© International Epidemiological Association 2000 Printed in Great Britain *International Journal of Epidemiology* 2000;**29**:722–729

An introduction to instrumental variables for epidemiologists

Sander Greenland

Instrumental-variable (IV) methods were invented over 70 years ago, but remain uncommon in epidemiology. Over the past decade or so, non-parametric versions of IV methods have appeared that connect IV methods to causal and measurement-error models important in epidemiological applications. This paper provides an introduction to those developments, illustrated by an application of IV methods to non-parametric adjustment for non-compliance in randomized trials.

Keywords Biometry, causal models, compliance, confounding, econometrics, epidemiological methods, instrumental variables, measurement error, misclassification, regression, regression calibration, statistics

Accepted 17 December 1999

Methods for control of confounding and measurement error are central to non-experimental research. One large class of such methods based on instrumental variables (IV) dates back to the 1920s. They have been an integral part of econometrics for decades[1,2] and have appeared in the health sciences,[3–5] yet they remain little known in epidemiology. Their absence from the field may in part be due to the fact that the methods were rarely presented outside of linear-regression contexts until the 1980s. The past two decades have seen extensions of IV methods to non-parametric causal models and to non-linear regression.[4–16] I here provide an elementary introduction to non-parametric IV methods, with a focus on showing how IV assumptions lead to corrections for confounding by non-compliance in randomized trials. This application is especially important because treatment assignment can provide a perfect instrumental variable for confounding control, and IV methods provide an alternative to intent-to-treat analysis. I will also briefly sketch how IV methods for misclassification correction are related to confounding control.

An intuitive basis for the methods discussed here is as follows: Suppose X and Y are the exposure and outcome of interest, and we can observe their relation to a third variable Z, called an *instrumental variable* or *instrument*, that is associated with X but not associated with Y except through its association with X. Then, under certain conditions, we can write the Z-Y association as a product of the Z-X and X-Y associations,

$$\text{Assoc}_{ZY} = \text{Assoc}_{ZX}\text{Assoc}_{XY}$$

and solve this equation for the X-Y association. This equation is of particular use when either (i) the observed X-Y association is confounded by unmeasured covariates, but the Z-X and Z-Y associations are not confounded; or (ii) the X-Y association cannot be observed directly because we cannot observe X directly, but Z is an observed surrogate for X whose association with X is known or estimable, and whose deviation from X is independent of other variables or errors. The precise conditions under which the equation holds vary with the problem, as will be discussed below.

Department of Epidemiology, UCLA School of Public Health, Los Angeles, CA 90095-1772, USA.

Instrumental variables for confounding control

Let U be the set of all variables that affect X and Y, and suppose Z has the following properties:

1) Z is independent of U;
2) Z is associated with X;
3) Z is independent of Y given X and U.

Note that assumption 3 implies that Z has no direct effect on Y. Figure 1 gives a causal diagram[17] that satisfies these assumptions, with labels from the example below.

The variables in U may be partly or entirely unmeasured or even unimagined. It might then appear that there is no way to estimate the effect of X on Y in an unconfounded manner. A fundamental insight of IV estimation is that the instrument Z provides a means to estimate bounds on the X effect;[6,7,10] with further assumptions, the upper and lower bounds may be narrowed or even equal, in which case IV methods provide a point estimate.[8] This estimate is perhaps most easily understood in the following special case, which is based on a now standard counterfactual (potential-outcomes) model for treatment effects in the presence of non-compliance.[8,18,19]

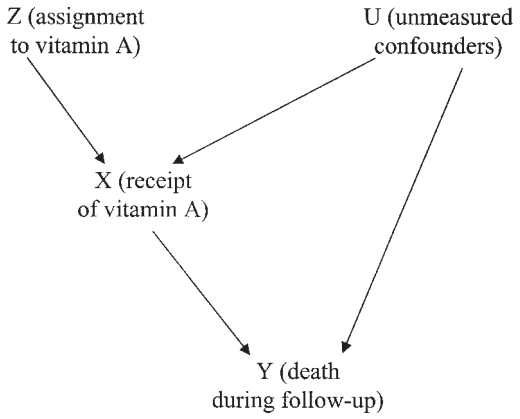

Figure 1

Intrumental variable methods for non-compliance

A paradigmatic example in which the IV conditions 1–3 are often satisfied is in a randomized trial with non-compliance: Z becomes treatment *assignment*, which is randomized and so fulfills assumption 1; X becomes treatment *received*, which is affected but not fully determined by assignment Z. To illustrate these concepts, Table 1 presents individual one-year mortality data from a cluster-randomized trial of vitamin A supplementation in childhood.[18,20] Of 450 villages, 229 were assigned to a treatment in which village children received two oral doses of vitamin A; children in the 221 control villages were assigned none. This protocol resulted in 12 094 children assigned to the treatment (Z = 1) and 11 588 assigned to the control (Z = 0). Only children assigned to treatment received the treatment; that is, no one had Z = 0 and X = 1. Unfortunately, 2419 (20%) of those assigned to the treatment did not receive the treatment (had Z = 1 and X = 0), resulting in only 9675 receiving treatment (X = 1). Nonetheless, assumption 1 is satisfied if the randomization was not subverted, while assumption 2 is supported by the data: Assignment to vitamin A increased the percentage receiving A from 0 to 80%.

