Innovations

Psychotherapy and Psychosomatics

Psychother Psychosom 2009;78:26–34
DOI: 10.1159/000162298

Published online: October 14, 2008

The Use of Propensity Score Methods in Psychotherapy Research

A Practical Application

Anna Bartak[a, b] Marieke D. Spreeuwenberg[c, d] Helene Andrea[a, e]
Jan J.V. Busschbach[a, e] Marcel A. Croon[d] Roel Verheul[a, b, f]
Paul M.G. Emmelkamp[b] Theo Stijnen[g]

[a]Viersprong Institute for Studies on Personality Disorders, Halsteren; [b]Department of Clinical Psychology, University of Amsterdam, and [c]Department of Clinical Epidemiology and Biostatistics, VU Medical Centre, Amsterdam; [d]Department of Methodology and Statistics, Tilburg University, Tilburg; [e]Department of Medical Psychology and Psychotherapy, Erasmus Medical Centre, Rotterdam; [f]Trimbos Institute, Utrecht; [g]Department of Medical Statistics and Bioinformatics, Leiden University Medical Centre, Leiden, The Netherlands

Key Words
Psychotherapy · Research design, propensity score · Quasi-experiment · Randomized controlled trials, selection bias

Abstract
Background: Randomized controlled trials are considered the best scientific proof of effectiveness. There is increasing concern, though, about their feasibility in psychotherapy research. We discuss a quasi-experimental study design for situations in which a randomized controlled trial is not feasible. Here, as an alternative strategy, the propensity score (PS) method is used to correct for selection bias. ***Methods:*** We used data from a Dutch research project, SCEPTRE (Study on Cost-Effectiveness of Personality Disorder Treatment). The sample consisted of 749 psychotherapy patients with personality pathology. We tested whether the PS method was useful and applicable. We examined differences between 2 treatment groups (short vs. long treatment duration) in pretreatment characteristics before and after PS correction. This revealed the impact of the PS on outcome differences. ***Results:*** The PS offered statistical control over observed pretreatment differences between patients in a non-randomized study. ***Conclusions:*** When a randomized controlled trial is not possible, this quasi-experimental design using the PS could be a feasible alternative. Its advantages and limitations are discussed. Implemented carefully, this method is promising for future effectiveness research.

Introduction

The first randomized study in medicine was conducted by Amberson [1] in 1931 by flipping a coin. Now, randomized controlled trials are considered the gold standard for comparing the effectiveness of psychotherapeutic treatment methods. Randomization assumes that all known and unknown characteristics of the participants are balanced between the experimental groups, except for the treatment condition. With randomization, treatment effects can theoretically be estimated by merely subtracting the mean responses of the treatment groups [2]. In many cases, though, randomization may be difficult, unethical or impossible [3], especially in psychotherapy research [4–7]. Here patients' and clinicians' personal preferences regarding treatment allocation may work against ran-

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

0033–3190/09/0781–0026$26.00/0

Accessible online at:
www.karger.com/pps

Anna Bartak
Viersprong Institute for Studies on Personality Disorders (VISPD)
PO Box 7
NL–4660 AA Halsteren (The Netherlands)
Tel. +31 164 632 200, Fax +31 164 632 220, E-Mail anna.bartak@deviersprong.nl

domization. The resulting high number of excluded subjects makes the generalization of such results difficult [8]. Hence, research on treatment effects in various (para)medical fields often requires well-designed and carefully conducted non-randomized studies [e.g. 9, 10]. Shadish et al. [11] called these studies 'quasi-experimental', based on their resemblance to true experiments, except for the random assignment of participants to treatments. In these quasi-experimental designs, the researcher has some influence on the manipulation of treatment and measurement. This is in contrast to pure observational studies, where the size and direction of a relationship among variables are simply observed [11]. In case of non-random allocation to treatment, persons with different treatments can differ on pretreatment characteristics. This 'selection bias' affects the estimates of the treatment effect. Rosenbaum [12] distinguishes 2 types of bias: hidden bias, due to unobserved differences in baseline characteristics, and overt bias, due to observed differences in baseline characteristics. Hidden bias is the most difficult to deal with. Overt bias can be corrected with various statistical methods, by incorporating known initial differences into the statistical analysis. The most widely used methods are matching, stratification and regression adjustment [13–15]. In matching, each individual in the treatment group is paired with the most similar individual in the control group. After matching, the groups as a whole are assumed to be as similar as possible on the matched characteristics. In stratification, subgroups of patients are formed based on baseline variables. In psychotherapy research, however, there is usually a large number of variables to match or stratify on, making it almost impossible to find patients or groups similar on all these variables. This is called the 'dimensionality problem'. Regression analysis with covariates, a third tool to compensate for overt bias, has limitations as well: when many pretreatment variables are used as covariates, statistical-modelling problems and a loss of power arise. A promising alternative method to correct for overt bias is the propensity score (PS) method [12, 16].

