Selection bias in (quasi-) experimental research

Spreeuwenberg, M.D.

Publication date:
2010

Selection Bias in (Quasi-)Experimental Research

Marieke D. Spreeuwenberg

Printed by: Ridderprint BV, Ridderkerk
ISBN/EAN: 978–90–5335–312–7

Selection Bias in (Quasi-)Experimental Research

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit van Tilburg, op gezag van rector magnificus, prof.dr. Ph. Eijlander, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit op vrijdag 29 Oktober 2010 om 10.15 uur

door

Marieke Dingena Spreeuwenberg
geboren op 26 augustus 1978 te Utrecht

Promotor: Prof.dr. J.A.P. Hagenaars
Copromotor: Dr. M.A.Croon

Contents

Chapter 1

Introduction: Causes, consequences and solutions of selection bias

The central topic of this thesis is selection bias in (quasi-)experimental research. Selection bias is the bias introduced into a (quasi-)experimental study by the selection of different types of individuals into experimental program(s) and reference program(s). Consequently, the pre-existing differences between treatment programs may explain the results of a study, as opposed to true treatment effects (Heckman, 1979). The thesis is organized in three parts. The first part focusses on statistical methods dealing with selection bias due to observed differences between treatment programs (chapter 2, 3 and 4). The second part of the thesis focusses on statistical methods dealing with selection bias due to unobserved differences between treatment programs (chapter 5) and selection bias due to non-compliance of patients within treatment programs (chapter 6). The last part discusses selection bias in diagnostic testing settings (chapter 7).

The aim of this chapter is to discuss, in a very general way, the nature, causes and consequences of selection bias problems in experimental and non-experimental studies and ways to overcome these problems. Because selection bias is studied here from the viewpoint of biased causal conclusions, Rubin's precise and well known model of causality provides an excellent starting point for this discussion (Rubin, 1974). Next, a few general and traditional methods for countering selection bias will be presented. These traditional approaches will be critically evaluated, improved and extended in the remaining chapters

of this dissertation.

1.1 Randomized experiments

1.1.1 Rubin's causal model of counterfactual means

Researchers in the social, behavioral, and life sciences are conducting studies to answer questions about the effectiveness of interventions such as educational programs, social reforms, therapy programs or medications. By means of these studies, researchers try to answer questions such as "Is medication A better than medication B?", "Does obesity lead to diabetes?", "Does measure A diminish the crime rate in a particular neighborhood more than approach B?", "Is teaching style A more effective than teaching style B?".

Consider a study comparing two psychotherapy programs for patients with personality disorders. (For a real world application of this example, see chapter 2). The first program offers high intensity short-term psychotherapy (less than six months with many therapeutic sessions) and the second program offers low intensity long-term psychotherapy (more than 6 months with less contacts per week). The psychotherapeutic institution providing both therapies is interested which of the two therapy programs is most effective in reducing the psychological problems of their patients. Formulated in a causal terminology, the researchers want to know whether the short-term psychotherapy program causes a higher reduction of psychological problems compared to the long-term psychotherapy program. To answer such a causal question, the researchers should design their study in a way that make causal interpretations of the results possible.

Causality is a much debated topic, not only in philosophical, ontological and epistemological discussions, but also in more down to earth methodological and statistical disputes. Several different frameworks have been offered to handle the notion of causality in empirical research, some of them widely diverging, others at least partly overlapping. Crucial and often controversial concepts in these causal accounts are manipulation, randomization, counterfactuals, potential outcomes, structural equation models, and graphical modeling (Rubin, 1974, 1978; Holland, 1986; Robins, 1986; Pearl, 1995; MacLachlan & Krishnan, 2000; Cox & Wermuth, 2001; Lauritzen, 2001; Rosenbaum, 1995; Morgan & Winship, 2007). Perhaps the most influential contribution in the

social and behavioral sciences is Rubin's causal model of potential outcomes and his model will be used to define more precisely, what is meant by causality and selection in this dissertation (Rubin, 1978).

Rubin's causal model of potential outcomes is a counterfactual account of causality and is based on the idea of assigning each treatment program or intervention to each research unit under otherwise identical circumstances for both assignments. Each treatment program potentially affects the outcome of interest such as psychological scores on a standardized psychological test. In our example, there are two populations of interest, namely patients following the short-term psychotherapy program and patients following the long-term psychotherapy program. Obviously, the reasoning of Rubin's model is also applicable when more than two treatment programs are compared or in a classical study comparing the effect of an experimental program with a reference program (no treatment at all, standard program or placebo). The key assumption of the counterfactual framework is that each individual in the population of interest has a potential outcome under each program, even though, in practice, each patient can only be observed under one psychotherapy program at any point in time. For example, patients that completed the long-term psychotherapy program have theoretical what-if psychological outcomes in the hypothetical situation when completed the short-term psychotherapy program, and the other way around. These what-if potential outcomes are counterfactual. The potential outcomes of each individual are defined as the true values of the outcome of interest that would result from exposure to the alternative causal states. Let D denote the psychotherapy program, with value zero referring to the long-term psychotherapy program ($D = 0$) and with the value one referring to the short-term psychotherapy program ($D = 1$) (Morgan & Winship, 2007). Let Y_{id} represent the outcome score Y (response) of individual i within psychotherapy program D. The potential outcomes of each individual i are Y_{i0} and Y_{i1}. Because both Y_{i0} and Y_{i1} exist in theory for each individual, an individual causal effect (ICE), referred as δ_i, can be defined as the difference between Y_{i0} and Y_{i1} as;

$$ICE = \delta_i = Y_{i1} - Y_{i0} \tag{1.1}$$

Because a patient cannot follow both psychotherapy programs under identical circumstances at the same time, the outcome Y_{id} of an individual i can only

be observed under one and not under both psychotherapy programs. For patients in the long-term psychotherapy program (reference condition), only the outcome Y_{i0} is observed and not the outcome Y_{i1}. For patients in the short-term psychotherapy program (experimental condition), only the outcome Y_{i1} is observed and not the outcome Y_{i0}. In essence, this is a missing data problem (Rubin, 1976). The observed values for the outcome variable Y is Y_0 for patients following the long-term psychotherapy program and Y_1 for patients following the short-term psychotherapy program. The observed variable Y is therefore defined as;

$$\begin{aligned} Y &= Y_0 \text{ if } D = 0 \\ Y &= Y_1 \text{ if } D = 1 \end{aligned} \tag{1.2}$$

Because of the missing data problem, estimating the ICE, as defined in equation 1.1, is impossible. Therefore, the causal effects of the psychotherapy programs cannot be observed or directly calculated at the individual level. Therefore, one should focus on estimating the average causal effects where not the individual scores are used to estimate the effects, but the expected or mean score (yet still individual and impossible), averaged over the number of 'imaginal' independent and identical replications of the experiment. Let $E(\delta)$ denote the expected treatment effect in the population, called the average causal effect. Since the expectation of a difference is equal to the difference of two expectation the average causal effect can be defined as;

$$\begin{aligned} E(\delta) &= E(Y_1 - Y_0) \\ &= E(Y_1) - E(Y_0) \end{aligned} \tag{1.3}$$

Since the expectation of the individual causal effect ($E(\delta_i)$) is equal to the average causal effect across individuals of a population ($E(\delta)$), the subscript i has been dropped in equation 1.3.
The average causal effect for the controls (ACC) is;

$$\begin{aligned} E(\delta|D=0) &= E(Y_1 - Y_0|D=0) \\ &= E(Y_1|D=0) - E(Y_0|D=0) \end{aligned} \tag{1.4}$$

The average causal effect for the treated (ACT) is;

$$\begin{aligned} E(\delta|D=1) &= E(Y_1 - Y_0|D=1) \\ &= E(Y_1|D=1) - E(Y_0|D=1) \end{aligned} \tag{1.5}$$

In our example, the ACT is the expected what-if difference of psychological problems if one could treat a randomly selected patient in both the long-term and short-term psychotherapy programs. However, only $E(Y_0|D=0)$ and $E(Y_1|D=1)$ are observed and not $E(Y_1|D=0)$ and $E(Y_0|D=1)$. Only with the assumptions that $E(Y_1|D=0) = E(Y_0|D=0)$ and $E(Y_0|D=1) = E(Y_1|D=1)$, the average causal effect can be estimated as in equation 1.4 and 1.5, by merely subtracting mean outcome of patients in the long-term psychotherapy program from the mean outcome of patients in the short-term psychotherapy program as;

$$\begin{aligned} E(\delta) &= E(Y_1 - Y_0) \\ &= E(Y_1) - E(Y_0) \end{aligned} \tag{1.6}$$

However, to make such inferences, patients in the long-term psychotherapy (reference) program should be completely comparable in all respects to patients in the short-term psychotherapy (experimental) program, except for the received psychotherapy program. This comparability implies that some explicit assumptions should be met. These assumptions are:

1. *Stable unit treatment value assumption (SUTVA)*: Individuals do not interfere with each other; the observation of one individual should not be affected by the treatment assignment of other individuals (Cox, 1958;

Holland, 1986). There is no interference across treatments and the treatment effect does not depend on the number of individuals receiving the treatment (Morgan, 2001).

2. *Additive effect assumption*: The administration of the treatment raises the response of an individual by a constant amount.

3. *(Strongly) ignorable assumption*: The assignment mechanism is strongly ignorable if the two following assumptions are fulfilled;

   (a) The responses Y_0 or Y_1 are independent of the treatment D, given the observed variables $\boldsymbol{X}$. In formula:

   $$Y_1, Y_0 \perp D \mid \boldsymbol{X} \tag{1.7}$$

   For randomized experiments, the treatment indicator D is forced by design to be independent of the potential outcomes Y_0 and Y_1. Treatment status is therefore independent of the potential outcomes and the treatment assignment is said ignorable.

   (b) Moreover, every individual has a known probability of receiving the treatment or the reference program;

   $$0 < (P(D = 1) < 1 \tag{1.8}$$

   If the assignment is strongly ignorable, as is the case in randomized studies, it follows that the mean of the reference program can be used as an estimate of the counterfactual mean of the experimental program, and the other way around. Thus;

   $$\begin{aligned} E(Y_1|D = 0) = E(Y_1|D = 1) \text{ and} \\ E(Y_0|D = 0) = E(Y_0|D = 1) \end{aligned} \tag{1.9}$$

   If the assignment is weakly ignorable, it follows that the mean of the reference program, controlled for the observed variables $\boldsymbol{X}$, can be used as an estimate of the counterfactual mean of the experimental program. Thus;

   $$\begin{aligned} E(Y_1|D = 0, \boldsymbol{X}) = E(Y_1|D = 1, \boldsymbol{X}) \text{ and} \\ E(Y_0|D = 0, \boldsymbol{X}) = E(Y_0|D = 1, \boldsymbol{X}) \end{aligned} \tag{1.10}$$

Randomization is a wonderful way to meet the above assumptions. In randomized studies participants are assigned to the treatment programs by random procedures such as flipping a coin. That implies that, with equal group sizes, the probability of assignment to the short-term psychotherapy program, for patients agreeing to participate in the study, is .50 for each patient. With randomization it is expected that, with large sample sizes, both observed and unobserved pre-treatment variables have, on average, the same values in all treatment programs. The probability that this is actually true increases as the sample size increases. Let us consider that, in our example, patients are randomly assigned to either the short-term psychotherapy program ($D = 1$) or long-term psychotherapy program($D = 0$). Let Y_{id} represent the psychological outcome score Y of patient i within therapy D. Since patients are randomized into the therapies, one expects, especially in large sample sizes, that the two psychotherapy groups are initially comparable on pre-treatment variables such as age, gender, social economic class, initial level of depression or motivation. With initial comparability, a significant difference in the mean outcome depression scores between the two patient groups can be attributed to the psychotherapy program received. The added value of the short-term psychotherapy program to the standard long-term psychotherapy program (δ), i.e. the average causal effect for the treated (ACT), can therefore be estimated by subtracting the mean outcome of participants following the short-term psychotherapy program ($E(Y_1)$) from the mean outcome of patients following the long-term psychotherapy program ($E(Y_0)$) (Rubin, 1974).

1.1.2 Feasibility and shortcomings of randomized studies

Although randomized studies are considered the best way to achieve comparability, and to obtain unbiased estimates of the average causal effect (ACE), randomization is not always feasible or even desirable for a large number of reasons:

1. *Randomization only works well in samples that are not too small*: Randomization only realizes balance in pre-treatment characteristics with a certain probability that will never be zero but also never reaches one. In general, for very small groups, this probability will be very low. The larger the sample sizes of the treatment programs, the higher this probability. In

practice, this implies that even when the randomization procedure is carried out flawlessly, no balance of pre-treatment variables may be achieved due to small sample sizes. In some research designs, a sufficiently large number of patients that are both willing and eligible to receive treatment is very difficult or impossible to get.

2. *Randomization may be unethical*: In some research designs, it may be unethical to assign individuals at random to different treatment programs. Consider, for example, a study about the effect of cigarette smoking on lung cancer. It may be very unethical to force individuals in one program to smoke for some years and prevent individuals in the other program from smoking.

3. *Randomization may be impossible*: It is not always possible to assign patients at random to the different treatment programs. For example, it is not possible to assign individuals at random to variables that cannot be manipulated such as age, gender or to variables that occurred in the past such as previous education.

4. *Randomization may be impractical*: It is not always practical to randomize individuals into the treatment programs. For example, in a study on the effect of a new teaching method on the reading skills of children in schools, it is hardly practicable to assign children within the same class at random to different teaching methods.

5. *Randomization may be very expensive and time-consuming*: Randomized studies require extensive planning and control and may therefore be very expensive. Randomized studies may be very time-consuming and not be desirable when quick answers are needed.

6. *Randomized studies may be very different from natural situations*: The programs of randomized experiments may differ from real world situations in which the treatment is actually applied. Therefore, the results from randomized experiments cannot always be generalized to natural situations and are therefore not always the preferable research design.

7. *Randomized studies may be imperfect*: Even when randomization is perfectly carried out using a rather large number of individuals, the intended

comparability between the treatment programs might not be realized because of what happens later during the implementation of the research design. Important examples of imperfect randomized experiments are studies in which non-compliance with the experimental instructions and drop-out occur. For example, patients may refuse to take their pre-described medication or forget to take the medication on a regularly basis. In psychotherapy research, it may happen that some patients do not show up at psychotherapy sessions or decide to drop out from therapy before the end of the study. It may even happen that patients switch to another treatment because they think the other treatment has a more positive effect on them. If, as to be expected, drop-out and non-compliance are not random phenomena, the intended randomization plan fails (Shadish & Cook, 2002).

1.2 Selection bias in quasi-experiments and observational studies

From the discussion in the previous section it follows that, for a number of ethical and practical considerations, randomized studies cannot always be carried out perfectly and are not always possible or even the best choice. Therefore, for investigating the causal consequences of interventions, one often has to rely on results from non-randomized studies, also named quasi-experiments or observational studies. In the literature, the terms quasi-experiments and observational studies are often used interchangeably, but others view them as distinct research designs. Both observational studies and quasi-experiments have in common that the assignment into treatment programs is not random. However, some scholars then make a difference in the sense that in quasi-experimental studies the researcher has control on the form and content of the intervention (manipulation) and on the 'experimental environment', while in observational studies this is not the case (Rosenbaum, 1995). In this dissertation, both observational and quasi-experimental studies are treated as failed experiments, where non-random allocation of individuals into treatment programs potentially causes selection bias problems.

Since in non-randomized studies patients are not randomly assigned to the treatment and reference program(s), the individuals in the programs may dif-

fer, on average, on important pre-treatment characteristics. Variables that influence treatment assignment are often called selection variables. As a consequence non-random treatment assignment, patients groups may be non-comparable before the start of the study. Consequently, the outcomes of a study may be potentially explained because of pre-existing differences between the programs, as opposed to the treatment itself. When one or more of these pre-treatment characteristics are related to both the outcome and treatment allocation, the estimate of the treatment effect by means of the ACE becomes confounded with these selection variable(s). Such a pre-treatment variable that is associated to both assignment and outcome is called a confounder. Within Rubin"s model, based on mean differences estimation, without adjustment for this confounder variable, the ACE estimated as in equation 1.3 might be biased, i.e. unequal to the (expected) individual causal effect. This is the essence of selection bias or confounding (Anderson et al., 1980). In terms of possibilities to correct for selection bias it is important to make the distinction between overt and hidden bias (Rosenbaum, 1995). Overt bias is bias due to observed and measured variables and hidden bias due to unobserved and unmeasured variables. In general, there exists a range of statistical methods to correct for overt bias such as matching, stratification and statistical control by regression analysis. The main idea behind these correction procedures is the following; Let $\boldsymbol{X}$ denote a vector of observed pre-treatment variables. Then, the average causal effect can be estimated, controlled for the observed pre-treatment variables $\boldsymbol{X}$ as;

$$\delta = E(Y_1|\boldsymbol{X}, D = 1) - E(Y_0|\boldsymbol{X}, D = 0) \tag{1.11}$$

The concept of overt bias closely relates to the ignorability assumption. Ignorability implies that there is no bias, given the confounding variables. When all confounding variables are observed and controlled for, overt bias is dealt with in the analysis. However, when the confounding variables also include unobserved (latent) variables, even after control on the observed selection variables, bias may arise (Dehejia & Wahba, 1999).

In the next section, three traditional methods are discussed that try to achieve comparability in quasi-experimental research. All three methods control for overt bias and assume that every confounding variable is measured. They all assume that treatment assignment is ignorable and does not depend further on unmeasured variables. The methods discussed below are match-

ing, stratification and statistical equating. In chapter 2, 3 and 4, an alternative method controlling for overt bias, named the propensity score method, is discussed, illustrated and extended. In chapter 5, methods dealing with the problem of hidden bias are discussed.

1.3 Basic remedies of overt bias

1.3.1 Matching

With matching one attempts to achieve comparability by pairing each individual from the experimental program with one (or more) individuals from the reference program, with respect to the individuals' observed characteristics $\boldsymbol{X}$. As a result, in the matched data, the distribution of the observed characteristics $\boldsymbol{X}$ are equally distributed across the treatment programs. Matching can be conducted with or without replacement and individuals in the experimental condition may be paired with more than one reference individual. With matching, the mean outcome of individuals from the reference program $E(Y_0)$ that are matched with individuals from the experimental program, can be used as an estimate of the counterfactual mean of the experimental program and visa versa. Therefore, in the matched data-set, the ACE can still be estimated as in equation 1.3.

In exact matching, one pairs each individual in the experimental program with an individual from the reference program with exactly the same value on $\boldsymbol{X}$. Sometimes, however, exact matching is not possible, because there is no similar individual available in the reference program. In that case, several alternative matching methods exists. The most common matching techniques are nearest available matching and caliper matching. The essence of these methods are described below.

In nearest available matching, a match is formed by finding the closest possible similar individual in the reference program for each individual in the experimental program. Let X_i and X_j denote the score on an independent variable X for an individual in the experimental condition i and a reference individual j, respectively. An individual i is then matched with the reference individual j which has the minimum distance on (a set of) some observed variable(s) X. In formula this is denoted as;

$$\min |X_i - X_j| \tag{1.12}$$

In random-order nearest available matching, individuals are ordered on (a set of) some observed variable(s) X from highest to lowest (or visa versa). For each individual in the experimental condition, the closest reference individual is found. When a reference individual is closest to two individuals in the experimental condition, the match is formed randomly. The main problem with nearest available matching is that pairs can still differ a lot on X, because there is no restriction on the distance within matched pairs. Therefore, a variant has been developed in which only a predefined difference on a variable X is tolerated, named caliper matching. With caliper matching, only reference individuals with a predefined difference on a variable are tolerated. A pair can only be matched if the difference on X is no more than a predefined tolerance σ as;

$$|X_i - X_j| < \sigma \tag{1.13}$$

As a consequence, some individuals in the experimental condition cannot be matched with reference individuals because they differ too much. The main advantage of caliper matching is that it allows to use more reference individuals when the matches are good and less when matches are poor. Thereby, this method ensures that pairs do not differ a lot from each other. This may lead to less bias compared to nearest available matching. However, when the tolerance is small, this method requires a large number of reference individuals to find matches for each individual in the reference program (Anderson et al., 1980). When no matches are found, the individual is excluded from the analysis and no full use is made of the available data (Cochran & Rubin, 1973). There also exists some alternative matching techniques such as discriminant matching or mahalanobis distance matching. More can be read about these methods in Heckman, Ichimura, and Todd (1997) and Heckman, Ichimura, Smith, and Todd (1998). The extent to which matching leads to bias reduction depends mainly on three factors; first on the distributional overlap regarding X between the two samples, second on the ratio of the population variances and third on the size of the reference sample. With almost no overlap, reference individuals differ a lot from individuals in the experimental condition. If matches can be

formed at all, the matched pairs will differ a lot from each other. Thereby, the further the population means are separated, the larger the number of reference individuals must be to find close matches, unless the variances are such that the two population distributions overlap substantially (Anderson et al., 1980). Note that with matching, internal validity problems might occur, since only the overlapping respondents are used to estimate the treatment effect.

1.3.2 Stratification

Stratification is an alternative strategy to control for observed baseline differences. With stratification, several groups of individuals are formed, based on the same set of observed variables $\boldsymbol{X}$. As a result, within each group or strata, individuals are more or less equal on $\boldsymbol{X}$. Note that also here, the key assumption is made that $\boldsymbol{X}$ includes all confounding variables and no variables are missed. In exact stratification, the strata are homogenous in the observed characteristics in $\boldsymbol{X}$. In practice, this implies that individuals from both the reference and the experimental programs, with exactly the same pre-treatment variables, are grouped in one stratum. This is, however, only attainable when the number of covariates and/or the number of categories are low (Rosenbaum, 1995). Therefore, exact stratification is not always possible and individuals within a stratum differ more or less from each other on $\boldsymbol{X}$. The more individuals differ within a stratum, the more bias will eventually arise.

After stratification, one way to estimate average causal effects is by weighting the mean responses of combination of the individual strata differences as;

$$\delta = \frac{1}{N} \sum_{k-1}^{k} n_k (E(Y_{1k}) - E(Y_{0k})) \tag{1.14}$$

where n_k denotes the number of treatment or reference patients in the kth stratum (k=1,2,3...K) and N the total number of patients in the study. Cochran (1968) showed in a simulation study that defining five strata is often sufficient to remove 90% of the bias.

The main problem of stratification is that when the number of covariates increases, the number of strata increases exponentially and the probability of finding good matches decreases. This is called the dimensionality problem (Dehejia & Wahba, 1999). The first part of this thesis discusses and illustrates the

propensity score (PS) method. The PS can be used to overcome the dimensionality problem. With the PS method, only a single score (the PS) can be used for matching and stratification, instead of a number of covariates. In chapter 2, the PS method for two-way comparisons will be discuss and illustrated. In chapters 3 and 4, the PS method is extended to studies comparing more than two treatments.

1.3.3 Statistical equating

Regression adjustment techniques try to achieve comparability in the statistical analysis through statistical equating. Multiple regression adjustment has the advantage over matching that it uses all available data and, in theory, can be used when the distributions of the two programs do not overlap completely (Campbell, 1999). The basic multiple regression model assumes that there exists a linear relationship between the outcome variable and the covariates $\boldsymbol{X}$ with identical slopes, but possible different intercepts for both the experimental program and the reference program; the variables in $\boldsymbol{X}$ do not interact with the effect of D on Y. The linear regression equation for the treatment effect δ is equal to;

$$Y_i = \alpha + \beta \boldsymbol{X} + \delta D + \varepsilon_i \tag{1.15}$$

where $\boldsymbol{X}$ is a set of confounding variables, α the mean of the reference program where all covariates have the value of zero, δ the effect of the intervention D and ε_i the individual error term.

Estimating causal effects with standard multiple regression adjustment implies some basic assumptions such as a constant and linear effect of Y on $\boldsymbol{X}$, without interactions. However, departures from these assumptions can be dealt with in more complex models. Since, in randomized studies, one assumes that there are no pre-treatment differences between treatment programs, multiple regression adjustment seems unnecessary. However, also in randomized studies, regression analysis with covariates reduces the error variance and more efficient estimates of the effect can be obtained. Statistical equating, for example by regression adjustment, may be preferable over matching and stratification methods if the relation between the covariates and the outcome is linear or if the researcher is confident that non-linearity can easily be accounted for in the

model. Non-linear models can become, however, very complex and matching and stratification methods are then preferable (Maxwell & Delaney, 2004).

1.4 Contents of this thesis

The goals of this thesis are to compare, illustrate and present statistical methods that reduce selection bias in social, behavioral and medical research.

Chapter 2 discusses the dimensionality problem of matching and stratification in situations where the number of pre-treatment variable is large. As an alternative method, the propensity score method (PS) is discussed. The propensity score is the probability of assignment into the experimental group, given a set of pre-treatment variables. The propensity score method is illustrated step-by step with data coming from a large a Dutch research project, named the "Study on Cost-Effectiveness of Personality Disorder Treatment" (*SCEPTRE*). Since the propensity score is mainly used in two-arm studies, for illustrative purposes, the data are divided into a short-term psychotherapy program (up to six months) and a long-term psychotherapy program (more than six months), although the original treatment variable contained more categories.

The standard propensity score method has been well developed for (quasi-) experiments comparing two treatment programs. In chapter 3, the multiple propensity score method is discussed for studies comparing multiple (more than two) programs. The method is illustrated step-by step using the data from the *SCEPTRE* study, where the effectiveness of five different therapies for patients with cluster C personality disorders are compared, differing in setting and duration. This application exemplifies how to handle selection bias in more complicated, but often occurring real world research situations.

In chapter 4, the results from the large and complicated *SCEPTRE* study are discussed from a more clinical point of view. The multiple propensity score is used to compare the effectiveness of five different therapies, differing in setting and duration, for patients with cluster C personality disorders. Since the study had a repeat testing structure, the multiple propensity scores are included into a random intercept multilevel model. In this model, the results are adjusted for both dependency of the data due to repeat testing and for the confounding effect of a large number of observed pre-treatment differences

across the psychotherapy programs.

In chapter 5, attention is paid to two statistical methods that control for hidden bias in observational studies. These methods are (1) the original Heckman two-step method and (2) its extended version using Structural Equation Modeling (SEM). In four artificial data-sets, the performances of both methods are compared to the results of regression analysis and the propensity score method. In addition, the *SCEPTRE* data are used to compare and illustrate the methods.

In chapter 6, a method is presented to prevent bias due to non-compliance in randomized studies. If, as to be expected, non-compliance is not a random phenomena, the intended randomization plan fails and bias may arise. In general, with non-compliance, researchers perform either an "Intention To Treat Analysis" or an "As Treated Analysis" or both. An alternative model based on a latent class extension of the instrumental variable approach is presented.

In the final chapter 7, a different kind of selection problem is discussed that occurs in diagnostic testing, named verification bias. When the verification of true disease status by a gold standard test is performed only for a part of the sample, based on previous testing results, the estimates of the sensitivity and specificity may be biased. Data coming from a large study performed in the Netherlands are used to illustrate how to account for verification problems in a repeat testing situation.

Chapter 2

The use of propensity score methods in psychotherapy research*

2.1 Summary

Randomized controlled trials are considered the best scientific proof of effectiveness. There is increasing concern, though, about their feasibility in psychotherapy research. A quasi-experimental study design is discussed 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. Data from a Dutch research project, named "Study on Cost-Effectiveness of Personality Disorder Treatment" (*SCEPTRE*), is used as an illustrative example. The sample consisted of 749 psychotherapy patients with personality pathology. It is tested whether the PS method was useful and applicable. Differences between 2 treatment groups (short vs. long treatment duration) in pre-treatment characteristics before and after PS correction is examined. This revealed the impact of the PS on outcome differences. The PS offered statistical control over observed pre-treatment differences between patients in a non-randomized study. 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. If implemented carefully, this method

*This chapter has been published as: Bartak, A., Spreeuwenberg, M.D., Andrea, H., Busschbach, J.J.V., Croon, M.A., Verheul, R., Emmelkamp, P.M.G. & Stijnen, T. (2009). *Psychotherapy and Psychosomatics*, 78, 26–34.

is promising for future effectiveness research.

2.2 Introduction

The first randomized study in medicine was conducted by Amberson, McMahon, and Pinner (1931) 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 (Rubin, 1997).

In many cases, though, randomization may be difficult, unethical or impossible, especially in psychotherapy research (Black, 1996; Westen, Novotny, & Thompson-Brenner, 2004; Leichsenring, 2004; Castonguay & Beutler, 2006; Maat, Dekker, Schoevers, & Jonghe, 2007). Here, patients' and clinicians' personal preferences regarding treatment allocation may work against randomization. The resulting high number of excluded subjects makes the generalization of such results difficult (Brewin & Bradley, 1989). Hence, research on treatment effects in various (para)medical fields often requires well-designed and carefully conducted non-randomized studies (Forstmeier & Rueddel, 2007; Chiesa & Fonagy, 2007). Shadish and Cook (2002) 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 (Shadish & Cook, 2002). In case of non-random allocation to treatment, persons with different treatments can differ on pre-treatment characteristics. This selection bias affects the estimates of the treatment effect.

Rosenbaum (1995) distinguishes 2 types of bias: hidden bias, due to unobserved differences in pre-treatment variables, and overt bias, due to observed differences in pre-treatment variables. 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 that deal with overt bias are matching, stratification and regression adjustment (Rosenbaum, 1995; Frangakis & Rubin, 2002; Rubin & Thomas, 1996). In matching, each individual in the experimental group is paired with the most similar individual in the reference 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 pre-treatment variables are used as covariates, statistical-modeling problems and a loss of power arise. A promising alternative method to correct for overt bias is the propensity score (PS) method (Rosenbaum, 1995; Rosenbaum & Rubin, 1983).

2.3 Propensity score

Rosenbaum and Rubin (1983) suggested using the PS method to reduce the dimensionality problem. The PS method reduces the entire collection of observed pre-treatment variables ($\boldsymbol{X}$) to a single score. The estimated PS is defined as the conditional probability of assignment to a particular treatment, given a set of observed pre-treatment characteristics. Let D denote treatment group membership, where $D = 0$ denotes the reference condition and $D = 1$ denotes the experimental condition. Then, PS is defined as:

$$PS = P(D = 1|\boldsymbol{X}) \tag{2.1}$$

Rosenbaum and Rubin (1983) 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 pre-treatment 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, social sciences and economics (Connors et al., 1996; Lieberman et al., 1996; Lytle

et al., 1999; Potosky et al., 2000; Stenestrand & Wallentin, 2001; Chan et al., 2002; Mehta, Pascual, Soroko, & Chertow, 2002; Wolfe & Michaud, 2004; Lechner, 1999; Jalan & Ravallion, 2003; Dranove & Lindrooth, 2003; Gibsons, 2003; Yoshikawa, Magnuson, Bos, & Hsueh, 2003; Leow, Marcus, Zanutto, & Boruch, 2004; Guo, R, & Gibbons, 2006). The United States Food and Drug Administration recommended the PS as a tool to overcome selection bias in treatment studies (Jung, Chow, & Chi, 2007). 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 (Kachele, Kordy, & Richard, 2001; Robinson, Harper, & Schoeny, 2003; Hill, Waldfogel, Brooks-Gunn, & Han, 2005; Golkaramnay, Bauer, Haug, Wolf, & Kordy, 2007).

2.4 Aim

The aims of this study 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. The PS method is applied to a case study, the research project *SCEPTRE* (Study on Cost-Effectiveness of Personality Disorder Treatment) (Bartak et al., 2010). Two treatment groups are compared from *SCEPTRE*, using the PS to correct for known baseline differences. The two 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 chapter are easily done in common statistical packages such as SPSS.

2.5 Method

2.5.1 Participants

Patients were recruited from 6 mental health care centers 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 centers from March 2003 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 semi-structured 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 analysis, leaving 749 patients. Of these, 507 (67.7 percent) were female. The mean age was 34.24 years (SD 9.93, range 1762). This sample is divided 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).

2.5.2 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 2.1 and 2.2).

Psychiatric symptomatology was measured with the Symptom Checklist 90 Revised, Dutch version (SCL-90) (Arrindell & Ettema, 2003; Derogatis, 1977, 1986). In this study, the Global Severity Index of the SCL-90 (GSI; the mean score of all 90 items) is used as the primary outcome measure, with higher scores indicating more distress. To measure the type and degree of personality pathology the 4 higher-order factors of the Dimensional Assessment of Personality Pathology Basic Questionnaire, Dutch version (DAPP-BQ): (1) emotional dysregulation, (2) dissocial behaviour, (3) inhibition and (4) compulsivity (Kampen, 2002; Livesley & Jackson, 2002) were used. Psychosocial functioning was measured with the Outcome Questionnaire 45, Dutch version (OQ-45) (Lambert et al., 1996). Of this self-report measure, 2 subscales were included: (1) interpersonal relations and (2) social-role functioning. Health-related quality of life was assessed with the EuroQoL EQ-5D (Brooks, R, & Charro, 2003). Personality disorders were assessed with the Structured Interview of DSM-IV Personality, Dutch version (SIDP-IV) (Pfohl, Blum, & Zimmerman, 1997; DeJong, Brink, Harteveld, & Wielen, 1993; DeJong, Derks, Oel, & Rinne, 1986). The severity of personality pathology was measured with 5 higher-order domains of the Severity Indices of Personality Problems (SIPP): self-control, social concordance, identity integration, relational functioning and responsibility (Andrea et al., 2007; Verheul et al., 2008). To measure patients

motivation for treatment, the two scales of the Motivation for Treatment Questionnaire (MTQ-8): need for help and readiness to change (Beek & Verheul, 2008) were used.

2.6 Results

2.6.1 Results of the case study

To avoid bias in the estimation of the treatment effect, the influence of known pre-treatment differences was corrected. This was done by stratification of the sample based on the PS. This process took 9 steps, described below.

Step 1: Effect estimation before correction

Before correction for known pre-treatment differences, the treatment effect is estimated 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

The 2 treatment groups were compared on pre-treatment variables before stratification. Note that this step is neither relevant for variable selection for the PS, nor for further analysis. 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, a comparison of overall regression coefficients is shown. A number of regression analysis are conducted with group membership as an independent variable and pre-treatment characteristics as dependent variables (linear regression analysis for continuous variables, see table 2.1 , and multinomial logistic regression analysis for categorical variables, see table 2.2). 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.

Table 2.1: 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 versus 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. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 2.2: Differences in categorical variables between short-term and long-term treatment groups

Variable	Demographic data, %		Odds ratio treatment duration (short versus 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 /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.

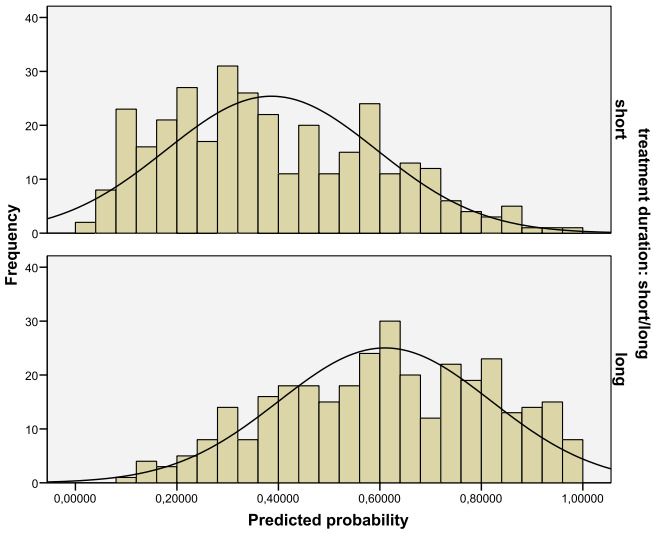

Figure 2.1: Overlap op the PS in the two treatment groups (short/long)

Step 3: Variable selection for PS estimation

To estimate the PS, all baseline variables related to outcome (GSI) are used. To identify these variables, a number of linear regression analysis are conducted 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 behavior 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 figure 2.1). 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, see table 2.3) (Cochran, 1968). 4 dummy variables were created based on these 5 groups.