Assumption 3 is plausible biologically, but must be reconciled with the fact that, among those both assigned to no vitamin A (Z = 0) and receiving no vitamin A (X = 0), mortality is only 639 per 100 000, versus 1406 per 100 000 for those assigned to vitamin A (Z = 1) but receiving no vitamin A (X = 0). Assuming that this difference is due to confounding by factors U that affect compliance (and hence X) and mortality (Figure 1), this illusory direct effect of assignment Z exemplifies the type of bias that arises when one attempts to estimate direct effects by stratifying on intermediates[21] (X is intermediate between Z and Y). There are many plausible explanations for such confounding. For example, perhaps families that fail to comply tend to be the poorest and so provide high-risk environments (poorer nutrition and sanitation); their low compliance would leave behind a low-risk group of compliers in the X = Z = 1 category, and thus confound an unadjusted comparison of the treated group (X = 1) with the untreated group (X = 0).

Confounding is a threat whenever people fail to comply with their assignment (i.e. have $X \neq Z$) for reasons (U) related to their outcome;[4–10,18,19,22,23] this problem is often referred to as one of biased selection for treatment. For example, patients assigned to a complex pill regimen may become lax in following that regimen. These non-compliers are often those who feel less ill and who have a better prognosis with respect to the outcome Y. In such situations, there will be confounding in a comparison of those complying with treatment to the other patients, because those complying are sicker than the others (i.e. there is self-selection for treatment that is related to prognosis).

Concerns of this sort have led to recommendations (often rigid) that *intent-to-treat* analysis be followed. To test and estimate effects, intent-to-treat compares those *assigned* to one treatment against those *assigned* to another treatment without regard to actual treatment received (X). Critics of this approach point out that treatment received is the source of biological efficacy, and that comparison of treatment assigned is biased for the effect of treatment received (furthermore, the bias is not always toward the null,[24] contrary to common lore). By recognizing treatment assignment as an instrument, IV methods provide an alternative to the biased extremes of analysing Z as the treatment (intent-to-treat) and analysing received treatment X in the conventional manner (which is likely to be confounded by determinants of compliance).

To see how the IV concept can be used to control for confounding due to non-compliance, let us refer to people who would always obey their treatment assignment as *co-operative*; among these people, X and Z are always equal. It is crucial to distinguish the concept of *co-operative* people from the concept of *compliance*. Co-operative people are those who will receive their assigned regimen, *no matter which regimen (treatment) they are assigned*. In the example, co-operative children have parents or guardians who will fully co-operate with the researchers, in that they will allow the researchers to give their child the vitamin if assigned to receive it, and will not give their child the vitamin if assigned to not receive it. Non-co-operative people are those who will not receive certain regimens if assigned to them. In the example, some parents may refuse to let their child receive the experimental treatment. These refusers have children who exhibit non-compliance if they are assigned to the vitamin A, but who exhibit compliance if they

Table 1 One-year mortality data from cluster-randomized trial of vitamin A supplementation in children.[20] Z = 1 if assigned A, 0 if not; X = 1 if received A, 0 if not

	Z = 1			**Z = 0**	
	X = 1	X = 0	Total	X = 1	X = 0
Deaths (Y = 1)	12	34	46	0	74
Total	9675	2419	12 094	0	11 588
Risk[a]	124	1406	380	undefined	639

[a] Deaths per 100 000 within one year.

are assigned to no vitamin. Thus, some non-co-operative people will be in compliance with their assignment, simply because they were not assigned to a treatment they would have refused.

Following earlier derivations of the correction below,[8,19] I will introduce this simplifying assumption:

4) Treatment assignment affects X only among co-operative people.

This assumption says that the participants may be divided into two groups: co-operative, for whom X always equals Z; and non-co-operative, for whom assignment Z has no effect on X. If the treatment variable has only two levels (1 and 0), assumption 4 reduces to an assumption that no trial participant would always (perversely) receive the opposite of what she was assigned, no matter which treatment she was assigned,[8] so that there are only two types of non-co-operators: Those who would always receive treatment and those who would never receive treatment. Table 1 provides evidence that assumption 4 is satisfied in the example: If there were participants who would always receive the opposite of their assignment ('defiers'[8]), we should expect to see some of them in the Z = 0, X = 1 column, but no one is seen there.

Now define

p_c = proportion of trial participants who are co-operative.

p_c is the effect that assignment to treatment 1 rather than 0 would have on the average value of the treatment indicator X. To see this, define

p_1 = proportion of participants who would receive treatment 1 (X = 1) if assigned treatment 1 (Z = 1)
= average of X if everyone were assigned treatment 1
p_0 = proportion of participants who would receive treatment 1 (X = 1) if assigned treatment 0 (Z = 0)
= average of X if no one were assigned treatment 1.