Propensity Score

Rosenbaum and Rubin [16] suggested using the PS method to reduce the 'dimensionality problem'. The PS method reduces the entire collection of observed pretreatment variables (X) to a single score. The estimated PS is defined as the conditional probability of assignment to a particular treatment, given a set of observed pretreatment characteristics. Let Z denote treatment group membership, where Z = 0 denotes the control condition and Z = 1 denotes the treatment condition. Then, PS is defined as:

$$PS = P(Z = 1 \mid X)$$

Rosenbaum and Rubin [16] proved that, given the value of the PS, assignment to treatment no longer depends on baseline variables. The PS is a score balancing all observed pretreatment variables among patients with the same value of the PS. In this way, the PS method can put overt bias under statistical control. Different from the conventional approach, i.e. controlling for or matching on many baseline variables, the PS enables researchers to deal with one composite, single variable which is much easier and, in regression analysis, preserves power. The PS has so far been used in medicine [e.g. 17–24], social sciences [e.g. 25–28] and economics [e.g. 29–31]. The United States Food and Drug Administration recommended the PS as a tool to overcome selection bias in treatment studies [32]. In psychotherapy research, however, the PS is not widely known. To the best of our knowledge, only a handful of pioneering studies have used this instrument for selection bias control in non-randomized studies [33–36].

Aim

The aims of this paper are (1) to investigate if the PS method is applicable in psychotherapy research and (2) to outline a step-by-step protocol for the psychotherapy researcher to facilitate use of the PS in comparative outcome studies when randomization is unfeasible. We applied the PS method to a case study, the research project SCEPTRE ('Study on Cost-Effectiveness of Personality Disorder Treatment') [37]. We compared 2 treatment groups from SCEPTRE, using the PS to correct for known baseline differences. The 2 treatment groups selected for comparison are short versus long psychotherapy duration, as this distinction is straightforward and simple to understand. Results should only be interpreted as an illustration, not as a relevant clinical message. All statistical techniques presented in this paper are easily done in common statistical packages, such as SPSS.

Method

Participants

Patients were recruited from 6 mental health care centres in the Netherlands offering outpatient, day hospital and/or inpatient psychotherapy for patients with personality pathology. Out of 2,540 patients who were admitted to the centres from March 2003

Table 1. Differences in continuous variables between short-term and long-term treatment groups

Variable	Short term (n = 331)	Long term (n = 328)	Unstandardized β treatment duration (short/long)	
			before PS correction	after PS correction
Age, years	36.83 ± 9.63	31.86 ± 9.49	–4.97***	–0.10
Personality pathology (DAPP-BQ)				
Emotional dysregulation	21.93 ± 4.02	22.87 ± 3.66	0.95**	0.05
Dissocial behaviour	17.35 ± 4.10	18.01 ± 4.40	0.66*	0.09
Inhibitedness	22.11 ± 5.06	22.51 ± 4.97	0.40	0.03
Compulsivity	24.29 ± 6.84	23.87 ± 7.29	–0.42	–0.18
Motivation (MTQ-8)				
Need for help	28.87 ± 5.23	28.46 ± 5.22	–0.41	–0.02
Readiness to change	30.70 ± 5.04	29.96 ± 5.16	–0.74	–0.53
Quality of life (EQ-5D)	0.59 ± 0.26	0.55 ± 0.26	–0.04	–0.00
Psychological capacities (SIPP)				
Self-control	4.65 ± 0.91	4.48 ± 0.90	–0.17*	–0.03
Social concordance	5.72 ± 0.78	5.63 ± 0.81	–0.09	–0.03
Identity integration	3.54 ± 0.71	3.38 ± 0.65	–0.16**	–0.01
Relational functioning	3.97 ± 0.84	3.79 ± 0.78	–0.17**	–0.02
Responsibility	4.67 ± 0.84	4.52 ± 0.88	–0.14*	–0.02
Psychiatric symptomatology (SCL-90)	2.39 ± 0.62	2.55 ± 0.65	0.16**	0.01
Functioning (OQ-45)				
Interpersonal functioning	20.07 ± 6.29	21.60 ± 6.01	1.54**	–0.01
Social role functioning	15.28 ± 4.86	15.59 ± 4.58	0.32	0.06
Axis-II diagnosis (SIDP-IV)				
Number of Axis-II cluster A disorders	0.04 ± 0.19	0.09 ± 0.29	0.05*	0.01
Number of Axis-II cluster B disorders	0.19 ± 0.48	0.34 ± 0.58	0.15***	0.03
Number of Axis-II cluster C disorders	0.65 ± 0.78	0.70 ± 0.79	0.05	0.03
Duration of psychological problems	3.59 ± 0.81	3.59 ± 0.79	0.00	0.04