Table 2.3: 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

Step 8: Balance check after correction

To know if the stratification of the sample based on the PS resulted in a balance of pre-treatment variables between the 2 treatment groups, differences in pre-treatment variables were checked again. This might be done for instance by comparing groups per stratum, but to keep in line with the illustrative analysis of step 2, the corrected differences between treatment groups was calculated by performing a number of regression analysis: this time with group membership and the 4 dummy variables indicating stratum membership as independent variables and pre-treatment characteristics as dependent variables. The regression coefficients in tables 2.1 and 2.2 (with stratum membership as covariate) indicated that - on average across all strata - there were no longer significant differences in pre-treatment variables. The estimated PS seemed to balance, in a satisfactory way, the observed significant pre-treatment differences between the short-term and the long-term groups. In case differences in pre-treatment 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 pre-treatment 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 the analysis in step 1, a linear regression analysis was used 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 pre-treatment differences were not taken into account, the treatment effect was overestimated. Stratification of the sample based on the PS reduced this bias.

2.6.2 Alternatives to stratification: PS in regression analysis and matching

The results of 2 alternative methods for adjusting a treatment effect estimation using the PS are presented below.

Regression analysis

A linear regression analysis was performed 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

Each subject from the long-term group (this was the smallest group) was matched 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, caliper matching was used, i.e. all

pairs with a PS difference larger than 0.10 were removed from the analysis (Quade, 1982). This meant only 179 matched pairs (358 individuals) remained in the analysis. After matching, the 2 groups showed no difference on any of the observed pre-treatment variables. To keep in line with the previous analysis, 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 the matching procedure was successful in balancing and correcting for observed pre-treatment differences, a substantial amount of information was lost 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.

2.7 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 pre-treatment differences. To overcome selection bias, the PS method offers advantages compared to traditional methods.

First, the PS provides better insight in the selection process. Modeling 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 pre-treatment 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 (Rosenbaum & Rubin, 1983). The PS method, like any statistical correction

method for selection bias, is only helpful given a considerable balance of pre-treatment variables. 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 pre-treatment 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, it is mentioned that the PS method can be applied in different ways (stratification, matching and 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 (Yue, 2007). A researcher using the PS should take into account the following recommendations.

First, the PS only corrects for observed pre-treatment 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 fields to gain insight into the most relevant pre-treatment 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 non-randomized studies can have similar outcomes (McKee et al., 1999; Benson & Hartz, 2000; Concato, Shah, & Horwitz, 1968).

Second, be careful when selecting variables to estimate the PS. Brookhart et al. (2006) 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, their advice was followed. However, in the field there is still discussion on which is the best method for selecting the variables for the PS model (Austin, Grootendorst, & Anderson, 2007).

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.

To conclude, the PS method is a powerful way of simultaneously adjusting for many observed confounders in non-randomized 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.

Chapter 3

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*

3.1 Summary

The propensity score method (PS) has proven to be an effective tool to reduce bias in non-randomized studies, especially when the number of (potential) confounders is large and dimensionality problems arise. The PS method introduced by Rosenbaum and Rubin is described in detail for studies with two treatment options. Since in clinical practice one is often interested in the comparison of multiple interventions, there was a need to extend the PS method to multiple treatments. It has been shown that, in theory, a multiple PS method is possible. So far, its practical application is rare and a practical introduction lacking. A practical guideline to illustrate the use of the multiple PS method is provided with data from a mental health study. The multiple PS is estimated

*This chapter has been published as: Spreeuwenberg, M.D., Bartak, A., Croon, M.A., Hagenaars, J.A., Busschbach, J.J.V., Andrea, H., Twisk, J., Stijnen, T. (2010). *Medical care*, 48(2), 166-174.

with a multinomial logistic regression analysis. The multiple PS is the probability of assignment to each treatment category. Subsequently, to estimate the treatment effects while controlling for initial differences, the multiple PSs, calculated for each treatment category, are included as extra predictors in the regression analysis. With the multiple PS method, balance was achieved in all relevant pre-treatment variables. The corrected estimated treatment effects were somewhat different from the results without control for initial differences. The results indicate that the multiple PS method is a feasible method to adjust for observed pre-treatment differences in non-randomized studies where the number of pre-treatment differences is large and multiple treatments are compared.

3.2 Introduction

Results from randomized controlled trials (RCTs) are considered the highest level of scientific evidence, since one can expect that with randomization, on average, all patient characteristics are balanced between treatment groups. Theoretically, this implies that, after randomization, treatment effects can be directly estimated, without control for initial differences (Shadish & Cook, 2002). Nevertheless, randomized study designs have the drawback that they are often difficult to conduct in clinical practice. Not only do medical ethical committees often object to randomization, random allocation is also often hampered by both clinicians and patients preferences (Maat et al., 2007; Westen et al., 2004; Mosis, Dieleman, Stricker, Lei, & Sturkenboom, 2006). In particular, when differences in treatment options are substantial, the intended randomization plan either fails or leaves the researcher with small research samples. Therefore, in most research fields, one regularly has to rely on results from non-randomized studies, also called quasi-experimental designs (Shadish & Cook, 2002).

In quasi-experimental designs, owing to non-random allocation, possible differences between pre-treatment variables of patients in treatment groups can lead to bias in the estimated treatment effect, also called selection bias or confounding (Winship & Mare, 1992). Rosenbaum (1995) distinguished between overt bias and hidden bias. Overt bias is bias owing to observed pre-treatment differences and hidden bias to unmeasured and unobserved differences (Rosenbaum, 1991). Traditionally, for two treatment comparisons, overt bias is con-

trolled statistically by means of regression analysis, matching or stratification. When many variables are present to match or stratify, however, it is impossible to find patients who are similar in terms of all these variables. This is called the dimensionality problem (Rosenbaum, 1995; DÁgostino, 1998). Moreover, the number of covariates one can afford in a regression model is limited and depends strongly on the number of observations. Therefore, there is a need to find statistical methods that are able to control for many pre-treatment characteristics. For two-way comparisons (for instance, placebo versus a new treatment) the propensity score (PS) has been described as a valid solution (Bartak et al., 2009; Rubin, 1974; Thomas, 1992). The PS method can be extended to multiple comparisons (e.g. treatment A, B and Placebo). For multi-valued treatments, Imbens (2000) suggested the use of multiple or generalised PS. Although, theoretically, the multiple PS has proven effectiveness, the method is not often encountered in clinical practice. In this chapter, it is illustrated how the multiple PS method can be used. The multiple PS method is demonstrated step-by-step with data from a mental health study.

3.3 The (multiple) propensity score method

With the PS method, a large collection of observed pre-treatment variables can be used to estimate a single score, the PS (Rosenbaum & Rubin, 1983). This score is the probability of assignment to the experimental condition, given a set of pre-treatment variables. In randomized studies, the PS is supposed known (mostly 50 percent). In non-randomized studies, the PS score can, for example, be estimated by a logistic or probit regression analysis. With the assumption of ignorability, meaning that $\boldsymbol{X}$ includes all important pre-treatment variables, it can be shown that treatment assignment and covariates are independent, given the PS (Rubin, 1974, 1997, 1976). This implies that control on the PS through regression adjustment, matching or stratification removes the bias associated with the differences in observed pre-treatment differences (DÁgostino, 1998; Bartak et al., 2009; Thomas, 1992; Heckman et al., 1997; Ho, Imai, King, & Stuart, 2007; Lu, Zanutto, Hornik, & Rosenbaum, 2001; Morgan & Harding, 2006; Rubin, 1997). It has been proven that the PS is a balancing score since, after control on the PS, the distribution of the covariates is assumed the same for the experimental group and the reference group. Accordingly, with the PS

method one can easily check this balancing effect and advise the investigator whether the causal question can be answered by the data at hand (Rubin, 1997).

Until recently, the PS method has been mostly used for two treatment settings. In many cases, however, one might be interested in the comparison of more than two treatments. Rubin proposed to create separate PS models for each paired treatment comparison (Rubin, 1997; Luellen, Shadish, & Clark, 2005). If these models are not constrained, however, the probability of choosing all treatment arms will end up greater than 1. In addition, the parameter estimates obtained in these separate models are less efficient than those obtained by fitting models simultaneously within a multinomial regression model (Agresti, 2002). For nominal treatments, as is often the case in mental health research, Imbens (2000) suggested the use of multiple PS, defined as the conditional probability of receiving a particular level of the treatment given a set of observed pre-treatment variables. The multiple PS can be estimated with a multinomial logistic or probit regression. Here, for each subject, the probability of receiving each treatment category given the observed covariates is estimated. Imbens (2000) proved theoretically that, just like the original PS, the multiple PS is a balancing score and that, instead of conditioning on the entire set of covariates $\boldsymbol{X}$, it is sufficient to condition on the multiple PS. Therefore, the multiple PS can be used to correct for initial baseline differences and leads to valid estimates in multiple treatment comparisons. Simulation studies using subclassification in the multiple propensity score support this finding (Imai & Dyk, 2004). Note that the assumption of ignorability is crucial in this aspect, since it is assumed that all possibly confounding variables are observed and used in the multiple PS estimation (Heitjan & Rubin, 1991). Recently, a few studies have used the multiple PS for matching and subclassification (Frisco, Muller, & Frank, 2007; Zanutto, Lu, & Hornik, 2005). Matching and stratification on many multiple PSs are difficult to conduct in clinical practice, however, especially when the number of treatments compared is large, and may result in very small groups. Therefore, in this study, a step-by-step application of the multiple PS method using regression analysis is presented.

3.4 Aim

The aims of this chapter are (1) to investigate if the multiple 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 multiple PS in comparative outcome studies when randomization is unfeasible. The multiple PS method is applied to a case study, the research project *SCEPTRE* (Study on Cost-Effectiveness of Personality Disorder Treatment) (Bartak et al., 2009). 5 treatment groups from *SCEPTRE* are compared, using the multiple PS to correct for known baseline differences. Results should only be interpreted as an illustration, not as a relevant clinical message.

3.5 Methods

3.5.1 Participants

To illustrate the use of a multiple PS, a sample of 361 patients is used, who all enrolled in different forms of psychotherapy in six mental health care institutes in the Netherlands. Patients were divided into five therapy groups, which differed in treatment duration (up to six months (short) or more than six months (long)), and treatment setting (outpatient, day hospital or inpatient). The five therapies were long outpatient, short day hospital, long day hospital, short inpatient, and long inpatient treatment.

3.5.2 Measures

The baseline assessment included measurements for all variables that were identified as potential confounders of the treatment-outcome association, i.e. age, gender, 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 DSM-IV Axis II cluster A disorders, number of DSM-IV Axis II cluster B disorders, number of DSM-IV Axis II cluster C disorders, and psychological capacities. This list of variables had been carefully chosen by both clinicians and researchers, and was based on the existing literature and clinical knowledge. Psychiatric symptomatology was measured

with the Global Severity Index (GSI) and used as primary outcome measure (Arrindell & Ettema, 2003; Derogatis, 1986). Three treatment institutes conducted their follow-up measures at 12, 24, and 36 months after baseline. The three remaining treatment institutions conducted their follow-up measures at the end of treatment, 6 and 12 months after the end, and again at 36 months after baseline. For specific details of this study the reader is referred to the literature (Bartak et al., 2009, 2010). For illustrative simplicity, the mean GSI score of all follow-up measures is used as primary outcome measure.

3.6 Statistical analysis and results

The analyses were done with SPSS for Windows, version 15.0 (SPSS Inc., Chicago, IL, USA). The multiple PS was applied to the data in the following steps.

Step 1: Effect estimation before correction

In the naïve model, the treatment effects were estimated in a multiple regression analysis without any correction for pre-treatment differences. The dependent variable was the GSI outcome score. As independent variables four dummy variables indicating treatment group membership were included, with the short inpatient treatment as reference category. See table 3.1 for the estimated pairwise treatment effects in the naïve model. Without control for initial baseline differences the mean GSI score in the short inpatient treatment was lower than the GSI score in the short day hospital ($p < 0.05$), long outpatient treatment ($p < 0.05$), the long day hospital treatment ($p < 0.05$) and the long inpatient treatment ($p < 0.05$). No other significant differences between the treatment groups were found.

Step 2: Balance check before correction

First is checked to what extent the five treatment groups differed initially. Note that this is not relevant for variable selection for the multiple PS, nor for further analysis; the reason why it is done here is that it gives us an idea of the initial comparability between the five treatment groups. For each continuous variable an ANCOVA is conducted with treatment group as fixed factor. For the dichotomous variables a logistic regression analysis is conducted with the

Table 3.1: Uncorrected estimated differences in treatment effects between the 5 treatment groups

	Long outpatient		Short day hospital		Long day hospital		Short inpatient	
Treatments	β	95% CI	β	95% CI	β	95% CI	β	95% CI
Long outpatient	Reference							
Short day hospital	-0.08	-0.28-0.13						
Long day hospital	-0.02	-0.23-0.19	0.06	-0.15-0.26				
Short inpatient	-0.27*	-0.49 to -0.05	-0.19*	-0.41 to -0.02	-0.25*	-0.47 to -0.03		
Long inpatient	-0.02	-0.22–0.18	0.05	-0.14-0.24	0.00	-0.20-0.19	0.25*	0.04-0.45

$* \; p < 0.05$, $** \; p < 0.01$, $*** \; p < 0.001$.
β indicates unstandardized regression coefficient; CI, confidence interval.

categorical treatment variable as independent variable. For nominal variables, a multinomial logistic regression analysis is conducted and the significance of the log-likelihood ratio test for treatment reported. See table 3.2 for the mean and standard deviation of the continuous variables in each treatment group and the p-values for significance before correction on the multiple PS. See table 3.3 for the percentages of levels of the categorical variables in each group and the p-values for significance. Sixteen out of 24 distributions of the continuous variables differed between the groups. For the categorical variables 8 out of 16 variables differed between the five treatment groups. This implies that, without correction for these differences, the five treatment groups initially differed in many pre-treatment variables and were not comparable in many ways.

Step 3: Variable selection for multiple PS estimation

As suggested, all baseline variables related to outcome were used for estimating the multiple PS (Lu et al., 2001; Brookhart et al., 2006). To identify these variables, several linear regression analysis are conducted with the GSI outcome score as dependent and each potential confounder as independent variable. All variables with a p-value smaller than 0.10 were selected for the estimation of the multiple PS. These variables are denoted with † in tables 3.2 and 3.3.

Step 4: Multiple PS estimation

Since in this study the treatment categories are nominal, the multiple PSs are estimated by multinomial regression analysis with all variables related to outcome as independent variables and group membership as dependent variable. The likelihood ratio test of the model, compared with the "empty" model was $\chi^2 = 220.20, df = 96, p < 0.001$. The pseudo R^2 of Nagelkerke was 45.7%. The multiple PSs are the estimated predicted probabilities of assignment to each treatment group, calculated for each subject. Because in this study five psychological treatments are compared, five multiple PSs are estimated as suggested by Imbens (2000). Since all these PSs add up to 1 and are complementary, only four out of five multiple PSs are needed in the further analysis. Note that a main assumption of multinomial regression analysis is the Independence

Table 3.2: Means and standard deviations of the continuous variables in the treatment groups and p-values for the differences before and after correction on the multiple PS

	Demographic data, mean±SD					P-value	
	Long outpatient (N=64)	Short day hospital (N=76)	Long day hospital (N=72)	Short inpatient (N=58)	Long inpatient (N=91)	before Multiple PS correction	after Multiple PS correction
Age, years	36.91 ±8.9	35.11 ±9.2	31.99 ±10.2	37.72 ±9.3	28.35 ±6.7	0.000*	0.000*
Personality pathology (DAPP-BQ)							
Emotional dysregulation†	23.53 ±3.7	22.50 ±3.7	23.82 ±4.2	23.52 ±3.7	24.10 ±3.7	0.093	1.000
Dissocial behaviour†	17.85 ±4.9	17.42 ±4.0	18.40 ±4.9	17.11 ±4.3	17.80 ±4.1	0.057	1.000
Inhibitedness†	22.27 ±4.3	22.78 ±4.6	22.29 ±4.6	24.08 ±4.6	24.79 ±4.9	0.001*	0.999
Compulsivity	26.82 ±7.2	24.53 ±6.6	24.46 ±6.9	26.78 ±6.0	25.75 ±7.1	0.117	0.123
Motivation (MTQ-8)							
Need for help†	27.21 ±5.1	28.03 ±6.1	29.61 ±4.5	31.07 ±3.8	30.45 ±3.5	0.000*	0.995
Readiness to change	28.31 ±5.5	29.00 ±6.1	31.01 ±4.8	31.88 ±3.9	30.21 ±5.1	0.001*	0.034*
Quality of life (EQ-5D)†	0.58 ±0.02	0.60 ±0.03	0.50 ±0.03	0.60 ±0.03	0.50 ±0.03	0.031*	0.999
Psychological capacities (SIPP)							
Self-control†	4.32 ±1.0	4.53 ±0.9	4.11 ±0.9	4.57 ±0.9	4.53 ±0.8	0.011*	0.999
Social concordance†	5.26 ±0.9	5.59 ±0.8	5.36 ±0.8	5.70 ±0.8	5.50 ±0.7	0.013*	1.000
Identity integration†	3.34 ±0.7	3.49 ±0.7	3.18 ±0.7	3.17 ±0.6	3.11 ±0.6	0.001*	1.000
Relational functioning†	3.57 ±0.8	3.91 ±0.8	3.68 ±0.8	3.57 ±0.8	3.48 ±0.7	0.008*	0.999
Responsibility†	4.67 ±0.9	4.63 ±0.8	4.52 ±0.9	4.64 ±0.8	4.47 ±0.9	0.538	1.000
GSI (SCL-90)	2.50 ±0.7	2.44 ±0.6	2.67 ±0.6	2.75 ±0.5	2.75 ±0.7	0.006*	1.000
Functioning (OQ-45)							
Symptom Distress	54.94 ±14.0	51.81 ±13.1	58.02 ±13.0	59.58 ±11.5	58.26 ±13.7	0.003*	1.000
Social role functioning†	15.82 ±4.1	15.22 ±4.6	16.73±4.7	17.85 ±3.8	16.89 ±4.6	0.007*	1.000
Interpersonal functioning†	54.94 ±14.0	51.81 ±13.1	58.02 ±13.0	59.58 ±11.5	58.26 ±13.7	0.003*	1.000
Axis-II diagnosis (SIDP-IV)							
No. Axis-II cluster A† disorders	0.13 ±0.33	0.07 ±0.25	0.15 ±0.40	0.03 ±0.18	0.10 ±0.30	0.187	1.000
No. Axis-II cluster B† disorders	0.47 ±0.71	0.34 ±0.56	0.38 ±0.62	0.21 ±0.64	0.35 ±0.55	0.220	1.000
No. Axis-II cluster C disorders	1.31 ±0.5	1.36 ±0.6	1.35 ±0.6	1.29 ±0.6	1.42 ±0.6	0.736	0.958
Dimensional score cluster A†	10.90 ±7.2	9.04 ±6.4	8.71 ±7.7	6.55 ±6.4	9.21 ±6.9	0.017*	0.998
Dimensional score cluster B†	18.80 ±13.8	16.39 ±11.2	16.60 ±11.1	10.65 ±10.4	16.91 ±10.2	0.002*	0.994
Dimensional score cluster C	27.34 ±7.5	29.33 ±8.8	27.39 ±8.3	27.49 ±7.3	28.50 ±9.0	0.522	1.000
Total dimensional score	79.43 ±27.1	79.30 ±24.8	75.54 ±27.1	64.44 ±5.3	77.82 ±23.9	0.006*	0.995

Values are presented as means ± SD. * $p < 0.05$. PS indicates propensity score
†Variables related to outcome with a $P < 0.10$.

Table 3.3: Percentages of levels of the categorical variables in the treatment groups and P-values for the differences before and after correction on the multiple PS

						P-value	
	Long outpatient (N=64)	Short day hospital (N=76)	Long day hospital (N=72)	Short inpatient (N=58)	Long inpatient (N=91)	before Multiple PS correction	after Multiple PS correction
Gender						0.080	0.435
Female	62.5	77.6	79.2	63.8	65.9		
Male	37.5	22.4	20.8	36.2	34.1		
Civil status						0.000*	0.000*
Married	28.1	21.1	15.5	33.7	11.0		
Widowed/divorced	20.3	14.5	10.7	8.4	2.2		
Never married	51.6	64.5	73.8	57.8	68.8		
Living situation						0.000*	0.001*
Alone	32.8	28.9	36.1	46.6	41.8		
With partner	48.4	43.4	36.1	44.8	18.7		
With child without partner	9.4	9.2	5.6	1.7	1.1		
With parent(s)	6.2	9.2	13.9	3.4	22.0		
With other people	3.1	9.2	8.3	3.4	16.5		
Childcare						0.000*	0.006*
No care for children	61.4	72.4	83.3	77.6	93.4		
Care for children	35.9	27.6	16.7	22.4	6.6		
Work situation†						0.474	1.000
Unemployed	40.6	40.8	27.8	37.9	35.2		
Study or paid work	59.4	59.2	72.2	62.1	64.8		
Level of education						0.318	0.814
Low	22.6	26.3	25.2	15.8	17.6		
Middle	23.4	22.4	21.5	14.0	16.5		
High	50.0	51.3	53.3	70.2	65.9		
Previous outpatient treatment†						0.001*	0.003*
No	31.3	11.8	20.8	13.8	5.5		
Yes	68.8	88.2	79.2	86.2	94.5		
Previous inpatient treatment†						0.360	0.999
No	81.3	86.8	84.7	74.1	79.1		
Yes	18.8	13.2	15.3	25.9	20.9		
Previous medication						0.010*	0.195
No	61.4	55.3	62.5	36.2	47.3		
Yes	35.9	44.7	37.5	63.8	52.7		
Alcohol abuse						0.067	0.046*
No	93.7	85.5	73.6	84.5	89.0		
Yes	6.3	14.5	26.4	15.5	11.0		
Drug abuse						0.047*	0.385
No	84.5	84.2	76.4	91.4	72.5		
Yes	15.6	15.8	23.6	8.6	27.5		
Preference setting†						0.000*	0.000*
Outpatient	53.1	6.6	8.3	1.7	1.1		
Day hospital	10.9	61.8	54.2	20.9	20.9		
Inpatient	1.6	14.5	15.3	61.5	61.5		
Do not know	34.3	34.4	22.2	16.5	16.5		
Preference duration†						0.000*	1.000
Up to 6 months	12.5	36.8	26.9	26.4	30.8		
$\geq$ 6 months	40.6	28.9	21.2	58.5	44.0		
Do not know	46.9	34.2	51.9	44.0	25.3		
Diagnosis avoidance						0.070	0.638
No	43.8	46.1	36.1	32.8	26.4		
Yes	56.3	53.9	63.9	67.2	73.6		
Diagnosis dependent						0.340	0.906
No	85.9	72.4	76.4	81.0	74.7		
Yes	14.1	27.6	23.6	19.0	25.3		
Diagnosis obsessive compulsivity						0.160	0.831
No	39.1	46.1	52.8	56.9	57.1		
Yes	60.9	53.9	47.2	43.1	42.9		

* $P < 0.05$. PS indicates propensity score. †Variables related to outcome with a $P < 0.10$.

of Irrelevant Alternatives Assumption (IIA) which means that adding irrelevant outcome categories does not affect the odds ratio among the remaining outcomes. With the module mlogtest of the computer package STATA this assumption is checked. In our case, adding irrelevant outcome categories did not influence the odds of treatment. When, however, the IIA assumption is violated, multinomial probit analysis can be used. In the case when treatment categories are defined by an ordinal value such as treatment dosage, ordinal logistic regression can be used as an alternative estimation method (Lu et al., 2001; Joffe & Rosenbaum, 1999; Wang, Donnan, Steinke, & MacDonald, 2001).

Step 5: Check for overlap of the distributions

Before the multiple PSs are included in the regression analysis, it is advisable to inspect the distributions of the multiple PSs, as non-overlapping distributions will make the analysis like comparing apples and oranges. That is, for treatment comparability, it is important that each patient in a therapy group also had a certain probability of assignment to the other therapy groups. A lack of overlap between the distributions of the multiple PSs can yield imprecise estimates of the treatment effect that is only applicable for a subgroup. Figure 3.1 shows the distributions of the multiple propensity scores. In the comparison of the ranges of the multiple propensity scores for subjects assigned to each treatment group there is considerable overlap. For the comparison in the overlap of two distributions Cochran and Rubin (1973) defined a distance score (d) where the value depends on the mean and the variance in two distributions. This method can be used for each pairwise comparison.

Step 6: Balance check after correction

The use of the multiple PS is considered successful when balance is achieved in the distribution of all observed covariates between the five treatment groups. The similarity of the covariates can be assessed with significance testing. For each continuous variable an ANCOVA is conducted with treatment group as fixed factor. To correct the comparison for the PS four out of five multiple PSs are added along with its mutual interactions as covariates. Table 3.2 shows the p-values for significance testing of the treatment groups differences before and after correction on the multiple PS. For the dichotomous variables a logistic regression analysis is carried out with the categorical treatment variable along

Table 3.4: Estimated differences in treatment effects between the 5 treatment groups after correction on the multiple PS

	Long outpatient		Short day hospital		Long day hospital		Short inpatient	
Treatments	β	95% CI	β	95% CI	β	95% CI	β	95% CI
Long outpatient	Reference							
Short day hospital	0.01	-0.21-0.24						
Long day hospital	0.01	-0.22-0.24	-0.01	-0.22-0.21				
Short inpatient	-0.28*	-0.53 to -0.03	-0.29*	-0.53 to -0.05	-0.26*	-0.52 to -0.05		
Long inpatient	-0.04	-0.26–0.18	-0.05	-0.26-0.15	-0.05	-0.25-0.16	0.24*	0.02-0.46

* $p < 0.05$.

β indicates unstandardized regression coefficient; CI, confidence interval.

Table 3.5: Estimated differences in treatment effects between the 5 treatment groups after correction on all variables relating to outcome

	Long outpatient		Short day hospital		Long day hospital		Short inpatient	
Treatments	β	95% CI	β	95% CI	β	95% CI	β	95% CI
Long outpatient	Reference							
Short day hospital	0.03	-0.18-0.23						
Long day hospital	-0.01	-0.21-0.19	-0.04	-0.23-0.16				
Short inpatient	-0.27*	-0.50 to -0.04	-0.30*	-0.53 to -0.06	-0.26*	-0.47 to -0.05		
Long inpatient	-0.07	-0.27–0.13	-0.10	-0.28-0.09	-0.06	-0.24-0.13	0.20*	0.01-0.40

* $p < 0.05$

β indicates unstandardized regression coefficient; CI, confidence interval.

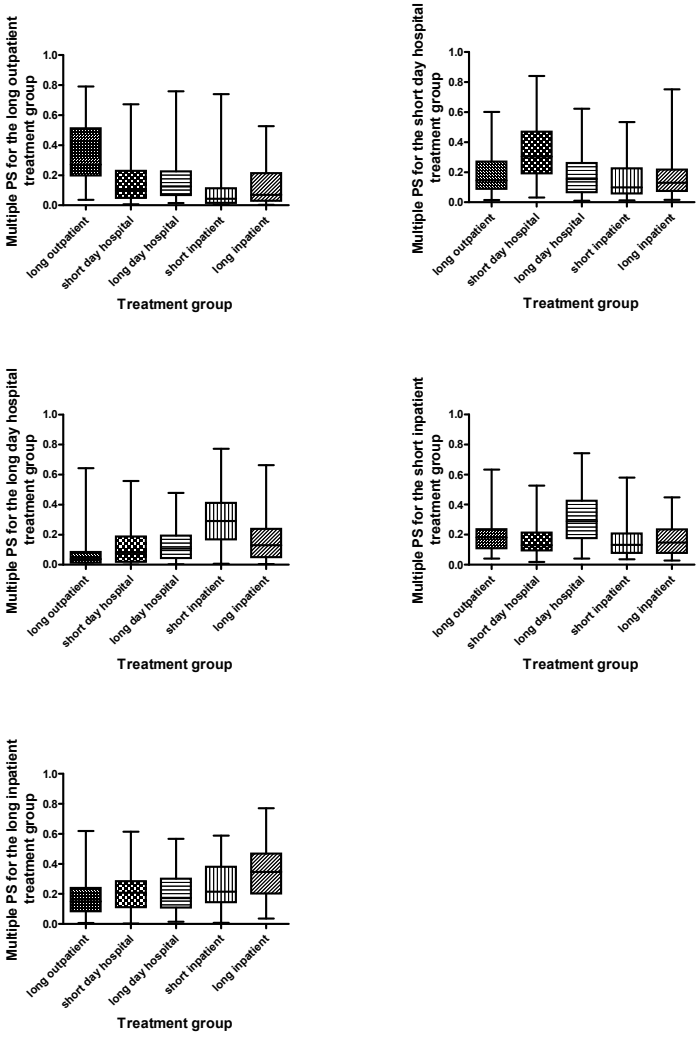

Figure 3.1: Box-plots for overlap of the multiple PS between the 5 treatments

with four multiple PSs as independent variables. For nominal variables multinomial logistic regression analysis is used with treatment as factor treatment and the four multiple PSs as covariates. Table 3.3 shows p-values for significance of treatment group differences for the categorical variables before and after correction on the multiple PS. Tables 3.2 and 3.3 reveal that control on the multiple PS satisfactorily balances the distribution of all covariates used for estimating the PS. After control on the multiple propensity score there was no difference between the variables included in the multiple PS. This implies that further analysis is possible.

Step 7: Effect estimation after correction

To estimate the treatment effect, taking into account the influence of pretreatment characteristics, the naïve model was extended by including the multiple propensity scores in the model. The GSI score was used as dependent variable and as independent variables the following covariates were included: four dummy variables indicating group membership, four multiple PSs and their product terms. See table 3.4 for each estimated pairwise treatment effect, after correction on the multiple PS. In accordance with the naïve model, the mean GSI score in the short inpatient treatment was lower than the GSI score in the short day hospital treatment ($p < 0.05$), long outpatient treatment ($p < 0.05$), the long day hospital treatment ($p < 0.05$) and the long inpatient treatment ($p < 0.05$). No other differences between the treatment groups were found. In comparison of the corrected and uncorrected treatment effects, it is seen that the uncorrected treatment effects were only slightly different (see tables 3.1 and 3.4). This implies that the role of overt bias was only small in this study. An explanation is that the variables included in the study were not strong confounders. As the propensity score method often yields the same results as traditional multiple regression analysis, also a traditional multiple regression analysis is done with all variables that were originally included in the multiple PS as extra predictors in the naïve model. The results are presented in table 3.5. As can be seen, the estimates from this analysis are comparable to the estimates provided by the multiple propensity scores.

3.7 Discussion

The present study introduces the multiple PS methodology by presenting a practical step-by-step approach using data from a mental health study. The study results indicate that the multiple PS can correct for observed pre-treatment differences, thereby reducing the influence of selection bias in non-randomized studies. The results indicate a superiority of the short-term inpatient treatment.

As Bartak et al. (2010) suggested, the superiority of the short-term inpatient treatment can be explained by the combination of short hospitalization, thereby preventing iatrogenic effects, and a high level of therapeutic intensity and pressure. This makes inpatient psychotherapeutic treatment an interesting option for patients with Cluster C personality disorder.

Even though the multiple PS is a strong tool for correction of initial differences, one has to keep in mind several considerations while using it. First, one has to take into account that the true multiple PSs remain unknown. Only when balance of confounders between treatment groups is observed, a researcher knows that the multiple PS succeeded in controlling for bias and further analysis can take place (Ho et al., 2007).

Second, the (multiple) PS, like virtually all statistical methods used for observational data, relies strongly on the ignorability assumption and thus on the assumption that hidden bias is absent. Researchers are often unaware of the presence of the influence of unobserved variables on the results of their study and can therefore not fully rely on the results. To reduce the risk of hidden bias, it is important to choose carefully a list of potential confounders that should be measured before the start of the study. In the present study, expert panels from both the clinical and statistical fields were used to identify possible confounders.

Third, it is emphasized that, although the multiple PS step-by-step approach seems straightforward, one has to keep in mind that the method should be used with care. This applies especially to the selection of variables included in the multiple PS. Brookhart et al. (2006) state that variables which are only related to treatment assignment should not be included and all variables related to outcome should be included in the estimation of the (multiple) PS. Rubin and Thomas (1996) state that no prognostic variable should be left out

unless this variable is clearly balanced and that it is advisable to include it in the PS model even when it is not statistically significant. In the present study, the advice of Brookhart et al. (2006) is followed and variables related to outcome are included. A conservative selection rule was followed with a p-value smaller than 0.10. Another consideration is the way the balancing effect of the variables is shown after control on the multiple propensity scores. With two treatments, it was straightforward to show this balancing effect, for example by making strata on the propensity score and by checking for differences within strata. With more than two (K) treatments there is no state of the art on how to show the balancing effect. In principle, balance can be shown by making K times K strata and showing the balancing effect within each stratum. In this study comparing five treatments, however, this was not doable. When the researcher is very careful, however, in estimating the multiple propensity score and has considered all possible interaction terms, at a certain point s/he should rely on the fact that the multiple propensity score was successful in balancing the important pre-treatment variables. P-values for significance testing are a useful way to illustrate the balancing effect but, owing to its dependency on the sample size, should be interpreted with care. Note that the purpose of adding variables and their interactions into the propensity score method is to obtain a better estimation. The purpose of (multiple) propensity score estimation is mainly for point estimation and the model selection process is superfluous in the propensity score model.

Fourth, in this study matching was not performed, as it was impossible to match on all five multiple propensity scores, which resulted in small treatment samples. Therefore, a regression analysis was adopted to correct for the multiple PSs.

The purpose of the present study was to provide hands-on guidelines for the clinical researcher who is faced with the impossibility of randomization in studies when trying to answer relevant clinical questions. It is illustrated that, in this study, the multiple PS could be implemented in the statistical process to control for a large set of confounders, and also when multiple treatments are compared. To ease the use of this promising method, an easy to follow step-by-step approach is presented. Hopefully, this will make this method accessible to a broad audience and foster its application, thereby enhancing the appreciation of well-conducted non-randomized studies. By using advanced statistical methods

to avoid bias such as the multiple PS method, researchers take responsibility for reducing the reasonable criticism of non-randomized studies. This might increase the scientific status of these studies and will help to answer relevant clinical questions in the mental health field and beyond.

Chapter 4

Effectiveness of different modalities of psychotherapeutic treatment for patients with cluster C personality disorder: Results of a large prospective multicentre study *

4.1 Summary

No previous studies have compared the effectiveness of different modalities of psychotherapeutic treatment, as defined by different settings and durations, for patients with cluster C personality disorders. The aim of this multicenter study was to compare the effectiveness of 5 treatment modalities for patients with cluster C personality disorders in terms of psychiatric symptoms, psychosocial functioning, and quality of life. The following treatment modalities were compared: long-term outpatient (more than 6 months), short-term day

*This chapter has been published as: Bartak, A., Spreeuwenberg, M.D., Andrea, H., Holleman, L., Rijnierse, P., van Rossum, B.V., Hamers, E.F.M., Meerman, A.M.M.A., Aerts, J., Busschbach, J.J.V., Verheul, R., Stijnen, T., & Emmelkamp, P.M.G. (2010). *Psychotherapy and Psychosomatics*, 79, 20-30.

