
Characteristics that define simulated datasets	**Categories**	
Correlation between covariates (X_1–X_3)	0; 0.3; 0.7	
Correlation Z—covariates (X_1–X_3)	Yes; no	
Sample size	250; 500; 1000	
Variables selected in PS		
Univariate PS	X_2, X_3; X_1, X_2, X_3; X_2, X_3, Z; X_1, X_2, X_3, Z; X_1, X_2; X_1, X_2, Z; X_2; X_2, Z	
Generalized PS	X_2, X_3; X_1, X_2, X_3; X_1, X_2; X_2	

*The parameter values related to the different variables in the description of the treatment assignment and outcome are arbitrary.
PS indicates propensity score.

related to each dummy variable. Because we were interested in 3 coefficients per regression model [Eqs (2) or (4)] and these coefficients were not comparable one-to-one, we averaged the bias, MSE, and SE over the 3 relevant coefficients per regression model [Eqs (2) or (4)]. We then used this value for the bias, MSE, and SE per PS method to compare the PS methods.

Case Study

Our sample consisted of a total of 841 patients with personality disorders[32] who had enrolled for different types of psychotherapy in 6 mental health care institutes in The Netherlands. The patients were selected for either short-term (up to 6 mo) or long-term (>6 mo) psychotherapy in various settings.[25,33–35] The mean age was 34.12 (SD = 9.83; range, 17–62 y) and 68.6 % were female. To compare the PS methods, we investigated whether the treatment effect was modified by the severity of problems, that is, having mild or severe problems. Although we were aware of more recent possible classifications of severity,[36] for comparison purposes we differentiated between the patients having personality difficulties or a simple personality disorder versus patients having complex or more severe personality disorders based on a classification of personality disorders by Tyrer et al.[37,38]

The primary outcome measure was psychiatric symptomatology and was measured with the Global Severity Index (GSI), which is the mean score of the 53 items of the Brief Symptom Inventory.[39,40] The GSI ranges from 0 to 4, with higher scores indicating more problems. Three treatment institutes conducted their follow-up measures on the GSI at 12, 24, 36, and 60 months after baseline. The 3 remaining treatment institutions conducted their follow-up measures at the end of treatment, 6 and 12 months after end of treatment, and again at 36 and 60 months after baseline. As in an earlier study by Spreeuwenberg et al,[8] we used the mean GSI score of all follow-up measures as a primary outcome measure (range, 0.01–3.17).[32] We excluded 114 cases that had missing values on one of the potential confounders, leaving 727 patients in the final sample. The excluded cases were not significantly different on the outcome GSI.

The potential confounders were assessed at baseline, that is, age, sex, civil status, living situation, care of children, employment, level of education, duration of psychological complaints, treatment history, alcohol and drug abuse, motivation, treatment preferences, level of psychiatric symptomatology, level of personality pathology, interpersonal functioning, social role functioning, quality of life, number of Diagnostic and Statistical Manual (DSM)-IV Axis II cluster A personality disorders, number of DSM-IV Axis II cluster B personality disorders, number of DSM-IV Axis II cluster C personality disorders, and psychological capacities. For specific details of this study, we refer the reader to the literature.[25,33,34]

Computation

The analyses were performed with IBM SPSS for Windows, version 20 (SPSS Inc., Chicago, IL). All simulations were performed in *R* programming language, version 2.13.0.[41]

RESULTS

Monte Carlo Simulation Results

We evaluated the bias, MSE, and SE of the relevant effects in the simulation study (Table, Supplemental Digital

