Methods to adjust for confounding

Propensity scores and instrumental variables

Edwin Martens

Druk: Universal Press, Veenendaal

Omslag: gebaseerd op een ets van Sjoerd Bakker (www.sjoerd-bakker.nl)

Bijschrift bij omslag: “In sommige omgevingen is een kam gewoon de meest aangewezen *ontwarringsmethode*” (EPM)

CIP-gegevens Koninklijke Bibliotheek, Den Haag

Martens, E.P.

Methods to adjust for confounding. Propensity scores and instrumental variables

Thesis Utrecht - with ref. - with summary in Dutch

ISBN 978-90-393-4703-4

METHODS TO ADJUST FOR CONFOUNDING

Propensity scores and instrumental variables

METHODEN OM VOOR CONFOUNDING TE CORRIGEREN

Propensity scores en instrumentele variabelen

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit Utrecht
op gezag van de rector magnificus, prof. dr. J.C. Stoof,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op
maandag 10 december 2007 des middags te 4.15 uur

door
Edwinus Paulus Martens

geboren op 4 november 1964 te Rotterdam

Promotor: Prof. dr. A. de Boer

Co-promotoren: Dr. O.H. Klungel

Dr. W.R. Pestman

CONTENTS

1 Introduction **7**

2 An overview of methods **11**

2.1 Methods to assess intended effects of drug treatment in observational studies are reviewed 13

3 Strengths and limitations of adjustment methods **31**

3.1 "Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: A Monte Carlo study" . . . 33

3.2 An important advantage of propensity score methods compared to logistic regression analysis 39

3.3 Instrumental variables: application and limitations 51

4 Application of adjustment methods **69**

4.1 Comparing treatment effects after adjustment with multivariable Cox proportional hazards regression and propensity score methods 71

4.2 A non-parametric application of instrumental variables in survival analysis 85

5 Improvement of propensity score methods **101**

5.1 The use of the overlapping coefficient in propensity score methods . . 103

5.2 Measuring balance in propensity score methods 119

6 Discussion **137**

Summary **149**

Samenvatting **157**

Dankwoord **167**

About the author / Over de schrijver **168**

CHAPTER 1

INTRODUCTION

RANDOMIZED AND OBSERVATIONAL STUDIES

Randomization is one of the most important attainments in the last century concerning research on the effect of treatment.[1,2] In medical research the randomized controlled trial is the gold standard for quantifying treatment effects. The major characteristic of this type of studies is the control over treatments by the researcher by means of randomization. Randomization can be defined as the random allocation of experimental units across treatment groups. An *observational study* on the other hand lacks such random allocation of subjects. William Cochran was the first to define an observational study:

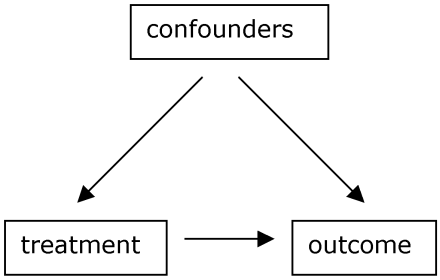
> "... the objective is to elucidate cause-and-effect relationships ... [in which] it is not feasible to use controlled experimentation, in the sense of being able to impose the procedures or treatments whose effects it is desired to discover or to assign subjects at random to different procedures."[3]

Barriers to random assignment are for instance ethical (one of the treatments is harmful), political (it is not possible for an individual researcher to influence political processes), personal (subjects do not want to change their habits) or economic (randomized control over treatments is too costly). Without such random assignment uncertainty about the true effect of treatment is in general greater and alternative explanations for the observed effect are easier to give. If such explanations are also plausible these should be subject to future investigations to either increase or decrease the evidence for the causal relationship between treatment and outcome. Principles on the *design and analysis* of studies are of utmost importance, such as the selection of covariates and the method of analysis.[4] Examples of major medical findings that came from observational studies are for instance the causal effect of smoking on lung cancer[5] and the causal effect of the use of DES on the presence of vaginal cancer.[6]

CONFOUNDING

The most important challenge in observational studies when treatment effect estimation is involved, is to combat *confounding*. Confounding is the phenomenon that an effect estimation (quantitative association between treatment and outcome) is distorted by prognostic factors of the outcome which are unequally distributed over treatment modalities. This is illustrated in Figure 1.1. While in randomized studies imbalances between known and unknown prognostic factors are largely suppressed by the randomization procedure, in observational studies these imbalances generally exist. Therefore, effort should be made to adjust the estimated treatment effect for these confounding factors as much as possible. Subject of this thesis are statistical methods that have the objective to adjust for the non-random assignment of individuals to treatments.

Figure 1.1: Illustration of the concept of confounding



OUTLINE OF THE THESIS

In Chapter 2 we give an overview of such methods, discuss their use, advantages and limitations. In the other chapters we focus on two of these methods, *propensity scores* and *instrumental variables*. Chapter 3 covers some specific advantages and limitations of these methods. An overlooked advantage of propensity scores is the subject of Sections 3.1 and 3.2, whereas the application, the assumptions and the limitations of instrumental variables are discussed in Section 3.3. In Chapter 4 we demonstrated how propensity scores and instrumental variables can be used with censored survival data. In Section 4.1 different propensity score methods were applied and in Section 4.2 tools are given for calculating survival probabilities based on an instrumental variable. In Chapter 5 we suggested improvements in propensity score applications, especially by proposing measures that quantify the balance reached in propensity score modelling. Chapter 6 contains the main results, the strengths and limitations of the chosen methods and implications and recommendations for future research.

REFERENCES

[1] Fisher RA. The arrangement of field experiments. *J Ministry Agric*, 33:503–513, 1926.

[2] Neyman J. On the application of probability theory to agricultural experiments. essay on principles. Section 9. *Roczniki Nauk Roiniczych, Tom X*, 19:1–51, 1923. Reprinted in *Statistical Science* 1990;5:463-480, with discussion by T. Speed and D. Rubin.

[3] Cochran WG. The planning of observational studies of human populations (with discussion). *J Royal Stat Society Series A*, 128:134–155, 1965.

[4] Rosenbaum PR. *Observational studies, 2nd edition*. Springer-Verlag, New York, 2002.

[5] United States Surgeon General's Advisory Committee Report. *Smoking and Health*. US Department of Health, Education and Welfare, 1964.

[6] Herbst A, Ulfelder H, Poskanzer D. Adenocarcinoma of the vagina: Association of maternal stilbestrol therapy with tumor appearance in young women. *N Engl J Med*, 284:878–881, 1971.

CHAPTER 2

AN OVERVIEW OF ADJUSTMENT METHODS

2.1 METHODS TO ASSESS INTENDED EFFECTS OF DRUG TREATMENT IN OBSERVATIONAL STUDIES ARE REVIEWED

Olaf H. Klungel[a], Edwin P. Martens[a,b], Bruce M. Psaty[c], Diederik E. Grobbee[d], Sean D. Sullivan[e], Bruno H.Ch. Stricker[f], Hubert G.M. Leufkens[a], A. de Boer[a]

[a] *Department of Pharmacoepidemiology and Pharmacotherapy, Utrecht Institute of Pharmaceutical Sciences (UIPS), Utrecht University, Utrecht, the Netherlands*
[b] *Centre for Biostatistics, Utrecht University, Utrecht, the Netherlands*
[c] *Cardiovascular Health Research Unit, Medicine, Health Services, and Epidemiology, University of Washington, Seattle, WA, USA*
[d] *Julius Centre for Health Sciences and Primary Care, Utrecht Medical Centre (UMC), Utrecht, the Netherlands*
[e] *Departments of Pharmacy and Health Services, University of Washington, Seattle, WA, USA*
[f] *Department of Epidemiology and Biostatistics, Erasmus University Rotterdam, Rotterdam, the Netherlands*

Journal of Clinical Epidemiology 2004; 57:1223-1231

ABSTRACT

Background and objective: To review methods that seek to adjust for confounding in observational studies when assessing intended drug effects.
Methods: We reviewed the statistical, economical and medical literature on the development, comparison and use of methods adjusting for confounding.
Results: In addition to standard statistical techniques of *(logistic) regression* and *Cox proportional hazards regression*, alternative methods have been proposed to adjust for confounding in observational studies. A first group of methods focuses on the main problem of nonrandomization by balancing treatment groups on observed covariates: *selection, matching, stratification, multivariate confounder score,* and *propensity score methods*, of which the latter can be combined with stratification or various matching methods. Another group of methods looks for variables to be used like randomization in order to adjust also for unobserved covariates: *instrumental variable methods, two-stage least squares*, and *grouped-treatment approach*. Identifying these variables is difficult, however, and assumptions are strong. *Sensitivity analyses* are useful tools in assessing the robustness and plausibility of the estimated treatment effects to variations in assumptions about unmeasured confounders.
Conclusion: In most studies regression-like techniques are routinely used for adjustment for confounding, although alternative methods are available. More complete empirical evaluations comparing these methods in different situations are needed.

Keywords: Review; Confounding; Observational studies; Treatment effectiveness; Intended drug effects; Statistical methods

INTRODUCTION

In the evaluation of intended effects of drug therapies, well-conducted *randomized controlled trials* (RCTs) have been widely accepted as the scientific standard.[1] The key component of RCTs is the randomization procedure, which allows us to focus on only the outcome variable or variables in the different treatment groups in assessing an unbiased treatment effect. Because adequate randomization will assure that treatment groups will differ on all known and unknown prognostic factors only by chance, probability theory can easily be used in making inferences about the treatment effect in the population under study (confidence intervals, significance). Proper randomization should remove all kinds of potential selection bias, such as physician preference for giving the new treatment to selected patients or patient preference for one of the treatments in the trial.[2,3] Randomization does not assure equality on all prognostic factors in the treatment groups, especially with small sample sizes, but it assures confidence intervals and p-values to be valid by using probability theory.[4]

There are settings where a randomized comparison of treatments may not be feasible due to ethical, economic or other constraints.[5] Also, RCTs usually exclude particular groups of patients (because of age, other drug usage or non-compliance); are mostly conducted under strict, protocol-driven conditions; and are generally of shorter duration than the period that drugs are used in clinical practice.[6,7] Thus, RCTs typically provide evidence of what can be achieved with treatments under the controlled conditions in selected groups of patients for a defined period of treatment.

The main alternatives are *observational studies*. Their validity for assessing intended effects of therapies has long been debated and remains controversial.[8–10] The recent example of the potential cardiovascular risk reducing effects of hormone replacement therapy (HRT) illustrates this controversy.[11] Most observational studies indicated that HRT reduces the risk of cardiovascular disease, whereas RCTs demonstrated that HRT increases cardiovascular risk.[12] The main criticism of observational studies is the absence of a randomized assignment of treatments, with the result that uncontrolled confounding by unknown, unmeasured, or inadequately measured covariates may provide an alternative explanation for the treatment effect.[13,14]

Along with these criticisms, many different methods have been proposed in the literature to assess treatment effects in observational studies. With all these methods, the main objective is to deal with the potential bias caused by the non-randomized assignment of treatments, a problem also known as *confounding*.[15]

Here we review existing methods that seek to achieve valid and feasible assessment of treatment effects in observational studies.

DESIGN FOR OBSERVATIONAL STUDIES

A first group of methods dealing with potential bias following from non-randomized observational studies is to narrow the treatment and/or control group in order to create more comparable groups on one or more measured characteristics. This can be done by selection of subjects or by choosing a specific study design. These methods can also be seen as a first step in removing bias, where a further reduction of bias will be attained by any data-analytical technique.

HISTORICAL CONTROLS

Before the introduction and acceptance of the RCT as the gold standard for assessing the intended effect of treatments, it was common to compare the outcome of treated patients with the outcome of *historical controls* (patients previously untreated or otherwise treated).[16] An example of this method can be found in Kalra *et al*.[17] The authors assessed the rates of stroke and bleeding in patients with atrial fibrillation receiving warfarin anticoagulation therapy in general medical clinics and compared these with the rates of stroke and bleeding among similar patients with atrial fibrillation who received warfarin in a RCT.

Using historical controls as a comparison group is in general a problematic approach, because the factor time can play an important role. Changes of the characteristics of a general population or subgroup over time are not uncommon.[18] Furthermore, there may exist differences in population definitions between different research settings.

CANDIDATES FOR TREATMENT

If current treatment guidelines exist, the comparison between the treated and the untreated group can be improved by choosing for the untreated group only those subjects who are candidates for the treatment under study according to these guidelines. As a preliminary selection, this method was used in a cohort study to estimate the effect of drug treatment of hypertension on the incidence of stroke in the general population by selecting candidates on the basis of their blood pressure and the presence of other cardiovascular risk factors.[19] The selection of a cohort of candidates for treatment can also be conducted by a panel of physicians after presenting them the clinical characteristics of the patients in the study.[20]

COMPARING TREATMENTS FOR THE SAME INDICATION

When different classes of drugs, prescribed for the same indication, have to be studied, at least some similarity in prognostic factors between treatment groups occurs naturally. This strategy was used in two case-control studies to compare the effects of different antihypertensive drug therapies on the risks of myocardial infarction and ischemic stroke.[21,22] Only patients who used antihypertensive drugs for the indication hypertension were included in these studies (and also some subgroups that had other indications such as angina for drugs that can be used to treat high blood pressure were removed).

CASE-CROSSOVER AND CASE-TIME-CONTROL DESIGN

The use of matched case-control (case-referent) studies when the occurrence of a disease is rather rare is a well-known research design in epidemiology. This type of design can also be adopted when a strong treatment effect is suspected[23] or when a cohort is available from which the subjects are selected (nested case-control study).[24] Variations of this design have been proposed to control for confounding due to differences between exposed and unexposed patients. One such variant is the *case-crossover study*, in which event periods are compared with control periods within cases of patients who experienced an event. This study design may avoid bias resulting from differences between exposed and nonexposed patients, but variations in the underlying disease state within individuals could still confound the association between treatment and outcome.[25] An extension of this design is the *case-time-control design*, which takes also into account changes of exposure levels over time. With this design and with certain assumptions confounding due to time trends in exposure can be removed, but variations in the severity of disease over time within individuals, although probably correlated with exposure levels, cannot be controlled.[26–28] In a study comparing the effect of high and moderate β-antagonist use on the risk of fatal or near-fatal asthma attacks, the odds ratio (OR) from a case-time-control analysis controlling for time trends in exposure, turned out to be much lower (OR$=1.2$, 95% confidence interval, CI 95%: $0.5, 3.0$) than in a conventional case-control analysis (OR$=3.1$, CI 95%: $1.8, 5.4$).[27]

Advantages of these designs in which each subject is its own control, are the considerably reduced intersubject variability and the exclusion of alternative explanations from possible confounders. These methods are on the other hand of limited use, because for only some treatments the outcome can be measured at both the control period and the event period, and thereby excluding possible carryover effects.

DATA-ANALYTICAL TECHNIQUES

Another group of bias reducing methods are the data-analytical techniques, which can be divided into model-based techniques (regression-like methods) and methods without underlying model assumptions (stratification and matching).

STRATIFICATION AND MATCHING

Intuitive and simple methods to improve the comparison between treatment groups in assessing treatment effects, are the techniques of *stratification* (subclassification) and *matching* on certain covariates as a data-analytical technique. The limitations and advantages of these methods are in general the same. Advantages are (i) clear interpretation and communication of results, (ii) direct warning when treatment groups do not adequately overlap on used covariates, and (iii) no assumptions about the relation between outcome and covariates (e.g., linearity).[29,30] The main limitation of these techniques is, that in general only one or two covariates or rough

strata or categories are possible. More covariates will easily result in many empty strata in case of stratification and many mismatches in case of matching. Another disadvantage is that continuous variables have to be classified, using (mostly) arbitrary criteria.

These techniques can easily be combined with methods like propensity scores and multivariate confounder score, as will be discussed below, using the advantages of clear interpretation and absence of assumptions about functional relationships.

ASYMMETRIC STRATIFICATION

A method found in the literature that is worth mentioning, is *asymmetric stratification*.[31] Compared to cross-stratification of more covariates, in this method each stratum of the first covariate is subdivided by the covariate that has highest correlation with the outcome within that stratum. For instance, men are subdivided on the existence of diabetes mellitus because of the strongest relationship with the risk of a stroke, and women are subdivided by the history of a previous cardiovascular disease. By pooling all treatment effects in the strata in the usual way, a corrected treatment effect can be calculated. Although by this method more covariates can be handled than with normal stratification, most of them will be partly used. We are unaware of any medical study in which this method has been used.

COMMON MULTIVARIABLE STATISTICAL TECHNIQUES

Compared to selection, restriction, stratification, or matching, more advanced multivariable statistical techniques have been developed to reduce bias due to differences in prognosis between treatment groups in observational studies.[32] By assessing a model with outcome as the dependent and type of treatment as the independent variable of interest, many prognostic factors can be added to the analysis to adjust the treatment effect for these confounders.Well known and frequently used methods are *multivariable linear regression, logistic regression*, and *Cox proportional hazards regression* (survival analysis). Main advantage over earlier mentioned techniques is that more prognostic variables, quantitative and qualitative, can be used for adjustment, due to a model that is imposed on the data. It is obvious that also in these models the number of subjects or the number of events puts a restriction on the number of covariates; a ratio of 10 to 15 subjects or events per independent variable is mentioned in the literature.[33,34]

An important disadvantage of these techniques when used for adjusting a treatment effect for confounding, is the danger of extrapolations when the overlap on covariates between treatment groups is too limited. While matching or stratification gives a warning or breaks down, regression analysis will still compute coefficients. Mainly when two or more covariates are used, a check on adequate overlap of the joint distributions of the covariates will be seldom performed. The use of a functional form of the relationship between outcome and covariates is an advantage for dealing with more covariates, but has its drawback, mainly when treatment groups have different covariate distributions. In that case, the results are heavily dependent on the chosen relationship (e.g., linearity).

PROPENSITY SCORE ADJUSTMENT

An alternative way of dealing with confounding caused by nonrandomized assignment of treatments in cohort studies, is the use of *propensity scores*, a method developed by Rosenbaum & Rubin.[35] D'Agostino found that "the propensity score for an individual, defined as the conditional probability of being treated given the individual's covariates, can be used to balance the covariates in observational studies, and thus reduce bias."[36] In other words, by this method a collection of covariates is replaced by a single covariate, being a function of the original ones. For an individual i ($i = 1, ..., n$) with vector x_i of observed covariates, the propensity score is the probability $e(x_i)$ of being treated ($Z_i = 1$) versus not being treated ($Z_i = 0$):

$$e(x_i) = \Pr(Z_i = 1 \big| X_i = x_i) \tag{2.1}$$

where it is assumed that the Z_i are independent, given the Xs.

By using logistic regression analysis, for instance, for every subject a probability (propensity score) is estimated that this subject would have been treated, on the basis of the measured covariates. Subjects in treatment and control groups with (nearly) equal propensity scores will tend to have the same distributions of the covariates used and can be considered similar. Once a propensity score has been computed, this score can be used in three different ways to adjust for the uncontrolled assignment of treatments: (i) as a matching variable, (ii) as a stratification variable, and (iii) as a continuous variable in a regression model (covariance adjustment). Examples of these methods can be found in two studies of the effect of early statin treatment on the short-term risk of death.[37,38]

The most preferred methods are stratification and matching, because with only one variable (the propensity score) the disadvantages noted earlier for matching and stratification disappear and the clear interpretation and absence of model-based adjustments remain as the main advantages. When classified into quintiles or deciles, a stratified analysis on these strata of the propensity score is most simple to adopt. Within these classes, most of the bias due to the measured confounders disappears. Matching, on the other hand, can be much more laborious because of the continuous scale of the propensity score. Various matching methods have been proposed. In all these methods, an important role is given to the distance matrix, of which the cells are most often defined as simply the difference in propensity score between treated and untreated patients. A distinction between methods can be made between *pair-matching* (one treated to one untreated patient) and *matching with multiple controls* (two, three, or four). The latter method should be used when the number of untreated patients is much greater than the number of treated patients; an additional gain in bias reduction can be reached when a variable number per pair, instead of a fixed number, is used.[39] Another distinction can be made between *greedy methods* and *optimal methods*. A greedy method selects at random a treated patient and looks for an untreated patient with smallest distance to form a pair. In subsequent steps, all other patients are considered for which a match can be made within a defined maximum dis-

tance. An optimal method, on the other hand, takes the whole distance matrix into account to look for the smallest total distance between all possible pairs. An optimal method combined with a variable number of controls should be the preferred method.[40]

The method of propensity scores was evaluated in a simulation study, and it was found that the bias due to omitted confounders was of similar magnitude as for regression adjustment.[41] The bias due to misspecification of the propensity score model was, however, smaller than the bias due to misspecification of the multivariable regression model. Therefore, propensity score adjustment is less sensitive to assumptions about the functional form of the association of a particular covariate with the outcome (e.g., linear or quadratic).[35] Recently, the propensity score method was compared to logistic regression in a simulation study with a low number of events and multiple confounders.[42] With respect to the sensitivity of the model misspecification (robustness) and empirical power, the authors found the propensity score method to be superior overall. With respect to the empirical coverage probability, bias, and precision, they found the propensity score method to be superior only when the number of events per confounder was low (say, 7 or less). When there were more events per confounder, logistic regression performs better on the criteria of bias and coverage probability.

MULTIVARIATE CONFOUNDER SCORE

The *multivariate confounder* score was suggested by Miettinen as a method to adjust for confounding in case-control studies.[43] Although Miettinen did not specifically propose this method to adjust for confounding in studies of intended effects of treatment, the multivariate confounder score is very similar to the propensity score, except that the propensity score is not conditional on the outcome of interest, whereas the multivariate confounder score is conditional on not being a case.[43]

The multivariate confounder score has been evaluated for validity.[44] Theoretically and in simulation studies, this score was found to exaggerate significance, compared to the propensity score. The point estimates in these simulations were, however, similar for propensity score and multivariate confounder score.

INSTRUMENTAL VARIABLES

A technique widely used in econometrics, but not yet generally applied in medical research, is the use of *instrumental variables* (IV). This method can be used for the estimation of treatment effects (the effect of treatment on the treated) in observational studies as an alternative to making causal inferences in RCTs.[45] In short, an instrumental variable is an observable factor associated with the actual treatment but not directly affecting outcome. Unlike standard regression models, two equations are needed to capture these relationships:

$$D_i = \alpha_0 + \alpha_1 Z_i + \nu_i \tag{2.2}$$

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i \tag{2.3}$$

where Y_i is outcome, D_i is treatment, Z_i is the instrumental variable or assignment, and $\alpha_1 \neq 0$. Both treatment D and assignment Z can be either continuous or dichotomous. In case of a dichotomous D, equation 2.2 can be written as $D_i^* = \alpha_0 + \alpha_1 Z_i + \nu_i$, where D_i^* is a latent index ($D_i^* > 0 \rightarrow D_i = 1$; otherwise $D_i = 0$).

By equation 2.2, it is explicitly expressed that it is unknown how treatments are assigned (at least we know it was not random) and that we like to explain why one is treated and the other is not by a variable Z. Substituting equation 2.2 into 2.3 gives:

$$Y_i = (\beta_0 + \beta_1 \alpha_0) + \beta_1 \alpha_1 Z_i + (\beta_1 \nu_i + \epsilon_i) \tag{2.4}$$

The slope $\beta_1 \alpha_1$ can be estimated by least squares regression and is, when Z is dichotomous, the difference in outcome between $Z = 0$ and $Z = 1$ (i.e., the intention-to-treat estimator). In order to estimate the direct treatment effect β_1 of treatment D on outcome Y, this estimator $\beta_1 \alpha_1$ must be divided by α_1, the effect of Z on D from equation 2.2. As an illustration, it can be seen that in case of a perfect instrument (e.g., random assignment), a perfect relationship exists between Z and D and the parameter $\alpha_1 = 1$, in which case the intention-to-treat estimator and the instrumental variable estimator coincide. By using two equations to describe the problem, the implicit but important assumption is made that Z has no effect on outcome Y other than through its effect on treatment D (cov$[Z_i, e_i]= 0$). Other assumptions are that $\alpha_1 \neq 0$ and that there is no subject i "who does the opposite of its assignment".[46] This is illustrated in the following example.

One of the earliest examples of the use of instrumental variables (simultaneous equations) in medical research was in the study of Permutt and Hebel, where the effect of smoking on birth weight was studied.[47] The treatment consisted of encouraging pregnant women to stop smoking. The difference in mean birth weight between the treatment groups, the intention-to-treat estimator ($\beta_1 \alpha_1$), was found to be 92 g, whereas the difference in mean cigarettes smoked per day was -6.4. This leads to an estimated effect $\beta_2 = \frac{92}{-6.4} = -15$, meaning an increase of 15 g in birth weight for every cigarette per day smoked less. The assumption that the encouragement to stop smoking (Z) does not affect birth weight (Y) other than through smoking behavior seems plausible. Also the assumption that there is no woman who did not stop smoking because she was encouraged to stop, is probably fulfilled.

Another example of the use of an instrumental variable can be found in the study of McClellan *et al.*, where the effect of cardiac catheterization on mortality was assessed.[48] The difference in distance between their home and the nearest hospital that performed cardiac catheterizations and the nearest hospital that did not perform this procedure, was used as an

instrumental variable. Patients with a relatively small difference in distance to both types of hospitals (< 2.5 miles) did not differ from patients with a larger difference in distance to both types of hospitals (≥ 2.5 miles) with regard to observed characteristics such as age, gender, and comorbidity; however, patients who lived relatively closer to a hospital that performed cardiac catheterizations more often received this treatment (26%) compared to patients who lived farther away (20%). Thus, the differential distance affected the probability of receiving cardiac catheterization, whereas it could reasonably be assumed that differential distance did not directly affect mortality.

As stated above, the main limitation of instrumental variables estimation is that it is based on the assumption that the instrumental variable only affects outcome by being a predictor for the treatment assignment and no direct predictor for the outcome (exclusion restriction). This assumption is difficult to fulfill; more important, it is practically untestable. Another limitation is that the treatment effect may not be generalizable to the population of patients whose treatment status was not determined by the instrumental variable. This problem is similar to that seen with RCTs, where estimated treatment effects may not be generalizable to a broader population. Finally, when variation in the likelihood of receiving a particular therapy is small between groups of patients based on an instrumental variable, differences in outcome due to this differential use of the treatment may be very small and, hence, difficult to assess.

SIMULTANEOUS EQUATIONS AND TWO-STAGE LEAST SQUARES

The method just described as instrumental variables is in fact a simple example of the more general methods of *simultaneous equations estimation*, widely used in economics and econometrics. When there are only two simultaneous equations and regression analysis is used this method is also known as *two-stage least squares* (TSLS).[49] In the first stage treatment D is explained by one or more variables that do not directly influence the outcome variable Y. In the second stage this outcome is explained by the predicted probability of receiving a particular treatment, which is adjusted for measured and unmeasured covariates. An example of this method is used to assess the effects of parental drinking on the behavioral health of children.[50] Parental drinking (the treatment) is not randomized, probably associated with unmeasured factors (e.g., parental skills) and estimated in the first stage by exogenous or instrumental variables that explain and constrain parents drinking behavior (e.g., price, number of relatives drinking).

Because the method of simultaneous equations and two-stage least squares covers the technique of instrumental variables, the same assumptions and limitations can be mentioned here. We have chosen to elaborate the instrumental variables approach, because in the medical literature these type of methods are more known under that name.

ECOLOGIC STUDIES AND GROUPED-TREATMENT EFFECTS

Ample warning can be found in the literature against the use of *ecologic studies* to describe relationships on the individual level (the ecologic fallacy); a correlation found at the aggre-

gated level (e.g., hospital) cannot be interpreted as a correlation at the patient level. Wen and Kramer, however, proposed the use of ecologic studies as a method to deal with confounding at the individual level when intended treatment effects have to be estimated.[51] In situations where considerable variation in the utilization of treatments exists across geographic areas independent of the severity of disease but mainly driven by practice style, the "relative immunity from confounding by indication may outweigh the *ecologic fallacy*" by performing an ecologic study.[51] Of course, such ecologic studies have low statistical power by the reduced number of experimental units and tell us little about the individuals in the compared groups. Moreover, Naylor argues that the limitations of the proposed technique in order to remove confounding by indication are too severe to consider an aggregated analysis as a serious alternative when individual level data are available.[52]

An alternative method described in the literature is known as the *grouped-treatment approach*. Keeping the analysis at the individual level, the individual treatment variable will be replaced by an ecological or grouped-treatment variable, indicating the percentage of treated persons at the aggregated level.[53] With this method the relative immunity for confounding by indication by an aggregated analysis is combined with the advantage of correcting for variation at the individual level. In fact this method is covered by the method of *two-stage least squares*, where in the first stage more variables are allowed to assess the probability of receiving the treatment. This method faces the same assumptions as the instrumental variables approach discussed earlier. Most important is the assumption that unmeasured variables do not produce an association between prognosis and the grouped-treatment variable, which in practice will be hard to satisfy.

VALIDATIONS AND SENSITIVITY ANALYSES

Horwitz *et al*.[54] proposed to validate observational studies by constructing a cohort of subjects in clinical practice that is restricted by the inclusion criteria of RCTs. Similarity in estimated treatment effects from the observational studies and the RCTs would provide empirical evidence for the validity of the observational method. Although this may be correct in specific situations,[17,55] it does not provide evidence for the validity of observational methods for the evaluation of treatments in general.[8]

To answer the question whether observational studies produce similar estimates of treatment effects compared to randomized studies, several authors have compared the results of randomized and nonrandomized studies for a number of conditions, sometimes based on meta-analyses.[56–58] In general, these reviews have concluded that the direction of treatment effects assessed in nonrandomized studies is often, but not always, similar to the direction of the treatment effects in randomized studies, but that differences between nonrandomized and randomized studies in the estimated magnitude of treatment effect are very common. Trials may under-

or overestimate the actual treatment effect, and the same is true for nonrandomized comparison of treatments. Therefore, these comparisons should not be interpreted as true validations.

A *sensitivity analysis* can be a valuable tool in assessing the possible influence of an unmeasured confounder. This method was probably first used by Cornfield *et al*.[59] when they attacked Fisher's[60] hypothesis that the apparent association between smoking and lung cancer could be explained by an unmeasured genetic confounder related to both smoking and lung cancer. The problem of nonrandomized assignment to treatments in observational studies can be thought of as a problem of unmeasured confounding factors. Instead of stating that an unmeasured confounder can explain the treatment effect found, sensitivity analyses try to find a lower bound for the magnitude of association between that confounder and the treatment variable. Lin *et al*. developed a general approach for assessing the sensitivity of the treatment effect to the confounding effects of unmeasured confounders after adjusting for measured covariates, assuming that the true treatment effect can be represented in a regression model.[61] The plausibility of the estimated treatment effects will increase if the estimated treatment effects are insensitive over a wide range of plausible assumptions about these unmeasured confounders.

SUMMARY AND DISCUSSION

Although randomized clinical trials remain the gold standard in the assessment of intended effects of drugs, observational studies may provide important information on effectiveness under everyday circumstances and in subgroups not previously studied in RCTs. The main defect in these studies is the incomparability of groups, giving a possible alternative explanation for any treatment effect found. Thus, focus in such studies is directed toward adjustment for confounding effects of covariates.

Along with standard methods of *appropriate selection of reference groups, stratification* and *matching*, we discussed multivariable statistical methods such as *(logistic) regression* and *Cox proportional hazards regression* to correct for confounding. In these models, the covariates, added to a model with 'treatment' as the only explanation, give alternative explanations for the variation in outcome, resulting in a corrected treatment effect. In fact, the main problem of balancing the treatment and control groups according to some covariates has been avoided. A method that more directly attacks the problem of imbalance between treatment and control group, is the method of *propensity scores*. By trying to explain this imbalance with measured covariates, a score is computed which can be used as a single variable to match both groups. Alternatively, this score can be used as a stratification variable or as a single covariate in a regression model.

In all these techniques, an important limitation is that adjustment can only be achieved for *measured* covariates, implicating possible measurement error on these covariates (e.g., the severity of a past disease) and possible omission of other important, unmeasured covariates. A method not limited by these shortcomings is a technique known as *instrumental variables*. In

this approach, the focus is on finding a variable (the instrument) that is related to the allocation of treatments, but is related to outcome only because of its relation to treatment. This technique can achieve the same effect as randomization in bypassing the usual way in which physicians allocate treatment according to prognosis, but its rather strong assumptions limit its use in practice. Related techniques are *two-stage least squares* and the *grouped-treatment approach*, sharing the same limitations. All these methods are summarized in Table 2.1.

Table 2.1: Strengths and limitations of methods to assess treatment effects in nonrandomized, observational studies

Method	Used	Strengths	Limitations
Design approaches			
Historical controls	Infrequently	• Easy to identify comparison group	• Treatment effect often biased
Candidates for treatment	Infrequently	• Useful for preliminary selection	• Difficult to identify not treated candidates
Treatments for the same indication	Infrequently, when possible	• Similarity of prognostic factors	• Only useful for diseases treated with several drugs • Only effectiveness of one drug compared to another
Case-crossover and case-time-control designs	Infrequently	• Reduced variability by intersubject comparison	• Only useful to assess time-limited effects • Possible crossover effects
Data-analytical approaches			
Stratification and (weighted) matching	Frequently	• Clear interpretation/no assumptions • Clarity of incomparability on used covariates	• Only a few covariates or rough categories can be used
Asymmetric stratification	Not used	• More covariates than with normal stratification	• Still limited number of covariates
Common statistical methods: linear regression, logistic regression, survival analysis	Standard, very often	• More covariates than matching or stratification • Easy to perform	• Focus is not on balancing groups • Adequate overlap between groups difficult to assess
Propensity scores	More often	• Many covariates possible	• Performs better with only a few number of events per confounder
Multivariate confounder score	Scarcely	• Less insensitive to misspecification	• Exaggerates significance
Ecologic studies	Scarcely	• Immune to confounding by indication	• Loss of power • Loss of information at the individual level
Instrumental variables (IV), two-stage least squares; grouped-treatment effects	Infrequently	• Large differences per area are needed	• Difficult to identify an IV • IV must be unrelated to factors directly affecting outcome

Given the limitations of observational studies, the evidence in assessing intended drug effects from observational studies will be in general less convincing than from well conducted RCTs. The same of course is true when RCTs are *not* well conducted (e.g., lacking double blinding or exclusions after randomization). This means that due to differences in quality, size or other characteristics disagreement among RCTs is not uncommon.[62,63] In general we subscribe to the view that observational studies including appropriate adjustments are less suited to assess new intended drug effects (unless the expected effect is very large), but can certainly be valuable for assessing the long-term beneficial effects of drugs already proven effective in short-term RCTs. For instance, the RCTs of acetylsalicylic acid that demonstrated the bene-

ficial effects in the secondary prevention of coronary heart disease were of limited duration, but these drugs are advised to be taken lifelong. Another purpose of observational studies is to investigate the causes of interindividual variability in drug response. Most causes of variability in drug response are unknown. Observational studies can also be used to assess the intended effects of drugs in patients that were excluded from RCTs (e.g., very young patients, or patients with different comorbidities and polypharmacy), or in patients that were studied in RCTs but who might still respond differently (e.g., because of genetic differences).

Comparison between the presented methods to assess adjusted treatment effects in observational studies is mainly based on theoretical considerations, although some empirical evidence is available. A more complete empirical evaluation that compares the different adjustment methods with respect to the estimated treatment effects under several conditions will be needed to assess the validity of the different methods. Preference for one method or the other can be expressed in terms of bias, precision, power, and coverage probability of the methods, whereas the different conditions can be defined by means of, for instance, the severity of the disease, the number of covariates, the strength of association between covariates and outcome, the association among the covariates, and the amount of overlap between the groups. These empirical evaluations can be performed with existing databases or computer simulations. Given the lack of empirical evaluations for comparisons of the different methods and the importance of the assessment of treatment effects in observational studies, more effort should be directed toward these evaluations.

REFERENCES

[1] Friedman LM, Furberg CD, DeMets DL. *Fundamentals of Clinical Trials*. St Louis: Mosby-Year Book, 1996.

[2] Chalmers I. Why transition from alternation to randomisation in clinical trials was made [Letter]. *BMJ*, 319:1372, 1999.

[3] Schulz KF, Grimes DA. Allocation concealment in randomised trials: defending against deciphering. *Lancet*, 359:614–618, 2002.

[4] Urbach P. The value of randomization and control in clinical trials. *Stat Med*, 12:1421–1431, 1993. discussion 143341.

[5] Feinstein AR. Current problems and future challenges in randomized clinical trials. *Circulation*, 70:767–774, 1984.

[6] Gurwitz JH, Col NF, Avorn J. The exclusion of the elderly and women from clinical trials in acute myocardial infarction. *JAMA*, 268:1417–1422, 1992.

[7] Wieringa NF, de Graeff PA, van der Werf GT, Vos R. Cardiovascular drugs: discrepancies in demographics between pre- and post-registration use. *Eur J Clin Pharmacol*, 55:537–544, 1999.

[8] MacMahon S, Collins R. Reliable assessment of the effects of treatment on mortality and major morbidity, II: observational studies. *Lancet*, 357:455–462, 2001.

[9] McKee M, Britton A, Black N, McPherson K, Sanderson C, Bain C. Methods in health services research. interpreting the evidence: choosing between randomised and non-randomised studies. *BMJ*, 319:312–315, 1999.

[10] Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med*, 342:1887–1892, 2000.

[11] Grodstein F, Clarkson TB, Manson JE. Understanding the divergent data on postmenopausal hormone therapy. *N Engl J Med*, 348:645–650, 2003.

[12] Beral V, Banks E, Reeves G. Evidence from randomised trials on the long-term effects of hormone replacement therapy. *Lancet*, 360:942–944, 2002.

[13] Messerli FH. Case-control study, meta-analysis, and bouillabaisse: putting the calcium antagonist scare into context [Editorial]. *Ann Intern Med*, 123:888–889, 1995.

[14] Grobbee DE, Hoes AW. Confounding and indication for treatment in evaluation of drug treatment for hypertension. *BMJ*, 315:1151–1154, 1997.

[15] Rosenbaum PR. *Observational studies, 2nd edition*. Springer-Verlag, New York, 2002.

[16] Sacks H, Chalmers TC, Smith H Jr. Randomized versus historical controls for clinical trials. *Am J Med*, 72:233–240, 1982.

[17] Kalra L, Yu G, Perez I, Lakhani A, Donaldson N. Prospective cohort study to determine if trial efficacy of anticoagulation for stroke prevention in atrial fibrillation translates into clinical effectiveness. *BMJ*, 320:1236–1239, 2000.

[18] Ioannidis JP, Polycarpou A, Ntais C, Pavlidis N. Randomised trials comparing chemotherapy regimens for advanced non-small cell lung cancer: biases and evolution over time. *Eur J Cancer*, 39:2278–2287, 2003.

[19] Klungel OH, Stricker BH, Breteler MM, Seidell JC, Psaty BM, de Boer A. Is drug treatment of hypertension in clinical practice as effective as in randomized controlled trials with regard to the reduction of the incidence of stroke? *Epidemiology*, 12:339–344, 2001.

[20] Johnston SC. Identifying confounding by indication through blinded prospective review. *Am J Epidemiol*, 154:276–284, 2001.

[21] Psaty BM, Heckbert SR, Koepsell TD, Siscovick DS, Raghunathan TE, Weiss NS, Rosendaal FR, Lemaitre RN, Smith NL, Wahl PW. The risk of myocardial infarction associated with antihypertensive drug therapies. *JAMA*, 274:620–625, 1995.

[22] Klungel OH, Heckbert SR, Longstreth WT Jr, Furberg CD, Kaplan RC, Smith NL, Lemaitre RN, Leufkens HG, de Boer A, Psaty BM. Antihypertensive drug therapies and the risk of ischemic stroke. *Arch Intern Med*, 161:37–43, 2001.

[23] Abi-Said D, Annegers JF, Combs-Cantrell D, Suki R, Frankowski RF, Willmore LJ. A case-control evaluation of treatment efficacy: the example of magnesium sulfate prophylaxis against eclampsia in patients with preeclampsia. *J Clin Epidemiol*, 50:419–423, 1997.