Any difference between p_1 and p_0 has to be due to the change in X among co-operative people (because, by assumption 4, only the co-operative people are affected by Z). More precisely, $p_1 - p_0$ must equal p_c because only co-operative people would go from X = 0 to X = 1 in response to a change from Z = 0 to Z = 1. Also, p_0 must equal the proportion of non-co-operators who always receive treatment, $1-p_1$ must equal the proportion of non-co-operators who never receive treatment, and $p_c + p_0 + 1 - p_1 = 1$.

Under assumption 1, $p_1 - p_0$ and hence p_c is validly estimated by the observed difference in the proportion receiving treatment 1 (X = 1) for the group assigned treatment 1 (Z = 1) versus the group assigned treatment 0 (Z = 0). In the example, this estimate is the difference in the proportions receiving vitamin A among those assigned and not assigned to A:

$$\hat{p}_c = \hat{p}_1 - \hat{p}_0 = \frac{9675}{12\,094} - \frac{0}{11\,588} = 0.80 - 0 = 0.80.$$

Next, define

$m_{\bullet 1}$ = average outcome ($\bar{Y}$) if everyone is assigned treatment 1 (Z = 1)

$m_{\bullet 0}$ = average outcome if everyone is assigned treatment 0 (Z = 0).

The average outcome difference $m_{\bullet 1} - m_{\bullet 0}$ is the effect that assignment to treatment 1 rather than 0 would have on the average outcome. (Recall that, when Y is a binary (0,1) disease indicator, the average outcome is the average risk or incidence proportion, and the average outcome difference is the average risk difference.[25]) Under assumption 1, this difference is validly estimated by the observed difference in average outcome for the group assigned treatment 1 versus the group assigned treatment 0 (which is the intent-to-treat [ITT] estimate of treatment effect):

$m_{\bullet 1}$ = 46/12 095 = 380 per 100 000
$m_{\bullet 0}$ = 74/11 588 = 639 per 100 000.

Next, consider two quantities that we do not observe directly:

m_{1c} = average outcome of co-operative people if everyone is assigned treatment 1,
m_{0c} = average outcome of co-operative people if everyone is assigned treatment 0.

$m_{1c} - m_{0c}$ is the effect that assignment to treatment 1 rather than 0 would have on the average outcome of *co-operative people*. In addition, because assignment Z and treatment received X are identical among co-operative people, $m_{1c} - m_{0c}$ is the effect that *receiving* treatment 1 rather than 0 would have on the average outcome among co-operative people. In other words, $m_{1c} - m_{0c}$ is both an intent-to-treat (Z) effect and a biological (X) effect on co-operative people.

Finally, consider two quantities that are estimable in a randomized trial:

m_{1n} = average outcome among non-co-operative people who receive treatment
m_{0n} = average outcome among non-co-operative people who do not receive treatment.

Under assumptions 3 and 4, these averages do not depend on assignment Z. Thus, m_{1n} can be estimated directly from the average outcome of those assigned to 0 but receiving 1 (Z = 0, X = 1); similarly, m_{0n} can be estimated directly from the average outcome of those assigned to 1 but receiving 0 (Z = 1, X = 0).

Table 2 summarizes the data expected from a study satisfying assumptions 1–4, using the above notation. It shows that we can write $m_{\bullet 1}$, the average outcome if everyone were assigned treatment 1, as a weighted average of the average outcomes for co-operative people and non-co-operative people, with weights equal to the respective proportions of each type of person:

$$m_{\bullet 1} = p_c m_{1c} + p_0 m_{1n} - (1 - p_1) m_{0n} \quad (1)$$

Similarly,

$$m_{\bullet 0} = p_0 m_{1n} + (1 - p_1) m_{0n} + p_c m_{0c} \quad (2)$$

Subtracting the second equation from the first yields

$$m_{\bullet 1} - m_{\bullet 0} = p_c(m_{1c} - m_{0c}) = (p_1 - p_0)(m_{1c} - m_{0c}) \quad (3)$$

Table 2 Expected data under assumptions 1–4 and text notation for a binary treatment variable X, with everyone assigned (Z = 1) or not assigned (Z = 0) treatment. C = 1 for co-operators, N = 1 for non-co-operators who always have X = 1, N = 0 for non-co-operators who always have X = 0

	If Z = 1:			**If Z = 0:**		
	X = 1		X = 0	X = 1	X = 0	
	C = 1	N = 1	N = 0	N = 1	N = 0	C = 1
Average Y	m_{1c}	m_{1n}	m_{0n}	m_{1n}	m_{0n}	m_{0c}
Proportion	p_c	p_0	$1 - p_1$	p_0	$1 - p_1$	p_c
Overall average Y:	$p_c m_{1c} + p_0 m_{1n} + (1 - p_1) m_{0n} = m_{\bullet 1}$			$p_0 m_{1n} + (1 - p_1) m_{0n} + p_c m_{0c} = m_{\bullet 0}$		

Thus, the effect $m_{\bullet 1} - m_{\bullet 0}$ of assignment on the overall average outcome can be viewed as the effect $p_1 - p_0$ of assignment Z on treatment received X times the effect $m_{1c} - m_{0c}$ of treatment received X on the outcome Y among co-operative people. Equation 3 exhibits the dilution of the Z effect produced by non-compliance; p_c quantifies this dilution on a scale from 0 (no Z effect if no compliance) to 1 (Z effect equals X effect if full compliance).

