CHAPTER 3

STRENGTHS AND LIMITATIONS OF ADJUSTMENT METHODS

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

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

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

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

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

ABSTRACT

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

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

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

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

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

DIFFERENT TREATMENT EFFECTS

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

TRUE CONDITIONAL TREATMENT EFFECT

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

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

MARGINAL OR CONDITIONAL TREATMENT EFFECTS?

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

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

REFERENCES

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

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

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

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

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

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

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

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

Provisionally accepted by International Journal of Epidemiology

ABSTRACT

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

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

INTRODUCTION

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

SYSTEMATIC DIFFERENCES BETWEEN TREATMENT EFFECT ESTIMATES

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

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

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

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

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

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

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

EXPLAINING DIFFERENCES IN TREATMENT EFFECT ESTIMATES

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

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

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

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

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

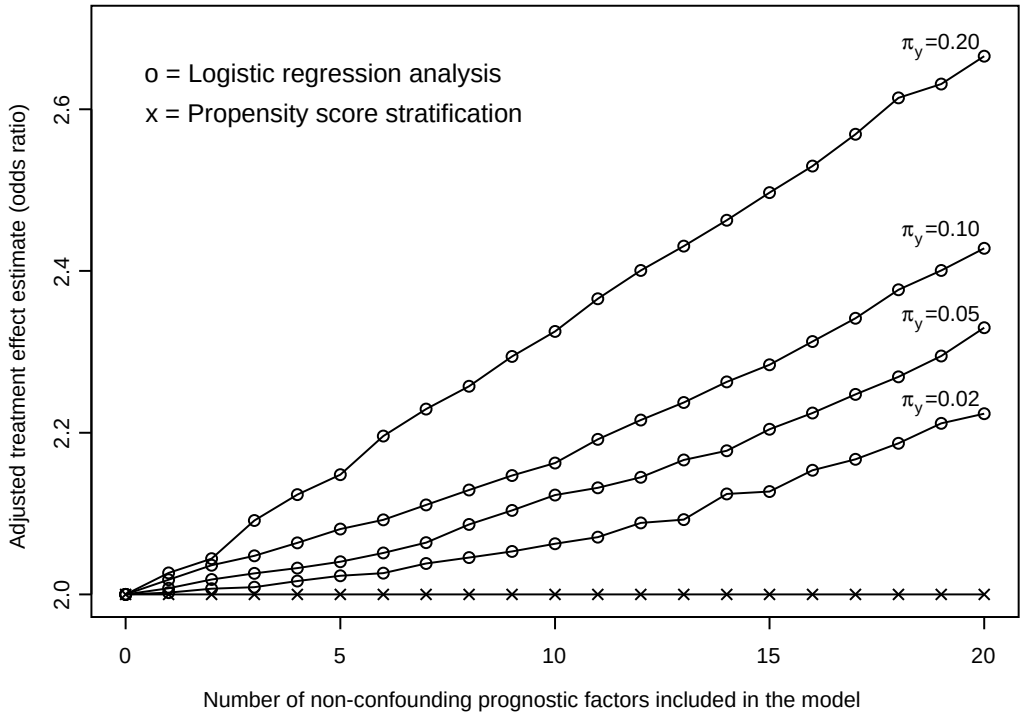
Suppose that in a certain situation only one prognostic factor exists (x_1) with a different distribution for both treatment groups. An adjusted treatment effect β_t^* will be interpreted as an estimate for the true average treatment effect, i.e. the effect that would be found when both treatment groups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

ADJUSTING FOR EQUALLY DISTRIBUTED PROGNOSTIC FACTORS

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

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

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



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

ADJUSTING FOR IMBALANCED PROGNOSTIC FACTORS

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

SIMULATIONS

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

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

and the propensity scores are calculated as:

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

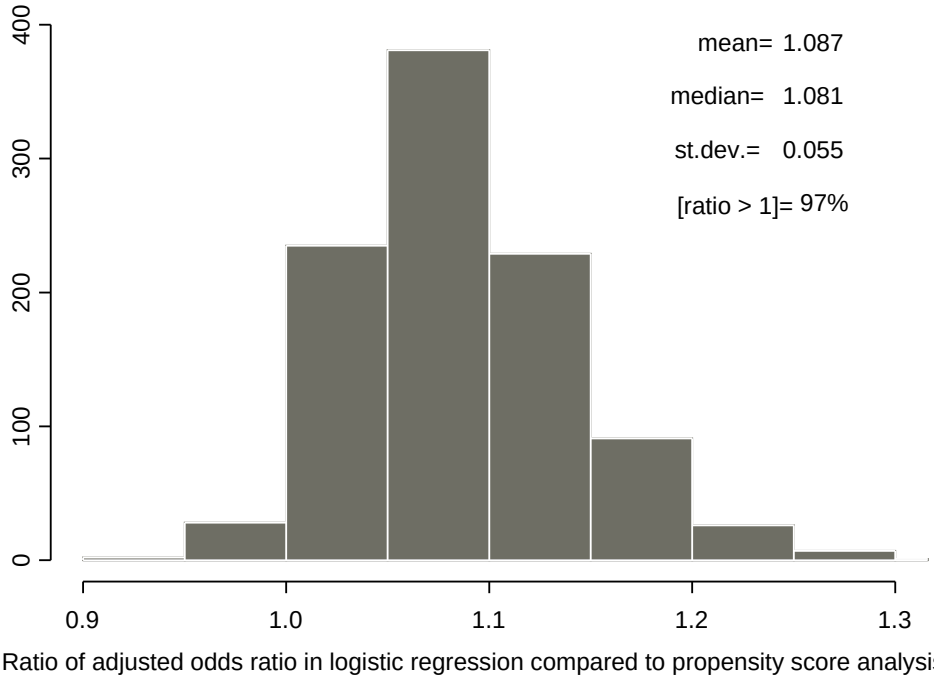
with $\text{logit}(t) = \alpha_t + \beta_{1t} x_1 + ... + \beta_{5t} x_5$.

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

COMPARISON OF ADJUSTED TREATMENT EFFECTS

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

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



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

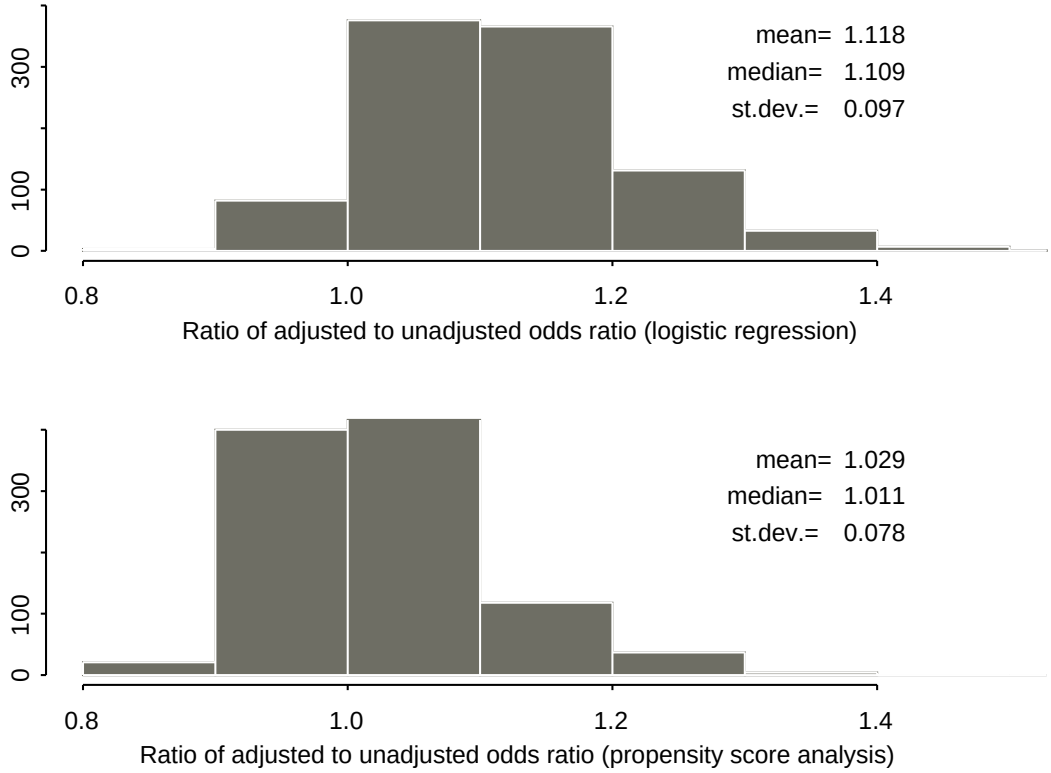
Table 3.2: Summary measures of the ratio of adjusted odds ratios of treatment effect in logistic regression compared to propensity score analysis in $1,000$ samples.