TABLE 2. MSE of Simulations

Sample Size	Correlation Covariates X_1, X_2, X_3	Correlation Z—X_1, X_2, X_3		Variables in PS Model: Univariate PS: X_2, X_3	X_1, X_2, X_3	X_2, X_3, Z	X_1, X_2, X_3, Z	X_1, X_2	X_1, X_2, Z	X_2	X_2, Z	Generalized PS: X_2, X_3	X_1, X_2, X_3	X_1, X_2	X_2
Scenario 1*															
N = 250	0	Absent	MSE	*0.0523*	0.0579	0.0642	0.0605	0.0600	0.0624	0.0541	0.0664	*0.0329*	0.0351	0.0428	0.0411
		Present	MSE	*0.0524*	0.0561	0.0641	0.0583	0.0608	0.0630	0.0551	0.0674	*0.0342*	0.0367	0.0451	0.0424
	0.3	Absent	MSE	*0.0491*	0.0562	0.0618	0.0608	0.0585	0.1322	0.0569	0.0675	*0.0348*	0.0375	0.0449	0.0472
		Present	MSE	*0.0556*	0.0683	0.0687	0.0724	0.0682	0.0722	0.0673	0.0709	*0.0356*	0.0381	0.0578	0.0725
	0.7	Absent	MSE	*0.0481*	0.0518	0.0593	0.0592	0.0560	0.0635	0.0561	0.0664	*0.0365*	0.0391	0.0422	0.0456
		Present	MSE	*0.0657*	0.0792	0.0926	0.0907	0.0741	0.0884	0.0805	0.0892	*0.0476*	0.0514	0.0719	0.0993
N = 500	0	Absent	MSE	*0.0261*	0.0294	0.0330	0.0311	0.0312	0.0329	0.0278	0.0350	*0.0162*	0.0183	0.0218	0.0209
		Present	MSE	0.0292	*0.0264*	0.1095	0.0306	0.0297	0.0310	0.0277	0.0334	*0.0163*	0.0186	0.0227	0.0213
	0.3	Absent	MSE	*0.0245*	0.0285	0.0318	0.0311	0.0285	0.0311	0.0285	0.0344	*0.0174*	0.0185	0.0217	0.0248
		Present	MSE	*0.0294*	0.0381	0.0328	0.0394	0.0385	0.0400	0.0403	0.0350	*0.0172*	0.0186	0.0361	0.0519
	0.7	Absent	MSE	0.0270	*0.0248*	0.0313	0.0315	0.0254	0.0299	0.0268	0.0326	*0.0183*	0.0198	0.0215	0.0250
		Present	MSE	*0.0331*	0.0432	0.0489	0.0494	0.0396	0.0469	0.0493	0.0450	*0.0238*	0.0249	0.0455	0.0776
N = 1000	0	Absent	MSE	*0.0132*	0.0148	0.0160	0.0154	0.0149	0.0155	0.0136	0.0164	*0.0081*	0.0090	0.0110	0.0105
		Present	MSE	*0.0130*	0.0146	0.0160	0.0152	0.0154	0.0161	0.0138	0.0168	0.0085	*0.0081*	0.0099	0.0094
	0.3	Absent	MSE	*0.0124*	0.0141	0.0158	0.0153	0.0145	0.0157	0.0156	0.0184	*0.0087*	0.0090	0.0108	0.0144
		Present	MSE	*0.0162*	0.0214	0.0174	0.0227	0.0232	0.0245	0.0262	0.0195	*0.0084*	0.0093	0.0235	0.0394
	0.7	Absent	MSE	*0.0121*	0.0131	0.0153	0.0964	0.0123	0.0145	0.0146	0.0174	*0.0093*	0.0095	0.0103	0.0149
		Present	MSE	*0.0174*	0.0256	0.0270	0.0290	0.0237	0.0277	0.0354	0.0238	*0.0116*	0.0122	0.0312	0.0643
Scenario 2*															
N = 250	0	Absent	MSE	*0.0537*	0.0575	0.0994	0.0678	0.0591	0.0698	0.0545	0.1030	*0.0329*	0.0344	0.0436	0.0414
		Present	MSE	*0.0576*	0.0622	0.1116	0.0770	0.0643	0.0787	0.0595	0.1152	*0.0348*	0.0362	0.0445	0.0430
	0.3	Absent	MSE	*0.0525*	0.0604	0.1117	0.0850	0.0618	0.0876	0.0589	0.1106	*0.0343*	0.0374	0.0458	0.0453
		Present	MSE	*0.0605*	0.0717	0.1382	0.0763	0.0745	0.0778	0.0739	0.1252	*0.0375*	0.0404	0.0627	0.0750
	0.7	Absent	MSE	*0.0512*	0.0558	0.0944	0.0872	0.0562	0.0883	0.0577	0.0977	*0.0373*	0.0388	0.0428	0.0465
		Present	MSE	*0.0684*	0.0809	0.1901	0.1042	0.0798	0.1045	0.0865	0.1732	*0.0490*	0.0531	0.0765	0.1042
N = 500	0	Absent	MSE	*0.0289*	0.0330	0.0682	0.0427	0.0335	0.0433	0.0294	0.0696	*0.0177*	0.0186	0.0220	0.0210
		Present	MSE	*0.0269*	0.0295	0.0660	0.0383	0.0302	0.0390	0.0275	0.0666	*0.0163*	0.0171	0.0208	0.0199
	0.3	Absent	MSE	*0.0245*	0.0279	0.0702	0.0467	0.0287	0.0478	0.0287	0.0675	*0.0172*	0.0183	0.0214	0.0245
		Present	MSE	*0.0317*	0.0394	0.0943	0.0390	0.0404	0.0392	0.0413	0.0784	*0.0183*	0.0196	0.0376	0.0517
	0.7	Absent	MSE	*0.0259*	0.0283	0.0634	0.0560	0.0287	0.0565	0.0303	0.0652	*0.0187*	0.0194	0.0209	0.0249
		Present	MSE	*0.0370*	0.0443	0.1483	0.0633	0.0445	0.0639	0.0528	0.1207	*0.0254*	0.0268	0.0471	0.0777
N = 1000	0	Absent	MSE	*0.0128*	0.0145	0.0453	0.0219	0.0149	0.0224	0.0133	0.0463	*0.0083*	0.0089	0.0105	0.0101
		Present	MSE	*0.0145*	0.0161	0.0502	0.0249	0.0166	0.0254	0.0148	0.0511	*0.0088*	0.0093	0.0114	0.0107
	0.3	Absent	MSE	*0.0127*	0.0146	0.0523	0.0308	0.0150	0.0315	0.0160	0.0491	*0.0087*	0.0092	0.0108	0.0141
		Present	MSE	*0.0173*	0.0220	0.0713	0.0203	0.0230	0.0205	0.0260	0.0555	*0.0090*	0.0098	0.0256	0.0404
	0.7	Absent	MSE	*0.0121*	0.0129	0.0468	0.0384	0.0131	0.0387	0.0153	0.0470	*0.0090*	0.0094	0.0102	0.0141
		Present	MSE	*0.0189*	0.0247	0.1193	0.0395	0.0260	0.0399	0.0356	0.0893	*0.0124*	0.0134	0.0342	0.0651