We can solve equation 3 for the effect of treatment X among co-operative people to get

$$m_{1c} - m_{0c} = (m_{\bullet 1} - m_{\bullet 0})/p_c = (m_{\bullet 1} - m_{\bullet 0})/(p_1 - p_0) \quad (4)$$

This equation shows that, subject to the assumptions, we can estimate the effect of treatment received X on the average outcome among co-operative people as the estimated effect of treatment assignment Z on the overall average outcome divided by the estimated effect of treatment assignment on the received treatment X. The Appendix gives a formula for the variance of this estimate. Because $m_{\bullet 1} - m_{\bullet 0}$ and $p_1 - p_0$ equal the Z-coefficients from the linear regressions of Y on Z and X on Z, the effect estimate from equation 4 equals the classical IV-regression estimate.[2]

Returning to the example, the unadjusted estimate of the risk difference produced by treatment (which is confounded by non-compliance) compares X = 1 to X = 0:

$$124 - 100\,000\,(34 + 74)/(2419 + 11\,588)$$
$$= -647 \text{ per } 100\,000 = -0.65\%$$

(95% CI : –0.81%, –0.49%). In contrast, the usual ITT estimate of the risk difference (which is a valid estimate of the risk difference caused by *assignment* to vitamin A) compares Z = 1 to Z = 0:

$$\hat{m}_{\bullet 1} - \hat{m}_{\bullet 0} = 380 - 639 = -259 \text{ per } 100\,000 = -0.26\%$$

(95% CI : –0.44%, –0.08%); this is a 259/639 = 41% risk reduction. The IV estimate is

$$(\hat{m}_{\bullet 1} - \hat{m}_{\bullet 0})/\hat{p}_c = -259/0.80 = -323 \text{ per } 100\,000 = -0.32\%$$

(95% CI : –0.55%, –0.10%), which is a 324/639 = 51% risk reduction. (Because the original data are unavailable, the confidence limits were computed assuming simple randomization, and so are incorrect to the extent that village effects are present.) It thus appears that the unadjusted estimate severely overestimates the treatment effect, but that the ITT estimate somewhat underestimates the effect.

Choice of target effect

In the above example the effect of vitamin A, $m_{1c} - m_{0c}$, is an effect restricted to co-operative people. This effect is useful if we think that co-operative people in the trial are typical (with respect to treatment effect) of people who will accept the treatment they are assigned. Aside from such generalizability issues (which arise in all trials), a conceptual problem is that we often cannot identify co-operative people with any certainty.[8,18] We can of course ask trial participants (or, above, their parents) assigned to and receiving a treatment if they would have obtained and taken that treatment had they not been assigned to it, but the reliability of the responses would ordinarily be unknown (even to the participant).

Under assumption 4 and the assumption that the treatment would never be received by those not assigned to it (which is plausible in the above example), we may identify as co-operative those people who receive treatment among those assigned to treatment. Otherwise, although we can estimate the *proportion* p_c of co-operative people (to whom our estimate of $m_{1c} - m_{0c}$ applies), we cannot characterize those people. The only behavioural effect we can always identify is the effect of assignment Z on treatment received X within our trial.

Sometimes, the effects of treatment under strictly enforced regimens are of central interest in planning mandatory programmes of otherwise unavailable prophylactics (such as vaccines). Define

m_{jf} = average outcome if everyone were forced to receive treatment j.

In the example, $m_{1f} - m_{0f}$ is the effect of forcing every child to take vitamin A versus withholding the supplement from every child. If co-operative and non-co-operative people differ in any way related to the effect of treatment received (i.e. if the variable co-operative/non-co-operative modifies the effect of X), the effect of X among co-operative people will not suffice to estimate the forcing effect. Unfortunately, we should expect the selective non-compliance that confounds the naïve X = 1 versus X = 0 comparison to also make $m_{1c} - m_{0c}$ unequal to $m_{1f} - m_{0f}$. Thus, one should never presume that the IV estimate of the effect is a valid estimate of the forcing effect, even when it is a valid estimate of the effect on co-operative people.