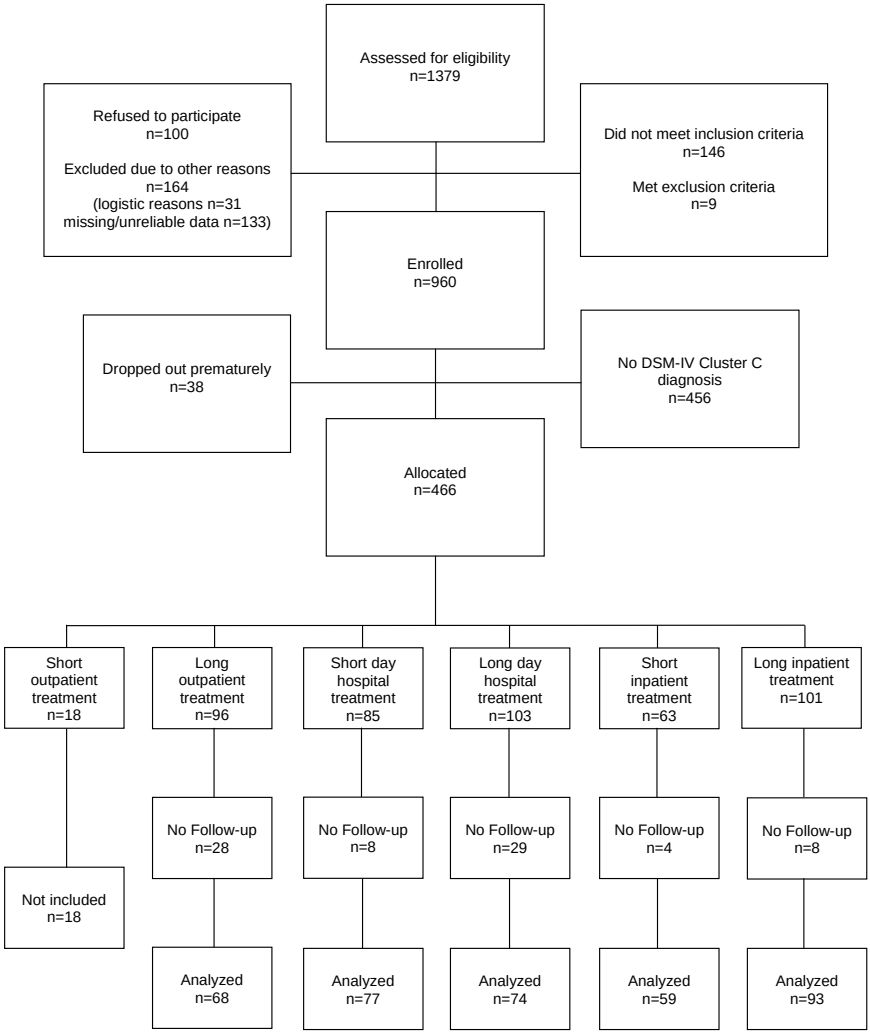

Figure 4.1: Patient flow

hospital (up to 6 months), long-term day hospital, short-term inpatient, and long-term inpatient psychotherapy. The study was conducted between March 2003 and June 2008 in 6 mental health care centers in the Netherlands, with a sample of 371 patients with a DSM-IV-TR axis-II cluster C diagnosis. Patients were assigned to 5 different modalities of psychotherapeutic treatment, and effectiveness was assessed at 12 months after baseline. An intention-to-treat analysis was conducted for psychiatric symptoms (Brief Symptom Inventory), psychosocial functioning (Outcome Questionnaire-45), and quality of life (EQ-5D), using multilevel statistical modeling. As the study was non-randomized, the propensity score method was used to control for initial differences. Patients in all treatment groups had improved on all outcomes 12 months after baseline. Patients receiving short-term inpatient treatment showed more improvement than patients receiving other treatment modalities. Psychotherapeutic treatment, especially in the short-term inpatient modality, is an effective treatment for patients with cluster C personality disorders.

4.2 Introduction

An estimated 2.6% of the general population is affected by cluster C personality disorders (PD): avoidant, dependent, and obsessive-compulsive PD (Coid, Yang, Tyrer, Roberts, & Ullrich, 2006). This cluster of PD is associated with significant functional impairment and a high economic burden, yet studies investigating treatment effectiveness in this patient population are scarce (Skodol et al., 2002; Skodol, Johnson, Cohen, Sneed, & Crawford, 2007; Grant et al., 2004; Skodol, Johnson, Cohen, Sneed, & Crawford, 2008; Duggan, Huband, Smailagic, Ferriter, & Adams, 2007). As in research on other psychological disorders, the available studies on cluster C PD typically compare treatments that are identical in treatment setting and duration. Investigators have compared different outpatient treatments, different day hospital treatments, and different inpatient treatments (Emmerik, Kamphuis, & Emmelkamp, 2008; Alden, 1989; Emmelkamp et al., 2006; Hellerstein et al., 1998; Stravynski, Belisle, Marcouiller, Lavallee, & Elie, 1994; Svartberg, Stiles, & Seltzer, 2004; Winston et al., 1994; Karterud et al., 2003; Wilberg et al., 1999; Gude & Vaglum, 2001; Teusch, Bohme, Finke, & Gastpar, 2001). One recent study in Norway compared outpatient and day hospital treatment for patients with all forms of

PD, and found no significant superiority of one treatment over another at 8 months after the start of treatment (Arnevik et al., 2009). However, so far, no study has compared the effectiveness of treatments across widely differing settings and durations. In this chapter, treatment modality was specified as a combination of treatment setting (i.e. outpatient, day hospital, or inpatient) and duration (i.e. short-term or long-term), as these are the most important aspects regarding treatment costs, a crucial aspect in times of restricted health care budgets.

It is likely that one of the reasons this comparison has not been undertaken previously is the difficulty of random assignment to different treatment modalities in clinical samples due to practical or ethical constraints (Black, 1996). Furthermore, even if researchers were successful in setting up and starting a randomized treatment modality study, its external validity would be doubtful because a high number of patients would refuse to participate (Zeeck et al., 2009). Therefore, quasi-experimental studies using statistical correction models to counter selection bias are increasingly being found in the literature (Facchinetti, Ottolini, Fazzio, Rigatelli, & Volpe, 2007; Forstmeier & Rueddel, 2007; Golkaramnay et al., 2007; Grossman, Tiefenthaler-Gilmer, Raysz, & Kesper, 2007).

The aim of the present quasi-experimental study was to compare the effectiveness of different treatment modalities for patients with cluster C PD in a naturalistic setting, thereby insuring high external validity. In fact, treatment modality might be an overlooked factor in psychotherapy effectiveness research.

4.3 Method

4.3.1 Participants

Participants (n = 371) were recruited from consecutive admissions to 6 mental health care centres in the Netherlands (Centre of Psychotherapy De Viersprong, Halsteren; Altrecht, Utrecht; Zaans Medical Centre, Zaandam; Centre of Psychotherapy De Gelderse Roos, Lunteren; GGZWNB, Bergen op Zoom & Roosendaal; Centre of Psychotherapy Centrum, Amsterdam). These institutions offer outpatient, day hospital, and/or inpatient psychotherapeutic treatment for patients with personality pathology. From March 2003 to March 2006, 1,379 patients completed the intake procedure and were selected for treatment

(figure 4.1).

Of these, 146 patients (10.6%) were excluded from the study for not meeting one of the following inclusion criteria: age between 18 and 70 (n = 13), significant personality pathology (n = 34), and referral for psychotherapeutic treatment aimed at personality problems (n = 99). Nine patients (0.7%) met one of the following exclusion criteria: insufficient command of the Dutch language (n = 6), organic cerebral impairment (n = 1), mental retardation (n = 1), and schizophrenia (n = 1). This left 1,224 participants, of whom 100 (8.2%) refused to participate. Another 31 patients (2.5%) could not participate due to logistic reasons (e.g. no appointment could be made to provide informed consent), and 133 patients (10.9%) were excluded due to missing or unreliable baseline data. Thirty-eight patients (3.1%) received less than 2 treatment sessions or less than 2 days of inpatient or day hospital therapy, and were therefore excluded. The remaining 922 patients were informed about the study and its procedure, provided written informed consent, and entered the study. Of those, 466 patients (50.5%) had 1 or more cluster C PD.

In the absence of explicit guidelines for treatment assignment in PD, the selection procedure was based on the expert opinion of clinicians who used their clinical experience combined with patient data from standardized instruments (Manen, 2008; Vervaeke & Emmelkamp, 1998). To elucidate the criteria used for the assignment process, the research group recently conducted a study with intake clinicians from the participating treatment centers. They found evidence of substantial (implicit) consensus among clinicians concerning the criteria used for treatment decision-making. For example, focality of problems (focal or broad spectrum of problems) and ego strength were found to be related to decisions about a short or long treatment duration for a substantial number of intake clinicians (Manen, 2008).

Patients were assigned to 1 of 6 treatment modality groups: 18 to short-term outpatient (up to 6 months), 96 to long-term outpatient (more than 6 months), 85 to short-term day-hospital, 103 to long-term day hospital, 63 to short-term inpatient, and 101 to long-term inpatient treatment. The short-term outpatient group was excluded from the analysis for 2 reasons: (1) only a minority of patients (3.9%) were assigned to this short and low-frequency treatment modality, as could be expected in a PD patient population; (2) these patients differed significantly from patients in the other treatment groups on

a high number of pre-treatment variables, indicating a dissimilar and most importantly a structurally less sick patient population, incomparable with the rest of the sample. A comparison with this treatment modality would most probably also fail when trying to design a randomized trial, as short-term outpatient therapy differs most from all other modalities in terms of its relatively low impact on patients lives compared to other treatment modalities. In the end, 448 participants were included in the study. Follow-up data were not available for 77 patients (17.2%; patients who did not respond to any follow-up assessment or patients where follow-up measurements were not yet available). There was no difference in psychiatric symptoms at baseline between patients with follow-up data and those without (this holds true for both the comparison in the total sample and the comparisons within the 5 treatment groups). The final sample consisted of 371 patients to be included in the analysis.

4.3.2 Treatment

The 6 mental health care centers offer a variety of psychotherapeutic treatments tailored to a PD patient population. Their treatments differ according to several features. As this study focused on different treatment modalities in terms of setting and duration, the following 5 treatment groups were compared:

* Patients in long-term outpatient treatment (n = 68, 18.3% of the study sample). These patients come for individual (76.5%) or group (23.5%) psychotherapy sessions, for up to 2 sessions per week (mean 0.8 sessions/week, SD 0.51, median 0.5) for more than 6 months (mean duration 15.4 months, SD 6.36, median 12.0).

* Patients in short-term day hospital treatment (n = 77; 20.8% of the study sample). These patients come to the institutions at least 1 morning/afternoon per week (mean 3.2 days/week, SD 1.51, median 3.0) for up to 6 months (mean duration 5.4 months, SD 1.32, median 6.0) and receive different forms of psychotherapeutic and psychosocial treatment, but sleep at home.

* Patients in long-term day-hospital treatment (n = 74, 19.9% of the study sample). These patients come to the institutions at least 1 morning/afternoon per week (mean 3.3 days/week, SD 1.42, median 3.0) for more than

6 months (mean duration 12.1 months, SD 2.41, median 12.0) and receive different forms of psychotherapeutic and psychosocial treatment, but sleep at home.

* Patients in short-term inpatient treatment (n = 59, 15.9% of the study sample). These patients stay at the institutions 5 days a week for up to 6 months (mean duration 4.2 months, SD 1.48, median 3.0) and receive different forms of psychotherapeutic and psychosocial treatment.

* Patients in long-term inpatient treatment (n = 93, 25.1% of the study sample). These patients stay at the institutions 5 days a week for more than 6 months (mean duration 10.2 months, SD 1.98, median 10.0) and receive different forms of psychotherapeutic and psychosocial treatment.

Day hospital and inpatient programs typically consist of group psychotherapy as a core element, mostly in combination with one or more non-verbal or expressive group therapies, individual psychotherapy, sociotherapy within the therapeutic community, coaching for social problems, community meetings, and/or pharmacological treatment. The psychotherapists are all licensed psychiatrists or psychologists. On average, they had 14.9 years (SD 10.1) of postgraduate clinical experience. The treatments under study can be considered highly representative of regular clinical practice in the Netherlands, as therapists did not receive specific training for this study and treatment integrity was not monitored.

4.3.3 Assesments

Baseline measures

An extensive standard assessment battery of instruments was administered to the patients before treatment assignment. PD were measured using the Dutch version of the Structured Interview for DSM-IV Personality (DeJong et al., 1986; Pfohl et al., 1997). This interview covers the 11 formal DSM-IV-TR axis II diagnoses including PD not otherwise specified, 2 appendix diagnoses (i.e. depressive and negativistic PD), and self-defeating PD. Interviewers were masters level psychologists, who were trained thoroughly by one of the authors (R.V.), and who received monthly booster sessions to avoid deviation from the interviewer guidelines. Inter-rater reliability was evaluated in 25 video-taped

interviews, which were rated by 3 observer-raters. Percentage of agreement between observer- raters ranged from 84 (avoidant PD) to 100% (schizoid) (median 95%). Intra-class correlation coefficients for the sum of DSM-IV PD traits present (i.e. scores 2 or 3) ranged from 0.60 (schizotypal) through 0.92 (antisocial) (median 0.74). To measure patient characteristics at baseline, the assessment battery also included 3 self-report instruments. The first of those was the Dutch version of the Dimensional Assessment of Personality Pathology-Basic Questionnaire (DAPP-BQ), for measuring the type and degree of personality pathology (Kampen, 2002; Livesley & Jackson, 2002). Patients scores on this questionnaire for the 4 higher-order factors were used: emotional dysregulation, dissocial behaviour, inhibition, and compulsivity. To measure the severity of personality pathology 5 higher- order domains of the Severity Indices of Personality Problems (SIPP) were used: self-control, social concordance, identity integration, relational capacities, and responsibility (Verheul et al., 2008). To measure patients motivation for treatment, the 2 scales of the Motivation for Treatment Questionnaire (MTQ-8) were used: need for help and readiness to change (Beek & Verheul, 2008).

Outcome measures

The primary outcome measure was general psychiatric symptomatology. This was measured using the Dutch version of the Brief Symptom Inventory, a validated self-report scale derived from the Symptom Checklist 90 Revised (Derogatis & Melisaratos, 1983; Beurs & Zitman, 2006; Derogatis, 1986; Arrindell & Ettema, 2003). In this study, the mean score of the 53 items of the Brief Symptom Inventory were used, i.e. the Global Severity Index (GSI), ranging from 0 to 4. Psychosocial functioning was measured with 2 subscales of the Outcome Questionnaire-45 (OQ- 45): (1) interpersonal relations and (2) social role functioning (Lambert et al., 1996). Health-related quality of life was measured using the EuroQol EQ-5D (EQ-5D) (Brooks et al., 2003). All 4 outcome measures, the GSI, OQ-45 interpersonal relations, OQ-45 social role, and EQ-5D, were assessed at baseline and several follow-up points. Three treatment centers conducted their follow-up at approximately 12, 24, and 36 months after baseline; the other 3 treatment centers conducted their follow-up at the end of treatment, approximately 6 and 12 months afterwards, and again at 36 months after baseline. The use of different assessment points was due to

logistic reasons, and was taken into account by choosing multilevel modeling as the statistical method for the analysis.

Table 4.1: Variables used for propensity score estimation, outcome GSI

Variable	Content
Age	patients age
DAPP-BQ Emotional dysregulation	unstable affective responding, interpersonal problems
DAPP-BQ Inhibition	deriving little enjoyment from intimate relationships
MTQ-8 Need for help	patients expressed desire for external help
MTQ-8 Readiness to change	willingness for treatment-seeking behaviour
EQ-5D	quality of life
SIPP Self-control	capacity to tolerate, use and control ones own emotions and impulses
SIPP Identity integration	coherence of identity; the ability to see oneself and ones own life as stable, integrated and purposive
SIPP Relational capacities	capacity to genuinely care about others as well as feeling cared for by them, to be able to communicate personal experiences, and to hear and engage with the experiences of others often but not necessarily in the context of a long-term intimate relationship
SIPP Responsibility	capacity to set realistic goals, and achieve these goals in line with the expectations generated in others
GSI	level of psychiatric symptoms
OQ-45 Symptom distress	level of symptom distress
OQ-45 Relational functioning	level of interpersonal functioning
OQ-45 Social role functioning	level of social and work functioning
Dimensional score cluster C PD	dimensional score of cluster C PD characteristics
Total dimensional score all PD	dimensional score of all PD characteristics
Avoidant PD	diagnosis of avoidant PD
Dependent PD	diagnosis of dependent PD
Obsessive-compulsive PD	diagnosis of obsessive-compulsive PD

Statistical analysis

First, the uncorrected results on all 4 outcome measures at 12 months after baseline were examined. Multilevel modeling was used to deal with: (1) the dependency of repeated measures on the same subject in time and (2) longitudinal data with observations unequally spaced in time (see Outcome measures). To estimate the uncorrected treatment effect at 12 months after baseline, a random intercept and random slope model was used with time as level I and patient number as level II. This resulted in a final best-fitting model with the following independent variables: dummy variables indicating group membership, time, and interaction between group membership and time. Subsequently, the within-group effect sizes (Cohens d) were calculated to describe change from

baseline to 12 months in each group (Cohen, 1988).

However, since this is a non-randomized study, the comparison of the groups had to be corrected for the influence of confounders, i.e. initial patient differences. To adjust for these differences and avoid bias in effect estimation, the 'multiple propensity score' was included in the analysis. The classic propensity score is defined as the conditional probability of assignment to 1 of 2 treatment groups given a set of observed pre-treatment variables (Rosenbaum & Rubin, 1983). The multiple propensity score is an extension of the classic propensity score to more than 2 treatment groups (Imbens, 2000). Statistical inclusion of possible confounders in the outcome analysis controls selection bias due to known confounders while comparing multiple groups. To identify relevant confounders, a long list of social, economic, and diagnostic variables was considered, carefully selected by both clinicians and researchers, based on the literature and clinical knowledge (Spreeuwenberg et al., 2010). All variables significantly related to a specific outcome were used to estimate the multiple propensity scores in a multinomial regression analysis, with group membership as a dependent variable (see table 4.1 for the variables included in the GSI propensity score; complete list of potential/identified confounders for all outcome variables available upon request).

One major advantage of the propensity score method, as compared to other correction techniques, is the fact that the overlap in propensity score distributions (and thus the overlap in relevant variables) between treatment groups can be easily judged and visualized. From looking at the overlap between the 5 treatment groups it appeared that, in spite of some differences, these groups were readily comparable. For a detailed description of this method and its use in psychotherapy research, see Bartak et al. (2009). A more sophisticated multilevel model, now including multiple propensity scores, was used to compare change in outcome variables across treatment groups. Dependent variables were the change scores (from baseline) observed during follow-up for each of the outcome measures. Independent variables were dummy variables indicating group membership, time, interaction between group membership and time, and the multiple propensity scores (with their mutual interactions). This model estimated differences in change scores at 12 months after baseline in pairwise comparisons of the 5 treatment groups. If significant differences in change scores were found, the between-group effect sizes were calculated.

To render the outcome estimates at 12 months more reliable, optimum use of the potential of this data-set was made by including all available data collected up to 800 days after baseline. Data collected after that point was not used in order to prevent bias of the 12-month data due to changes much later in the process. The number of available follow-up measures was as follows: up to 800 days, 30.5% of the total sample had 1 follow-up measure, 36.7% had 2 follow-up measures, and 32.9% had 3 follow-up measures. The analysis were performed using SPSS 15.0 for data preparation and Proc Mixed of SAS 9.1.3 for multilevel modeling (SASS Institute, Cary, N.C., USA).

4.4 Results

Sample characteristics

Of the 371 patients, 29.6% were male and 70.4% were female. The mean age was 33.5 years (SD 9.5). The highest level of education was low for 22.9%, medium for 19.4%, and high for 57.7%. Furthermore, 70.4% were unmarried, 21.3% were married, and 8.4% were divorced or widowed. The majority, 66.6%, had pure cluster C PD (i.e. no comorbid cluster A or B PD), 23.7% had a combination of cluster C PD and cluster B PD, 4.0% had a combination of cluster C PD and cluster A PD, and 5.7% had a combination of cluster C PD and both cluster A and B PD. A majority (63.3%) had a diagnosis of avoidant PD, 49.3% had a diagnosis of obsessive-compulsive PD, and 22.6% a diagnosis of dependent PD.

Uncorrected outcome

One year after baseline, patients in all treatment groups showed improvement in terms of psychiatric symptoms (GSI), the primary outcome measure. This is shown in table 4.2 and figure 4.2. Within-group effect sizes of the uncorrected scores ranged from 0.62 (medium effect, short-term day hospital group) to 1.78 (huge effect, short-term inpatient group).

Improvements were also seen in terms of psychosocial functioning and quality of life (table 4.2). Effect sizes for these outcome measures were somewhat lower compared to psychiatric symptoms, but a positive change in psychosocial functioning and quality of life was evident.

Corrected comparison

After correction for all relevant pre-treatment differences, improvement between baseline and assessment at 12 months proved to be significant for patients in all treatment groups on all 4 outcome measures ($p < 0.001$).

The short-term inpatient group showed significantly more improvement in psychiatric symptoms (GSI) than 3 other groups: the short-term day hospital group ($\beta = 0.38$, p = 0.0059, 95% CI 0.11-0.65), the long-term day hospital group ($\beta = 0.43$, p = 0.0032, 95% CI 0.15-0.71), and the long-term inpatient group ($\beta = 0.31$, p = 0.0248, 95% CI 0.04-0.57) (table 4.3). Between-group effect sizes (Cohens d) were 0.54, 0.57, and 0.40, respectively. This indicates medium effect sizes for the between-group comparisons of short-term inpatient

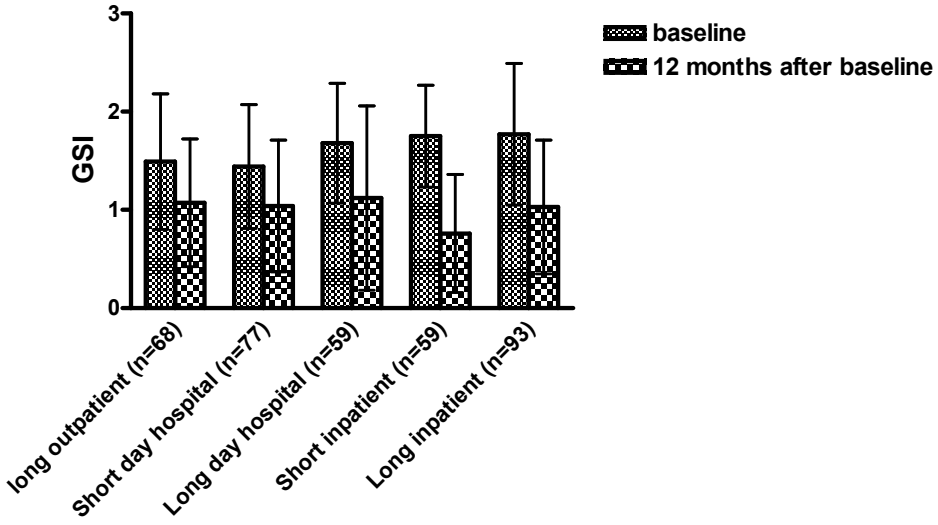

Figure 4.2: GSI uncorrected mean scores ± SD at baseline and 12-month follow-up

Table 4.2: Uncorrected outcomes (mean ± SD) and effect sizes in 5 treatment groups for all outcome variables

Variable	Treatment group	Baseline	12 months	Within-group effect size, Cohen's d
GSI	Long outpatient (n = 68)	1.49 ±0.69	1.07 ±0.65	0.63
	Short day hospital (n = 77)	1.44 ±0.63	1.04 ±0.67	0.62
	Long day hospital (n = 74)	1.68 ±0.61	1.12 ±0.94	0.71
	Short inpatient (n = 59)	1.75 ±0.52	0.76 ±0.60	1.78
	Long inpatient (n = 93)	1.77 ±0.72	1.03 ±0.68	1.06
OQ-45 Social role	Long outpatient (n = 68)	15.84 ±4.27	12.98 ±4.42	0.66
	Short day hospital (n = 77)	15.20 ±4.52	13.59 ±4.53	0.36
	Long day hospital (n = 74)	16.79 ±4.75	13.39 ±5.29	0.68
	Short inpatient (n = 59)	17.78 ±3.84	12.41 ±4.83	1.24
	Long inpatient (n = 93)	16.97 ±4.64	12.42 ±5.31	0.92
OQ-45 Inter-personal relations	Long outpatient (n = 68)	22.22 ±5.98	19.37 ±6.43	0.46
	Short day hospital (n = 77)	20.93 ±5.24	18.17 ±5.90	0.50
	Long day hospital (n = 74)	22.89 ±6.41	18.41 ±8.05	0.62
	Short inpatient (n = 59)	23.97 ±5.63	17.54 ±6.77	1.04
	Long inpatient (n = 93)	24.09 ±5.24	18.38 ±6.59	0.96
EQ-5D	Long outpatient (n = 68)	0.58 ±0.24	0.73 ±0.16	0.74
	Short day hospital (n = 77)	0.60 ±0.25	0.69 ±0.24	0.37
	Long day hospital (n = 74)	0.50 ±0.27	0.72 ±0.22	0.90
	Short inpatient (n = 59)	0.49 ±0.27	0.78 ±0.21	1.21
	Long inpatient (n = 93)	0.51 ±0.26	0.68 ±0.25	0.67

treatment versus other treatment groups.

In terms of social role functioning, the short-term inpatient group improved significantly more than 2 other groups the short-term day hospital group (β = 2.51, p = 0.0067, 95% CI 0.71-4.31) and the long-term day hospital group (β = 2.05, p = 0.0476, 95% CI 0.02-4.07) with between- group effect sizes of 0.49 and 0.38, respectively. The improvement in interpersonal functioning was significantly higher in the short-term inpatient group than in one other group the short-term day hospital group (β = 2.54, p = 0.0319, 95% CI 0.22-4.86) with a between group effect size of 0.39. Quality of life improved significantly more in the short-term inpatient group than in 2 other groups: the short-term day-hospital group (β = 0.15, p = 0.0009, 95% CI 0.06-0.23) and the long-term inpatient group (0.15, p = 0.0009, 95% CI 0.06-0.23) and the long-term inpatient group (β = 0.11, p = 0.0113, 95% CI 0.03-0.19). Between-group effect sizes were 0.6 and 0.42, respectively.

All results were based on intention-to-treat analysis (ITT), whereby ITT is defined as assignment and a minimal exposure to the intended treatment modality. The analysis were repeated with the treatment completers, i.e. those

Table 4.3: Difference scores (β) of 5 treatment groups 12 months after baseline, corrected for propensity score (all outcome variables)

Variable	Treatment group	n	β			
			Long outpatient	Short day hospital	Long day hospital	Short inpatient
GSI	Long outpatient	68				
	Short day hospital	77	0.078			
	Long day hospital	74	0.128	0.050		
	Short inpatient	59	0.302	0.380**	0.430**	
	Long inpatient	93	0.004	0.075	0.124	0.306*
OQ-45 Social role	Long outpatient	68				
	Short day hospital	77	1.632			
	Long day hospital	74	1.123	0.460		
	Short inpatient	59	0.876	2.508**	2.048*	
	Long inpatient	93	0.169	1.463	1.003	1.045
OQ-45 Interpersonal relations	Long outpatient	68				
	Short day hospital	77	0.836			
	Long day hospital	74	0.611	0.225		
	Short inpatient	59	1.704	2.540*	2.315	
	Long inpatient	93	0.084	0.752	0.527	1.788
EQ-5D	Long outpatient	68				
	Short day hospital	77	0.060			
	Long day hospital	74	0.001	0.061		
	Short inpatient	59	0.089	0.149***	0.088	
	Long inpatient	93	0.021	0.039	0.022	0.110*

Positive coefficients indicate that the treatment group shown in the left column is superior, negative coefficients indicate that the treatment group in the above row is superior.
* P-value < 0.05, ** P-value < 0.01, *** P-value < 0.001.

who actually stayed in the intended treatment modality group during their treatment (n = 298, 80.3% of the ITT sample, ranging from 66.2% for short-term day hospital to 89.7% for long-term outpatient treatment). These results followed the same pattern as the results from the ITT analysis: significant change within all treatment groups and a superiority of short-term inpatient treatment across all outcome measures (data available on request).

4.5 Discussion

This is the first study comparing the effectiveness of 5 modalities of psychotherapeutic treatment in a large population of patients with cluster C PD, as a contribution to the search for effective treatments for this patient group. Patients in all treatment groups had improved psychiatric symptoms, psychosocial functioning, and quality of life after 12 months. Most improvement was observed in the short-term inpatient group. This finding held when pre-treatment differences were controlled for with the propensity score.

Strengths and limitations

A clear strength of the present study is its external validity and clinical utility: it was conducted in regular clinical practice, not under experimental conditions (Hodgson, Bushe, & Hunter, 2007). A second strength is the rigorous statistical control of potential confounders, using the multiple propensity score methodology. Finally, a major asset of this study is its large number of patients. All this enabled the comparison of different psychotherapeutic treatment modalities while keeping sufficient statistical power.

Despite these strengths, the present findings have to be interpreted considering several limitations. First, even though all observed pre-treatment differences were controlled for, it cannot be ruled out that results have been influenced by unobserved confounders. To diminish this constraint as much as possible, a broad range of possible confounders was carefully selected and measured, based on both clinical and empirical knowledge, including variables identified in the literature as significant predictors of therapy outcome or process such as severity of baseline psychopathology, previous hospitalization, and substance misuse (Bartak et al., 2009; Gunderson et al., 2006; Links, Mitton, & Steiner, 1993; McGlashan, 1985; Ogrodniczuk et al., 2008; Plakun, 1991;

Ryle & Golynkina, 2000). In line with these earlier findings, previous hospitalization and substance misuse for example were significantly related to one of the secondary outcome measures, interpersonal functioning, and were therefore included in the propensity score for this measure. However, even when considerably reducing the possibility of important confounders being overlooked, not all possible variables could be covered in interviews and questionnaires at baseline, and therefore several variables, such as self-harm, were not measured (Chiesa & Fonagy, 2007).

Second, for ethical reasons, a reference group receiving no treatment at all was not included. Yet, several previous studies showed that specialized psychotherapeutic treatment yields better outcomes than various control conditions (for example waiting list controls) (Alden, 1989; Emmelkamp et al., 2006; Winston et al., 1994).

Third, research compliance differed between the treatment groups compared with most missing follow-up observations in the long-term treatment groups (figure 4.1). This might cause a problem of internal validity if non-response is not random, but related to systematic bias in effect estimation (positive or negative). However, there are 2 reasons why systematic bias seems unlikely: (1) responders and non-responders did not differ in psychiatric symptoms at baseline, and therefore it seems that they do not represent 2 structurally different groups of patients; (2) during the frequent telephone contact the authors had with non-responding patients to remind them to send back their questionnaires, these patients reported both negative and positive outcomes as reasons why they did not respond: some of them argued that their problems had worsened and that therefore they felt they did not have enough energy to fill in the questionnaires, others argued that their life had changed in a positive way and that therefore they did not want to be reminded of their time in therapy by filling in the questionnaires. Keeping this in mind, it seems unlikely that non-response was related to systematic negative or positive bias.

Fourth, this study does not rule out the possibility that treatment characteristics other than setting and duration played a role in the differential effectiveness of the 5 treatment modalities, e.g. frequency of sessions or theoretical orientation of treatment. This might represent a potential threat to internal validity. This is especially true for the role of theoretical orientation as a possible factor in the superiority of short-term inpatient treatment: most short-term

inpatient programs were based on psychodynamic principles. This concern is somewhat mitigated by previous studies comparing different theoretical orientations where no differences were found (Svartberg et al., 2004). However, to test the differential effect of modality and other treatment characteristics, a combined research design combining all these factors is needed.

Future directions and implications

What are the implications of the present results for future research, for practice guidelines, and for everyday clinical practice?

For patients with cluster C personality pathology, the short-term inpatient treatment clearly was associated with the highest improvement within 12 months. For this patient group, this modality of therapy seems to be the treatment backed up by the best available evidence in absence of long-term follow-up data. Replication of these results in a long(er)-term follow-up study is of vital importance to draw final conclusions. There might be a bias in favor of short-term treatment because patients in the long-term treatment groups might still be in therapy at 12 months. Long-term follow-up after termination of all treatment programs is therefore warranted. Another question is whether the benefit in terms of effectiveness is worth the potential cost differences when evaluated with recently upcoming state-of-the-art cost-effectiveness analysis (Leichsenring et al., 2009; McCrone et al., 2007). From these analysis within this study sample, it appeared that the mean direct treatment costs of the 5 treatment modalities were EUR 10,005 (SE 1,134) for long-term outpatient treatment, EUR 16,813 (SE 1,361) for short-term day hospital treatment, EUR 27,648 (SE 2,654) for long-term day hospital treatment, EUR 25,933 (SE 859) for short-term inpatient treatment, and EUR 49,260 (SE 2,435) for long-term inpatient treatment (Skodol et al., 2008). It would be interesting to compare the cost-effectiveness of short-term inpatient psychotherapeutic treatment with that of manual-based outpatient treatments such as cognitive-behavioral therapy (Emmelkamp et al., 2006). A state-of-the-art cost-effectiveness analysis would include medical costs incurred outside the treatment institution, productivity costs, and other indirect costs. This kind of analysis and its economic interpretation is beyond the range of this study and needs considerable research in the future.

If the superiority of short-term inpatient psychotherapeutic treatment holds

at long-term follow-up, in cost-effectiveness analysis, and in comparison with other evidence- based manual-based treatments, this treatment modality might be considered as the treatment of choice for this patient group. This would be a thought-provoking finding, as previous studies in cluster B PD patients have found outpatient and day hospital treatments to be very effective in this population (Chiesa, Fonagy, & Gordon, 2009; Clarkin, Levy, Lenzenweger, & Kernberg, 2007; Giesen-Bloo et al., 2006; Bateman & Fonagy, 2001). Even though no study compared one of these modalities directly with inpatient therapy, one might speculate that different therapy modalities are effective for different groups of patients. It could be that the success of short-term inpatient treatment in a cluster C PD sample is embedded in the combination of only short hospitalization, thereby preventing iatrogenic effects, and a high level of therapeutic intensity and pressure. Patients with cluster C personality pathology might be able to handle the high pressure of this treatment modality better than (pure) cluster B PD patients, who probably have a lower tolerance for therapeutic pressure, resulting in more early drop-outs and thus a less effective treatment. They might instead need less pressure with a longer treatment duration (Bateman & Fonagy, 2001; Lorentzen & Hoglend, 2008). Future studies may verify this hypothesis. However, even when superiority of short-term inpatient treatment for cluster C PD patients has been confirmed in the literature, patients caring for children might still not be assigned to inpatient treatment. Also, patients with a high severity of psychiatric symptoms or a low level of ego strength might not be able to handle the pressure of intensive inpatient treatment. It is recommended to investigate these potential matching factors further as this would enable clinicians to make specific treatment recommendations for different subgroups of cluster C PD patients and to develop new clinical practice guidelines.

In conclusion, this study suggests that psychotherapy, especially in a short-term inpatient modality, is an effective treatment for patients with cluster C PD. This makes inpatient psychotherapeutic treatment an interesting option for patients with avoidant, dependent, and obsessive-compulsive PD. The present findings can contribute to more adequate and tailored health care for this vulnerable patient group, as implementing effective treatments may reduce the considerable burden to individuals and society as a whole.

Chapter 5

Countering hidden bias in psychotherapy research: Extending the Heckman method*

5.1 Summary

In psychotherapeutic research, traditional methods that counter for overt bias are often used in quasi-experimental study designs. As an alternative, to account for possible hidden bias, the original two-step Heckman method and its extended version using structural equation modeling are discussed. The performances of multiple regression analysis, the propensity score method, the original Heckman maximum likelihood method and its extended version using structural equation modeling (SEM) are compared in four artificial data-sets. In addition, to illustrate the methods, data from a mental health study are used as a real world example. The original Heckman method is very sensitive to mis-specification of the selection model and to violations of the normality of error-terms assumption. When a randomized controlled trial is not possible, methods other than those dealing with overt bias could be considered. When good indicators for a 'latent tendency to participate in the study' are available, the extended version of the Heckman method using SEM analysis is preferred over the Heckman method.

*This chapter is under review at Evaluation & the Health Professions.

5.2 Introduction

An important research area in the social sciences is the comparison of treatment programs. In comparative research studies, randomized control trials (RCT's) are considered gold standard. In RCT's, participants are assigned to the treatment programs by random procedures such as flipping a coin. With randomization, it is expected that both observed and unobserved pre-treatment variables have, on average, the same values in all treatment groups. The probability that this is actually true increases as the sample size increases. Let us consider a study comparing two therapies for depression (D). Each patient i is randomly allocated to either the new therapy ($D = 1$) or standard therapy ($D = 0$). Let Y_{id} represent the depression outcome score Y of patient i within therapy D. Since patients are randomized into the therapies, one expects, especially in large sample sizes, that the two treatment groups are initially comparable on variables such as age, gender, social economic class, initial level of depression, or motivation. With initial comparability, a significant difference in the mean outcome depression scores between the two patient groups can be attributed to the therapy program received. The added value of the new therapy against the standard therapy (δ), i.e. the average causal effect for the treated (ACT) can therefore be estimated by subtracting the mean outcome of participants in the new therapy ($E(Y_1)$) from the mean outcome of patients in the standard therapy ($E(Y_0)$) (Rubin, 1974). In the formula the ACT is:

$$ACT = E(Y_1) - E(Y_0) \tag{5.1}$$

In clinical practice, however, it is not always possible to randomly assign patients to treatments, as randomization may be unethical, difficult, costly, or impossible. Therefore, quasi-experimental or observational studies are often conducted where patients select themselves into the treatment options (Shadish & Cook, 2002). When mainly male patients with a low social economic background choose the standard therapy, but mainly women with a high social economic background choose the new therapy, selection has occurred. Variables related to this selection process such as gender or social economic status, are named selection variables. Variables that influence the outcome value are usually named independent variables. Confounding variables are variables that influence both the selection process and the outcome value, which without con-

trol, will lead to bias in the estimated treatment effect. With confounding, estimating the ACT as in equation 5.1 will lead to biased estimated treatment effects. Selection bias is the bias introduced into a (quasi-)experimental study by the selection of different types of subjects into the reference and experimental condition(s) (Heckman, 1979; Winship & Mare, 1992). Rosenbaum (1995) distinguishes between overt and hidden bias. Overt bias results from differences in observed and measured pre-treatment variables, whereas hidden bias results from differences in unobserved and unmeasured characteristics between treatment groups.