*Results in italics indicate that MSE was closest to zero.
MSE indicates mean squared error; PS, propensity score.

Content 1, http://links.lww.com/MLR/A877 for the bias, MSE, and SE in scenario 1; for scenario 2, see Table, Supplemental Digital Content 2, http://links.lww.com/MLR/A878). Because taking an average over 3 estimated bias values related to the 3 relevant coefficients per PS method can average out positive and negative bias values, the MSE was used to find which PS estimations was most efficient (Table 2). In almost all simulated datasets within scenario 1, when the subgroup was not related to the treatment assignment, the MSE was closest to zero if the variables related to the outcome only were included in the univariate PS and the generalized PS (Table 2). If the subgroup was related to the treatment assignment, as in scenario 2, the MSE was closest to zero when the variables related to the outcome were included in the PS model in all simulated datasets (Table 2). In both the scenarios, including the subgroup variable in the univariate, PS estimation gave larger MSE values.

When comparing the univariate and generalized PS methods, in which only the covariates related to the outcome were included, the MSE was smaller for the generalized PS in all simulated datasets in scenario 1 and scenario 2 (Table 2).

In addition, if the sample size increased, the MSE decreased in all simulations. If the correlation increased when there was a correlation between the subgroup and covariates, the MSE increased. However, if the correlation between the subgroup and covariates was absent, the MSE showed a rather inconsistent pattern. Comparing the simulations when the correlation between the subgroup and covariates was either present or absent, gave overall lower MSE values if this correlation was absent.

Univariate PS: Case Study

The univariate PS was applied according to the protocol described by Bartak et al.[4] In total, 28 covariates related to outcome ($P = 0.10$) were selected in the PS estimation. We added 4 sociodemographic variables as these are considered highly relevant in psychotherapy research.[4] The subgroup of interest, that is the severity of problems, was related to treatment assignment ($\chi^2_1 = 10.80$, $P = 0.001$) and to the outcome ($B = 0.318$, $P = 0.000$; in a linear regression on the outcome). Thus, in applying the PS methods in the case study, we followed the results of scenario 2: for the univariate PS, we excluded the variable that reflected the severity of problems from the PS estimation.