Estimation of ratio measures

A problem arises if one wishes to estimate measures that are ratios of average outcomes, reflective of more general non-collapsibility problems of ratio measures:[25,26] Because logs of

averages are not averages of logs, one cannot simply take Y as (say) the log rate or log odds and apply the above formulas to differences in these logs to obtain corrected log relative risk estimates; one needs additional strong assumptions about homogeneity of risks within levels of X and all controlled covariates to use the formulas. Such assumptions are implicit in IV methods based on regression modelling.[2,11–16]

Alternatively, under assumptions 1–3 and one more assumption one may derive risk ratio (RR) estimators for the effect of treatment among co-operative people, as well as alternative risk-difference (RD) estimators[18,23] (note, however, that the formulas in Sommer and Zeger[18] have serious misprints). For example, suppose that $p_0 = 0$, i.e. one cannot get treatment 1 (X = 1) without assignment to treatment 1 (Z = 1), so that the only non-co-operators are those who do not receive treatment when assigned to treatment (as in the example). Then, from assumptions 1 and 3,

The proportion receiving treatment among the proportion assigned to treatment, $\hat{p}_1$, is a valid estimate of p_c.

People assigned to and receiving treatment (Z = X = 1) are representative of all co-operative people in the trial, and so their average outcome provides a valid estimate of m_{1c}; call it $\hat{m}_{1c}$.

People assigned to but *not* receiving treatment (Z = 1, X = 0) are representative of non-co-operative people in the trial, and so their average outcome provides a valid estimate of m_{0n}; call it $\hat{m}_{0n}$.

As before, we can also estimate $m_{\bullet 0}$ by the average outcome among those assigned to 0 (Z = 0), $\hat{m}_{\bullet 0}$.

We can now solve equation 2 for m_{0c} to obtain

$$m_{0c} = [m_{\bullet 0} - (1 - p_c)m_{0n}]/p_c \qquad (5)$$

and substitute the above estimates into this equation to get $\hat{m}_{0c}$. The ratio and difference of $\hat{m}_{1c}$ and $\hat{m}_{0c}$ are now the estimated relative and absolute effects of received treatment on the average outcome. In the example, $\hat{p}_c = \hat{p}_1 = 0.80$, $\hat{m}_{1c}$ = 12/9675 = 124 per 100 000, $\hat{m}_{\bullet 0}$ = 639 per 100 000, $\hat{m}_{0n}$ = 34/2419 = 11 406 per 100 000, and so

$$\hat{m}_{0c} = [639 - (0.20)1406]/0.80 = 447 \text{ per } 100\,000$$

$$\hat{RR} = \hat{m}_{1c}/\hat{m}_{0c} = 124/447 = 0.28,$$

$$\hat{RD} = \hat{m}_{1c} - \hat{m}_{0c} = 124 - 447 = -323 \text{ per } 100\,000.$$

The latter estimate is the same as the estimate obtained from equation 4, although it was derived from the assumption that $p_0 = 0$, which is stronger than assumption 4 (from which equation 4 was derived).

Cuzick *et al.*[23] derived a more general risk-ratio estimator under an alternative assumption that the ratio is the same (homogeneous) for co-operators and for all types of non-compliers; when $\hat{p}_1 = 0$ (as in the example), their estimator reduces to that just given. As with the other estimators, their estimator also requires assumptions 1–3. Connor *et al.*[27] also derived formula 5 in the context in which treatment is screening and compliance is acceptance of screening, using assumptions equivalent to 1–3 and $p_0 = 0$.

Insufficiency of the instrumental assumptions

Assumptions 1–3, which here define Z as an instrument, are not sufficient to yield a point estimate of effects of the received treatment X. They do allow setting of bounds for X effects,[4,6,7,10] but these bounds can be uselessly wide.[9] Using only assumptions 1–3 in the above example, Balke and Pearl[10] derived non-parametric bounds for the forcing risk difference $m_{1f} - m_{0f}$ of –0.5% and 19%. The details of their derivation are beyond the present paper, but their upper bound suggests vitamin A might kill up to a fifth of the children, an absurd value in light of the extensive background information on the non-toxicity of vitamin A in the doses administered.[20] Thus, assumptions 1–3 plus the example data provide almost no additional information beyond what is already known to place an upper bound on the risk difference.

To obtain useful results, one will often need further plausible biological (causal) or parametric statistical assumptions beyond those embodied in assumptions 1–3. In the above example, assumption 4 is quite plausible. An even more plausible assumption is that vitamin A would not kill any of the children. With this 'no harm' assumption replacing assumption 4, the ITT estimate (-0.26%) is an upper bound for the forcing risk difference; this bound is obtained by assuming that no death in the group with Z = 1 and X = 0 (who are non-co-operative people) would have been prevented by forcing every child to take vitamin A. A companion lower bound of 12/12 094 – 74/11 584 = –0.54% is obtained by assuming that every death in the group with Z = 1 and X = 0 would have been prevented by forcing treatment on these children. These bounds are only for the *point estimate* of effect that would be obtained under a forced regimen; they do not account for random errors in the numbers of deaths. Nonetheless, the lower bound is already implausibly low on biological grounds, for we can be sure that treatment could not have prevented every death in the Z = 1, X = 0 group.