Values are presented as means ± SD. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

to March 2006, 1,047 were selected for treatment, i.e. short- or long-duration psychotherapy in various settings. Before treatment allocation, all patients were assessed with a routinely distributed assessment battery including self-report questionnaires. A semistructured interview was conducted to diagnose personality disorders with DSM-IV criteria. Of the 1,047 patients selected for treatment, 298 patients had not yet completed a follow-up measure, so no outcome score could be calculated. These were excluded from the analyses, leaving 749 patients. Of these, 507 (67.7%) were female. The mean age was 34.24 years (SD 9.93, range 17–62). We divided this sample into 2 groups: one group allocated to short-term therapy (up to 6 months), the other group allocated to long-term therapy (more than 6 months).

Measures

The baseline assessment measured a long list of social, economic and diagnostic variables carefully selected by both clinicians and researchers, based on literature and clinical knowledge (see tables 1 and 2).

Psychiatric symptomatology was measured with the Symptom Checklist 90 Revised, Dutch version (SCL-90) [38–40]. In this study, we used the Global Severity Index of the SCL-90 (GSI; the mean score of all 90 items) as the primary outcome measure, with higher scores indicating more distress. To measure the type and degree of personality pathology we used the 4 higher-order factors of the Dimensional Assessment of Personality Pathology Basic Questionnaire, Dutch version (DAPP-BQ) [41, 42]: (1) emotional dysregulation, (2) dissocial behaviour, (3) inhibition and (4) compulsivity. Psychosocial functioning was measured with the Outcome Questionnaire 45, Dutch version (OQ-45) [43]. Of this self-report measure, we used 2 subscales: (1) interpersonal relations and (2) social-role functioning. Health-related quality of life was assessed with the EuroQoL EQ-5D [44]. Personality disorders were assessed with the Structured Interview of DSM-IV Personality, Dutch version (SIDP-IV) [45–47]. The severity of personality pathology was measured with 5 higher-order domains of the Severity Indices of Personality Problems (SIPP) [48, 49]: self-control, social concordance, identity integration, relational functioning and responsibility. To measure patients' motivation for treatment, we used the two scales of the Motivation for Treatment Questionnaire (MTQ-8) [50]: need for help and readiness to change.

Table 2. Differences in categorical variables between short-term and long-term treatment groups

Variable	Demographic data, %		Odds ratio treatment duration (short/long)	
	short term (n = 331)	long term (n = 328)	before PS correction	after PS correction
Gender				
Female	65.3	68.9	1.00[a]	1.00[a]
Male	34.7	31.1	1.18	1.01
Civil status				
Married	27.5	18.0	1.00[a]	1.00[a]
Widowed or divorced	13.3	10.1	0.86	1.07
Never married	59.2	72.0	0.54**	1.04
Living situation				
Alone	39.0	38.4	1.00[a]	1.00[a]
With partner (with or without child)	44.4	29.3	1.50*	0.98
With child without partner	5.7	6.4	0.88	1.02
With parent(s)	4.2	17.7	0.24***	1.14
With other people	6.6	8.2	0.80	1.02
Childcare				
No care for children	72.5	80.5	1.00[a]	1.00[a]
Care for children	27.5	19.5	1.56*	0.95
Work situation				
Unemployed	33.2	36.3	1.00[a]	1.00[a]
Study or paid work	66.8	63.7	1.14	0.99
Level of education				
Low	19.3	28.0	1.00[a]	1.00[a]
Middle	22.7	17.7	1.86**	0.94
High	58.0	54.3	1.55*	0.89
Previous outpatient treatment				
No	17.2	22.6	1.00[a]	1.00[a]
Yes	82.8	77.4	1.40	1.00
Previous inpatient treatment				
No	83.4	79.9	1.00[a]	1.00[a]
Yes	16.6	20.1	0.79	1.03
Previous medication treatment				
No	53.8	52.7	1.00[a]	1.00[a]
Yes	46.2	47.3	0.96	1.17
Alcohol abuse				
No	84.5	87.2	1.00[a]	1.00[a]
Yes	15.5	12.8	1.25	0.80
Drug abuse				
No	86.1	77.4	1.00[a]	1.00[a]
Yes	13.9	22.6	0.55**	1.10
Preference for treatment setting				
Outpatient	12.1	22.9	1.00[a]	1.00[a]
Day hospital	30.9	24.8	2.36***	0.68
Inpatient	35.5	29.4	2.29**	0.85
Do not know	21.5	22.9	1.78*	0.67
Preference for treatment duration				
Up to 6 months	43.5	25.3	1.00[a]	1.00[a]
Longer than 6 months	26.9	37.2	0.42***	0.99
Do not know	29.6	37.5	0.46***	1.04
Treatment setting				
Outpatient	18.7	34.1	1.00[a]	1.00[a]
Day hospital	31.7	30.2	1.92**	0.99
Inpatient	49.5	35.7	2.53***	0.96

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

[a] Category is reference category; for regression purposes all categorical variables were translated into dummy variables, whereby the first category always serves as a reference category with an odds ratio of 1.00.