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

COMPARISON OF ADJUSTED AND UNADJUSTED TREATMENT EFFECTS

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

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



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

CONCLUSION AND DISCUSSION

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

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

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

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.3 INSTRUMENTAL VARIABLES: APPLICATION AND LIMITATIONS

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

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

Epidemiology 2006; 17:260-267

ABSTRACT

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

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

INTRODUCTION

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

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

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

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

A SIMPLE LINEAR IV MODEL

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

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

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

THE IV MODEL AND ITS ESTIMATOR

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

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

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

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

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

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

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

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

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

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

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

HOW TO OBTAIN A VALID INSTRUMENTAL VARIABLE

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

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

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

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

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

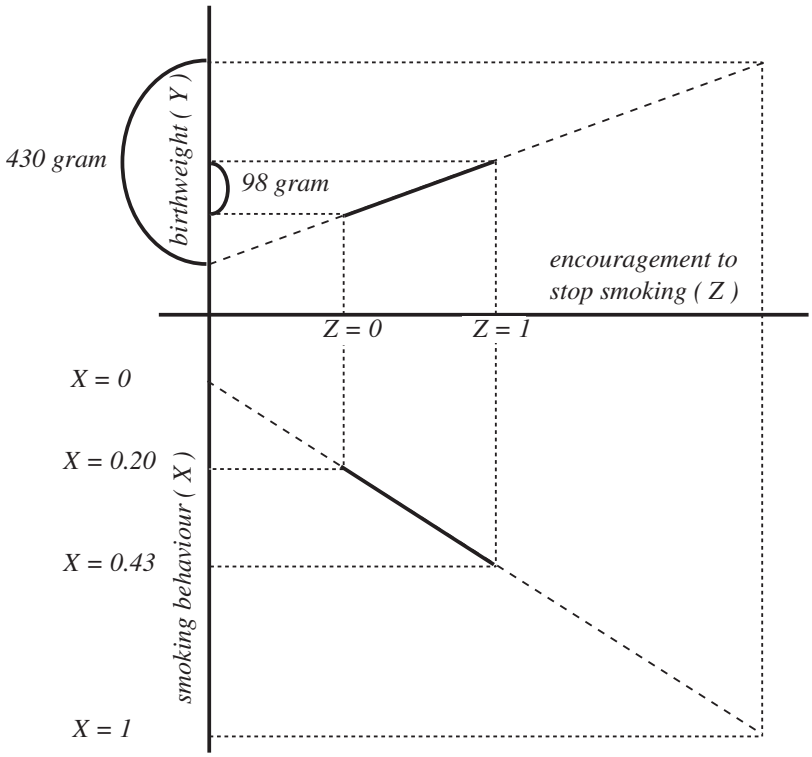
It should be noted that confounders of the X-Y relation are not explicitly mentioned in these assumptions, and that these confounders are part of both errors E and F. In the special case that $\rho_{E,F} = 1$, only confounders can be used to formulate the assumptions.[6]