Instrumental variables for misclassification correction

One obtains a different perspective on IV methods by considering a surrogate or noisy measure Z for the exposure of interest X. The chief difference from the confounding problem is that no causal interpretation of the associations is required; the methods apply even for a purely descriptive (associational) analysis. Thus, for simplicity, in this section I will set aside U; I will also assume Y is measured without error. Assumption 3 then simplifies to

3′) Z is independent of Y given X

This assumption corresponds to the notion that error in Z as a measure of X is non-differential with respect to the outcome; this error may have systematic as well as random components, as long as neither component is associated with Y.

Given assumptions 2 and 3′, we can validly estimate the association of Y with X using IV formulas, provided we have validation data that show how Z predicts X in our study (i.e. that provide estimates of the predictive values for Z as a measure of X). To see this relation for a binary exposure X in a cohort study, for each exposure level j (j = 1 or 0) define

$m_{X=j}$ = average outcome among those with X = j

$m_{Z=j}$ = average outcome among those with Z = j

p_j = probability that X = 1 when Z = j for the entire cohort;

p_1 and $1 - p_0$ are then the positive and negative predictive values for the entire cohort (without regard to the outcome).

Under assumption 3′, Z has no effect on Y other than through Z; hence, the average outcomes within levels of X do not change across levels of Z, and in particular equal the $m_{X=j}$. Furthermore, the average outcome $m_{Z=j}$ within a given level j of Z is just the average of the average outcomes within levels of X, weighted by the probabilities of the X levels within the Z level; that is,

$$\begin{aligned} m_{Z=1} &= p_1 m_{X=1} + (1 - p_1) m_{X=0} \text{ and} \\ m_{Z=0} &= p_0 m_{X=1} + (1 - p_0) m_{X=0} \end{aligned} \quad (6)$$

Subtracting the second equation from the first yields

$$\begin{aligned} m_{Z=1} - m_{Z=0} &= (p_1 - p_0) m_{X=1} - (p_1 - p_0) m_{X=0} \\ &= (p_1 - p_0)\ (m_{X=1} - m_{X=0}) \end{aligned} \quad (7)$$

We now solve this equation to get

$$m_{X=1} - m_{X=0} = (m_{Z=1} - m_{Z=0})/(p_1 - p_0). \quad (8)$$

This equation shows that, subject to the assumptions, we can estimate the exposure-specific outcome difference as the ratio of the surrogate-specific outcome difference (the Z-Y association) and the difference of the surrogate-specific probabilities of X = 1 (the Z-X association).

Use of validation data

Equation 8 does not require any data that directly relate the true exposure X to the outcome Y: We can correct the estimate from a study relating the surrogate Z to the outcome provided we can construct accurate estimates of p_1 and p_0 from a validation study relating the true exposure to the surrogate. If, however, the validation data also show the relation of the true exposure to the outcome (as would be the case if they were a random sample from the study relating surrogate to outcome), the IV-corrected estimate from applying equation 8 to the unvalidated data can and should be combined with the direct estimate from the validation data. Such data combination is done most efficiently with regression modelling.[15,28] Validation data that are not restricted on the outcome (e.g. that include both cases and non-cases when Y is a disease indicator) also allow one to test assumption 3 and to use methods that do not require that assumption.[28–30]

Relation to confounding control

The resemblance of equation 8 to 4 reflects an underlying common feature of the two situations. In both, we lack complete data on the relation of X to Y. In the non-compliance problem, we lack data on who is co-operative; the IV method uses the X-Z data to correct for the dilution of the Z-Y association as a measure of the X effect that results from inclusion of non-co-operative people in the intent-to-treat comparison. Similarly, in the classification problem we lack data on who is correctly classified; the IV method uses X-Z data to correct for the dilution resulting from the inclusion of the misclassified data in our comparison.

There is a distinction, however: For misclassification, assumption 3′ implies that the X-Y *association* is the same among the correctly and incorrectly classified; hence, the corrected estimate applies to the entire study cohort. In contrast, for confounding control, the assumptions do not imply that the *effect* of X on Y would be the same for co-operative and non-co-operative people; hence, the corrected estimate applies only to co-operative people (who may be difficult to identify). This distinction does not appear in classical IV methods, as these methods are based on models in which the X effect on Y is homogeneous across covariates U. (Note that, by assumption 3, Z has no effect and so the X effect *must* be homogeneous within levels of Z.)

Discussion

The above corrections extend directly to situations requiring adjustment for measured covariates by applying them within covariate strata and summarizing; nonetheless, a more efficient approach is supplied by IV methods based on regression models.[15] Those methods are often presented under the heading of regression-calibration or linear-imputation methods[13,29] but are special cases of general IV regression formulas. Bashir and Duffy[29] provide an elementary introduction to linear imputation and other measurement-error corrections, while Carroll *et al.*[15,28] provide advanced and thorough coverage of model-based corrections, including general IV corrections; the latter allow both Y and X to be measured with error in both the main and validation sample, as long as that error is uncorrelated with Z.

