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.