The aim of this study is to discuss the essence and assumptions of two statistical methods dealing with hidden bias in their statistical analysis, namely the traditional Heckman two-step method and its extended version using structural equation modeling (SEM). The results from these two methods are compared with traditional methods for overt bias, such as multiple regression analysis and propensity score methods. The remainder of the chapter is organized as follows: in section 5.2, the methods for overt bias such as matching, stratification and propensity score methods are discussed. In section 5.3, the traditional Heckman two-step method is discussed and its extended version based on SEM modeling presented. Section 5.4 discusses the results of the analysis of artificial data where the performances of multiple regression analysis and the propensity score method are compared to the Heckman two-step method and SEM analysis. In section 5.5, data from the Dutch research project *SCEPTRE* ('Study on cost-effectiveness of personality disorder treatment') is used as a real world example (Bartak et al., 2009).

5.3 Overt bias

Traditionally, most statistical methods focus on reducing overt bias. The basic idea underlying all these methods is to make treatment groups as comparable as possible on all observed pre-treatment variables. The most common methods are matching, stratification, statistical control by means of multiple regression analysis, and propensity score methods. With matching, one attempts to achieve comparability by pairing each patient in the experimental group with one or more similar patient(s) in the reference group. With stratification, several groups of patients are formed based on the same set of observed

pre-treatment variables. As a result of stratification, within each group or strata patients have, in principle, the same distribution of the pre-treatment variables. With multiple regression analysis, overt bias may be reduced by adding extra covariates into the multiple regression equation. Traditional multiple regression analysis is, however, a statistical method that implies a linear relationship between the independent variables and the dependent variable, with no interaction effects. When the relationships become more complex with non-linear relations or many interaction effects, matching or stratification may be more simple and easier-to-use methods to handle this complexity. However, matching and stratification will become difficult when the number of variables to match or stratify on increases. In that case, it may be impossible to find patients who are similar on all these variables. To reduce this 'dimensionality problem', Rosenbaum and Rubin (1983) proposed the use of a single score, the propensity score (PS), which can be used for matching, stratification and multiple regression adjustment. The PS is defined as the probability of assignment to the experimental condition, given all observed pre-treatment variables. The PS can be estimated by means of a probit analysis as:

$$P(D = 1|\boldsymbol{X}) = \Phi(\alpha_0 + \boldsymbol{\alpha_1 X}) \tag{5.2}$$

where Φ is the cumulative distribution function of the standard normal distribution, α_0 the multiple regression constant, α the multiple regression coefficients, and $\boldsymbol{X}$ a vector of observed pre-treatment variables. There exist alternative estimation methods for the estimation of PS such as logit or discriminant analysis. With balance, conditional on the PS, the assignment into the treatment programs does not depend further on pre-treatment variables and is treated as random. For that reason the ACT can, in quasi-experimental studies, be estimated after control on the PS as:

$$ACT = E(Y_1|PS, D = 1) - E(Y_0|PS, D = 0) \tag{5.3}$$

The main disadvantage of all statistical methods described above is, however, that they strongly rely on the assumption that the assignment into the treatment programs only relies on measured variables and does not depend on variables that are unmeasured (Rubin, 1976). This implies that one has to know and measure all confounding variables. When the PS does not include all

important variables, even after control of the PS, the assignment cannot be treated as random. Then, estimating the ACT as in equation 5.3 will yield biased estimated treatment effects. Therefore, it is important that researchers select and measure all potential confounding variables before the start of their study. Careful selection and measurement of the baseline variables from both a practical and statistical point of view is therefore very important. Nonetheless, researchers are never sure that all important confounder variables are included in their studies.

5.4 Hidden bias

The traditional Heckman two-step method is most influential in taking hidden bias into account in the analysis (Rubin, 1976). Based on the correlation between the error-terms of the selection and outcome models, an extra term, meant to capture the influence of the unknown variables, is estimated and included as an extra predictor in the outcome model. Because of the strong reliance on the assumption of normally distributed error-terms, however, the method has received some criticism. Therefore, as an alternative method, a modified version of the Heckman two-step method using structural equation modeling (SEM) is presented in this chapter. In this section, both the Heckman two-step method and its extended version using SEM are explained (Heckman, 1979; Bollen, 1989).

5.4.1 The original Heckman two-step method model

The basic idea of the Heckman two-step method is that assignment into treatment programs is not random, but depends on a latent, unobserved variable, which can be seen as a latent variable reflecting the 'tendency to participate in the experimental condition instead of in the reference condition' (Heckman, 1979). The original Heckman two-step model distinguishes a selection and an outcome model. In the selection model, this latent variable D^* is explained by a set of pre-treatment variables $\boldsymbol{X}$ such as need for help or motivation. In formula the relation between D^* and $\boldsymbol{X}$ is expressed as:

$$D^* = \alpha_0 + \boldsymbol{\alpha}\boldsymbol{X} + \varepsilon_{d^*} \tag{5.4}$$

where $\boldsymbol{\alpha}$ represent the set of the multiple regression coefficients which explain

the strength of the direct effects of $\boldsymbol{X}$ on D^* and ε_{d^*} the error-term; the variance of the error-term is denoted as $\sigma^2_{\varepsilon_{d^*}}$. To determine the measurement scale of D^*, the mean value of the error-term of D^* (ε_{d^*}) is, without loss of generality, arbitrarily set equal to zero with a standard deviation equal to one. It is assumed that when an individual i participates in the experimental condition (e.g. the new therapy) ($D_i = 1$), its value on D^* is larger than zero ($D_i^* \geq 0$). When an individual does not participate in the experimental condition but in the standard therapy ($D_i = 0$), its value on D^* is smaller than zero ($D_i^* < 0$). When $\boldsymbol{X}$ includes all selection variables, the error-term ε_{d^*} is uncorrelated with all variables in $\boldsymbol{X}$. However, when hidden bias is present, it implies that one or more selection variables are unmeasured and not included in $\boldsymbol{X}$. These unmeasured selection variables will correlate with the error-term ε_{d^*}.

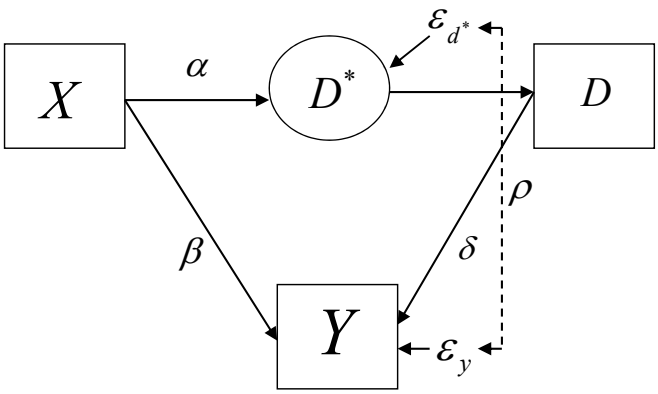

Figure 5.1: Graphical representation of the Heckman method

In the outcome model, the outcome score (Y) of participants in either the experimental or the reference condition is explained by the therapy program (D) and the values on $\boldsymbol{X}$. The outcome equation can therefore be written as:

$$Y = \beta_0 + \delta D + \boldsymbol{\beta X} + \varepsilon_y \tag{5.5}$$

where β_0 is the multiple regression constant, δ the estimated average causal effect (ACE), β a set of multiple regression coefficients related to the pre-treatment variables $\boldsymbol{X}$, and ε_y the error-term of the equation. The model assumes that ε_{d^*} and ε_y may be correlated by a factor rho (ρ).
For participants in the reference condition, the expected value of the outcome score Y is:

$$E(Y|\boldsymbol{X}, D = 0) = \beta_0 + \boldsymbol{\beta X} + E(\varepsilon_y|D = 0) \tag{5.6}$$

and for participants in the experimental condition, the expected value of the outcome score Y is:

$$E(Y|X, D = 1) = \beta_0 + \delta D + \boldsymbol{\beta X} + E(\varepsilon_y|D = 1) \tag{5.7}$$

As $D = 0$ is an indicator of a negative value for the latent variable 'tendency to participate in the experimental condition', equation 5.6 can be rewritten as:

$$E(Y|\boldsymbol{X}, D = 0) = E(Y|\boldsymbol{X}, D^* < 0) = \beta_0 + \boldsymbol{\beta X} + E(\varepsilon_y|D^* < 0) \tag{5.8}$$

Because in the selection model D^* is modeled as $D^* = \alpha_0 + \boldsymbol{\alpha X} + \varepsilon_{d^*}$, this equation can again be rewritten as:

$$\begin{aligned} E(Y|\boldsymbol{X}, D^* < 0) &= \beta_0 + \boldsymbol{\beta X} + E(\varepsilon_y|\alpha_0 + \boldsymbol{\alpha X} + \varepsilon_{d^*} < 0) = \\ &= \beta_0 + \boldsymbol{\beta X} + E(\varepsilon_y|\varepsilon_{d^*} < -(\alpha_0 + \boldsymbol{\alpha X})) \end{aligned} \tag{5.9}$$

Since $D = 1$ is an indicator of a positive value for the latent variable 'tendency to participate in the experimental condition', for participants in the experimental condition, equation 5.7 can be rewritten as:

$$E(Y|\boldsymbol{X}, D = 1) = E(Y|\boldsymbol{X}, D^* \geq 0) = \beta_0 + \delta D + \boldsymbol{\beta X} + E(\varepsilon_y|D^* \geq 0) \quad (5.10)$$

which can be rewritten as:

$$\begin{aligned} E(Y|\boldsymbol{X}, D^* \geq 0) &= \beta_0 + \delta D + \boldsymbol{\beta X} + E(\varepsilon_y|\alpha_0 + \boldsymbol{\alpha X} + \varepsilon_{d^*} \geq 0) = \\ &= \beta_0 + \delta D + \boldsymbol{\beta X} + E(\varepsilon_y|\varepsilon_{d^*} \geq -(\alpha_0 + \boldsymbol{\alpha X})) \end{aligned} \quad (5.11)$$

When all important variables are included in the selection and outcome models and no hidden bias is present, then it follows that; (a) the error-terms of the selection equation (ε_{d^*}) and outcome equation (ε_y) are not correlated, (b) the expected value of (ε_{d^*}) is equal to zero, and (c) the joint distribution of the error-terms in the outcome models ε_y is assumed to follow a bivariate normal distribution with a mean expected value of zero and a standard deviation of one.

When hidden bias is present and confounding variables are missed, however, it follows that; (a) these variables are correlated with the error-term ε_{d^*}, (b) the two error-terms of the selection and the outcome model are correlated by a factor rho (ρ), (c) the expected value of (ε_{d^*}) is not equal to zero, and (d) estimating the ACT as in equation 5.6 leads to biased estimated treatment effects.

With the assumption that the error-terms in the outcome models follow a strict bivariate normal distribution, the expected value of the error-terms of the outcome equation of the participants in the experimental condition and the reference condition can be rewritten as:

$$E(\varepsilon_y|\varepsilon_{d^*} < -(\alpha_0 + \boldsymbol{\alpha X})) = -\rho(\frac{\phi(\alpha_0 + \boldsymbol{\alpha X})}{1 - \Phi(\alpha_0 + \boldsymbol{\alpha X})}) \quad (5.12)$$

and

$$E(\varepsilon_y|\varepsilon_{d^*} \geq -(\alpha_0 + \boldsymbol{\alpha X})) = \rho(\frac{\phi(\alpha_0 + \boldsymbol{\alpha X})}{\Phi(\alpha_0 + \boldsymbol{\alpha X})}) \quad (5.13)$$

where ϕ denotes the standard normal density and Φ the cumulative distribution function (Heckman, 1979). When the error-terms of the selection and outcome

equation are uncorrelated (ρ =0), the expected error-terms are both zero for all subjects. When the error-terms are correlated, due to a variable that is omitted from the analysis, the expected error-terms are non-zero.

The main idea of the original Heckman two-step method is to estimate an extra term for all subjects, named lambda, and add this to the outcome model to protect against the biasing effect of an unmeasured variable. By adding this term in the multiple regression equation, the expected value of the error-term is zero again for all subjects. Here, a main assumption is that all confounding variables that are omitted from the analysis are uncorrelated to the other variables in $\boldsymbol{X}$. Then, for each participant i, the term lambda (γ) is estimated in the following ways;
For reference participants ($D = 0$):

$$\gamma_i = -(\frac{\phi(\alpha_0 + \boldsymbol{\alpha X})}{1 - \Phi(\alpha_0 + \boldsymbol{\alpha X})}) \tag{5.14}$$

and for experimental participants ($D = 1$):

$$\gamma_i = (\frac{\phi(\alpha_0 + \boldsymbol{\alpha X})}{\Phi(\alpha_0 + \boldsymbol{\alpha X})}) \tag{5.15}$$

To correct for hidden bias, the estimated lambda term is added in the outcome equation as an extra variable in the model. Then, according to the model of Heckman, the new outcome equation is;

$$Y = \beta_0 + \delta D + \boldsymbol{\beta X} + \rho\gamma + \varepsilon_y \tag{5.16}$$

where ρ represents the partial regression coefficient of the lambda term. With overt bias, the correlation between the error-terms ε_{d^*} and ε_y is zero and the lambda term will disappear from equation 5.16. With hidden bias, the correlation between the error-terms is unequal to zero and the influence of the lambda term increases. A t-test for the regression coefficient of lambda (ρ) can be used to test if hidden bias is present. See Heckman (1979) and Greene (1981) for the corrected standard errors of the estimated treatment effect δ. The estimation method of the original two-step approach of Heckman is limited information. First the selection is modeled, followed by the outcome equation. As an alternative, a full information maximum likelihood estimation method can be

applied to estimate the selection and the outcome model simultaneously. Statistical packages such as (StataCorp, 2001) offer this full information maximum likelihood (ML) approach.

5.4.2 Evaluation of the original Heckman two-step method

A first consideration when using the original Heckman two-step method is its strong reliance on the assumption of normal distributed error-terms and on linearity assumptions (Winship & Mare, 1992). A key assumption of the Heckman method is that the error-terms in the selection and outcome equation follow a bivariate normal distribution. This assumption is needed for a consistent estimate of the lambda term and implies a linear relationship between both error-terms. If the relation between ε_{d^*} and ε_y is non-linear or ε_{d^*} is not normally distributed, the lambda term is misspecified and the model will yield biased estimated treatment effects. Goldberger (1983) showed that, with the Heckman method, even small violations of the normality assumption lead to bias. Because error-terms represent unmeasured variables, a normal distribution cannot be guaranteed and never be validated. For that reason, some researchers have questioned the use of the Heckman method. In the econometric field, semi-parametric and non-parametric adaptations of the Heckman method have been developed (Olsen, 1980).

A second limitation concerns the two multicollinearity problems that arise when using the Heckman method. The first multicollinearity problem is a consequence of the high correlation between the dummy variable indicating treatment and the lambda term in the outcome equation (Puhani, 2000). Due to this multicollinearity, statistical tests will have power problems. The second multicollinearity problem arises when exactly the same variables are used in the selection equation and in the outcome equation. To avoid this problem, it is advisable to use at least one variable in the selection equation that is not in the outcome equation. This is called the exclusion restriction (Yamagata & Orme, 2005),.

Third, just like other statistical methods, the Heckman method depends largely on the specification of the model. If the outcome equation is not well specified it will result in biased estimated treatment effects.

Fourth, the Heckman method relies strongly on the assumption that the confounder variables omitted from the analysis are uncorrelated with the in-

dependent or confounding variables that are explicitly included in the model. This assumption will often not be met in practice.

In the Heckman approach, the dummy variable (D) that indicates whether a participant receives either the reference condition or the experimental condition, is used as the only indicator for the latent variable D^*. In the next section, an alternative model, SEM analysis, is presented in which more indicators for the latent variable D^* are used. As a consequence, a more stable and flexible model can be obtained in which relationships between the variables can be modeled in a more complex way.

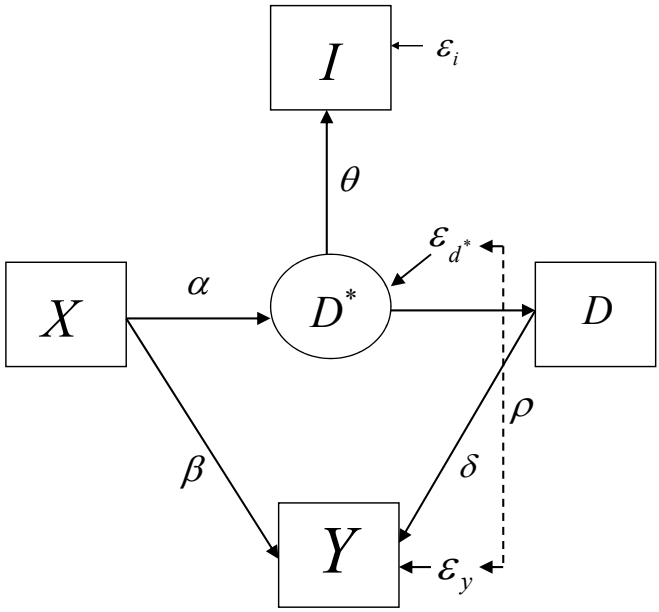

Figure 5.2: Graphical representation of the SEM analysis

5.4.3 Structural equation modeling (SEM)

In this section, structural equation modeling (SEM) is discussed as an alternative for estimating treatment effects when hidden bias is present. SEM is a general statistical modeling technique with which to establish 'causal' relationships among variables. It consists of a system of (logit) multiple regression equations that are analyzed simultaneously in one model. A key feature of SEM is that some indicator variables are understood to represent a 'latent construct', that cannot be directly measured but only inferred from the observed measured variables. Both the independent and dependent variables in the model can be continuous, discrete or present a latent (unobserved) variable. SEM analysis is an estimation model that minimizes the difference between observed sample covariances and the covariances predicted by the model (Bollen, 1989).

Figure 5.2 shows a path diagram of an SEM analysis applicable to observational studies. It is an extension of the Heckman method in the sense that it allows for more than one indicator variable for the latent variable D^*. These indicator variables $\boldsymbol{I}$ are observed variables that may contain a set of proxy variables for the tendency to participate such as motivational aspects or a wish for treatment. Again, there are some pre-treatment variables $\boldsymbol{X}$ influencing the latent variable D^*. The latent variable D^* can be seen as a latent variable measuring the tendency to participate in the experimental condition. This latent variable is not directly measured, but rather assessed indirectly by some indicator variables $\boldsymbol{I}$. See Bollen (1989) for ways to obtain an identifiable latent variable model. In our model, the latent variable determines which treatment a subject receives, as indicated by a dummy variable D. The outcome variable Y is influenced by the dummy variable indicating treatment and possibly by some other independent variables $\boldsymbol{X}$. Notice that latent variables are graphically given by circles and observed variables by rectangles. The lines in the path diagram represent direct relations between variables: lack of a line between variables implies that no direct relationship between these variables is hypothesized, given that all appropriate prior and intervening variables are hold constant. Lines have either single or double headed arrows. A line with one arrow represents a direct relationship between two variables, controlled for the other variables in the model. The variable with one arrow pointing to it is the dependent variable. A line with a double headed arrow allows for covariance between the two variables with no implied direction of effect.

The model graphically displayed in figure 5.2 can also be represented in a series of multiple regression and logit equations as:

$$\begin{aligned} D^* &= \alpha_0 + \boldsymbol{\alpha X} + \varepsilon_{d^*} \\ \boldsymbol{I} &= \theta_0 + \theta D^* + \varepsilon_i \\ \text{if } (D^* < 0), \text{ then } D &= 0 \\ \text{if } (D^* \geq 0), \text{ then } D &= 1 \\ Y &= \beta_0 + \delta D + \boldsymbol{\beta X} + \varepsilon_y \end{aligned} \tag{5.17}$$

where α represents the direct effects of $\boldsymbol{X}$ on D^*, θ the direct effect of D^* on $\boldsymbol{I}$, δ the estimated treatment effect of D on Y, $\boldsymbol{\beta}$ the set of partial regression coefficients of $\boldsymbol{X}$ on Y, and ε_{d^*}, ε_i and ε_y represent the error-terms. A common model assumption is that all error-terms have an expected value of zero for all subjects and are uncorrelated to each other. However, when confounding variables are omitted from the analysis that are independent of the other variables, it fuses into the disturbance terms ε_{d^*} and ε_y. For that reason, a covariance between the error-terms ε_{d^*} and ε_y is allowed for. In figure 5.2, this covariance is represented by a line with a double headed arrow, connecting the error-term ε_{d^*} of the latent variable 'tendency to participate in a study' with the error-term ε_y of the outcome equation of Y. In SEM all equations can be simultaneously analyzed and the parameters of the model in SEM are estimated by Maximum Likelihood. There exist some statistical packages that facilitate SEM analysis with dichotomous outcome variables such as M*plus* or Latent Gold (Muthén & Muthén, 2008; Vermunt & Magidson, 2000). In M*plus*, a step-function cannot be determined. However, the deterministic relationship between D and D^* can be approached by a logit equation with the regression coefficient of D on D^*, fixed to a very large value. M*plus* also provides robust ML estimates when the assumption of normally distributed error-terms is violated. Fit measures such as chi-square, RMSEA or CFI are provided to asses the fit of models or to compare models with each other. A main advantage of SEM analysis over the Heckman method is that it provides a more flexible model. The model can be extended to more indicators for D^* (either continuous or dichotomous) or with more complex relationships between variables such as a direct relationship

between indicators and the outcome.

5.5 Analysis on artificial data

The aim of the analysis on the artificial data-sets is to compare the results from the original Heckman method and its extended version using SEM analysis with the results from multiple regression analysis and the propensity score method. Analysis on the artificial data is performed to answer the following questions; Which method is best to use when a variable is omitted in the analysis that influences the outcome, the selection or both? Do these methods show much variation in their performance?

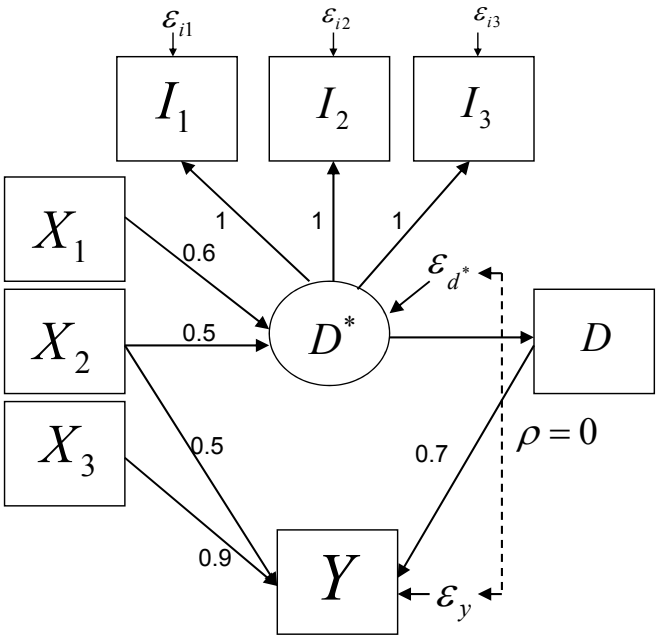

Figure 5.3: Graphical representation of the population model used to simulate the artificial data

Table 5.1: Summary of the four artificial data-sets

	Correlation between independent variables X	Distribution of the error-terms
data-set 1	0	Normal
data-set 2	0.5	Normal
data-set 3	0	Kurtosis
data-set 4	0	Skewed

5.5.1 Description of the four artificial data-sets

In this study four very large artificial data-sets are generated. The sample size of a single data-set was 10,000. See figure 5.3 for the graphical representation of the population model that is used to generate the four artificial data-sets. In each data-set, three independent variables X_1, X_2 and X_3 were generated. These variables were normally distributed with a mean of zero and a variance of one. Variables X_1 and X_2 influenced a continuous latent variable D^* which represents the 'tendency to participate in the experimental condition'. This latent variable influenced three indicator variables I_1, I_2 and I_3 with an effect coefficient of one. Indicators I_1, I_2 and I_3 are also influenced by an error term, but not D^*. The latent variable D^* determines the dummy variable D completely: when $D^* < 0$, the participant i is assigned to the reference condition ($D = 0$), when $D^* \geq 0$, the participant is assigned to the experimental condition ($D = 1$). The outcome Y is directly influenced by the treatment variable D and independent variables X_2 and X_3. The relationship between D and Y is seen as the treatment effect δ which is equal to 0.7. In the population model, the correlation between the error-terms ε_{d^*} and ε_y was equal to zero.

In table 5.1 the four artificial data-sets are described. In the first artificial data-set the values of the inter-correlations among the independent variables X_1, X_2 and X_3 were equal to zero. In the second artificial data-set these values were equal to 0.5. The performances of the methods between these two data-sets are compared to investigate the influence of omitting a variable that relates to other independent variables in the model. In the third artificial data-set the correlation between the independent variables was equal to zero, but with non-normally distributed error-terms ε_{d^*} and ε_y with a kurtosis distribution of 0.75. In the fourth artificial data-set the correlation between the independent variables is equal to zero, but the error-terms ε_{d^*} and ε_y were simulated with a skewed distribution of 0.75. The method of Fleishman (1978) was used to

Table 5.2: Results of the analysis in the four artificial data-sets

	Regression analysis		Propensity score		Heckman two-step		Heckman ML		SEM analysis	
	δ	Abs.	δ	Abs.	δ	Abs.	δ	Abs.	δ	Abs.
	Se(δ)	Bias	Se(δ)	Bias	Se(δ)	Bias	Se(δ)	Bias	Se(δ)	Bias
data-set 1: cor(X)=0										
Naïve model	1.026									
	(0.029)									
Correct model	0.709	0.009	0.697	0.003	0.744	0.044	0.743	0.043	0.730	0.030
	(0.023)		(0.032)		(0.051)		(0.028)		(0.028)	
X1 missing	0.715	0.015	0.716	0.016	0.996	0.296	0.807	0.107	0.728	0.028
	(0.021)		(0.029)		(0.566)		(0.313)		(0.040)	
X2 missing	1.084	0.384	1.084	0.384	0.743	0.043	0.748	0.048	0.781	0.081
	(0.024)		(0.031)		(0.059)		(0.058)		(0.036)	
X3 missing	0.698	0.002	0.697	0.003	0.703	0.003	0.703	0.003	0.735	0.035
	(0.031)		(0.032)		(0.070)		(0.069)		(0.043)	
data-set 2: cor(X)=0.5										
Naïve model	1.714									
	(0.030)									
Correct model	0.682	0.018	0.674	0.026	0.695	0.005	0.695	0.005	0.667	0.033
	(0.024)		(0.031)		(0.064)		(0.064)		(0.035)	
X1 missing	0.685	0.015	0.677	0.023	0.967	0.267	0.931	0.231	0.657	0.043
	(0.023)		(0.027)		(0.235)		(0.218)		(0.040)	
X2 missing	0.930	0.230	0.931	0.231	1.314	0.614	1.316	0.616	0.979	0.279
	(0.025)		(0.031)		(0.048)		(0.048)		(0.033)	
X3 missing	0.677	0.023	0.675	0.025	2.114	1.414	1.892	1.192	1.116	0.416
	(0.030)		(0.032)		(0.077)		(0.064)		(0.043)	
data-set 3: Kurtosis=0.75										
Naïve model	0.950									
	(0.095)									
Correct model	0.711	0.011	0.705	0.005	0.776	0.076	0.769	0.069	0.748	0.048
	(0.100)		(0.103)		(0.310)		(0.272)		(0.030)	
X1 missing	0.720	0.020	0.720	0.020	4.540	3.840	5.898	5.198	0.748	0.048
	(0.095)		(0.097)		(4.760)		(0.761)		(0.040)	
X2 missing	0.975	0.275	0.973	0.273	0.829	0.129	0.835	0.135	1.001	0.301
	(0.097)		(0.099)		(0.313)		(0.292)		(0.030)	
X3 missing	0.704	0.004	0.705	0.005	0.726	0.029	0.724	0.024	0.767	0.067
	(0.102)		(0.103)		(0.316)		(0.273)		(0.031)	
data-set 4: Skewness=0.75										
Naïve model	0.906									
	(0.102)									
Correct model	0.704	0.004	0.723	0.023	0.371	0.329	0.118	0.582	0.724	0.024
	(0.106)		(0.109)		(0.363)		(0.856)		(0.035)	
X1 missing	0.687	0.013	0.687	0.013	-1.790	2.490	-5.396	6.096	0.770	0.070
	(0.102)		(0.104)		(5.692)		(0.577)		(0.032)	
X2 missing	0.930	0.230	0.932	0.232	0.441	0.259	0.155	0.545	0.968	0.268
	(0.104)		(0.106)		(0.368)		(1.072)		(0.035)	
X3 missing	0.719	0.019	0.723	0.023	0.293	0.407	-0.146	0.846	0.727	0.027
	(0.108)		(0.109)		(0.369)		(1.493)		(0.036)	

δ= estimated treatment effect; se (δ) = standard error of the estimated treatment effect;
Abs. bias = absolute value of the bias of the estimated treatment effect; ML = maximum likelihood;
cor X = correlation between the independent variables.

generate a distribution with non-normal distributions.

Each artificial data-set was analyzed using five statistical methods: (1) traditional multiple regression analysis, (2) multiple regression analysis with the propensity score, (3) the original Heckman two-step method, (4) the original Heckman method using maximum likelihood estimation and (5) the extended version of the Heckman method using SEM analysis. With traditional multiple regression analysis, an ordinary least squares multiple regression was used to estimate the parameters of the model, with Y as the dependent variable and D and $\boldsymbol{X}$ as independent variables. With the propensity score method, the estimated the propensity score was estimated using logistic multiple regression analysis with a set of $\boldsymbol{X}$ as independent variables and D as the dependent variable. Then, the PS was included in a multiple regression analysis with Y as the dependent variable and D and the PS as independent variables. See Bartak et al. (2009) for more details about this procedure. In the Heckman two-step method, the lambda was estimated and included into the multiple regression equation with Y as the dependent variable and D, $\boldsymbol{X}$ and lambda as independent variables. In the Heckman maximum likelihood estimation, the parameters of both the selection model and outcome model are estimated simultaneously using (robust) ML. With SEM, the data is analyzed simultaneously using (robust) ML, as described in the previous section with three indicators for D^*. Since the relation between D and D^* is deterministic and resembles a step-function, the partial regression coefficient γ was fixed to a very large value (γ=20). The results of all these methods are compared in situations where (1) no additional variables in $\boldsymbol{X}$ or $\boldsymbol{I}$ are included in the analysis (the naïve model), (2) all additional variables in $\boldsymbol{X}$ or $\boldsymbol{I}$ are included in the analysis (the correct model), (3) a variable related to the selection was omitted from the analysis (X_1 missing), (4) a variable related to both selection and outcome was omitted from the analysis (X_2 missing), and (5) a variable related to only the outcome was omitted from the analysis (X_3 missing).

The estimated treatment effect, the standard error of the estimated treatment effect, and the absolute bias were used to compare the performance of all methods. R was used to simulate the three artificial data-sets. R, M*plus*, version 5.21 (Muthén & Muthén, 2008) and STATA (StataCorp, 2001) were used to analyze the data (R Development Core Team, 2005; Muthén & Muthén, 2008; StataCorp, 2001).

5.5.2 Results of the analysis on the four artificial data-sets

The results of the analysis on the four artificial data-sets are given in table 5.2. In artificial data-set 1, the correlations between the independent variables were equal to zero. The naïve estimate of the treatment effect, without including independent variables, was 1.026 (se 0.029) and not equal to the true treatment effect of 0.7. As expected, both multiple regression adjustment and the propensity score give unbiased results when all independent variables are included in the model or when a variable is missed in the analysis that relates only to the selection or only to the outcome. However, when a true confounder variable is missed in the model that relates both to the selection and to the outcome value, these methods give biased results. As expected, the Heckman method and SEM analysis give rather unbiased results when a true confounder variable is missed in the analysis. The Heckman method is more sensitive to misspecification of the selection model than SEM analysis.

In artificial data-set 2, the correlations between the three independent variables X_1, X_2 and X_3 were equal to 0.5. Since both the Heckman method and SEM analysis rely on the assumption that the missing confounder variables is independent of the other independent variables in the model, it was no surprise that, by analyzing data-set 2, both methods provide biased results when a confounder variable (X_2 missing) or a variable only related to the outcome (X_3 missing) are omitted from the analysis. This bias and the standard error of the estimated treatment effects are, however, much smaller for SEM analysis then for the Heckman method.

In artificial data-sets 3 and 4 the distribution of the error-terms were simulated with a non-normal distribution and with no correlation between the independent variables X_1, X_2 and X_3. In these situations, traditional multiple regression analysis and the propensity score method follow the same pattern of results as in artificial data-set 1, with slightly larger standard errors of the estimates. Both the Heckman method and SEM analysis fail when a true confounder is missed in the analysis and for the Heckman method when the selection model is misspecified. The effects of non-normal distributed error-terms is much larger for the Heckman method compared to SEM analysis.

To summarize, in situations where a true confounder is missed in the analysis that is unrelated to other variables in the model, both the original Heckman method and the Heckman method provide unbiased results. Compared to the

original Heckman method, its extended version using SEM analysis provides much more robust estimators of the treatment effect in situations where the missed confounder variable relates to other variables in the model. This is explained by the fact that a part of rho is explained in the model by the inter-correlations. As in the Heckman model the estimation of the lambda term depends heavily on the normality assumption of the error-terms; this method is less robust compared to SEM analysis regarding violations of this assumption. SEM analysis relies less heavily on this assumption and the model can easily be extended and changed according to the specific design of the research.

In the next section, a real world example will be discussed that compares the results of traditional multiple regression analysis and matching on the PS with the results of the robust maximum likelihood Heckman method and the extended version using SEM analysis. The data comes from a large a Dutch research project named the 'Study on cost-effectiveness of personality disorder treatment' (*SCEPTRE*) (Bartak et al., 2009, 2010).

5.6 Case study

5.6.1 Method

Patients were recruited from six mental health care centers 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 centers from March 2003 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 of tests including self-report questionnaires. A semi-structured 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 could be calculated. For illustrative purposes this sample is divided into two groups: one allocated to short-term therapy (up to six months), the other group to long-term therapy (more than six months) (Bartak et al., 2009).

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. In this study, the Global Severity Index

(GSL) of the SCL-90 (the mean score of all 90 items) is used as the primary outcome measure, with higher scores indicating more distress (Arrindell & Ettema, 2003; Derogatis, 1986). To measure the type and degree of personality pathology, the four higher-order factors of the 'Dimensional assessment of personality pathology basic questionnaire', Dutch version (DAPP-BQ) are used : emotional dysregulation, dissocial behavior, inhibition and compulsivity (Livesley & Jackson, 2002; Kampen, 2002). Psychosocial functioning was measured with the 'Outcome questionnaire 45', Dutch version (OQ-45) (Lambert et al., 1996).

Of this self-report measure, two sub-scales were used: interpersonal relations and social-role functioning. Health-related quality of life was assessed with the EuroQoL EQ-5D (Brooks et al., 2003). Personality disorders were assessed with the 'Structured interview of DSM-IV personality', Dutch version (SIDP-IV) (Pfohl et al., 1997). The severity of personality pathology was measured by five higher-order domains of the 'Severity indices of personality problems' (SIPP): self-control, social concordance, identity integration, relational functioning and responsibility (Andrea et al., 2007; Verheul et al., 2008). For the indicators of the latent variable 'tendency to participate in long-term psychotherapy', two scales of the 'Motivation for treatment questionnaire' (MTQ-8) were included which both measure motivation to treatment: need for help and readiness to change, and a measurement of the individual wish for treatment duration (short versus long) (Beek & Verheul, 2008). The indicator variable 'wish for treatment duration' was not available for 311 patients. These patients were excluded from the analysis, leaving 438 patients.