NUMERICAL EXAMPLE OF IV APPLICATION

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

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

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



IMPLICATIONS OF WEAK INSTRUMENTS

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

LARGE STANDARD ERROR

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

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

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

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

BIAS WHEN SAMPLE SIZE IS SMALL

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

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

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

BIAS WHEN THE MAIN ASSUMPTION IS ONLY SLIGHTLY VIOLATED

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

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

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

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

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

A LIMIT ON THE STRENGTH OF INSTRUMENTS

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

WHEN THE MAIN ASSUMPTION HAS BEEN FULFILLED

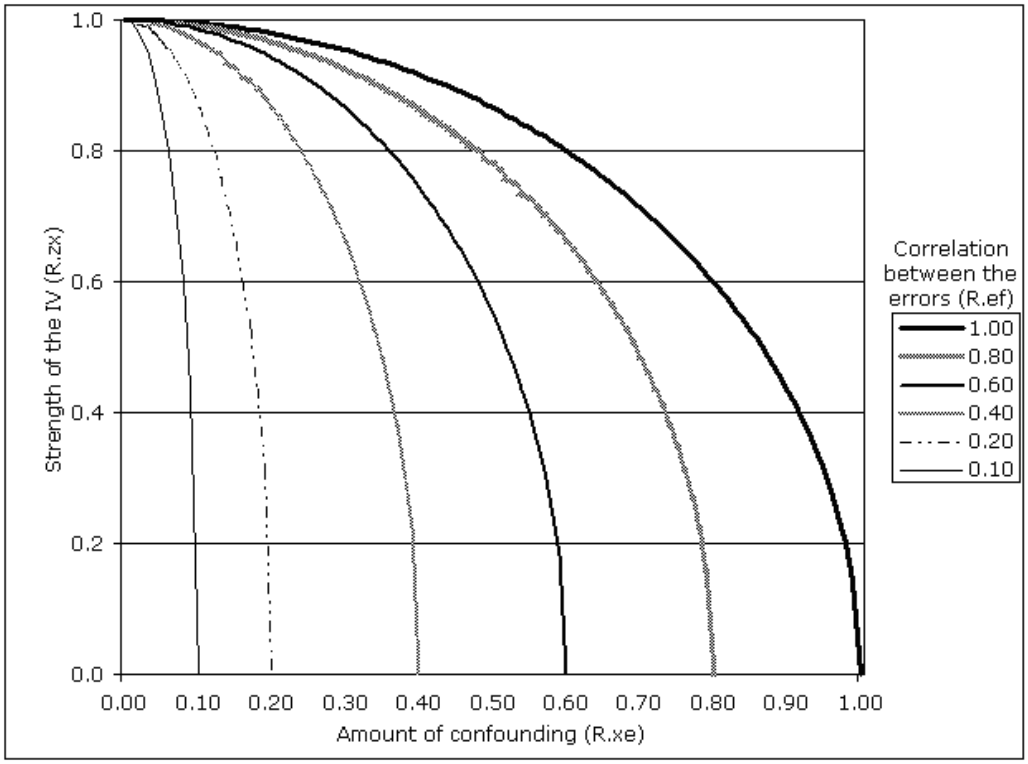
In case the main assumption of IV has been fulfilled, which means that the IV changes the outcome only through its relationship with the exposure, it can be shown that

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

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

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

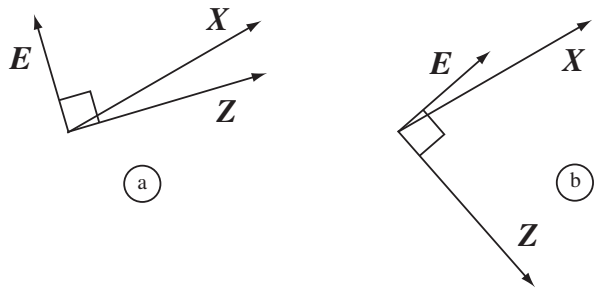
Figure 3.5: Relationship between strength of an instrumental variable ($\rho_{Z,X}$) and amount of confounding ($\rho_{X,E}$) for different error correlation levels ($\rho_{E,F}$), when main assumption has been fulfilled ($\rho_{Z,E} = 0$)