The PS was estimated in a logistic regression analysis on the treatment assignment and no interaction terms between the covariates were added. The distributions of the estimated PS scores showed considerable overlap (Fig. 1). A lack of overlap would yield imprecise estimates of the treatment effect.[8]

The PS was then added to a regression model on the outcome GSI, in which treatment duration, severity of problems, and an interaction term were the independent variables:

$$\begin{aligned}\text{Outcome} = {} & \beta_0 + \beta_1 \text{PS} + \beta_2\,\text{Treatment} + \beta_3\,\text{Severity} \\ & + \beta_4\,\text{Treatment} \times \text{Severity}.\end{aligned} \quad (5)$$

If patients had mild problems, long-term treatment yielded more favorable results than short-term treatment (standardized coefficient of 0.092; Table 3). For patients having severe problems, both treatment options were equally effective: for patients having severe problems in the short-term treatment the standardized coefficient was 0.240, whereas for the long-term treatment, the final standardized coefficient was 0.248. The interaction effect was, however, not significant. Excluding this coefficient indicated that long-term treatment was preferred for patients with severe problems (Table 3).

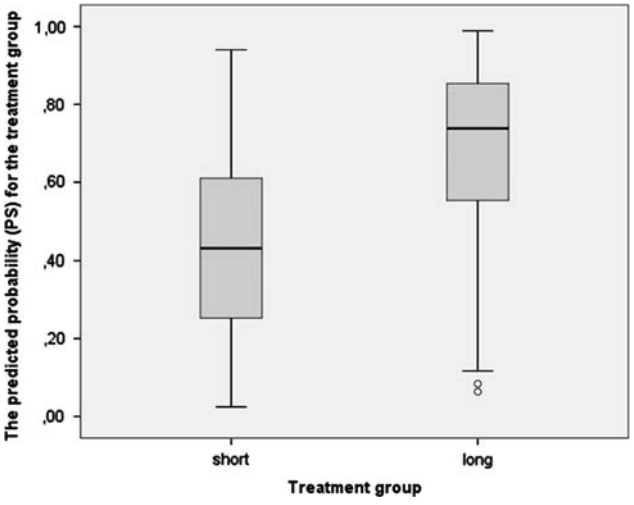

FIGURE 1. Boxplots of the overlap of the univariate propensity score (PS) distributions.

Generalized PS: Case Study

The generalized PS was applied according to the protocol described by Spreeuwenberg et al.[8] The generalized PS was estimated as a combination variable of treatment duration and severity of problems: short-term treatment for patients having mild problems (reference category, N = 268), short-term treatment for patients having severe problems (N = 34), long-term treatment for patients having mild problems (N = 338), and long-term treatment for patients having severe problems (N = 87). The same list of covariates selected in the univariate PS was included in the generalized PS estimation. Here, we followed the simulation results of scenario 2. The PS was estimated by multinomial regression analysis, with the combination variable of treatment duration and severity of problems as dependent variable and not including interaction terms between the covariates. However, the number of cases in 2 groups was small and validity of the model fit was therefore uncertain.[9,16,26,27] Because the 4 estimated generalized PSs add up to 1 and are complementary, only 3 of 4 were used in further analyses. As required when using the PS, the ranges of the estimated PS scores showed overlap (Fig. 2).

In the final regression model on the outcome GSI, 3 generalized PSs and 3 dummies indicating group membership were included [Eq. (6)]:

$$\begin{aligned}\text{Outcome} = {} & \beta_0 + \beta_1 \text{PS}_1 + \beta_2 \text{PS}_2 + \beta_3 \text{PS}_3 \\ & + \beta_4\,\text{Long Mild}_1 + \beta_5\,\text{Short Severe}_2 \\ & + \beta_6\,\text{Long Severe}_3.\end{aligned} \quad (6)$$

These results indicated that long-term treatment was more favorable for patients having mild problems. For patients having severe problems, both treatment options were almost equally effective, just as was presented when the univariate PS was applied while taking the interaction effect into account (Table 3).