[24] Concato J, Peduzzi P, Kamina A, Horwitz RI. A nested case-control study of the effectiveness of screening for prostate cancer: research design. *J Clin Epidemiol*, 54:558–564, 2001.

[25] Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. *Am J Epidemiol*, 133:144–153, 1991.

[26] Greenland S. Confounding and exposure trends in case-crossover and case-time-control designs. *Epidemiology*, 7:231–239, 1996.

[27] Suissa S. The case-time-control design. *Epidemiology*, 6:248–253, 1995.

[28] Suissa S. The case-time-control design: further assumptions and conditions. *Epidemiology*, 9:441–445, 1998.

[29] Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24:295–313, 1968.

[30] Rubin DB. Estimating causal effects from large data sets using the propensity score. *Ann Intern Med*, 127:757–763, 1997.

[31] Cook EF, Goldman L. Asymmetric stratification: an outline for an efficient method for controlling confounding in cohort studies. *Am J Epidemiol*, 127:626–639, 1988.

[32] Psaty BM, Koepsell TD, Lin D, *et al*. Assessment and control for confounding by indication in observational studies. *J Am Geriatr Soc*, 47:749–754, 1999.

[33] Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *J Clin Epidemiol*, 48:1503–1510, 1995.

[34] Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*, 49:1373–1379, 1996.

[35] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

[36] D'Agostino, RB Jr. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med*, 17:2265–2281, 1998.

[37] Stenestrand U, Wallentin L. Early statin treatment following acute myocardial infarction and 1-year survival. *JAMA*, 285:430–436, 2001.

[38] Aronow HD, Topol EJ, Roe MT, Houghtaling PL, Wolski KE, Lincoff AM, Harrington RA, Califf RM, Ohman EM, Kleiman NS, Keltai M, Wilcox RG, Vahanian A, Armstrong PW, Lauer MS. Effect of lipid-lowering therapy on early mortality after acute coronary syndromes: an observational study. *Lancet*, 357:1063–1068, 2001.

[39] Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am Stat*, 39:33–38, 1985.

[40] Ming K, Rosenbaum PR. Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics*, 56:118–124, 2000.

[41] Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, 49:1231–1236, 1993.

[42] Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol*, 158:280–287, 2003.

[43] Miettinen OS. Stratification by a multivariate confounder score. *Am J Epidemiol*, 104:609–620, 1976.

[44] Pike MC, Anderson J, Day N. Some insights into miettinen's multivariate confounder score approach to case-control study analysis. *Epidemiol Community Health*, 33:104–106, 1979.

[45] Newhouse JP, McClellan M. Econometrics in outcomes research: the use of instrumental variables. *Ann Rev Public Health*, 19:17–34, 1998.

[46] Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *JASA*, 91:444–455, 1996.

[47] Permutt Th, Hebel JR. Simultaneous-equation estimation in a clinical trial of the effect of smoking on birth weight. *Biometrics*, 45:619–622, 1989.

[48] McClellan M, McNeil BJ, Newhouse JP. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *JAMA*, 272:859–866, 1994.

[49] Angrist JD, Imbens GW. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *JASA*, 90:431–442, 1995.

[50] Snow Jones A, Miller DJ, Salkever DS. Parental use of alcohol and children's behavioural health: a household production analysis. *Health Econ*, 8:661–683, 1999.

[51] Wen SW, Kramer MS. Uses of ecologic studies in the assessment of intended treatment effects. *J Clin Epidemiol*, 52:7–12, 1999.

[52] Naylor CD. Ecological analysis of intended treatment effects: caveat emptor. *J Clin Epidemiol*, 52:1–5, 1999.

[53] Johnston SC, Henneman T, McCulloch CE, van der Laan M. Modeling treatment effects on binary outcomes with grouped-treatment variables and individual covariates. *Am J Epidemiol*, 156:753–760, 2002.

[54] Horwitz RI, Viscoli CM, Clemens JD, Sadock RT. Developing improved observational methods for evaluating therapeutic effectiveness. *Am J Med*, 89:630–638, 1990.

[55] Hlatky MA, Califf RM, Harrell FE Jr, Lee KL, Mark DB, Pryor DB. Comparison of predictions based on observational data with the results of randomized controlled clinical trials of coronary artery bypass surgery. *J Am Coll Cardiol*, 11:237–245, 1988.

[56] Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med*, 342:1878–1886, 2000.

[57] Ioannidis JP, Haidich AB, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, Contopoulos-Ioannidis DG, Lau J. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA*, 286:821–830, 2001.

[58] Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ*, 317:1185–1190, 1998.

[59] Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL. Smoking and lung cancer: recent evidence and a discussion of some questions. *J Natl Cancer Inst*, 22:173–203, 1959.

[60] Fisher RA. Lung cancer and cigarettes? *Nature*, 182:108, 1958.

[61] Lin DY, Psaty BM, Kronmal RA. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*, 54:948–963, 1998.

[62] LeLorier J, Gregoire G, Benhaddad A, Lapierre J, Derderian F. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *N Engl J Med*, 337:536–542, 1997.

[63] Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA*, 273:408–412, 1995.

CHAPTER 3

STRENGTHS AND LIMITATIONS OF ADJUSTMENT METHODS

3.1 "Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: A Monte Carlo study"

Edwin P. Martens[a,b], Wiebe R. Pestman[b] and Olaf H. Klungel[a]

[a] *Department of Pharmacoepidemiology and Pharmacotherapy, Utrecht Institute of Pharmaceutical Sciences (UIPS), Utrecht University, Utrecht, the Netherlands*
[b] *Centre for Biostatistics, Utrecht University, Utrecht, the Netherlands*

Statistics in Medicine 2007, 26;16:3208-3210

Letter to the editor as reaction on: Austin PC, Grootendorst P, Sharon-Lise TN, Anderson GM. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: A Monte Carlo study. *Statistics in Medicine* 2007, 26;4:754-768, 2007

ABSTRACT

In medical research logistic regression and Cox proportional hazards regression analysis, in which all the confounders are included as covariates, are often used to estimate an adjusted treatment effect in observational studies. In the last decade the method of propensity scores has been developed as an alternative adjustment method and many examples of applications can be found in the literature. Frequently this analysis is used as a comparison for the results found by the logistic regression or Cox proportional hazards regression analysis, but researchers are insufficiently aware of the different types of treatment effects that are estimated by these analyses.

This is emphasized by a recent simulation study by Austin *et al.* in which the main objective was to investigate the ability of propensity score methods to estimate conditional treatment effects as estimated by logistic regression analysis. Propensity score methods are in general incapable of estimating conditional effects, because their aim is to estimate marginal effects like in randomized studies. Although the conclusion of the authors is correct, it can be easily misinterpreted. We argue that in treatment effect studies most researchers are interested in the marginal treatment effect and the many possible conditional effects in logistic regression analysis can be a serious overestimation of this marginal effect.

For studies in which the outcome variable is dichotomous we conclude that the treatment effect estimate from propensity scores is in general closer to the treatment effect that is of most interest in treatment effect studies.

Keywords: Confounding; Propensity scores; Logistic regression analysis; Marginal treatment effect; Conditional treatment effect; Average treatment effect

In a recent simulation study Austin *et al.* conclude that conditioning on the propensity score gives biased estimates of the true conditional odds ratio of treatment effect in logistic regression analysis. Although we generally agree with this conclusion, it can be easily misinterpreted because of the word bias. From the same study one can similarly conclude that logistic regression analysis will give a biased estimate of the treatment effect that is estimated in a propensity score analysis. Because propensity score methods aim at estimating a marginal treatment effect, we believe that the last statement is more meaningful.

DIFFERENT TREATMENT EFFECTS

The authors raise an important issue, which is probably unknown to many researchers, that in logistic regression analysis a summary measure of conditional treatment effects will in general not be equal to the marginal treatment effect. This phenomenon is also known as non-collapsibility of the odds ratio,[1] but is apparent in all non-linear regression models and generalized linear models with a link function other than the identity link (linear models) or log-link function.[2] In other words, even if a prognostic factor is equally spread over treatment groups, the inclusion of this variable in a logistic regression model will increase the estimated treatment effect. This increasing effect of a conditional treatment effect compared to the overall marginal effect is larger when more prognostic factors are added, but lower when the treatment effect is closer to OR=1 and also lower when the incidence rate of the outcome is smaller.[3] In general, it can be concluded that in a given research situation many different conditional treatment effects exist, depending on the number of prognostic factors in the model.

TRUE CONDITIONAL TREATMENT EFFECT

The true treatment effect is the effect on a specific outcome of treating a certain population compared to not treating this population. In randomized studies this can be estimated as the effect of the treated group compared to the non-treated group. The true conditional treatment effect as defined in Austin *et al.* is the treatment effect in a certain population given the set of six prognostic factors and given that the relationships in the population can be captured by a logistic regression model. Two of the six prognostic factors were equally distributed between treatment groups and included in the equation for generating the data. But there are also non-confounding prognostic factors excluded from this equation, because not all of the variation in the outcome is captured by the six prognostic factors. That means that it seems to be at least arbitrary how many and which of the non-confounding prognostic factors were included or excluded to come to a 'true conditional treatment effect'. Because of the non-collapsibility of the odds ratio, all these conditional treatment effects are in general different from each other, but which of these is the one of interest remains unclear. The only thing that is clear,

is that application of the model that was used to generate the data will find on average this 'true conditional treatment effect', while all other models, including less or more prognostic factors, will in general find a 'biased' treatment effect. It should be therefore no surprise that propensity score models will produce on average attenuated treatment effects, for propensity score models correct for only one prognostic factor, the propensity score. This implies that the treatment effect estimates from propensity score models are in principal closer to the overall marginal treatment effect than to one of the many possible conditional treatment effects.

MARGINAL OR CONDITIONAL TREATMENT EFFECTS?

The authors give two motivations why a conditional treatment effect is more interesting than the overall marginal treatment effect (which is the effect that would be found if treatments were randomized). Firstly, they indicate that a conditional treatment effect is more interesting to physicians, because it allows physicians to make appropriate treatment effect decisions for specific patients. Indeed, in clinical practice treatment decisions are made for individual patients, but these decisions are better informed by subgroup analyses with specific treatment effects for subgroups: a specific conditional treatment effect is still some kind of 'average' over all treatment effects in subgroups. Another argument is that treatment decisions on individual patients should be based on the absolute risk reduction and not on odds ratios or relative risks.[4] Secondly, the authors suggest that in practice researchers use propensity scores for estimating conditional treatment effects. However, in most studies in which propensity scores and logistic regression analysis are both performed, researchers rather have an overall marginal treatment effect in mind than one specific conditional treatment effect.[5] Furthermore, the overall marginal treatment effect is one well-defined treatment effect, whereas conditional treatment effects are effects that are dependent on the chosen model. The reason for comparing propensity score methods with logistic regression analysis is probably not because the aim is to estimate conditional effects, but simply because logistic regression is the standard way of estimating an adjusted treatment effect with a dichotomous outcome.

In conclusion, propensity score methods aim to estimate a marginal effect, which in general is not a good estimate of a conditional effect in logistic regression analysis because of the non-collapsibility of the odds ratio. An overall marginal treatment effect is better defined and seems to be of more interest than all possible conditional treatment effects. Finally, these conditional effects are dependent on the number of non-confounders, which is not the case for propensity score methods.

REFERENCES

[1] Greenland S, Robins MR, Pearl J. Confounding and collapsibility in causal inference. *Stat Science*, 14:29–46, 1999.

[2] Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, 71:431444, 1984.

[3] Rosenbaum PR. *Propensity score. In: Armitage P, Colton T, eds. Encyclopedia of biostatistics*. Wiley, Chichester, United Kingdom, 1998.

[4] Rothwell PM, Mehta Z, Howard SC, Gutnikov SA, Warlow CP. From subgroups to individuals: general principles and the example of carotid endarterectomy. *The Lancet*, 365 (9455):256–265, 2005.

[5] Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *J Clin Epidemiol*, 58:550–559, 2005.

3.2 AN IMPORTANT ADVANTAGE OF PROPENSITY SCORE METHODS COMPARED TO LOGISTIC REGRESSION ANALYSIS

Edwin P. Martens[a,b], Wiebe R. Pestman[b], Anthonius de Boer[a], Svetlana V. Belitser[a] and Olaf H. Klungel[a]

[a] *Department of Pharmacoepidemiology and Pharmacotherapy, Utrecht Institute of Pharmaceutical Sciences (UIPS), Utrecht University, Utrecht, the Netherlands*
[b] *Centre for Biostatistics, Utrecht University, Utrecht, the Netherlands*

Provisionally accepted by International Journal of Epidemiology

ABSTRACT

In medical research propensity score (PS) methods are used to estimate a treatment effect in observational studies. Although advantages for these methods are frequently mentioned in the literature, it has been concluded from literature studies that treatment effect estimates are similar when compared with multivariable logistic regression (LReg) or Cox proportional hazards regression. In this study we demonstrate that the difference in treatment effect estimates between LReg and PS methods is systematic and can be substantial, especially when the number of prognostic factors is more than 5, the treatment effect is larger than an odds ratio of 1.25 (or smaller than 0.8) or the incidence proportion is between 0.05 and 0.95. We conclude that PS methods in general result in treatment effect estimates that are closer to the true average treatment effect than a logistic regression model in which all confounders are modeled. This is an important advantage of PS methods that has been frequently overlooked by analysts in the literature.

Keywords: Confounding; Propensity scores; Logistic regression analysis; Marginal treatment effect; Conditional treatment effect; Average treatment effect

INTRODUCTION

A commonly used statistical method in observational studies that adjusts for confounding, is the method of propensity scores (PS).[1,2] This method focusses on the balance of covariates between treatment groups before relating treatment to outcome. In contrast, classical methods like linear regression, logistic regression (LReg) or Cox proportional hazards regression (Cox PH) directly relate outcome to treatment and covariates by a multivariable model. Advantages to use PS methods that are frequently mentioned in the literature are the ability to include more confounders, the better adjustment for confounding when the number of events is low and the availability of information on the overlap of covariate distributions.[1–7] In two recent literature studies it is concluded that treatment effects estimated by both PS methods and regression techniques are in general fairly similar to each other.[8,9] Instead of a focus on the similarity in treatment effects between both methods, we will illustrate that the differences between PS methods and LReg analysis are systematic and can be substantial. We will also demonstrate that treatment effect estimates from PS methods are in general closer to the true average treatment effect than from LReg, which results in an important advantage of PS methods over LReg.

SYSTEMATIC DIFFERENCES BETWEEN TREATMENT EFFECT ESTIMATES

In the literature review of Shah *et al.* the main conclusion was that propensity score methods resulted in similar treatment effects compared to traditional regression modeling.[8] This was based on the agreement that existed between the significance of treatment effect in PS methods compared to LReg or Cox PH methods in 78 reported analyses. This agreement was denoted as excellent ($\kappa = 0.79$) and the mean difference in treatment effect was quantified as 6.4%. In the review of Stürmer *et al.* it was also stressed that PS methods did not result in substantially different treatment effect estimates compared to LReg or Cox PH methods.[9] They reported that in only 9 out of 69 studies (13%) the effect estimate differed by more than 20%.

The results of these reviews can also be interpreted differently: the dissimilarity between methods is systematic resulting in treatment effect estimates that are on average stronger in LReg and Cox PH analysis. In Shah *et al.* the disagreement between methods was in the same direction: all 8 studies that disagreed resulted in a significant effect in LReg or Cox PH methods and a non-significant effect in PS methods ($p = 0.008$, McNemar's test). Similarly, the treatment effect in PS methods was more often closer to unity than in LReg or Cox PH (34 versus 15 times, $p = 0.009$, binomial test with $\pi_0 = 0.5$). In the review of Stürmer *et al.* it turned out that substantial differences between both methods only existed when the estimates in LReg or Cox PH were *larger* than in PS methods. Because both reviews were partly based on the same studies, we summarized the results in Table 3.1 by taking into account studies that

were mentioned in both reviews. We included all studies that reported treatment effects for PS methods (matching, stratification or covariate adjustment) and regression methods (LReg or Cox PH), even when the information was that effects were 'similar'.

Table 3.1: Comparison of treatment effect estimates between propensity score methods (PS) and logistic regression (LReg) or Cox proportional hazards regression (Cox PH)[8,9]

	number of studies	percentage
Treatment effect is stronger in PS methods	24	25.0%
Treatment effects are equal or reported as 'similar'	22	22.9%
Treatment effect is stronger in LReg or Cox PH	50	52.1%

From all 96 studies (Table 3.1) there were twice as many studies in which the treatment effect from LReg or Cox PH methods was stronger than from PS methods: 50 versus 24 ($= 68\%$). Testing the null hypothesis of equal proportions (binomial test, $\pi_0 = 0.5$, leaving out the category when effects were reported to be equal or similar) resulted in highly significant differences ($p = 0.003$). The mean difference in the logarithm of treatment effects (δ)[8] between both methods was calculated at 5.0%, significantly different from 0 ($p = 0.001$, 95% confidence interval (CI): 2.0, 7.9). In studies with treatment effects larger than an odds ratio (OR) of 2.0 or smaller than 0.5 this mean difference was even larger: $\delta = 19.0\%$, 95% CI: 10.3, 27.6.

We conclude that PS methods result in treatment effects that are significantly closer to the null hypothesis of no effect than LReg or Cox PH methods. The larger the treatment effects, the larger the differences.

EXPLAINING DIFFERENCES IN TREATMENT EFFECT ESTIMATES

The reason for the systematic differences between treatment effect estimates from PS methods and LReg or Cox PH methods can be found in the *non-collapsibility* of the odds ratio and hazard ratio used as treatment effect estimators. In the literature this phenomenon has been recognized and described by many authors.[10–18] To understand this, we start by defining a *true average treatment effect* as the effect of treating a certain population instead of not treating a *similar* population, where similarity is defined in terms of prognostic factors. In general, this is the treatment effect in which we are primarily interested and equals the average effect in randomized studies. Note that this treatment effect is defined without using any outcome model with covariates. When treated and untreated populations are similar on prognostic factors, this true average treatment effect can be simply estimated by an *unadjusted treatment effect*, for instance a difference in means, a risk ratio or an odds ratio. When on the other hand both populations are not similar on prognostic factors, as is to be expected in observational studies, one should estimate an *adjusted treatment effect*, trying to correct for all potential confounders. This can be done for instance by any multivariable regression model or by PS methods using stratification, matching or covariate adjustment. When treated and untreated populations are

exactly similar on all covariates, unadjusted and adjusted treatment effects should coincide, because the primary objective of adjustment is to adjust for dissimilarities in covariate distributions: if there are none, ideally adjustment should have no effect. Unfortunately, this is not generally true, for instance when odds ratios from LReg analysis are used to quantify treatment effects. Consider two LReg models:

$$\text{logit}(y) = \alpha_1 + \beta_t t \tag{3.1}$$

$$\text{logit}(y) = \alpha_2 + \beta_t^* t + \beta_1 x_1 \tag{3.2}$$

where y is a dichotomous outcome, t a dichotomous treatment, x_1 a dichotomous prognostic factor and α_1 and α_2 constants, e^{β_t} the unadjusted treatment effect, $e^{\beta_t^*}$ the adjusted treatment effect and e^{β_1} the effect of x_1.

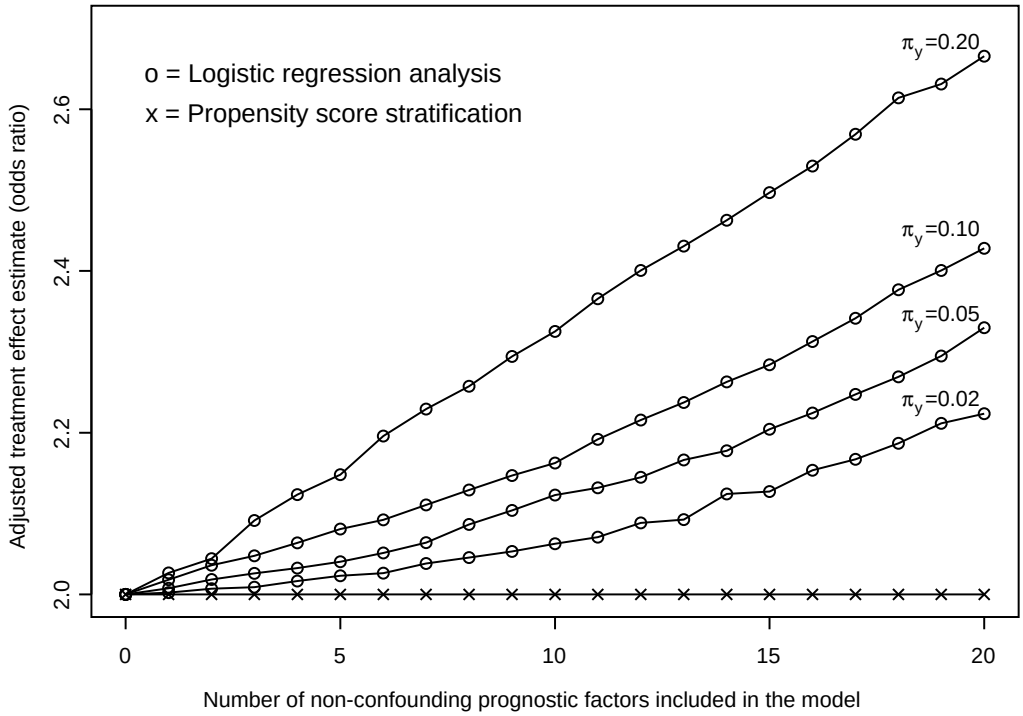
Suppose that in a certain situation only one prognostic factor exists (x_1) with a different distribution for both treatment groups. An adjusted treatment effect β_t^* will be interpreted as an estimate for the true average treatment effect, i.e. the effect that would be found when both treatment groups had *similar* distributions of x_1. But when in reality the distribution of x_1 is similar for both treatment groups and model 3.2 is applied, it turns out that the adjusted treatment effect estimate β_t^* does not equal the unadjusted treatment effect β_t. More generally, when both treatment groups are similar with respect to their covariate distributions, the adjusted and unadjusted treatment effects will not coincide in non-linear regression models or generalized linear models with another link function than the identity link (equalling a linear regression analysis) or log-link. We refer to the literature for a mathematical explanation of this phenomenon[10, 11, 19] and will illustrate in the next paragraph its implications for the comparison between LReg and PS methods in epidemiological research.

ADJUSTING FOR EQUALLY DISTRIBUTED PROGNOSTIC FACTORS

To illustrate the non-collapsibility of the OR, we created a large population of $n = 100,000$, a binary outcome y (π_y varying from 0.02 to 0.20), a treatment t ($\pi_t = 0.50$) and 20 binary prognostic factors x_1, ..., x_{20} with π_{x1}= ... =$\pi_{x20} = 0.50$ and $e^{\beta_{x1}}$= ... =$e^{\beta_{x20}} = 2.0$. These factors, which we will call *non-confounders*, were exactly equally distributed across treatments $t = 1$ and $t = 0$. The true average treatment effect is therefore known and equals the unadjusted effect of treatment on outcome e^{β_t} in equation 3.1, which was set to 2.0. First we included the factor x_1 in the LReg model of equation 3.2 and calculated an adjusted treatment effect $e^{\beta_t^*}$. We extended this model by including the factors x_2 to x_{20} and calculated the corresponding adjusted treatment effects. In Figure 3.1 all these adjusted treatment effects were plotted for various incidence proportions. For example, with an incidence proportion of $\pi_y = 0.10$ the adjusted treatment effect is estimated as nearly 2.16 in a LReg model with 10

non-confounders and as 2.43 in a model with 20 non-confounders. Its increase is stronger when the incidence proportion is higher. Also an increase in the strength of the treatment effect (here fixed at 2.0) or an increase in the strength of the association between non-confounders and outcome (also fixed at 2.0) will increase the difference between adjusted and unadjusted treatment effect estimates (data not shown).[20]

Figure 3.1: Adjusted treatment effects for 1 to 20 non-confounding prognostic factors and various incidence proportions in logistic regression and propensity score stratification ($n = 100,000$, $e^{\beta_t} = 2.0$)



This is in sharp contrast with PS methods for which treatment effects remain unchanged, irrespective of the number of covariates in the PS model, the incidence proportion, the strength of the treatment effect and the strength of the association between non-confounders and outcome. The reason is that all prognostic factors are equally distributed between treatment groups (univariate as well as multivariate), which means that the calculated propensity score is constant for every individual. Stratification on the PS or including it as a covariate will leave the unadjusted treatment effect unchanged. Although it seems obvious, it illustrates an important advantage of PS methods compared to LReg: PS methods leave the unadjusted treatment effect unchanged when prognostic factors are equally distributed between treatment groups. In contrast, this is not true for LReg analysis.

ADJUSTING FOR IMBALANCED PROGNOSTIC FACTORS

Perfectly balanced treatment groups, as used in the previous paragraph, are quite exceptional in practice. In general, treatment groups will differ from each other with respect to covariate distributions, in observational studies (systematic and random imbalances), but also in randomized studies (random imbalances). In this paragraph we will explore the differences between LReg and PS analysis when adjustment takes place for imbalanced prognostic factors. In simulation studies it is common to create imbalance between treatment groups by first modeling treatment as a function of covariates and then outcome as a function of treatment and covariates.[5,21–23] Unfortunately, the treatment effect that is defined in such studies as the *true treatment effect* does not match the effect that is commonly of interest when treatment effect studies are performed. It is an *adjusted* treatment effect which is conditional on the covariates that has been chosen in the true model. So, in such simulation studies the true average treatment effect as defined in the third section will be unknown.[24] One solution is to calculate such a true treatment effect with an iterative procedure,[25] but still all data are based on logistic regression models, one of the methods to be evaluated. These problems can be circumvented when one starts with a balanced population with a known true treatment effect in which no outcome model is involved in generating the data. By using the imbalances on prognostic factors that appear in random samples, the effects of adjustment between LReg and PS methods can be fairly compared. Random imbalances are indistinguishable from systematic model-based imbalances at the level of an individual data set: they only differ from one another by the fact that random imbalances will cancel out when averaged over many samples. For illustrating the differences between LReg and PS methods when adjusting for imbalances it is not important *how* imbalances have arisen.

SIMULATIONS

We created a population of $n = 100,000$, a binary outcome y ($\pi_y = 0.30$), treatment t ($\pi_t = 0.50$) and 5 normally distributed prognostic factors x_1, ..., x_5 with mean$= 0.50$, standard deviation$= 0.4$ and $e^{\beta_{x1}}$= ... =$e^{\beta_{x5}} = 2.0$. The true treatment effect in the population was set to $e^{\beta_t} = 2.5$. To randomly create imbalance, we took $1,000$ random samples with varying sample sizes ($n = 200$, 400, 800 and $1,600$). The LReg model used for adjustment is:

$$\text{logit}(y) = \alpha_y + \beta_t^* t + \beta_{1y} x_1 + ... + \beta_{5y} x_5 \tag{3.3}$$

and the propensity scores are calculated as:

$$PS = \frac{e^{\text{logit}(t)}}{1 + e^{\text{logit}(t)}} \tag{3.4}$$

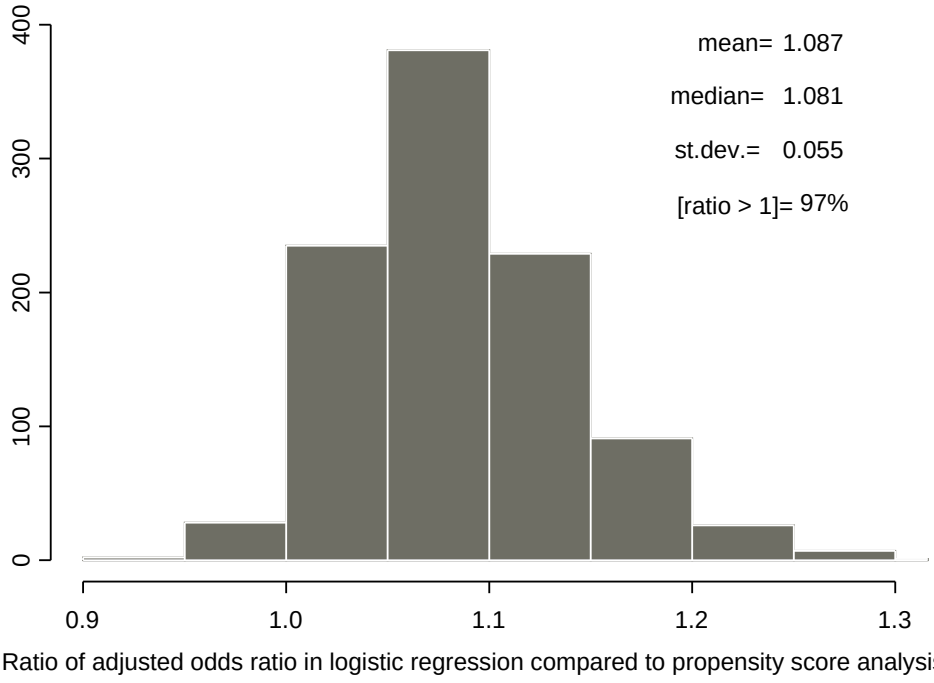
with $\text{logit}(t) = \alpha_t + \beta_{1t} x_1 + ... + \beta_{5t} x_5$.

To adjust for confounding we stratified subjects on the quintiles of the PS and calculated a common treatment effect using the Mantel-Haenszel estimator.

COMPARISON OF ADJUSTED TREATMENT EFFECTS

In Figure 3.2 it is illustrated that the adjusted odds ratios in a LReg analysis with $n = 400$ are nearly 9% larger than those in PS analysis: in nearly all samples (97%) the ratio of adjusted treatment effects from both analysis is larger than 1. This confirms the results found in the reviews and presented in Table 3.1 that LReg or Cox PH result in general higher treatment effects than PS analysis ($50/74 = 68\%$). The difference between both percentages is due to the diversity in models, treatment effects, sample sizes and number of confounders that were found in the literature.

Figure 3.2: Histogram of the ratio of adjusted odds ratios of treatment effect in logistic regression compared to propensity score analysis, 1,000 samples of $n = 400$



In Table 3.2 the results are summarized for various sample sizes. Between sample sizes of 400, 800 and $1,600$ there are only minor differences in the mean and median ratio. Overall it can be concluded that with the chosen associations and number of covariates, the adjusted treatment effect in LReg is $8 - 10\%$ higher than in PS analysis, slightly decreasing with sample size.

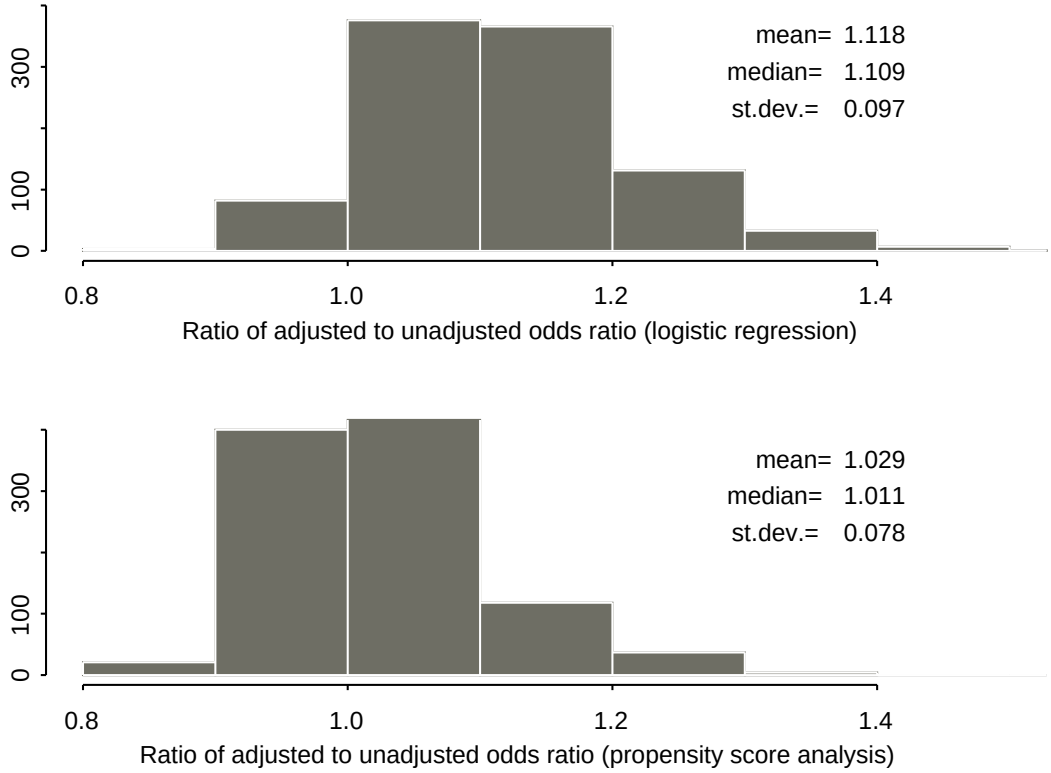
Table 3.2: Summary measures of the ratio of adjusted odds ratios of treatment effect in logistic regression compared to propensity score analysis in $1,000$ samples.

	n=200	n=400	n=800	n=1,600
Mean	1.102	1.087	1.085	1.082
Median	1.094	1.081	1.082	1.082
Standard deviation	0.096	0.055	0.038	0.030
Fraction > 1	0.887	0.970	0.994	0.999

COMPARISON OF ADJUSTED AND UNADJUSTED TREATMENT EFFECTS

Apart from a comparison between LReg and PS methods, it is relevant to compare the adjusted effect in both methods to the unadjusted effect, which in our setting equals on average the true treatment effect. Ideally, the average of the ratio of adjusted to unadjusted effect should be located around 1, because then the adjusted effect is an unbiased estimator of the true treatment effect.

Figure 3.3: Histograms of the ratio of adjusted to unadjusted odds ratios of treatment effect in logistic regression and propensity score analysis, $1,000$ samples of $n = 400$



The results are presented in Figure 3.3 for sample sizes of 400. From the upper panel it can be concluded that when the adjusted treatment effect is used as treatment effect estimate instead of the unadjusted treatment effect (in this setting known on average to be true), LReg systematically overestimates the effect by 12%. In contrast, the center of the histogram for PS stratification is much closer to 1 with an overestimation of only 3%. Another difference is the smaller standard deviation in PS analysis (0.078) compared to LReg (0.097). When the number of prognostic factors, the incidence proportion, the strength of the treatment effect or the strength of the association between prognostic factors and outcome increase, the overestimation in LReg compared to PS methods also increases.[20]

CONCLUSION AND DISCUSSION

In medical studies logistic regression analysis and propensity score methods are both applied to estimate an adjusted treatment effect in observational studies. Although effect estimates of both methods are classified as 'similar' and 'not substantially different', we stressed that differences are systematic and can be substantial. With respect to the objective to adjust for the imbalance of covariate distributions between treatment groups, we illustrated that the estimate of propensity score methods is in general closer to the true average treatment effect than the estimate of logistic regression analysis. The advantage can be substantial, especially when the number of prognostic factors is more than 5, the treatment effect is larger than an odds ratio of 1.25 (or smaller than 0.8) or the incidence proportion is between 0.05 and 0.95. This implies that there is an advantage of propensity score methods over logistic regression models that is frequently overlooked by analysts in the literature.

We showed that the number of included factors in the outcome model is one of the explanations for the difference in treatment effect estimates between the studied methods in which odds ratios are involved. For PS methods without further adjustment, this is only 2 (i.e. the propensity score and treatment), while for LReg this is in general much larger (the number of included covariates plus 1). For that reason it is to be expected that the main results are not largely dependent on the specific PS method used (stratification, matching, covariate adjustment or weighting), except when PS methods are combined with further adjustment for confounding by entering some or all covariates separately in the outcome model. Besides PS stratification we also used covariate adjustment using the PS. We hardly found any differences and speculate that the same is true for other PS methods like matching or weighing on the PS.

We used only the most simple PS model (all covariates linearly included) and did not make any effort to improve the PS model in order to minimize imbalances.[26] The advantage of PS methods is expected to be larger when a more optimal PS model will be chosen.

We conclude that PS methods in general result in treatment effect estimates that are closer to the true average treatment effect than a logistic regression model in which all confounders are modeled.

REFERENCES

[1] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

[2] D'Agostino, RB Jr. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med*, 17:2265–2281, 1998.

[3] Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *JAMA*, 387:516–524, 1984.

[4] Braitman LE, Rosenbaum PR. Rare outcomes, common treatments: analytic strategies using propensity scores. *Ann Intern Med*, 137:693–695, 2002.

[5] Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol*, 158:280–287, 2003.

[6] Glynn RJ, Schneeweiss S, Stürmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic and Clinical Pharmacology and Toxicology*, 98:253–259, 2006.

[7] Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiol Drug Saf*, 13(12):841–853, 2004.

[8] Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *J Clin Epidemiol*, 58:550–559, 2005.

[9] Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol*, 59:437–447, 2006.

[10] Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, 71:431444, 1984.

[11] Gail MH. The effect of pooling across strata in perfectly balanced studies. *Biometrics*, 44:151163, 1988.

[12] Robinson LD, Jewell NP. Some surprising results about covariate adjustment in logistic regression models. *Int Stat Rev*, 58:227–240, 1991.

[13] Guo J, Geng Z. Collapsibility of logistic regression coefficients. *J R Statist Soc B*, 57:263–267, 1995.

[14] Greenland S, Robins MR, Pearl J. Confounding and collapsibility in causal inference. *Stat Science*, 14:29–46, 1999.

[15] Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *Am J Epidemiol*, 125:761–768, 1987.

[16] Wickramaratne PJ, Holford ThR. Confounding in epidemiologic studies: The adequacy of the control group as a measure of confounding. *Biometrics*, 43:751–765, 1987. Erratum in: Biometrics 45:1039, 1989.

[17] Bretagnolle J, Huber-Carol C. Effects of omitting covariates in Coxs model for survival data. *Scand J Stat*, 15:125–138, 1988.

[18] Morgan TM, Lagakos SW, Schoenfeld DA. Omitting covariates from the proportional hazards model. *Biometrics*, 42:993–995, 1986.

[19] Robins JM, Tsiatis AA. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Commun Statist -Theory Meth*, 20(8):2609–2631, 1991.

[20] Rosenbaum PR. *Propensity score. In: Armitage P, Colton T, eds. Encyclopedia of biostatistics*. Wiley, Chichester, United Kingdom, 1998.

[21] Austin PC, Grootendorst P, Normand ST, Anderson GM. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. *Stat Med*, 26:754–768, 2007.

[22] Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Stat Med*, 26:734–753, 2007.

[23] Negassa A, Hanley JA. The effect of omitted covariates on confidence interval and study power in binary outcome analysis: A simulation study. *Cont Clin trials*, 28:242–248, 2007.

[24] Martens EP, Pestman WR, Klungel OH. Letter to the editor: 'Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study, by Austin PC, Grootendorst P, Normand ST, Anderson GM'. *Stat Med*, 26:3208–3210, 2007.

[25] Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. *Stat Med*, 2007. On line: DOI: 10.1002/sim.2781.

[26] Rubin DB. On principles for modeling propensity scores in medical research (Editorial). *Pharmacoepidemiol Drug Saf*, 13:855–857, 2004.