Results

To avoid bias in the estimation of the treatment effect, we corrected for the influence of known pretreatment differences. We did this by stratification of the sample based on the PS. This process took 9 steps, described below.

Stratification

Step 1: Effect Estimation before Correction

Before correction for known pretreatment differences, we estimated the treatment effect by conducting a linear regression analysis. In this 'naïve' estimate the only independent variable was 'group membership' (short vs. long), the dependent variable was outcome, being defined here as the level of psychiatric symptomatology (GSI) at the first measurement following baseline. The uncorrected treatment effect β was 0.20 (SE = 0.05; $p < 0.001$).

Step 2: Balance Check before Correction

We compared the 2 treatment groups on pretreatment variables before stratification. Note that this step is neither relevant for variable selection for the PS, nor for further analyses. It is only important here to be able to demonstrate the influence of propensity correction on the balance between groups. This demonstration can be done in several ways. For illustration purposes, we chose to show a comparison of overall regression coefficients. We conducted a number of regression analyses with 'group membership' as an independent variable and pretreatment characteristics as dependent variables (linear regression analyses for continuous variables, see table 1, and multinomial logistic regression analyses for categorical variables, see table 2). The 2 patient groups (short- vs. long-term treatment) differed significantly on 19 of the 34 baseline variables. This implies that, without correction for these differences, the 2 groups were not readily comparable – a problem that may be dealt with using the PS.

Step 3: Variable Selection for PS Estimation

To estimate the PS, we used all baseline variables related to outcome (GSI). To identify related variables, we conducted a number of linear regression analyses with the GSI as the dependent variable and each potential confounder as an independent variable. The following variables emerged as primary candidates for the estimation of the PS: level of personality pathology (i.e. emotional dysregulation, dissocial behaviour and inhibitedness), motivation for treatment (i.e. need for help), quality of life, psychological capacities (i.e. self-control, social concordance, identity integration, relational functioning and responsibility), level of psychiatric symptomatology, functioning (i.e. interpersonal and social-role functioning), number of cluster A, B and C personality disorders, working situation, level of education, previous inpatient treatment, patient preferences for treatment duration and setting of treatment. Sociodemographic variables were added to the PS model as well, because they are considered highly relevant in psychotherapy research: age, gender, marital status, living situation and responsibility for the care of children.

Step 4: Exclusion of Incomplete Cases

In this example, only patients with no missing values on the selected potential confounders (see 'Step 3') were included in the PS analysis. The final sample therefore consisted of 659 patients. Alternatively, imputation techniques might be used to fill in the missing values in estimation variables.

Step 5: PS Estimation

The PS was estimated in a logistic regression analysis. All selected potential confounders were used as independent variables, and 'group membership' as the dependent variable. One can estimate and save these probabilities for each subject, e.g. by using the option 'save predicted probability' in SPSS.

Step 6: Inspection of Overlap and Exclusion of Non-Overlapping Cases

For the short-term treatment group (n = 331), the PS ranged between 0.03 and 0.98; for the long-term treatment group (n = 328), the PS ranged between 0.10 and 0.99 (see fig. 1). The PS range that both groups cover is between 0.10 and 0.98. Patients with a PS outside this common range (n = 24) were excluded from the stratification, leaving a sample of 635 patients.

Step 7: Stratification of the Sample Based on the PS

The sample of 635 patients was divided into 5 equal subgroups with similar PS (so-called 'strata' [51], see table 3). We then created 4 dummy variables based on these 5 groups.

Step 8: Balance Check after Correction

We needed to know if the stratification of the sample based on the PS resulted in a balance of pretreatment variables between the 2 treatment groups. Therefore, we checked again for differences in pretreatment variables.