To compare the relative effects of this model to the results of using the univariate PS, we used the standardized coefficients. These coefficients can be interpreted independent of the intercept and PS scores added in each model. The coefficient for patients having severe problems in short-term treatment was 0.240 using the univariate versus 0.129 using the generalized PS. The coefficient of patients having mild problems in long-term treatment was almost equal to 0.092 using the univariate PS versus 0.099 using the generalized PS. For patients having severe problems in long-term treatment, we combined the standardized coefficients of the model in which the univariate PS was applied and compared it with the corresponding coefficient in the model of the generalized PS. For these patients, the combined coefficient was 0.248 using the univariate versus 0.139 using the generalized PS (Table 3).

DISCUSSION

The present study illustrates the use of the univariate and generalized PS in subgroup analyses in nonrandomized outcomes research, and describes how the generalized PS could be used in subgroup analysis. The results indicate that the generalized PS—estimated by crossing the treatment

TABLE 3. Linear Regression on GSI Outcome

Variables (N = 727)	*B*	95% CI	Standardized Coefficient
Using the univariate propensity score (PS)			
Intercept	0.512**	0.408–0.614	—
Long duration treatment group	0.106**	0.006–0.206	0.092
Severe problems	0.366**	0.168–0.564	0.240
Interaction	−0.147	−0.382 to 0.088	−0.084
Using the generalized PS			
Intercept	0.536**	0.426–0.646	—
Short duration—mild problems	Reference	—	—
Short duration—severe problems	0.348**	0.091–0.605	0.129
Long duration—mild problems	0.113*	0.011–0.215	0.099
Long duration—severe problems	0.244**	0.064–0.424	0.139

*$P < 0.05$.
**$P < 0.01$.
CI indicates confidence interval; GSI, Global Severity Index; PS, propensity score.

