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

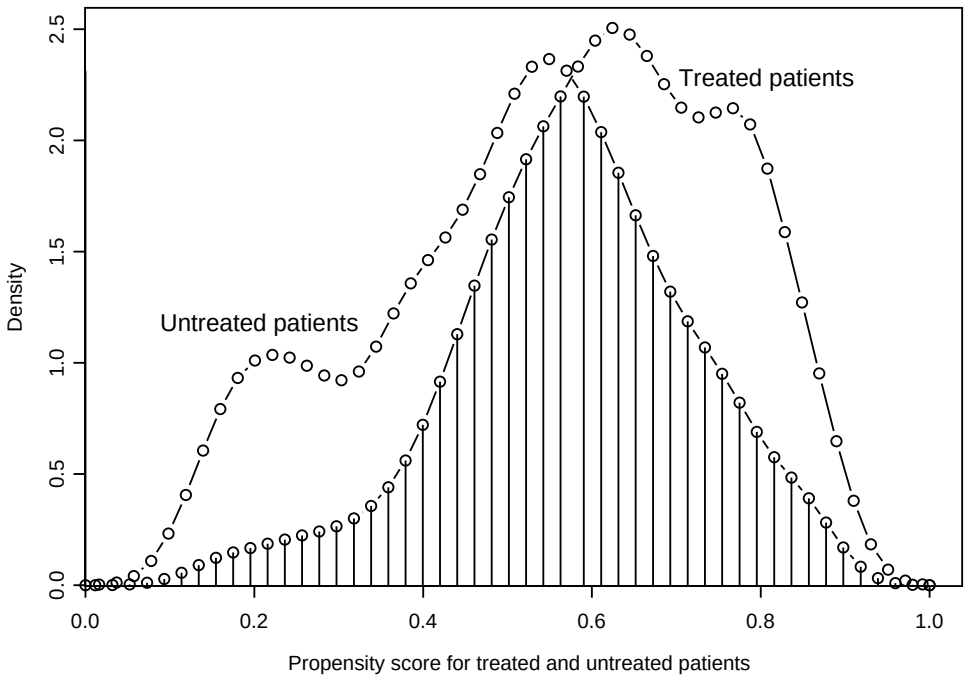
covariate	treated (n=1,293)	untreated (n=954)	odds ratio	95% confidence interval
History of CVA (%)	7.7	5.1	1.53	1.08, 2.18 *
Sex (% men)	34.3	47.5	0.58	0.49, 0.69 *
Smoking (%)	21.0	24.1	0.84	0.69, 1.02
Diabetes (%)	7.3	7.0	1.05	0.76, 1.45
Previous CVD (%)	24.1	17.1	1.54	1.24, 1.90 *
Previous MI (%)	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]

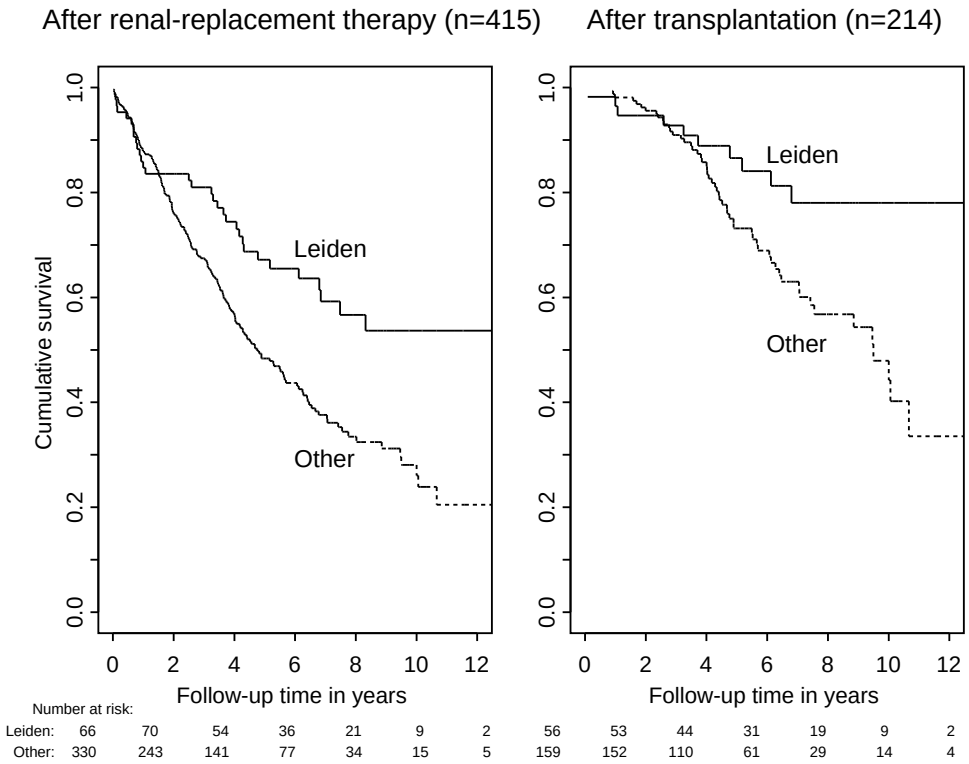
Table 4.8: Patient characteristics for Leiden and other areas