3.3 INSTRUMENTAL VARIABLES: APPLICATION AND LIMITATIONS

Edwin P. Martens[a,b], Wiebe R. Pestman[b], Anthonius de Boer[a], Svetlana V. Belitser[a] and Olaf H. Klungel[a]

[a] *Department of Pharmacoepidemiology and Pharmacotherapy, Utrecht Institute of Pharmaceutical Sciences (UIPS), Utrecht University, Utrecht, the Netherlands*
[b] *Centre for Biostatistics, Utrecht University, Utrecht, the Netherlands*

Epidemiology 2006; 17:260-267

ABSTRACT

To correct for confounding, the method of instrumental variables (IV) has been proposed. Its use in medical literature is still rather limited because of unfamiliarity or inapplicability. By introducing the method in a non-technical way, we show that IV in a linear model is quite easy to understand and easy to apply once an appropriate instrumental variable has been identified. We also point at some limitations of the IV estimator when the instrumental variable is only weakly correlated with the exposure. The IV estimator will be imprecise (large standard error), biased when sample size is small, and biased in large samples when one of the assumptions is only slightly violated. For these reasons it is advised to use an IV that is strongly correlated with exposure. However, we further show that under the assumptions required for the validity of the method, this correlation between IV and exposure is limited. Its maximum is low when confounding is strong, for instance in case of confounding by indication. Finally we show that in a study where strong confounding is to be expected and an IV has been used that is moderately or strongly related to exposure, it is likely that the assumptions of IV are violated, resulting in a biased effect estimate. We conclude that instrumental variables can be useful in case of moderate confounding, but are less useful when strong confounding exists, because strong instruments cannot be found and assumptions will be easily violated.

Keywords: Confounding; Instrumental variables; Adjustment method; Structural equations; Non-compliance

INTRODUCTION

In medical research randomized controlled trials (RCTs) remain the gold standard in assessing the effect of one variable of interest, often a specified treatment. Nevertheless, observational studies are often used in estimating such an effect.[1] In epidemiologic as well as sociological and economic research, observational studies are the standard for exploring causal relationships between an exposure and an outcome variable. The main problem of estimating the effect in such studies is the potential bias resulting from confounding between the variable of interest and alternative explanations for the outcome (confounders). Traditionally, standard methods such as stratification, matching, and multiple regression techniques have been used to deal with confounding. In the epidemiologic literature some other methods have been proposed,[2,3] of which the method of propensity scores is best known.[4] In most of these methods, adjustment can be made only for observed confounders.

A method that has the potential to adjust for all confounders, whether observed or not, is the method of *instrumental variables* (IV). This method is well known in economics and econometrics as the estimation of *simultaneous regression equations*[5] and is also referred to as structural equations and two-stage least squares. This method has a long tradition in economic literature, but has entered more recently into the medical research literature with increased focus on the validity of the instruments. Introductory texts on instrumental variables can be found in Greenland[6] and Zohoori and Savitz.[7]

One of the earliest applications of IV in the medical field is probably the research of Permutt and Hebel,[8] who estimated the effect of smoking of pregnant women on their child's birth weight, using an encouragement to stop smoking as the instrumental variable. More recent examples can be found in Beck *et al.*,[9] Brooks *et al.*,[10] Earle *et al.*,[11] Hadley *et al.*,[12] Leigh and Schembri,[13] McClellan[14] and McIntosh.[15] However, it has been argued that the application of this method is limited because of its strong assumptions, making it difficult in practice to find a suitable instrumental variable.[16]

The objectives of this paper are first to introduce the application of the method of IV in epidemiology in a non-technical way, and second to show the limitations of this method, from which it follows that IV is less useful for solving large confounding problems such as confounding by indication.

A SIMPLE LINEAR IV MODEL

In a randomized controlled trial (RCT) the main purpose is to estimate the effect of one explanatory factor (the treatment) on an outcome variable. Because treatments have been randomly assigned to individuals, the treatment variable is in general independent of other explanatory factors. In case of a continuous outcome and a linear model, this randomization procedure allows one to estimate the treatment effect by means of ordinary least squares with a well

known unbiased estimator (see for instance Pestman[17]). In observational studies, on the other hand, one has no control over this explanatory factor (further denoted as *exposure*) so that ordinary least squares as an estimation method will generally be biased because of the existence of unmeasured *confounders*. For example, one cannot directly estimate the effect of cigarette smoking on health without considering confounding factors such as age and socioeconomic position.

One way to adjust for all possible confounding factors, whether observed or not, is to make use of an instrumental variable. The idea is that the causal effect of exposure on outcome can be captured by using the relationship between the exposure and another variable, the instrumental variable. How this variable can be selected and which conditions have to be fulfilled, is discussed below. First we will illustrate the model and its estimator.

THE IV MODEL AND ITS ESTIMATOR

A simple linear model for IV-estimation consists of two equations

$$Y = \alpha + \beta X + E \tag{3.5}$$

$$X = \gamma + \delta Z + F \tag{3.6}$$

where Y is the outcome variable, X is the exposure, Z is the instrumental variable and E and F are errors. In this set of structural equations the variable X is *endogenous*, which means that it is explained by other variables in the model, in this case the instrumental variable Z. Z is supposed to be linearly related to X and *exogenous*, i.e. explained by variables outside the model. For simplicity we restrict ourselves to one instrumental variable, two equations and no other explaining variables. Under conditions further outlined in the next section, it can be proved that expression 3.7 presents an asymptotically unbiased estimate of the effect of X on Y:[18]

$$\hat{\beta}_{iv} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})} = \frac{\hat{\sigma}_{Z,Y}}{\hat{\sigma}_{Z,X}} \tag{3.7}$$

where $\hat{\sigma}_{Z,Y}$ is the sample covariance of Z and Y and $\hat{\sigma}_{Z,X}$ is the sample covariance of Z and X. It is more convenient to express the IV estimator in terms of two ordinary least squares estimators:

$$\hat{\beta}_{iv} = \frac{\hat{\sigma}_{Z,Y}}{\hat{\sigma}_{Z,X}} = \frac{\hat{\sigma}_{Z,Y}/\hat{\sigma}_Z^2}{\hat{\sigma}_{Z,X}/\hat{\sigma}_Z^2} = \frac{\hat{\beta}_{ols(Z\rightarrow Y)}}{\hat{\beta}_{ols(Z\rightarrow X)}} \tag{3.8}$$

The numerator equals the effect of the instrumental variable on the outcome, whereas in the denominator the effect of the IV on the exposure is given. In case of a dichotomous IV, the numerator equals simply the difference in mean outcome between $Z = 0$ and $Z = 1$ and the denominator equals the difference in mean exposure. When the outcome and exposure variable are also dichotomous and linearity is still assumed, this model is known as a linear

probability model. In that case the IV estimator presented above can be simply expressed as probabilities:[18]

$$\hat{\beta}_{iv} = \frac{P(Y=1|Z=1) - P(Y=1|Z=0)}{P(X=1|Z=1) - P(X=1|Z=0)} \tag{3.9}$$

where $P(Y=1|Z=1) - P(Y=1|Z=0)$ equals the risk difference of an event between $Z=1$ and $Z=0$.

HOW TO OBTAIN A VALID INSTRUMENTAL VARIABLE

One can imagine that a method that claims to adjust for all possible confounders without randomization of treatments puts high requirements on the IV to be used for estimation. When this method is applied, three important assumptions have been made. The first assumption is the existence of at least some correlation between the IV and the exposure, because otherwise equation 3.6 would be useless and the denominator of equation 3.8 would be equal to zero. In addition to this formal condition it is important that this correlation should not be too small (see *Implications of weak instruments*).

The second assumption is that the relationship between the instrumental variable and the exposure is not confounded by other variables, so that equation 3.6 is estimated without bias. This is the same as saying that the correlation between the IV and the error F must be equal to zero. One way to achieve this, is to use as IV a variable that is *controlled by the researcher*. An example can be found in Permutt and Hebel,[8] where a randomized encouragement to stop smoking was used as the IV to estimate the effect of smoking by pregnant women on child's birth weight. The researchers used two encouragement regimes, an encouragement to stop smoking versus no encouragement, randomly assigned to pregnant smoking women. Alternatively, in some situations a *natural randomization process* can be used as the IV. An example, also known as Mendelian randomization, can be found in genetics where alleles are considered to be allocated at random in offspring with the same parents.[19,20] In a study on the causality between low serum cholesterol and cancer a genetic determinant of serum cholesterol was used as the instrumental variable.[21,22] When neither an active randomization nor a natural randomization is feasible to obtain an IV, the only possibility is to select an IV on *theoretical grounds*, assuming and reasoning that the relationship between the IV and the exposure can be estimated without bias. Such an example can be found in Leigh and Schembri[13] where the observed cigarette price per region was used as the IV in a study on the relationship between smoking and health. The authors argued that there was no bias in estimating the relationship between cigarette price and smoking because the price elasticities in their study (the percentage change in number of cigarettes smoked related to the percentage change in cigarette price) matched the price elasticities mentioned in the literature.

The third assumption for an IV is most crucial, and states that there should be no correlation between the IV and the error E (further referred to as *the main assumption*). This means that the instrumental variable should influence the outcome neither directly, nor indirectly by

its relationship with other variables. Whether this assumption is valid can be argued only theoretically, and cannot be tested empirically.
These three assumptions can be summarized as follows:

1) $\rho_{Z,X} \neq 0$, no zero-correlation between IV and exposure
2) $\rho_{Z,F} = 0$, no correlation between IV and other factors explaining X (error F)
3) $\rho_{Z,E} = 0$, no correlation between IV and other factors explaining Y (error E),
main assumption

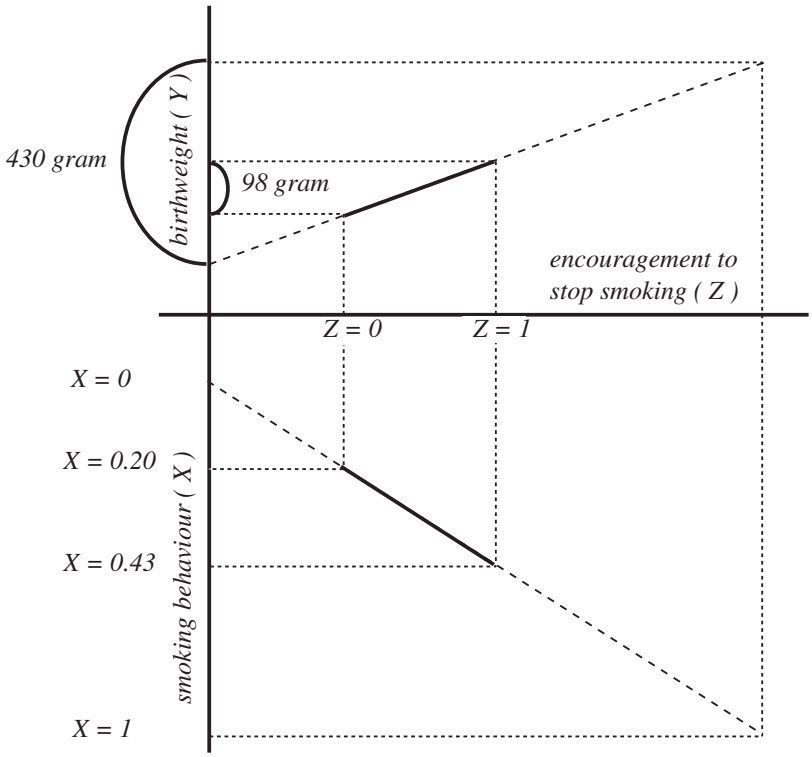
It should be noted that confounders of the X-Y relation are not explicitly mentioned in these assumptions, and that these confounders are part of both errors E and F. In the special case that $\rho_{E,F} = 1$, only confounders can be used to formulate the assumptions.[6]

NUMERICAL EXAMPLE OF IV APPLICATION

As an example of IV estimation we will use the research of Permutt and Hebel.[8] Here the effect of smoking (X) by pregnant women on child's birth weight (Y) was studied. The instrumental variable (Z) was the randomization procedure used to assign women to an encouragement program to stop smoking, which fulfills the second assumption. To apply IV estimation, first the intention-to-treat estimator $\hat{\beta}_{ols(Z \rightarrow Y)}$ needs to be calculated. In case of a dichotomous IV this simply equals the difference in mean birth weight between women who were encouraged to stop smoking and women who were not ($\hat{\beta}_{ols(Z \rightarrow Y)} = 98$ gram). Next we calculate the difference between encouragement groups in the fraction of women who stopped smoking ($\hat{\beta}_{ols(Z \rightarrow X)} = 0.43 - 0.20 = 0.23$). The ratio equals the IV-estimator ($= \frac{98}{0.43-0.20} = 430$ gram), indicating that stopping smoking raises average birth weight by 430 gram. Figure 3.4 illustrates this calculation, where "actually stopped smoking" is denoted as $X = 1$ and "continued to smoke" as $X = 0$.

The encouragement-smoking relationship and the encouragement-birth weight relationship are represented by the solid lines in the lower and upper panel respectively. Under the assumptions of IV estimation, the effect of smoking on birth weight is known only when smoking is changed from 0.43 to 0.20, where in fact interest is in a change from $X = 0$ to $X = 1$. Extending this difference to a difference from 0 to 1, indicated by the dotted line in the lower panel, and using the relationship between Z and Y in the upper panel, the intention-to-treat estimator of 98 gram is 'extended' to become the IV estimator of 430 gram. Reminding that our second assumption has been fulfilled by randomization, the possible bias of the IV estimator mainly depends on the assumption that there should be no effect from encouragement on child's birth weight other than by means of changing smoking behavior. Such an effect can not be ruled out completely, for instance because women who were encouraged to stop smoking, could become also more motivated to change other health related behavior as well (for instance nutrition). Birth weight will then be influenced by encouragement independently of smoking, which will lead to an overestimation of the effect of stopping smoking.

Figure 3.4: The instrumental variable estimator in the study of Permutt and Hebel[8]



IMPLICATIONS OF WEAK INSTRUMENTS

In the previous sections the method and application of instrumental variables in a linear model was introduced in a non-technical way. Here we will focus on the implications when the correlation between the instrumental variable and the exposure is small, or when the instrument is weak. We will refer to this correlation as $\rho_{Z,X}$.

LARGE STANDARD ERROR

A weak instrument means that the denominator in equation 3.8 is small. The smaller this covariance, the more sensitive the IV estimate will be to small changes. This sensitivity is mentioned by various authors[16,23] and can be deduced from the formula for the standard error:

$$\hat{\sigma}_{\beta_{iv}} = \frac{\sigma_Z \, \sigma_E}{\sigma_{Z,X}} \tag{3.10}$$

where σ_Z is the standard deviation of Z, σ_E is the standard deviation of E and $\sigma_{Z,X}$ is the covariance of Z and X. This covariance in the denominator behaves as a multiplier, which

means that a small covariance (and hence a small correlation) will lead to a large standard error. In Figure 3.4 this sensitivity is reflected by the fact that the slope estimate in the lower panel becomes less reliable when the difference in X between $Z = 0$ and $Z = 1$ becomes smaller.

BIAS WHEN SAMPLE SIZE IS SMALL

An important characteristic of an estimator is that it should equal on average the true value (*unbiasedness*). Assuming that the assumptions of IV are not violated, the IV estimator is only *asymptotically* unbiased, meaning that on average bias will exist when the estimator $\hat{\beta}_{iv}$ is used in smaller samples. This bias appears because the relationship between the instrumental variable and the exposure is in general unknown and has to be estimated by equation 3.6. As is usual in regression, overfitting generates a bias that depends on both the sample size and the correlation between the IV and the exposure. With moderate sample size and a weak instrument, this bias can become substantial.[24] It can be shown that this bias will be in the direction of the ordinary least squares estimator $\hat{\beta}_{ols}$ calculated in the simple linear regression of outcome on exposure.[23,25] Information on the magnitude of the small-sample bias is contained in the F-statistic of the regression in equation 3.6, which can be expressed as

$$F = \frac{\hat{\rho}^2_{Z,X}(n-2)}{1-\hat{\rho}^2_{Z,X}} \tag{3.11}$$

An F-value not far from 1 indicates a large small-sample bias, whereas a value of 10 seems to be sufficient for the bias to be negligible.[16] For example, in a sample of 250 independent observations the correlation between Z and X should be at least 0.20 to reach an F-value of 10. Another solution to deal with possible small-sample bias is to use other IV estimators.[16,26]

BIAS WHEN THE MAIN ASSUMPTION IS ONLY SLIGHTLY VIOLATED

Every violation of the main assumption of IV will naturally result in a biased estimator. More interesting is that only a small violation of this assumption will result in a large bias in case of a weak instrument because of its multiplicative effect in the estimator. Bound *et al*.[23] expressed this bias in infinitely large samples (inconsistency) as a relative measure compared with the bias in the ordinary least squares estimator

$$\frac{\lim \hat{\beta}_{iv} - \beta}{\lim \hat{\beta}_{ols} - \beta} = \frac{\rho_{Z,E}/\rho_{X,E}}{\rho_{Z,X}} \tag{3.12}$$

where *lim* is the limit as sample size increases. From this formula it can be seen that even a small correlation between the instrumental variable and the error ($\rho_{Z,E}$ in the denominator) will produce a large inconsistency in the IV estimate relative to the ordinary least squares estimate when the instrument is weak, i.e. when $\rho_{Z,X}$ is small. Thus, when Z has some small direct

effect on Y, or an indirect effect other than through X, the IV estimate will be increasingly biased when the instrument becomes weaker, even in very large samples.

It can be concluded that a small correlation between the IV and the exposure can be a threat for the validity of the IV method, mainly in combination with a small sample or a possible violation of the main assumption. Although known from the literature, this aspect is often overlooked.

A LIMIT ON THE STRENGTH OF INSTRUMENTS

From the last section it follows that the correlation between a possible instrumental variable and exposure (the strength of the IV $\rho_{Z,X}$) has to be as strong as possible, which also intuitively makes sense. However, in practice it is often difficult to obtain an IV that is strongly related to exposure. One reason can be found in the existence of an upper bound on this correlation, which depends on the amount of confounding (indicated by $\rho_{X,E}$), the correlation between the errors in the model ($\rho_{E,F}$) and the degree of violation of the main assumption ($\rho_{Z,E}$). We will further explore the relationship between these correlations, and will distinguish between a situation where the main assumption is fulfilled and one in which it is not.

WHEN THE MAIN ASSUMPTION HAS BEEN FULFILLED

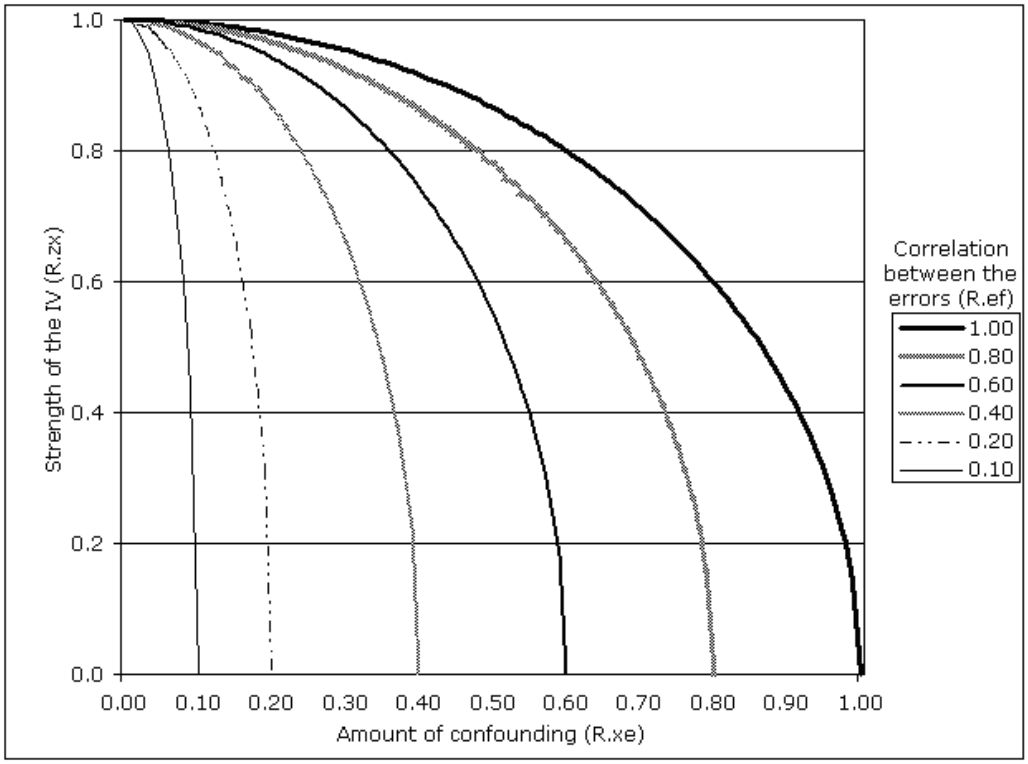
In case the main assumption of IV has been fulfilled, which means that the IV changes the outcome only through its relationship with the exposure, it can be shown that

$$|\rho_{Z,X}| = \sqrt{1 - \frac{\rho_{X,E}^2}{\rho_{E,F}^2}} \tag{3.13}$$

of which the proof is given in Appendix A. Equation 3.13 indicates that there is a maximum on the strength of the instrumental variable, and that this maximum decreases when the amount of confounding increases. In case of considerable confounding, the maximum correlation between IV and exposure will be quite low. This relationship between the correlations is illustrated in Figure 3.5.

The relation between the strength of the IV $\rho_{Z,X}$ and the amount of confounding $\rho_{X,E}$ is illustrated by curves representing various levels of the correlation between the errors $\rho_{E,F}$. It can be seen that the maximum correlation between the potential instrumental variable and exposure becomes smaller when the amount of confounding becomes larger. When for example there is considerable confounding by indication ($\rho_{X,E} = 0.8$), the maximum strength of the IV is 0.6. Probably this maximum will be even lower because the correlation between the errors will generally be less than 1.0. When for instance $\rho_{E,F} = 0.85$ this maximum drops to only 0.34.

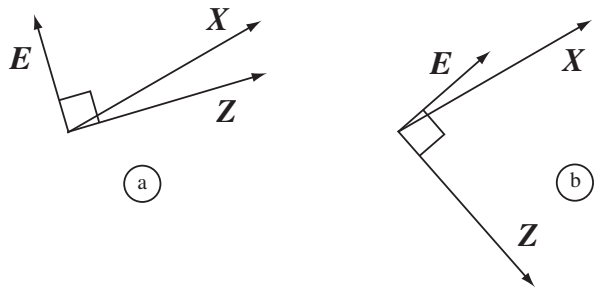
Figure 3.5: Relationship between strength of an instrumental variable ($\rho_{Z,X}$) and amount of confounding ($\rho_{X,E}$) for different error correlation levels ($\rho_{E,F}$), when main assumption has been fulfilled ($\rho_{Z,E} = 0$)



Of the three correlations presented in equation 3.13 and Figure 3.5, the correlation between the errors is most difficult to understand. For the main message, however, its existence is not essential, as is illustrated in Figure 3.6 using vectors.

In panel a of Figure 3.6 the angle between X and E is close to $90°$, meaning that their correlation is small (small confounding). Because Z has to be uncorrelated with E according to the third IV assumption (perpendicular), the angle between X and Z will be automatically small, indicating a strong IV. In contrast, panel b of Figure 3.6 shows that a large confounding problem (small angle between X and E) implies a weak instrument (large angle and small correlation between X and Z). The trade-off between these correlations is an important characteristic of IV estimation. (Note that we simplified the figure by choosing Z in the same plane as X and Y in order to remove $\rho_{E,F}$ from the figure because it equals its maximum of 1.0. See Appendix B for the situation in which Z is not in this plane.)

Figure 3.6: Relationship among X, Z and E expressed in vectors



As has been said, the correlation between the errors $\rho_{E,F}$ also plays a role. To better understand its meaning we give two examples. In Permutt and Hebel,[8] it is likely that this correlation will be small. Other reasons for birth weight variation besides smoking include socioeconomic conditions, inadequate nutrition, abuse, genetic factors, ethnic factors, physical work conditions and chronic diseases. Because these explanatory factors for birth weight will be only partly overlapping with the reasons for *non-compliance*, i.e. to continue smoking while encouraged to stop, $\rho_{E,F}$ is expected to be small. When, on the other hand, this correlation approaches 1, it means that the set of variables accounting for the unexplained variation in the outcome Y (error E) is strongly correlated with the unexplained instrumental variance (error F). An example of such a large correlation is a case of strong confounding by indication, where unobserved health problems are the main reason for getting an illness and also for receiving preventive treatment. That causes variables E and F to be strongly correlated and the maximum strength of the IV to be relatively small (see the right side of Figure 3.5).

WHEN THE MAIN ASSUMPTION HAS NOT BEEN FULFILLED

When the main assumption has not been (completely) fulfilled, the correlation between Z and E is not equal to 0. Because the correlation between the errors plays a minor role, this correlation has been set to its maximum value of 1. In that case the next inequality holds:

$$\rho_{Z,X} \leq |\rho_{Z,E}|\,|\rho_{X,E}| + \sqrt{1-\rho_{Z,E}^2}\,\sqrt{1-\rho_{X,E}^2} \tag{3.14}$$

Like equation 3.13, this expression states that in case of considerable confounding the strength of the instrumental variable is bound to a relatively small value. It further states that a trade-off exists between $\rho_{Z,X}$ and $\rho_{Z,E}$: given a certain degree of confounding, the strength of the IV can be enlarged by relaxing the main assumption. In practice this means that when IV is applied to a situation in which a considerable amount of confounding is to be expected and a very strong instrument has been found, it is very likely that the main assumption has been violated.

THE EFFECT ON BIAS

The limit of the correlation between exposure and instrumental variable has an indirect effect on the bias, because the correlation to be found in practice will be low. This has several disadvantages that can be illustrated using some previous numerical examples. Suppose we deal with strong confounding by indication, say $\rho_{X,E} = 0.80$. As has been argued before, this will naturally imply a strong but imperfect correlation between the errors, say $\rho_{E,F} = 0.85$. In that case, the limit of the correlation between exposure and IV will be $\rho_{Z,X} = 0.34$. Restricting ourselves to instrumental variables that fulfill the main assumption ($\rho_{Z,E} = 0$), it will be practically impossible to find an IV that possess the characteristic of being maximally correlated with exposure, which implies that this correlation will be lower than 0.34, for instance 0.20. With such a small correlation, the effect on the bias will be substantial when sample size falls below 250 observations. Because we cannot be sure that the main assumption has been fulfilled, care must be taken even with larger samples sizes.

DISCUSSION

We have focused on the method of instrumental variables for its ability to adjust for confounding in non-randomized studies. We have explained the method and its application in a linear model and focused on the correlation between the IV and the exposure. When this correlation is very small, this method will lead to an increased standard error of the estimate, a considerable bias when sample size is small and a bias even in large samples when the main assumption is only slightly violated. Furthermore, we demonstrated the existence of an upper bound on the correlation between the IV and the exposure. This upper bound is not a practical limitation when confounding is small or moderate because the maximum strength of the IV is still very high. When, on the other hand, considerable confounding by indication exists, the maximum correlation between any potential IV and the exposure will be quite low, resulting possibly in a fairly weak instrument in order to fulfill the main assumption. Because of a trade-off between violation of this main assumption and the strength of the IV, the presence of considerable confounding and a strong instrument will probably indicate a violation of the main assumption and thus a biased estimate.

This paper serves as an introduction on the method of instrumental variables demonstrating its merits and limitations. Complexities such as more equations, more instruments, the inclusion of covariates and non-linearity of the model have been left out. More equations could be added with more than two endogenous variables, although it is unlikely to be useful in epidemiology when estimating an exposure (treatment) effect. In equation 3.6, multiple instruments could be used; this extension does not change the basic ideas behind this method.[27] An advantage of more than one instrumental variable is that a test on the exogeneity of the instruments is possible.[16] Another extension is the inclusion of measured covariates in both equations.[27]

We limited the model to linear regression, assuming that the outcome and the exposure are

both continuous variables, while in medical research dichotomous outcomes or exposures are more common. The main reason for this choice is simplicity: the application and implications can be more easily presented in a linear framework. A dichotomous outcome or dichotomous exposure can easily fit into this model when linearity is assumed using a *linear probability model*. Although less known, the results from this model are practically indistinguishable from logistic and probit regression analyses, as long as the estimated probabilities range between 0.2 and 0.8.[28,29] When risk ratios or log odds are to be analyzed, as in logistic regression analysis, the presented IV-estimator cannot be used and more complex IV-estimators are required. We refer to the literature for IV-estimation in such cases or in non-linear models in general.[6,30,31] The limitations when instruments are weak, and the impossibility of finding strong instruments in the presence of strong confounding, apply in a similar way.

When assessing the validity of study results, investigators should report both the correlation between IV and exposure (or difference in means) and the F-value resulting from equation 3.6 and given in equation 3.11. When either of these are small, instrumental variables will not produce unbiased and reasonably precise estimates of exposure effect. Furthermore, it should be made clear whether the IV is randomized by the researcher, randomized by nature, or is simply an observed variable. In the latter case, evidence should be given that the various categories of the instrumental variable have similar distributions on important characteristics. Additionally, the assumption that the IV determines outcome only by means of exposure is crucial. Because this can not be checked, it should be argued theoretically that a direct or indirect relationship between the IV and the outcome is negligible. Finally, in a study in which considerable confounding can be expected (e.g. strong confounding by indication), one should be aware that the existence of a very strong instrument within the IV assumptions is impossible. Whether the instrument is sufficiently correlated with exposure depends on the number of observations and the plausibility of the main assumption.

We conclude that the method of IV can be useful in case of moderate confounding, but is less useful when strong confounding (by indication) exists, because strong instruments can not be found and assumptions will be easily violated.

APPENDIX A

Theorem 1
The correlation between Z and X, $\rho_{Z,X}$ is bound to obey the equality

$$|\rho_{Z,X}| = \sqrt{1 - \frac{\rho^2_{X,E}}{\rho^2_{E,F}}} \tag{3.15}$$

Proof: According to the model one has

$$\begin{cases} Y = \alpha + \beta X + E \\ X = \gamma + \delta Z + F \end{cases}$$

with

$$\sigma_{Z,E} = 0 \quad \text{and} \quad \sigma_{Z,F} = 0$$

It follows from this that $\sigma_{X,E} = \sigma_{\gamma,E} + \delta\,\sigma_{Z,E} + \sigma_{F,E} = 0 + 0 + \sigma_{E,F} = \sigma_{E,F}$. Using this expression for $\sigma_{X,E}$ one derives that

$$\begin{aligned} \rho_{X,E} &= \frac{\sigma_{X,E}}{\sigma_X \sigma_E} = \frac{\sigma_{E,F}}{\sigma_X \sigma_E}\frac{\sigma_F}{\sigma_F} = \rho_{E,F}\frac{\sigma_F}{\sigma_X} \\ &= \pm\sqrt{\rho^2_{E,F}\frac{\sigma^2_F}{\sigma^2_X}} = \pm\sqrt{\rho^2_{E,F}(1-\rho^2_{Z,X})} \end{aligned}$$

Squaring, rearranging terms and taking square roots will give

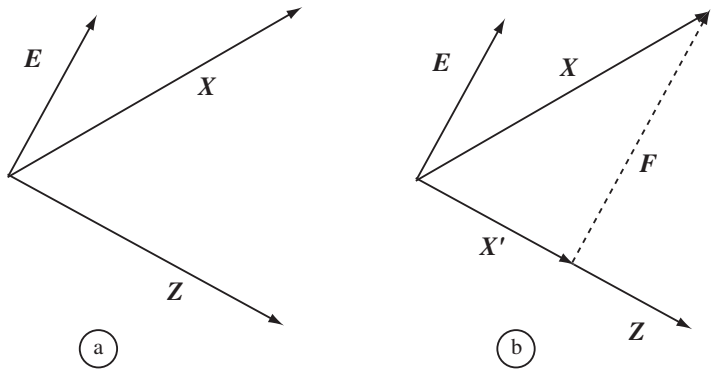
$$|\rho_{Z,X}| = \sqrt{1 - \frac{\rho^2_{X,E}}{\rho^2_{E,F}}}$$

which proves the theorem. □

APPENDIX B

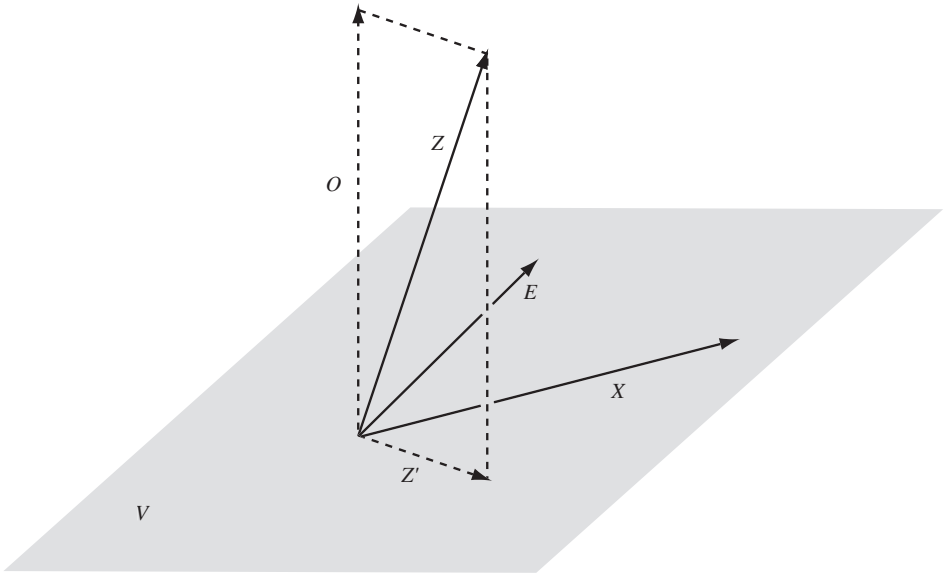
The condition $\rho_{E,F} = 1$ is equivalent to the condition that Z is in the same plane as X and E as can be seen in Figure 3.7. For simplicity we assume that the expectation values of the variables X, Y and Z are all equal to zero.

Figure 3.7: Relationship between X, Z, E and F expressed in vectors



According to the IV condition that $\rho_{Z,E} = 0$ (these are perpendicular in panel a of Figure 3.7) and the condition that $\rho_{Z,F} = 0$, it follows from panel b of Figure 3.7 that E and F necessarily point in the same or opposite direction, implying $\rho_{E,F} = 1$. In this situation there is (up to scalar multiples) only one instrumental variable Z possible in the plane spanned by E and X. As has been argued in the text, it is not likely that this correlation equals 1. This is visualized in Figure 3.8 where Z is not in the plane spanned by X and E, meaning that F, which is in the plane spanned by X and Z and perpendicular to Z, can impossibly point in the same direction as E. Consequently one then has $\rho_{E,F} < 1$. Here Z' is the projection of Z on the plane spanned by E and X. The vector Z can now be decomposed as $Z = Z' + O$ where Z' is in the plane spanned by E and X and where O is perpendicular to this plane. The vector O can be referred to as *noise* because it is uncorrelated to both X and Y. Note that the variable Z' is an instrumental variable itself.

Figure 3.8: Three dimensional picture of X, Z, E and noise O expressed in vectors



REFERENCES

[1] Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med*, 342:1887–1892, 2000.

[2] McMahon AD. Approaches to combat with confounding by indication in observational studies of intended drug effects. *Pharmacoepidemiol Drug Saf*, 12:551–558, 2003.

[3] Klungel OH, Martens EP, Psaty BM, *et al*. Methods to assess intended effects of drug treatment in observational studies are reviewed. *J Clin Epidemiol*, 57:1223–1231, 2004.

[4] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

[5] Theil H. *Principles of Econometrics*. Wiley, 1971.

[6] Greenland S. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol*, 29:722–729, 2000.

[7] Zohoori N, Savitz DA. Econometric approaches to epidemiologic data: relating endogeneity and unobserved heterogeneity to confounding. *Ann Epidemiol*, 7:251–257, 1997. Erratum in: Ann Epidemiol 7:431, 1997.

[8] Permutt Th, Hebel JR. Simultaneous-equation estimation in a clinical trial of the effect of smoking on birth weight. *Biometrics*, 45:619–622, 1989.

[9] Beck CA, Penrod J, Gyorkos TW, Shapiro S, Pilote L. Does aggressive care following acute myocardial infarction reduce mortality? Analysis with instrumental variables to compare effectiveness in Canadian and United States patient populations. *Health Serv Res*, 38:1423–1440, 2003.

[10] Brooks JM, Chrischilles EA, Scott SD, Chen-Hardee SS. Was breast conserving surgery underutilized for early stage breast cancer? Instrumental variables evidence for stage II patients from Iowa. *Health Serv Res*, 38:1385–1402, 2003. Erratum in: Health Serv Res 2004;39(3):693.

[11] Earle CC, Tsai JS, Gelber RD, Weinstein MC, Neumann PJ, Weeks JC. Effectiveness of chemotherapy for advanced lung cancer in the elderly: instrumental variable and propensity analysis. *J Clin Oncol*, 19:1064–1070, 2001.

[12] Hadley J, Polsky D, Mandelblatt JS, *et al*. An exploratory instrumental variable analysis of the outcomes of localized breast cancer treatments in a medicare population. *Health Econ*, 12:171–186, 2003.

[13] Leigh JP, Schembri M. Instrumental variables technique: cigarette price provided better estimate of effects of smoking on SF-12. *J Clin Epidemiol*, 57:284–293, 2004.

[14] McClellan M, McNeil BJ, Newhouse JP. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *JAMA*, 272:859–866, 1994.

[15] McIntosh MW. Instrumental variables when evaluating screening trials: estimating the benefit of detecting cancer by screening. *Stat Med*, 18:2775–2794, 1999.

[16] Staiger D, Stock JH. Instrumental variables regression with weak instruments. *Econometrica*, 65:557–586, 1997.

[17] Pestman WR. *Mathematical Statistics*. Walter de Gruyter, Berlin, New York, 1998.

[18] Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *JASA*, 91:444–455, 1996.

[19] Thomas DC, Conti DV. Commentary: the concept of 'Mendelian Randomization'. *Int J Epidemiol*, 33:21–25, 2004.

[20] Minelli C, Thompson JR, Tobin MD, Abrams KR. An integrated approach to the meta-analysis of genetic association studies using Mendelian Randomization. *Am J Epidemiol*, 160:445–452, 2004.

[21] Katan MB. Apolipoprotein E isoforms, serum cholesterol, and cancer. *Lancet*, 1:507–508, 1986.

[22] Smith GD, Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. *Int J Epidemiol*, 33:30–42, 2004.

[23] Bound J, Jaeger DA, Baker RM. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *JASA*, 90:443–450, 1995.

[24] Sawa T. The exact sampling distribution of ordinary least squares and two-stage least squares estimators. *J Am Stat Ass*, 64:923–937, 1969.

[25] Nelson CR, Startz R. Some further results on the exact small sample properties of the instrumental variable estimator. *Econometrica*, 58:967–976, 1990.

[26] Angrist JD, Krueger AB. Split sample instrumental variables. *J Bus and Econ Stat*, 13:225–235, 1995.

[27] Angrist JD, Imbens GW. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *JASA*, 90:431–442, 1995.

[28] Cox DR, Snell EJ. *Analysis of Binary Data*. Chapman and Hall, 1989.

[29] Cox DR, Wermuth N. A comment on the coefficient of determination for binary responses. *The American Statistician*, 46:1–4, 1992.

[30] Bowden RJ, Turkington DA. A comparative study of instrumental variables estimators for nonlinear simultaneous models. *J Am Stat Ass*, 76:988–995, 1981.

[31] Amemiya T. The nonlinear two-stage least-squares estimator. *Journal of econometrics*, 2:105–110, 1974.

CHAPTER 4

APPLICATION OF ADJUSTMENT METHODS

4.1 COMPARING TREATMENT EFFECTS AFTER ADJUSTMENT WITH MULTIVARIABLE COX PROPORTIONAL HAZARDS REGRESSION AND PROPENSITY SCORE METHODS

Edwin P. Martens[a,b], Anthonius de Boer[a], Wiebe R. Pestman[b], Svetlana V. Belitser[a], Bruno H. Ch. Stricker[c] and Olaf H. Klungel[a]

[a] *Department of Pharmacoepidemiology and Pharmacotherapy, Utrecht Institute of Pharmaceutical Sciences (UIPS), Utrecht University, Utrecht, the Netherlands*
[b] *Centre for Biostatistics, Utrecht University, Utrecht, the Netherlands*
[c] *Department of Epidemiology and Biostatistics, Erasmus MC, Rotterdam, The Netherlands*

Accepted by Pharmacoepidemiology and Drug Safety

ABSTRACT

Purpose: To compare adjusted effects of drug treatment for hypertension on the risk of stroke from propensity score methods with a multivariable Cox proportional hazards regression in an observational study with censored data.