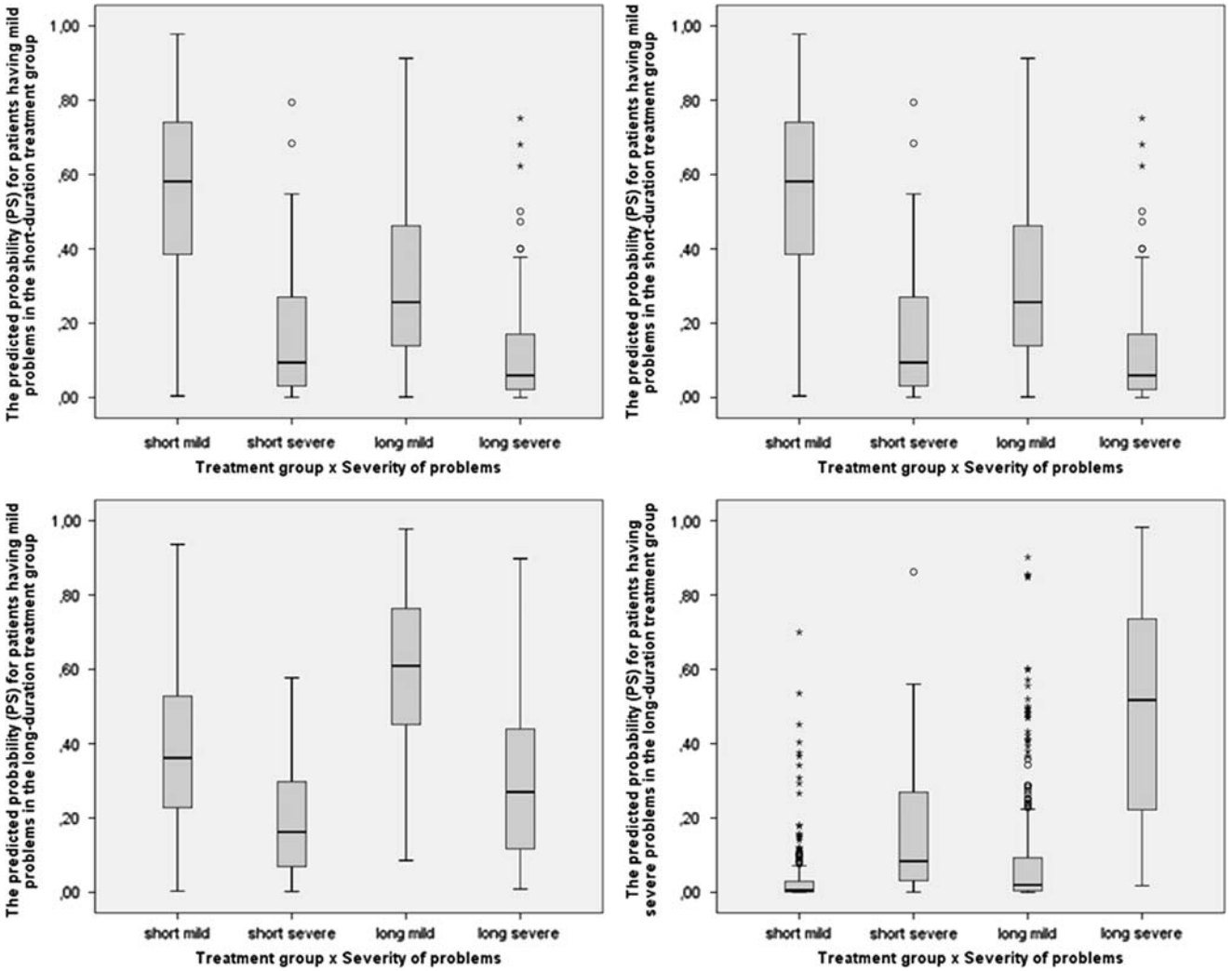

FIGURE 2. Boxplots of the overlap of the generalized propensity score (PS) distributions. ○ is a mild outlier (defined when the value > 3rd quartile +1.5 • interquartile range or < 1st quartile −1.5 • interquartile range). * an extreme outlier (defined when the value > 3rd quartile +3 • interquartile range or < 1st quartile −3 • interquartile range).

options with a subgroup variable—could be a feasible option and should be seriously considered when assessing subgroup effects while correcting for observed pretreatment differences. In the Monte Carlo simulation study, the generalized PS gave more efficient results overall than the univariate PS, regardless of whether there was a relationship between the subgroup and treatment assignment. In both PS methods, the variables related to the outcome should be included in the PS estimation. These results follow earlier studies of Brookhart et al[30] and Austin et al[31] in selecting only the covariates related to the outcome. Furthermore, when the univariate PS was used, the subgroup of interest should be excluded from the PS estimation. Applying the 2 PSs estimations on real-world data produced almost equal model results, illustrating the modifying effect of the severity of problems on the differential effectiveness of 2 psychotherapy treatment arms.

In applying the generalized PS when analyzing subgroups effects, a researcher should take into account additional characteristics of their datasets. Firstly, the characteristics of the subgroup variable should be taken into account. For example, the independence of irrelevant alternatives assumption can be violated. This assumption will be violated if, for example, short-term psychotherapy for patients having mild problems is no longer available and this influences the relative risks of the remaining categories. We tested this assumption in our study and it was not violated. However, when it is violated, a nested structure can overcome this violation by first defining the probability that a patient belongs to a particular subgroup and is subsequently assigned to a treatment option. Fuji et al[29] focused on a 2-stage structure (ie, a nested structure), in which patients could be assigned sequentially to 2 treatment options, each consisting of 2 suboptions. Yet, if for example a patient characteristic evolves after treatment, it could mediate the relationship between the independent and dependent variable (ie, a mediator) and should be analyzed differently from the proposed methods.[42]

Secondly, for various reasons, other application methods when using the PS could be more advantageous.[12,43] For example, the PS can be estimated in each subgroup separately,[23,24] requiring a large study population to have sufficient power. If the sample size is large enough, a multivariable-adjusted model including interaction terms for the subgroup effects can also be a valid alternative.[22] In this study, we applied the PS using covariate adjustment, as sample sizes in clinical practice can be small and this method uses the complete sample size. However, covariate adjustment inherently assumes a correctly specified outcome regression model,[44] whereas for matching this is not required. Inverse probability weighting on the PS is a third and efficient method to control for selection bias.[13] The latter 2 methods can indeed eliminate most systematic differences between treated and untreated subjects,[9] but matching for example can result in very small comparison groups.[8] To improve precision of the effect estimates, those methods can also be used in combination with regression analysis.[28,45]