5.6.2 Statistical analysis

To obtain a 'naïve' idea of the results before adjustment for both overt and hidden bias, the mean outcomes between the two treatment groups were compared using a multiple regression analysis. Here, the GSI score was used as the dependent variable and treatment duration as the independent variable. Correction of the treatment effect for observed pre-treatment differences and thus for overt bias, was done by; (1) multiple regression analysis using a dummy variable of treatment duration (short versus long) along with all variables relating to outcome as independent variables, and (2) the PS method as described by Bartak et al. (Bartak et al., 2009). Here, the PS is estimated by a logistic multiple

Table 5.3: Overview of variables relating to the selection and/or to the outcome value

	Variables relating to selection P-value < 0.10	Variables relating to outcome P-value < 0.10
Age, years	X	
Personal Pathology (DAPP-BQ)		
Emotional dysregulation	X	X
Dissocial behavior		
Inhibitedness		X
Compulsivity		
Quality of Life (EQ-5D)		X
Psychological capacities (SIPP)		
Self-control	X	X
Social concordance	X	X
Identity integration	X	X
Relational functioning	X	X
Responsibility	X	
Psychiatric symptomatology (SCL-90)	X	X
Functioning (OQ-45)		
Interpersonal functioning	X	X
Social role functioning		X
Axis II diagnosis (SIDP-IV)		
Number of cluster A disorders	X	X
Number of cluster B disorders	X	X
Number of cluster C disorders	X	X
Gender	X	
Civil Status	X	
Living situation	X	
Childcare	X	
Work situation		X
Level of education		
Previous outpatient treatment		
Previous inpatient treatment		X
Previous medication treatment		
Alcohol abuse	X	
Drugs abuse	X	

regression analysis with the dummy variable 'treatment duration' as the dependent variable and all variables related to outcome as independent variables. This PS is included into a multiple regression analysis with treatment duration and the PS as independent variables and the GSI outcome value as the dependent variable. To correct for possible hidden and overt bias, both the original Heckman two-step method using robust maximum likelihood (Heckman, 1979) and the extended version using SEM analysis was used. The statistical package STATA (StataCorp, 2001) with modules psmatch2 for matching, treatreg for the Heckman method and the statistical package M*plus* (Muthén & Muthén, 2008) for the SEM analysis were used.

5.6.3 Results

The 'naïve' effect of short versus long psychotherapy was 0.240 (95% CI: 0.120–0.361, p-value < 0.001). With multiple regression analysis including all variables relating to the GSI score, the estimated treatment effect was 0.131 (95% CI: 0.014-0.249, p-value < 0.05). With the propensity score method, all vari-

ables relating to the outcome value (see table 5.3) were used for the PS estimation. After correction on the PS all important pre-treatment variables were balanced among the treatment groups. For further information see (Bartak et al., 2009). The estimated treatment effect using multiple regression analysis with PS was 0.13 (95% CI: 0.129-0.250, p-value < 0.001). With the robust ML Heckman method, all variables related to the selection were included in the selection model and the variables relating to the outcome value were included in the outcome model (see table 5.3 for the variables relating to the outcome and/or to the selection). The multiple regression coefficient of lambda was 0.15 (95% CI: -0.055–0.346, p-value > 0.05), but not significant ($p > 0.05$). The correlation between the error-terms of the selection and outcome equation (rho) was 0.243 ($\chi^2_{df=1}$=1.99, p-value > 0.05). After including lambda in the regression model, the estimated treatment effect was -0.08 (95% CI: -0.395-0.238, p-value > 0.05), but not significant. According to the VIF-index (VIF > 10) there was multicollinearity between the lambda term and the dummy variable indicating treatment, resulting in very high standard errors. The exclusion of one relevant variable in the selection equation influenced the estimated treatment effect to a large extent and even, in some cases, changed the sign of the treatment effect. This implies that the Heckman method, in this case, was very sensitive to the specification of the selection model. With SEM analysis, indicators for the latent variable where two scales of the 'Motivation for treatment questionnaire' (MTQ-8), namely 'need for help' and 'readiness to change', and the measurement of the individual wish for treatment duration (short versus long). All variables related to the selection influenced this latent variable. The latent variable, in turn, influenced the three indicator variables and the treatment variable. To approach a deterministic step-function, the partial (logistic) regression coefficient of the latent variable and the treatment variable was, arbitrarily, equal to the high value of 20. The GSI outcome value was, in turn, influenced by the treatment variable and all other variables that influenced the outcome value (see table 5.3). The estimated treatment effect was -0.112 (95% CI: -0.394-0.170, p-value > 0.05). The correlation between the latent variable and the outcome value given all other variables was estimated as 0.027 (p-value > 0.05). To summarize, although with the traditional methods favored the short-term therapy, the methods countering for hidden bias did not confirm this conclusion. However, these results should only be interpreted as

an illustration rather than a relevant clinical message.

5.7 Conclusions

In psychotherapeutic research, a lot of attention is paid to methods that control for overt bias in quasi-experimental study designs, such as multiple regression analysis and propensity score methods. Nevertheless, even though researchers may be very careful in the selection of variables they want to include in the study, it seems almost impossible to account for every possible confounder. Reasons such as financial constraints, time and, even more importantly, no knowledge of variable effects mean that important variables go unmeasured. This chapter discusses statistical methods to overcome hidden bias. It discussed the traditional Heckman approach and presented its extended version of using structural equation modeling. By analyzing artificial data-sets and a real world example, the performance of both methods were compared to the traditional methods.

The simulation study confirms that methods dealing with overt bias, such as traditional multiple regression analysis and the PS method, fail when a true confounder is missed in the analysis. In the unique situation where this missing confounder does not correlate with all other independent variables in the model, both the Heckman method and SEM analysis provide unbiased results. When this correlation exists, however, both the Heckman method and its extended version using SEM fail to correct for the hidden bias in the study. The Heckman method is, however, especially sensitive to misspecification of the selection model and to violations of the assumption of normal distributed error-terms. Overall, SEM is less sensitive to these assumptions and provides less bias compared to the Heckman method. As in SEM analysis all relations between the variables can easily be modelled, it is a much more flexible model. In the illustrative example, methods assuming overt bias revealed a positive effect for short-term treatment compared to long-term treatment. Nevertheless, both the robust Heckman maximum likelihood method and the extended version using SEM could not reject the null-hypothesis of no treatment effect. Based on the simulation study and the illustrative example, it is concluded that the strong reliance on normally distributed error-terms along with the independency assumption implies that the Heckman method is difficult to use in

practice. When good indicators for the latent tendency to participate in the study are available, the extended version of the Heckman method using SEM analysis could be considered as a feasible alternative.

Chapter 6

Latent class analysis of experimental data under non-compliance*

6.1 Summary

In randomized experiments, the equivalence of the treatment and control groups may be threatened when subjects fail to comply to their instructions in the various groups. Traditional methods for handling differential non-compliance behavior like Intention-to-treat, Analysis-as-treated, or Per-protocol analysis have been shown to be defective in several respects. An alternative is the Instrumental variable approach which yields an unbiased estimate of the complier average causal effect. This approach can be recast in terms of a latent class model. In the present chapter several extensions of that latent class model are presented. These extensions pertain to situations in which (a) the outcome variable is only measured indirectly via indicator variables, (b) the experimental intervention has more than two levels, and/or (c) a factorial design has been implemented. The methods proposed in this chapter are applied to data from an experiment that studied the effects of various physical programs on the cognitive functioning in the elderly.

*This chapter has been submitted.

6.2 Introduction

In experimental settings where subjects are assigned to either a reference or an experimental group, participants do not always strictly follow their treatment instructions or requirements. Non-compliance occurs when a patient fails to fulfill the requirements of a prescribed treatment condition. Because of some patients' non-compliance, an originally randomized experiment may loose the characteristics of a true experiment, and more appropriate statistical methods than the standard ones are needed in order to estimate the effect of a treatment.

Consider an experiment with one experimental and one reference group, and assume that each participant in the study is randomly allocated to one of the two groups. As a consequence of this randomization process, one may expect that participants in the experimental and reference group are, on average, comparable at the start of the experiment. Let R denote this random assignment to the experimental groups, where $R = 0$ denotes random assignment into the reference group, and $R = 1$ random assignment into the experimental group. Within the context of Rubin's causal model (Rubin, 1974), one may define two potential outcomes which would be observed if a subject were assigned to a particular condition. Let Y_0 denote the potential outcome when a subject assigned to the reference group, and let Y_1 represent the same subject's potential outcome when assigned to the experimental group. The two variables Y_0 and Y_1 are not both observed: for subjects in the reference condition only the scores on Y_0 are observed, whereas for subjects in the experimental condition only the scores on Y_1 are available.

For estimating the effect of the intervention, the assumption is made that the potential outcomes (Y_0, Y_1) are independent of the assignment R, given covariates $\boldsymbol{X}$. This ignorability assumption (Rubin, 1974) can be represented as

$$(Y_1, Y_0) \perp R | \boldsymbol{X}. \tag{6.1}$$

Because each participant is only observed in one condition, individual effects of the treatment cannot be determined. In order to estimate the average causal effect (ACE), inferences have to made about the mean outcome of participants in the condition they were not assigned to. This mean is called the counterfactual mean. When the ignorability assumption is valid, the average causal effect

(ACE) can be determined by comparing the mean outcome of the reference group with the mean outcome of the experimental group. The mean outcome of the reference group is then used as an estimate of the counterfactual mean of the experimental group. Let δ denote the ACE of a treatment, then

$$\delta = E(Y_0) - E(Y_1) \ . \tag{6.2}$$

A major problem is, however, that patients do not always follow their treatment requirements. For example, participants who are randomly assigned into the experimental condition may not show up in the intervention, or reference participants may obtain the treatment outside the experimental set-up. As a consequence, the treatment actually received by participants is not always the same as the treatment they were assigned to, and the estimation of the ACE as in equation 6.2 could yield biased results.

Intention-to-treat analysis (ITT) is the most frequently used method to estimate treatment effects in a study with non-compliance. ITT is a strategy for the analysis of randomized controlled studies that compares patients in the groups to which they where originally randomly assigned. This means that all the patients are included in the statistical analysis, regardless of the (amount of) treatment actually received, or even regardless of whether they withdrew or not from the study. Let $E(Y|R = 0)$ and $E(Y|R = 1)$ denote the means or expected values of the outcome variable of participants assigned to the reference group and experimental group, respectively. ITT then estimates the ACE as

$$\delta_{ITT} = E(Y|R = 1) - E(Y|R = 0). \tag{6.3}$$

Although ITT is a frequently used method, it has some clear drawbacks (Nagelkerke, Fidler, Bernsen, & Borgdorff, 2000). For example, when people who experience better results are more likely to follow their treatment requirements, ITT may yield biased estimates of the treatment effect. Bias may also occur when different side-effects, resulting from different treatment conditions, influence compliance behavior.

A first alternative to ITT is the 'as-treated' analysis (AT), where participants are classified by the treatment actually received, instead of by the treatment they were assigned to. Let D denote the treatment actually received by participants, where $D = 0$ denotes that they actually received the reference

condition and $D = 1$ that they actually received the treatment. In AT the ACE is estimated as

$$\delta_{AT} = E(Y|D = 1) - E(Y|D = 0) \tag{6.4}$$

where $E(Y|D = 0)$ and $E(Y|D = 1)$ denote the mean outcomes in the groups receiving the reference and the experimental condition, respectively. Unfortunately, this method is also not without its own problems. For example, when the most treatment resistant patients in the less effective treatment groups switch to the more effective treatment, they may decrease the average level of improvement for the more effective treatment and the ACE will be underestimated.

In stead of the ITT or AT estimator of the causal effect, the 'per protocol' estimator (PP) is occasionally being used. This estimator, which is defined as

$$\delta_{PP} = E(Y|R = D = 1) - E(Y|R = D = 0) \ , \tag{6.5}$$

only considers data from patients who completed the treatment protocol as originally planned.

As an alternative to the ITT, AT, and PP analysis, Angrist, Imbens, and Rubin (1996) proposed an instrumental variables approach for estimating causal effects when assignment to a binary treatment is randomized, but compliance with the treatment is not perfect. Suppose that in a regression analysis of Y on X, the error term may be correlated with X so that ordinary regression analysis yields a biased estimate of the regression coefficient. However, suppose that a third variable Z is available that it is uncorrelated with the error term but has a non-zero covariance with X. Such a variable is called an instrumental variable. An unbiased estimate of the regression coefficient of X is then given by

$$\frac{cov(Y, Z)}{cov(X, Z)} \ .$$

In Angrist et al. (1996) the binary variable R is treated as an instrumental variable for estimating the average causal effect of D on Y in the subpopulation of compliers, the complier average causal effect (CACE). Their estimator is given by

$$\delta_{CACE} = \frac{E(Y|R=1) - E(Y|R=0)}{E(D|R=1) - E(D|R=0)} \ . \tag{6.6}$$

Angrist et al. (1996) prove that under some assumptions their approach provides an unbiased estimate of the CACE. These assumptions are:

* The stable unit treatment value assumption, which states that the potential outcomes for a particular individual are unrelated to the treatment status of other individuals;

* Random assignment: Individuals are randomly assigned to the control and treatment conditions;

* Exclusion restriction: The response distribution of Y is independent of R given D, so that the only effect that R has on Y is via D;

* The variable R has an effect on D so that they are not completely independent;

* Monotonicity: No individual does exactly the opposite of his assignment.

In Angrist and Imbens (2005), similar results were obtained when the treatment variable has several intensity levels.

The aim of the present chapter is to discuss the non-compliance model from a latent variable point of view. After discussing how the instrumental variable approach of Angrist et al. (1996) can be rephrased as a latent class model, several extensions of this basic model will be introduced. First, the situation in which the outcome status is not measured by a single variable but by several outcome variables will be discussed. In this extension of the model, the basic outcome variable will be treated as a latent variable measured via a set of indicator variables. Next, the situation in which the treatment variable has more than two levels will be considered. Finally, it will be shown how the basic latent class model can be extended to cover non-compliance in factorial experiments. In a final section a combination of the various extensions of the latent class model will be applied to a real data set.

6.3 The instrumental variables approach as a latent class model

That the instrumental variables approach can be formulated in terms of a latent class model was already made clear by Imbens and Rubin (1997a), Little and Yau (1998), and Forcina (1975). Angrist et al. (1996) distinguished between four types of subjects with respect to their compliance behavior: compliers, never-takers, always-takers, and defiers. *Compliers* are subjects who always strictly follow their treatment instructions and for which then $D = R$ holds. When compliers are assigned to the experimental group ($R = 1$), they actually receive the experimental treatment ($D = 1$); when they are assigned to the reference condition ($R = 0$), they receive the reference treatment ($D = 0$). *Never-takers* are subjects who never undergo the experimental treatment, even when assigned to the experimental condition; for never-takers always $D = 0$ holds, whatever the value of R. *Always-takers*, on the other hand, are subjects who always undergo the experimental treatment, even when assigned to the reference condition; for these subjects always $D = 1$ holds, whatever the value of R. Finally, *defiers* are those individuals who do exactly the opposite of their assignment; for these subjects $D = 1 - R$ holds.

In the most general latent class formulation, these four types of subjects are considered as the values of a nominal latent variable Compliance type C; they represent four latent classes or sub-populations of subjects. Table 6.1 shows how the conditional probability $p(D = 1|C = c, R = r)$ of receiving the treatment varies as a function of latent class membership C and assigned treatment R, where the first latent class($C = 1$) denotes compliers, the second latent class ($C = 2$) never-takers, the third class ($C = 3$) always-takers and the fourth class ($C = 4$) defiers.

Table 6.1: Conditional probability $p(D = 1|C = c, R = r)$ as a function of latent class membership C and assigned treatment R

	R=0	R=1
C=1	0	1
C=2	0	0
C=3	1	1
C=4	1	0

All the conditional probabilities $\Pr(D = 1|C = c, R = r)$ are known constants being equal to either zero or one. The conditional probabilities $\Pr(D = 0|C = c, R = r)$ of not receiving the experimental treatment are the complements of the former.

Let π_1, π_2, π_3 and π_4 denote the population proportion of compliers, never-takers, and always-takers, and defiers, respectively. These latent class probabilities π_i for $i = 1, \cdots, 4$ are unknown parameters and, hence, have to be estimated.

The variables R and D are observed but the variable C, representing latent class membership, is not. Moreover, knowledge of a participant's values on R and D does not allow to unequivocally determine the latent class he belongs to. Table 6.2 contains the theoretical allocation probabilities $\Pr(C = c|R = r, D = d)$, i.e. the probability that someone belongs to a particular compliance status class given his response pattern on R and D.

Table 6.2: Allocation probabilities $p(C = c|R = r, D = d)$

R	D	C=1	C=2	C=3	C=4
0	0	$\frac{\pi_1}{\pi_1+\pi_2}$	$\frac{\pi_2}{\pi_1+\pi_2}$	0	0
0	1	0	0	$\frac{\pi_3}{\pi_3+\pi_4}$	$\frac{\pi_4}{\pi_3+\pi_4}$
1	0	0	$\frac{\pi_2}{\pi_2+\pi_3}$	$\frac{\pi_3}{\pi_2+\pi_3}$	0
1	1	$\frac{\pi_1}{\pi_1+\pi_4}$	0	0	$\frac{\pi_4}{\pi_1+\pi_4}$

From table 6.2 it is clear that knowing the value of R and D for a subject is not sufficient to determine his compliance status. For each combination of values for R and D a subject may still belong to two of the four classes so that the compliance status of any subject can only partially be deduced from his compliance behavior. For example, a subject who was randomized into the reference group ($R = 0$) but did not receive the treatment ($D = 0$) can either be a complier or a never-taker.

To obtain an applicable latent class model for non-compliance in experimental studies, further assumptions about the distribution of the response variable Y and the latent class structure are needed. Here it is assumed that Y is a metric variable, so that its mean or expected value can be defined. In principle, the expected value of Y for a particular participant may depend on the value of the latent class C that participant belongs to, the treatment R he was assigned

to, and the treatment D he actually underwent. So, in general $4 \times 2 \times 2$ different means μ_{crd} could be considered. However, the exclusion restriction made by Angrist et al. (1996) states that any effect that R has on Y is an indirect one over D. So, the expected values of Y only depend on the latent class C and effective treatment D, and can be represented by the symbol μ_{cd}. Moreover, in the class of the never-takers (Class 2) only μ_{20} is identified since a participant in this class is never observed under $D = 1$. Similarly, participants in the class of always-takers (Class 3) are never observed under $D = 0$, so that for this class only μ_{31} is identified. Table 6.3 shows which expected values of Y remain to be estimated.

Table 6.3: Expected values outcome variable as a function of class and effective treatment

Class	D=0	D=1
Compliers	μ_{10}	μ_{11}
Never-takers	μ_{20}	-
Always-takers	-	μ_{31}
Defiers	μ_{40}	μ_{41}

In the class of compliers both the mean μ_{10} of the reference condition and the mean μ_{11} of the experimental condition can be estimated. The difference $\mu_{11} - \mu_{10}$ is an estimate of the complier average causal effect. In a similar way one could define the DACE (defier average causal effect) as $\mu_{41} - \mu_{40}$. However, the monotonicity condition made by Angrist et al. (1996) is equivalent to the assumption that the class of defiers is empty, so that $\pi_4 = 0$ in the present notation, and the DACE is not identified. In general then, one is only interested in estimating the CACE. In the approach of Angrist et al. (1996) the randomized assignment variable R acts as an instrumental variable for estimating the effect of D on Y.

The full four-class formulation was already discussed by Imbens and Rubin (1997b) who proposed a Bayesian approach to estimate the causal effect in experiments with non-compliance. Also Little and Yau (1998) and Yau and Little (2001) used the latent class framework for the estimation of the CACE. Their analysis is based on a latent class model that is both a simplification and an extension of the model described above. It is a simplification because only two of the four classes are retained in the model: both the classes of always-takers

and defiers are assumed to be empty. On the other hand, their model is an extension of the latent class model described above since it also incorporates explanatory covariates on which the expected value of the outcome variables Y and the compliance probability π_1 may depend. Assuming multivariate normality for the conditional distribution of Y given the covariates, maximum likelihood estimates of the model parameters are obtained by means of the EM algorithm treating latent class membership as missing data. Imbens and Rubin (1997a) showed that, assuming a normally distributed outcome variable with equal variance in all classes, the instrumental variable approach is also capable of estimating the marginal distributions of the outcome distributions of treatment and control subjects in the subpopulation of compliers. However, neither, the joint distribution of Y_0 and Y_1 nor the distribution of their difference $Y_1 - Y_0$ are identified, but, as the authors remarks, this is also not possible in randomized between-subjects experiment where all subjects comply to their treatment assignment.

Nagelkerke et al. (2000) also used assignment treatment R as an instrumental variable to estimate the effect of the actual treatment D on the outcome variable Y in the presence of confounding variables that affect both D and Y. Let E be the residual variable from the regression analysis of D on R. They show that, if there is no other effect of R on Y other than via D, and if the confounders do not moderate the effect of D on Y, the effect of D on Y can be estimated in a regression analysis by including E as an additional covariate in the model. A similar approach was taken by Ten Have, Joffe, and Cary (2003) for estimating the marginal causal log-odds ratio for binary outcomes under treatment non-compliance in a randomized trial. As Little and Yau (1998), they assumed a two latent class model for non-compliance: compliers and never-takers. In their model they also incorporated covariates for response and compliance type latent class membership.

6.4 Fitting latent class models for data with non-compliance

In this chapter, several extensions of the basic latent class model for non-compliance will be presented. All these models can be formulated as mixture regression models and be fitted to real data by means of M*plus* 5.0 (Muthén

& Muthén, 2008). By considering the model depicted in figure 6.1 the basic principles of this estimation procedure that makes use of so-called training variables are illustrated.

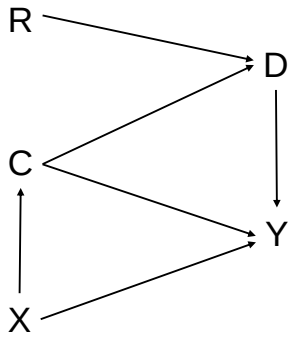

Figure 6.1: The instrumental variable model as a latent class model

In figure 6.1, R represents the randomly assigned treatment condition with $R = 0$ for the reference and $R = 1$ for the experimental condition. C represents the latent class variable representing compliance behavior. Although later some of these classes will be considered empty, it is initially assumed that the nominal variables C has four different values representing compliers, never-takers, always-takers and defiers, respectively. The symbol $\boldsymbol{X}$ represents a set of explanatory variables, some of which may affect the probability of belonging to a particular latent class while others may affect only the distribution of the outcome variable Y. Since its values are randomly defined, R is independent of both C and $\boldsymbol{X}$. The variable D represents the effective treatment a participant was subjected to with $D = 0$ for no intervention and $D = 1$ for intervention. As shown in table 6.1 above, D is a deterministic function of C and R. The outcome variable Y is affected by D and some of the explanatory variables in

the set represented by $\boldsymbol{X}$. The subset of variables from $\boldsymbol{X}$ having an effect on Y may partially or completely overlap with those affecting latent class membership. Since the variable R has no direct effect on Y, but only an indirect one over D, the model depicted here satisfies the exclusion restriction.

For a continuous outcome variable Y, it is generally assumed that its conditional distribution is normal with its expected value a linear function of D and the relevant predictor variables from $\boldsymbol{X}$. If D is taken as a binary variable with $D = 0$ for the reference condition and $D = 1$ for the experimental condition, the relationships between the expected value of Y and D and $\boldsymbol{X}$ are class specifically given by a series of four regression equations:

$$\begin{aligned} E(Y|\boldsymbol{X} = x, C = 1) &= \alpha_{01} + \delta_1 D + x'\alpha_1 \ , \\ E(Y|\boldsymbol{X} = x, C = 2) &= \alpha_{03} + x'\alpha_2 \ , \\ E(Y|\boldsymbol{X} = x, C = 3) &= \alpha_{03} + x'\alpha_3 \ , \\ E(Y|\boldsymbol{X} = x, C = 4) &= \alpha_{04} + \delta_4 D + x'\alpha_4 \ . \end{aligned} \tag{6.7}$$

Since the participants in the second and third class are only observed under one condition (either the reference or experimental condition, respectively), the effect of D is not defined in these two classes. The effect of D can only be assessed in the first class of compliers and the fourth class of defiers. The estimate of the effect in the class of compliers is given by δ_1 and is actually the complier average causal effect (CACE) as referred to above. Similarly, δ_4 is the defier average causal effect (DACE), which is not necessarily equal to the CACE. Under the monotonicity condition, which implies that there are no defiers, the DACE is not identified; in this case, only the CACE can be estimated. Note also that in this general model the effects of the auxiliary explanatory variables $\boldsymbol{X}$ on Y may vary over the classes. Neither have the regression models to be homoscedastic.

The probability of belonging to a particular latent class may also depend on some or all of the explanatory variables $\boldsymbol{X}$. In general, a multinomial regression model is postulated for this relationship:

$$\Pr(C = c|\boldsymbol{X} = x) = \frac{\exp(\beta_{0c} + x'\beta_c)}{\sum_k \exp(\beta_{0k} + x'\beta_c)} \ .$$

To obtain an identified model the β-parameters for one of the classes are all set equal to zero.

The latent class model as described here is an example of a mixture regression model (Wedel & DeSarbo, 2002). In the present application, the observed variables R and D provide partial information about class membership, since each response pattern (R, D) can only occur in two of the four classes. In order to fit the latent class model by means of a software program like M*plus* (Muthén & Muthén, 2008), a binary training or learning variable T_c has to be defined for each class as a function of the response pattern (R, D). The training variable for a particular latent class is equal to 1 for those response patterns (R, D) that can occur in that class, and is equal to 0 for response patterns that cannot occur. In order to apply M*plus*, users have to define the training variables themselves before entering them in the analysis together with the assigned treatment variable R.

Table 6.4 specifies the definition of the four training variables in terms of the observed response patterns (R, D).

Table 6.4: Definition of training variables in the general case of four latent classes

R	D	T_1 (C)	T_2 (Nt)	T_3 (At)	T_4 (D)
0	0	1	1	0	0
0	1	0	0	1	1
1	0	0	1	0	1
1	1	1	0	1	0

From this table it is seen that the response pattern $(R = 0, D = 0)$ can only occur in the classes of compliers and never-takers, but not in the classes of always-takers or defiers. Analogous observations can be drawn for the three other response pattern, showing that each response pattern can only occur in exactly two classes. If the number of latent classes is reduced, for instance, by assuming that there are no defiers or always-takers, the corresponding training variables are removed from the analysis.

The four-class model without auxiliary explanatory variables $\boldsymbol{X}$ is not identified, but the model with three classes obtained by removing the fourth class

of defiers is. A limited simulation study indicated that adding explanatory variables that determine latent class members and the conditional distribution of Y seems to resolves the identification issue for the four-class model.

For the analysis of an ordinal categorical outcome variable Y, Muthén's threshold model (Muthén, 1984) can be used. In this model a categorical variable Y is derived from a continuous unobserved variable Y^*, which is categorized via a threshold model. Suppose Y has m ordered response categories. Then for $m-1$ threshold values θ_k, the relationship between Y^* and Y is as follows:

$$\begin{aligned} Y = 1 &\Leftrightarrow Y^* \leq \theta_1 \\ Y = k &\Leftrightarrow \theta_{k-1} < Y^* \leq \theta_k \text{ for } 2 \leq k \leq m-1 \\ Y = m &\Leftrightarrow \theta_{m-1} \leq Y^* \ . \end{aligned}$$

In this case, the regression models are formulated for the continuous latent variable Y^* rather than for the categorical observed variable Y.

6.5 Extensions of the basic latent class model

In this section several extensions of the basic latent class model for non-compliance are discussed:

1. The latent class model with an indirectly measured outcome variable;
2. A non-compliance latent class model for multiple treatments;
3. A non-compliance latent class model for a factorial design.

6.5.1 Estimating the CACE when the outcome status is indirectly measured via indicator variables

In many experimental studies the final outcome variable is not directly measured by a single variable, but by a set of several indicator variables which can be thought of as imperfect measures of a underlying latent construct. In such studies one is not interested in the effect of the treatment on each of the indicator variables, but in its effect on the underlying latent construct. In order to accommodate for data collected in such experimental design, the basic latent

class model can be extended by including a measurement model for the latent variable. This extended latent class model is shown in figure 6.2.

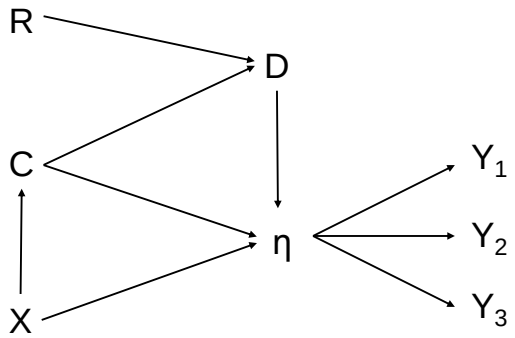

Figure 6.2: The non-compliance latent class model with latent outcome variable

In this figure the unobserved latent outcome variable is represented by η, for which three observed indicator variables Y_1, Y_2 and Y_3 are available. These three indicator variables are independent conditional on η; moreover, they are also conditionally independent of $R, C, \boldsymbol{X}$ and D given η. This implies that effective treatment D has no direct effects on any of the indicator variables, but only indirect effects that run over η. In general, the latent variable η will be treated as a continuous variable, but some or all of the indicator variables may be categorical. In this adaptation of the basic latent class model the effective treatment D and the explanatory $\boldsymbol{X}$ have a direct effect on the latent outcome variable η, but not on the indicator variables Y_k themselves. The regression equations from equation 6.7 have to be modified accordingly. When the indicator variables are continuous, their relationship with the latent outcome variable is usually specified as a factor analytic model with one common factor:

$$Y_k = \alpha_k \eta + \upsilon_k \ .$$

In order to obtain an identified measurement model, the scale of the latent variable η has to be fixed, either by setting the factor loading of one of the indicator variables equal to 1, or by setting the variance of η equal to 1. Not all indicator variables need to be continuous. For categorical indicator variables this factor analytic model has to be combined with the same threshold model as described above.

6.5.2 A non-compliance latent class model for multiple treatments

In many experimental studies several experimental conditions are compared to a reference condition. Suppose that the experimental manipulation of the independent variable resulted in three different conditions: $R = 0$ for a reference or placebo condition, $R = 1$ for a first experimental condition, and $R = 2$ for a second one. The two experimental treatments may, for instance, differ with respect to the intensity with which a particular stimulus is applied to the participants, or with respect to the dose of a pharmaceutical drug administered to patients in a clinical study. On the other hand, the two treatments may represent qualitatively different interventions that cannot be ordered along a quantitative intensity continuum. The effective treatment D assumes the same three values as R, but which value D assumes for a particular participants depends on the latter's compliance status. If the participant belongs to the class of compliers, $D = R$ holds. If, on the other hand, a participant is a never-taker one may assume $D = 0$, whatever the value of R. It is less straightforward to define D for the classes of always-takers or defiers, since R assumes more than two values. What, for instance, would an always-taker do in this context? Would he always opt for $R = 2$ or $R = 1$? It is even less clear how a defier would behave in this situation.

So, in a multiple treatment situation it may be wise to consider only models with the two latent classes consisting of compliers and never-takers. Only two training variables T_1 and T_2 have then to be defined given the response patterns (R, D). This is shown in the following table 6.5.

Note first that the response patterns (0,1), (0,2), (1,2), and (2,1) for (R, D) can never occur when there are no always-taker or defiers. Moreover, the response patterns (1,1) and (2,2) can only occur in the class of compliers, whereas the patterns ((1,0) and (2,0) can only be observed in the class of never-takers.

Table 6.5: Definition of training variables for a two-class model in an experiment with three condition

R	D	T_1	T_2
0	0	1	1
1	0	0	1
1	1	1	0
2	0	0	1
2	2	1	0

Finally, the response pattern (0,0) is possible in both latent classes.

Let $\boldsymbol{X}$ represent a set of explanatory variables that may affect the probability of belonging to a particular latent class as well as the expected value of the outcome variable Y, the model can be further specified by the equations

$$logit(\Pr(C = 1|x_i)) = \alpha_0 + x_i'.\alpha \ , \tag{6.8}$$

$$E(Y_i|C = 1, x_i) = \beta_0 + \delta_1 D_1 + \delta_2 D_2 + x_i'.\beta \ , \tag{6.9}$$

$$E(Y_i|C = 2, x_i) = \gamma_0 + x_i'\gamma \ . \tag{6.10}$$

The first equation defines the allocation model by indicating how the probability of being a complier varies as a function of the explanatory variables in $\boldsymbol{X}$. A potential explanatory variables can be excluded from this allocation model by setting its α parameter equal to zero. The second equation specifies how in the class of compliers the expected value of the outcome variable Y varies as a function of the received treatment, after controlling for the relevant explanatory variables. In this equation the effective treatment is represented by two dummy variables D_1 and D_2 with D_1 equal to 1 for $D = 1$ and zero otherwise, and D_2 equal to 1 for $D = 2$ and zero otherwise. In this way the reference condition is treated as the reference category with which the effects of both experimental treatments are compared. In the class of never-takers, the expected value of Y only depends on the explanatory variables, but their effects on Y may be different from those in the first class.

6.5.3 A non-compliance latent class model for a factorial design

Non-compliance problems may also arise in factorial experiments in which two or more experimental factors are crossed. Consider a 2×2 design in which two factors A and B are orthogonally crossed. In principle, one could now define $4 \times 4 = 16$ different classes by combining the four compliance classes for factor A with the four compliance classes for factor B. Such an extended latent class model will not work in practice, even if it is identified by including additional explanatory variables in the model. A more applicable latent class model is obtained by only allowing a complier and a never-taker class per factor. Let R_A and R_B denote the randomly assigned levels for factor A and B, respectively. In a similar way, let D_A and D_B be the received levels for factors A and B, respectively. The two values of the variables R and D are continually denoted as 0 and 1.

By combining the two compliance status classes for A and B, four different classes can be defined:

* Class 1 consists of participants who comply to both the A and B treatment and for which then $D_A = R_A$ and $D_B = R_B$ hold;
* Class 2 consists of participants who comply to the A treatment, but never take B: $D_A = R_A$ and $D_B = 0$;
* Class 3 consists of participants who comply to the B treatment, but never take A: $D_A = 0$ and $D_B = R_B$;
* Class 4 consists of the participants who never take A or B: $D_A = 0$ and $D_B = 0$.

In order to fit this model, four training variables have to be defined. They are defined as follows:

$$\begin{aligned} T_1 &= (D_A = R_A) \wedge (D_B = R_B) \\ T_2 &= (D_A = R_A) \wedge (D_B = 0) \\ T_3 &= (D_A = 0) \wedge (D_B = R_B) \\ T_4 &= (D_A = 0) \wedge (D_B = 0) \ . \end{aligned}$$

For the model that only allows compliers or never-takers for each factor, seven of the 16 possible response patterns(R_A, D_A, R_B, D_B) cannot occur. Table 6.6 describes the four training variables for the nine remaining response patterns that can occur.

Table 6.6: Definition of training variables for the two-class model in a 2×2 design

R_A	D_A	R_B	D_B	T_1	T_2	T_3	T_4
0	0	0	0	1	1	1	1
0	0	1	0	0	1	0	1
0	0	1	1	1	0	1	0
1	0	0	0	0	0	1	1
1	1	0	0	1	1	0	0
1	0	1	0	0	0	0	1
1	0	1	1	0	0	1	0
1	1	1	0	0	1	0	0
1	1	1	1	1	0	0	0

It is interesting to note that four of the eight response patterns can only appear in one latent class. For participants with these response patterns latent class membership is observed. The zero response pattern, on the other hand, can occur in all four classes, whereas the remaining four patterns can arise in two different classes.

The most obvious way to define an allocation model in this situation is by postulating a multinomial regression model for the latent class probabilities. An alternative model might consist of treating the four latent classes as the result of the crossing of two dichotomous latent variables U_A and U_B which take on the value 1 for compliance and the value 0 for non-compliance on the corresponding factor. Latent class membership can then be modeled by a separate model for $\Pr(U_A = 1)$ and $\Pr(U_B = 1)$. In the model for the outcome variable, a different regression equation has to be specified for each of the four latent classes. In the first class the regression model has to include the main effects both factor A and B, and eventually their interaction effect. In the second and third class only the main effect of either A or B is implied according to the compliance status of the subjects belonging to these classes. Finally, in the fourth class none of the effects of A and B have to be taken into account. Moreover, in all four classes partially or completely overlapping subsets of covariates may enter

the regression equations but their regression coefficients need not be invariant over the classes.