Methods: From two prospective population-based cohort studies in the Netherlands a selection of subjects was used who either received drug treatment for hypertension ($n = 1,293$) or were untreated "candidates" for treatment ($n = 954$). A multivariable Cox proportional hazards was performed on the risk of stroke using eight covariates along with three propensity score methods.

Results: In multivariable Cox proportional hazards regression the adjusted hazard ratio hazard ratio for treatment was 0.64 (CI 95%: 0.42, 0.98). After stratification on the propensity score the hazard ratio was 0.58 (CI 95%: 0.38, 0.89). Matching on the propensity score yielded a hazard ratio of 0.49 (CI 95%: 0.27, 0.88), whereas adjustment with a continuous propensity score gave similar results as Cox regression. When more covariates were added (not possible in multivariable Cox model) a similar reduction in hazard ratio was reached by all propensity score methods. The inclusion of a simulated balanced covariate gave largest changes in HR using the multivariable Cox model and matching on the propensity score.

Conclusions: In propensity score methods in general a larger number of confounders can be used. In this data set matching on the propensity score is sensitive to small changes in the model, probably because of the small number of events. Stratification, and covariate adjustment, were less sensitive to the inclusion of a non-confounder than multivariable Cox proportional hazards regression. Attention should be paid to propensity score model building and balance checking.

Keywords: Confounding; Propensity scores; Cox proportional hazards regression; Hypertension; Observational studies

INTRODUCTION

Cox proportional hazards regression (Cox PH) has been widely used as an adjustment technique in observational studies with censored data.[1] Often there is one variable of interest (the 'treatment' effect) and a set of covariates (confounders) that are used as independent variables to explain a dichtomous outcome variable. When these covariates are included in the model it can be said that the treatment effect is adjusted for the influence of the observed confounders. An alternative approach in such cases is to use the propensity score (PS), a method originally proposed by Rosenbaum & Rubin in 1983.[2] With this approach the focus is on the *imbalance of covariates* between treatment groups, which can be seen as a result of the non-random assignment of treatments to patients. Therefore, in the PS method first attention is directed to balance treatment groups with respect to the observed covariates and second to estimate the treatment. In fact, a randomized controlled trial (RCT) has a similar two-step procedure: first balancing treatment groups and second estimating treatment effect. Of course, a randomization procedure aims at balancing treatment groups on all confounders, where the PS can only handle confounders that are observed.

This approach is theoretically different from a Cox PH, linear or logistic regression model where an adjusted treatment effect is estimated by using the observed covariates as *additional explanations* for the variation in the outcome variable. This means that the method of PS is an alternative for model-based methods as far as estimation of a treatment effect is concerned; it is no alternative when the objective is to model and estimate the influence of the observed confounders on the outcome variable.

The propensity score is defined as the conditional probability of being treated given the values of covariates. In general this probability is unknown but can be estimated using logistic, probit or discriminant analysis, where treatment is considered the dependent variable. It has been shown that a treated patient and an untreated control with the same PS or classes of subjects with the same PS tend to have the same distribution of covariates.[3] This means that the PS can be used as a single matching or stratification variable to reduce confounding due to observed covariates. Furthermore, the distribution of the PS can be compared between treatment groups, revealing for which part of the treated patients no controls are available and vice versa. This possible lack of overlap is essential information when treatment groups are to be compared on some outcome variable, something that is seldom done or reported when a Cox PH, linear or logistic regression analysis has been performed.

PS methods are increasingly used in the medical literature, but different PS methods and model-based adjustment techniques have been less frequently compared. In a recent simulation study PS stratification was compared to logistic regression analysis[4] and in some other studies a PS analysis was performed together with a regression-based method (among others[5,6]). Our study objective was to systematically compare the effect of drug treatment for hypertension on the risk of stroke between a multivariable Cox PH regression and three PS methods.

MATERIALS AND METHODS

DATA

The data we used have been described by Klungel *et al.*[7]and come from two prospective population-based cohort studies in The Netherlands. Briefly, the first study, the Monitoring Project on Cardiovascular Risk Factors, was conducted from 1987 through 1991 as a cross-sectional study in Amsterdam, Maastricht and Doetinchem (62% agreed to participate). In Doetinchem, subjects were followed up through general practice records. The second study, the Rotterdam Study, was started in 1990 in Rotterdam as a population-based prospective follow-up study. All residents of a suburb of Rotterdam aged 55 years or older were invited to participate (78% agreed). The baseline measurements continued until 1993. In total 1,293 treated hypertensives and 954 untreated "candidates" for treatment were used for analysis, where the incidence of stroke was the outcome. The overall incidence rate was 4.2%, with 42 cases in the treated and 53 cases in the untreated patients. The selection of untreated controls was based on high blood pressure and the existence of other common cardiovascular risk factors. The following confounding factors were available for analysis: history of cerebrovascular disease (CVA), age, sex, diabetes, total cholesterol, body mass index, smoking, previous cardiovascular disease (CVD), previous myocard infarction (MI), previous transient ischemic attack (TIA), family history of MI and HDL-cholesterol.

Three sets of covariates were defined. The first set, motivated by Klungel *et al.*,[7] consists of a selection of eight covariates (history of CVA, age, sex, diabetes, total cholesterol, body mass index, smoking and previous CVD). The second set consists of all available covariates. In order to investigate the sensitivity of the estimated treatment effect for the inclusion of a non-confounder, we created a third set of covariates. This simulated binary non-confounder was not correlated with treatment (equally balanced over treatment groups) nor with all other covariates in the model, but strongly associated with outcome (the incidence of stroke). Inclusion of such a risk factor will not change the estimated treatment effect in linear models, but it will change the effect in models like logistic regression or Cox PH regression.[8] By including this non-confounder we are able to compare the sensitivity to the results of the various methods.

MULTIVARIABLE COX PROPORTIONAL HAZARDS REGRESSION

We used a multivariable Cox PH regression to model the time and the incidence of stroke (see for instance Therneau,[9] SPSS 14.0). By adding the covariates to the model adjustment for confounding is achieved and an adjusted treatment effect is estimated. As the number of events per covariate was too low to use all covariates with this method, only the first and third set of covariates were used; a maximum of 10 events per covariate is advised in the literature.[10]

PROPENSITY SCORE METHODS

Achieving balance

With treatment as the dependent and the three different sets covariates we used logistic regression analysis to estimate the propensity score (SPSS 14.0). Some interactions and higher-order terms were added in order to improve the balance. In this model the number of 'events' (i.e. the lower of the number of treated and untreated patients) was sufficient to include these extra terms, in contrast to the multivariable Cox PH regression where the number of events (i.e. the number of strokes) is rather limited. Even when overfitting takes place in the propensity score model by a large number of terms, this is not of great concern, because it is not the intention to make inferential statements concerning the relationship between treatment and covariates. Instead we will focus on the balance of covariates between groups that will result when propensity score methods are used.

For a similar reason we did not check goodness-of-fit (GOF) or the discrimination of the propensity score model (as is frequently done by reporting the Hosmer-Lemeshow GOF or the area under the receiver operator characteristic curve or *c*-statistic): the issue is not to predict treatment or to estimate coefficients.[11,12] By adding interactions and higher-order terms to the propensity score model we selected only potential confounding factors, i.e. those terms that showed at least a moderate relationship with the outcome. By this strategy we clearly express that we focus on the problem of confounding and not on making the best predictive model for treatment. On the other hand, inclusion of some other terms or misspecification of the model does not seem to be of major concern.[13]

Checking balance on covariates

A check on the balance on covariates achieved by the propensity score is essential for this method, although not always done or reported in the literature.[14] To perform this check we used a stratified logistic regression analysis with treatment as the dependent, covariates as independents (LogXact 2.1) and with strata based on the quintiles of the PS (strata referred to as 'fifths'). We also applied the standard method where for every covariate and every stratum of the PS the difference between treatment groups is assessed and tested. We prefer the stratified multivariable method because many separate comparisons, having reduced power within strata, will then be avoided. Another reason is that balance should be checked conditional on other covariates, which can be achieved when using a stratified multivariable check. Ideally, within subclasses of the propensity score all covariates should be balanced and differences between treatment groups should disappear.

Estimating adjusted treatment effects

We estimated an adjusted treatment effect in three ways: stratification on the PS (1), matching on the PS (2) and using the PS as a covariate (3).

(1). Stratification on the PS was based on its quintiles. The resulting categorical variable was used in a Cox PH regression with stroke as the dependent and treatment as the only independent (S-Plus 6.2). The interaction between treatment and PS was tested in order to compare differences in treatment effect within strata.
(2). Matching on the PS was based on pair-matching. This means that for every treated subject only one control was selected. A greedy algorithm was used (SAS 8.0) and resulted in such pairs of subjects by randomly selecting a case and matching this to the control with the smallest difference in PS.[15] This process was continued until no more controls could be found that differed less than 0.1 in propensity score.
(3). The third estimation method is to use the PS as a continuous covariate in a Cox PH regression replacing all single covariates. Although this method has often been used in practice,[14] it is not recommended because too much weight is given to the absolute value of estimation of the PS. Another reason is that assumptions have to be made about the functional relationship between the PS and the outcome.[11] These three sets of covariates are combined with the four different adjustment methods, as is summarized in Table 4.1.

Table 4.1: Overview of different methods of analysis and different sets of covariates used in this paper

Method of analysis	Set 1: 8 covariates	Set 2: 12 covariates	Set 3: set 1 plus balanced covariate
Multivariable Cox PH regression	*	x	*
Cox PH regression, stratification on PS	*	*	*
Cox PH regression, matching on PS	*	*	*
Cox PH regression, PS as covariate	*	*	*

* = analysed

x = not analysed because of small number of events per covariate

Cox PH = Cox proportional hazards; PS = propensity score

RESULTS

ACTUAL IMBALANCE ON COVARIATES

In Table 4.2 the means or percentages of all covariates for both treatment groups are given, including the univariate test result on their differences. Most of the odds ratios are not close to 1 indicating imbalance on these covariates between groups. The covariates diabetes, family history of MI and total cholesterol are reasonably well balanced between treatment groups, whereas sex, previous TIA and previous CVD are clearly imbalanced between treatment groups.

BALANCE CREATING PROPERTIES OF THE PROPENSITY SCORE

As a first impression of the balance created by the PS we investigated the overlap in PS distributions between both treatment groups using the first set of covariates (Figure 4.1).

Table 4.2: Means or percentages of covariates for treated and untreated candidates for treatment, odds ratios and univariate significance tests

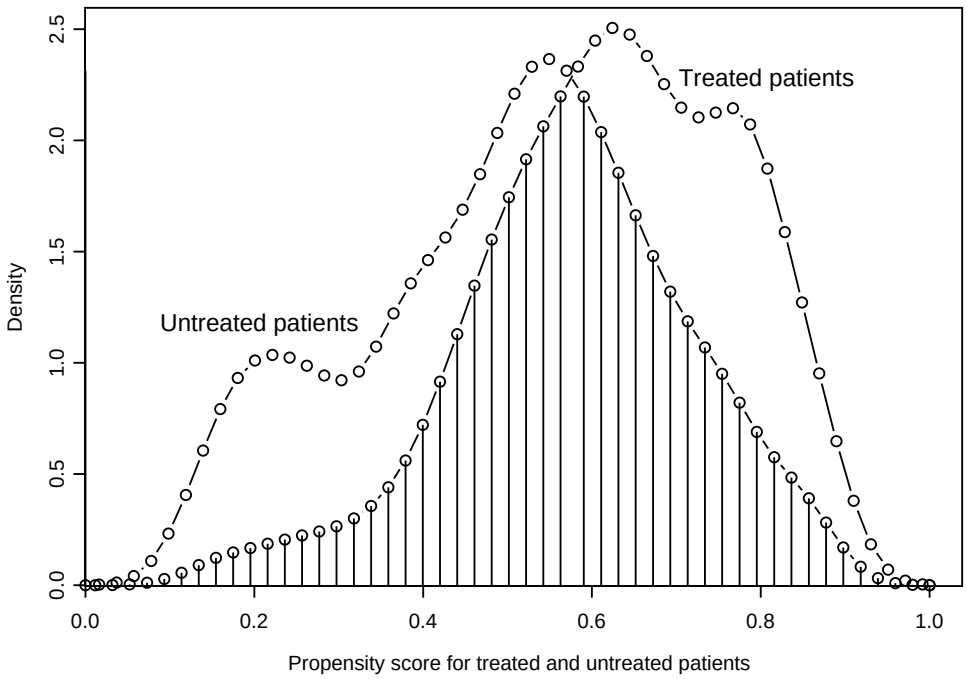
covariate	treated (n=1,293)	untreated (n=954)	odds ratio	95% confidence interval
History of CVA (%)	7.7	5.1	1.53	1.08, 2.18 *
Sex (% men)	34.3	47.5	0.58	0.49, 0.69 *
Smoking (%)	21.0	24.1	0.84	0.69, 1.02
Diabetes (%)	7.3	7.0	1.05	0.76, 1.45
Previous CVD (%)	24.1	17.1	1.54	1.24, 1.90 *
Previous MI (%)	11.4	9.2	1.26	0.96, 1.67
Previous TIA (%)	4.4	2.5	1.79	1.10, 2.90 *
Family history of MI (%)	10.5	9.4	1.14	0.87, 1.51
Age	65.1	65.8	1.00	0.99, 1.00
Body mass index	27.8	26.8	1.07	1.05, 1.09 *
Total cholesterol	6.56	6.61	0.97	0.90, 1.04
HDL-cholesterol	1.26	1.32	0.62	0.49, 0.79 *

* =different from an odds ratio of 1 at significance level 0.05, two-sided, using likelihood ratio test

CVA =cerebrovascular accident; CVD =cardiovascular disease; MI =myocard infarction;

TIA =transient ischemic attack

Figure 4.1: Distribution of the propensity score within both treatment groups



As could be expected, the untreated group tends to have lower scores: 18% of the untreated patients compared to only 3% of the treated patients have a probability of being treated of less than 0.30, whereas propensity scores higher than 0.70 are found for 13% of the untreated and for more than 35% of the treated patients. On the other hand there is considerable overlap: only 1.2% of the subjects have a propensity scores outside the range of the other group.

For a further check on the balance of the covariates and some interactions we used a stratified logistic regression with treatment as the dependent variable. The results are given in Table 4.3. Most of the odds ratios are near one and none reached a significance level of 0.10. The relatively low odds ratio (OR) of 0.57 for sex (with very large confidence interval) is mainly due to the inclusion of two interaction terms with sex, giving this coefficient a less straightforward interpretation (in a model without these interactions the OR for sex is 0.98).

Table 4.3: Check for balance between treatment groups on all covariates, stratified on the propensity score in a multivariable logistic regression

covariate	odds ratio	95% confidence interval	*p*-value*
History of CVA	1.09	0.60, 1.95	0.78
Sex	0.57	0.18, 1.81	0.34
Smoking	0.99	0.79, 1.24	0.93
Diabetes	1.02	0.72, 1.46	0.90
Previous CVD	1.08	0.83, 1.40	0.57
Age	1.05	0.98, 1.12	0.20
Body mass index	1.00	0.98, 1.03	0.74
Total cholesterol	0.98	0.91, 1.07	0.67
Age x Sex	1.01	0.99, 1.02	0.39
History of CVA x Sex	0.94	0.43, 2.08	0.88
Age squared	1.00	1.00, 1.00	0.17

* = from Wald test, two-sided

CVA = cerebrovascular accident; CVD = cardiovascular disease

The check for balance for sex is given in Table 4.4. All odds ratios within the five strata of the PS are non-significant and closer to one than the highly significant OR for the total sample.

Table 4.4: Check for balance between treatment groups on the covariate sex within fifths of the propensity score

	n	odds ratio	95% confidence interval	*p*-value*
Total sample	2,247	0.58	0.49, 0.69	0.00
1^{st} fifth of propensity score	449	1.03	0.66, 1.60	0.90
2^{nd} fifth of propensity score	450	1.19	0.82, 1.73	0.36
3^{rd} fifth of propensity score	449	0.82	0.55, 1.21	0.32
4^{th} fifth of propensity score	450	0.79	0.50, 1.25	0.32
5^{th} fifth of propensity score	449	1.04	0.58, 1.87	0.90

* = from likelihood ratio test, two-sided

n = number of observations

FIRST SET OF COVARIATES

The estimated hazard ratio (HR) in the Cox PH regression adjusted for the eight covariates was 0.64 with 95% confidence interval (CI 95%) from 0.42 to 0.98 (Table 4.5). When stratification on the PS was used a slightly smaller HR was found (0.58 versus 0.64), indicating a slightly larger treatment effect, estimated somewhat more precisely (CI95%: $0.38, 0.89$). The treatment effects within the five strata did not differ significantly from each other ($p = 0.89$). Matching on the PS leads to an even larger treatment effect (0.49), somewhat less precisely estimated mainly because of a reduced number of observations in the analysis. Using the PS as a covariate gives similar results as the multivariable Cox PH regression.

Table 4.5: Unadjusted treatment effects and adjusted effects with the first set of covariates* using multivariable Cox PH, stratification on the PS, matching on the PS and PS as covariate

Method of analysis	hazard ratio	95% confidence interval	*n*
Unadjusted	0.54	0.36, 0.82	2,246
Multivariable Cox PH regression	0.64	0.42, 0.98	2,134
Cox PH regression, stratification on PS	0.58	0.38, 0.89	2,136
Cox PH regression, matching on PS	0.49	0.27, 0.88	1,490
Cox PH regression, PS as covariate	0.64	0.41, 0.99	2,134

* = stratified on history of cerebrovascular accident, age and sex and further adjusted for age, diabetes, total cholesterol, body mass index, smoking and history of cardiovascular disease

Cox PH = Cox proportional hazard; PS = propensity score; *n* = number of observations

SECOND SET OF COVARIATES

In the second set an adjusted treatment effect is estimated for the three different PS methods when four covariates were added to the first set. Because of the low number of events per covariate a multivariable Cox PH regression was not performed. For all propensity score methods we found a similar downward shift in the hazard ratio of around 7% compared to the first set of covariates, as well as a smaller confidence interval (Table 4.6).

Table 4.6: Adjusted treatment effects with the second set of covariates* using stratification on the PS, matching on the PS and PS as covariate

Method of analysis	hazard ratio	95% confidence interval	*n*
Cox PH regression, stratification on PS	0.53	0.35, 0.83	2,037
Cox PH regression, matching on PS	0.45	0.25, 0.84	1,488
Cox PH regression, PS as covariate	0.57	0.36, 0.89	2,122

* = stratified on history of cerebrovascular accident, age and sex and further adjusted for age, diabetes, total cholesterol, body mass index, smoking, history of cardiovascular disease, previous myocard infarction, previous transient ischemic attack, family history of myocard infarction and HDL-cholesterol

Cox PH = Cox proportional hazard; PS = propensity score; *n* = number of observations

THIRD SET OF COVARIATES: FIRST SET PLUS BALANCED COVARIATE

In the third set a balanced covariate was added to the first set to check the sensitivity of the various methods to the inclusion of a non-confounder. In the multivariable Cox PH regression we found a large downward change in the hazard ratio (from 0.64 to 0.54), whereas stratification on the PS induced only a minor change in the treatment effect (Table 4.7). Also with covariate adjustment on the PS the change in treatment effect was small. Matching on the PS lead to an upward change in the treatment effect (from 0.49 to 0.57) and to a wider confidence interval.

Table 4.7: Adjusted treatment effects with the third set of covariates* using multivariable Cox PH, stratification on the PS, matching on the PS and PS as covariate

Method of analysis	% change in hazard ratio**	hazard ratio	95% confidence interval	*n*
Multivariable Cox PH regression	-15.8	0.54	0.35, 0.85	2,134
Cox PH regression, stratification on PS	-1.8	0.57	0.37, 0.87	2,136
Cox PH regression, matching on PS	+14.3	0.55	0.31, 0.97	1,536
Cox PH regression, PS as covariate	-4.3	0.59	0.38, 0.91	2,134

* = stratified on history of cerebrovascular accident, age and sex and adjusted for age, diabetes, total cholesterol, body mass index, smoking, history of cardiovascular disease and a simulated balanced covariate

** = percentage change in hazard ratio compared to the first model in which 8 covariates were used

Cox PH = Cox proportional hazard; PS = propensity score; *n* = number of observations

DISCUSSION

Three propensity score methods were compared with a multivariable Cox PH regression to estimate an adjusted effect of drug treatment for hypertension on the incidence of stroke. Matching and stratification on the PS gave a somewhat larger treatment effect than when a multivariable Cox PH regression was used or when the PS was used as covariate. Propensity score methods had the possibility to include more covariates (not performed in the multivariable Cox model), which gave a similar shift in the treatment effect in all propensity score methods. When a balanced covariate was added, the smallest change was found by stratification on the PS and when the PS was used as covariate; when the multivariable Cox PH regression or matching on the PS was used the change was large.

We contributed to the application of propensity score methods in medical science by giving a systematic comparison between these methods and a model based adjustment approach in a real life data set with many covariates and a relatively low number of events. Furthermore we pointed at the difficulties in finding the best propensity score model and in checking the balance between treatment groups. We also tested the sensitivity of the models against the addition of more covariates, including a balanced one.

In the medical literature application of propensity score methods is becoming more

widespread. In most studies only one of the methods has been used, whereas only some compare the results with a model-based approach. Because in most of these studies the same set of covariates was used in the PS together with covariate adjustment, the conclusion that 'no differences were found when a propensity score method was used' is not surprising. Often it is unclear how the model was created and whether the balance was sufficient.[14]

A recent systematic comparison of a propensity score method and multivariable logistic regression analysis with a low number of events can be found in Cepeda *et al.*[4] In a simulated data set the number of confounding variables, the strength of associations, the number of events and the strength of the exposure were varied. It was concluded that the estimation of the treatment effect by means of propensity scores was less biased, more robust and more precise than logistic regression when there were seven or fewer events per confounder. With more than seven events logistic regression analysis was recommended. Unfortunately they used only the known propensity score model, the one that was used for generating the data, so that the step of reaching balance could be skipped. Furthermore they used the propensity score only as categorical variable in the final analysis, where covariate adjustment or matching on the PS could have been used.

Our study has some limitations. First, the data set used was already to some extent balanced by choosing a control group that consisted of untreated candidates for treatment. A more general control group would produce less overlap in the distributions of covariates and could lead to larger differences between the methods. On the other hand, the more comparable the groups are, the more the differences in treatment effect can be contributed to the methods instead of the specific data set used. A second limitation is that only greedy pair-matching was used. Unfortunately a more optimal method could not be used because no large pool of controls was available.[16] To use five instead of another number of classes goes back to Cochran,[17] who stated that a 90% bias reduction is expected when stratifying was based on the quintiles.[18] Also 7 and 10 strata were used, but this didn't change the main results.

Furthermore, one can comment that the multivariable way of checking balance will leave questions whether this balance is sufficient and whether imbalances within strata exist. It can be shown that the balance on all these observed covariates is even better than could be expected in a randomized controlled trial (RCT). In a RCT it is expected that on average one in 10 of the terms is significant at the 0.10 levels, where in our model none was found. Of course, randomization takes care of all covariates, also the unobserved ones. We also checked the balance on all of the eight covariates separately within the five strata of the PS. We found that only one out of 40 comparisons turned out to be significant at the 0.10 level, where four are to be expected in case of randomization.

A last comment concerns the precision of the differences found between the different methods. No confidence intervals are given for these differences, so that it is unclear to what extent the results are sensitive for the specific sample.

Application of propensity score methods is not a straightforward task. There exist some practical difficulties in applying this intuitively appealing method. The first is the check for balance; a crucial step after a propensity score model has been made. There are no general rules available for the practical user how this check needs to be performed. We used a multivariable stratified analysis, but it remains unclear whether this is the best way to check balance. Another difficulty is when to stop adding interactions and higher-order terms to the propensity score model when an acceptable balance has not yet been reached. There is hardly any limit in the number of terms that can be added, because estimating coefficients is not an objective. Therefore measures of goodness-of-fit, area under the ROC and predictability of the model should not be used as a guideline. The PS model is meant to adjust for confounding, which means that terms are to be considered that have a relationship with treatment as well as the outcome. The relationship with treatment can be checked in the PS model itself (as usual in logistic regression analysis), but the relationship with the outcome should come from outside this model. In general only terms should be included in the PS model that have an empirical or logical relationship with the outcome, because otherwise effort is wasted in attempting to balance non-confounders. This contradicts the idea that the same PS model can be used for different outcome variables.

Concerning the sensitivity of the inclusion of a non-confounder, stratification on the PS (and covariate adjustment) performed better than matching and multivariable Cox PH. Matching on the PS seems also to be a rather sensitive method when there is a small number of events, like in our data set. It is recommended to perform propensity score methods, but special attention should be given to the PS model building and balance checking phases.

REFERENCES

[1] Cox DR. Regression models and life tables. *J Royal Stat Society Series B*, 34:187–220, 1972.

[2] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

[3] D'Agostino, RB Jr. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med*, 17:2265–2281, 1998.

[4] Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol*, 158:280–287, 2003.

[5] Petersen LA, Normand SL, Daley J, McNeil BJ. Outcome of myocardial infarction in Veterans Health Administration patients as compared with medicare patients. *N Engl J Med*, 343:1934–1941, 2000.

[6] Wijeysundera DN, Beattie WS, Rao V, Ivanov J, Karkouti K. Calcium antagonists are associated with reduced mortality after cardiac surgery: a propensity analysis. *J Thorac Cardiovasc Surg*, 127:755–762, 2004.

[7] Klungel OH, Stricker BH, Breteler MM, Seidell JC, Psaty BM, de Boer A. Is drug treatment of hypertension in clinical practice as effective as in randomized controlled trials with regard to the reduction of the incidence of stroke? *Epidemiology*, 12:339–344, 2001.

[8] Hauck WW, Anderson S, Marcus SM. Should we adjust for covariates in nonlinear regression analyses of randomized trials? *Control Clin Trials*, 19:249–256, 1998.

[9] Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York, 2000.

[10] Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *J Clin Epidemiol*, 48:1503–1510, 1995.

[11] Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Weaknesses of goodness-of-fit tests for evaluating propensity score models: the case of the omitted confounder. *Pharmacoepidemiol Drug Saf*, 14(4):227–238, 2005.

[12] Rubin DB. On principles for modeling propensity scores in medical research (Editorial). *Pharmacoepidemiol Drug Saf*, 13:855–857, 2004.

[13] Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, 49:1231–1236, 1993.

[14] Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiol Drug Saf*, 13(12):841–853, 2004.

[15] A SAS macro is available on http://www2.sas.com/proceedings/sugi26/p214-26.pdf.

[16] Rosenbaum PR. *Observational studies, 2nd edition*. Springer-Verlag, New York, 2002.

[17] Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24:295–313, 1968.

[18] Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *JAMA*, 387:516–524, 1984.

4.2 A NON-PARAMETRIC APPLICATION OF INSTRUMENTAL VARIABLES IN SURVIVAL ANALYSIS

Edwin P. Martens[a,b], Anthonius de Boer[a], Wiebe R. Pestman[b], Svetlana V. Belitser[a], Yves F.C. Smets[c], Rudi G.J. Westendorp[d] and Olaf H. Klungel[a]

[a] *Department of Pharmacoepidemiology and Pharmacotherapy, Utrecht Institute of Pharmaceutical Sciences (UIPS), Utrecht University, Utrecht, the Netherlands*
[b] *Centre for Biostatistics, Utrecht University, Utrecht, the Netherlands*
[c] *Onze Lieve Vrouwe Gasthuis, Amsterdam, The Netherlands*
[d] *Leiden University Medical Centre, Leiden, The Netherlands*

Submitted for publication

ABSTRACT

Background: The application of instrumental variables is not widespread in medical research with censored survival outcomes.
Objectives: To show how instrumental variables can be combined with survival analysis in a non-parametric way and to compare the estimated treatment effect with other estimators.
Design and methods: In a sample of 214 patients with type-1 diabetes who started renal-replacement therapy in the Netherlands ($1985-1996$), the effect of pancreas-kidney transplantation versus kidney transplantation alone on mortality was analyzed using hospital admission area as instrumental variable.
Results: The instrumental variables estimate of the difference in survival probabilities between pancreas-kidney and kidney changed from a non-significant -0.03 (95% CI: -0.19, 0.13) after 2 years of follow-up to a significant 0.42 (95% CI: 0.04, 0.80) after 6 years, favoring pancreas-kidney transplantation. This is substantially larger than the intention-to-treat estimate: after 6 years the difference was 0.15 (95% CI: 0.01, 0.29).
Conclusion: Instrumental variables can be an alternative method for estimating treatment effects in the presence of censored survival outcomes in observational studies with unobserved confounding. A suitable instrument, fulfilling the strong assumptions involved in instrumental variable estimation, should be present. It leads in general to a larger treatment effect than intention-to-treat, with wider confidence intervals.

Keywords: Instrumental variables; Survival analysis; Observational studies; Unobserved confounding; All-or-none compliance

INTRODUCTION

Well-conducted randomized controlled trials (RCTs) have been widely accepted as the scientific standard to estimate the effect of one treatment against another.[1] There are settings where a randomized comparison of treatments may not be feasible due to ethical, economic or other constraints. The main alternative is an observational study in which a randomized assignment of treatments is absent and in which confounding factors may provide an alternative explanation for the treatment effect.[2] To adjust for these confounding factors several methods have been proposed.[3,4] The more traditional model-based approaches of regression (*Cox proportional hazards regression* (Cox PH),[5] linear or logistic regression) and methods that are primary focussed on treatment probabilities (*propensity score methods and inverse probability weighting*[6–11]), all aim to adjust only for observed confounders. In contrast, methods that use *instrumental variables (IV)* aim to adjust for observed and unobserved confounders. Literature on the method of instrumental variables can be found in Angrist, Imbens and Rubin,[12] Baker and Lindeman[13] and Imbens and Angrist.[14] Originating from the economic literature, applications of these methods can be found in the medical literature.[15–25]

The claim in IV methods to adjust for unobserved confounders has an important drawback. The quality of the estimate is dependent on the existence of an instrumental variable or instrument that satisfies the strong assumptions underlying the method. The first assumption is that a substantial correlation exists between the instrument and the treatment variable.[26–28] The second assumption is that the relationship between the instrumental variable and the exposure is not confounded by other variables. This assumption is fulfilled when individuals are (or considered to be) randomly assigned to the different categories of the instrumental variable. The third and most important assumption is that the instrument will have only an effect on the outcome by means of the treatment variable of interest, and not directly nor indirectly by means of other variables. This assumption, known as the *exclusion restriction*, can not be tested but should be evaluated in the specific research situation.

Assuming there exists an acceptable instrument, application of the method is quite straightforward in a linear model with a continuous outcome or in a linear probability model when the outcome is dichotomous. For non-linear models in general several IV-estimators have been proposed.[29,30] Bijwaard and Ridder developed an IV-estimator which allows for censoring but assumes perfect compliance in the control group.[31] Baker[32] extended the work of Angrist et all.[12] to estimate life years saved using the difference in hazards. Robins developed several models for the analysis of time-varying exposures using G-estimation,[10,33,34] which can also be used for IV-estimation.[35–37] Abbring and Van den Berg[38] gave a non-parametric IV-estimator and its standard error for the difference in survival outcomes which allows for censoring in a context of social experiments.

The aim of this study is to show how this non-parametric method of Abbring and Van den Berg can be applied in a non-experimental, medical context. In the next section we will give

an overview of treatment effect estimators in general, where the third section provides the formulas needed for IV-estimation on survival data. In the fourth section the IV-method will be applied to a medical observational data set used in the research of Smets et al.[39] in which the effect of an additional pancreas transplantation has been estimated on the survival of patients with type-1 diabetes mellitus and end-stage renal failure. The IV-estimate will also be compared with other known estimators from clinical trials: the intention-to-treat, the per-protocol and the as-treated estimates.

TREATMENT EFFECT ESTIMATORS

Dependent on the type of outcome variable an effect estimator of a dichotomous treatment X on the outcome can be defined in several ways. For survival outcomes that are possibly censored, the mean survival time doesn't make sense because for censored observations survival time is unknown. This implies that treatment effects can only be defined for different points in time, except when additional assumptions are made. To analyze the time to an event that is possibly censored, different methods of survival analysis can be used.[40] To define treatment effects one should concentrate on so-called contrasts of the survival or hazard function of both treatment groups. Possible treatment effect estimators are the ratio or the difference of hazards, or the ratio or difference of survival probabilities, all evaluated at time t.

In IV estimation the important assumption should be made that the instrument will not influence directly nor indirectly the outcome (exclusion restriction). Suppose that this assumption has been fulfilled at $t = 0$, this will be in general not true for $t > 0$ because the subgroup of survivors will differ from the group at $t = 0$. When treatment effects are estimated on these subgroups of survivors (like the ratio or difference of hazards), these estimates will capture both the treatment effect of interest and a selection effect. We will use the difference of survival probabilities as the treatment effect, because these probabilities are based on the total group for all t. It can be shown that all other treatment effects will not represent any meaningful treatment effect when estimated in a non-parametric way.[38,41]

IV-ESTIMATOR WHEN SURVIVAL OUTCOMES ARE CENSORED

In a clinical trial with all-or-none compliance different types of analyses can be distinguished: *as-treated* (AT), *per-protocol* (PP) and *intention-to-treat* (ITT). In general it is common practice to perform an ITT analysis when non-compliance is present. The ITT-estimate can be seen as the effect of *policy or assignment*, a mixture of the actual effect of treatment and the effect of all-or-none compliance. With a focus on the difference of survival probabilities as the

treatment effect, these estimators can be written as follows:

$$\widehat{\Delta}_{AT}(t) = \widehat{\overline{F}}_{X=1}(t) - \widehat{\overline{F}}_{X=0}(t) \tag{4.1}$$

$$\widehat{\Delta}_{PP}(t) = \widehat{\overline{F}}_{Z=1,X=1}(t) - \widehat{\overline{F}}_{Z=0,X=0}(t) \tag{4.2}$$

$$\widehat{\Delta}_{ITT}(t) = \widehat{\overline{F}}_{Z=1}(t) - \widehat{\overline{F}}_{Z=0}(t) \tag{4.3}$$

where $\widehat{\Delta}_{AT}(t)$, $\widehat{\Delta}_{PP}(t)$, $\widehat{\Delta}_{ITT}(t)$ is the estimated treatment effect for a censored survival outcome at time t in a AT, PP, ITT analysis, $X \in (0,1)$ is treatment, $Z \in (0,1)$ is assignment to treatment, $\widehat{\overline{F}}_{X=x} = \widehat{\Pr}(T > t|X = x)$ equals the Kaplan-Meier estimate for the survival function for $X = x$[40,42] and analogously for $\widehat{\overline{F}}_{Z=z}$.

IV POINT ESTIMATOR

Another estimator is known as the instrumental variables (IV) estimator. This estimator gives an estimate of the average effect of treating instead of not treating a certain population and adjusts for observed and unobserved confounding. This comes at the cost of making assumptions, which can be quite strong in some situations. The three assumptions that define the instrumental variable has already been mentioned. A further assumption has to be made to identify the estimator. For that reason we assume that there exists *monotonicity*, which implies in the case of dichotomous treatment and dichotomous IV that there are no *defiers*. Defiers are persons who always do the opposite of their assignment and can not be identified empirically.[12]

The IV estimator can be used when survival outcomes are censored in clinical trials with all-or-none compliance, where the instrumental variable is simply the original assignment. This method can also be applied in observational studies, but then a suitable instrument should be found. The non-parametric IV-estimator in case of a dichotomous treatment and dichotomous instrumental variable, can be written as[38]

$$\widehat{\Delta}_{IV}(t) = \frac{\widehat{\overline{F}}_{Z=1}(t) - \widehat{\overline{F}}_{Z=0}(t)}{\widehat{\Pr}(X = 1|Z = 1) - \widehat{\Pr}(X = 1|Z = 0)} \tag{4.4}$$

where $\widehat{\Pr}(X = 1|Z = z)$ is the estimated probability of being treated for $Z = z$ at $t = 0$. In short, the numerator equals the ITT-estimate of the survival difference between $Z = 1$ and $Z = 0$ and the denominator equals the difference in treatment probabilities between $Z = 1$ and $Z = 0$. This structure is similar to the IV-estimator for a linear probability model with binary outcome Y[12]

$$\widehat{\beta}_{IV} = \frac{\widehat{\Pr}(Y = 1|Z = 1) - \widehat{\Pr}(Y = 1|Z = 0)}{\widehat{\Pr}(X = 1|Z = 1) - \widehat{\Pr}(X = 1|Z = 0)} \tag{4.5}$$

IV VARIANCE ESTIMATOR

The variance of the non-parametric IV-estimator $\widehat{\Delta}_{IV}(t)$ asymptotically equals[38]

$$var\big[\widehat{\Delta}_{IV}(t)\big] = \frac{1}{(p_1-p_0)^2}\Big\{\frac{p_1(1-p_1)}{n_1}\big[\overline{F}_{11}(t)-\overline{F}_{01}(t)-\Delta_{iv}(t)\big]^2+ \frac{p_0(1-p_0)}{n_0}\big[\overline{F}_{10}(t)-\overline{F}_{00}(t)-\Delta_{iv}(t)\big]^2+ p_1^2\sigma_{11}^2(t)+(1-p_1)^2\sigma_{01}^2(t)+ p_0^2\sigma_{10}^2(t)+(1-p_0)^2\sigma_{00}^2(t)\Big\} \quad (4.6)$$

where $p_z = \Pr(\mathrm{X}=1|\mathrm{Z}=\mathrm{z})$, n_z is the number of observations for $Z=z$ at $t=0$, $\overline{F}_{xz}(t) \equiv \overline{F}_{X=x;Z=z}(t) = \widehat{\Pr}(T>t|X=x,Z=z)$, $\sigma_{xz}^2(t) \equiv \sigma_{X=x,Z=z}^2(t)$ which is the variance for the survival function at time t for $X=x$ and $Z=z$ obtained by Greenwood's formula.[43]

The variance $var\big[\hat{\Delta}_{IV}(t)\big]$ can be consistently estimated by appropriate sample estimates for p_z, $\overline{F}_{xz}$ and σ_{xz}^2 and the number of observations n_z. To find these quantities one should calculate the Kaplan-Meier estimate for four subgroups defined by the treatment $(0,1)$ and the instrumental variable $(0,1)$.

CONFIDENCE INTERVALS FOR THE DIFFERENCE IN SURVIVAL CURVES

It has been proven that the Kaplan-Meier estimator is asymptotically normally distributed,[44] which also means that the difference between two independent estimates is normally distributed. A pointwise 95% confidence interval for the IV estimator can be obtained in the usual way

$$\Delta_{IV}(t) = \widehat{\Delta}_{IV}(t) \pm 1.96\sqrt{var\big[\hat{\Delta}_{IV}(t)\big]} \quad (4.7)$$

Apart from this pointwise confidence interval, a simultaneous confidence band for the difference of two survival curves has been proposed by Parzen et al.,[45] which can be seen as an extension of the one-sample Hall-Wellner type confidence band.[46] More recently, simultaneous confidence bands have been developed using the empirical likelihood method, which seems to be an improvement in small samples.[47,48] We restrict ourselves to pointwise confidence intervals based on the normal approximation and validated these intervals in bootstrap samples.