Thirdly, when using the PS, researchers should assess carefully which method is most appropriate for their specific research question. For example, when the treatment effect itself is more important, using the univariate PS and adjusting for extra covariates additionally reduces the bias of the treatment effect estimation.[15,16] However, incorporating effect modification reduces the direct interpretability of the main treatment effect.[20,22] Furthermore, if the distributions of the effect modifying variables vary highly, the overall estimates may differ across PS quantiles.[19] Different adjustment methods can thus result in divergent results, which all may be correct, but strongly depends on the research question and the population in which the estimation is most suitable.[18,22]

Our study has several limitations. First, only a basic simulation study was designed. The characteristics of the simulated data were known in advance to the analyzer, which could have influenced the analysis and method chosen. Testing the methods using new simulated data that could be based on a real dataset is recommended to further investigate the performance of the methods. Furthermore, the number of simulated datasets was rather small, which could have caused the small differences and inconsistencies in the simulation results, due to Monte Carlo error. Secondly, the overlap for the generalized PSs in the case study appeared to be less than optimal (Fig. 2). This could have caused the difference in the estimated coefficients when the PS methods were compared. A distance score defined by Cochran and Rubin[46] can be used to precisely test and define the overlap. A third limitation deals with the selection of variables into the PS. We left out the subgroup variable in the univariate PS, whereas Rubin and Thomas[15] state that no prognostic variable should be left out. Although the results of our simulations and the case study only slightly changed when we added the subgroup variable to the univariate PS, we recommend investigating its influence in more detail. Fourth, although we controlled for observed pretreatment variables, hidden bias due to unobserved confounders could not be controlled for in the case study. As we did not include hidden bias in the simulated datasets either, we do not know the effect of hidden bias in using the PSs in subgroup analysis.

This study supports the idea that the generalized PS can be used in estimating the treatment effect when this is modified by a subgroup variable. As patient-tailored treatment becomes more and more important in outcomes research,[14] this study contributes to the literature on how to handle effect estimation in nonrandomized outcome studies of patient subgroups using the PS.

REFERENCES

1. Winship C, Mare RD. Models for sample selection bias. *Ann Rev Sociol*. 1992;18:327–350.
2. D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med*. 1998;17:2265–2281.
3. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70:41–55.
4. Bartak A, Spreeuwenberg MD, Andrea H, et al. The use of propensity score methods in psychotherapy research: a practical application. *Psychother Psychosom*. 2009;78:26–34.
5. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66:688–701.