	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]

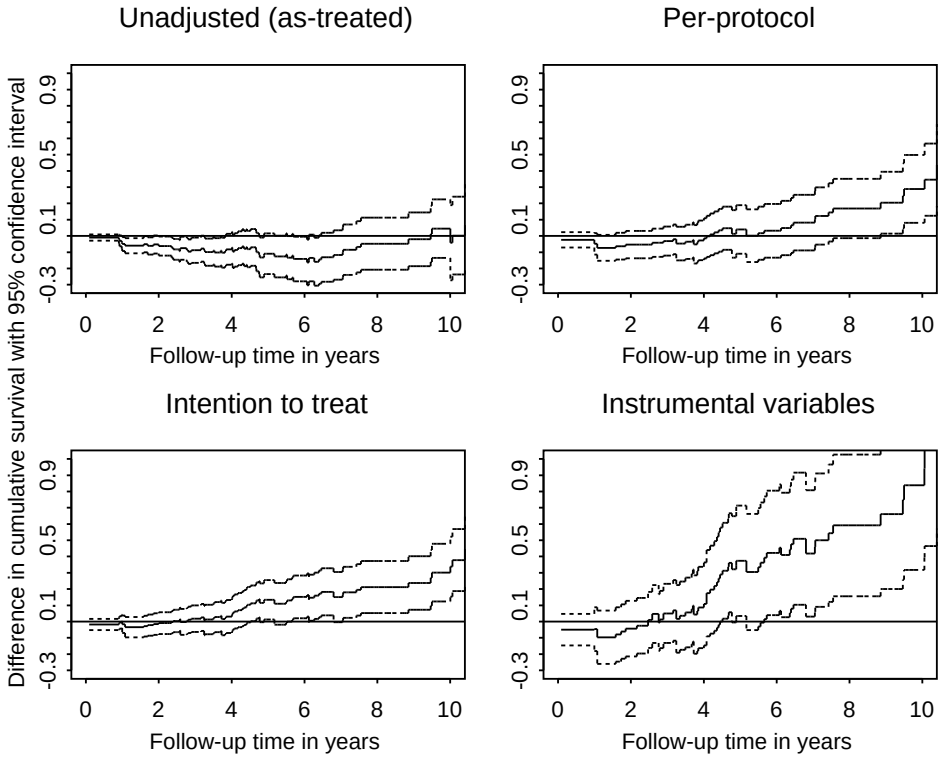
Figure 4.2: Patient survival after start of renal-replacement therapy and after transplantation, Leiden versus other areas



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

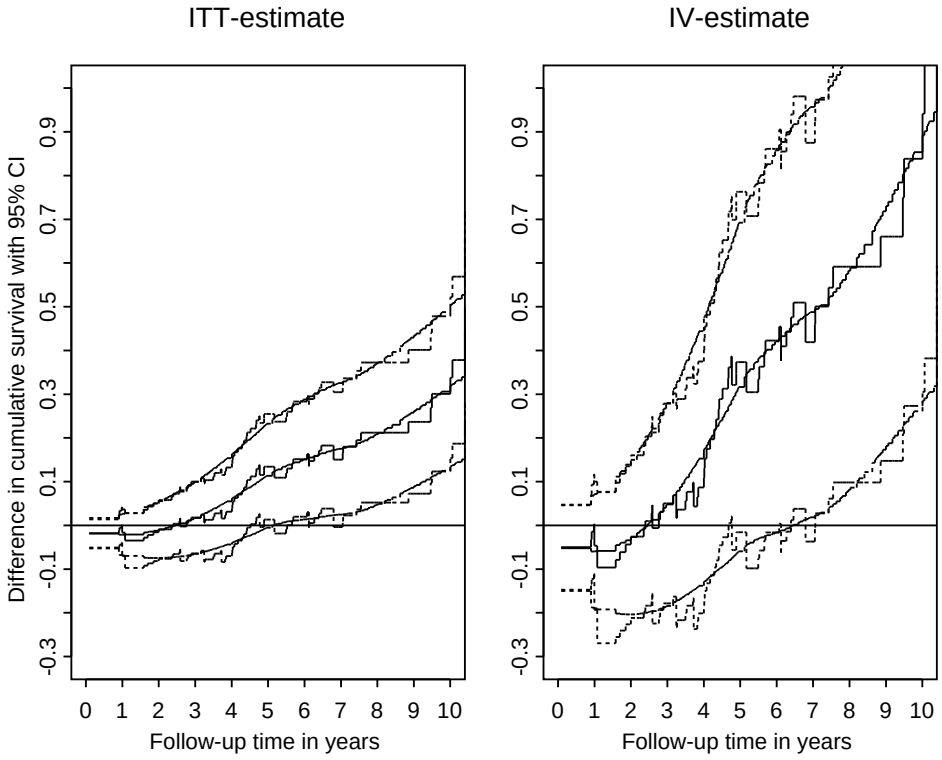


INSTRUMENTAL VARIABLES APPROACH

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.