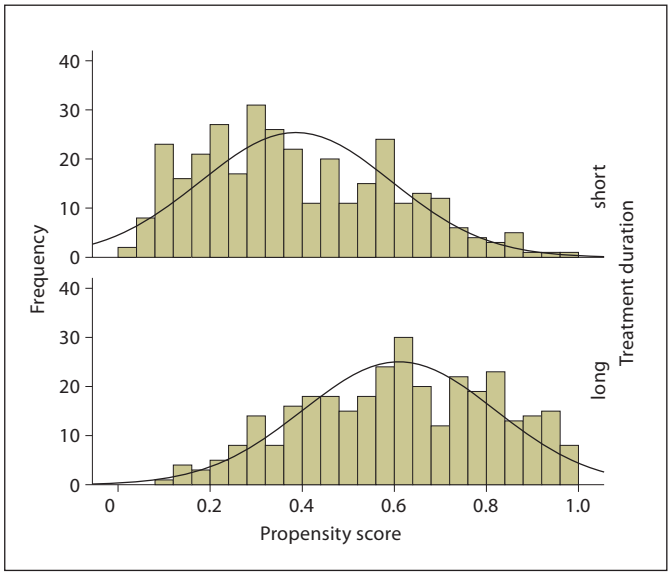

Fig. 1. Overlap of PS in the 2 treatment groups (short/long).

Table 3. Distribution of patients across the 5 strata

Stratum	Short term	Long term	Total
1	104	23	127
2	78	49	127
3	62	65	127
4	48	79	127
5	17	110	127
Total	309	326	635

This might be done for instance by comparing groups per stratum, but to keep in line with the illustrative analyses of step 2, we calculated the corrected differences between treatment groups by performing a number of regression analyses: this time with 'group membership' and the 4 dummy variables indicating stratum membership as independent variables and pretreatment characteristics as dependent variables. The regression coefficients in tables 1 and 2 (with stratum membership as covariate) indicated that – on average across all strata – there were no longer significant differences in pretreatment variables. Our estimated PS seemed to balance, in a satisfactory way, the observed significant pretreatment differences between the short-term and the long-term groups. In case differences in pretreatment variables between groups are more persistent, one can try to re-estimate the PS, for instance by including interaction terms or non-linear relationships and restart at step 5.

Step 9: Effect Estimation after Correction

After taking into account the influence of known pretreatment characteristics using the PS, a corrected estimate of the treatment effect can be calculated. This can be done in different statistical ways, for instance by weighting the 5 treatment effects of the different strata. To keep in line with our analysis in step 1, we used a linear regression analysis with the GSI as the dependent variable, but this time 'group membership' and the 4 dummy variables indicating stratum membership were the independent variables. The effect of the treatment group on outcome was reduced from $\beta = 0.20$ (SE = 0.05; $p < 0.001$) before PS correction to $\beta = 0.15$ (SE = 0.06; $p < 0.05$) after PS correction. This shows that, when observed pretreatment differences were not taken into account, the treatment effect was overestimated. Stratification of the sample based on the PS reduced this bias.

Alternatives to Stratification: PS in Regression Analysis and Matching

We present the results of 2 alternative methods for adjusting a treatment effect estimation using the PS.

Regression Analysis

We performed a linear regression analysis with the GSI as the dependent variable, and the PS (as a continuous covariate) and the variable 'treatment group' as independent variables. After controlling for the PS by including it as a covariate in the regression analysis, the effect of treatment group membership was reduced from $\beta = 0.20$ (SE = 0.05; $p < 0.001$) before the correction to $\beta = 0.14$ (SE = 0.06; $p < 0.05$) after the PS correction. This is similar to the result of adjustment by stratification.

Matching

We matched each subject from the long-term group (this was the smallest group) with a subject from the short-term group, based on nearest available PS. Each subject from the short-term group only served once as matching partner for a subject from the long-term group ('sampling without replacement'). To ensure similarity in the matched pairs we used 'caliper matching' [52], i.e. all pairs with a PS difference larger than 0.10 were removed from the analysis. This meant only 179 matched pairs (358 individuals) remained in the analysis. After matching, the 2 groups showed no difference on any of the ob-

served pretreatment variables. To keep in line with our previous analyses, a regression analysis was conducted in the matched sample, with the GSI as the dependent variable, and the variable 'group membership' as the independent variable. The effect of treatment group membership was reduced from $\beta = 0.20$ (SE = 0.05; $p < 0.001$) before matching to $\beta = 0.15$ (SE = 0.07; $p < 0.05$) after matching (alternatively, a paired t test might be conducted in the matched sample). Though our matching procedure was successful in balancing and correcting for observed pretreatment differences, we lost a substantial amount of information due to a reduced sample size. In other (bigger) samples, matching might still be a useful strategy to correct for overt bias, especially when the control pool is large.