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

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

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



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

WHEN THE MAIN ASSUMPTION HAS NOT BEEN FULFILLED

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

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

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

THE EFFECT ON BIAS

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

DISCUSSION

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

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

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

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

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

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

APPENDIX A

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

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

Proof: According to the model one has

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

with

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

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

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

Squaring, rearranging terms and taking square roots will give

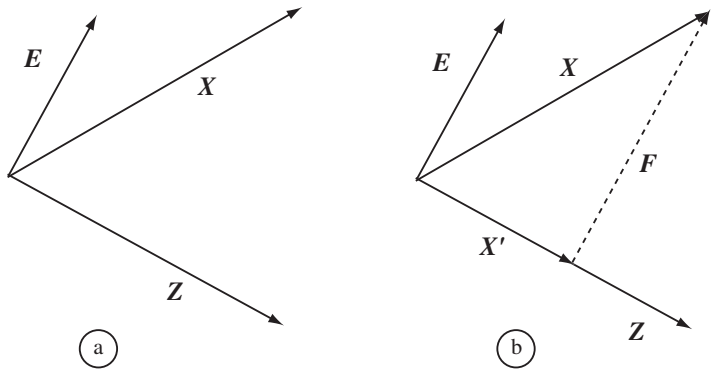
$$|\rho_{Z,X}| = \sqrt{1 - \frac{\rho^2_{X,E}}{\rho^2_{E,F}}}$$

which proves the theorem. □

APPENDIX B

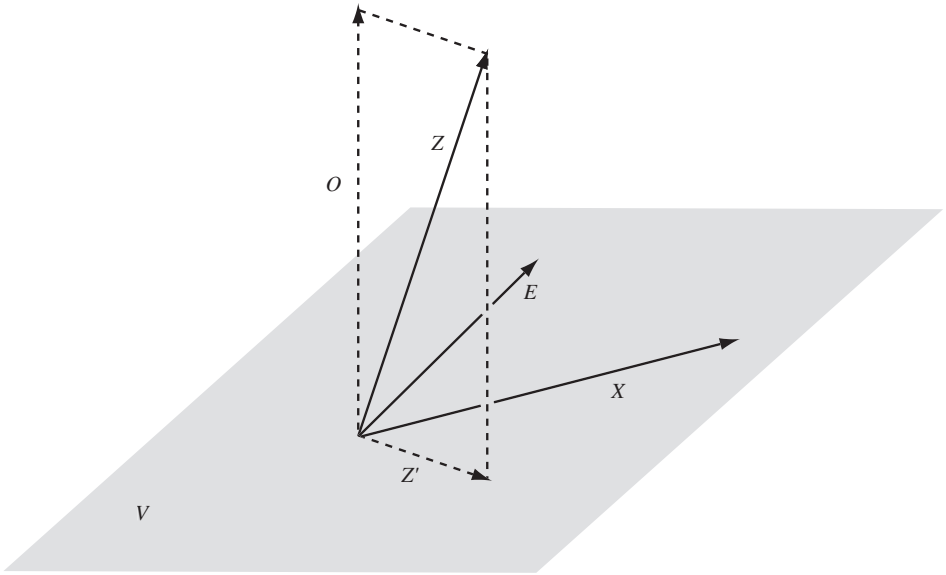
The condition $\rho_{E,F} = 1$ is equivalent to the condition that Z is in the same plane as X and E as can be seen in Figure 3.7. For simplicity we assume that the expectation values of the variables X, Y and Z are all equal to zero.

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



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

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



REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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