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*

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 multi-variate 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 ($Z_i = 0$):

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

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	● Only a few covariates or rough categories can be used
		● Clarity of incomparability on used covariates	
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	● Focus is not on balancing groups
		● Easy to perform	● 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.