APPLICATION OF THE METHOD

DATA SET

For an application of the method we used a data set from the renal replacement registry in the Netherlands (RENINE).[49] This data set consists of 415 patients with type-1 diabetes who started renal-replacement therapy in the Netherlands between 1985 and 1996. All patients were followed up to July, 1997. The objective of the research of Smets et al.[39] was to assess the impact on survival of an additional pancreas transplantation next to a kidney transplantation. Because it was expected that a direct comparison of the treatments kidney and pancreas-kidney (by means of an AT-estimate) was strongly confounded by unobserved factors, the researchers performed an ITT-analysis by comparing two areas, Leiden versus other areas. In the Leiden-area the primary intention to treat was a simultaneous pancreas-kidney transplantation, whereas in the other areas kidney transplantation alone was the predominant type of treatment. Of all transplanted patients 73% in Leiden and 37% in the other areas received the simultaneous pancreas-kidney transplantation (see Table 4.8). The incidence of renal-replacement therapy was quite similar in both areas, as was the initiation of renal-replacement therapy limited to dialysis. The age and sex distributions for all patients did not differ significantly.[39] Because of the strict allocation of patients to a center and these similarities, the authors considered it as unlikely that patients would differ markedly between the areas with respect to factors influencing survival.[39]

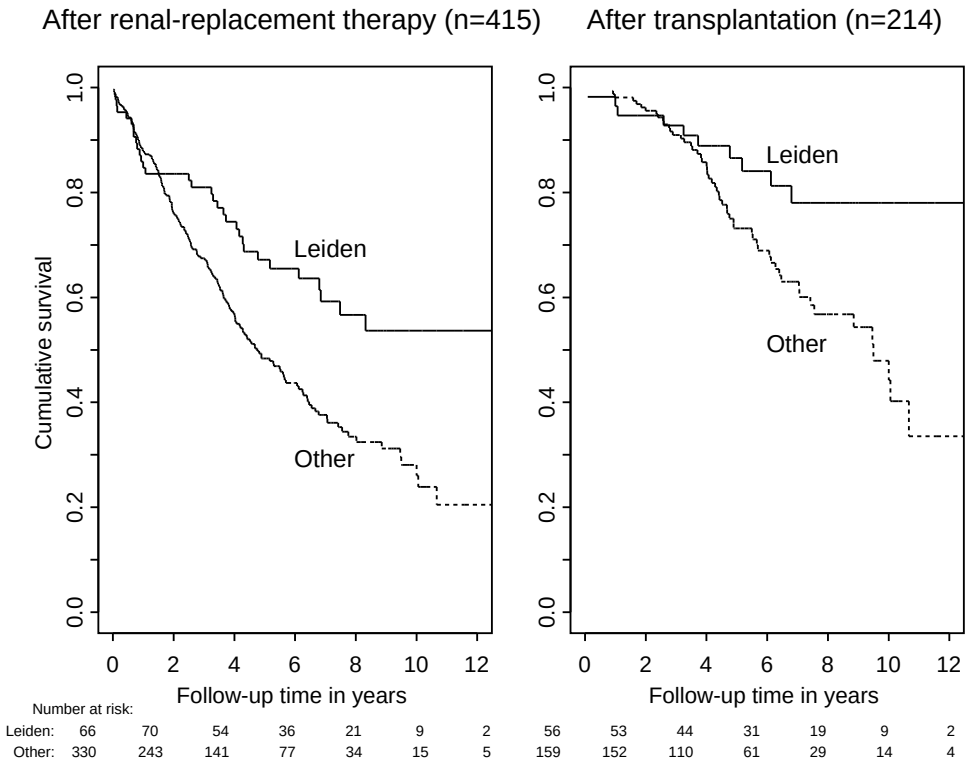
Table 4.8: Patient characteristics for Leiden and other areas

	Leiden (n=85)	Other areas (n=330)
Mean age	40.2	41.5
Male patients	47 (55%)	205 (62%)
Number of patients with dialysis only	29 (34%)	172 (52%)
Number of transplants	56 (66%)	158 (48%)
Pancreas-kidney transplant	41 (73%)	59 (37%)
Kidney transplant alone	15 (27%)	99 (63%)
Number of deaths in transplanted patients	10 (18%)	55 (35%)
Pancreas-kidney transplant	7 (17%)	24 (41%)
Kidney transplant alone	3 (20%)	31 (31%)

In Figure 4.2 the survival curves of all patients who started renal-replacement therapy (left panel) and transplanted patients (right panel) are presented for Leiden and other areas. These curves are estimated by the Kaplan-Meier method. As has been stated in Smets et al.[39] the survival for all patients who started renal-replacement therapy was significantly higher in the Leiden area than in the other areas (log rank test, $p < 0.001$, unadjusted hazard ratio 0.53, 95% CI $0.36, 0.77$). For transplanted patients (right panel) a similar result was found (log rank test, $p = 0.008$, unadjusted hazard ratio, 0.41, 95% CI $0.21, 0.81$), although the difference in the earlier years is somewhat smaller and in the later years somewhat larger than for all patients. Because the treatment effect of interest concerns transplantation methods, we will

further restrict ourselves to the group of transplanted patients: in Leiden 56 and in the other areas 158 patients. Although the overall proportion of transplants in Leiden was higher than in the other areas (66% versus 48%), this did not result in selection bias because no difference in survival was found for all patients who were on dialysis only (log-rank test, $p = 0.94$).[39]

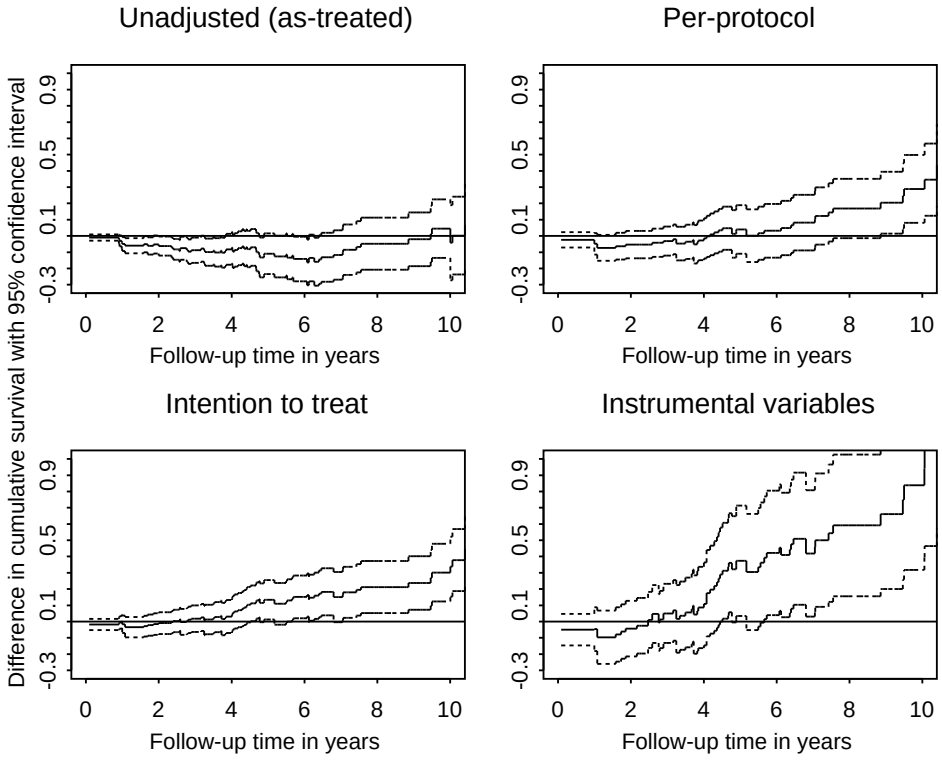
Figure 4.2: Patient survival after start of renal-replacement therapy and after transplantation, Leiden versus other areas



AS-TREATED, PER-PROTOCOL, INTENTION-TO-TREAT ANALYSIS

An analysis in which only treatments are compared without any adjustment (AT-analysis in clinical trials), gives until 6 years of follow-up a marginally significant result in favor of the kidney transplantation method and after 6 years a non-significant effect (see panel 1 Figure 4.3). As Smets et al. argue, this estimate will be clearly biased because of selection; patients for simultaneous pancreas-kidney transplantation are selected on good as well as poor health indicators.[39] The PP estimate shows a treatment effect in favor of the pancreas-kidney transplantation after 4 years, which become significant only after 9 years of follow-up. In the ITT analysis we found after 6 years of follow-up a significant difference in survival probabilities between Leiden and other areas of 15% (95% CI: $0.01, 0.29$), favoring the pancreas-kidney transplantation method.

Figure 4.3: Treatment effect as the difference in survival probabilities (Leiden minus other areas) for various methods of analysis and 95% confidence interval

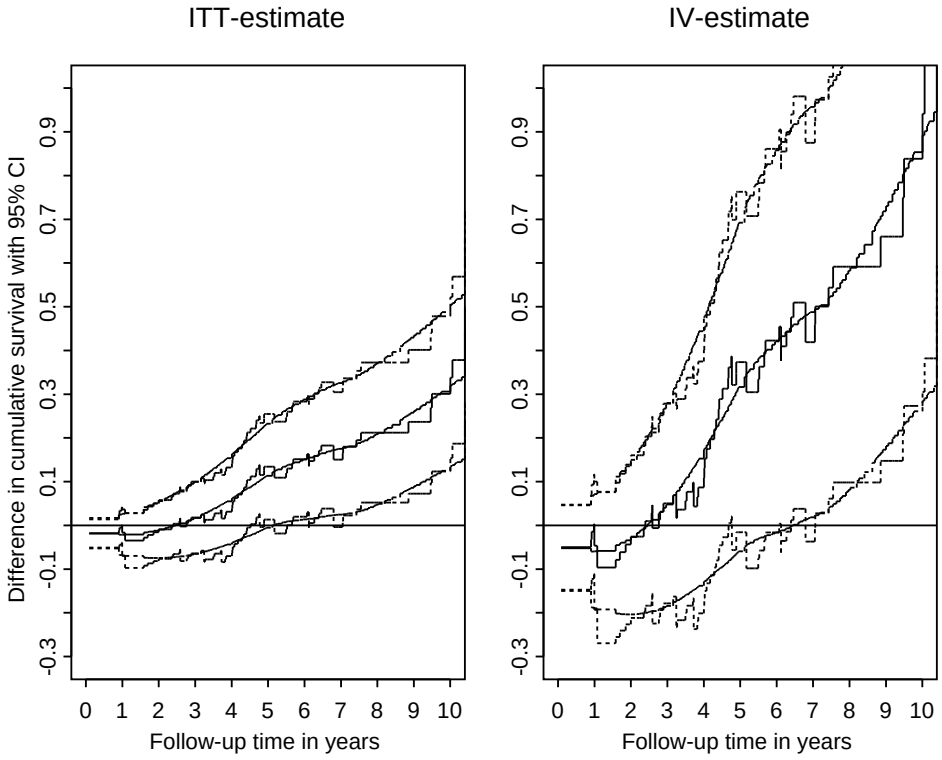


INSTRUMENTAL VARIABLES APPROACH

In order to disentangle the *effect of transplantation policy* (that is common in that center) and the *effect of actual treatment* (pancreas-kidney versus kidney), we apply the method of IV. As the instrument we use the dichotomous variable indicating the area: Leiden versus other areas. The assumptions justifying its use as an instrument, are probably fulfilled. First, there is a substantial effect of area on treatment i.e. a difference of $36\%(= 73\% - 37\%)$. The associated F-statistic from the regression of treatment on instrumental variable is 23.5, much larger than the proposed minimum of 10.[26] Second, the distribution of patient characteristics between the areas can be considered to be similar, because patients were allocated to their treatment center by place of residence.[39] Third, it is unlikely that the instrument, the area in which the transplantation took place, will have a direct or indirect effect on the outcome. Furthermore, we assume monotonicity[12] which implies here that the pancreas-kidney transplants in the other areas would also have resulted in pancreas-kidney transplants when these patients were living in Leiden.

In Figure 4.4 the result of the IV analysis is shown. To calculate point estimates, we used formula 4.5, whereas for the construction of a 95% confidence interval formula 4.6 was used. The IV analysis is compared with the ITT result for transplanted patients. We also used local regression method (LOESS) to fit a somewhat smoother curve through the data.[50] We restricted the comparative analysis to a follow-up period of 10 years, because for a longer period the estimates become unreliable.

Figure 4.4: Original and smoothed differences in survival robabilities after transplantation (Leiden minus other areas), ITT-estimate versus IV-estimate and 95% confidence interval



The difference in survival for the IV analysis is much larger than for the ITT analysis, indicating that the ITT-estimate underestimates the advantage of treating patients with a pancreas-kidney transplantation. The larger confidence intervals for the IV-estimate indicates that more uncertainty exists around these parameters. In Table 4.9 these differences are summarized, including the estimates of the AT and PP analyses. After 6 years of follow-up, the ITT point estimate is 0.15, whereas the IV estimate is 0.42 in favor of the pancreas-kidney transplantation. The associated confidence interval for the IV analysis is approximately 3 times as large, covering a large part of the outcome space. Both methods lead to a similar pattern of significance across the years; after 5 to 6 years of follow-up a significant difference was reached between these two methods. The results for the PP analysis are quite different with small and insignificant

effects in the first 8 years.

Table 4.9: Estimated treatment effects (95% CI) expressed as survival probability differences (Leiden minus other areas in ITT analysis, pancreas-kidney minus kidney in AT, PP and IV analysis)

	AT analysis	PP analysis	ITT analysis	IV analysis
2 years	-0.06 (-0.12,-0.01)	-0.05 (-0.13, 0.03)	-0.01 (-0.07, 0.06)	-0.03 (-0.19, 0.13)
4 years	-0.09 (-0.19, 0.01)	-0.01 (-0.13, 0.11)	0.06 (-0.04, 0.16)	0.17 (-0.09, 0.43)
6 years	-0.13 (-0.27, 0.01)	0.05 (-0.11, 0.22)	0.15 (0.01, 0.29)	0.42 (0.04, 0.80)
8 years	-0.07 (-0.23, 0.09)	0.16 (-0.03, 0.34)	0.21 (0.05, 0.37)	0.58 (0.14, 1.00)
10 years	0.01 (-0.20, 0.21)	0.30 (0.09, 0.51)	0.32 (0.14, 0.50)	0.89 (0.35, 1.00)

STANDARD ERRORS IN INSTRUMENTAL VARIABLES APPROACH

The standard errors and associated confidence intervals of the IV estimates are fairly large. To verify these asymptotic quantities we performed a bootstrap procedure: standard errors and confidence intervals only differed by less than 3%.

The influence of the sample size on the width of the confidence interval found in IV analysis can be approximated by $\frac{1}{\sqrt{k}}$, which for instance means that the width decreases by 30% when sample size is doubled ($k = 2$). More events or a more equal distribution of patients between Leiden and other areas, reduces the interval width of the IV estimates even more.

We investigated the influence of the 'compliance rates' in both areas on standard errors and interval width. By weighing the data set we changed the original difference in pancreas-kidney transplants between the areas 36% to 60% (80% in Leiden and 20% in other areas). As can be expected, the influence on standard errors is fairly large, reducing the width of the confidence intervals by approximately 45%. Note that in case of full compliance in both areas (in Leiden 100% pancreas-kidney and in other areas 100% kidney) the IV-estimator and its confidence intervals coincides with those from the ITT-estimator.

DISCUSSION AND CONCLUSION

The method of instrumental variables finds its way into medical literature as an adjustment method for unobserved confounding in observational studies and in randomized studies with all-or-none compliance. This method can also be applied in survival analysis with censored outcomes by using the difference in survival probabilities as the treatment effect of interest.[38] As an example we used a data set to analyze the beneficial effect of an additional pancreas transplantation for patients with type-1 diabetes who started renal-replacement therapy in the Netherlands between 1985 and 1996. We conclude that the additional pancreas transplantation has a significant positive effect on survival. The 6 year difference in survival probabilities between the two transplantation methods, adjusted for observed and unobserved confounding, was 0.42 (95% CI: $0.04, 0.80$) in favor of the pancreas-kidney transplantation. Compared to the intention-to-treat estimate of 0.15 (95% CI: $0.01, 0.29$) this is substantially larger, which indicates that the comparison of policies (ITT-estimate) dilutes the differences between both

treatment methods. A direct comparison of these treatment methods, as has been given by the as-treated estimate, can be considered as clearly biased because of selection processes.

As is inherent to IV-analysis the estimate is quite sensitive to the underlying assumptions of the method. It could be argued that the main assumption (i.e. the instrument has no direct nor indirect influence on the outcome) has not been fulfilled in this data set. This could be for example the capability of specialists or the quality of equipment. It has been shown by Smets et al.[39] that graft survival, an outcome parameter closer related to the performance of the transplantation than overall survival, was fairly identical between the two areas. Another indication on the fulfillment of this assumption is the similarity of patient survival during dialysis between the two areas.

It could also be argued that the variable used as the instrument was not randomized, violating our second assumption of IV estimation. Although patients were indeed not randomized over areas, the treatment allocation was determined by place of residence of the patient. Therefore, it is unlikely that the possible difference of patient characteristics with type-1 diabetes between both areas is large enough to cause substantial bias. Overall survival of patients on dialysis only did not differ between areas. Also age and sex differences between these areas turned out to be fairly similar.

From all type-1 diabetes patients arriving at the hospital, not all received a transplant and if so, a certain time elapsed before the actual transplantation took place. It is therefore possible that difference in patient survival could be partly explained either by the percentage of pre-emptive transplants or by a shorter time on dialysis before transplantation. The Leiden area differed from the other areas in a higher percentage of pre-emptive transplantations and a shorter duration of dialysis before transplantation. It is not likely that this will explain a large portion of the difference, because the hazard ratio remained identical when all pre-emptive transplant recipients were excluded and duration as a covariate did not change the results.

In the data set that was used to illustrate the method, large confidence intervals were calculated for the IV analysis. Although in IV analysis intervals are usually large, in this data set it is mainly due to the small number of transplanted patients in one of the areas (56 in Leiden), the associated small number of deaths (10) and the limited overall sample size ($n = 214$).

The difference in survival probabilities is naturally restricted to the interval $[-1, 1]$. The way the IV estimate has been calculated does not ensure the estimate to fall within this interval. In our data set we faced this problem at the end of the follow-up period, where the confidence intervals are widest. We restricted therefore the analysis to a follow-up time of 10 years, leaving out the least reliable estimates.

We conclude that this non-parametric method can easily combine IV-estimation with survival analysis in observational data to estimate a treatment effect that is adjusted for unobserved confounding or in randomized studies with all-or-none compliance. Confidence intervals of these estimates can be large, mainly at the end of the survival curve. As in all IV applications, careful attention should be paid to the fulfillment of the assumptions.

REFERENCES

[1] Friedman LM, Furberg CD, DeMets DL. *Fundamentals of Clinical Trials*. St Louis: Mosby-Year Book, 1996.

[2] Vandenbroucke JP. When are observational studies as credible as randomised trials? *Lancet*, 363:17281731, 2004.

[3] McMahon AD. Approaches to combat with confounding by indication in observational studies of intended drug effects. *Pharmacoepidemiol Drug Saf*, 12:551–558, 2003.

[4] Klungel OH, Martens EP, Psaty BM, *et al*. Methods to assess intended effects of drug treatment in observational studies are reviewed. *J Clin Epidemiol*, 57:1223–1231, 2004.

[5] Cox DR. Regression models and life tables. *J Royal Stat Society Series B*, 34:187–220, 1972.

[6] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

[7] Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *JAMA*, 387:516–524, 1984.

[8] D'Agostino, RB Jr. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med*, 17:2265–2281, 1998.

[9] Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiol Drug Saf*, 13(12):841–853, 2004.

[10] Robins JM. Marginal structural models. *Proceedings of the section on Bayesian statistical science, American Statistical Association*, pages 1–10, 1998.

[11] Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health*, 60:578–586, 2006.

[12] Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *JASA*, 91:444–455, 1996.

[13] Baker SG, Lindeman KS. The paired availability design: a proposal for evaluating epidural analgesia during labor. *Stat Med*, 13:2269–2278, 1994. Correction: 1995;14:1841.

[14] Imbens G, Angrist J. Identification and estimation of local average treatment effects. *Econometrica*, 62:467–476, 1994.

[15] Greenland S. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol*, 29:722–729, 2000.

[16] Zohoori N, Savitz DA. Econometric approaches to epidemiologic data: relating endogeneity and unobserved heterogeneity to confounding. *Ann Epidemiol*, 7:251–257, 1997. Erratum in: Ann Epidemiol 7:431, 1997.

[17] Permutt Th, Hebel JR. Simultaneous-equation estimation in a clinical trial of the effect of smoking on birth weight. *Biometrics*, 45:619–622, 1989.

[18] Beck CA, Penrod J, Gyorkos TW, Shapiro S, Pilote L. Does aggressive care following acute myocardial infarction reduce mortality? Analysis with instrumental variables to compare effectiveness in Canadian and United States patient populations. *Health Serv Res*, 38:1423–1440, 2003.

[19] Brooks JM, Chrischilles EA, Scott SD, Chen-Hardee SS. Was breast conserving surgery underutilized for early stage breast cancer? Instrumental variables evidence for stage II patients from Iowa. *Health Serv Res*, 38:1385–1402, 2003. Erratum in: Health Serv Res 2004;39(3):693.

[20] Earle CC, Tsai JS, Gelber RD, Weinstein MC, Neumann PJ, Weeks JC. Effectiveness of chemotherapy for advanced lung cancer in the elderly: instrumental variable and propensity analysis. *J Clin Oncol*, 19:1064–1070, 2001.

[21] Hadley J, Polsky D, Mandelblatt JS, *et al*. An exploratory instrumental variable analysis of the outcomes of localized breast cancer treatments in a medicare population. *Health Econ*, 12:171–186, 2003.

[22] Leigh JP, Schembri M. Instrumental variables technique: cigarette price provided better estimate of effects of smoking on SF-12. *J Clin Epidemiol*, 57:284–293, 2004.

[23] McClellan M, McNeil BJ, Newhouse JP. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *JAMA*, 272:859–866, 1994.

[24] McIntosh MW. Instrumental variables when evaluating screening trials: estimating the benefit of detecting cancer by screening. *Stat Med*, 18:2775–2794, 1999.

[25] Brookhart MA, Wang PS, Solomon DH, Schneeweiss S. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiol*, 17:268275, 2006.

[26] Staiger D, Stock JH. Instrumental variables regression with weak instruments. *Econometrica*, 65:557–586, 1997.

[27] Bound J, Jaeger DA, Baker RM. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *JASA*, 90:443–450, 1995.

[28] Martens EP, de Boer A, Pestman WR, Belitser SV, Klungel OH. Instrumental variables: application and limitations. *Epidemiology*, 17:260–267, 2006.

[29] Amemiya T. The nonlinear two-stage least-squares estimator. *Journal of econometrics*, 2:105–110, 1974.

[30] Bowden RJ, Turkington DA. A comparative study of instrumental variables estimators for nonlinear simultaneous models. *J Am Stat Ass*, 76:988–995, 1981.

[31] Bijwaard G, Ridder G. Correcting for selective compliance in a re-employment bonus experiment. *Journal of Econometrics*, 125:77–111, 2004.

[32] Baker SG. Analysis of survival data from a randomized trial with all-or-none compliance: estimating the cost-effectiveness of a cancer screening program. *JASA*, 93:929–934, 1998.

[33] Robins JM. The analysis of randomized and nonrandomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. *Health Service Research Methodology: a focus on AIDS*, pages 113–159, 1989.

[34] Mark SD, Robins JM. Estimating the causal effect of smoking cessation in the presence of confounding factors using a rank preserving structural failure time model. *Stat Med*, 12:1605–1628, 1993.

[35] Joffe , Brensinger C. Weighting in instrumental variables and G-estimation. *Stat Med*, 22:12851303, 2003.

[36] Robins JM, Tsiatis AA. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Commun Statist -Theory Meth*, 20(8):2609–2631, 1991.

[37] Robins JM. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics - Theory and Methods*, 23:2379–2412, 1994.

[38] Abbring JH, Van den Berg GJ. Social experiments and instrumental variables with duration outcomes. *Tinbergen Institute Discussion Papers*, 05-047/3, 2005. http://www.tinbergen.nl/discussionpapers/05047.pdf.

[39] Smets YFC, Westendorp RGJ, van der Pijl JW, de Charro FTh, Ringers J, de Fijter JW, Lemkes HHPJ. Effect of simultaneous pancres-kidney transplantation on mortality of patients with type-1 diabetes mellitus and end-stage renal failure. *Lancet*, 353:1915–1919, 1999.

[40] Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York, 2000.

[41] Ham JC, LaLonde RJ. The effect of sample selection andinitial conditions in duration models: Evidence from experimental data on training. *Econometrica*, 64:175–205, 1996.

[42] Hosmer DW, Lemeshow S. *Applied Survival Analysis*. Wiley Interscience, 1999.

[43] Greenwood M. The natural duration of cancer. *Reports on Public Health and Medical Subjects*, 33:1–26, 1926.

[44] Fleming TR, Harrington DP. *Counting Processes and Survival Analysis*. John Wiley and Sons, 1991.

[45] Parzen MI, Wie LJ, Ying Z. Simultaneous confidence intervals for the difference of two survival functions. *Scand J Statist*, 24:309–314, 1997.

[46] Hall WJ, Wellner JA. Confidence bands for a survival curve from censored data. *Biometrika*, 67:133–143, 1980.

[47] Shen J, He S. Empirical likelihood for the difference of two survival functions under right censorship. *Statistics and Probability Letters*, 76:169–181, 2006.

[48] McKeague IW, Zhao Y. Comparing distribution functions via empirical likelihood. *Int J Biostat*, 1:1–18, 2005.

[49] de Charro Fth, Ramsteyn PG. Renine, a relational registry. *Nephrol Dial Transplant*, 10:436–441, 1995.

[50] Cleveland WS, Devlin SJ, Grosse E. Regression by local fitting. *J Econometr*, 37:87–114, 1988.

CHAPTER 5

IMPROVEMENT OF PROPENSITY SCORE METHODS

5.1 THE USE OF THE OVERLAPPING COEFFICIENT IN PROPENSITY SCORE METHODS

Edwin P. Martens[a,b], Wiebe R. Pestman[b], Anthonius de Boer[a], Svetlana V. Belitser[a] and Olaf H. Klungel[a]

[a] *Department of Pharmacoepidemiology and Pharmacotherapy, Utrecht Institute of Pharmaceutical Sciences (UIPS), Utrecht University, Utrecht, the Netherlands*
[b] *Centre for Biostatistics, Utrecht University, Utrecht, the Netherlands*

Provisionally accepted by American Journal of Epidemiology

ABSTRACT

Propensity score methods focus on balancing confounders between groups to estimate an adjusted treatment or exposure effect. However, there is a lack of attention in actually measuring and reporting balance and also in using balance for model selection. We propose to use the overlapping coefficient in propensity score methods. First to report achieved balance on covariates and second to use it as an aid in selecting a propensity score model.

We demonstrated how the overlapping coefficient can be estimated in practical settings and performed simulation studies to estimate the association between the weighted average overlapping coefficient and the amount of bias. For various incidence rates and strengths of treatment effect we found an inverse relationship between the overlapping coefficient and bias, strongly increasing with sample size. For samples of 400 observations Pearson's correlation was only -0.10, while for $6,000$ observations -0.83 was found. Mainly for large samples the overlapping coefficient can be used as a model selection tool because its value is predictive for the amount of bias. For smaller data sets other methods can better be used to help selecting propensity score models, although an absolute quantification of balance will not be given by these methods.

Keywords: Confounding; Propensity scores; Observational studies; Measures for balance; Overlapping coefficient

INTRODUCTION

A commonly used statistical method to assess treatment effects in observational studies, is the method of propensity scores (PS).[1,2] PS methods focus on creating balance on covariates between treatment groups by first creating a PS model to estimate the conditional probability to be treated given the covariates (the propensity score). In the second step an adjusted treatment effect is estimated, using the propensity score as matching variable, as stratification variable, as continuous covariate or inverse probability weight. Traditionally, in the literature on PS methods there has been more attention for the second step (how to use the PS) than for the first (creating balance with the PS model). This is confirmed in recent literature reviews, where a lack of attention to building the PS model has been noticed.[3–5] Building such a PS model involves the (theoretical) selection of potential confounders and possibly transformations of these variables, higher-order terms or interactions with other covariates to include in the model. Because the objective of the PS model is to balance treatment groups on covariates and not to find the best estimates of coefficients, a check on balance on important prognostic covariates is important.[1,6] Unlike prediction models the selection of variables for a PS model (in which treatment is the dependent variable) is more complex: both the relationship with treatment and outcome has to be taken into account. In a recent study on variable selection it was confirmed that the PS model should contain variables related to both treatment and outcome and that it is better *not* to include variables that are only related to treatment because this will increase the standard error of the estimate.[7] Any stepwise regression method to build the PS model only selects on significance of the relationship with treatment, but does not use information on the strength of the relationship with the outcome. A strong relationship between treatment and covariates is not necessary for having a good PS model or good balance.[6,8] Such an example would be a PS model in a randomized trial: there will be a weak association between treatment and covariates, but still good balance exists.

In applications of PS modeling the balance on covariates that has been reached has not been reported frequently or systematically.[4,5,9] When information on balance is given, this is mostly done by performing significance tests within strata of the PS to assure that the mean (or proportion) on covariates do not differ significantly between treatment groups. An early method proposed by Rosenbaum & Rubin to check the balance per covariate using the F-statistic from analysis of variance (ANOVA) is not often used.[10] More frequently the *c*-statistic (area under the receiver operating curve) is reported, but does not give the information needed: also a low value of the *c*-statistic can indicate good balance on important covariates (for instance in a randomized trial).

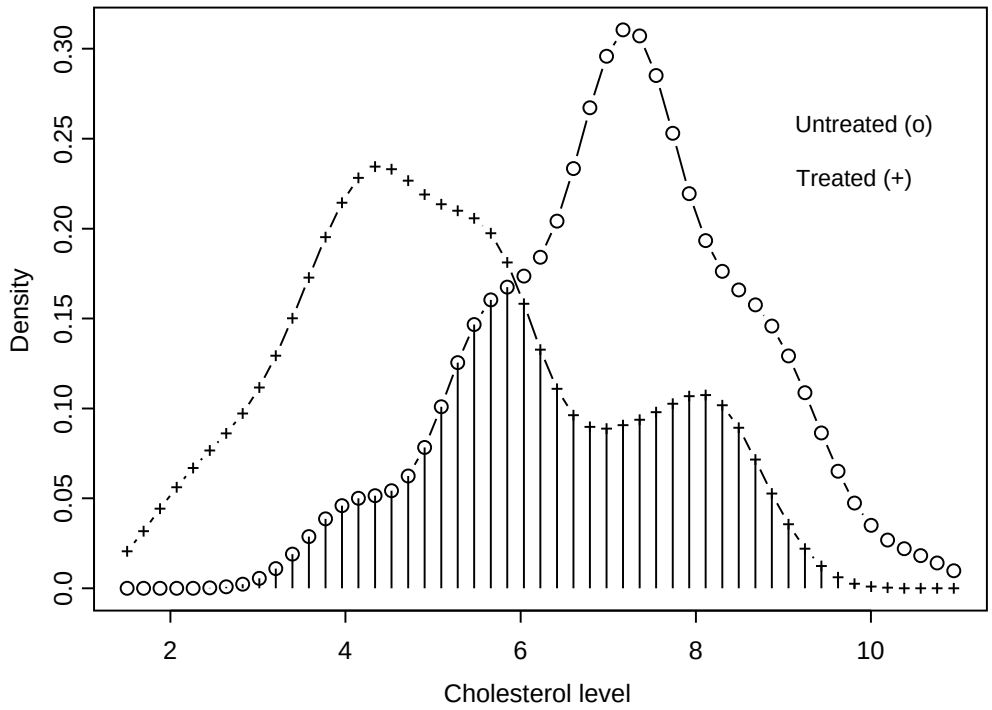
In this paper we propose to use a measure for balance in PS methods, known in the literature as the *overlapping coefficient* (OVL)[11,12] or proportion of similar responses.[13] This measure directly quantifies the overlap between a covariate distribution of the treated and the untreated. Its value gives information on the amount of balance that has been reached in a certain PS

model. In the next section we give the definition of the non-parametric OVL. In the third section we show how this measure can be used in PS methods to check and report the amount of balance. In the fourth section we perform a simulation study in which the OVL has been used as a model selection tool and compare it with other approaches.

NON-PARAMETRIC OVERLAPPING COEFFICIENT

The concept of overlap of probability distributions fits the objectives of PS analysis. Without assuming any prediction model, one seeks to create balance on covariates. To understand the meaning of balance in this context, one can look at randomized studies. The randomization process guarantees that the *whole distribution* of all covariates is 'on average' similar between treatment groups. Any departure from similarity of distributions between treatment groups should be measured. In general, a statistical test on differences of group means is only a limited way of collecting information on the similarity of whole distributions. The question whether covariate distributions between treatment groups are similar (see Figure 5.1), can bet-

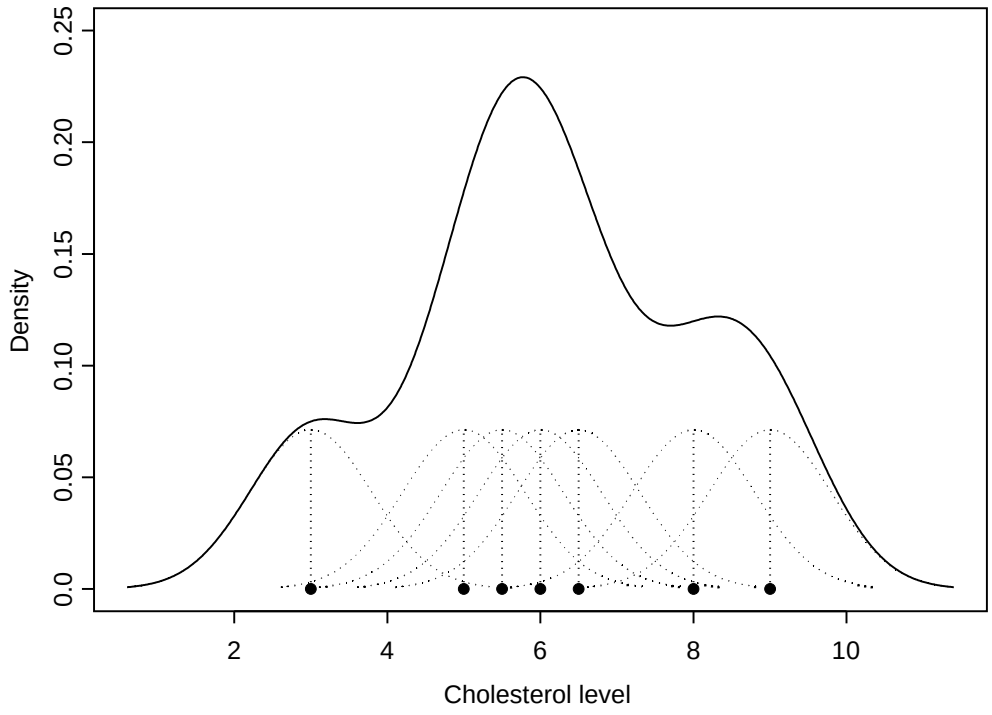
Figure 5.1: Illustration of the concept of overlap for the covariate cholesterol level in two random samples, one from a Gamma distribution (treated group, $n = 50$, $\mu = 6$ and $\lambda = 1$) and one from a normal distribution (untreated group, $n = 50$, $\mu = 7$ and $\sigma = 1.5$), using kernel density estimation



ter be answered by a measure for balance, the overlapping coefficient: it measures the amount of overlap in two distributions in a direct way and has a clear interpretation. The OVL is an estimate of that part of the distribution that overlaps with the other distribution.

In our situation the interest is in an estimate of the overlap between the covariable distributions of treated and untreated individuals. To estimate this overlap we first need to estimate the density of both distributions. It is not reasonable to assume any known theoretical distribution of covariates within subclasses of the propensity score. Therefore, we will estimate the densities in a non-parametrical way[14,15] by using kernel density estimation.[16,17] This can be seen as an alternative for making a histogram of the data with n observations. A kernel density is the sum of n density functions K, located at each observation with a chosen bandwidth h. With larger bandwidths the density function will be more smooth. There are different methods to find an optimal bandwidth. In Figure 5.2 kernel density estimation is illustrated for a small sample of 7 observations, using the normal density function for the kernel and a bandwidth determined by the normal reference rule method.[16]

Figure 5.2: Illustration of kernel density estimation in a sample of 7 cholesterol levels ($3, 5, 5.5, 6, 6.5, 8$ and 9) using the normal density kernel and the normal reference rule bandwidth



When for both treatment groups the density functions $\hat{f}(x|t=0)$ and $\hat{f}(x|t=1)$ are estimated, the OVL is the proportion of the density that overlaps with the other. Numerically we calculated

this proportion with Simpson's rule using a 101 grid.[14] In Appendices B and C the S-Plus and SAS codes are given to implement the estimation of the OVL in practical settings.

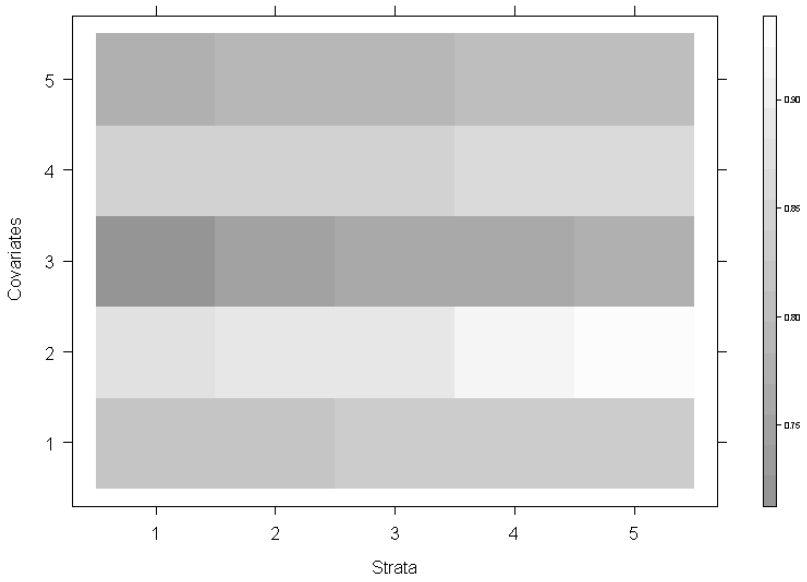
$$\widehat{OVL} = \int_{-\infty}^{\infty} \min\{\hat{f}(x|t=0), \hat{f}(x|t=1)\}dx \tag{5.1}$$

The influence on the OVL estimate of choosing other functions for the kernel, like Epanechnikov's kernel or fourth-order kernel,[18] other bandwidth methods or other grids is quite small.[14] It should be noted that in case of perfect overlap of both treatment groups in the population, the expectation of the OVL in a sample will be less than its maximum of 1. The variance of the OVL estimator can best be approximated by bootstrap methods, because even the derived formulas for normal distributions are in general too optimistic.[12]

OVERLAPPING COEFFICIENT TO CHECK AND REPORT BALANCE

In a PS analysis it is common practice to divide the sample in five groups based on the quintiles of the propensity score (strata). The OVL can be used to quantify the balance within strata on each of the prognostic factors, which gives a *distribution of estimated OVLs*. This information can be summarized in a cross table or in a graph like Figure 5.3.

Figure 5.3: Visualization of estimated overlapping coefficients for 5 normally distributed covariates within 5 strata of the propensity score, in a simulated data set of $n = 400$



The information as presented in this figure (where dark cells indicate low OVLs) can be used in several ways. First, one will be alerted on covariates or strata in which the balance is comparatively low, which could be improved for instance by including higher-order terms or interactions with other covariates. Second, the average OVL per covariate could be compared with the crude balance that exists on that covariate in the data set to see what *improvement in balance* has been reached by using PS stratification. Third, the information can give an answer to the important question whether the overall balance is *acceptable* in order to continue the analysis and estimate an adjusted treatment effect using the specified PS model. One way to do this, is to compare the whole distribution of estimated OVLs with a *reference distribution of OVLs*. A quite natural reference is the distribution of OVLs under the null hypothesis of similar covariate distributions between treatment groups, which is in fact the expected balance in a randomized trial. In Table 5.2 in Appendix A only the first decile of this distribution is given as a reference for various distributions and sample sizes. Fortunately, the OVL distribution is only slightly dependent on the shape of the covariate distribution. As an example we can use the data underlying Figure 5.3. The first decile estimated on these data equals 0.83 with an average number of observations of 40 per group and stratum. When this value is compared to the closest relevant figure in Table 5.2 (0.82, $n = 50$, normal distributions), it can be concluded that the balance after propensity score modeling is approximately comparable with the balance that would have been found if groups were in fact similar. Although one should strive for even better balance when possible, it gives at least an indication whether any acceptable balance has been reached.