An important limitation of corrections based on regression models is their model dependence, especially on the models for error distributions. For example, *P*-values and confidence limits from the basic regression-calibration (linear-imputation) form of IV correction assumes that the regression of the true (biologically relevant) exposure X on the instrument Z follows a linear model with normal errors.[13,29] This model is highly implausible in many situations and is mechanically impossible to satisfy if X and Z are discrete. This limitation is not shared by non-parametric methods[4–10] and special methods for categorical variables.[30–32]

In purely observational studies (in which neither the instrument Z nor the treatment X has been subject to experimental manipulation), a major limitation of all IV methods is their strong dependence on assumptions 1 and 3. The corrections may even be harmful if Z is associated with other errors or with unmeasured confounders, as might be expected if (say) Z is self-reported alcohol consumption, X is true consumption, Y is cognitive function, and U is use of illegal drugs. Thus, IV corrections are no cure for differential errors; they address only independent non-differential errors, as embodied by assumptions 1 and 3. If there are violations of the assumptions, bias due to measurement error (using Z as a surrogate for X) will no longer act multiplicatively (as in formula 7) and adjustment will require more complex formulas.[15,28]

The sensitivity of IV corrections to the assumptions increases with the amount of non-compliance or the amount of error in Z as a measure of X; the corrections will be especially unreliable if Z is a very noisy measure of X, so that the association of Z and X is weak.[33] The key to successful IV correction is thus to find

an instrument Z that is strongly associated with the exposure X but not otherwise associated with uncontrolled factors affecting Y or with other sources of error. Such instruments may be difficult to find when only self reports of sensitive personal characteristics are available, but can sometimes be found or created from records or physical measurements.[13,15]

Finally, one should be aware that the IV corrections (like most epidemiological statistics) are large-sample procedures, which means they include some bias due to sample-size limitations, even if all the assumptions are met. For linear-regression IV corrections this bias is in the direction of undercorrection,[33] and so should be of less concern than the effects of biased measurement. Furthermore, this bias decreases with increasing association between Z and X,[33] which provides further impetus to find instruments highly associated with X.

Despite the aforementioned cautions, IV corrections can be valuable in many situations. When the IV assumptions are questionable, the corrections can still serve as part of sensitivity analysis or external adjustment.[34] When the assumptions are more defensible, as in field trials and in studies that obtain validation or reliability data, IV methods can form an integral part of the analysis.

Acknowledgements

The author thanks Wendy McKelvey, Hal Morgenstern, and the referees for helpful comments.

References

[1] Wright S. Appendix. In: Wright PG (ed.). *The Tariff on Animal and Vegetable Oils*. New York: Macmillan, 1928.

[2] Goldberger AS. Structural equations methods in the social sciences. *Econometrica* 1972;**40:**979–1001.

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

[4] Robins JM. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In: Sechrest L, Freeman H, Mulley A (eds). *Health Service Research Methodology: A Focus on AIDS*. Washington, DC: US Public Health Service, 1989, pp.113–59.

[5] Sheiner LB, Rubin DB. Intention-to-treat analysis and the goals of clinical trials. *Clin Pharm Therap* 1995;**57:**6–15.

[6] Manski CF. Nonparametric bounds on treatment effects. *Am Econ Rev, Papers Proc* 1990;**80:**319–23.

[7] Pearl J. On the testability of causal models with latent and instrumental variables. In: Besnard P, Hanks S (eds). *Uncertainty in Artificial Intelligence 11*. San Francisco: Morgan Kaufmann, 1994, pp.435–43.

[8] Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables (with discussion). *J Am Statist Assoc* 1996;**91:**444–72.

[9] Imbens GW, Rubin DB. Bayesian inference for causal effects in randomized experiments with noncompliance. *Ann Stat* 1997;**25:** 305–27.

[10] Balke A, Pearl J. Bounds on treatment effects from studies with imperfect compliance. *J Am Statist Assoc* 1997;**92:**1171–76.

[11] Hansen LP, Singleton RJ. Generalized instrumental variable estimation of non-linear rational expectations models. *Econometrica* 1982;**50:**1269–86.

[12] Amemiya Y. Two-stage instrumental variable estimators for the nonlinear errors in variables model. *J Econometrics* 1990;**44:**311–32.

[13] Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Stat Med* 1989;**8:**1051–69.

[14] Carroll RJ, Stefanski LA. Measurement error, instrumental variables and corrections for attenuation with applications to meta-analysis. *Stat Med* 1994;**13:**1265–82.

[15] Carroll RJ, Ruppert D, Stefanski LA. *Measurement Error in Nonlinear Models*. New York: Chapman and Hall, 1995.

[16] Buzas JS, Stefanski LA. Instrumental variable estimation in generalized linear measurement error models. *J Am Statist Assoc* 1996;**91:**999–1006.

[17] Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999;**10:**37–48.