Discussion

Randomization in general and its application in psychotherapy research have been criticized by different authors for various reasons. Non-randomized studies, however, face the serious problem of selection bias. As a result, a need is felt for alternative and complementary research designs in the field of psychotherapy, like quasi-experimental designs. The PS method offers a solution to one part of the problem, overt bias, by balancing the treatment groups with regard to observed pretreatment differences. To overcome selection bias, the PS method offers advantages compared to traditional methods. First, the PS provides better insight in the selection process. Modelling treatment selection in a logistic regression analysis clarifies which variables affect selection and to what degree. Second, it is easier to match or stratify on a single score (like the PS) than on a range of pretreatment characteristics. The same holds true for regression adjustment techniques. Use of the single score PS enhances statistical power, as compared to many covariates in a regression analysis. Third, both the overlap in the distribution of the PS and balance of baseline variables after correction can be investigated and used as a descriptive tool [16]. The PS method, like any statistical correction method for selection bias, is only helpful given a considerable balance of baseline characteristics. After all, comparing very different subject groups in an outcome study is irrelevant, both scientifically and clinically. The PS helps to identify subjects differing widely on their pretreatment characteristics (and, as a consequence, on their PS). Determining the (essential) overlap of the distributions and balance with classical covariate regression analysis is cumbersome and therefore probably rarely done. As a last advantage, we would mention that the PS method can be applied in different ways (stratification, matching, in a regression analysis). Therefore, it can be tailored to sample characteristics and researchers' insights and decisions. Obviously, the PS method is not without limitations and has to be used responsibly [53]. A researcher using the PS should take into account the following recommendations. First, the PS only corrects for observed pretreatment characteristics, not for unobserved (unknown) variables, hampering true cause-effect analysis. This is called the 'ignorability' or 'no unobserved confounders' assumption. Even when using the PS carefully, results may still be biased due to unobserved variables. This is why, before starting a study, as many confounders as possible should be identified and measured in a reliable way. This reduces the risk that important variables are overlooked. It is recommended to consult several experts from both the clinical and statistical field to gain insight into the most relevant pretreatment variables. Experts' consensus and statistical relevance should guide the choice for potential confounders. Interestingly, when prognostic factors are well understood and controlled for, and inclusion/exclusion criteria are the same, randomized and nonrandomized studies can have similar outcomes [54–56]. Second, be careful when selecting variables to estimate the PS. Brookhart et al. [57] tested several ways of selecting relevant variables in a simulation study. Their findings suggest that all variables related to study outcome should be included in the PS model, whether or not these variables influence treatment assignment. In this study, we followed their advice. However, in the field there is still discussion on which is the best method for selecting the variables for the PS model [e.g. 58]. Third, the sample size of a study has to be sufficiently large, especially for stratification purposes, to allow for a meaningful correction of bias by means of the PS. Otherwise, several strata might be populated exclusively by patients with the same treatment condition, making comparison impossible. A high number of missing values on baseline variables causes problems as well. As the PS method uses a combination of many variables, just one missing variable leads to a missing PS, excluding this patient from all further analysis. Well-chosen imputation methods can be used to fill in missing values and guarantee a sufficient sample size without losing statistical precision. The availability of all essential data is the first condition for a meaningful application of the PS method, just as for any other statistical correction method. We conclude that the PS method is a powerful way of simultaneously adjusting for many observed confound-

ers in nonrandomized studies, thereby most probably reducing bias in treatment comparisons. If used in a responsible and thoughtful way, the PS method used in quasi-experimentation offers a strong research design in situations where randomization is not possible. Therefore, the PS method is a promising tool for future psychotherapy research.

Acknowledgements

Anna Bartak, Helene Andrea, Jan J.V. Busschbach and Roel Verheul are employees of the Viersprong Institute for Studies on Personality Disorders (www.vispd.nl). We would like to thank Els Havermans for her assistance in the data collection and Justus van Oel for his constructive comments on the manuscript. Furthermore, we are very grateful to Mark Powers for editing the text.