A fourth way to use the balance information is to calculate a summary measure for the distribution of estimated OVLs on a wide range of possible PS models for helping to select a good PS model, a model with high overlap. To use this overall measure of balance for model selection there should exist a strong association with bias. In the next paragraph we give the results of a simulation study in which we investigated the strength of the relationship between a summary measure of balance and bias in estimating a treatment effect.

OVERLAPPING COEFFICIENT TO SELECT A PS MODEL

For a variety of PS models a distribution of estimated OVLs can be estimated which gives information on the amount of balance for that particular model. By relating a summary measure of this distribution to the amount of bias, we assessed the ability of the OVL to serve as a model selection tool in PS analysis. We also explored three other methods that could be used for model selection in PS analysis, but which are hardly used in practical settings. First, we used the *p*-values from *t*-tests, performed on all covariates within strata of the PS to detect a difference between treated and untreated individuals. Second, we used the method described in Rosenbaum & Rubin,[10] who regressed each of the covariates on treatment alone and on both treatment and PS, in order to compare both models with the *F*-statistic from ANOVAs. Third,

we calculated the *c*-statistic for any PS model. Although its absolute value among different data sets is not indicative for the amount of balance, its relative value within data sets could be used to choose among various PS models.

METHODS

We simulated a population of $100,000$ individuals, a dichotomous outcome y, a dichotomous treatment t and 10 normally distributed covariates of which five were prognostic for the outcome ($x_1 - x_5$) and five were not ($x_6 - x_{10}$). First we simulated the distribution of the outcome ($\pi_y = 0.30$) and treatment ($\pi_t = 0.50$) and their relationship by means of the odds ratio ($OR_{ty} = 2.0$), making sure that all covariates were perfectly unrelated to treatment and moderately related to outcome ($x_1 - x_5, OR_{xy} = 1.3$) or not related to outcome ($x_6 - x_{10}, OR_{xy} = 1.0$). This enabled us to know the true marginal treatment effect without assuming any true outcome model (in the population no confounding). When sampling from this population, sample-specific confounding will appear which is in general small. To create stronger confounding in samples, which is common in observational studies, we gave to individuals different sampling probabilities that are related to treatment and to covariates x_1, x_2, x_4, x_6 and x_7. This resulted in unadjusted treatment effects OR_{unadj} between 1.5 and 4.1 with a mean of 2.55. An adjusted treatment effect was estimated with PS stratification, dividing the propensity score into five strata. When averaged over simulations, bias was defined as the percentage difference between the average adjusted treatment effect and the true treatment effect by means of the odds ratio.

To summarize the distribution of OVLs, and similarly the distributions of *p*-values from *t*-tests and ANOVAs, we used the first decile (10%), the first quintile (20%) and the median of these distributions. Thereby, we calculated for the OVL distribution an unweighted and a weighted average. The weighted average takes into account the estimated association between covariate i and outcome y ($\widehat{OR}_{x_i y}$) and is defined as:

$$\widehat{OVL}_w = \frac{1}{JI} \sum_{i=1}^{I} \sum_{j=1}^{J} w_i \widehat{OVL}_{ij} \tag{5.2}$$

where I is the number of covariates, J the number of strata, $\widehat{OVL}_{ij}$ the estimated OVL for covariate i in stratum j and

$$w_i = 1 + \widehat{OR}_{x_i y} - \frac{1}{I} \sum_{k=1}^{I} \widehat{OR}_{x_k y} \tag{5.3}$$

The weighted average OVL quantifies the idea that it is more important for strong prognostic factors to be balanced than for covariates that are weakly or not related with outcome.

To evaluate the relationship between bias and measure for balance we created 20 different PS models within each simulated data set, ranging from only three covariates to all 10 covariates with interactions. The simulations were done with various sample sizes ($n = 400, 800, 1,200, 1,600, 2,000, 4,000$ and $6,000$). Because results between simulations were fairly similar, a comparatively small number of 100 simulations per sample size was sufficient for a reliable estimate of the correlation.

For the final conclusion on the strength of the association between the measure for balance and bias, we used Pearson's correlation coefficients within simulations and sample sizes. We also performed an overall analysis on the results, i.e. a linear mixed-effects model (S-Plus function `lme`) with bias as the dependent variable, measure for balance as the fixed effect and simulation number as the random effect (random intercept only). As measure for model improvement we used Akaike's Information Criterion (AIC).

RESULTS

There exists an inverse relationship between the summary measures for the OVL distribution and the percentage of bias when estimating a treatment effect: the higher the measure for balance, the smaller the difference between estimated and true effect. The strength of this relationship is dependent on the chosen summary measure and on sample size. For example, for $n = 400$ Pearson's correlation was $R = -0.11$, while for $n = 6,000$ a correlation of $R = -0.83$ was found (see Table 5.1). The unweighted average and the percentile measures for the OVL showed somewhat smaller correlations (only first quintile shown in Table 5.1 column 2) than the weighted average OVL.

Table 5.1: Average Pearson's correlations (standard deviation) between bias and various summary measures of balance: weighted average and 1^{st} quintile (p_{20}) from OVL distribution, 1^{st} quintile from *p*-value distribution from *t*-tests, 1^{st} quintile from *p*-value distribution from ANOVAs and the *c*-statistic, 100 simulations per sample size

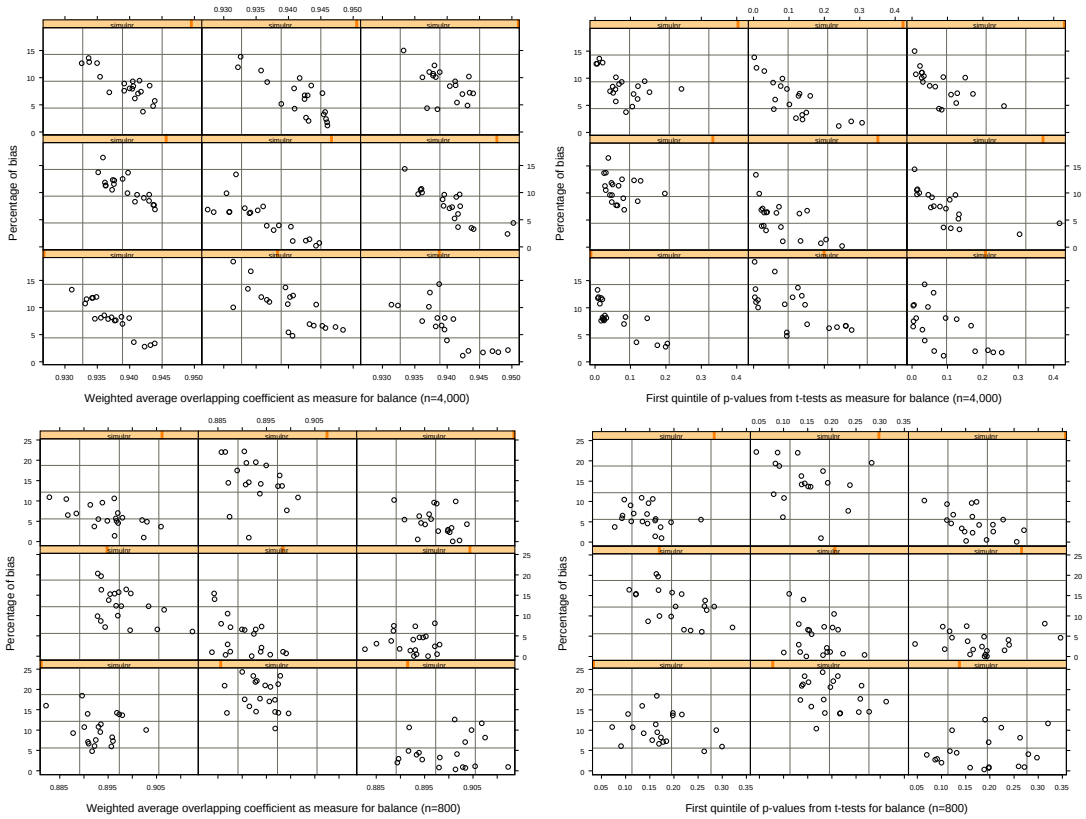
	OVL weighted average	OVL p_{20}	*t*-tests p_{20}	ANOVAs p_{20}	*c*-statistic
n=400	-0.11 (0.24)	-0.06 (0.22)	-0.28 (0.25)	-0.28 (0.31)	-0.26 (0.29)
n=800	-0.30 (0.29)	-0.17 (0.31)	-0.34 (0.28)	-0.27 (0.28)	-0.35 (0.25)
n=1,200	-0.38 (0.28)	-0.28 (0.28)	-0.42 (0.26)	-0.28 (0.33)	-0.39 (0.20)
n=1,600	-0.56 (0.23)	-0.42 (0.25)	-0.55 (0.23)	-0.40 (0.27)	-0.42 (0.25)
n=2,000	-0.59 (0.20)	-0.44 (0.24)	-0.59 (0.16)	-0.43 (0.24)	-0.34 (0.20)
n=4,000	-0.81 (0.09)	-0.68 (0.14)	-0.59 (0.18)	-0.32 (0.22)	-0.37 (0.22)
n=6,000	-0.83 (0.09)	-0.70 (0.15)	-0.58 (0.17)	-0.42 (0.22)	-0.44 (0.18)

In general, it can be concluded that other measures show higher correlations for sample sizes below 800 and lower correlations for sample sizes exceeding $1,600$. The correlations for the summary measure of the *p*-value distribution from ANOVAs range from -0.27 to -0.43 (first quintile shown in column 4), whereas the summary measures based on the distribution of *p*-values from *t*-tests are somewhat higher, ranging from -0.28 to -0.59 (first quintile shown

in column 3). For sample sizes of 400 observations this measure is more predictive for bias than the OVL measure, while for samples between 800 and $1,600$ the differences between methods are small. We also performed linear mixed-effects models on the simulation results and found a similar pattern when the AICs were compared (results not shown).

As an illustration for these results, the relationship between bias and measure for balance is captured in Figure 5.4 for a random selection of nine samples of 800 and $4,000$. For samples of $4,000$ observations the predictive power for the weighted OVL (left upper panel) is larger than for the first quintile of the *p*-value distribution from *t*-tests (right upper panel). For sample sizes of 800 the fit is worse for both methods and slightly better for the *p*-values (right lower panel) than for the weighted OVL measure (left lower panel).

Figure 5.4: Association between average percentage of bias and weighted average OVL (left panels) and first quintile of *p*-value distribution from *t*-tests (right panels), within 9 random chosen samples of $n = 4,000$ (upper panels) and $n = 800$ (lower panels)



DISCUSSION

In observational studies that use propensity score analysis to estimate the effect of treatment or any exposure, we propose to focus more on the stage of creating the PS model by using a measure for balance, the overlapping coefficient. In the first place this measure for balance can be used to quantify balance in an absolute sense. Second, it can be used to judge whether the balance created on covariates with propensity score modeling was successful and sufficient to continue the analysis and estimate an adjusted treatment effect. Third, due to its inverse association with bias this measure can also be a help for model selection. The weighted average OVL calculated on the set of available covariates show strongest association with bias for larger data sets. For smaller data sets the *p*-values from significance tests and the *c*-statistic have higher predictive power for the bias than the OVL measure.

A disadvantage of the OVL is that it is estimated per covariate and per stratum. For small sample sizes and a large number of covariates this implies a great number of calculations with a small number of observations per stratum which can make the estimated overlap in densities unreliable. This effect will be partly diminished because the focus is on the whole distribution of OVLs. Also when the propensity score is divided in a larger number of strata, although five is common practice, estimation of OVLs becomes less reliable.

We focussed on the use of the OVL in propensity score stratification. When matching on the propensity score is chosen as method to estimate treatment effects, one could similarly estimate OVLs and compare models by using strata before matching takes place. After the matching one could also estimate OVLs on all covariates between both matched sets, but because this does not take into account the dependency of the data, it is not recommended.

We presented the results for a true population treatment effect of $OR_{ty} = 2.0$ and an incidence rate $\pi_y = 0.30$. The results were fairly robust against small changes in these values ($1.5 < OR_{ty} < 3.0$ and $0.10 < \pi_y < 0.40$). In all situations an increasing predictive power for the OVL measures was seen with increasing sample size. Compared to the other methods, the OVL measures performed best with larger data sets.

We summarized the distribution of *p*-values and OVLs over covariates and strata by using different methods: first decile, first quintile and median. Because the first quintile was most predictive, we only gave these figures in Table 5.1.

To know the true treatment effect in simulation studies it is common to first simulate the covariates, then the treatment variable as a function of the covariates (the propensity score model) and finally the outcome variable as a function of treatment and covariates (the outcome model). Our setting on the other hand was less restrictive because we did not need to specify any model for the outcome nor did we specify any propensity score model. Furthermore, specification of the outcome model using logistic regression can lead to a divergence in the type of effect to be estimated[19–22] with possibly misleading conclusions when comparison with propensity score methods is the main objective.[7,23,24] Finally, using the true propensity score

model if it were known, is recommended, because a sample-specific propensity score model generates on average less bias than the true propensity score model.[10,25–27]

We propagate that in studies using propensity score methods, more attention should be paid to creating, measuring and reporting the balance that has been reached by the chosen propensity score model. The use of the overlapping coefficient could be a great help, also as a tool to select a PS model among a variety of possible models.

APPENDIX A: THE OVL DISTRIBUTION UNDER THE NULL

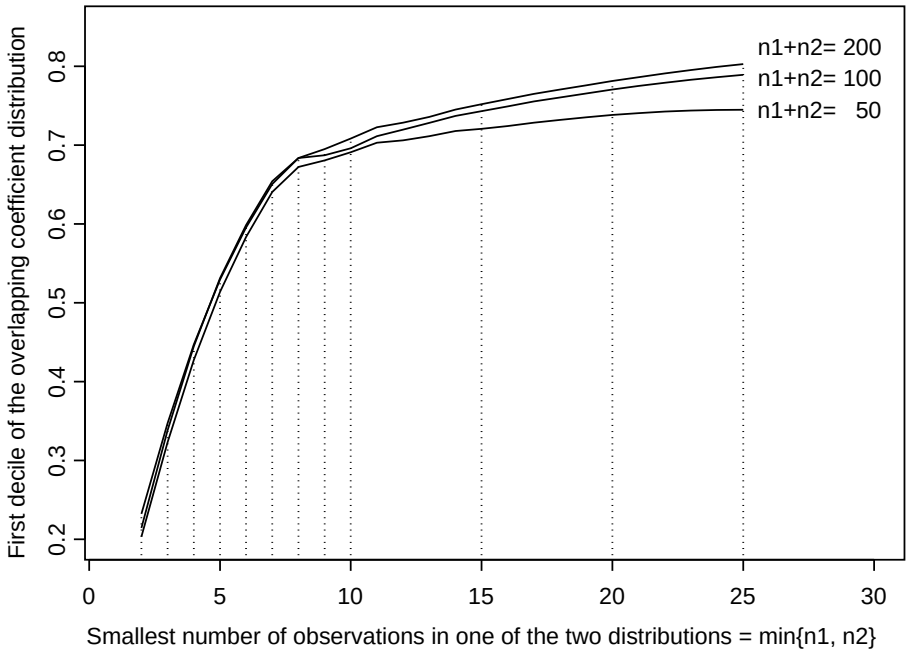
Under the null hypothesis that two samples come from populations with similar covariate distributions, we determined the OVL distribution by simulating $1,000$ samples with an equal number of observations from both populations. In Table 5.2 the first decile (10%) of this OVL distribution is given.

Table 5.2: First decile of the expected distribution of OVLs when both covariate distributions are similar for normal, chi-square, uniform, gamma and exponential distributions and various number of observations (n_i=number of observations in distribution i)

	normal	chi-square	uniform	gamma	exponential
$n_1 = n_2$	μ=0, σ=1	df=2	min=0, max=1	λ=3, μ=1	rate=2
25	0.74	0.70	0.78	0.74	0.71
50	0.82	0.79	0.83	0.81	0.79
100	0.87	0.85	0.88	0.86	0.84
200	0.90	0.89	0.91	0.90	0.89
400	0.93	0.92	0.93	0.92	0.92
800	0.94	0.94	0.95	0.94	0.94
1,600	0.96	0.96	0.96	0.96	0.96
3,200	0.97	0.97	0.97	0.97	0.97

When the first decile of the OVL distribution calculated on own data is higher than the tabulated value, this indicates that the balance is at least as good as could be expected when both groups are similar. Because the underlying distribution of covariates within strata is usually unknown, it is convenient that the values in Table 5.2 are quite similar among different distributions. Note that in this table the number of observations concern the numbers per treatment group within strata and are assumed to be equal ($n_1 = n_2$). When the number of observations is not the same for both groups, the expected OVLs will be lower. Below eight observations in one of the groups (irrespective of the number of observations in the other group), the left tail of the OVL distribution quickly reaches low values. This can be seen in Figure 5.5 for normal distributions. For instance, in case of similarity of groups with sample sizes of $n_1 = 4$ and $n_2 = 46$ in 10% of the cases an OVL will be found lower than 0.45. With such a low number of observations estimation of the overlapping coefficient is questionable.

Figure 5.5: First decile of overlapping coefficient distribution, calculated in 500 simulations with unequal sample sizes in both normal distributions, for $n_1 + n_2 = 50, 100$ and 200



APPENDIX B: S-PLUS CODE FOR ESTIMATING THE OVL

Calculating the OVL involves estimation of two density functions evaluated at the same x-values and then calculating the overlap. The function `ovl` needs two input vectors of observations on the covariate for both groups (`group0` and `group1`). We used the S-Plus built-in function `density` using the normal density rule `bandwidth.nrd`. For calculation of the overlap we used Simpsons rule on a grid of 101. A plot of the two densities and the overlap is optional (`plot=T`).

```
# S-Plus Function to calculate the non-parametric overlapping coefficient
#                 (plus optional figure)
ovl <- function(group0, group1,plot=F){
    wd1     <- bandwidth.nrd(group1)
    wd0     <- bandwidth.nrd(group0)
    from    <- min(group1,group0) - 0.75 * mean(c(wd1,wd0))
    to      <- max(group1,group0) + 0.75 * mean(c(wd1,wd0))
    d1      <- density(group1, n = 101, width=wd1, from=from, to=to)
    d0      <- density(group0, n = 101, width=wd0, from=from, to=to)
    dmin    <- pmin(d1$y,d0$y)
    ovl     <- ((d1$x[(n<-length(d1$x))]-d1$x[1])/(3*(n-1)))*
               (4*sum(dmin[seq(2,n,by=2)])+2*sum(dmin[seq(3,n-1,by=2)])
                    +dmin[1]+dmin[n])
    if(plot){
        maxy    <- max(d0$y, d1$y)
```

```
        minx    <- min(d0$x)
        plot(d1, type="l", lty=1, ylim=c(0,maxy), ylab="Density",xlab="")
        lines(d0, lty=3)
        lines(d1$x, dmin, type="h")
        text(minx, maxy, "  OVL =")
        text(minx+0.085*(max(d1$x)-minx), maxy, round(ovl,3))
        }
    round(ovl,3)
    }

# Example
treated     <- rnorm(100,10,3)
untreated   <- rnorm(100,15,5)
ovl(group0=untreated, group1=treated, plot=T)
```

APPENDIX C: SAS CODE FOR ESTIMATING THE OVL

```
%macro OVL(group0, group1);
  proc univariate data=&group0 noprint;
    var var;
    output out=res0 n=n0 mean=mean0 var=var0 min=min0 max=max0 q1=var0_q1 q3=var0_q3;
  run;
  proc univariate data=&group1 noprint;
    var var;
    output out=res1 n=n1 mean=mean1 var=var1 min=min1 max=max1 q1=var1_q1 q3=var1_q3;
  run;
  data res;
    merge res0 res1;
    bandwidth0= 4*1.06 * min(sqrt(var0), (var0_q3 - var0_q1)/1.34) * n0**(-1/5);
    bandwidth1= 4*1.06 * min(sqrt(var1), (var1_q3 - var1_q1)/1.34) * n1**(-1/5);
    from = min(min0,min1) - 0.75 * mean(bandwidth0,bandwidth1);
    to   = max(max0,max1) + 0.75 * mean(bandwidth0,bandwidth1);
    call symput('from',from);
    call symput('to',to); run;
  PROC KDE data=&group0 ;
    univar var (ngrid=101 gridl=&from gridu=&to) /method=SNR out=dens0; run;
  PROC KDE data=&group1 ;
    univar var ( ngrid=101 gridl=&from gridu=&to) /method=SNR out=dens1; run;
  data dens;
    merge dens0(rename=(var=var0 value=val0 density=dens0 count=n0))
          dens1(rename=(var=var1 value=val1 density=dens1 count=n1)); run;
  data ovl;
    set   dens nobs=last;
    retain two_sums 0 x_first dens_min_first; keep ovl;
    dens_min=min(dens0,dens1);
    if _n_>=2 AND mod(_n_,2)=0 then two_sums = two_sums + 4 * dens_min;
    if _n_>=3 AND mod((_n_-1),2)=0 then two_sums = two_sums + 2 * dens_min;
    if _n_=1 then do;
        x_first = val0;
        dens_min_first = dens_min;
        end;
    if _n_=last then do;
        ovl = ((val0-x_first)/(3*(_n_-1)))*(two_sums + dens_min_first + dens_min);
        output;
    end; run;
  proc print data=ovl; run;
%mend OVL;

%OVL(untreated0, treated1);
```

REFERENCES

[1] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

[2] D'Agostino, RB Jr. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med*, 17:2265–2281, 1998.

[3] Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Weaknesses of goodness-of-fit tests for evaluating propensity score models: the case of the omitted confounder. *Pharmacoepidemiol Drug Saf*, 14(4):227–238, 2005.

[4] Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *J Clin Epidemiol*, 58:550–559, 2005.

[5] Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol*, 59:437–447, 2006.

[6] Rubin DB. On principles for modeling propensity scores in medical research (Editorial). *Pharmacoepidemiol Drug Saf*, 13:855–857, 2004.

[7] Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection in propensity score models. *Am J Epidemiol*, 163:1149–1156, 2006.

[8] Rubin DB, Thomes N. Matching using estimated propensity score: relating theory to practice. *Biometrics*, 52:249–264, 1996.

[9] Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiol Drug Saf*, 13(12):841–853, 2004.

[10] Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *JAMA*, 387:516–524, 1984.

[11] Bradley EL. *Overlapping coefficient, in: Encyclopedia of Statistical Sciences, vol. 6*. Wiley, New York, 1985.

[12] Inman HF, Bradley EL. The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. *Commun Statist -Theory Meth*, 18(10):3851–3874, 1989.

[13] Rom DM, Hwang E. Testing for individual and population equivalence based on the proportion of similar responses. *Stat Med*, 15:1489–1505, 1996.

[14] Stine RA, Heyse JF. Non-parametric estimates of overlap. *Stat Med*, 20:215–236, 2001.

[15] Cleveland WS, Devlin SJ, Grosse E. Regression by local fitting. *J Econometr*, 37:87–114, 1988.

[16] Silverman BW. *Density estimation for statistics and data analysis*. Chapman and Hall: London, 1986.

[17] Wand MP, Jones MC. *Kernel Smooting*. Chapman and Hall, 1995.

[18] Mammitzsch V Gasser T, Müller H-G. Kernels for nonparametric curve estimation. *J Royal Stat Soc, Series B*, 47:238–252, 1985.

[19] Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, 71:431444, 1984.

[20] Robins JM, Tsiatis AA. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Commun Statist -Theory Meth*, 20(8):2609–2631, 1991.

[21] Hauck WW, Neuhaus JM, Kalbfleisch JD, *et al*. A consequence of omitted covariates when estimating odds ratios. *J Clin Epidemiol*, 44:77–81, 1991.

[22] Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *Am J Epidemiol*, 125:761–768, 1987.

[23] Austin PC, Grootendorst P, Normand ST, Anderson GM. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. *Stat Med*, 26:754–768, 2007.

[24] Martens EP, Pestman WR, Klungel OH. Letter to the editor: 'Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study, by Austin PC, Grootendorst P, Normand ST, Anderson GM'. *Stat Med*, 26:3208–3210, 2007.

[25] Robins JM, Mark SD, Newey WK. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48:479–495, 1992.

[26] Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. *Am J Epideiol*, 150:327–333, 1999.

[27] Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol*, 158:280–287, 2003.

5.2 MEASURING BALANCE IN PROPENSITY SCORE METHODS

Edwin P. Martens[a,b], Wiebe R. Pestman[b], Anthonius de Boer[a], Svetlana V. Belitser[a] and Olaf H. Klungel[a]

[a] *Department of Pharmacoepidemiology and Pharmacotherapy, Utrecht Institute of Pharmaceutical Sciences (UIPS), Utrecht University, Utrecht, the Netherlands*
[b] *Centre for Biostatistics, Utrecht University, Utrecht, the Netherlands*

Submitted for publication

ABSTRACT

Propensity score methods focus on balancing confounders between groups to estimate an adjusted treatment or exposure effect. However, there is a lack of attention in actually measuring, reporting and using the information on balance, for instance for model selection. We propose to use a measure for balance in propensity score methods and describe three such measures: the overlapping coefficient, the Kolmogorov-Smirnov distance and the Lévy metric.

We performed simulation studies to estimate the association between these measures for balance and the amount of bias. For all three measures we found an inverse relationship with bias increasing with sample sizes. The simulations further suggest that the predictive power for the overlapping coefficient was highest: for samples of 800 observations the average Pearson's correlation was -0.23, while for $2,000$ observations -0.63 was found. Mainly for large samples the overlapping coefficient can be used as a model selection tool because its value is predictive for the amount of bias. The mean squared error for these balancing strategies is quite similar among these methods, for the overlapping coefficient ranging from 0.031 for $n = 2,000$ to 0.197 for $n = 400$. This is much smaller than when a standard PS model including all covariates is applied (0.076 to 0.302). We conclude that these measures for balance are useful to report the amount of balance reached in any propensity score analysis and can be a help in selecting the final propensity score model.

Keywords: Confounding; Propensity scores; Observational studies; Measures for balance; Overlapping coefficient; Kolmogorov-Smirnov distance; Lévy metric

INTRODUCTION

A commonly used statistical method to assess treatment effects in observational studies, is the method of propensity scores (PS).[1,2] PS methods focus on creating balance on covariates between treatment groups by first creating a PS model to estimate the conditional probability to be treated given the covariates (the propensity score). In the second step an adjusted treatment effect is estimated, using the propensity score as matching variable, as stratification variable, as continuous covariate or inverse probability weight. In the literature in which PS methods are applied there is a lack of attention to building the PS model.[3–5] Building such a PS model involves the selection (and transformations) of variables and possibly interactions or higher-order terms to include in the model and a check whether the chosen model creates balance on the important prognostic covariates. Unlike prediction models the selection of variables for a PS model (in which treatment is the dependent variable) is more complex: both the relationship with treatment and outcome has to be taken into account. That means that model-building strategies like stepwise regression are not very useful in deciding whether a certain PS model is acceptable or not. Any stepwise regression method to build the PS model only selects on significance of the relationship *with treatment*, but this does not use information on the strength of the relationship *with outcome*. A strong relationship between treatment and covariates is not necessary for having a good PS model.[6,7] What should be important in PS models is the *balance on prognostic covariates* that has been reached by using the chosen PS model.

When information on balance is given, this is mostly done by performing significance tests within strata of the covariates to assure that the mean (or proportion) on covariates do not differ significantly between treatment groups. An early method proposed by Rosenbaum & Rubin to check the balance per covariate using the F-statistic from analysis of variance (ANOVA) is not often used.[8] More frequently the c-statistic (area under the receiver operating curve) is reported, but does not give the information needed: also a low value of the c-statistic can indicate good balance on prognostic factors (for instance in a randomized trial). We propose to use an overall weighted measure for balance to report the amount of balance reached and to select the final PS model.

In this paper we describe three measures for balance that can be used in PS methods: the *overlapping coefficient* (OVL),[9,10] also known as the *proportion of similar responses*,[11] the *Kolmogorov-Smirnov distance* (D)[12,13] and the *Lévy metric* (L).[14,15] In the next section we will define these measures. In the third section we give the results of a simulation study in which these measures are compared on their relationship with bias and on their ability in correctly estimating the treatment effect compared to a standard PS model.

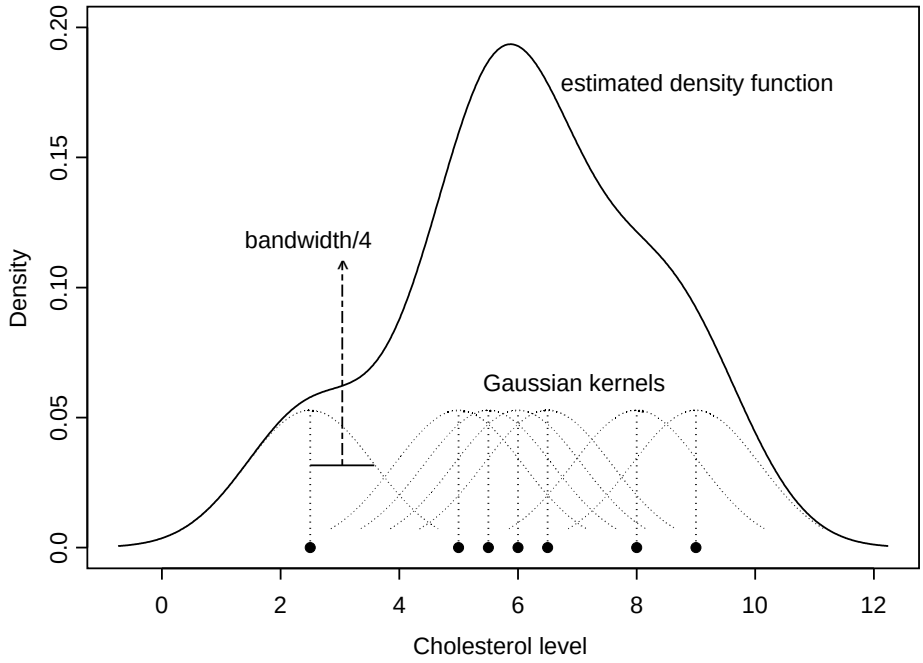
THREE MEASURES FOR BALANCE

The objective of PS methods is to create balance on the covariates that confound the relationship between outcome and treatment in observational studies. In randomized experiments this balance implies that the *whole distribution* of all covariates is 'on average' similar between treatment groups, not only the mean of the distribution or other summary measures. Whether or not covariate distributions of treatment groups are similar can best be approached by actually measuring the balance instead of testing whether the means of both distributions are significantly different. Any departure from similarity on prognostic factors could cause a difference in the outcome not caused by treatment. We will discuss three possible measures for balance.

NON-PARAMETRIC OVERLAPPING COEFFICIENT

The overlapping coefficient measures the amount of overlap in two distributions and is an estimate of that part of the distribution that overlaps with the other distribution. To estimate the overlapping coefficient we first need to estimate the density of both distributions. Because it is not reasonable to assume any known theoretical distribution of covariates within subclasses of the propensity score, we will estimate the densities in a non-parametrical way[16,17] by using kernel density estimation.[18,19] This can be seen as an alternative for making a histogram of the

Figure 5.6: Illustration of kernel density estimation in a sample of 7 cholesterol levels ($2.5, 5, 5.5, 6, 6.5, 8$ and 9) using the normal density function for the kernel and the normal reference rule bandwidth

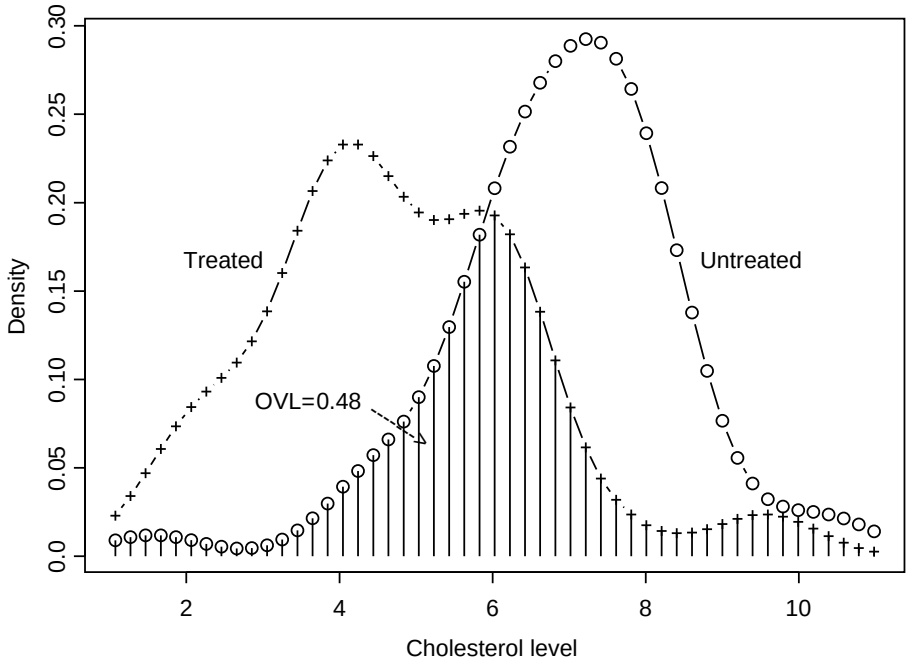


data with n observations. A kernel density is the sum of n density functions K, located at each observation with a chosen bandwidth h. Increasing the bandwidth will make the density function more smooth. In Figure 5.6 kernel density estimation is illustrated for a small sample of 7 observations, using the normal density function for the kernel and a bandwidth determined by the normal reference rule method.[18] When for both treatment groups ($t = 0$ is untreated, $t = 1$ is treated) the density functions $\hat{f}(x|t = 0)$ and $\hat{f}(x|t = 1)$ are estimated, the OVL is the proportion of one density that overlaps with the other. Numerically we calculated this proportion with Simpson's rule using a 101 grid.[16]

$$\widehat{OVL} = \int_{-\infty}^{\infty} \min\{\hat{f}(x|t = 0), \hat{f}(x|t = 1)\}dx \tag{5.4}$$

The influence on the OVL estimate of choosing other functions for the kernel, like Epanechnikov's kernel or fourth-order kernel,[20] other bandwidth methods or other grids is quite small.[16] Note that in case of perfect overlap of both treatment groups in the population, the expectation of the OVL in a sample will be less than 1. The variance of the OVL estimator can best be approximated by bootstrap methods, because even the derived formulas for normal distributions are in general too optimistic.[10] In Figure 5.7 the overlapping coefficient is illustrated for a

Figure 5.7: Illustration of the overlapping coefficient for cholesterol level in two random samples, treated group drawn from a Gamma distribution ($n = 50$, $\mu = 6$ and $\lambda = 1$) and untreated group from a normal distribution ($n = 50$, $\mu = 7$ and $\sigma = 1.5$), using kernel density estimation



ted group of 50 observations drawn from a Gamma distribution with $\mu = 6$ and $\lambda = 1$ and an untreated group of 50 observations drawn from a normal distribution with $\mu = 7$ and $\sigma = 1.5$. The OVL is calculated at 0.48. In Appendix A the S-Plus code for the OVL is given.

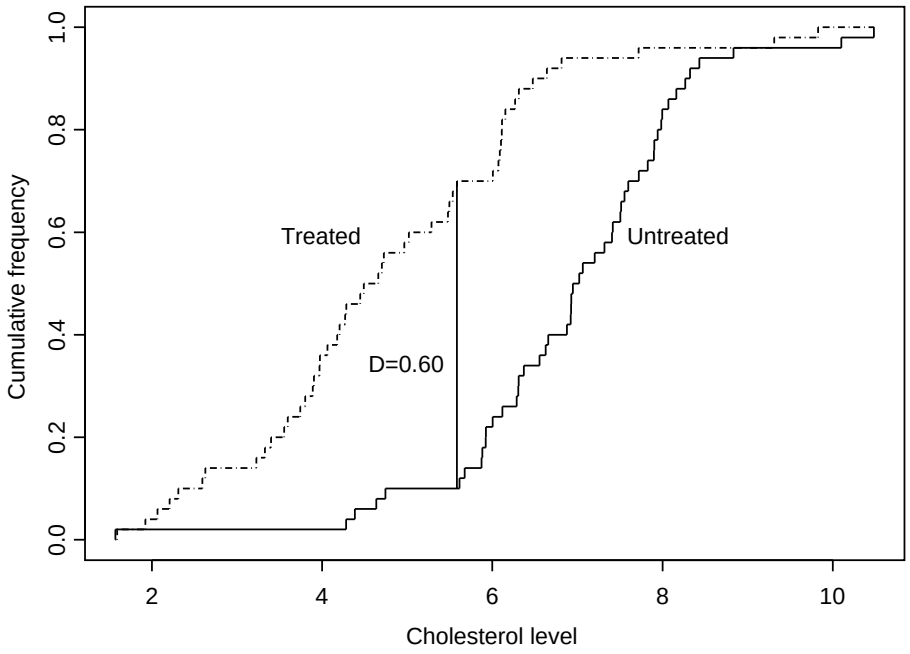
KOLMOGOROV-SMIRNOV DISTANCE

The Kolmogorov-Smirnov distance D can be described as the maximum of all vertical distances between two cumulative distribution functions, expressed as relative frequencies. The minimum distance of 0 will be reached when both distributions are exactly similar. The larger this measure, the less similar distributions are, with a maximum of 1. This distance is also used for the difference between an empirical and a known theoretical distribution. The Kolmogorov-Smirnov distance D between untreated and treated individuals is defined as

$$\hat{D} = \max\left\{\left|\hat{F}(x|t=0) - \hat{F}(x|t=1)\right|\right\} \tag{5.5}$$

where $\hat{F}(x|t=0)$ is the estimated cumulative distribution function for untreated individuals and $\hat{F}(x|t=1)$ for treated individuals. An illustration of D as a measure for balance is given in Figure 5.8 for the same data as used for Figure 5.7. In Appendix A the S-Plus code for the Kolmogorov-Smirnov distance is given.