6.6 A real data application

In this section, an application of the latent class model for non-compliance in situations is discussed in which a latent outcome variable is indirectly observed via several indicators in a 2×2 factorial design. In this application, two of the extensions discussed in the previous section are combined in a single analysis. Moreover, two explanatory variables are included in the model which might have a potential effect on latent class membership as well as on the distribution of the latent outcome variable. The data come from a randomized controlled factorial trial conducted by the EMGO Institute for Health and Care Research in the Netherlands (van Uffelen, Chin A Paw, van Mechelen, & Hopman-Rock, 2008).

This study examined the effects of two different treatments, vitamin B supplementation (R_1) and aerobic exercise (R_2) on cognitive function in older adults with mild cognitive impairment. A sample of 152 participants were randomly assigned to the interventions: (1) a twice-weekly, group-based, moderate-intensity walking program (n = 77) or a low-intensity placebo activity program (n = 75) for one year; and (2) daily vitamin pill containing 5 mg folic acid, 0.4 mg vitamin B-12, 50 mg vitamin B-6 (n = 78) or placebo pill (n = 74) for one year. Cognitive functioning was measured with eight neuropsychological tests at baseline and after six and 12 months. Here, only the data is used collected at 12 months.

The Abridged Stroop color word test (SCWT-A) is used as a measure of complex processing (Klein, Ponds, & Jolles, 1997). The SCWT-A consist of three tasks; 1) SCWT-A1: word reading, 8 rows of 5 written colors; 2) SCWT-A2: color naming, naming the colors of 8 rows of 5 red, green, blue or yellow colored rectangles; 3) SCWT-A3: combination task, the words red, green, blue or yellow have been printed in a different color of ink, the subject is asked to name the color of the ink. The Auditory Verbal Learning Test (AVLT) is used as a measure of memory for direct and delayed recall (Rey & Muthén, 1964). During this test, a list of monosyllabic words is read aloud by the examiner for 5 times. After each trial, the subject is asked to repeat the words

he or she remembers. After fifteen minutes with other questions, delayed recall is assessed by asking the participant which words he or she still remembers. Both the versions AVLT15 and AVLT6 were used. General cognitive function is measured with the Mini Mental State Examination (Folstein, Folstein, & McHugh, 1975). The MMSE consists of 11 questions concerning orientation, registration, attention and calculation, recall and language. The maximum score is 30 and a score below 24 is considered abnormal for dementia screening. The digit symbol substitution test (DSST) is used as a measure of attention, perceptual speed, motor speed, visual scanning and memory (Uiterwijk, 2001). The subject is given a piece of paper with nine symbols corresponding with nine digits. Next on this piece of paper are three rows of digits with empty spaces below them. The subject is asked to fill in as many corresponding symbols as possible in 90 seconds. Expressive language was assessed using the verbal fluency test (VFT) (Lezak, 2004). The subject is given a letter and is asked to name words beginning with the particular letter in one minute.

In our model, also information about the age and gender of participants were included. Subjects with missing data were excluded, leaving a sample of 131 subjects. Thirty (30) subjects were assigned into the walking program with a placebo for vitamin intake, 33 subjects were assigned into the low-intensity placebo activity program with a placebo for vitam intake, 34 subjects were assigned into the walking program with a daily vitamin intake, and 33 subjects were assigned into the low-intensity placebo activity program with a daily vitamin intake. At baseline, the patients from the four study groups did, on average, not significantly differ on age (F=0.271, df=3, p =0.885) and gender (χ^2=6.522, p=0.089). Compliance with the walking program is assessed as the percentage of attended lessons (less than 75%). Compliance with the vitamin supplementation is verified by pill counts and determining blood vitamin levels.

6.6.1 The model

The analysis reported in this section is based on the model schematically shown in figure 6.3.

Since all elderly were randomized to either the reference or the experimental group of both experimental factors, the distributions of the randomized treatments R_1 and R_2 are perfectly known. R_1 represents allocation to the condition of the vitamin factor: $R_1 = 0$ for assignment to the placebo pill

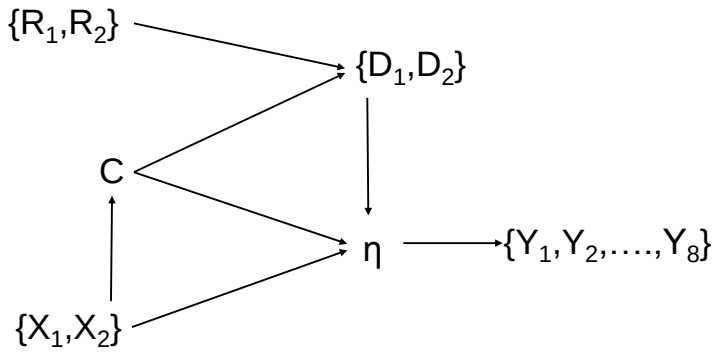

Figure 6.3: The non-compliance latent class model with two manipulated factors and eight indicators

condition, $R_1 = 1$ for assignment to the daily vitamin pill condition on this factor. The dummy variable R_2 represents random assignment to the Exercise factor: $R_2 = 0$ for the low-intensity placebo activity program condition, $R_2 = 1$ for the moderate-intensity walking program condition. For both factors compliance the effective treatments D_1 and D_2 can be treated as completely observed, given the operational definitions of compliance as stated above.

The latent outcome η represents cognitive functioning and is measured via the eight cognitive tests represented by Y_1 to Y_8. Two explanatory variables X_1 (Gender) and X_2 (Age) are included in the model. Those two explanatory variables may first determine class membership of the participants; they may also have direct effects on the latent outcome variable η. The present analysis is based on the assumption that two latent classes may be defined. The first latent class $C = 1$ contains the participants who comply on both factors; the second class $C = 2$ contains people who are never-taker on the second factor but always comply to their assignment on the first one. Note that here the latent class consisting of compliers on the first factor and never-takers for the second factor are not included. Due to the limited number of subjects in this study, analysis with three or four latent classes did not converge properly. From a substantive point of view, a selection of the two latent classes reflects the fact that compliance on the second factor requires more effort than compliance on the first one: for the vitamin intake compliance was almost 100 %, whereas for the activity program it was only 66 %.

The model being fitted to the data was specified by a series of specific assumptions:

1. The probability of belonging to the class of compliers is given by a logit model with X_1 and X_2 as explanatory variables:

$$logit(C = 1|X_1, X_2) = p_0 + p_1X_1 + p_2X_2 \ .$$

   The second latent class is treated here as the reference class.

2. The two training variables T_1 and T_2 are defined by the following logical operations:

$$\begin{aligned} T_1 &= (D_1 = R_1) \wedge (D_2 = R_2) \\ T_2 &= (D_1 = R_1) \wedge (D2 = 0) \ . \end{aligned}$$

3. In latent class $C = 1$ the expect value of η depends on both the status of the two manipulated variables and their interaction, and the explanatory variables:

$$E(\eta) = a_0 + a_1 D_1 + a_2 D_2 + a_3 D_1 \times D_2 + a_4 X_1 + a_5 X_2 \ .$$

4. In latent class $C = 2$ the expected value of η only depends on the explanatory variables:

$$E(\eta) = b_0 + b_1 D_1 + b_4 X_1 + b_5 X_2 \ .$$

5. No equality constraints are imposed on the regression coefficients of the common explanatory variables in the different class: the effects of D_1, X_1, and X_2 may be different in the two classes.

6. The latent outcome variable η is measured by eight indicator variables $Y_k, k = 1, ..., 8$. Here a factor model with one common factor is postulated:

$$Y_k = \lambda_k \eta + \epsilon_k$$

In order to obtain an identified measurement model the factor loading of the first indicator was set equal to 1. It is assumed that the same measurement model applies in all three classes.

6.6.2 Results

Predicting latent class membership Table 6.7 shows the estimates of the parameters of the logistic regression equation by means of which latent class membership is predicted on the basis of Gender and Age.
The results indicate that none of these explanatory variables has a significant effect on the probability of belonging to a particular latent class. Whether someone is a complier or not does not depend on his age or gender.

Table 6.7: Estimates for the logit equation for belonging to the class of compliers

Variable	Estimate	SE	Estimate/SE	p-value
Constant	-4.46	6.98	-0.64	0.52
Gender	0.16	0.63	0.26	0.80
Age	0.05	0.09	0.51	0.61

Measurement model for latent outcome variable Table 6.8 contains the unstandardized factor loadings of the eight indicator variables for measuring the latent outcome variable η. In order to obtain an identified measurement model, the factor loading of the first indicator (SCWT-A1) was set equal to 1. These parameters describe the measurement model for η that applies for both compliers and non-compliers.

Table 6.8: Unstandardized estimates of the measurement model for the latent outcome variable

Indicator	Estimate	SE	Estimate/SE	p-value
SCWT-A1	1.00	-	-	-
SCWT-A2	1.27	0.18	7.14	0.00
SCWT-A3	4.73	0.89	5.32	0.00
AVLT15	-1.25	0.33	-3.79	0.00
AVLT16	-0.29	0.12	-2.49	0.01
MMSE	-0.19	0.06	-3.04	0.00
DST	-2.18	0.33	-6.65	0.00
LFT	-1.81	0.39	-4.63	0.00

All indicators have significant factor loadings on the latent variable η.

Predicting the latent outcome variable Table 6.9 gives the unstandardized estimates of the regression equations for η for compliers and non-compliers. The effects of both experimental manipulations and their interaction are only defined in the class of compliers. In the class of partial compliers only the effect of the first factor is defined. Note also that the covariates Gender (X_1) and Age (X_2) may have both a different effect in the two classes. For the class of compliers (C=1) there is no main effect for the vitamin B supplementation

($b = 0.81, p > 0.05$), nor for aerobic exercise ($b = 0.93, p > 0.05$). Moreover, there is also no interaction between these two factors ($b = 0.56, p > 0.05$). For the class of compliers Gender does have a slight effect on the cognitive functioning of the elderly patients, with men having a better cognitive functioning ($b = 1.94, p < 0.10$). In the class of partial compliers Age influences the cognitive functioning, with older participants having a better cognitive functioning ($b = 0.03, p < 0.05$).

The researchers from the EMGO institute for Health Care Research (van Uffelen et al., 2008) have analyzed the data in an intention to treat analysis for each outcome variable separately. As outcome measures they used the results from eight neuropsychological tests, namely the SCWT-A, SCWT-B, SCWT-C, DSST, VFT, AVLT1-5 and the AVLT6. They concluded that neither the walking program nor vitamin supplementation improved cognition in the community-dwelling older adults with mild cognitive impairment. They suggested that the lack of a main effect of exercise may have been caused by the moderate adherence to the exercise programs. Since they analyzed the data according to the intention-to-treat principle, all randomized participants with available data were included in the analysis, irrespective of exercise adherence. Even data from participants which did not attend a single exercise session were included, thereby underestimating the actual intervention effect. In the re-analysis of the data still no improved cognition was found for either the walking program or the vitamin supplementation. Even treating all outcome variables as indicator for cognitive functioning was of no help in yielding significant treatment effects. It seems that it is not the lack of treatment adherence that explains the negative results, but that alternative explanations have to be found. Maybe the relatively high baseline physical activity level, or the small contrast between the programs may be responsible for the absence of treatment effects.

6.7 Discussion

A randomized control trial (RCT) is certainly the optimal approach for testing the effectiveness of treatment interventions. In research practice, however, planning an RCT is not a guarantee that at the end of the study the intended causal inferences can be made. Too often, participants in a study think for themselves and may all have their own personal reasons to act in a different

Table 6.9: Standardized estimates of the outcome model

	Compliers				Never-takers			
Variable	Estimate	SE	Estimate/SE	p-value	Estimate	SE	Estimate/SE	p-value
R1	0.81	2.86	0.28	0.78	-	-	-	-
R2	0.93	3.33	0.28	0.78	-	-	-	-
R1 x R2	0.56	3.07	0.18	0.86	-	-	-	-
Gender	1.94	1.08	1.94	0.07	-1.07	0.93	-1.16	0.25
Age	0.03	0.24	0.13	0.89	0.32	0.14	2.27	0.02

way than the researcher had in mind. Some patients expect the other treatment which they did not receive to have a more positive impact and try to obtain it. Other participants may experience negative side-effects of the treatment and only partially comply to their treatment regimen. With non-compliance of this kind, the results of a study cannot be simply causally interpreted anymore. Researchers often try to solve the non-compliance problems by carrying out both an intention-to-treat and an as-treated-analysis, and argue afterwards that the true treatment effect lies somewhere in the middle of the two estimates obtained in this way. This strategy is flawed, however.

In this chapter, the instrumental variable approach has been discussed from a latent class point of view. By classifying participants into compliers, always-takers, defiers and never-takers, and by making additional assumptions about the data generation process, true treatment effects can be estimated for the class of compliers. It was shown that, with only a little effort, this way of latent categorization of participants, can easily be modeled in advanced statistical packages. Suggestions are made about how some extensions of the basic model can be dealt with by incorporating so called training variables in M*plus*. In the real world example it is shown that the former conclusion of the researchers of no treatment effect still holds, even after correction for non-compliance in the latent class model. Hopefully this chapter will help researchers to deal with non-compliance in a different way in their future research, and give them a guideline for dealing with non-compliance in a more sophisticated way in their future analysis. The chapter has shown that the basic model can be extended in a flexible way, so that non-compliance with multiple treatments and in factorial designs can be dealt with. It proved to be very difficult to convince researchers to make their data available to use them as an illustrative example. Too often, researchers were suspicious and afraid that a re-analysis of the data would contradict their own conclusions. Therefore plead is made in favor of a policy where data of previously published studies are made available for secondary analysis.

All the extensions of the basic latent class model considered in this chapter assumed that the effective treatment D was perfectly known. As already suggested by the real data application, this may not always be the case. In the real data example patients were seen as complying to their assigned exercise treatment if they behave according to their treatment protocol in 75 % of the

cases. But why 75 %, and not 60 % or 90%? Sharpening or weakening the cut off point for compliance may alter the results of the analysis. In many other applications, it may not even be possible to deduce compliance in this way. Instead, researchers have to be content with some imperfect and unreliable indicators of complying behavior. For instance, patients may be asked to fill in a questionnaire for measuring compliance behavior. An example of such a questionnaire is the Compliance Questionnaire for drug taking behavior of patients with rheumatic diseases (de Klerk, van der Heijde, & Landerwé, 2003). On the basis of such questionnaire data it will, in general, not be possible to define D unambiguously, so that the specific approach described in this chapter is no longer applicable. However, a more drastic modification of the basic latent class model can be conceived. In this approach, two different latent classes are postulated: a first one containing the subjects who can be considered as complying to their treatment, and a second one which contains the never-takers. In the class of compliers, the expected value of the outcome variable Y depends on the assigned treatment R and the covariates; in the class of never-takers, it only depends on the covariates. Since D is not directly observed, no training variables T_1 or T_2 can be defined, but using the questionnaire data it might possible to obtain some compliance indicators whose distributions differ among the classes. Those indicator variables then provide additional information for separating the latent classes. A model of this kind is another instance of a mixture regression model, but without any partial observation of latent class membership. The only information one now has about latent class membership resides in the distributions of the indicator variables.

Although in this chapter only the case of continuous outcome variables is discussed, it is made clear that the case of categorical indicators can also be handled by means of the threshold model proposed by Muthén (1984). This approach, implemented in M*plus*, is however quite restrictive since it assumes that the categorical variables arise as a consequence of categorizing a set of multinormally distributed continuous variables. If these assumptions are not met in a particular data set, alternative analysis based on a loglinear formulation of the models described here could be considered. Such mixture latent class models can be fitted by software as ℓ EM (Vermunt, 1997) or by Latent Gold (Vermunt & Magidson, 2008).

Chapter 7

Adjusting for non-verification in screening studies with repeat testing

7.1 Summary

In medical screening, subjects are often pre-screened by one or multiple non-invasive diagnostic tests and only subjects with at least one positive test are verified for disease status. This strategy may lead to verification bias in estimating the performances of the non-invasive tests. Several methods have been developed to adjust for verification bias in cross-sectional studies. A repeat testing setting is considered where some subjects are directly verified and some are invited for non-invasive retesting at a later time point depending on the baseline test results. A path model is presented which accounts for non-verification and dependencies among the non-invasive tests. For parameter estimation, an expectation maximization (EM) algorithm is presented. The model is applied to data collected in a large cervical cancer screening trial in the Netherlands. A main goal of this trial was to compare the accuracy of cytological testing to human papillomavirus (HPV) DNA testing. It is illustrated how the cross-sectional and longitudinal dependencies of the two tests can be modeled and non-verification can be studied by fitting missing at random (MAR) and not missing at random (NMAR) models.

7.2 Introduction

An important area in medical research is the evaluation of diagnostic tests. To evaluate the performance of a diagnostic test, the test results are compared to the measurements of a gold standard test and the false negative and false positive rates are computed. Here, the false negative rateis the probability of a negative test result in a diseased subject and the false positive rate is the probability of a positive result in a healthy subject. Unfortunately, in many studies, the gold standard test, which provides a definite disease verification, is available only for a subset of the subjects because verification may be burdensome. In a naïve approach where the analysis is based only on the subjects with a gold standard verification, the results are biased because subjects with a positive test result are more often verified than subjects with a negative test result. This type of bias is known in the literature as verification bias (Begg & Greenes, 1983).

For cross-sectional data, several methods for countering verification bias are available. In the setting of only one diagnostic test, Begg and Greenes (1983) propose a stratified estimator where the strata are defined on the basis of the results of the diagnostic test. For each test result, a conditional disease probability is estimated and the false negative and false positive rates are estimated by combining these disease probabilities. The method gives unbiased estimates of the false positive and false negative rates if the probability of verification varies only with the test results. Kosinski and Barnhart (2003) choose a different approach and formulate a model where the test results are defined conditional on the (partially unobserved) disease status. The false positive and negative rates are regression parameters in a logistic regression model that may contain both categorical and continuous covariates. Kosinski and Barnhart (2003) present an expectation maximization (EM) algorithm for estimating the model parameters. Baker (1995), Zhou (1998) and Alonzo (2005) consider the problem of verification bias in the assessment of two diagnostic tests. As in Begg and Greenes (1983), Baker (1995), Zhou (1998) and Alonzo (2005) formulate the disease probability. Baker (1995), Kosinski and Barnhart (2003) and Zhou and Castelluccio (2004) allow the verification probability to depend on test results and on the disease status.

In the present study, verification bias in a repeat testing setting is consid-

ered. This is a common situation in medical screening and occurs when some of the subjects who are not verified immediately, are retested at a later time point. The motivating data have been collected in a cervical cancer screening trial in the Netherlands (Bulkmans et al., 2007). In this trial, two different screening tests were compared. The data have a repeated testing structure because women without severe abnormalities on the baseline tests were not immediately verified but were retested after 6 and 18 months.

For modeling disease verification, different missingness mechanisms are assumed including not missing at random (NMAR) mechanisms (Rubin, 1976). In the NMAR models, the probability of disease verification depends not only on observed variables but also on the partially unobserved disease status. Regarding the relation between test results and disease, the same approach as Kosinski and Barnhart (2003) will be followed and the test results conditional on the disease status are modeled. The repeat testing data structure will be explicitly incorporated in the model by formulating a path model. The association among the variables in the model will be described by a series of logistic regression equations (Kosinski & Barnhart, 2003; Goodman, 1978; Baker & Laird, 1988).

The remainder of the chapter is organized as follows. In section 7.3, the data will be described and the verification problem will be explained. In section 7.4, path models for verification bias will be presented. In section 7.5, the data will be analyzed and the results will be presented. The final section 7.6 contains a discussion.

7.3 Cervical cancer screening study

7.3.1 Data description

In a Dutch screening trial (Bulkmans et al., 2007), 44,102 women aged 30-60 years were screened by a cytologic inspection of the cervix or by a combination of cytology and a molecular test. The molecular test checks for DNA of human papillomavirus (HPV) which is the causal agent of cervical cancer. Because a direct comparison is made between cytology and the HPV DNA test, only the data from the experimental group (N=21,950) is used. The screening management protocol is presented by the flow chart in figure 7.1. It can be read from the figure that participants can be immediately referred to the gynecologist,

dismissed from further follow-up or invited for retesting. The first and second retest are targeted at 6 and 18 months after baseline. Positive and negative HPV DNA test results are denoted by HPV+ and HPV-. Normal, mild, and severe cytologic abnormalities are denoted by cyt-, cyt+, and cyt++, respectively. If a woman is referred to the gynecologist, a cervical specimen will be taken from cellular abnormal tissue which will be analyzed by the pathologist for disease verification (gold standard test). A woman is considered as diseased if she presents with a cervical lesion grade 3 or worse (CIN3+).

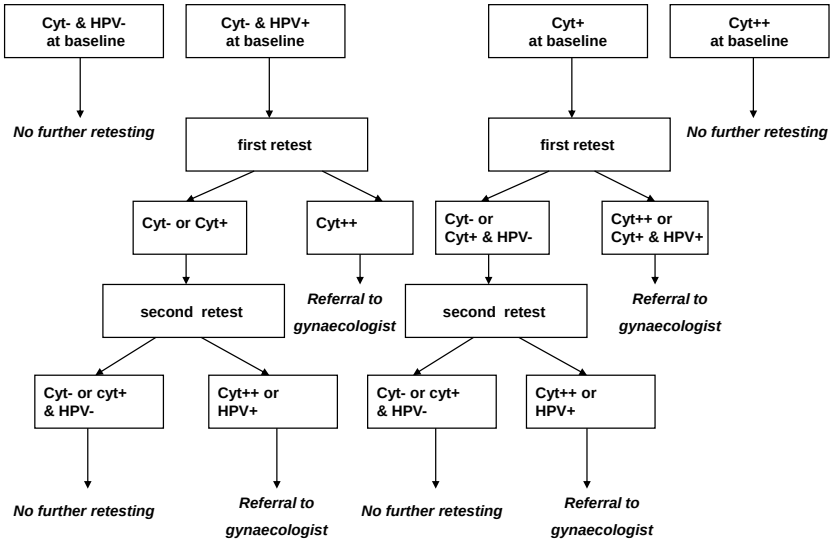

Figure 7.1: Management flow chart of the trial in the experimental group

7.3.2 Verification

Verification is incomplete for three reasons. First, according to the screening protocol, some women will not be verified depending on the test results at baseline and/or after repeat testing. Second, some women do not show up at retesting at 6 and 18 months. Third, not every woman who is referred to the gynecologist actually receives the gold standard test. The latter type of non-verification may occur for two reasons: a woman does not show up at the

gynecologist or the gynecologist decides not to take a sample for verification. The three different types of non-verification are illustrated in figure 7.2.

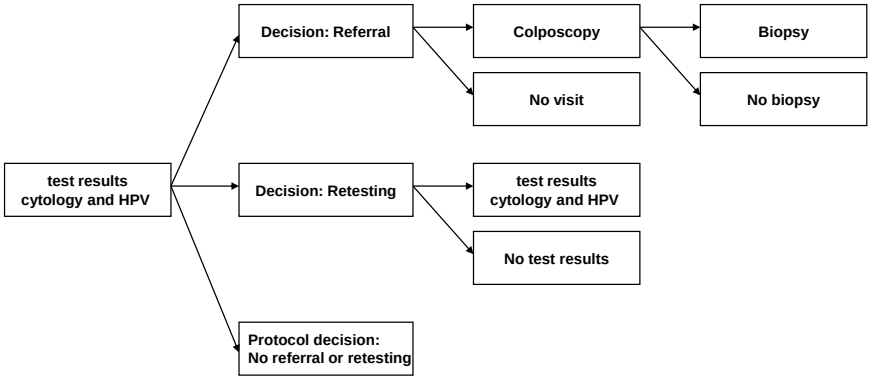

Figure 7.2: Graphical representation of the reasons for incomplete verification

7.4 Path models

7.4.1 Model for cross-sectional data

Incomplete verification can be seen as a missing data problem. For variables that are partially observed, Fay (1986) and Baker and Laird (1988) propose a log-linear path model, consisting of a series of logit models, where missingness is specified by an indicator variable (Goodman, 1978).

First, the path model for cross-sectional data is explained within the context of the cervical screening example. The outcomes of the cytology and HPV test are denoted by categorical variables A (scores 0, 1, and 2 for cyt-, cyt+, and cyt++, respectively) and B (scores 0 and 1 for HPV- and HPV+, respectively). The true disease status is denoted by Y (0 if healthy and 1 if diseased). Also the verification variable R is introduced (1 if verified and 0 if not verified). The

model parameters are in vector $\boldsymbol{\theta}$. The joint probability of observing $Y = y$, $A = a$, $B = b$, and $R = r$, i.e. $p(y, a, b, r|\boldsymbol{\theta})$ is modeled as the product of $p(y, a, b|\boldsymbol{\theta})$ and the conditional probability of verification $p(r|y, a, b; \boldsymbol{\theta})$:

$$p(y, a, b, r|\boldsymbol{\theta}) = p(y, a, b|\boldsymbol{\theta})\, p(r|y, a, b; \boldsymbol{\theta}) \ . \tag{7.1}$$

The i-th subject is identified by index i. The log-likelihood of the data is obtained by adding the subject-specific log-probabilities:

$$\begin{aligned} \log\, L(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \log p(y_i, a_i, b_i, r_i|\boldsymbol{\theta}) \\ &= \sum_{i=1}^{N} \{R_i \log p(y_i, a_i, b_i, R_i = 1|\boldsymbol{\theta}) + \\ &= (1 - R_i) \log p(a_i, b_i, R_i = 0|\boldsymbol{\theta}) \ \} \end{aligned} \tag{7.2}$$

.

The right-hand side of equation 7.2 can be further evaluated. Following equation 7.1, the probability of observing $Y_i = y_i$, $A_i = a_i$, $B_i = b_i$ can be written as

$$p(y_i, a_i, b_i, R_i = 1|\boldsymbol{\theta}) = p(y_i, a_i, b_i|\boldsymbol{\theta})\, p(R_i = 1|y_i, a_i, b_i; \boldsymbol{\theta}) \ ,$$

and the probability of observing only test results $A_i = a_i$ and $B_i = b_i$ can be written as

$$p(a_i, b_i, R_i = 0|\boldsymbol{\theta}) = \sum_{j=0}^{1} p(Y_i = j, a_i, b_i|\boldsymbol{\theta})\, p(R_i = 0|Y_i = j, a_i, b_i; \boldsymbol{\theta}) \ .$$

The probability $p(y, a, b|\boldsymbol{\theta})$ can be modeled in two ways. Baker (1995) and Zhou (1998) use

$$p(y, a, b|\boldsymbol{\theta}) = p(a, b|\boldsymbol{\theta})\, p(y|a, b; \boldsymbol{\theta}), \tag{7.3}$$

where $p(y|a, b; \boldsymbol{\theta})$ is the probability of disease outcome y given test results a and b. Kosinski and Barnhart (2003) use

$$p(y, a, b|\boldsymbol{\theta}) = p(y|\boldsymbol{\theta})\, p(a, b|y; \boldsymbol{\theta}) \ . \tag{7.4}$$

Kosinski and Barnhart (2003)'s approach is the most useful one for this situation with two possibly dependent tests because the relation between the

test results can be modeled in a sensible way after conditioning on the disease status (Sasieni, 2001). As put forward by Kosinski and Barnhart (2003) the components from equation 7.4 can be modeled by a series of logistic equations. Because cytology scores are trichotomous, a variant of Kosinski and Barnhart's model for nominal responses needs to be set up by pairing response categories to a baseline category (Agresti, 2002). If it is assumed that the outcome B of the HPV test depends on the outcome A of the cytology test, and also assumed that verification depends on both cytology and the disease status (NMAR), then a model might be set up with the following 3 components.
Disease component:

$$\text{logit } p(Y_i = 1|\phi) = \phi \ . \tag{7.5}$$

Diagnostic test components:

$$\begin{aligned} \log \frac{p(A_i = m|y_i;\boldsymbol{\alpha})}{p(A_i = 0|y_i;\boldsymbol{\alpha})} &= \alpha_{0m} + \alpha_{1m} y_i \ , \qquad m = 1,\ 2, \\ \text{logit } p(B_i = 1|y_i;\boldsymbol{\beta}) &= \beta_0 \ + \ \beta_1 y_i \ + \ \beta_2 a_{i0} \ + \ \beta_3 a_{i0} \times y_i \ . \end{aligned}$$

Verification component:

$$\text{logit } p(R_i = 1|y_i, a_i, b_i, \boldsymbol{\gamma}) = \gamma_0 + \gamma_1 a_{i1} \ + \ \gamma_2 a_{i2} \ + \ \gamma_3 b_i \ + \ \gamma_4 y_i \ , \tag{7.6}$$

where a_{i0}, a_{i1}, and a_{i2} are dummies with score 1 for cytologic results cyt-, cyt+, and cyt++, respectively. Model parameter vector $\boldsymbol{\theta}$ consists of ϕ, $\boldsymbol{\alpha}$ $= (\alpha_{01}, \alpha_{11}, \alpha_{02}, \alpha_{12})'$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_3)'$, and $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_4)'$. Independence between cytology and the HPV test can be imposed by setting β_2 and β_3 equal to zero. Setting γ_4 equal to zero leads to a Missing at Random (MAR) model, and setting $\gamma_1, \ldots, \gamma_4$ equal to zero leads to a Missing Completely at Random (MCAR) model.

7.4.2 Model for repeat testing data

To set up the model for the cervical screening trial with repeated testing, time index variable $t = 0, 1, 2$ are introduced corresponding to respectively the baseline, the first repeat testing moment, and the second repeat testing moment. The test and the verification variable are extended with the time index and have cytology variable $\boldsymbol{A}_i = (A_{i(0)}, \ldots, A_{i(t_i)})'$ for subject i, where t_i is the time index of the last test. Also the HPV test variable $\boldsymbol{B}_i = (B_{i(0)}, \ldots, B_{i(t_i)})'$ and verification variable $\boldsymbol{R}_i = (R_{i(0)}, \ldots, R_{i(t_i)})'$ will be defined. Furthermore

variable $\boldsymbol{S}_i = (\boldsymbol{S}_{i(0)}, \ldots, \boldsymbol{S}_{i(t_i)})'$ will be defined, the t-th element of which indicates whether, according to the screening protocol in figure 7.1, woman i is referred to the gynecologist at time index t (score 1 if decision is referral and 0 if decision is no referral). Note that $\boldsymbol{S}_i$ is determined by the variables $\boldsymbol{A}_i$ and $\boldsymbol{B}_i$.

If logistic regression link functions are imposed, a model for subject i with the following 3 components may be defined as:

Disease Component:

$$\text{logit } P(Y_i = 1|\boldsymbol{\phi}) = \boldsymbol{v}_i\,\boldsymbol{\phi}\ . \tag{7.7}$$

Diagnostic test components:

$t = 0, \ldots, t_i$:

$$\log \frac{p(A_{i(t)} = m|\boldsymbol{z}_{i(t)}; \boldsymbol{\alpha}_{(t)})}{p(A_{i(t)} = 0|\boldsymbol{z}_{i(t)}; \boldsymbol{\alpha}_{(t)})} = \boldsymbol{z}_{i(t)}\boldsymbol{\alpha}_{m(t)}, \qquad m = 1,\ 2, \tag{7.8}$$

$$\text{logit}\, p(B_{i(t)} = 1|\boldsymbol{x}_{i(t)}; \boldsymbol{\beta}_{(t)}) = \boldsymbol{x}_{i(t)}\boldsymbol{\beta}_{(t)}\ . \tag{7.9}$$

Verification component:

$t = 0, \ldots, t_i$:

$$\begin{aligned} p(R_{i(t)} = 1|S_{i(t)} = 0) &= 0\ , \\ \text{logit}\, p(R_{i(t)} = 1|\boldsymbol{w}_{i(t)}, S_{i(t)} = 1; \boldsymbol{\gamma}_{(t)}) &= \boldsymbol{w}_{i(t)}\boldsymbol{\gamma}_{(t)}\ . \end{aligned} \tag{7.10}$$

The predictor matrices $\boldsymbol{v}_i$, $\boldsymbol{z}_{i(t)}$, $\boldsymbol{x}_{(i)t}$, and $\boldsymbol{w}_{i(t)}$ may consist of manifest continuous and categorical predictors but also of the partially observed disease status y_i. The regression coefficients $\boldsymbol{\alpha}_{m(t)}$ $(m = 1, 2)$, $\boldsymbol{\beta}_{(t)}$, and $\boldsymbol{\gamma}_{(t)}$ may vary across time. In the verification component, the verification probability can be positive only for women that are referred to the gynecologist (i.e. $S_{i(t)} = 1$).

The model parameters can be estimated with the EM algorithm (Dempster, Laird, & Rubin, 1977). In the E-step, the expectation of the logarithm of

the complete data likelihood at the current estimate of the model parameters $\boldsymbol{\theta}^* = \{\ \boldsymbol{\phi}^*, \boldsymbol{\alpha}^*_{(t)}, \boldsymbol{\beta}^*_{(t)}, \boldsymbol{\gamma}^*_{(t)}, t = 1, 0, 2\}$:

$$\log L_c(\boldsymbol{\theta}^*) = \sum_{i=1}^{N} \log p(y_i, \boldsymbol{a}_i, \boldsymbol{b}_i, r_i | \boldsymbol{\theta}^*)\ ,$$

is taken with respect to the unobserved disease scores. Let us denote y_i and $(1-y_i)$ by y_{i1} and y_{i0}, respectively. The complete data likelihood can be written as

$$\log L_c(\boldsymbol{\theta}^*) = \sum_{i=1}^{N} \left\{ r_i \log p(y_i, \boldsymbol{a}_i, \boldsymbol{b}_i, r_i | \boldsymbol{\theta}^*) + (1 - r_i) \sum_{j=0}^{1} y_{ij} \log p(Y_i = j | \boldsymbol{a}_i, \boldsymbol{b}_i, r_i; \boldsymbol{\theta}^*) \right\}\ .$$

The E-step involves calculating the conditional expectation of the disease status y_{ij}, yielding $\mathrm{E}\left[y_{ij} | \boldsymbol{a}_i, \boldsymbol{b}_i, r_i; \boldsymbol{\theta}^*\right] = \tilde{p}_{ij} = p(Y_i = j | \boldsymbol{a}_i, \boldsymbol{b}_i, r_i; \boldsymbol{\theta}^*)$. The conditional expectation $\tilde{p}_{ij}$ can be computed by writing $\tilde{p}_{ij}$ as

$$\tilde{p}_{ij} = \frac{p(\boldsymbol{a}_i, \boldsymbol{b}_i, r_i | Y_i = j; \boldsymbol{\theta}^*)\, p(Y_i = j | \boldsymbol{\theta}^*)}{\sum_{j=0}^{1} p(\boldsymbol{a}_i, \boldsymbol{b}_i, r_i | Y_i = j; \boldsymbol{\theta}^*)\, p(Y_i = j | \boldsymbol{\theta}^*)}\ ,$$

and substituting the current estimates retrieved from the disease, diagnostic, and verification component. In the M-step, the expected complete data likelihood is maximized with respect to $\boldsymbol{\theta}$, that is, maximization on:

$$Q(\boldsymbol{\theta}) = \sum_{i=1}^{N} \left\{ r_i \log p(y_i, \boldsymbol{a}_i, \boldsymbol{b}_i, r_i | \boldsymbol{\theta}) + (1 - r_i) \sum_{j=0}^{1} \tilde{p}_{ij} \log p(Y_i = j, \boldsymbol{a}_i, \boldsymbol{b}_i, r_i | \boldsymbol{\theta}) \right\}\ .$$

Estimated standard errors for the model parameters can be obtained by inverting the observed information matrix which is minus the matrix of second-order derivatives of the log-likelihood function with regard to the model parameters.