References

1 Amberson JB, McMahon BT, Pinner M: A clinical trial of sancrosyn in pulmonary tuberculosis. Am Rev Tuberc 1931;24:401–435.
2 Rubin DB: Estimating causal effects from large data sets using propensity scores. Ann Intern Med 1997;127:757–763.
3 Black N: Why we need observational studies to evaluate the effectiveness of health care. BMJ 1996;312:1215–1218.
4 Westen D, Novotny CM, Thompson-Brenner H: The empirical status of empirically supported psychotherapies: assumptions, findings, and reporting in controlled clinical trials. Psychol Bull 2004;130:631–663.
5 Leichsenring F: Randomized controlled versus naturalistic studies: a new research agenda. Bull Menninger Clin 2004;68:137–151.
6 Castonguay LG, Beutler LE: Principles of Therapeutic Change That Work. New York, Oxford University Press, 2006.
7 de Maat S, Dekker J, Schoevers R, de Jonghe F: The effectiveness of long-term psychotherapy: methodological research issues. Psychother Res 2007;17:59–65.
8 Brewin CR, Bradley C: Patient preferences and randomized clinical trials. BMJ 1989; 299:313–315.
9 Forstmeier S, Rueddel H: Improving volitional competence is crucial for the efficacy of psychosomatic therapy: a controlled clinical trial. Psychother Psychosom 2007;76:89–96.
10 Chiesa M, Fonagy P: Prediction of medium-term outcome in cluster B personality disorder following residential and outpatient psychosocial treatment. Psychother Psychosom 2007;76:347–353.
11 Shadish WR, Cook TD, Campbell DT: Experimental and Quasiexperimental Designs for Generalized Causal Inference. Boston, Houghton Mifflin, 2002.
12 Rosenbaum PR: Observational Studies, ed 2. New York, Springer Publishing, 2002.
13 Rosenbaum PR, Rubin DB: Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Assoc 1984;79:516–524.
14 Frangakis CE, Rubin DB: Principal stratification in causal inference. Biometrics 2002; 58:21–29.
15 Rubin DB, Thomas N: Matching using estimated propensity scores: relating theory to practice. Biometrics 1996;52:249–264.
16 Rosenbaum PR, Rubin DB: The central role of the propensity score in observational studies for causal effects. Biometrika 1983; 70:41–55.
17 Connors AF Jr, Speroff T, Dawson NV, Thomas C, Harrell FE Jr, Wagner D, Desbiens N, Goldman L, Wu AW, Califf RM, Fulkerson WJ Jr, Vidaillet H, Broste S, Bellamy P, Lynn J, Knaus WA: The effectiveness of right heart catheterization in the initial care of critically ill patients. Support investigators. JAMA 1996;276:889–897.
18 Lieberman E, Lang JM, Cohen A, D'Agostino R Jr, Datta S, Frigoletto FD Jr: Association of epidural analgesia with cesarean delivery in nulliparas. Obstet Gynecol 1996;88:993–1000.
19 Lytle BW, Blackstone EH, Loop FD, Houghtaling PL, Arnold JH, Akhrass R, McCarthy PM, Cosgrove DM: Two internal thoracic artery grafts are better than one. J Thorac Cardiovasc Surg 1999;117:855–872.
20 Potosky AL, Legler J, Albertsen PC, Stanford JL, Gilliland FD, Hamilton AS, Eley JW, Stephenson RA, Harlan LC: Health outcomes after prostatectomy or radiotherapy for prostate cancer: results from the prostate cancer outcomes study. J Natl Cancer Inst 2000;92:1582–1592.
21 Stenestrand U, Wallentin L: Early statin treatment following acute myocardial infarction and 1-year survival. JAMA 2001; 285:430–436.
22 Chan AW, Bhatt DL, Chew DP, Quinn MJ, Moliterno DJ, Topol EJ, Ellis SG: Early and sustained survival benefit associated with statin therapy at the time of percutaneous coronary intervention. Circulation 2002; 105:691–696.
23 Mehta RL, Pascual MT, Soroko S, Chertow GM: Diuretics, mortality, and nonrecovery of renal function in acute renal failure. JAMA 2002;288:2547–2553.
24 Wolfe F, Michaud K: Heart failure in rheumatoid arthritis: rates, predictors, and the effect of anti-tumor necrosis factor therapy. Am J Med 2004;116:305–311.
25 Gibson C: Privileging the participant: the importance of sub-group analysis in social welfare evaluations. Am J Eval 2003;24:443–469.
26 Yoshikawa H, Magnuson KA, Bos JM, Hsueh J: Effects of earnings supplement policies on adult economic and middle-childhood outcomes differ for the 'hardest to employ'. Child Dev 2003;74:1500–1521.
27 Leow C, Marcus S, Zanutto E, Boruch R: Effects of advanced course-taking on math and science achievement: addressing selection bias using propensity scores. Am J Eval 2004; 25:461–478.
28 Guo S, Barth R, Gibbons C: Propensity score matching strategies for evaluating substance abuse services for child welfare clients. Child Youth Serv Rev 2006;28:357–383.
29 Lechner M: Earnings and employment effects of continuous off-the-job training in East Germany after unification. J Bus Econ Statist 1999;17:74–90.
30 Jalan J, Ravallion M: Estimating the benefit incidence of an antipoverty program by propensity-score matching. J Bus Econ Statist 2003;21:19–30.
31 Dranove D, Lindrooth R: Hospital consolidation and costs: another look at the evidence. J Health Econ 2003;22:983–997.
32 Jung SH, Chow SC, Chi EM: A note on sample size calculation based on propensity analysis in nonrandomized trials. J Biopharm Stat 2007;17:35–41, discussion 43.
33 Kachele H, Kordy H, Richard M: Therapy amount and outcome of inpatient psychodynamic treatment of eating disorders in Germany: data from a multicenter study. Psychother Res 2001;11:239–257.
34 Robinson WL, Harper GW, Schoeny ME: Reducing substance use among African American adolescents: effectiveness of school-based health centers. Clin Psychol Sci Pract 2003;10:491–504.
35 Hill JL, Waldfogel J, Brooks-Gunn J, Han WJ: Maternal employment and child development: a fresh look using newer methods. Dev Psychol 2005;41:833–850.
36 Golkaramnay V, Bauer S, Haug S, Wolf M, Kordy H: The exploration of the effectiveness of group therapy through an internet chat as aftercare: a controlled naturalistic study. Psychother Psychosom 2007;76:219–225.