6. Feng P, Zhou XH, Zou QM, et al. Generalized propensity score for estimating the average treatment effect of multiple treatments. *Stat Med*. 2012;31:681–697.
7. Imbens GW. The role of the propensity score in estimating dose-response functions. *Biometrika*. 2000;87:706–710.
8. Spreeuwenberg MD, Bartak A, Croon MA, et al. The multiple propensity score as control for bias in the comparison of more than two treatment arms: an introduction from a case study in mental health. *Med Care*. 2010;48:166–174.
9. Austin PC. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Med Decis Making*. 2009;29: 661–677.
10. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med*. 1997;127:757–763.
11. Rosenbaum PR. Discussing hidden bias in observational studies. *Ann Intern Med*. 1991;115:901–905.
12. Austin PC. An Introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*. 2011;46:399–424.
13. Hirano K, Imbens GW, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*. 2003;71:1161–1189.
14. Norcross JC, Wampold BE. What works for whom: tailoring psychotherapy to the person. *J Clin Psychol*. 2011;67:127–132.
15. Rubin DB, Thomas N. Combining propensity score matching with additional adjustments for prognostic covariates. *J Am Stat Assoc*. 2000;95:573–585.
16. Stürmer T, Joshi M, Glynn RJ, et al. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol*. 2006;59: 437–447.
17. Glynn RJ, Schneeweiss S, Sturmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic Clin Pharmacol Toxicol*. 2006;98:253–259.
18. Kurth T, Walker AM, Glynn RJ, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *Am J Epidemiol*. 2006;163:262–270.
19. Lunt M, Solomon D, Rothman K, et al. Different methods of balancing covariates leading to different effect estimates in the presence of effect modification. *Am J Epidemiol*. 2009;169:909–917.
20. Sturmer T, Rothman KJ, Glynn RJ. Insights into different results from different causal contrasts in the presence of effect-measure modification. *Pharmacoepidemiol Drug Saf*. 2006;15:698–709.
21. Ye Y, Bond JC, Schmidt LA, et al. Toward a better understanding of when to apply propensity scoring: a comparison with conventional regression in ethnic disparities research. *Ann Epidemiol*. 2012;22: 691–697.
22. Liem YS, Wong JB, Hunink MM, et al. Propensity scores in the presence of effect modification: a case study using the comparison of mortality on hemodialysis versus peritoneal dialysis. *Emerg Themes Epidemiol*. 2010;7:1–8.
23. Radice R, Ramsahai R, Grieve R, et al. Evaluating treatment effectiveness in patient subgroups: a comparison of propensity score methods with an automated matching approach. *Int J Biostat*. 2012;8:25. (1-43).
24. Kreif N, Grieve R, Radice R, et al. Methods for estimating subgroup effects in cost-effectiveness analyses that use observational data. *Med Decis Making*. 2012;32:750–763.
25. Bartak A, Spreeuwenberg MD, Andrea H, et al. Effectiveness of different modalities of psychotherapeutic treatment for patients with cluster C personality disorders: results of a large prospective multicentre study. *Psychother Psychosom*. 2010;79:20–30.
26. Shah BR, Laupacis A, Hux JE, et al. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *J Clin Epidemiol*. 2005;58:550–559.
27. Weitzen S, Lapane KL, Toledano AY, et al. Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiol Drug Saf*. 2004;13:841–853.
28. Hirano K, Imbens GW. Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. *Health Serv Outcomes Res Methodol*. 2001;2:259–278.
29. Fujii Y, Henmi M, Fujita T. Evaluating the interaction between the therapy and the treatment in clinical trials by the propensity score weighting method. *Stat Med*. 2012;31:235–252.
30. Brookhart MA, Schneeweiss S, Rothman KJ, et al. Variable selection for propensity score models. *Am J Epidemiol*. 2006;163:1149–1156.
31. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Stat Med*. 2007;26:734–753.
32. Association AP. *Diagnostic and Statistical Manual of Mental Disorders (Text Revision)*. 4th ed. Washington, DC: American Psychiatric Association; 2000.
33. Bartak A, Andrea H, Spreeuwenberg MD, et al. Patients with cluster A personality disorders in psychotherapy: an effectiveness study. *Psychother Psychosom*. 2011;80:88–99.
34. Bartak A, Andrea H, Spreeuwenberg MD, et al. Effectiveness of outpatient, day hospital, and inpatient psychotherapeutic treatment for patients with Cluster B personality disorders. *Psychother Psychosom*. 2011;80:23–38.
35. Soeteman DI, Verheul R, Meerman AM, et al. Cost-effectiveness of psychotherapy for cluster C personality disorders: a decision-analytic model in the Netherlands. *J Clin Psychiatry*. 2011;72:51–59.
36. Crawford MJ, Koldobsky N, Mulder R, et al. Classifying personality disorder according to severity. *J Pers Disord*. 2011;25:321–330.
37. Tyrer P. New approaches to the diagnosis of psychopathy and personality disorder. *J R Soc Med*. 2004;97:371–374.
38. Tyrer P, Johnson T. Establishing the severity of personality disorder. *Am J Psychiatry*. 1996;153:1593–1597.
39. Arrindell WA, Ettema JHM. *Herziene Handleiding bij de Multidimensionele Psychopathologie-Indicator SCL-90-r [Revised Manual for a Multidimensional Indicator of Psychopathology]*. Lisse, The Netherlands: Swets & Zeitlinger; 2003.
40. Derogatis LR. *SCL-90 (R): Administration, Scoring and Procedure Manual-II for the Revised Version*. Townson, MD: Clinical Psychometric Research; 1986.
41. R Development Core Team. *R: A Language and Environment for Statistical Computing Version 2130*. Vienna, Austria: R Foundation for Statistical Computing; 2010.
42. VanderWeele TJ, Vansteelandt S. Conceptual issues concerning mediation, interventions and composition. *Stat Interface*. 2009;2:457–468.
43. Rubin DB. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J Am Stat Assoc*. 1979;74:318–324.
44. Rubin DB. On principles for modeling propensity scores in medical research. *Pharmacoepidemiol Drug Saf*. 2004;13:855–857.
45. Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: a review. *Rev Econ Stat*. 2004;86:4–29.
46. Cochran WG, Rubin DB. Controlling bias in observational studies: a review. *Sankhya*. 1973;35:417–446.