The EM algorithm can be run in standard packages for latent categorical data such as M*plus* (Muthén & Muthén, 2008) or Lem (Vermunt, 1997). However, estimation can become difficult as there is a limitation to the number of repeated testing moments and the number of restrictions imposed on the model parameters. The inclusion of both continuous and categorical predictors in the logistic regression equations may also lead to computational difficulties. Fast computation of the M-step is possible by defining pseudo-data. The idea of constructing a pseudo data set was put forward by Lambert (1992) and was

suggested for cross-sectional data by Kosinski and Barnhart (2003) for estimating a model with partial verification. Suppose that only the first $N-U$ subjects are verified. Then, $Q(\boldsymbol{\theta})$ can be written as

$$\begin{aligned} Q(\boldsymbol{\theta}) \quad = \quad & \sum_{i=1}^{N-U} \log p(y_i, \boldsymbol{a}_i, \boldsymbol{b}_i, r_i|\boldsymbol{\theta}) + \\ & \sum_{i=N-U+1}^{N} \tilde{p}_{i1} \log p(Y_i = 1, \boldsymbol{a}_i, \boldsymbol{b}_i, r_i|\boldsymbol{\theta}) + \\ & \sum_{i=N-U+1}^{N} \tilde{p}_{i0} \log p(Y_i = 0, \boldsymbol{a}_i, \boldsymbol{b}_i, r_i|\boldsymbol{\theta}) \ . \end{aligned} \tag{7.11}$$

It is easy to see that the function $Q(\boldsymbol{\theta})$ is equal to a function that would be obtained by weighted summing of the log-likelihoods of pseudo-data of $N+U$ subjects. The pseudo data of the $N+U$ subjects in the weighted log-likelihood function are as follows. The first $N-U$ subjects correspond to the verified subjects from the original data ($Y_i = y_i$, $R_i = 1$) and have weight 1. The next U subjects are the non-verified subjects of the original data and they are labeled as diseased ($Y_i = 1$, $R_i = 0$) and receive weight $\tilde{p}_{i1}$. The last U subjects are again the non-verified subjects of the original data but now they are labeled as healthy ($Y_i = 0$, $R_i = 0$) and receive weight $\tilde{p}_{i0}$. Thus, the non-verified subjects of the original data appear twice in the pseudo-data, once as diseased and once as healthy subjects. The three subgroups correspond to the three terms at the right-hand side of equation 7.11. Although the connection between the log-likelihood of the observed data and the weighted log-likelihood of the pseudo-data of $N+U$ subjects is purely technical, the connection is useful computationally as it enables us to fit the data using a weighted generalized linear regression module. Such a module is incorporated in many statistical packages. Moreover, because the disease, testing and verification components do not share model parameters, separate regressions can be fit to the different model components. R-code of the EM algorithm is available from the authors upon request.

7.5 Example: results

Five nested models are fitted to the cervical screening trial data. In model 1 (the simplest model), the regression equation for disease status (equation 7.7)

only has an intercept. The disease status y_i is included as a covariate in the regression equations for cytology (equation 7.8) and the HPV test (equation 7.9). Besides, time-invariant regression coefficients of cytology are included in the regression equation for the verification variable (equation 7.10). Because cytology has three levels, the verification model component contains two cytology dummies. All regression coefficients are restricted to be time-invariant.

In models 2 to 4, dependencies among repeated cytology and HPV test outcomes are modeled. Only model dependencies in healthy women are modeled with the assumption that the outcomes of the cytology and HPV test are independent in diseased women. The independence between the test outcomes in healthy women is likely to be violated for at least two reasons. First, a woman may carry a transient HPV infection that will eventually disappear without causing cervical disease. Second, a woman may present with abnormal cytology because of a non-HPV related lesion that will eventually regress to normal. Both types of dependencies will lead to a positive association between consecutive test results. The dependencies between the test results cannot be explained by the disease status and can only be captured by explicitly modeling the dependency.

In model 2, the dependency among repeated HPV test results is modeled by including the previous HPV test result in the regression equation for the HPV test (equation 7.9). Two lagged covariates $b_{i(t-1)}(1-y_i)$ and $(1-b_{i(t-1)})(1-y_i)$ are included. The value of $b_{i(t-1)}$ at $t=0$ is set equal to 0. In model 3, the dependency between repeated cytology scores is modeled by including the lagged covariates $b_{i(t-1)}(1-y_i)$ and $(1-b_{i(t-1)})(1-y_i)$ in the regression equations (equation 7.8) for the multinomial cytology scores. Also the following two lagged cytology covariates are included in equation 7.8: $a_{i0(t-1)}(1-y_i)$ and $(1-a_{i0(t-1)})(1-y_i)$ where $a_{i0(t-1)}$ is 1 if cytology is normal at the previous time point. At $t=0$, $a_{i0(t-1)}$ is set equal to 0. In model 4, the cross-sectional dependency between HPV and cytology test result is modeled by including two dummy variables for the trichomotous cytology score $a_{i0(t)}(1-y_i)$ in the regression equation at time t (equation 7.9). The regression coefficients of the current and lagged variables are restricted to be time-invariant. Model 5 is an extension of model 4 where the verification regression equation (equation 7.10) is extended with covariate disease status y_i. Therefore, models 1-4 are MAR and model 5 is NMAR.

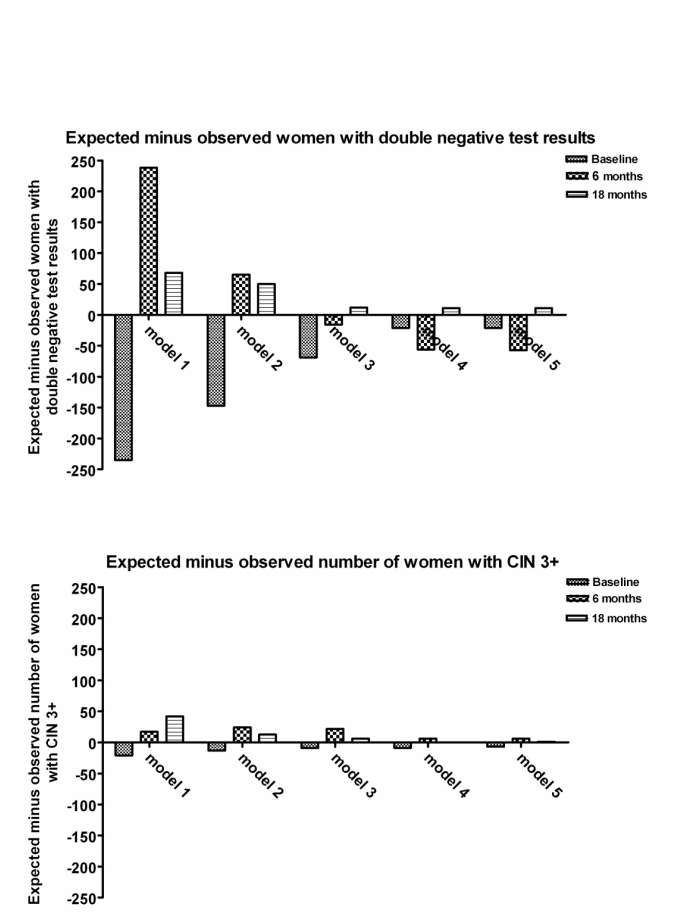

Figure 7.3: Expected minus observed frequencies of women with double negative test results and women with CIN3+

The parameters of key interest are the disease prevalence (i.e. the probability that the woman has CIN3+) and the sensitivities and specificities of cytology and the HPV test. The sensitivity and specificity of a test are widely used in the medical field and are 1 minus the false negative rate and 1 minus the false positive rate, respectively. For cytology and the HPV test, the sensitivities are $p(a > 0|y = 1; \boldsymbol{\theta})$ and $p(b = 1|y = 1; \boldsymbol{\theta})$, respectively, and the specificities are $p(a = 0|y = 0; \boldsymbol{\theta})$ and $p(b = 0|y = 0; \boldsymbol{\theta})$.

The disease prevalence is presented in figure 7.4 together with the 95% confidence intervals. The CIN3+ prevalences obtained under models 1 to 5 are presented in figure 7.5. It is seen that the prevalences are much higher in models 1-3 than in models 4 and 5. The CIN3+ prevalences in models 4 and 5 are similar. This indicates that the decision about the dependency structure has a stronger effect on the CIN3+ prevalence than the decision about the verification mechanism. The test sensitivities are presented in figure 7.5. Both for cytology and the HPV test, the sensitivities are the lowest for models 1 and 2. Again, the estimates are sensitive to decision about the dependency structure but are not sensitive to the decision about the verification mechanism. Finally, the estimated (marginal) specificity of cytology ranges from 97.4% to 98.0% in models 1 to 5 and the (marginal) specificity of the HPV test ranges from 95.7% to 97.1%.

If the models are compared by likelihood ratio testing, it follows that the models differ significantly in fit. The smallest improvement ($\chi^2(1) = 7.86$, $p = .005$) is obtained when comparing model 5 (NMAR verification) to model 4 (MAR verification). To check whether the models are consistent with the data, the predicted number of women with CIN3+ as well as the number of women with double negative test results are computed (negative on cytology and the HPV test) at 0, 6, and 18 months (figure 7.3). It is seen that models 4 and 5 predict the number of women with CIN3+ well, but the other models overestimate the number of women with CIN3+ at 6 and 18 months. The number of double negative women is reasonably well predicted by models 3 to 5. Models 1 and 2 give a poor prediction of the number of double negative women observed at baseline.

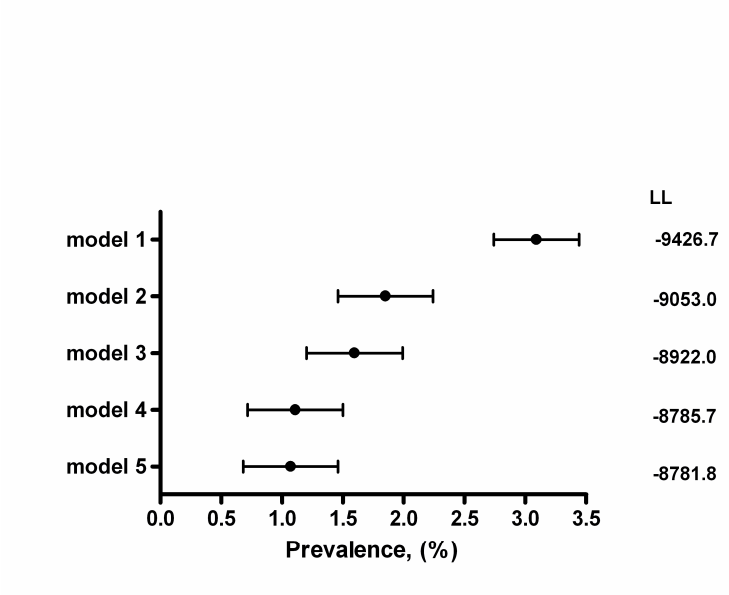

Figure 7.4: Estimated CIN3+ prevalences and log-likelihood values

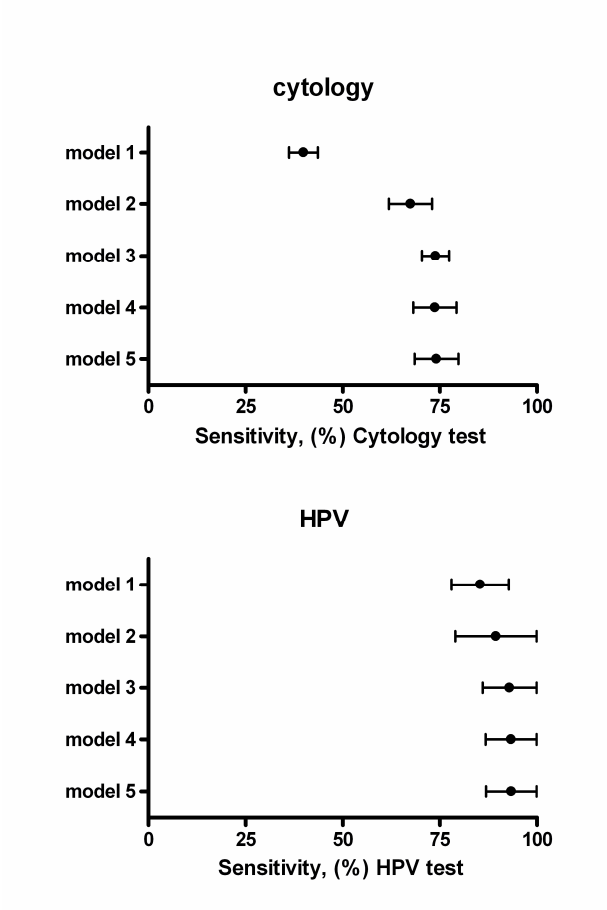

Figure 7.5: Estimated test sensitivities

7.6 Discussion

The development of the model has been motivated by the field of medical screening, where there is an increasing awareness of the patient burden and costs induced by invasive verification techniques. Consequently, diagnostic guidelines are more and more often developed where patients that have positive test results are not immediately verified, but are retested at a later time. It is shown how to analyze the outcomes of studies in which such a repeat testing strategy is adopted. The focus of this chapter was on the setting of two screening tests, but the model can also be set up for a different number of screening tests. The model has large flexibility as the disease status, the test results, and the verification status may depend on both continuous and categorical covariates. The estimation of the model parameters is likelihood-based which enables us to draw inferences about the disease prevalence and the test sensitivity and specificity.

In this model, the test outcome is defined conditional on the disease status. This approach is useful because it enables us to define dependencies between the test results separately for diseased and healthy subjects. In the cervical screening example, it was assumed that the cytology and HPV test results were independent in diseased subjects. This assumption was necessary to obtain an identifiable model as subjects with a double negative test result (negative on cytology and the HPV test) at baseline or at 18 months were not verified. The independence assumption seems reasonable as a working assumption although it is unlikely to hold exactly. Therefore, the interpretation of the estimated parameters should be done with care recognizing the underlying independence assumption.

The cervical screening example showed that it is important to accurately model the dependencies among the screening tests, also when the test results are formulated conditional on the disease status. Two types of dependencies were considered: a longitudinal dependency between the outcomes at consecutive testing moments and a cross-sectional dependency between the cytology and HPV test. Omitting either type of dependency had a large effect on the estimates of the test sensitivity and the disease prevalence. The model predictions were consistent with observed test results and CIN3+ prevalences. If the number of repeated measurements had been larger than three, it would

have been easier to violate the observed data. For such data, it may be worthwhile to include higher-order lagged effects in the diagnostic testing regression equations (Diggle, 1994).

In this model loss-to-follow up was not modeled. A loss-to-follow-up component is ignorable when loss-to-follow-up is independent of the disease status and the loss-to-follow up component does not share parameters with the other components (Rubin, 1976). In the example, there was no necessity to model loss-to-follow-up because cervical screening takes place in an asymptomatic population and loss-to-follow up is only related to the screening test results. However, in a setting where the attendance at follow-up tests depends on the manifestation of clinical symptoms, then it makes sense to formulate a loss-to-follow-up component in addition to the disease, testing, and verification component. This component could for example be defined by taking a selection model approach where follow-up is predicted by the test results and the disease status (Little, 1993).

References

Agresti, A. (2002). *Categorical data analysis [2nd edition].* New York: Wiley Interscience.

Alden, L. (1989). Short-term structured treatment for avoidant personality disorder. *Journal of Consulting and Clinical Psychology*, *57*, 756-764.

Alonzo, T. A. (2005). Verification bias-corrected estimators of the relative true and false positive rates of two binary screening tests. *Statistics in Medicine*, *24*, 403–417.

Amberson, J. B., McMahon, B. T., & Pinner, M. (1931). A clinical trial of sancrosyn in pulmonary tuberculosis. *American Review of Tuberculosis*, *24*, 401-435.

Anderson, S., Auquier, A., Hauck, W. W., Oakes, D., Vandaele, W., & Weisberg, H. I. (1980). *Statistical methods for comparative studies: Techniques for bias reduction.* New York: John Wiley Sons.

Andrea, H., Verheul, R., Berghout, C. C., Dolan, C., Kroft, P. J. A. V. der, Busschbach, J. J. V., Bateman, A. W., & Fonagy, P. (2007). *Measuring the core components of maladaptive personality: severity indices of personality problems (sipp- 118)* (Tech. Rep.). Viersprong Institute for Studies on Personality Disorders (VISPD) in cooperation with the department of Medical Psychology & Psychotherapy, Erasmus University Rotterdam.

Angrist, J. D., & Imbens, G. W. (2005). Two-stage least squares estimation of average causal effects in model with variable treatment intensity. *Journal of the American Statistical Association*, *90*, 431–442.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 444–472.

Arnevik, E., Wilberg, T., Urnes, O., Johansen, M., Monsen, J. T., & Karterud, S. (2009). Psychotherapy for personality disorders: short-term day hospital psychotherapy versus outpatient individual therapy a randomized controlled study. *European Psychiatry*, *24*, 71-78.

Arrindell, W. A., & Ettema, J. H. M. (2003). *Herziene handleiding bij een multidimensionele psychopathologie-indicator.* Lisse: Swets & Zeitlinger.

Austin, P. C., Grootendorst, P., & Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a monte carlo study. *Statistics in Medicine*, *26*, 734-753.

Baker, S. G. (1995). Evaluating multiple diagnostic tests with partially verification. *Biometrics*, *51*, 330–337.

Baker, S. G., & Laird, N. M. (1988). Regression-analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, *83*, 62–69.

Bartak, A., Spreeuwenberg, M. D., Andrea, H., Busschbach, J. J. V., Croon, M. A., Verheul, R., Emmelkamp, P. M. G., & Stijnen, T. (2009). The use of propensity score methods in psychotherapy research: a practical application. *Pychotherapy and Psychosomatics*, *78*, 26-34.

Bartak, A., Spreeuwenberg, M. D., Andrea, H., Holleman, L., Rijnierse, P., Rossum, B. V. van, Hamers, E. F. M., Aerts, A. M. M. A., Busschbach, J. J. V., Verheul, R., Stijnen, T., & Emmelkamp, P. M. G. (2010). Effectiveness of different modalities of psychotherapeutic treatment for patients with cluster c personality disorder: results of a large prospective multicentre study. *Psychotherapy and Psychosomatics*, *79*, 20-30.

Bateman, A., & Fonagy, P. (2001). Treatment of borderline personality disorder with psychoanalytically oriented partial hospitalization: an 18-month follow-up. *American Journal of Psychiatry*, *158*, 36-42.

Beek, N. van, & Verheul, R. (2008). Motivation for treatment in patients with personality disorders. *Journal of Personality Disorders*, *22*, 89-100.

Begg, C. B., & Greenes, R. A. (1983). Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*, *39*, 207–215.

Benson, K., & Hartz, A. J. (2000). A comparison of observational studies and randomized, controlled trials. *The New England Journal of Medicine*, *342*, 1878-1886.

Beurs, E. de, & Zitman, F. G. (2006). De brief symptom inventory (bsi): de betrouwbaarheid en validiteit van een handzaam alternatief voor de scl-90 (the brief symptom inventory (bsi): the reliability and validity of a brief alternative of the scl-90). *Maandblad GeestelijkeVolksgezondheid*, *61*, 120-141.

Black, N. (1996). Why we need observational studies to evaluate the effectiveness of health care. *British Medical Journal*, *312*, 1215-1218.

Bollen, K. A. (1989). *Structural equations with latent variables.* New-York: Wiley.

Brewin, C. R., & Bradley, C. (1989). Patient preferences and randomized clinical trials. *British Medical Journal*, *299*, 313-315.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Sturmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, *163*, 1149-1156.

Brooks, R., R, R., & Charro, F. de. (2003). *The measurement and valuation of health status using eq-5d: A european perspective. evidence from the euroqol biomed research programme.* Dordrecht: Kluwer Academic Publishers.

Bulkmans, N. W. J., Berkhof, J., van Kemenade, L. R. F. J., Boeke, A., Bulk, S., Voorhorst, F., Verhijen, R., van Groningen, K., & Boon, M. (2007). Human papillomavirus dna testing for the detection of cervical intraepithelial neoplasia grade 3 and cancer: 5-year follow-up of a randomised controlled implementation trial. *Lancet*, *370*, 1764–1772.

Castonguay, L. G., & Beutler, L. E. (2006). *Principles of therapeutic change that work.* New York: Oxford University Press.

Chan, A. W., Bhatt, D. L., Chew, D. P., Quinn, M. J., Moliterno, D. J., Topol, E. J., & Ellis, S. G. (2002). Early and sustained survival benefit associated with statin therapy at the time of percutaneous coronary intervention. *Circulation*, *105*, 691-696.

Chiesa, M., & Fonagy, P. (2007). Prediction of mediumterm outcome in cluster b personality disorder following residential and outpatient psychosocial treatment. *Pychotherapy and Psychosomatics*, *76*, 347-353.

Chiesa, M., Fonagy, P., & Gordon, J. (2009). Community-based psychodynamic treatment program for severe personality disorders: clinical description and naturalistic evaluation. *Journal of Psychiatric Practice*, *15*, 12-24.

Clarkin, J. F., Levy, K. N., Lenzenweger, M. F., & Kernberg, O. F. (2007). Evaluating three treatments for borderline personality disorder: A multiwave study. *American Journal of Psychiatry*, *164*, 922-928.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, *24*, 295-313.

Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational studies: a review. *Sankhya*, *35*, 417–446.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale: Lawrence Erlbaum Associates.

Coid, J., Yang, M., Tyrer, P., Roberts, A., & Ullrich, S. (2006). Prevalence and correlates of personality disorder in great britain. *British Journal of Psychiatry*, *188*, 423-431.

Concato, J., Shah, N., & Horwitz, R. I. (1968). Randomized, controlled trials, observational studies, and the hierarchy of research designs. *The New England Journal of Medicine*, *342*, 1887-1892.

Connors, A. F., Speroff, T., Dawson, N. V., Thomas, C., Harrell, F. E., Wagner, D., Desbiens, N., Goldman, L., Wu, A. W., Califf, R. M., Fulkerson, W. J., Vidaillet, H., Broste, S., Bellamy, P., Lynn, J., & Knaus, W. A.

(1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. *Journal of the American Medical Association*, *276*, 889-897.

Cox, D. R. (1958). *Planning of experiments.* New York: Wiley.

Cox, D. R., & Wermuth, N. (2001). Some statistical aspects of causality. *European Sociological Review*, *17*, 65-74.

DÁgostino, R. B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, *17*, 2265–22817.

Dehejia, R., & Wahba, S. (1999). Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, *94*, 1053-1062.

DeJong, C. A., Brink, W. V. den, Harteveld, F. M., & Wielen, E. V. der. (1993). Personality disorders in alcoholics and drug addicts. *Comprehensive Psychiatry*, *34*, 87-94.

DeJong, C. A. J., Derks, F. C. H., Oel, C. J. V., & Rinne, T. (1986). *Gestructureerd interview voor de dsmiv persoonlijkheidsstoornissen (sidp-iv).* Sint Oedenrode: Stichting VerslavingszorgOost Brabant.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, *39*, 1–38.

Derogatis, L. R. (1977). *Scl-90 (r): Administration, scoring, and procedures manual i for the revised version.* Baltimore: Johns Hopkins University School of Medicine, Clinical Psychometrics Research Unit.

Derogatis, L. R. (1986). *Scl-90-r: Administration, scoring and procedure. manual ii for the revised version.* Townson: Clinical Psychometric Research.

Derogatis, L. R., & Melisaratos, N. (1983). The brief symptom inventory: an introductory report. *Psychological Medicine*, *13*, 595-605.

Diggle, K. Y. (1994). *Analysis of longitudinal data.* New York: Oxford University Press.

Dranove, D., & Lindrooth, R. (2003). Hospital consolidation and costs: another look at the evidence. *Journal of Health Economics*, *22*, 983-997.

Duggan, C., Huband, N., Smailagic, N., Ferriter, M., & Adams, C. (2007). The use of psychological treatments for people with personality disorder: a systematic review of randomized controlled trials. *Personality and Mental Health*, *1*, 95-180.

Emmelkamp, P. M. G., Benner, A., Kuipers, A., Feiertag, G. A., Koster, H. C., & Apeldoorn, F. J. van. (2006). Comparison of brief dynamic and cognitive behavioural therapies in avoidant personality disorder. *British Journal of Psychiatry*, *189*, 60–64.

Emmerik, A. A. P. van, Kamphuis, J. H., & Emmelkamp, P. M. G. (2008). Treating acute stress disorder and posttraumatic stress disorder with cognitive behavioral therapy or structured writing therapy: a randomized controlled trial. *Psychotherapy and Psychosomatics*, *77*, 93–100.

Facchinetti, F., Ottolini, F., Fazzio, M., Rigatelli, M., & Volpe, A. (2007). Psychosocial factors associated with preterm uterine contractions. *Psychotherapy and Psychosomatics*, *76*, 391-394.

Fay, R. E. (1986). Causal models for patterns of nonresponse. *Journal of the American Statistical Association*, *81*, 354–365.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*(4), 521–532.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). Mini-mental state. a preactical method for grading the cognitive state of patients for the clinician. *Journal of Pychiatric Research*, *12*, 189–198.

Forcina, A. (1975). Causal effects in the presence of noncompliance: A latent variable interpretation. *Metron: International Journal of Statistics*, *64*, 275–301.

Forstmeier, S., & Rueddel, H. (2007). Improving volitional competence is crucial for the efficacy of psychosomatic therapy: a controlled clinical trial. *Pychotherapy and Psychosomatics*, *76*, 89-96.

Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification and causal inference. *Biometrics*, *58*, 21–29.

Frisco, M. L., Muller, C., & Frank, K. (2007). Parents' union dissolution and adolescents śchool performance: Comparing methodological approaches. *Journal of Marriage and the Family*, *69*, 721–741.

Gibsons, C. (2003). Privileging the participant: the importance of sub-group analysis in social welfare evaluations. *American Journal of Evaluation*, *24*, 443-469.

Giesen-Bloo, J., Dyck, R. van, Spinhoven, P., Tilburg, W. van, Dirksen, C., Asselt, T. V., Kremers, I., Nadort, M., & Arntz, A. (2006). Outpatient psychotherapy for borderline personality disorder: randomized trial of schema-focused therapy vs. transference-focused psychotherapy. *Archives of General Psychiatry*, *63*, 649-658.

Goldberger, A. (1983). Abnormal selection bias. In S. Karlin, T. Amemiya, & L. Goodman (Eds.), *Studies in econometrics, time series and multivariate statistics.* New York: Academic Press.

Golkaramnay, V., Bauer, S., Haug, S., Wolf, M., & Kordy, H. (2007). The exploration of the effectiveness of group therapy through an internet chat as aftercare: a controlled naturalistic study. *Psychotherapy and Psychosomatics*, *76*, 219-225.

Goodman, L. A. (1978). *Analyzing qualitative/categorical data: Log-linear models and latent structure analysis.* London: Addison Wesley.

Grant, B. F., Hasin, D. S., Stinson, F. S., Dawson, D. A., Chou, S. P., Ruan, W. J., & Pickering, R. P. (2004). Prevalence, correlates, and disability of personality disorders in the united states: results from the national epidemiologic survey on alcohol and related conditions. *Journal of Clinical Psychiatry*, *65*, 948-958.

Greene, W. (1981). Sample selection bias as a specification error: Comment. *Econometrica*, *49*, 795-798.

Grossman, P., Tiefenthaler-Gilmer, U., Raysz, A., & Kesper, U. (2007). Mindfulness training as an intervention for fibromyalgia: evidence of postintervention and 3-year follow-up benefits in well-being. *Psychotherapy and Psychosomatics*, *76*, 226-233.

Gude, T., & Vaglum, P. (2001). One-year follow-up of patients with cluster c personality disorders: a prospective study comparing patients with pure and comorbid conditions within cluster c, and pure c with pure cluster a or b conditions. *Journal of Personality Disorders*, *15*, 216-228.

Gunderson, J. G., Daversa, M. T., Grilo, C. M., McGlashan, T. H., Zanarini, M. C., Shea, M. T., Skodol, A. E., Yen, S., Sanislow, C. A., Bender, D. S., Dyck, I. R., Morey, L. C., & Stout, R. L. (2006). Predictors of 2-year outcome for patients with borderline personality disorder. *American Journal of Psychiatry*, *163*, 822-826.

Guo, S., R, R. B., & Gibbons, C. (2006). Propensity score matching strategies for evaluating substance abuse services for child welfare clients. *Children and Youth Services Review*, *28*, 357-383.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*, 153–162.

Heckman, J. J., Ichimura, H., Smith, J., & Todd, P. E. (1998). Characterizing selection bias using experimental data. *Econometrica*, *66*, 1017–1098.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies*, *64*, 605–654.

Heitjan, D. F., & Rubin, D. B. (1991). Ignorability and coarse data. *Annals of Statistics*, *19*, 2244–2253.

Hellerstein, D. J., Rosenthal, R. N., Pinsker, H., Samstag, L. W., Muran, J. C., & Winston, A. (1998). A randomized prospective study comparing supportive and dynamic therapies: outcome and alliance. *Journal of Psychotherapy Practice and Research*, *7*, 261–271.

Hill, H. J., Waldfogel, J., Brooks-Gunn, J., & Han, W. J. (2005). Maternal employment and child development: a fresh look using newer methods. *Developmental Psychology*, *41*, 833-850.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*, 199–236.

Hodgson, R., Bushe, C., & Hunter, R. (2007). Measurement of long-term outcomes in observational and randomised controlled trials. *British Journal of Psychiatry*, *191*, 78-84.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945-960.

Imai, K., & Dyk, D. A. van. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, *99*, 854–866.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, *87*, 706-710.

Imbens, G. W., & Rubin, D. B. (1997a). Bayesian inference for causal effects in randomized experiments with non-compliance. *The Annals of Statistics*, *25*, 305–327.

Imbens, G. W., & Rubin, D. B. (1997b). Estimating outcome distributions for compliers in instrumental variables models. *Review of Economic Studies*, *64*, 555–574.

Jalan, J., & Ravallion, M. (2003). Estimating the benefit incidence of an antipoverty program by propensity-score matching. *Journal of Business and Economic Statistics*, *21*, 19-30.

Joffe, M. M., & Rosenbaum, P. R. (1999). Invited commentary: Propensity scores. *American Journal of Epidemiology*, *150*, 327–333.

Jung, S. H., Chow, S. C., & Chi, E. M. (2007). A note on sample size calculation based on propensity analysis in nonrandomized trials. *Journal of Biopharmaceutical Statistics*, *17*, 35-41.

Kachele, H., Kordy, H., & Richard, M. (2001). Therapy amount and outcome of inpatient psychodynamic treatment of eating disorders in germany: data from a multicenter study. *Psychotherapy Research*, *11*, 239-257.

Kampen, D. van. (2002). The dapp-bq in the netherlands: factor structure and relationship with basic personality dimensions. *Journal of Personal Disorders*, *16*, 235-254.

Karterud, S., Pedersen, G., Bjordal, E., Brabrand, J., Fris, S., Haaseth, O., Haavaldsen, G., Irion, T., Leirvag, H., Torum, E., & Urnes, O. (2003). Day treatment of patients with personality disorders: experiences from a norwegian treatment research network. *Journal of Personal Disorders*, *17*, 243-262.

Klein, M., Ponds, R. W., & Jolles, P. J. H. J. (1997). Effects of test duration on age-related differences in stroop interference. *Journal of Clinical Experimental Neuropsychology*, *19*, 77–82.

Kosinski, A. S., & Barnhart, H. X. (2003). Accounting for nonignorable verification bias in assessment of diagnostic tests. *Biometrics*, *59*, 163–171.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*, 1-14.

Lambert, M. J., Burlingame, G. M., Umphress, V., Hansen, N. B., Vermeersch, D. A., Clouse, G. C., & Yanchar, S. C. (1996). The reliability and validity of the outcome questionnaire. *Clinical Psychology and Psychotherapy*, *3*, 249-258.

Lauritzen, S. L. (2001). Causal inference from graphical models. In Chapman & Hall (Eds.), *Complex stochastic systems* (pp. 63–107). London: Academic Press.

Lechner, M. (1999). Earnings and employment effects of continuous off-the-job training in east germany after unification. *Journal of Business and Economic Statistics*, *17*, 74-90.

Leichsenring, F. (2004). Randomized controlled versus naturalistic studies: a new research agenda. *Bulletin of the Menninger Clinic*, *68*, 137-151.

Leichsenring, F., Hoyer, J., Beutel, M., Herpertz, S., Hiller, W., Irle, E., Joraschky, P., Konig, H. H., Liz, T. M. de, Nolting, B., Pohlmann, K., Salzer, S., Schauenburg, H., Stangier, U., Strauss, B., Subic-Wrana, C., Vormfelde, S., Weniger, G., Willutzki, U., Wiltink, J., & Leibing, E. (2009). The social phobia psychotherapy research network. the first multicenter randomized controlled trial of psychotherapy for social phobia: rationale, methods and patient characteristics. *Psychotherapy and Psychosomatics*, *78*, 35-41.

Leow, C., Marcus, S., Zanutto, E., & Boruch, R. (2004). Effects of advanced course-taking on math and science achievement: addressing selection bias using propensity scores. *American Journal of Evaluation*, *25*, 461-478.

Lezak, M. D. (2004). *Neurospychological assessment. 4th eddition.* New York: Oxford University Press.

Lieberman, E., Lang, J. M., Cohen, A., D'Agostino, R., Datta, S., & Frigoletto, F. D. (1996). Association of epidural analgesia with cesarean delivery in nulliparas. *Obstetrics and Gynecology*, *88*, 993–1000.

Links, P. S., Mitton, M. J. E., & Steiner, M. (1993). Stability of borderline personality-disorder. *Canadian Journal of Psychiatry*, *38*, 255-259.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, *88*, 125–134.

Little, R. J. A., & Yau, L. (1998). Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using rubin's causal model. *Psychological Methods*, *3*, 147–159.

Livesley, W. J., & Jackson, D. N. (2002). *Manual for the dimensional assessment of personality pathology-basic questionnaire (dapp-bq).* Port Huron: Sigma Press.

Lorentzen, S., & Hoglend, P. A. (2008). Moderators of the effects of treatment length in long-term psychodynamic group psychotherapy. *Psychotherapy and Psychosomatics*, *77*, 321-322.

Lu, B., Zanutto, E., Hornik, R., & Rosenbaum, P. R. (2001). Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association*, *96*, 1245–1253.

Luellen, J. K., Shadish, W. R., & Clark, M. H. (2005). Propensity scores - an introduction and experimental test. *Evaluation Review*, *29*, 530–558.

Lytle, B. W., Blackstone, E. H., Loop, F. D., Houghtaling, P. L., Arnold, J. H., Akhrass, R., McCarthy, P. M., & Cosgrove, D. M. (1999). Two internal thoracic artery grafts are better than one. *The Journal of Thoracic and Cardiovascular Surgery*, *117*, 855-872.

Maat, S. de, Dekker, J., Schoevers, R., & Jonghe, F. de. (2007). The effectiveness of long-term psychotherapy: methodological research issues. *Psychotherapy Research*, *17*, 59-65.

MacLachlan, G. J., & Krishnan, T. (2000). *Causality: models, reasoning, and inference.* Cambridge: Cambridge University Press.

Manen, J. van. (2008). How do intake clinicians use patient characteristics to select treatment for patients with personality disorders? *Psychotherapy and Psychosomatics*, *18*, 711-718.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective.* London: Lawrence Erlbaum Associates.

de Klerk, E., van der Heijde, D., & Landerwé, R. (2003). A compliance questionnaire could discriminate among patients for drug taking behaviour and correct dosing in rheumatic diseases. *Evidence Based Medicine*, *30*, 2469–2475.

R Development Core Team. (2005). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Ten Have, T. R., Joffe, M., & Cary, M. (2003). Causal logistic models for noncompliance under randomized treatment with univariate binary response. *Statistics in Medicine*, *22*, 1255–1283.

van Uffelen, J. G. Z., Chin A Paw, M. J. M., van Mechelen, W., & Hopman-Rock, M. (2008). Walking or vitamin b for cognition in older adults with mild cognitive impairment. *British Journal of Sports Medicine*, *42*, 344–351.

McCrone, P., Marks, I. M., Greist, J. H., Baer, L., Kobak, K. A., Wenzel, K. W., & Hirsch, M. J. (2007). Costeffectiveness of computer-aided behaviour therapy for obsessive-compulsive disorder. *Psychotherapy and Psychosomatics*, *76*, 249-250.

McGlashan, T. H. (1985). The prediction of outcome in borderline personality disorders; in mcglashan th (ed): The borderline: Current empirical research. In T. H. McGlashan (Ed.), *The borderline: Current empirical research* (p. 61-98). Washington: American Psychiatric Press.

McKee, M., Britton, A., Black, N., McPherson, K., Sanderson, C., & Bain, C. (1999). Methods in health services research. interpreting the evidence: choosing between randomised and non-randomised studies. *British Medical Journal*, *319*, 312-315.

Mehta, R. L., Pascual, M. T., Soroko, S., & Chertow, G. M. (2002). Diuretics, mortality, and nonrecovery of renal function in acute renal failure. *Journal of the American Medical Association*, *288*, 2547-2553.