37 Viersprong Institute for Studies on Personality Disorders: SCEPTRE study on cost-effectiveness of personality disorder treatment. http://www.vispd.nl/projects.htm#Sceptre.
38 Arrindell WA, Ettema JHM: SCL-90-R: Herziene Handleiding bij een Multidimensionele Psychopathologie-Indicator. Lisse, Swets & Zeitlinger, 2003.
39 Derogatis LR: SCL-90 (R): Administration, Scoring, and Procedures Manual I for the Revised Version. Baltimore, Johns Hopkins University School of Medicine, Clinical Psychometrics Research Unit, 1977.
40 Derogatis LR: SCL-90-R: Administration, Scoring and Procedure. Manual II for the Revised Version. Townson, Clinical Psychometric Research, 1986.
41 van Kampen D: The DAPP-BQ in The Netherlands: factor structure and relationship with basic personality dimensions. J Personal Disord 2002;16:235–254.
42 Livesley WJ, Jackson DN: Manual for the Dimensional Assessment of Personality Pathology-Basic Questionnaire (DAPP-BQ). Port Huron, Sigma Press, 2002.
43 Lambert MJ, Burlingame GM, Umphress V, Hansen NB, Vermeersch DA, Clouse GC, Yanchar SC: The reliability and validity of the outcome questionnaire. Clin Psychol Psychother 1996;3:249–258.
44 Brooks R, Rabin R, de Charro F: The Measurement and Valuation of Health Status Using EQ-5D: A European Perspective. Evidence from the EuroQoL Biomed Research Programme. Dordrecht, Kluwer Academic Publishers, 2003.
45 Pfohl B, Blum N, Zimmerman M: Structured Interview for DSM-IV Personality (SIDP-IV). Washington, American Psychiatric Press, 1997.
46 DeJong CA, Van den Brink W, Harteveld FM, Van der Wielen E: Personality disorders in alcoholics and drug addicts. Compr Psychiatry 1993;34:87–94.
47 DeJong CAJ, Derks FCH, Van Oel CJ, Rinne T: Gestructureerd Interview voor de DSM-IV Persoonlijkheidsstoornissen (SIDP-IV). Sint Oedenrode, Stichting Verslavingszorg Oost Brabant, 1996.
48 Andrea H, Verheul R, Berghout CC, Dolan C, Van der Kroft PJA, Busschbach JJV, Bateman AW, Fonagy P: Measuring the core components of maladaptive personality: severity indices of personality problems (SIPP-118). Report of the Viersprong Institute for Studies on Personality Disorders (VISPD) in cooperation with the department of Medical Psychology & Psychotherapy, Erasmus University Rotterdam, The Netherlands, 2007. http://hdl.handle.net/1765/10066.
49 Verheul R, Andrea H, Berghout CC, Dolan C, Busschbach JJV, Van der Kroft PJA, Bateman AW, Fonagy P: Severity Indices of Personality Problems (SIPP-118): development, factor structure, reliability and validity. Psychol Assess 2008;20:23–34.
50 van Beek N, Verheul R: Motivation for treatment in patients with personality disorders. J Personal Disord 2008;22:89–100.
51 Cochran WG: The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics 1968; 24:295–313.
52 Quade D: Nonparametric analysis of covariance by matching. Biometrics 1982;38:597–611.
53 Yue LQ: Statistical and regulatory issues with the application of propensity score analysis to nonrandomized medical device clinical studies. J Biopharm Stat 2007;17:1–13, discussion 15–17, 19–21, 23–27 passim.
54 McKee M, Britton A, Black N, McPherson K, Sanderson C, Bain C: Methods in health services research. Interpreting the evidence: choosing between randomised and non-randomised studies. BMJ 1999;319:312–315.
55 Benson K, Hartz AJ: A comparison of observational studies and randomized, controlled trials. N Engl J Med 2000;342:1878–1886.
56 Concato J, Shah N, Horwitz RI: Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 2000;342:1887–1892.
57 Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T: Variable selection for propensity score models. Am J Epidemiol 2006;163:1149–1156.
58 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.