Figure 5.8: Illustration of the Kolmogorov-Smirnov distance D for cholesterol level in two random samples, treated group drawn from a Gamma distribution ($n = 50$, $\mu = 6$ and $\lambda = 1$) and untreated group from a normal distribution ($n = 50$, $\mu = 7$ and $\sigma = 1.5$)



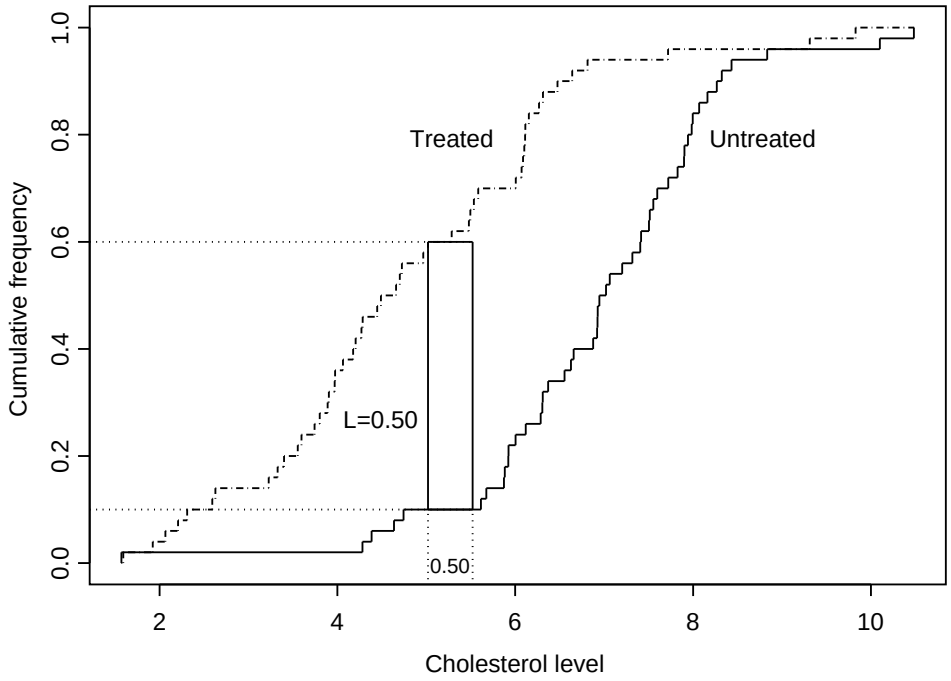
LÉVY METRIC

The Lévy metric L can be considered as a variant on the Kolmogorov-Smirnov distance that takes into account both horizontal and vertical distances between two cumulative distribution functions. The Lévy metric L is defined as

$$\hat{L} = \min\{\epsilon > 0 | \hat{F}(x-\epsilon|t=0) - \epsilon \leq \hat{F}(x|t=1) \leq \hat{F}(x+\epsilon|t=0) + \epsilon \text{ for all x in } \mathbb{R}\} \quad (5.6)$$

where $\hat{F}(x|t=0)$ is the estimated cumulative distribution function for untreated individuals and $\hat{F}(x|t=1)$ for treated individuals. Intuitively this measure can be understood as follows: if one inscribes squares between the two curves with sides parallel to the coordinate axes, then the side-length of the largest such square is equal to L. An illustration of L as a measure for balance is given in Figure 5.9 where the same data have been used as in Figures 5.7 and 5.8. From this figure it becomes clear that this distance measure is sensitive for the unit of measurement of the covariate. When different covariates are involved one should therefore use some kind of standardization of the covariates before these measures can be compared. In Appendix A the S-Plus code for the Lévy metric is given.

Figure 5.9: Illustration of the Lévy metric L for cholesterol level in two random samples, treated group drawn from a Gamma distribution ($n = 50$, $\mu = 6$ and $\lambda = 1$) and untreated group from a normal distribution ($n = 50$, $\mu = 7$ and $\sigma = 1.5$)



BALANCE MEASURES AS MODEL SELECTION TOOLS

In the first place the described measures for balance can be used to quantify and report the amount of balance reached in any PS analysis, something that is not often done when PS methods are applied.[3–5] Because theory on PS states that within strata of the propensity score distributions of covariates tend to be similar,[8] insight in the actual balance can be given by reporting this balance per covariate and stratum.

Another way to use the information on balance is when a selection among several PS models has to be made. The best PS model can be defined as the model that estimates a treatment effect as close as possible to the true treatment effect in the population. In practical settings this PS model is unknown and a selection of variables and additional terms (for instance interactions and/or quadratic terms) must be made for choosing the final PS model. Standard variable selection methods like forward or backward regression can not be used for PS models, because first the association between outcome and covariates is not taken into account and second, *balance* is the final objective of a PS model and not significance. When balance on prognostic covariates in one model is better than in the other, the one with the best balance should be preferred because in theory the estimated treatment effect has been better adjusted for imbalance of covariates.

In the previous paragraph we generally described three measures that quantify either the degree of overlap or the distance between cumulative distribution functions. To use these measures for model selection we calculated these for all covariates within strata of the propensity score, where the strata were based on the quintiles of the PS.[8] To get an overall measure for balance for every fitted PS model, we calculated a *weighted average* of the measures per covariate and stratum, with weights equal to the strength of the association between covariate and the outcome (on the log-odds scale). These weights express the idea that balance on strong prognostic factors is more important in estimating an adjusted treatment effect than on factors that are only weakly related to the outcome. This implies that a *high* value on the weighted average overlapping coefficient and *low* values on the weighted average Kolmogorov-Smirnov distance and the Lévy metric indicate good balance on prognostic factors. To find out to what degree this balance is related to bias, we performed a simulation study to compare these measures for balance on their ability to select PS models.

METHODS

For our simulations we used the framework of Austin,[21] also extensively described in Austin *et al.*[22,23] First nine normally distributed covariates $x_1 - x_9$ were simulated of which six were related to treatment t according to the following logistic model, including four interactions and

two quadratic terms:

$$\text{logit}(p_{i,t}) = \beta_{0,t} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_4 + \beta_4 x_5 + \beta_5 x_7 + \beta_6 x_8 + \\ \beta_7 x_2 x_4 + \beta_8 x_2 x_7 + \beta_9 x_7 x_8 + \beta_{10} x_4 x_5 + \beta_{11} x_1^2 + \beta_{12} x_7^2 \quad (5.7)$$

and where treatment was simulated by a Bernouilli distribution with $\pi = p_{i,t}$. The outcome y was simulated by the following logistic model, including covariates $x_1 - x_6$ and treatment t, including four interactions and two quadratic terms:

$$\text{logit}(p_{i,y}) = \alpha_{0,y} + \beta_t t + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_4 + \alpha_5 x_5 + \alpha_6 x_6 + \\ \alpha_7 x_2 x_4 + \alpha_8 x_3 x_5 + \alpha_9 x_3 x_6 + \alpha_{10} x_4 x_5 + \alpha_{11} x_1^2 + \alpha_{12} x_6^2 \quad (5.8)$$

Compared to the model of Austin we added four interactions and two quadratic terms to both the treatment model and the outcome model. The dichotomous outcome was generated by a Bernouilli distribution with $\pi = p_{i,y}$.
From equations 5.7 and 5.8 it can be deduced that:

the true confounding factors were $x_1, x_2, x_4, x_5, x_2x_4, x_4x_5$ and x_1^2,
the factors only related to treatment were x_7, x_8, x_2x_7, x_7x_8 and x_7^2,
the factors only related to outcome were x_3, x_6, x_3x_5, x_3x_6 and x_6^2.

The strength of the associations was:

for β_7, β_{10}, α_7, α_{10} equal to log(1.2),
for β_1, β_3, β_5, α_1, α_2, α_3, β_8, β_{11}, α_8, α_{11} equal to log(1.4),
for β_9, β_{12}, α_9, α_{12} equal to log(1.6)
for β_2, β_4, β_6, α_4, α_5, α_6 equal to log(2.0).

To assure that half of the subjects were treated and that an event occurred ($y = 1$) for approximately 25% of the untreated individuals, $\beta_{0,t}$ was set to -0.65 and $\alpha_{0,y}$ to -1.8.

An important feature of these simulations is that only a *conditional* effect β_t can be inserted as the true treatment effect, while with PS methods we aim to estimate a *marginal* treatment effect. For the difference between marginal and conditional treatment effects in logistic regression analysis, we refer to the literature.[24–28] Because we wanted to restrict the simulations to a true marginal treatment effect of $OR_{ty} = 2.0$, we had to find the corresponding conditional effect in this setting. We used the iterative process described in Austin[21] to calculate the true conditional treatment effect β_t to be equal to 0.8958, which equals an OR of 2.449. For our simulations we varied sample size ($n = 400, 800, 1,200, 1,600$ and $2,000$).

From the large number of possible PS models we sampled for every simulated data set at random 40 models, calculated the PS, stratified on the PS and calculated for the overlapping coefficient, the Kolmogorov-Smirnov distance and the Lévy metric the weighted average for all

these PS models. After calculating the relative bias as $\widehat{OR}_{ty}/OR_{ty} - 1$, we first used Pearson's correlation coefficient within simulations to determine the strength of the association between the measure for balance and bias. We also performed an overall analysis on the results, i.e. a linear mixed-effects model (S-Plus function `lme`) with bias as the dependent variable, measure for balance as the fixed effect and simulation number as the random effect (random intercept only). As measure for model improvement we used Akaike's Information Criterion (AIC).

In the second part we concentrated on the selection of only one PS model, i.e. the model that gave the best balance according to the measure for balance. We used the mean squared error because it directly combines average bias and the spread of the estimator, defined as:

$$\frac{1}{S}\sum_{s=1}^{S}(\widehat{OR}_{ty} - OR_{ty})^2 \tag{5.9}$$

where S is the number of simulations, $\widehat{OR}_{ty}$ the estimated treatment effect and OR_{ty} the true marginal treatment effect which was set in our simulations to 2.0. We compared the mean squared error among the three measures for balance, but also used three other PS models as a reference that include:

1. *all covariates* that are related to either treatment or outcome ($x_1 - x_8$),
2. all *true confounding factors* (x_1, x_2, x_4, x_5, interactions x_2x_4, x_4x_5, quadratic term x_1^2),
3. *all prognostic factors* as given in formula 5.8.

The last two PS models are in practical settings unknown and are used in this simulation only as theoretical models. On the other hand, knowing the true propensity score model is in general not very interesting: including factors that are only related to treatment can be disadvantageous in practice and an estimated propensity score performs better than the true propensity score.[8,29–32]

RESULTS

For all sample sizes the average Pearson's correlation coefficient between the weighted overlapping coefficient and bias is higher than for the other measures, except for sample sizes of 400 observations for which neither of the measures is rather predictive for the amount of bias (Table 5.3). For example, for a sample size of $1,600$ the average correlation for the weighted OVL was -0.61, for the weighted Kolmogorov-Smirnov distance 0.40 and for the weighted Lévy metric 0.46. As a comparison we used the *c*-statistic, for which much smaller correlations were found (at most -0.27). We also checked the correlations for the method proposed by Rosenbaum & Rubin using *F*-statistics from ANOVAs,[8] which were quite similar as for the *c*-statistic (ranging from -0.07 to -0.23). Apart from averaging correlations among simulations, we also performed an overall linear mixed-effects model and calculated AICs. We have chosen to present only the results for the correlations because it gave similar results and has a more direct interpretation of association than the AIC.

Table 5.3: Average Pearson's correlations between bias and various summary measures of balance: weighted average overlapping coefficient, weighted average Kolmogorov-Smirnov distance and weighted average Lévy metric, the *c*-statistic, 200 simulations per sample size

	overlapping coefficient	Kolmogorov-Smirnov distance	Lévy metric	*c*-statistic
n= 400	-0.06	-0.06	-0.04	-0.09
n= 800	-0.23	0.08	0.10	-0.12
n=1,200	-0.39	0.16	0.20	-0.24
n=1,600	-0.61	0.39	0.43	-0.27
n=2,000	-0.63	0.40	0.46	-0.24

In Table 5.4 the results are given when these measures for balance are used to select the model that has best balance according to that measure (columns 1-3). For the overlapping coefficient the mean squared error is 0.031 when the number of observations is $2,000$ and 0.197 for a sample size of 400. The differences between the other two measures for balance are quite small. When we compare the results of the fixed model strategy of including all covariates, either related to outcome or treatment (column 4) with the OVL measure, the mean squared error is considerably larger, ranging from 0.076 to 0.302. This PS model is commonly chosen when PS methods are adopted in practice.[33] The model that contains all confounding factors (column 5) has a slightly higher mean squared error (around 20%) than for the OVL measure, while for the fixed model strategy containing all prognostic factors (column 6) is comparable to the OVL measure. The conclusion is that the measures for balance have lower mean squared error than a commonly used PS model and are slightly less or comparable to models that contain the true factors. Because the true confounding and true prognostic factors are usually unknown in practice, it means that these methods are capable of selecting PS models that are at least as good as when the true confounding and true prognostic factors were known.

Table 5.4: Mean squared error of different methods: weighted average overlapping coefficient, weighted average Kolmogorov-Smirnov distance, weighted average Lévy metric, covariates x_1 to x_8, all confounding factors, all prognostic factors, 200 simulations per sample size

	overlapping coefficient	Kolmogorov-Smirnov	Lévy metric	covariates $x_1 - x_8$	confounding factors	prognostic factors
n= 400	0.197	0.200	0.228	0.302	0.208	0.187
n= 800	0.076	0.085	0.083	0.161	0.103	0.089
n=1,200	0.047	0.061	0.058	0.119	0.065	0.057
n=1,600	0.035	0.038	0.038	0.094	0.051	0.045
n=2,000	0.031	0.035	0.034	0.076	0.036	0.032

When the PS model has been chosen by the *c*-statistic or the *F*-statistic approach the mean squared error was approximately 30% larger (ranging from 0.05 to 0.25).

Another frequently adopted approach to adjust for confounding that is not based on PS

modeling is a multivariable logistic regression analysis. For a model that included treatment and all prognostic factors, the mean squared error ranged from 0.099 for $n = 2,000$ to 0.470 for $n = 400$, which is considerably larger than when PS methods are applied. This should be no surprise, because with logistic regression analysis the conditional treatment effect (=$exp(\beta_t) = 2.449$) is estimated, which is in general not the effect of interest and an overestimation of the marginal treatment effect.[24–28]

DISCUSSION

In observational studies that use propensity score analysis to estimate the effect of treatment or any exposure, we propose to focus more on the stage of creating the PS model by directly quantifying the amount of balance. Examples of such measures are the overlapping coefficient, the Kolmogorov-Smirnov distance and the Lévy metric. These measures can be used to report the amount of balance and can be useful for selecting the final PS model. For all three measures we showed an inverse association with bias, which was strongest for the weighted average overlapping coefficient. This association was stronger for larger than for smaller samples, for the overlapping coefficient $R = -0.06$ for n=400 and $R = -0.63$ for $n = 2,000$. Selecting the PS model with the overlapping coefficient seems to be most effective, because the mean squared error for this method was in general smallest (ranging from 0.031 to 0.197). The differences with the Kolmogorov-Smirnov distance and the Lévy metric were only minor. The PS model that contained all covariates had a considerable larger mean squared error, while the PS model that contained all true, but usually unknown, confounding factors had somewhat larger mean squared error.

The use of these measures should not replace the common sense of epidemiologists who should select, observe and measure those covariates that are potentially confounding factors. When faced with a choice of functional form or possible interactions, it can be worthwhile to use one of these measures to select the final PS model in order to have best balance and probably least bias.

Some remarks can be made about the choice among the three presented measures. First, it seems that the Lévy metric and the Kolmogorov-Smirnov distance give quite similar results, which makes the latter to be preferred because no standardization is needed. The choice between the overlapping coefficient and the Kolmogorov-Smirnov distance is more difficult to make. The overlapping coefficient has a clearer interpretation and performed best in these simulations. On the other hand, for its calculation a bandwidth and a kernel has to be chosen which may influence the estimated value.

A disadvantage of the proposed methods is that these measures must be estimated per covariate and per stratum. For small sample sizes and a large number of covariates this implies a great number of calculations with a small number of observations per stratum which can make the estimated overlap in densities unreliable. Previously we showed that estimation of

the overlapping coefficient is not valid when there are less than 8 observations in one of the distributions.[34]

We focused on the use of the OVL in propensity score stratification. When matching on the propensity score is chosen as method to estimate treatment effects, one could similarly calculate OVLs and compare models by using strata before matching takes place. After the matching one could also calculate OVLs on all covariates between both matched sets, but because this does not take into account the dependency of the data, it is not recommended.

We propagate that in studies using propensity score methods, more attention should be paid to creating, measuring and reporting the balance that has been reached by the chosen propensity score model. The use of the overlapping coefficient and the Kolmogorov-Smirnov distance could be a great help, also as a tool to select a PS model among a variety of possible models.

APPENDIX A: S-PLUS CODE FOR OVERLAPPING COEFFICIENT, KOLMOGOROV-SMIRNOV DISTANCE AND LÉVY METRIC

OVERLAPPING COEFFICIENT

Calculating the OVL involves estimation of two density functions evaluated at the same x-values and then calculating the overlap. The function `ovl` needs two input vectors of observations on the covariate for both groups (`group0` and `group1`). We used the S-Plus built-in function `density` using the normal density rule `bandwidth.nrd`. For calculation of the overlap we used Simpsons rule on a grid of 101. A plot of the two densities and the overlap is optional (`plot=T`).

```
# S-Plus Function to calculate the non-parametric overlapping coefficient
#                 (plus optional figure)
ovl <- function(group0, group1,plot=F){
    wd1     <- bandwidth.nrd(group1)
    wd0     <- bandwidth.nrd(group0)
    from    <- min(group1,group0) - 0.75 * mean(c(wd1,wd0))
    to      <- max(group1,group0) + 0.75 * mean(c(wd1,wd0))
    d1      <- density(group1, n = 101, width=wd1, from=from, to=to)
    d0      <- density(group0, n = 101, width=wd0, from=from, to=to)
    dmin    <- pmin(d1$y,d0$y)
    ovl     <- ((d1$x[(n<-length(d1$x))]-d1$x[1])/(3*(n-1)))*
               (4*sum(dmin[seq(2,n,by=2)])+2*sum(dmin[seq(3,n-1,by=2)])
                    +dmin[1]+dmin[n])
    if(plot){
        maxy    <- max(d0$y, d1$y)
        minx    <- min(d0$x)
        plot(d1, type="l", lty=1, ylim=c(0,maxy), ylab="Density",xlab="")
        lines(d0, lty=3)
        lines(d1$x, dmin, type="h")
        text(minx, maxy, "  OVL =")
        text(minx+0.085*(max(d1$x)-minx), maxy, round(ovl,3))
        }
    round(ovl,3)
    }

# Example
treated     <- rnorm(100,10,3)
untreated   <- rnorm(100,15,5)
ovl(group0=untreated, group1=treated, plot=T)
```

KOLMOGOROV-SMIRNOV DISTANCE

Within S-Plus the Kolmogorov-Smirnov distance can be simply extracted from the object generated by the function `ks.gof` by `ks.gof(group0,group1)$stat`. A function that optionally plots the two cumulative densities is given below (`plot=T`).

```
# S-Plus function to calculate the Kolmogorov-Smirnov distance using
# function 'ks2' from within function 'ks.gof' (plus optional figure)
ksdist <- function(group0, group1, plot=F){
          n0      <- length(group0)
          n1      <- length(group1)
          total   <- sort(unique(c(group0, group1)))
          ma0     <- match(group0, total)
          ma1     <- match(group1, total)
          F0      <- cumsum(tabulate(ma0, length(total)))/n0
          F1      <- cumsum(tabulate(ma1, length(total)))/n1
          diff    <- abs(F0-F1)
          ks      <- max(diff)
          if(plot){
             x.ks    <- order(ks-diff)[1]
             plot(F1, type="l", lty=1, ylab="Cumulative density", xlab="")
             lines(F0, lty=3)
             lines(c(x.ks, x.ks), c(F0[x.ks],F1[x.ks]), lty=2)
             text(0.08*(n0+n1), 1, "K-S distance =")
             text(0.20*(n0+n1), 1, ks)
             }
             ks
         }

# Example
treated     <- rnorm(100,10,3)
untreated   <- rnorm(100,15,5)
ksdist(group0=untreated, group1=treated, plot=T)
```

LÉVY METRIC

The Lévy metric can be calculated using the next two functions `mecdf` and `Levy`.

```
# S-Plus functions to calculate the L\'evy metric
mecdf <- function(group0,group1) {
           n0      <- length(group0)
           n1      <- length(group1)
           total   <- sort(unique(c(group0, group1)))
           ma0     <- match(group0, total)
           ma1     <- match(group1, total)
           F0      <- cumsum(tabulate(ma0, length(total)))/n0
           F1      <- cumsum(tabulate(ma1, length(total)))/n1
           min     <- min(F1-F0)
           max     <- max(F1-F0)
           m       <- c(min,max)
           return(m)
           }
Levy <- function(u,v){
         f <- function(s,u,v){
             t <- mecdf(u,v-s)+s
             return(t[1])
             }
         g <- function(s,u,v){
             t <- mecdf(u,v+s)-s
             return(t[2])
         }
         a <- min(c(u,v))
         b <- max(c(u,v))
         c <- b-a
         z1 <- uniroot(f,low=-c,up=c,tol=0.00000001,u=u,v=v)
         z2 <- uniroot(g,low=-c,up=c,tol=0.00000001,u=u,v=v)
         z <- max(z1$root,z2$root)
         return(z)
         }

# Example
treated     <- rnorm(100,10,3)
untreated   <- rnorm(100,15,5)
# mecdf(untreated,treated)
Levy(group0=untreated, group1=treated)
```

REFERENCES

[1] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

[2] D'Agostino, RB Jr. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med*, 17:2265–2281, 1998.

[3] Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Weaknesses of goodness-of-fit tests for evaluating propensity score models: the case of the omitted confounder. *Pharmacoepidemiol Drug Saf*, 14(4):227–238, 2005.

[4] Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *J Clin Epidemiol*, 58:550–559, 2005.

[5] Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol*, 59:437–447, 2006.

[6] Rubin DB, Thomes N. Matching using estimated propensity score: relating theory to practice. *Biometrics*, 52:249–264, 1996.

[7] Rubin DB. On principles for modeling propensity scores in medical research (Editorial). *Pharmacoepidemiol Drug Saf*, 13:855–857, 2004.

[8] Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *JAMA*, 387:516–524, 1984.

[9] Bradley EL. *Overlapping coefficient, in: Encyclopedia of Statistical Sciences, vol. 6*. Wiley, New York, 1985.

[10] Inman HF, Bradley EL. The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. *Commun Statist -Theory Meth*, 18(10):3851–3874, 1989.

[11] Rom DM, Hwang E. Testing for individual and population equivalence based on the proportion of similar responses. *Stat Med*, 15:1489–1505, 1996.

[12] Stephens MA. Use of the Kolmogorov-Smirnov, Cramer-von Mises and related statistics without extensive tables. *J Royal Stat Soc Series B*, 32:115–122, 1970.

[13] Pestman WR. *Mathematical Statistics*. Walter de Gruyter, Berlin, New York, 1998.

[14] Lévy P. *Théorie de l'addition des variables aléatoires*. Gauthier-Villars, 1937.

[15] Zolotarev VM. Estimates of the difference between distributions in the Lévy metric. *Proc Steklov Inst Math*, 112:232240, 1973.

[16] Stine RA, Heyse JF. Non-parametric estimates of overlap. *Stat Med*, 20:215–236, 2001.

[17] Cleveland WS, Devlin SJ, Grosse E. Regression by local fitting. *J Econometr*, 37:87–114, 1988.

[18] Silverman BW. *Density estimation for statistics and data analysis*. Chapman and Hall: London, 1986.

[19] Wand MP, Jones MC. *Kernel Smooting*. Chapman and Hall, 1995.

[20] Mammitzsch V Gasser T, Müller H-G. Kernels for nonparametric curve estimation. *J Royal Stat Soc, Series B*, 47:238–252, 1985.

[21] Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. *Stat Med*, 2007. On line: DOI: 10.1002/sim.2781.

[22] Austin PC, Grootendorst P, Normand ST, Anderson GM. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. *Stat Med*, 26:754–768, 2007.

[23] Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Stat Med*, 26:734–753, 2007.

[24] Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, 71:431444, 1984.

[25] Gail MH. The effect of pooling across strata in perfectly balanced studies. *Biometrics*, 44:151163, 1988.

[26] Robins JM, Tsiatis AA. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Commun Statist -Theory Meth*, 20(8):2609–2631, 1991.

[27] Hauck WW, Anderson S, Marcus SM. Should we adjust for covariates in nonlinear regression analyses of randomized trials? *Control Clin Trials*, 19:249–256, 1998.

[28] Martens EP, de Boer A, Pestman WR, Belitser SV, Klungel OH. An important advantage of propensity score methods compared to logistic regression analysis. *in review*.

[29] Robins JM, Mark SD, Newey WK. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48:479–495, 1992.

[30] Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. *Am J Epideiol*, 150:327–333, 1999.

[31] Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol*, 158:280–287, 2003.

[32] Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection in propensity score models. *Am J Epidemiol*, 163:1149–1156, 2006.

[33] Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiol Drug Saf*, 13(12):841–853, 2004.

[34] Martens EP, de Boer A, Pestman WR, Belitser SV, Klungel OH. The use of the overlapping coefficient in propensity score methods. *in review*.

CHAPTER 6

DISCUSSION

HISTORICAL PERSPECTIVE

Statistical concepts and methods have been strongly developed in the last century with major contributions of Francis Galton, Karl Pearson, Ronald Fisher and Jerzy Neyman at the end of the 19^{th} and the beginning of the 20^{th} century. For instance the concept of the correlation coefficient can be traced back to Karl Pearson,[1] whereas regression analysis goes back to Francis Galton[2] and George Yule.[3] The introduction of two other important regression methods can be contributed to David Cox: *logistic regression* analysis in 1958[4] and the *proportional hazards model* in 1972.[5] All these regression-based techniques are still extensively used in many research areas for the prediction and explanation of various phenomena.

When the main objective is to estimate a single treatment effect, as is common in medical and pharmaceutical research, the *randomized experiment* is the gold standard. Very influential on the design of such experiments was the work of Fisher, who is in general credited for the invention of randomized experiment in 1925.[6,7] The concept of random assignment of treatments goes back to Neyman,[8,9] but even in 1885 randomization was used by Charles Peirce.[10]

When an experiment is not possible or when subjects are not randomly assigned to treatments, it can still be the main objective to estimate a single treatment effect or exposure and to adjust for the *confounding* influence of other factors that are prognostic for the outcome. It is common practice to use the earlier mentioned conventional regression-based methods for this purpose, but alternative methods have specifically been developed to adjust for confounding. Two of these methods are investigated in more detail in this thesis: the method of *propensity scores* and *instrumental variables*. The first is a relatively recent method, developed by Rosenbaum & Rubin in 1983.[11] It has its fundament in matching on a continuous variable in the work of Rubin[12,13] and Cochran & Rubin.[14] The other main topic of this thesis is the method of *instrumental variables*. The related problem of *solving the identification problem* in simultaneous equation models goes back to Philip Wright in 1928,[15,16] whereas the term instrumental variable first appeared in work of Reiersøl in 1945.[17,18] The first appearance in medical research was probably in 1989 by Permutt and Hebel.[19]

The development and the improvement of methods to adjust for confounding are important, because results from observational studies can only be used to inform clinical practice if the effect estimates of treatment are valid. In the last decade an *increasing use* of propensity scores in medical studies can be noticed, but still there are questions on how these methods are best applied in different settings. Although instrumental variables as an adjustment method is less known and in general *less applicable*, the same is true for this method: how and when to apply this method. Therefore, the aim of this thesis was to assess the strengths and limitations of alternative adjustment methods (Chapter 2), to compare these methods to conventional regression-based methods (Chapter 3), to demonstrate less straightforward applications (Chapter 4) and to further develop these alternative methods (Chapter 5).

WHY TO USE ALTERNATIVE ADJUSTMENT METHODS?

To adjust for confounding in non-randomized studies the factor of interest and all possible confounders are usually included in a regression model. Alternatively one can use methods of propensity scores and instrumental variables to adjust for confounding. Contrary to linear, logistic or Cox proportional hazards regression, these methods have been developed with a randomized controlled trial in mind, which is the preferred research design to estimate intended treatment effects. Propensity score methods and instrumental variables have one variable of main interest (the treatment or exposure variable) and are primarily concerned with *similarity of treatment groups*. This is not true for regression-based methods in which all variables, confounders and treatment variable, have technically the same place in the outcome model. Although methods of propensity scores and instrumental variables can also use regression methods to finally estimate treatment effects, the philosophy is quite distinct from directly estimating adjusted treatment effects with a linear, logistic or Cox proportional hazards regression model.

PERFECT SIMILARITY

Important questions in medical research are whether a certain drug is effective to prevent or cure a specific disease and whether exposure to some environmental factor influences health. Such questions are primarily directed to find an *average causal effect* of a factor of interest, the treatment or exposure variable. One hypothetical way to answer such questions is to observe all individuals of a certain sample in two different states at the same moment. For instance, observe the cholesterol level when a patient is treated by a drug and compare it to its level when the *same patient* is not treated by this drug over the *same time period*. The direct causal effect of treatment for all individuals will then be exactly known because all other possible explanations except treatment did not change. Unfortunately it is physically impossible to measure the same person at the same time in two different states, treated and untreated, or exposed and not exposed. This problem could be solved by using two exactly identical individuals in order to observe this pair at the same time, one as treated and the other as untreated. For experiments with animals two identical (e.g. genetically) subjects could be used to be assigned to two different treatments. However, for humans this is virtually impossible, even when we restrict similarity to only those factors that are prognostic for the outcome.

AVERAGE SIMILARITY ON GROUP LEVEL

To make a valid assessment of an average treatment effect it is too restrictive to demand that *individuals* should be similar between both treatment groups. It should be sufficient for assessing an average treatment effect to reach on average similarity on group level. This can be achieved by a procedure known as *randomization*, a term contributed to Fisher.[6,20] Two groups of patients are on average similar on all characteristics if the assignment to these groups was completely at random, or in other words, if all individuals had the same probability to be in

one of the treatment groups. Nowadays this procedure is the scientific standard for medical research of interventions. This procedure assumes that the researcher has *control over treatments* or exposures in order to let the toss of a coin decide which of the treatments are given to the individuals.

NO CONTROL OVER TREATMENTS OR EXPOSURES

Unfortunately it is not always possible for researchers to have control over treatments, exposures or, more general, the factor of interest. This means that when observing a certain population many other factors can be explanations for the average difference in outcome between treatment groups, because treatment groups do not differ only by treatment. To adjust for factors that differ between treatment groups and at the same time are prognostic for the outcome, a distinction can be made by three possible approaches. The first is a regression-based approach in which all confounders and the treatment are included in the same model. The second is to create similarity of treatment groups within subcategories based on the confounders (propensity score methods) and the third is to identify groups that are similar with regard to confounders, but differ with regard to treatment (method of instrumental variables). These approaches are further explained in the next paragraphs.

REGRESSION-BASED APPROACH

A first approach is based on the simple general principle *'if you can't beat them, join them'*. If you can not get rid of competing, alternative explanations beforehand by controlling treatments or exposures, then combine them in a joint model with the factor of interest in order to adjust for their confounding influence on the treatment effect estimate. Such statistical models are for instance linear regression analysis, logistic regression analysis and Cox proportional hazards regression. Although one is primarily interested in the treatment or exposure effect, such models *simultaneously* estimate an adjusted treatment effect as well as all individual effects of the potential confounders in the model.

CREATE SIMILARITY OF GROUPS

A second approach of handling the problem of confounding is based on the idea that the undesirable differences between treatment groups are due to *different probabilities* of belonging to one of the treatment groups. If all factors that both influence treatment and outcome are known, these probabilities (the true *propensity scores*) can be determined. For individuals or groups of individuals who have the same probability to be treated, it is as if the toss of a coin has decided who was actually treated and who was not. In practice, all true confounders are unknown and the propensity scores have to be estimated with only *observed confounders*. This implies that the propensity score does not replace a randomized experiment because adjustment for unobserved factors is not possible. Using the estimated propensity score to create *subgroups or pairs* of subjects, the original comparison between *all* treated and *all* untreated subjects is now

replaced by a comparison within those subgroups or pairs. Although regression-based methods are also frequently used in propensity score methods, it is different from the first approach because in a propensity score model confounders are *not directly related to the outcome*.

USE RELATED 'TREATMENT GROUPS'

A third approach is based on the idea that if we can not directly relate treatment or exposure to outcome without facing the influence of other factors, we create or identify *a slightly different 'treatment' variable* that has *not* been influenced by other factors and is related to the original treatment of interest. For example, the original exposure to smoking can not be controlled by the researcher, but the related variable 'encourage to stop smoking' can be controlled by randomization before collecting the data on smoking.[19] In this situation the method of *instrumental variables* can be used, where the 'alternative treatment variable' is called the instrument or instrumental variable. For treatments with only two classes, this is similar to saying that it is not necessary for the treated group to contain *only treated* individuals and for the untreated group to contain *only untreated* individuals in order to compare these groups. The amount of 'noise' or better, the association between the original and alternative treatment groups, should be known in order to use this method for validly assessing treatment effects. Application of this method is not limited to situations in which a suitable instrumental variable has to be *created* by the researcher (as in the previous example of encouragement to stop smoking), but also in situations where a variable is available in the data and can be used as instrumental variable. Examples of such variables are the distance to a hospital that performed cardiac catheterizations[21] and the physician-specific prescribing preference to certain drugs.[22]

When the assumptions of this method are fulfilled it has the potential to adjust for *all confounders*, whether observed or not. This clearly distinguishes this approach from the other two in which it is only possible to adjust for observed confounders.

STRENGTHS AND LIMITATIONS OF ADJUSTMENT METHODS

The first approach in which the factor of interest and all possible confounders are included in a regression model, is still the standard in observational studies, while the methods of propensity scores and instrumental variables can be considered as *alternative methods* to adjust for confounding. Whenever a conventional regression-based technique can be adopted, the method of propensity scores is also applicable. This is not the case for instrumental variable methods, because at least one suitable instrumental variable is needed. An important strength of both alternative methods is that these are developed with a randomized controlled experiment in mind. That means that before relating treatment to outcome, these methods direct their attention towards the relationship between treatment and potential confounders. An advantage of regression-based methods is that also the effects of other factors on the outcome can be estimated instead of just the effect of treatment. However, when the focus is to estimate a valid

treatment effect, the effect estimates of other factors are usually not of interest.

A specific advantage of propensity score methods is its *transparency* on the similarity of treatment groups, which informs the user whether the method was successful in creating this similarity. Another advantage is the *larger number of confounders* that can be adjusted for in the analysis compared to regression-based methods, because only the factor of interest and the propensity score are used in the final outcome model.

In Section 3.1 and 3.2 we pointed at another, frequently overlooked advantage of propensity scores when dealing with a dichotomous outcome. With such data it is common to use logistic regression (or Cox proportional hazards regression with survival data) and express treatment effects as odds ratios (or hazard ratios). The important advantage of propensity scores is that the treatment effect estimator is closer to the true average treatment effect than in logistic or Cox proportional hazards regression. The reason is that in a multivariable logistic regression or Cox proportional hazards regression analysis the effect of treatment, averaged over for instance men and women, is *not equal* to the treatment effect for the whole population, even when the proportion of men is similar in both treatment groups. The difference is systematic and can lead to a serious overestimation of the average treatment effect, especially when the number of prognostic factors is more than 5, the treatment effect is larger than an odds ratio of 1.25 (or smaller than 0.8) or the incidence proportion is between 0.05 and 0.95.

The most important advantage of instrumental variable methods is that it adjusts for *all possible confounders*, whether observed or not. In fact, this is a similar objective as in randomized experiments, but is in observational studies rather ambitious. In situations where little information on confounding factors is available, instrumental variables might be considered. The price to pay are the strong assumptions for an instrumental variable to be valid, which means that 1. the instrumental variable should have no direct causal effect on the outcome, 2. its categories are similar with respect to other characteristics of individuals (for instance by randomization), and 3. there exists a relationship between the instrumental variable and the original treatment variable that should not be weak. While it is difficult to test whether the two first assumptions are fulfilled, the third assumption implies also that a strong instrument is wanted. In Section 3.3 we showed that in cases of strong confounding (for instance confounding by indication), a strong valid instrumental variable *can not be found* because of its inherently strong relationship with the confounders. One should either rely on a weak instrument or one is likely to violate the first assumption. This limitation is inherent to this method.

IMPROVEMENT OF PROPENSITY SCORE METHODS

In applying the method of propensity scores (Section 4.1) we recognized that in the methodological literature and in applications the step of creating and checking the balance between treatment groups has not been given much attention. Often no effort has been made to improve a certain propensity score model in order to reach a better balance on prognostic factors and

consequently a better estimate of the treatment effect. Therefore, we focused in Chapter 5 on the similarity or balance on confounders between treatment groups. In Section 5.1 and 5.2 we introduced different ways to *measure the amount of balance* in propensity score methods to inform the reader about the quality of the model and to help the researcher to choose among a number of possible propensity score models. One of these measures is the *overlapping coefficient*, which suits the objective of propensity scores: it directly measures the overlap in two distributions and that is exactly what this method tries to achieve. The better the overlap of covariate distributions within strata of the propensity score, the better the balance on those distributions. When compared to a reference distribution of expected values of the overlapping coefficient, we showed that this measure can be used to assess whether sufficient balance on covariates has been reached by propensity score modelling. In simulation studies we showed that there exists an inverse relationship between the weighted average overlapping coefficient (the balance) and the bias that remains after adjustment. For larger data sets this relationship is stronger than for other commonly used methods to check the balance. Compared to simple propensity score models a better reduction of the mean squared error can be reached when the overlapping coefficient is used to improve the propensity score model. We also studied some alternative ways of measuring balance (Section 5.2), i.e. the Kolmogorov-Smirnov distance and the Lévy metric, of which the first has, in the chosen setting, similar characteristics as the overlapping coefficient.

IMPLICATIONS AND FUTURE RESEARCH

This work is relevant to researchers in medical sciences and many other disciplines who face the problem of confounding. We showed the advantages, limitations and applications of two alternative methods to support researchers who want to apply these methods. For the method of instrumental variables we focused on the strength of the correlation between the instrumental variable and treatment: this correlation ought to be strong, but has a maximum which can be a problem for using this method in case of strong confounding. For propensity scores we pointed at a general overlooked advantage of the method compared to logistic regression analysis and Cox proportional hazards regression: it gives in general treatment effect estimates that are closer to the true average treatment effect.

In this thesis there are several leads for future research. One intriguing question is the effect of the *simulation procedure* on the results. In Section 3.2 and 5.1 a method was used in which the average marginal effect was known in advance and where weights were used to create confounding. In Section 5.2 the average marginal effect was calculated using the a model-based procedure described by Austin.[23] More research is needed to compare both simulation procedures in order to assess its influence on the results.

A limitation of our work is that we mainly used *stratification* on the propensity score although also matching, covariate adjustment or inverse probability weighting are also possible ways

to use the propensity score. This choice was partly based on the literature and on the study in Section 4.1, but there are still no convincing and conclusive research results that indicate which method is to be preferred. Conceptually stratification or matching is preferred, but more research is needed to investigate the situations in which these methods are favored.

Another limitation is the *fixed number of five strata* that were mostly used in this thesis for stratification on the propensity score. In Section 4.1 also 7 and 10 strata were explored, but the information is still insufficient for general remarks on the use of the number of strata. Obviously, the number of strata should be dependent on the number of observations, but guidelines are not available for determining the optimal number of strata. In Chapter 3 and 5 we conformed ourselves to the convention of using five strata.

In our work we recognized another 'problem' when propensity score methods are evaluated with simulated data. This is the situation in which there are *no or only a few* treated or untreated subjects in one of the strata. Such situations are difficult to handle in simulation studies because then results apply only to a subgroup of the original population. Comparison with another method that estimates a treatment effect for the whole population, will be difficult. In real data sets those situations are common when strong confounding is present and more research is needed to provide guidelines in case strata are 'empty' for one the groups.

In Sections 5.1 and 5.2 we mainly concentrated on *continuous covariates*, although the overlapping coefficient also can be calculated for dichotomous covariates. How the overlapping coefficient will behave in case of a mixture of continuous and dichotomous covariates, could be subject for further research.

For the method of instrumental variables we recognized that application in survival analysis is scarce (Section 4.2). We concentrated on *differences in survival probabilities* and did not estimate the more common hazard ratio using instrumental variables. More research and more medical research examples are needed concerning instrumental variables in general and the use with survival analysis in particular.

Except for Section 2.1 in which many methods are reviewed, we only focused on *propensity scores* and *instrumental variables* as methods to adjust for confounding. The most important reasons for this choice are that these methods are increasingly used in the medical literature in the last decade and that there are several questions to be answered when applying these methods. This choice does not mean that other possible methods are not interesting or not applicable. We mention for example *sensitivity analyses*, *propensity score calibration*, *G-estimation*, propensity score methods when treatment is *time dependent* and the application of propensity score methods in *case-control studies*.

Another subject that is not covered by this thesis is a *simulation study* in which methods of propensity scores and instrumental variables *are compared*. This is an interesting challenge for future research, although practical situations in which both methods are applicable are limited.