Morgan, S. L. (2001). Counterfactuals, causal effect heterogeneity, and the catholic school effect on learning. *Sociology of Education*, *74*, 341–373.

Morgan, S. L., & Harding, D. J. (2006). Matching estimators of causal effects - prospects and pitfalls in theory and practice. *Sociological Methods and Research*, *35*, 3–60.

Morgan, S. L., & Winship, C. (2007). *Counter-factuals and causal inference.* Cambridge: Cambridge University Press.

Mosis, G., Dieleman, J. P., Stricker, B. C., Lei, J. van der, & Sturkenboom, M. C. J. M. (2006). A randomized database study in general practice yielded quality data but patient recruitment in routine consultation was not practical. *Journal of Clinical Epidemiology*, *59*, 497–502.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*, 115–132.

Muthén, L. K., & Muthén, B. O. (2008). *M-plus: Statistical analysis with latent variables; users guide. version 5.* Los Angeles: Muthén & Muthén.

Nagelkerke, N., Fidler, V., Bernsen, R., & Borgdorff, M. (2000). Estimating treatment effects in randomized clinical trials in the presence of non-compliance. *Statistics in Medicine*, *19*, 1849–1864.

Ogrodniczuk, J. S., Joyce, A. S., LD, L. D. L., Piper, W. E., Steinberg, P. I., & Richardson, K. (2008). Predictors of premature termination of day treatment for personality disorder. *Psychotherapy and Psychosomatics*, *77*, 365-371.

Olsen, R. J. (1980). A least squares correction for selectivity bias. *Econometrica*, *48*(7), 1815–1820.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*, 669–710.

Pfohl, B., Blum, N., & Zimmerman, M. (1997). *Structured interview for dsm-iv personality (sidpiv).* Washington: American Psychiatric Press.

Plakun, E. M. (1991). Prediction of outcome in borderline personality disorder. *Journal of Personality Disorders*, *5*, 93-101.

Potosky, A. L., Legler, J., Albertsen, P. C., Stanford, J. L., Gilliland, F. D., Hamilton, A. S., Eley, J. W., Stephenson, R. A., & Harlan, L. C. (2000). Health outcomes after prostatectomy or radiotherapy for prostate cancer: results from the prostate cancer outcomes study. *Journal of the National Cancer Institute*, *92*, 1582-1592.

Puhani, P. (2000). The heckman correction for sample selection and its critique - a short survey. *Journal of Economic Surveys*, *14*, 153–68.

Quade, D. (1982). Nonparametric analysis of covariance by matching. *Biometrics*, *38*, 597-611.

Rey, A. L. K., & Muthén, B. O. (1964). *Léxamen clinique en psychologie.* Paris: Presses Universitaires de France.

Robins, J. M. (1986). new approach to casual inference in mortality studies with sustained exposure-application to control of the healthy worker survivor effect. *Mathematical Modelling*, *7*, 1393–1512.

Robinson, W. L., Harper, G. W., & Schoeny, M. E. (2003). Reducing substance use among african american adolescents: effectiveness of school-based health centers. *Clinical Psychology: Science and Practice*, *10*, 491-504.

Rosenbaum, P. R. (1991). Discussing hidden bias in observational studies. *Annals of Internal Medicine*, *115*, 901–905.

Rosenbaum, P. R. (1995). *Observational studies, 2nd edition.* New York: Springer Publishing.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41-55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688–701.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.

Rubin, D. B. (1978). Bayesian inference for causal effects. *Annals of Statistics*, *6*, 581–592.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, *127*, 757-763.

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: relating theory to practice. *Biometrics*, *52*, 249-264.

Ryle, A., & Golynkina, K. (2000). Effectiveness of timelimited cognitive analytic therapy of borderline personality disorder: factors associated with outcome. *British Journal of Medical Psychology*, *73*, 197-210.

Sasieni, P. (2001). Estimating prevalence when the true disease status is incompletely ascertained. *Statistics in Medicine*, *20*, 935–949.

Shadish, W. R., & Cook, T. D. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Skodol, A. E., Gunderson, J. G., McGlashan, T. H., Dyck, I. R., Stout, R. L., Bender, D. S., CM, C. M. G., Shea, M. T., Zanarini, M. C., Morey, L. C., Sanislow, C. A., & Oldham, J. M. (2002). Functional impairment in patients with schizotypal, borderline, avoidant, or obsessive-compulsive personality disorder. *American Journal of Psychiatry*, *159*, 276-283.

Skodol, A. W. E., Johnson, J. G., Cohen, P., Sneed, J. R., & Crawford, T. N. (2007). Personality disorder and impaired functioning from adolescence to adulthood. *British Journal of Psychiatry*, *190*, 415-420.

Skodol, A. W. E., Johnson, J. G., Cohen, P., Sneed, J. R., & Crawford, T. N. (2008). The economic burden of personality disorders in mental health care. *Journal of Clinical Psychiatry*, *69*, 259-265.

Spreeuwenberg, M. D., Bartak, A., Croon, M. A., Hagenaars, J. A., Busschbach, J. J. V., Andrea, H., Twisk, J., & Stijnen, T. (2010). 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. *Medical Care*.

StataCorp. (2001). *Statistical software: Release 7.0.* TX: Stata Corporation: College Station.

Stenestrand, U., & Wallentin, L. (2001). Early statin treatment following acute myocardial infarction and 1-year survival. *Journal of the American Medical Association*, *285*, 430-436.

Stravynski, A., Belisle, M., Marcouiller, M., Lavallee, Y. J., & Elie, R. (1994). The treatment of avoidant personality disorder by social skills training in the clinic or in real-life settings. *Canadian Journal of Psychiatry*, *39*, 377-383.

Svartberg, M., Stiles, T. C., & Seltzer, M. H. (2004). Randomized, controlled trial of the effectiveness of short-term dynamic psychotherapy and cognitive therapy for cluster c personality disorders. *American Journal of Psychiatry*, *161*, 810–817.

Teusch, L., Bohme, H., Finke, J., & Gastpar, M. (2001). Effects of client-centered psychotherapy for personality disorders alone and in combination with psychopharmacological treatment: an empirical follow-up study. *Psychotherapy and Psychosomatics*, *70*, 328-336.

Thomas, D. B. R. N. (1992). Characterizing the effect of matching using linear propensity score methods with normal distributions. *Biometrika*, *79*, 797–809.

Uiterwijk, J. M. (2001). *Wais-iii-nl/v.* Lisse: Swets & Zeitlinger.

Verheul, R., Andrea, H., Berghout, C. C., Dolan, C., Busschbach, J. J. V., Kroft, P. J. A. V. der, Bateman, A. W., & Fonagy, P. (2008). Severity indices of personality problems (sipp-118): development, factor structure, reliability and validity. *European Journal of Psychological Assessment*, *20*, 23-34.

Vermunt, J. K. (1997). *Lem 1.0: A general program for the analysis of categorical data.* Tilburg: Tilburg University.

Vermunt, J. K., & Magidson, J. (2000). *Latent gold user's manual.* Belmont: Statistical Innovations Inc.

Vermunt, J. K., & Magidson, J. (2008). *Lg-syntax user's guide: Manual for latent gold 4.5 syntax module.* Belmont: Statistical Innovations Inc.

Vervaeke, G. A. C., & Emmelkamp, P. M. G. (1998). Treatment selection: What do we know? *European Journal of Psychological Assessment*, *14*, 50-59.

Wang, J. X., Donnan, P. T., Steinke, D., & MacDonald, T. M. (2001). The multiple propensity score for analysis of dose-response relationships in drug safety studies. *Pharmacoepidemiology and Drug Safety*, *10*, 105–111.

Wedel, M., & DeSarbo, W. S. (2002). Mixture regression models. In J. A. Hagenaars & A. L. M. (Eds.) (Eds.), *Applied latent class analysis* (p. 366-382). Cambridge: Cambridge University Press.

Westen, D., Novotny, C. M., & Thompson-Brenner, H. (2004). The empirical status of empirically supported psychotherapies: assumptions, findings,and reporting in controlled clinical trials. *Psychological Bulletin*, *130*, 631-663.

Wilberg, T., Urnes, O., Friis, S., Irion, T., Pedersen, G., & Karterud, S. (1999). One-year follow-up of day treatment for poorly functioning patients with personality disorders. *Psychiatric Services*, *50*, 1326-1330.

Winship, C., & Mare, R. D. (1992). Models for sample selection bias. *Annual Review of Sociology*, *18*, 327–350.

Winston, A., Laikin, M., Pollack, J., Samstag, L. W., McCullough, L., & Muran, J. C. (1994). Short-term psychotherapy of personality disorders. *American Journal of Psychiatry*, *151*, 190-194.

Wolfe, F., & Michaud, K. C. (2004). Heart failure in rheumatoid arthritis: rates, predictors, and the effect of anti-tumor necrosis factor therapy. *American Journal of Medicine*, *116*, 305-311.

Yamagata, T., & Orme, C. D. (2005). On testing sample selection bias under the multicollinearity problem. *Econometric Reviews*, *24*, 467–481.

Yau, L. H. Y., & Little, R. J. (2001). Inference for the complier–average causal effect from longitudinal data subject to non–compliance and missing data, with application to a job training assessment for the unemployed. *Journal of the American Statistical Association*, *96*, 1232–1244.

Yoshikawa, H., Magnuson, K. A., Bos, J. M., & Hsueh, J. (2003). Effects of earnings supplement policies on adult economic and middle-childhood outcomes differ for the hardest to employ. *Child Development*, *74*, 1500-1521.

Yue, L. Q. (2007). Statistical and regulatory issues with the application of propensity score analysis to nonrandomized medical device clinical studies. *Journal of Biopharmaceutical Statistics*, *17*, 1-13, discussion 15-17, 19-21, 23-27 passim.

Zanutto, E., Lu, B., & Hornik, R. (2005). Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *Journal of Educational and Behavioral Statistics*, *30*, 59–73.

Zeeck, A., Weber, S., Sandholz, A., Wetzler-Burmeister, E., Wirsching, M., & Hartmann, A. (2009). Inpatient versus day clinic treatment for bulimia nervosa: a randomized trial. *Pychotherapy and Psychosomatics*, *78*, 152-160.

Zhou, X. H. (1998). Comparing accuracies of two screening tests in a two-phase study for dementia. *Journal of the Royal Statistical Society, series C*, *47*, 135–147.

Zhou, X. H., & Castelluccio, P. (2004). Adjusting for non-ignorable verification bias in clinical studies for alzheimer's disease. *Statistics in Medicine*, *23*, 221–230.

Summary

Randomized controlled trials are considered the best proof of effectiveness. Randomization assumes that all known and unknown characteristics of participants are balanced between 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. However, ethical, practical or financial considerations often force researchers in the (para)medical fields to look for alternative research designs, such as a quasi-experimental design. In the case of non-random allocation of participants into experimental groups, there is a large risk that persons in different treatments conditions differ, on average, on pre-treatment characteristics such as age, or motivation. These differences may lead to a selection bias. Selection bias is the bias introduced into a (quasi-)experimental study by the selection of different types of individuals into treatment program(s) and reference program(s). Consequently, the pre-existing differences between participants in the different treatment programs may explain the results of a study, as opposed to true treatment effects.

There are two forms of bias; overt bias and hidden bias. Overt bias is bias due to observed differences, and hidden bias is due to unobserved differences between experimental groups.

Also, when randomization is perfectly carried out using a rather large number of individuals, the intended randomization plan may fail because of what happens later during the implementation of the research design. Selective dropout or non-compliance can also result in selection problems. Researchers in the (para)medical field may therefore encounter a range of selection bias problems in their research and are forced to use statistical techniques that take possible biasing effects in account. The focus of this thesis is on discussing a range of selection bias problems and on presenting statistical techniques that counter

for these selection bias problems.

Chapter 1 discusses in a very general way, the nature, causes and consequences of selection bias problems in experimental and non-experimental studies and ways to overcome these problems. Since, in this thesis, selection bias is studied from the viewpoint of biased causal conclusions, Rubin's precise and well known model of causality is first presented. Next, a few general, well-known methods for countering overt bias are discussed such as matching, stratification and regression adjustment. These traditional approaches are critically evaluated, improved and extended in the remaining chapters of this dissertation.

Chapter 2 discusses the dimensionality problem of matching and stratification in situations where the number of pre-treatment variables is large. As an alternative method, the propensity score method (PS) is discussed. The propensity score is the probability of assignment into the experimental group, given a set of pre-treatment variables. The propensity score method is illustrated step-by step with data coming from a large a Dutch research project named the "Study on Cost-Effectiveness of Personality Disorder Treatment" (*SCEPTRE*). Since the propensity score is mainly used in two-arm studies, the data are divided into a short-term therapy program (up to six months) and a long-term therapy program (more than six months). This has been done for illustrative purposes, although the original treatment variable contained more categories. Differences between the two treatment groups (short versus long treatment duration) in pre-treatment characteristics before and after PS correction is examined to reveal the impact of the PS on outcome differences. In this quasi-experimental study, the PS offered statistical control over the observed pre-treatment differences. When randomization is not possible, a quasi-experimental study using the PS could be a feasible alternative. If implemented carefully, this method is promising for future effectiveness research.

The standard propensity score method has been well developed for (quasi-)experiments with two treatment programs. Since clinicians are often interested in the comparison of multiple (more than two) treatments, there is a need to extend the PS method to multiple treatments. It has been shown that the (multiple) propensity method is possible. So far, its practical application is rare and a practical introduction is lacking. Chapter 3, provides a practical guideline to illustrate the (multiple) multiple propensity score. The method is illustrated step-by-step using data from the *SCEPTRE* study, where the effectiveness

of five different therapies or patients with cluster C personality disorders are compared, differing in setting and duration. With the multiple propensity method, balance was achieved in all relevant pre-treatment variables. The corrected estimated treatment effect was somewhat different than the 'naïve' results. The results indicate that the multiple PS is a feasible method to adjust for observed pre-treatment differences in non-randomized studies where the number of covariates is large and multiple treatments are compared.

In chapter 4, the results from the large and complicated *SCEPTRE* study are discussed from a more clinical point of view. The multiple propensity score is used to compare the effectiveness of five different therapies, differing in setting and duration, for patients with cluster C personality disorders. Since the study had a repeat testing structure, the multiple propensity scores are included in a random intercept multilevel model. In this model, the results are adjusted for (1) the dependency of the data due to repeat testing and for (2) the confounding effect of a large number of observed pre-treatment differences across the psychotherapy programs. Patients in all treatment programs improved on all outcomes 12 months after baseline. Patients receiving short-term inpatient treatment showed more improvement than patients receiving other treatment modalities.

Whereas chapters 1 to 4 discuss statistical methods that counter for overt bias, in chapter 5, attention is paid to two statistical methods that control for hidden bias in quasi-experimental studies. Dealing with hidden bias is more difficult. Existing methods that control for hidden bias are rather unknown in the (para) medical research field. The methods discussed in chapter 5 are (1) the original Heckman two-step method and (2) an extended version using Structural Equation Modeling (SEM). By using four artificial data-sets, the performances of both methods are compared to the results of regression analysis and the propensity score method. In addition, the *SCEPTRE* data are used to compare and illustrate both methods. It is concluded that, especially the Heckman method is very sensitive to misspecification of the selection model and to violations of the normality assumption. When good indicators for a latent tendency to participate in the study are available, SEM analysis is preferred over the Heckman method.

In the occurrence of a perfectly carried out randomization using a rather large number of individuals, the intended randomization plan may fail because

of what happens later during the implementation of the research design. Selective drop-out or non-compliance may also lead to selection problems. Traditional methods for handling differential non-compliance behavior like Intention-To-Treat, Analysis-As-Treated or Per-Protocol-Analysis have been shown to be defective in several aspects. Chapter 6, discusses a latent class version of the instrumental variable approach which yields an unbiased estimate of the complier average causal effect. The chapter presents a number of elaborations of this latent class model. These elaborations pertain to situations in which (a) the outcome variable is only measured indirectly via indicator variables, (b) the experimental interventions has more than two levels and/or (c) a factorial designs is implemented. These methods are applied to data from an experiment that has studied the effects of various physical programs on cognitive functioning in the elderly.

Chapter 7 discusses a different type of selection problem often occurring in diagnostic testing, named verification bias. In medical screening, subjects are often pre-screened by one or multiple non-invasive diagnostic tests and only subjects with at least one positive test are verified for disease status. This strategy may lead to verification bias in estimating the sensitivity and specificity of the tests. Several methods have been developed to adjust for verification bias in cross-sectional studies. In chapter 7, a repeat testing setting is considered where some subjects are directly verified and some other subjects are invited for non-invasive retesting at a later point of time, depending on baseline results. A path model is presented which accounts for non-verification and dependencies among the non-invasive tests. For parameter estimation, an expectation maximization (EM) algorithm is presented. The model is applied to data collected in a large Dutch cervical cancer screening trial. A main goal of this trial was to compare the accuracy of cytological testing to human papillomavirus (HPV) DNA testing. It is illustrated how the cross-sectional and longitudinal dependencies of the two tests can be modeled. Non-verification is studied by fitting missing at random (MAR) and not missing at random (NMAR) models.

Samenvatting (Summary in Dutch)

Gerandomiseerde studies worden beschouwd als het beste bewijs van effectiviteit. Door het toepassen van randomisatie wordt ervan uit gegaan dat alle bekende en onbekende kenmerken van de deelnemers gelijk verdeeld zijn tussen de verschillende behandelingen, behalve voor de ontvangen behandeling. Na randomisatie kan, in theorie, het effect van de behandeling worden geschat door de gemiddelde uitkomsten in behandelgroepen direct met elkaar te vergelijken. Echter, vanwege ethische, praktische of financiële overwegingen, zijn onderzoekers binnen het (para)medische onderzoeksveld vaak aangewezen op alternatieve onderzoeksdesigns, zoals een quasi-experimenteel onderzoek. Wanneer deelnemers op een onwillekeurige manier zijn toegewezen aan de experimentele groepen, is het risico aanzienlijk dat personen op groepsniveau van elkaar verschillen, zoals in de gemiddelde leeftijd of motivatie. Dit verschil in baseline kenmerken kan leiden tot selectie bias. Selectie bias is de vertekening die in een (quasi-)experimenteel onderzoek ontstaat door de selectie van verschillende typen van personen in het behandelprogramma en het referentieprogramma. Als gevolg hiervan kunnen eventuele verschillen in behandeluitkomsten verklaard worden door het verschil in baseline karakteristieken, in plaats van door verschillen in de behandelingen zelf.

Er zijn twee vormen van vertekening, namelijk open (overt) bias en verborgen (hidden) bias. Open bias is vertekening door waargenomen en gemeten verschillen tussen de behandelgroepen en verborgen bias is vertekening door niet-waargenomen verschillen tussen behandelgroepen.

Ook wanneer in studies de randomisatie procedure perfect is uitgevoerd met een vrij groot aantal personen, kan het randomisatieplan mislukken door hetgeen gebeurt gedurende de uitvoering van het onderzoek. Te denken valt

aan selectieve drop-out of aan een gebrek aan therapietrouw van de deelnemers. Deze kunnen leiden tot selectieproblemen. Onderzoekers in het (para) medisch onderzoeksgebied kunnen een aantal selectieproblemen ondervinden tijdens het uitvoeren van hun onderzoek. Dit noodzaakt ze statistische technieken ter voorkoming van vertekening van de resultaten te hanteren. De focus van dit proefschrift ligt op de bespreking en presentatie van een reeks van statistische technieken die het oogmerk hebben om voor verschillende vormen van selectie bias in de analyse te corrigeren.

Hoofdstuk 1 geeft een algemeen overzicht van de aard, oorzaken en gevolgen van selectieproblemen in experimentele en niet-experimentele studies en methoden om deze problemen aan te pakken. Aangezien dit proefschrift selectie bias vanuit het oogpunt van vertekende causale conclusies bespreekt, wordt allereerst Rubin's causale model gepresenteerd. Vervolgens worden enkele algemene en traditionele methoden voor het tegengaan van open bias besproken, zoals matching, stratificatie en regressie analyse. Deze traditionele benaderingen worden kritisch geëvalueerd, verbeterd en uitgebreid in de hoofdstukken 2 tot en met 7 van dit proefschrift.

Hoofdstuk 2 bespreekt het dimensionaliteit probleem dat kan ontstaan bij het gebruik van matching en stratificatie methoden als het aantal baseline verschillen erg groot is. Als een alternatieve methode wordt de propensity score methode (PS) besproken. De propensity score is de kans op toewijzing aan de experimentele groep, gegeven een set van baseline variabelen. De propensity score methode wordt stap–voor–stap geïllustreerd met gegevens afkomstig uit een groot Nederlands onderzoek genaamd "Studie over de kosten-effectiviteit van Persoonlijkheid Stoornissen" (*SCEPTRE*). Aangezien de propensity score voornamelijk wordt gebruikt bij studies die twee behandeleffecten vergelijken, zijn de gegevens ter illustratie ingedeeld in (1) korte-termijn therapie (maximaal zes maanden) en (2) lange termijn therapie (meer dan zes maanden). De oorspronkelijke behandelingen bestonden echter uit meerdere categorien. Verschillen tussen de twee behandelde groepen (korte versus lange duur van de behandeling) in baseline kenmerken vóór en na correctie met de PS zijn onderzocht om de impact van de PS op de resultaten zichtbaar te maken. In deze quasi-experimentele studie heeft de PS tot statistische controle van de waargenomen baseline karakteristieken geresulteerd. Wanneer randomisatie niet mogelijk is, is een quasi-experimentele studie gebruik makend van de

PS een goed alternatief. Indien zorgvuldig uitgevoerd, is deze methode veelbelovend voor toekomstig effectiviteit onderzoek.

De standaard propensity score methode is voornamelijk ontwikkeld voor (quasi-)experimenten waarin twee behandelprogramma's vergeleken worden. Aangezien men in de klinische praktijk vaak geïnteresseerd is in de vergelijking van meerdere (meer dan twee) behandelingen, is er behoefte aan het geschikt maken van de PS-methode voor onderzoek dat meerdere behandelingen vergelijkt. Het is aangetoond dat de meervoudige PS mogelijk is. Tot nu toe is de praktische toepassing hiervan echter zeldzaam en ontbreekt hiervoor een praktische handleiding. In hoofdstuk 3 wordt een praktische handleiding gegeven voor het gebruik van de meervoudige propensity score. De methode wordt stap-voor-stap geïllustreerd, gebruik makende van de gegevens van de *SCEPTRE* studie, waarin de effectiviteit van vijf verschillende therapien of patiënten met cluster C persoonlijkheidsstoornissen, variërend in setting en duur, worden vergeleken. Met de meervoudige PS methode wordt statische controle op alle relevante baseline variabelen verkregen. Het gecorrigeerde geschatte effect van de behandeling bleek enigszins af te wijken van de ongecorrigeerde resultaten. De resultaten geven aan dat de meervoudige PS een haalbare methode is in niet-gerandomiseerde studies waarbij het aantal baseline verschillen groot is en er meerdere behandelingen worden vergeleken.

In hoofdstuk 4 worden de resultaten van het grote en uitgebreide *SCEPTRE* onderzoek besproken vanuit een klinisch oogpunt. De meervoudige propensity score wordt gebruikt om de effectiviteit van vijf verschillende therapieën voor personen met cluster C persoonlijkheidsstoornissen, variërend in duur en setting, te vergelijken. Aangezien de studie uit herhaalde metingen bestond, zijn de meervoudige (PS) meegenomen in een random intercept multilevel model. In dit model zijn de resultaten gecorrigeerd voor (1) de afhankelijkheid van de gegevens door herhaalde metingen en (2) het vertekende effect van een groot aantal baseline verschillen tussen de deelnemers in de verschillende psychotherapie programma's. In alle programma's verbeterden de patiënten op alle uitkomstwaarden 12 maanden na start van de therapie. Patiënten die de korte termijn intramurale behandeling volgden, vertoonden een grotere verbetering dan patiënten die andere vormen van behandeling volgden.

Hoofdstuk 1 tot en met 4 bespreken statistische methoden die corrigeren voor open bias. Echter, voor verborgen bias is het moeilijker om te corrigeren.

Bovendien worden reeds bestaande methoden voor de correctie voor verborgen bias maar zelden toegepast in het (para)medische onderzoeksgebied. In hoofdstuk 5 wordt aandacht besteedt aan twee statistische methoden die controleren voor verborgen bias in quasi-experimentele studies. Deze methoden zijn (1) de traditionele Heckman methode en (2) een alternatieve versie gebruik makend van Structural Equation Modeling (SEM). In vier kunstmatig gegenereerde data-sets worden de prestaties van beide methoden vergeleken met de resultaten van regressieanalyse en de propensity score methode. Daarnaast worden de *SCEPTRE* gegevens gebruikt om de methoden te vergelijken en te illustreren. Geconcludeerd wordt dat vooral de Heckman methode zeer gevoelig is voor misspecificatie van het selectiemodel en voor schendingen van de normaliteit assumptie. Wanneer goede indicatoren voor de neiging tot deelname aan de studie beschikbaar zijn, heeft SEM analyse de voorkeur vergeleken met de Heckman methode.

Zelfs wanneer de randomisatie procedure perfect wordt uitgevoerd met een groot aantal personen, kan het voorgenomen randomisatieplan mislukken gedurende de uitvoering van het onderzoek, vanwege selectieve drop-out of een gebrek aan therapietrouw van de patiënten. Ook dit kan leiden tot selectie problemen. Traditionele methoden ter correctie van differentiële therapietrouw, zoals intention-to-treat analyse, analyse-as-treated of per-protocol analyse, blijken op verschillende aspecten gebreken te vertonen. Hoofdstuk 6 bespreekt een latente klasse model dat gebaseerd is op de instrumentele variabele aanpak. Dit resulteert in een onvertekende schatting van het causale effect voor de groep van compliers. In dit hoofdstuk worden diverse uitbreidingen van dat latente klasse model besproken. Deze uitbreidingen hebben betrekking op situaties waarin (a) de uitkomstvariabele alleen indirect gemeten wordt via de zogenaamde indicator variabelen, (b) de experimentele interventies uit meer dan twee niveaus bestaan en / of (c) een factorieel design wordt gebruikt. De methoden zijn toegepast op gegevens uit een experiment dat de effecten van verschillende fysieke programma's op het cognitief functioneren bij ouderen onderzocht.

Hoofdstuk 7 bespreekt het probleem van verificatie bias, een selectie probleem ten gevolge van het selectief verifiëren van de ware ziektestatus van de deelnemers met een gouden standaard test, op basis van diagnostische screening test uitslagen. In screening onderzoek worden deelnemers vaak vooraf gescreend

met één of meerdere niet-invasieve diagnostische testen. Alleen patiënten met tenminste één positieve testuitslag worden vervolgens gecontroleerd op de aanwezigheid van de ziekte met een invasieve test. Deze strategie kan leiden tot verificatie bias (vertekening) bij de schatting van de sensitiviteit en specificiteit van de diagnostische testen. Verschillende methoden zijn ontwikkeld om verificatie bias in cross-sectioneel onderzoek tegen te gaan. In dit hoofdstuk wordt een herhaalde metingen design besproken waarin een aantal deelnemers direct wordt geverifieerd op ziektestatus na screening en een aantal andere deelnemers niet direct wordt geverifieerd op ziektestatus. Deze deelnemers worden, afhankelijk van de baseline resultaten, uitgenodigd voor een niet-invasieve nacontrole op een later tijdstip. In dit hoofdstuk wordt een pad model gepresenteerd waarin gecorrigeerd wordt voor non-verificatie en afhankelijkheden tussen de niet-invasieve tests. Voor het schatten van parameters, is een EM algoritme ontwikkeld. Het model wordt toegepast op gegevens die zijn verzameld in een groot Nederlands bevolkingsonderzoek naar baarmoederhalskanker. Eén van de belangrijkste doelen van dit onderzoek was om de precisie van de cytologische test te vergelijken met die van de human papillomavirus (HPV) DNA-test. De manier waarop cross-sectionele en longitudinale afhankelijkheden van de twee testen kunnen worden gemodelleerd, is geïllustreerd. Non-verificatie is bestudeerd aan de hand van Missing At Random (MAR) en Not Missing At Random (NMAR) modellen.

Dankwoord

Toen ik afstudeerde gaf John mij een zelfgemaakt schilderijtje met daarop de tekst "In een wereld vol ambitie, zit ik fluitend op mijn fietsie!". Zo zag hij mij. Helaas kan John mijn promotie niet meer meemaken. Aan hem draag ik mijn proefschrift op (John, † 2005).

Mijn proefschrift is een feit! Hoog tijd voor een dankwoord. Iedereen die hoe dan ook heeft bijgedragen aan de totstandkoming van mijn proefschrift wil ik bedanken en een aantal mensen in het bijzonder.

Allereerst richt ik me tot mijn promotor Jacques Hagenaars en copromotor Marcel Croon.

Jacques, ik dank je voor alle begeleiding en steun die je me de afgelopen jaren hebt gegeven. Ik wil je bedanken voor de vrijheid die je mij hebt gegeven om mezelf tijdens mijn promotie verder te ontwikkelen, met name door een combinatie mogelijk te maken van promoveren en werken bij de vakgroep epidemiologie en biostatistiek van het VU Medisch Centrum. Je hebt mij door de laatste loodjes heen gesleept.

Marcel, jouw onuitputtelijke kennis van methodologie en statistiek en je sterke analytische vermogen hebben mij zicht gegeven op waar mijn proefschrift echt over ging en moest gaan. Ik heb veel geleerd van jouw eigenschap je vast te bijten aan een methodologisch probleem. Ook jou wil ik uitdrukkelijk bedanken voor alle begeleiding die je me hebt gegeven en het vertrouwen dat je in mij hebt getoond.

Ik dank de vakgroep Epidemiologie en Biostatistiek van het VU Medisch Centrum.

Maarten, Dick en Bernard, bedankt dat ik tijdens mijn promotie in deeltijd

bij jullie mocht komen werken. Jullie hebben mij aangenomen als biostatisticus en hebben een groot vertrouwen in mij gehad. De combinatie was voor mij ideaal.

Jos, bedankt voor je rol als mijn mentor en voor alle methodologische kennis die je met mij hebt gedeeld.

Hans, samen hebben we veel energie gestoken in het analyseren van de HPV data-set. Het laatste hoofdstuk heb ik helemaal te danken aan de samenwerking met jou. Ik heb veel geleerd van je kritische blik en "helicopter view". Ik heb veel aan jouw hulp gehad en mijn waardering voor jou is groot.

De overige collega's bedank ik voor hun interesse en voor de goede, warme en motiverende sfeer waarin ik heb mogen werken.

De Hogeschool Zuyd en de Universiteit van Maastricht wil ik ook bedanken. Daar heb ik de kans gekregen om de laatste punten op de 'i' te zetten van mijn proefschrift.

Luc, bedankt dat je op dit moment voor mij de weg vrijmaakt om een mooie combinatie van praktisch en academisch werk mogelijk te maken.

Voor mijn proefschrift was ik grotendeels afhankelijk van de bereidheid van onderzoekers om aan mij hun data beschikbaar te stellen. Het was niet altijd gemakkelijk om dit voor elkaar te krijgen. Sommige onderzoekers waren bang dat, met de 'nieuwe' methoden, hun oorspronkelijke onderzoeksresultaten zouden veranderen en waren daarom huiverig om deze data aan mij beschikbaar te stellen. Iedereen bedankt die deze stap wel heeft gezet. Allereerst wil ik het psychotherapeutisch centrum 'De Viersprong' bedanken voor het beschikbaar stellen van de *SCEPTRE* data. In het bijzonder wil ik hierbij Anna noemen.

Anna, jij hebt mij altijd weer weten te inspireren in tijden waarin ik het zelf niet meer zag zitten. Dat ik jou ben tegengekomen bij de EMGO cursus van Jos Twisk is een geschenk uit de hemel geweest. Jouw gedrevenheid en energie werkten voor mij erg stimulerend. Het leek voor mij of wij samen een puzzel aan het maken waren. Het eerste deel van mijn proefschrift is dan ook geheel in samenwerking met jou tot stand gekomen. Het is fijn om je te kennen en ik hoop je in de toekomst in de privésfeer te mogen blijven ontmoeten.

Ik wil ook graag Jan, Helene en Theo bedanken voor hun vertrouwen, steun en altijd vriendelijke woorden.

Tot slot, wil ik Marijke Chin a Paw en Evert Verhagen van het EMGO instituut en Chris Meijer van het VU Medisch Centrum bedanken voor het beschikbaar stellen van hun data-sets.

Natuurlijk kan ik nog een hele lijst mensen noemen die de afgelopen jaren belangrijk voor mij zijn geweest, zowel binnen als buiten mijn promotieproject. Ik noem enkelen speciaal bij naam, maar ik besef dat nog vele anderen belangrijk zijn geweest.

Joost, je bent jarenlang mijn kamergenoot geweest en je was mijn steun en toeverlaat. Bedankt voor alle kopjes thee die je voor mij hebt gemaakt, de leuke en fijne gesprekken op ons kamertje en je leuke humor. Jammer dat we elkaar nog maar zo zelden zien.

Meike en Carmen, bedankt voor alle genegenheid en gesprekjes over 'ditjes en datjes'.

Luc, je bent in de loop der jaren mijn maatje geworden. Vandaar dat je vandaag ook mijn paranimf bent. Ik zal al onze gesprekken nooit vergeten. Ik zal blijven proberen je over te halen om naar Zuid–Limburg te komen!

Liesbeth, van jou en Luc heb ik het lesgeven geleerd. Ik heb jouw lieve en open karakter erg gewaardeerd en hoop je toch wat vaker tegen te komen.

Wilco, Andries, Wobbe, Marcel en John, bedankt voor jullie uitputtende interesse en de leuke liedjes die we samen hebben gemaakt. Ik verwacht dan ook een geweldig lied voor mijn promotie!

Marieke, bedankt voor alle secretariële ondersteuning die je de afgelopen jaren hebt gegeven. Alle overige collega's van de Universiteit van Tilburg, vakgroep methoden en technieken van onderzoek, dank ik voor hun collegialiteit.

Tenslotte nog een paar persoonlijke opmerkingen.

Cor en Dingena, mijn dank aan jullie kan ik moeilijk onder woorden brengen. Ik heb van jullie een goede basis meegekregen en die zal ik altijd met me mee dragen. Bedankt dat jullie al naar mijn "eigen wijsje" wilden luisteren toen ik nog heel jong was. Papa, bedankt voor je liefde, je vertrouwen, je luisterende oor en je kritische blik. Mama, bedankt voor alle moederlijke adviezen die je me de afgelopen jaren hebt gegeven, je onuitputtelijke liefde en voor alle hulp in mijn "huishouden van Jan Steen".

Maarten & Kristel en Jaap & Mam, jullie belangstelling en liefde voor mij

is groot.

Pier & Monique, Joep & Fleur, Han en René, jullie zijn echt mijn familie geworden!

Ze zeggen dat echte vrienden maar op één hand te tellen zijn, maar: Nancy & Robert, Judith & Guido, Elke & Martijn, Floris, Marcel, Bas & Vivian, Meke, Koen, Kasper, Marrigje & Co, Orm, Moon, Dino & Stefan, Carlijn, Faisca & Dave en Tanja: bedankt voor jullie vriendschap! Jullie zorgen altijd voor een goed gesprek of feestje en houden mij met beide benen op de aarde!

Sabine, je bent vandaag mijn paranimf, maar eigenlijk sta je altijd achter me op alle mooie en moeilijke momenten. Bedankt voor je vriendschap!

Kasper, bedankt voor alle avonden en nachten dat je, zo trouw als een hond kan zijn, bij mijn voeten hebt gelegen terwijl ik aan het werken was, zelfs bij het schrijven van dit dankwoord. Met jou zit ik nooit alleen. De wandelingen met jou vind ik heerlijk!

Geert, je bent mijn lief. Ik wil je bedanken voor het jou-mij zijn. We hebben samen aan ons thuis gebouwd (letterlijk en figuurlijk) waar het warm, veilig en fijn is. Je houdt voor mij de weg vrij voor de toekomst. Praktisch gezien bedank ik je voor alle heerlijke maaltijden die je voor mij hebt gekookt en alle kopjes thee die vervolgens niet opgedronken en koud bleven staan naast mijn computer. Ik geef je nog antwoord op alle onbeantwoorde vragen die je me hebt gesteld. Je bent het belangrijkste in mijn leven. Ik ben dik tevreden en trots op ons!

Marieke

Index