CONCLUSION

In conclusion, propensity scores to adjust for confounding have several advantages compared to conventional regression-based techniques. One of these is that the estimated treatment effect is closer to the true average treatment effect, mainly when there are numerous confounders, the treatment effect is substantial and incidence proportions are not too low. Therefore, this method should be considered more often when adjustment for confounding is needed. When propensity scores are considered, more attention should be given to the building of the propensity score model, based on a measure of balance such as the overlapping coefficient. For instrumental variables, researchers should be more aware of the possibility of using this method to adjust for confounding, while awareness of its limitations is equally important. Future research should be directed to further assess in different situations how propensity scores perform and to formulate concrete recommendations on when and how to apply this method. With respect to instrumental variables, it would be helpful when good examples of applications come available as researchers realize that randomization could be used in non-randomized settings. Also simulation studies that show the valid treatment effect estimates that can be reached, are important for future development and application of this method.

REFERENCES

[1] Pearson K. Regression, heredity and panmixia. *Phil Transactions of the Royal Society of London, Series A*, 187:253–318, 1896.

[2] Galton F. Types and their Inheritance [Presidential address, Section H, Anthropology]. *Nature*, 32:507–510, 1885.

[3] Yule GU. On the theory of correlation. *J Royal Stat Society*, 60:812–854, 1897.

[4] Cox DR. The regression analysis of binary sequences. *J Royal Stat Society Series B*, 20:215–242, 1958.

[5] Cox DR. Regression models and life tables. *J Royal Stat Society Series B*, 34:187–220, 1972.

[6] Fisher RA. *Statistical Method for Research Workers*. Oliver and Boyd, Edinburgh, 1925.

[7] Fisher RA. The arrangement of field experiments. *J Ministry Agric*, 33:503–513, 1926.

[8] Neyman J. On the application of probability theory to agricultural experiments. essay on principles. Section 9. *Roczniki Nauk Roiniczych, Tom X*, 19:1–51, 1923. Reprinted in *Statistical Science* 1990;5:463-480, with discussion by T. Speed and D. Rubin.

[9] Neyman J. Statistical problems in agricultural experimentation. *Sup J Royal Stat Society*, 2:107–180, 1935.

[10] Peirce CS, Jastrow J. On small differences in sensation. *Memoirs of the National Academy of Sciences*, 3:73–83, 1885. Reprinted in Burks AW (ed.). Collected Papers of Charles Sanders Peirce. Cambridge: Harvard University Press, 1958; 7:1334.

[11] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

[12] Rubin DB. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 29:184203, 1973.

[13] Rubin DB. *The use of matched sampling and regression adjustment in observational studies (Ph.D. Thesis)*. Department of Statistics, Harvard University: Cambridge, 1970.

[14] Cochran WG, Rubin DB. Controlling bias in observational studies: a review. *Sankhya-A*, 35:417–446, 1973.

[15] Wright PhG. *The tariff on Animal and vegetable oils*. Macmillan, New York, 1928.

[16] Stock JH, Trebbi F. Who invented IV regression? *J of Economic Perspectives*, 17:177–194, 2003.

[17] Reiersøl O. Confluence analysis by means of instrumental sets of variables. *Arkiv for Mathematik, Astronomi och Fysik*, 32:1–119, 1945.

[18] Aldrich J. Reiersøl, Geary and the idea of instrumental variables. *Economic and Social Review*, 24:247–274, 1993.

[19] Permutt Th, Hebel JR. Simultaneous-equation estimation in a clinical trial of the effect of smoking on birth weight. *Biometrics*, 45:619–622, 1989.

[20] Fisher RA. *The design of experiments*. Oliver and Boyd, Edinburgh, 1935.

[21] McClellan M, McNeil BJ, Newhouse JP. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *JAMA*, 272:859–866, 1994.

[22] Brookhart MA, Wang PS, Solomon DH, Schneeweiss S. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiol*, 17:268275, 2006.

[23] Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. *Stat Med*, 2007. On line: DOI: 10.1002/sim.2781.

SUMMARY

INTRODUCTION

A study in which the subjects are randomly assigned to the factor of interest, for instance a certain drug treatment, is called a *randomized controlled trial* and is generally accepted as the best way to assess the effect of such treatment. Studies that lack such random assignment but also aim at estimating the effect of a treatment (or exposure) are called *observational studies*. The evidence in assessing treatment effects from observational studies will be in general less convincing than from well conducted randomized experiments because prognostic factors are in general *not equally distributed* across treatment groups. Despite this deficit, observational studies can certainly be valuable for assessing the effect of treatments, for instance the long-term effects of drugs already proven effective in short-term randomized studies or the effects for patients that were excluded from randomized studies. Therefore, the focus of observational studies should lie on *the adjustment of the treatment* effect for the disturbing influence of *confounding factors*. There are different methods available that seek to adjust for such confounding.

AN OVERVIEW OF METHODS

As is discussed in Chapter 2, these methods consist mainly of methods that concentrate on the *design* of the study and *analytical methods* that estimate an adjusted treatment effect. *Restriction* to a certain subpopulation and specific designs like *case-crossover* and *case-time-control* designs are examples of adjustment methods in the design phase of the research. Traditional analytical methods include firstly standard methods like *stratification* and *matching*, but also multivariable statistical methods such as *(logistic) regression* and *Cox proportional hazards regression*. In these methods, measured covariates are added to a model with 'treatment' as explaining factor, in order to estimate a treatment effect that is adjusted for the confounding influence of the added covariates. That means that no specific effort has been made to directly correct the inequalities of covariate distributions between treatment groups. A method that do focus on the balance of covariates between treatment groups, is the method of *propensity scores*. First, a propensity score model is created to estimate the conditional probability of being treated given the covariates. In the second step an adjusted treatment effect is estimated, using the propensity score as matching variable, as stratification variable, as continuous covariate or inverse probability weight.

In all these techniques, an important limitation is that adjustment can only be achieved for *measured* covariates, so that no correction takes place for other possibly important, unmeasured covariates. A method not limited by these shortcomings is a technique known as *instrumental variables*. In this approach, the focus is on finding a variable (the instrument) that is related to treatment, and is related to outcome only because of its relation to treatment. This technique can achieve the same effect as randomization in bypassing the usual way in which

physicians allocate treatment according to prognosis, but its rather strong assumptions limit its use in practice. Related techniques are *two-stage least squares* and the *grouped-treatment approach*, sharing the same limitations.

Except for this general overview of methods this thesis is limited to the methods of propensity scores and instrumental variables. We discussed strengths and limitations, special applications and improvements of propensity score methods.

STRENGTHS AND LIMITATIONS OF ADJUSTMENT METHODS

IMPORTANT ADVANTAGE OF PROPENSITY SCORES

In Section 3.1 and Section 3.2 of Chapter 3 we compared the use of logistic regression analysis and propensity score methods. In medical research both methods are applied to estimate an adjusted treatment effect from observational studies. Although in the literature the effect estimates of both methods are classified as *'similar'* and 'not substantially different', we stressed that differences are *systematic* and can be *substantial*. With respect to the objective to adjust for the imbalance of covariate distributions between treatment groups, we illustrated that the estimate of propensity score methods is in general *closer to the true treatment effect* of interest than the estimate of logistic regression analysis. The advantage can be substantial, especially when the number of prognostic factors is more than 5, the treatment effect is larger than an odds ratio of 1.25 (or smaller than 0.8) or the incidence proportion is between 0.05 and 0.95. This implies that there is an advantage of propensity score methods over logistic regression models that is frequently *overlooked* by analysts in the literature.

This overestimation of treatment effects in logistic regression analysis is due to the use of the *odds ratio* as measure for treatment effect. The same is true for the hazard ratio, used in for instance Cox proportional hazards regression, but do not apply to other less frequently used effect measures like the risk difference. The advantage of propensity scores is that the odds ratio can still be used, but that the *overestimation* of the true average treatment effect is *smaller*, independent of whether stratification, matching or covariate adjustment is used. This advantage will disappear when propensity score methods are combined with further adjustment for confounding by entering some or all covariates separately in the model.

We conclude that propensity scores are in general better suited to estimate an average treatment effect than logistic regression analysis (by means of the odds ratio) and Cox proportional hazards regression (by means of the hazard ratio).

LIMITATION OF INSTRUMENTAL VARIABLES

In Section 3.3 we focused on the method of instrumental variables for its ability to adjust for confounding in non-randomized studies. An important strength of this method is its potential ability to *adjust for all confounders*, whether observed or not. Its weakness lies in the

main assumption, which states that the instrumental variable should influence the outcome neither directly, nor indirectly by its relationship with other variables than the treatment variable. Whether this assumption is valid can be argued only theoretically, and cannot be tested empirically.

We further focused on the *correlation* between the instrumental variable and the treatment (or exposure). When this correlation is very small, this method will lead to an increased standard error of the estimate, a considerable *bias when sample size is small* and a bias even in *large samples* when the main assumption is only slightly violated. Furthermore, we demonstrated the existence of an *upper bound* on the correlation between the instrumental variable and the exposure. This upper bound is not a practical limitation when confounding is small or moderate because the maximum strength of the instrumental variable is still very high. When, on the other hand, considerable *confounding by indication* exists, the maximum correlation between any potential instrumental variable and the exposure will be quite low, resulting possibly in a fairly *weak instrument* in order to fulfill the main assumption. Because of a trade-off between violation of this main assumption and the strength of the instrumental variable, the presence of considerable confounding and a strong instrument will probably indicate a *violation* of the main assumption and thus a biased estimate.

We conclude that the method of instrumental variables can be useful in case of moderate confounding, but is less useful when strong confounding (by indication) exists, because strong instruments can not be found and assumptions will be easily violated.

APPLICATION OF ADJUSTMENT METHODS

Propensity score methods and to a lesser extent methods that use an instrumental variable, are increasingly applied in the medical literature. In Chapter 4 we applied both adjustment methods on *survival data*.

PROPENSITY SCORES AND SURVIVAL DATA

In Section 4.1 we showed the application of propensity score methods in a study on the effect of treating hypertension for prevention of stroke. We also compared three propensity score methods for their ability to adjust for confounding and compared the results with the traditional Cox proportional hazards regression. The number of events in our data set was low, which gives an advantage to propensity score methods for its ability to handle more covariates.

From literature reviews it is known that in most propensity score applications covariate adjustment is used (which is not recommended), the selection of the propensity score model is routinely performed by entering all covariates and attention for and reporting of the attained balance are in general *insufficient*. We showed some methods to *check and report the balance*, but recognized that there is no uniform method that is used in practice. Finally we included a non-confounder and concluded that the treatment effect was less sensitive for stratifying on the

propensity score than for Cox proportional hazards regression. This confirms the results found in Section 3.2.

We conclude that application of propensity score methods is worthwhile when survival data are involved, but more attention should be given to the model building and balance checking phases.

INSTRUMENTAL VARIABLES AND SURVIVAL DATA

Although the application of instrumental variable methods finds its way into medical literature as an adjustment method for unobserved confounding in observational studies and in randomized studies with all-or-none compliance, it is hardly applied in case of *censored survival data*. In Section 4.2 we applied this method by using the *difference in survival probabilities* as the treatment effect of interest. We used a data set in which the survival times are compared between patients with type-1 diabetes that had a kidney transplantation alone with those who had a combined kidney-pancreas transplantation. A direct comparison of these treatment methods, as has been given by the *as-treated* estimate, can be considered as clearly biased because of selection processes. The initially reported *intention-to-treat* estimate rather compares treatment policies between hospitals and is substantially smaller than when *instrumental variables* is used. This indicates that the intention-to-treat effect *dilutes the differences* between both treatment methods. The standard errors with the method of instrumental variables were also *larger*.

As is inherent to instrumental variable methods the effect estimate is quite sensitive to the underlying assumptions. An additional weakness of instrumental variables application with survival data is the *number of events*, which can be quite low when rare events are studied. Furthermore, when time passes, the estimation of the survival curve will be *less reliable* because the risk set is reduced, which is an even larger problem when an instrumental variable is used.

We conclude that instrumental variable methods can be considered when survival data are involved, but its application involves considerable attention to its main assumptions. Standard errors of the estimates should be reported, mainly when the number of events is low, and can be quite high at the end of the survival curve.

IMPROVEMENT OF PROPENSITY SCORE METHODS

As was apparent from literature reviews and from the application of the method in Section 4.1, there is need for information how to apply propensity score methods, mainly when it concerns the building of the propensity score model and the check for balance on covariates. In Chapter 5 we focused on *measures for balance* in order to have a uniform measure for balance for reporting purposes and to be able to select objectively the propensity score model that balances covariates between treatment groups.

THE OVERLAPPING COEFFICIENT

In Section 5.1 we proposed the use of the *overlapping coefficient*, an existing measure for the overlap between two distributions. This measure can be used within strata of the propensity score to indicate the *balance* that has been reached on covariates by applying a propensity score model. The measure can be compared with an *expected value* to have an idea about the quality of the correction by means of the propensity score. This information is of crucial importance in order to interpret the reported adjusted treatment effect. Often only statements like "the treatment effect has been adjusted for x_1, x_2, x_3, ..." can be found in the literature and information on *the extent of adjustment* is missing. This measure can also be a help for model selection, because we showed its inverse association with bias in a simulation study. The *weighted average overlapping coefficient* calculated on the set of available covariates show strongest association with bias for larger data sets. For smaller data sets the *p*-values from significance tests and the *c*-statistic have higher predictive power for the bias than the overlapping coefficient.

We conclude that the use of the overlapping coefficient can be useful in propensity score methods. For smaller sample sizes the method does not seem to work well because of the difficulties in estimating the densities and therefore the overlap.

MORE MEASURES FOR BALANCE

A direct quantification of the amount of balance in propensity score methods is an attractive property of the overlapping coefficient. In Section 5.2 we also looked for some *other measures* that could possibly be used as measures for balance and to indicate the importance of the stage of creating the propensity score model. In a simulation study (which had a different setup than in Section 5.1) we compared the overlapping coefficient with the *Kolmogorov-Smirnov distance* and the Lévy metric. For all three measures we showed an inverse association with bias, which was strongest for the weighted average overlapping coefficient. In general this association was stronger for larger than for smaller samples. For the overlapping coefficient de correlation was $R = -0.06$ for n=400 and $R = -0.63$ for $n = 2,000$. Selecting the propensity score model with the overlapping coefficient seems to be *most effective*, because the mean squared error for this method was in general smallest (ranging from 0.031 to 0.197). The differences with the Kolmogorov-Smirnov distance and the Lévy metric were only minor. A propensity score model that contained *all covariates* had a considerable larger mean squared error, while the propensity score model that contained all true, but usually unknown, confounding factors had somewhat larger mean squared error.

We conclude that the use of the presented measures be useful in propensity score methods. We prefer the overlapping coefficient because of its easy interpretation and the slightly smaller mean squared error when estimating a treatment effect. Although estimation of the overlapping coefficient is more difficult for smaller sample sizes, the choice of model by using the overlapping coefficient will give better results than when simply all covariates are used in the propensity score model (in practice often seen).

FINALLY

In conclusion, propensity scores to adjust for confounding have several advantages compared to conventional regression-based techniques. An important advantage is that propensity score methods give treatment effects that are in general closer to the true average treatment effect than logistic or Cox proportional hazards regression analysis. Therefore, this method should be considered more often when adjustment for confounding is needed. Thereby, more attention should be paid to the way how the propensity score model is built and to measuring and reporting the amount of balance on covariates, for instance by using the overlapping coefficient.

The method of instrumental variables is much less common in epidemiological research. It should attract more the attention of researchers for its ability to completely adjust for confounding. At the same time its limitations will prevent a general application of the method. Good examples of instrumental variables applications in the medical literature will help to show its merits.

SAMENVATTING

INLEIDING

Als je geïnteresseerd bent in de vraag welke van twee behandelingen of geneesmiddelen het beste werkt, zijn er grofweg twee manieren om dat te onderzoeken: *experimenteel* en *observationeel*. Bij een experiment heeft de onderzoeker controle over wie welke behandeling krijgt. Vaak zal de onderzoeker dat *door het toeval* laten bepalen (randomiseren), zodat er gemiddeld genomen geen andere factoren van invloed zijn op de relatie tussen behandeling en uitkomst. Dit houdt in dat de behandelingsgroepen gemiddeld genomen 'gelijk' zijn en dus een directe relatie kan worden gelegd tussen behandeling en uitkomst. Deze manier van onderzoek is over het algemeen geaccepteerd als de beste methode om het effect van een behandeling of geneesmiddel vast te stellen.

Bij een observationele studie heeft de onderzoeker *geen invloed* op het toewijzen van behandelingen aan de personen in het onderzoek. Het directe gevolg hiervan is dat de behandelingsgroepen niet alleen 'behandeling' als verschil hebben, maar ook op *andere kenmerken* zullen verschillen. Hierdoor is het veel lastiger om het directe effect van de behandeling te bepalen, met name als deze kenmerken ook invloed hebben op de uitkomst (*prognostische factoren*). Bijvoorbeeld: behandeling A wordt vaker aan jongeren gegeven en behandeling B wat vaker aan ouderen. Als blijkt dat behandeling A over het geheel beter werkt, hoeft dat niet alleen aan de behandeling te liggen, maar kan dat ook komen doordat jongeren over het algemeen gezonder zijn dan ouderen. In zo'n geval is het effect van de behandeling op de uitkomst *verstrengeld* of *verward* met het effect van leeftijd op de uitkomst. De Engelse term hiervoor is *confounding*.

Ondanks de *verstrengeling* of *verwarring* met mogelijke andere factoren, kunnen observationele studies van grote waarde zijn om het effect van een behandeling te schatten, zeker als experimentele studies niet mogelijk zijn. Dit is bijvoorbeeld bij de lange-termijneffecten van medicijnen of bij patiënten die bij de experimentele studies zijn uitgesloten van deelname. Als observationele studies inderdaad gebruikt worden om het effect van een behandeling te schatten, zal de belangrijkste aandacht moeten uitgaan naar het *ontwarren* van behandelingseffect en effecten van verstorende factoren. Hiervoor zijn verschillende *correctie-* of *ontwarringsmethoden* beschikbaar die in Hoofdstuk 2 zijn besproken. Twee daarvan zijn in de andere hoofdstukken nader aan de orde geweest.

EEN OVERZICHT VAN CORRECTIEMETHODEN

Deze correctiemethoden kunnen in twee groepen worden verdeeld: methoden waarbij *voorafgaand aan de studie* wordt gecorrigeerd en methoden die corrigeren *nadat de data is verzameld*. *Restrictie* en specifieke onderzoeksdesigns zoals *case-crossover* en *case-time-control* designs zijn voorbeelden van correctiemethoden voorafgaand aan de studie. Tot de traditionele (analytische) methoden die achteraf corrigeren behoren naast standaardmethoden als *stratificatie*

en *matching*, ook de multivariabele statistische methoden *logistische regressie* en *Cox proportional hazards regressie*. Bij deze methoden worden geobserveerde variabelen toegevoegd aan een model met alleen behandeling als verklarende factor, met als doel het geschatte behandelingseffect te corrigeren voor de verstorende invloed van de toegevoegde variabelen (covariaten). Dit betekent dat hierbij geen speciale aandacht wordt besteed aan de ongelijke verdelingen van covariaten tussen behandelingsgroepen. Een methode die zich wel richt op de balans in covariaten tussen behandelingsgroepen is de methode van *propensity scores*. Eerst wordt een propensity score model gemaakt om de conditionele kans te schatten dat een individu tot de behandelde groep behoort (t.o.v. onbehandelde groep), gegeven een set covariaten. Vervolgens wordt dan een gecorrigeerd behandelingseffect geschat waarbij de propensity score wordt gebruikt als matchingvariabele, stratificatievariabele, covariaat of als gewicht (inverse probability weight).

Bij al deze methoden is een belangrijke beperking dat correctie alleen plaatsvindt voor variabelen die ook daadwerkelijk zijn gemeten en dat voor andere belangrijke ongemeten covariaten niet kan worden gecorrigeerd. Een methode die deze beperking niet kent, is de methode van *instrumentele variabelen*. Bij deze methode ligt het accent op het vinden van een variabele (het instrument) die samenhangt met de variabele 'behandeling' en die alleen aan de uitkomstvariabele gerelateerd is via de samenhang met 'behandeling'. Deze methode kan hetzelfde effect bereiken als het randomiseren in een experiment, maar de sterke aannames beperken het gebruik in de praktijk. Gerelateerde methoden met een soortgelijke beperking zijn *two-stage least squares* and *grouped-treatment approach*.

In de rest van de hoofdstukken hebben we ons beperkt tot de methoden van propensity scores en instrumentele variabelen. Hiervan hebben we de sterke en zwakke punten en speciale toepassingen besproken en hebben een balansmaat voorgesteld bij gebruik van propensity score methoden.

STERKE EN ZWAKKE PUNTEN VAN CORRECTIEMETHODEN

BELANGRIJK VOORDEEL VAN PROPENSITY SCORES

In Paragraaf 3.1 en Paragraaf 3.2 van Hoofdstuk 3 is het gebruik van logistische regressieanalyse vergeleken met propensity score methoden. In medisch onderzoek worden beide methoden toegepast om een gecorrigeerd behandelingseffect te schatten in observationele studies. Hoewel de effectschattingen van beide methoden in de literatuur als 'vergelijkbaar' en 'niet echt verschillend' worden bestempeld, is in deze paragrafen aangetoond dat de verschillen *systematisch* zijn en *wel substantieel verschillend* kunnen zijn. Met betrekking tot het doel om voor ongelijkheid in de verdelingen van covariaten te corrigeren, is laten zien dat het geschatte behandelingseffect bij propensity score methoden over het algemeen *dichter bij het werkelijke te schatten effect* ligt dan bij logistische regressie-analyse. Het verschil kan groot zijn, vooral

als het aantal prognostische factoren groter is dan 5, het behandelingseffect groter is dan een odds ratio van 1.25 (of kleiner dan 0.8) of als de incidentie groter is dan 5%. Dit betekent dat er een voordeel is van propensity score methoden vergeleken met logistische regressiemodellen, hetgeen door onderzoekers en in de literatuur vaak over het hoofd wordt gezien.

Deze overschatting van het behandelingseffect in logistische regressie-analyse komt door het gebruik van de *odds ratio* als effectmaat. Hetzelfde geldt voor de *hazard ratio* zoals deze bijvoorbeeld gebruikt wordt in Cox proportional hazards regresie; het geldt niet voor minder vaak gebruikte effectmaten zoals het risicoverschil. Het voordeel van propensity score methoden is dus dat de odds ratio kan worden gebruikt zonder dat daarbij het werkelijke gemiddelde behandelingseffect ernstig wordt overschat, zoals bij logistische regressie het geval kan zijn. Deze bevinding is onafhankelijk van de manier waarop de propensity score wordt gebruikt (voor stratificatie, voor matching of als covariaat). Dit voordeel verdwijnt echter als naast correctie via de propensity score ook nog extra gecorrigeerd wordt door covariaten apart in het uitkomstmodel op te nemen.

We concluderen dat de methode van propensity scores over het algemeen *beter geschikt is* om een gemiddeld behandelingseffect te schatten dan logistische regressie-analyse en Cox proportional hazards regressie.

BEPERKING VAN DE METHODE VAN INSTRUMENTELE VARIABELEN

In Paragraaf 3.3 is gekeken naar de mogelijkheden om via de methode van instrumentele variabelen te corrigeren voor vertekening in niet-gerandomiseerde studies. Het sterke punt van deze methode is de potentie om voor *alle verstorende invloeden* te corrigeren, zowel geobserveerde als niet-geobserveerde. Het zwakke punt van deze methode ligt in de *belangrijkste voorwaarde* die er op neerkomt dat de instrumentele variabele geen directe invloed mag hebben op de uitkomstvariabele, en ook geen indirecte invloed via de relatie met andere variabelen behalve behandeling. Of aan deze voorwaarde is voldaan kan in de praktijk alleen theoretisch worden beoordeeld en kan niet empirisch getest worden.

Verder is ingegaan op de *correlatie* tussen de instrumentele variabele en behandeling (of blootstelling). Als deze correlatie erg klein is, zal de methode niet alleen tot een grote standaardfout van de effectschatting leiden, maar ook tot een aanzienlijke *onzuiverheid* van de schatter bij kleine steekproeven; ook bij grote steekproeven zal een dergelijke onzuiverheid optreden als niet volledig aan de belangrijkste voorwaarde is voldaan.

Daarnaast is een *bovengrens* aangetoond van de correlatie tussen de instrumentele variabele en de behandeling (of blootstelling). Deze bovengrens is echter geen praktische belemmering als er geringe of matige vertekening bestaat, omdat deze bovengrens behoorlijk hoog ligt. Als de mate van vertekening wel groot is, bijvoorbeeld in geval van *vertekening door indicatie*, dan zal de maximale correlatie tussen de potentiële instrumentele variabele en de behandeling erg laag zijn, hetgeen resulteert in een erg zwakke instrumentele variabele. Omdat er een wisselwerking is tussen het voldoen aan de voorwaarde en de sterkte van de instrumentele variabele,

zal het bestaan van sterke vertekening naast een sterk instrument zeer waarschijnlijk leiden tot schending van deze veronderstelling en een onzuivere schatting van het behandelingseffect geven.

We concluderen dat de methode van instrumentele variabelen bruikbaar kan zijn *bij matige vertekening*, maar dat deze methode minder bruikbaar is bij sterke vertekening (bijv. door indicatie) omdat sterke instrumenten niet gevonden kunnen worden en voorwaarden gemakkelijk worden geschonden.

TOEPASSING VAN CORRECTIEMETHODEN

Propensity score methoden en in mindere mate instrumentele variabele methoden, worden in toenemende mate toegepast in de medische literatuur. In Hoofdstuk 4 zijn beide methoden toegepast op *overlevingsdata*.

PROPENSITY SCORES EN OVERLEVINGSDATA

In Paragraaf 4.1 is de methode van propensity scores toegepast op een studie naar het effect van het behandelen van hypertensie voor het voorkomen van een hersenberoerte. Wat betreft de mogelijkheid om voor vertekening te corrigeren, zijn drie propensity score methoden met elkaar vergeleken en vergeleken met de traditionele Cox proportional hazards regressie. Het aantal personen met een hersenberoerte in onze dataset was gering hetgeen een voordeel is voor propensity score methoden omdat dan meer covariaten kunnen worden opgenomen in vergelijking met standaard regressiemethoden.

Uit de literatuur is bekend dat bij de meeste propensity score toepassingen de propensity score als covariaat wordt gebruikt (is niet aanbevolen), dat de selectie van het propensity score model *routinematig* wordt gedaan door alle covariaten in het model te stoppen en dat aandacht voor en het *rapporteren van de bereikte balans* over het algemeen onvoldoende is. We hebben enkele methoden laten zien om inzicht te krijgen in de balans en om deze te rapporteren. Het blijkt dat er geen uniforme manier is om dat te doen in de praktijk. Tot slot hebben we een variabele toegevoegd die niet met de behandeling samenhangt waaruit geconcludeerd kan worden dat het geschatte behandelingseffect kleiner is bij stratificatie op de propensity score dan bij Cox proportional hazards regressie. Dit bevestigt de resultaten in Paragraaf 3.2.

We kunnen concluderen dat propensity score methoden bruikbaar zijn in combinatie met overlevingsdata, maar dat er meer aandacht moet zijn voor het maken van het propensity score model en voor het checken en rapporteren van de balans.

INSTRUMENTELE VARIABELEN EN OVERLEVINGSDATA

Hoewel toepassingen van instrumentele variabelen langzaamaan wat vaker in de medische literatuur verschijnen als methoden om voor vertekening te corrigeren, wordt deze methode nauwelijks toegepast bij gecensureerde overlevingsdata. In Paragraaf 4.2 is deze methode

toegepast met als behandelingseffect het verschil in overlevingskansen. Hiervoor hebben we gegevens gebruikt van type-1 diabetespatiënten die zijn verzameld om het verschil in overleving te bepalen tussen patiënten die alleen een niertransplantatie hebben ondergaan en patiënten bij wie daarnaast ook een alvleesklier is getransplanteerd. Als deze twee groepen patiënten direct met elkaar zouden worden vergeleken (*as-treated* of *zoals-behandeld*), zal het werkelijke verschil tussen beide behandelingen verkeerd worden geschat vanwege vertekening, ofwel omdat bij het bepalen van het type transplantatie selectieprocessen een rol spelen. Als daarentegen de oorspronkelijk gerapporteerde methode gebruikt zou worden om het behandelingseffect te schatten (*intention-to-treat* of *zoals-bedoeld*), dan worden in feite niet de behandelingen zelf vergeleken maar het beleid van ziekenhuizen om in principe de ene dan wel de andere methode te kiezen. Er is dan geen vertekening, maar het werkelijke verschil tussen de typen operaties wordt onderschat. Een derde mogelijkheid is de methode van instrumentele variabelen. Hiermee wordt het werkelijke effect van het tegelijkertijd transplanteren van de alvleesklier geschat, waarbij gecorrigeerd wordt voor vertekening. Het blijkt dan dat het effect veel groter is dan bij de methode van intention-to-treat, zij het dat de onbetrouwbaarheid van de schatting ook groter is.

Een nadeel van het gebruik van deze methode is de gevoeligheid van de effectschatting voor de onderliggende veronderstellingen. Een ander nadeel is dat bij weinig waarnemingen of bij een gering aantal gebeurtenissen de overlevingscurve die via de methode van instrumentele variabelen is geschat *onbetrouwbaarder* is dan bij andere schattingsmethoden. Dit geldt dan voornamelijk aan het einde van de curve, als er nog maar weinig personen tot de risicoset behoren.

We concluderen dat de methode van instrumentele variabelen ook gebruikt kan worden in geval van overlevingsdata. Aandacht voor de voorwaarden die ten grondslag liggen aan deze methode, is van groot belang. Standaardfouten van de schattingen moeten worden vermeld en kunnen erg groot zijn aan het einde van de overlevingscurve, zeker als het aantal gebeurtenissen gering is.

VERBETERING VAN PROPENSITY SCORE METHODEN

Zoals uit de literatuur en uit de toepassing in Paragraaf 4.1 is gebleken, is er behoefte aan informatie over hoe de methode van propensity scores precies moet worden toegepast, met name als het gaat om het maken van het propensity score model en het checken van de balans. In Hoofdstuk 5 zijn we nader ingegaan op balansmaten die kunnen helpen om uniformiteit te creëren bij het rapporteren van de balans die is bereikt en die kunnen helpen bij het selecteren van de variabelen voor het propensity score model om zo groot mogelijke gelijkheid op covariaten tussen behandelingsgroepen te verkrijgen.

DE OVERLAPPINGSCOËFFICIËNT

In Paragraaf 5.1 is voorgesteld om de *overlappingscoëfficiënt* te gebruiken bij propensity score toepassingen. Dit is een bestaande maat om de overlap tussen twee verdelingen te schatten. Als de propensity score in strata is verdeeld, kan met deze maat aangegeven worden hoe binnen deze strata de overlap is op de covariaten tussen beide behandelingsgroepen. Verder kan deze maat vergeleken worden met een waarde die men zou kunnen *verwachten* in geval van perfecte overlap in de populatie om zo de kwaliteit van de correctie voor vertekening te beoordelen. Deze informatie is van cruciaal belang om het gerapporteerde behandelingseffect te kunnen interpreteren. Vaak genoeg wordt alleen vermeld *dat er gecorrigeerd is* voor een set covariaten zonder daarbij te vermelden in welke mate dat ook daardwerkelijk *gelukt* is. Deze overlappingscoëfficiënt kan ook gebruikt worden om uit de vele mogelijke modellen een model te selecteren met een acceptabele balans; met een simulatiestudie hebben we namelijk aangetoond dat er een relatie is tussen deze maat en de zuiverheid van de effectschatter. De gewogen gemiddelde overlappingscoëfficiënt die berekend is op een set van beschikbare covariaten heeft de sterkste samenhang met de zuiverheid van de schatter in grotere datasets. Voor kleinere datasets hebben andere maten zoals de p-waarden uit significantietesten of de c-statistic een betere voorspellende kracht dan de overlappingscoëfficiënt.

We concluderen dat het gebruik van de overlappingscoëfficiënt nuttig kan zijn bij het toepassen van propensity score methoden. Bij kleinere steekproeven lijkt de methode minder goed te werken omdat dan het schatten van de overlap moeilijker is.

MEER BALANSMATEN

Een prettige eigenschap van de overlappingscoëfficiënt is dat het een directe maat is voor de overlap tussen twee verdelingen en daarmeee de balans in een propensity score toepassing. In Paragraaf 5.2 hebben we ook naar andere maten gekeken om deze balans vast te stellen. In een simulatiestudie (die anders is qua opzet dan in Paragraaf 5.1) hebben we de overlappingscoëfficiënt vergeleken met de *Kolmogorov-Smirnov afstand* en de *Lévy metriek*. Bij alle drie was een relatie aantoonbaar met de zuiverheid van de schatter; deze was het sterkste voor de gewogen overlappingscoëfficiënt en was sterker voor grotere steekproeven. Voor de overlappingscoëfficiënt was de samenhang $R = -0.06$ bij 400 waarnemingen en $R = -0.63$ bij 2000 waarnemingen. Een propensity score model selecteren met behulp van de overlappingscoëfficiënt lijkt het meest effectief te zijn, omdat de *gemiddelde gekwadrateerde fout* voor deze methode over het algemeen het kleinst was (variërend van 0.031 tot 0.197). De verschillen met de twee andere balansmaten waren echter niet groot. Een standaard propensity score model dat alle covariaten bevat, had een veel grotere gemiddelde gekwadrateerde fout. Ook als het propensity score model gekozen wordt met alle werkelijke (en in de praktijk altijd onbekende) verstorende factoren, dan was de gemiddelde gekwadrateerde fout groter.

We concluderen dat het gebruik van een van de gepresenteerde maten nuttig is bij het toepassen van propensity score methoden. Voorkeur gaat uit naar de overlappingscoëfficiënt

vanwege de directe interpretatie en de geringere gemiddelde gekwadrateerde fout bij het schatten van het behandelingseffect. Ondanks dat het schatten lastiger is bij kleinere steekproeven, levert de modelkeuze via de overlappingscoëfficiënt betere resultaten op dan de in de praktijk vaak gebruikte keuze voor het model met alle covariaten.

TOT SLOT

Propensity score methoden hebben verschillende voordelen ten opzichte van traditionele regressiemethoden als het gaat om het corrigeren voor vertekening (*confounding*). Een belangrijk voordeel is dat propensity score methoden behandelingseffecten geven die over het algemeen dichter bij de werkelijke gemiddelde behandelingseffecten liggen dan logistische of Cox proportional hazards regressie. Daarom zou deze methode vaker moeten worden toegepast als het gaat om correctie voor vertekening. Daarbij komt wel dat er meer aandacht voor het maken van het propensity score model moet zijn en voor het meten en rapporteren van de bereikte balans op covariaten, bijvoorbeeld door het gebruik van de overlappingscoëfficiënt.

De methode van instrumentele variabelen wordt veel minder toegepast in epidemiologisch onderzoek. Het zou wel meer aandacht verdienen omdat het een goede mogelijkheid is om volledig voor vertekening te corrigeren. Tegelijkertijd zullen de beperkingen van deze methode zeer algemeen gebruik ervan in de weg staan. Goede voorbeelden van toepassingen in de medische literatuur zullen ertoe bijdragen dat de methode op juiste waarde wordt geschat.

DANKWOORD

Eind 2002 werd het Centrum voor Biostatistiek door de afdeling Farmaco-epidemiologie en Farmacotherapie benaderd met de vraag of een van onze statistici als copromotor kon optreden bij het onderzoeksproject “Ontwikkeling en validatie van methoden om de effectiviteit van farmacotherapie in observationele studies vast te stellen”. Nadat een copromotor snel was gevonden in de persoon van Wiebe Pestman en duidelijk was dat het zoeken naar een promovendus nog niet was afgerond, bracht Ingeborg van der Tweel mij op het idee hierop te solliciteren. Ingeborg, bedankt hiervoor, want zelf was ik daar niet op gekomen. Ook wil ik je bedanken voor je stimulerende woorden gedurende mijn promotie-onderzoek.

Door de halftijds detachering gedurende een periode van bijna vierenhalf jaar kon ik niet al mijn onderwijs en de statistische consultatie blijven uitvoeren. Voor een deel kwamen deze werkzaamheden terecht bij mijn collega’s van het Centrum voor Biostatistiek, hetgeen soms een behoorlijke belasting bleek te zijn. Cas, Henk, Ingeborg, Jan, Maria, Paul, Rebecca, Wiebe en Wim, bedankt hiervoor.

Gedurende deze onderzoeksperiode zijn verschillende artikelen totstandgekomen onder begeleiding van mijn promotor Ton de Boer, copromotoren Olaf Klungel en Wiebe Pestman, en statistisch adviseur Svetlana Belitser. Sommige studies verliepen vlot, maar er waren ook momenten dat ik op onze regelmatige bijeenkomsten weinig te melden had. Al met al zijn deze bijeenkomsten zeer nuttig geweest en hebben ertoe bijgedragen dat we, ondanks onze verschillende achtergronden, steeds beter over de onderwerpen van dit proefschrift konden communiceren. Ton, naast het meedenken en het becommentariëren van de teksten, bedankt voor het stellen van de juiste vragen die me steeds weer dwongen op het ‘rechte’ pad te blijven. Olaf, bedankt voor het formuleren van het project, het meedenken en het kritisch lezen en becommentariëren van de teksten. Wiebe, bedankt voor je ideeën op het wiskundige vlak, het meedenken bij de uitgevoerde simulatiestudies en het schrijven van wiskundige programma’s. Svetlana, graag wil ik je bedanken voor je betrokkkenheid bij dit project, ondanks dat je geen formele begeleider bent. Dit heeft zich met name geuit in een toenemend aantal prettige discussies over de artikelen in dit proefschrift en in het schrijven van statistische programma’s.

Verder zijn er nog de volgende personen die ik graag wil bedanken: Jaap Abbring, Rick Grobbee, Bert Leufkens, Bruce Psaty, Yves Smets, Bruno Stricker, Sean Sullivan en Rudi Westendorp voor de bijdragen aan de verschillende artikelen, en: Ineke Dinzey, Suzanne de Visser, Addy Veeninga, Bert Agterhuis en Yvonne Weijers voor overige bijdragen.

Tot slot wil ik eenieder bedanken die in de afgelopen jaren zijn of haar interesse heeft getoond in de vorderingen in dit proefschrift: familie, vrienden, collega’s bij het Centrum voor Biostatistiek en collega’s bij de afdeling Farmaco-epidemiologie en Farmacotherapie, waaronder mijn ex-kamergenoten Ewoudt, Frank, Harald, Hedi, Ingeborg, Pearl, Tanja en Wim.

ABOUT THE AUTHOR

Edwin Martens was born on November 4, 1964 in Rotterdam where he finished in 1983 his secondary school. After graduating for his study Economy at the Erasmus University in Rotterdam in 1989, he worked until 2000 at the Institute for Sociologic-Economic Research (ISEO) at the Erasmus University Rotterdam as quantitatively oriented researcher and statistical advisor. From September 2000 he is working at the Centre for Biostatistics at Utrecht University, where teaching students statistical skills and advising researchers on statistical issues are his main tasks. In 2003 he was offered a half time PhD position at the Department of Pharmacoepidemiology and Pharmacotherapy at the Utrecht University.

OVER DE SCHRIJVER

Edwin Martens is geboren op 4 november 1964 in Rotterdam waar hij in 1983 zijn middelbare school afsloot. Na het behalen van zijn bull voor de studie economie aan de Erasmus Universiteit Rotterdam in 1989, werkte hij tot 2000 bij het Instituut voor Sociologisch-Economisch Onderzoek (ISEO) aan de Erasmus Universiteit in Rotterdam als kwantitatief onderzoeker en statistisch adviseur. Vanaf september 2000 is hij werkzaam bij het Centrum voor Biostatistiek aan de Universiteit Utrecht, waar hij zich vooral bezighoudt met statistiekonderwijs aan studenten en het adviseren van onderzoekers op statistisch gebied. In 2003 kreeg hij de mogelijkheid om halftijds te werken aan zijn promotie-onderzoek bij de afdeling Farmaco-epidemiologie en Farmacotherapie aan de Universiteit Utrecht.