[18] Sommer A, Zeger SL. On estimating efficacy from clinical trials. *Stat Med* 1991;**10:**45–52.

[19] Baker SG, Lindeman KS. The paired availability design: a proposal for evaluating epidural analgesia during labor. *Stat Med* 1994;**13:** 2269–78, correction **14:**1841.

[20] Sommer A, Tarwotjo I, Djunaedi E *et al.* Impact of vitamin A supplementation on childhood mortality: a randomized controlled community trial. *Lancet* 1986;**i:**1169–73.

[21] Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 1992;**3:**143–55.

[22] Baker SG. All-or-nothing compliance. In: Kotz S, Read C, Banks D (eds). *The Encyclopedia of Statistical Sciences*, **update 1:**134–38. New York: Wiley, 1997.

[23] Cuzick J, Edward R, Segnan N. Adjusting for non-compliance and contamination in randomized clinical trials. *Stat Med* 1997;**16:** 1017–29.

[24] Robins JM. Correction for non-compliance in equivalence trials. *Stat Med* 1998;**17:**269–302.

[25] Greenland S, Rothman KJ. Measures of effect and measures of association. In: Rothman KJ, Greenland S (eds). *Modern Epidemiology*. Philadelphia: Lippincott-Raven, 1998.

[26] Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. *Stat Sci* 1999;**14:**29–46.

[27] Connor RJ, Prorok PC, Weed DL. The case-control design and the assessment of the efficacy of cancer screening. *J Clin Epidemiol* 1991;**44:**1215–21.

[28] Carroll RJ, Gail MH, Lubin JH. Case-control studies with errors in predictors. *J Am Statist Assoc* 1993;**88:**177–91.

[29] Bashir SA, Duffy SW. The correction of risk estimates for measurement error. *Ann Epidemiol* 1997;**7:**154–64.

[30] Rao P. Double sampling. In: Armitage P, Colton T (eds). *The Encyclopedia of Biostatistics*. New York: Wiley, 1998, pp.1223–29.

[31] Reade-Christopher SJ, Kupper LL. Effects of exposure misclassification on regression analyses of epidemiologic follow-up study data. *Biometrics* 1991;**47:**535–48.

[32] Liu X, Liang KY. Adjustment for non-differential misclassification error in the generalized linear model. *Stat Med* 1991;**10:**1197–211.

[33] Bound J, Jaeger DA, Baker RM. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J Am Statist Assoc* 1995;**90:**443–50.

[34] Greenland S. Basic methods for sensitivity analysis. In: Rothman KJ, Greenland S (eds). *Modern Epidemiology. 2nd Edn.* Philadelphia: Lippincott-Raven, 1998.

[35] Mood AM, Graybill FA, Boes DC. *Introduction to the Theory of Statistics. 3rd Edn*. New York: McGraw-Hill, 1974.

Appendix

To derive the variance of the estimator based on equation 4, let $\hat{m}_{jk}$ denote the sample outcome mean at X = j, Z = k, $m_{jk} = E(\hat{m}_{jk})$, $m_{\bullet k} = m_{1k} + m_{0k}$, $V_{jk} = \text{var}(\hat{m}_{jk})$, $\hat{p}_k$ the sample proportion of X = 1 among Z = k, $\hat{q}_k = 1 - \hat{p}_k$, $p_k = E(\hat{p}_k)$, $p_c = p_1 - p_0$, $V_{pk} = \text{var}(\hat{p}_k)$. The exact formulas for V_{jk} and V_{pk} depend on the type of randomization used (e.g., simple, stratified, or cluster). The following derivation assumes that each randomized unit contributes only a small portion of each estimated quantity, so that approximately the $\hat{m}_{jk}$ are jointly independent, the $\hat{p}_k$ are jointly independent, and $\hat{m}_{jk}$ is independent of $\hat{p}_{l-k}$; it also uses the fact that $\text{var}(\hat{p}_k) = \text{var}(\hat{q}_k) = -\text{cov}(\hat{p}_k, \hat{q}_k)$. We then have

$$\begin{aligned} V_{\bullet k} &\equiv \text{var}(\hat{m}_{\bullet k}) = \text{var}(\hat{p}_k \hat{m}_{1k} + \hat{q}_k \hat{m}_{0k}) \\ &= \text{var}(\hat{p}_k \hat{m}_{1k}) + \text{var}(\hat{q}_k \hat{m}_{0k}) + 2\text{cov}(\hat{p}_k \hat{m}_{1k}, \hat{q}_k \hat{m}_{0k}) \\ &= V_{pk} V_{1k} + p_k^2 V_{1k} V_{pk} + m_{1k} V_{pk} + V_{pk} V_{0k} + q_k^2 V_{0k} \\ &\quad + m_{0k} V_{pk} - 2m_{1k} m_{0k} V_{pk} = V_{pk}\left[V_{1k} + V_{0k} + (m_{1k} - m_{0k})^2\right] \\ &\quad + p_k^2 V_{1k} + p_k^2 V_{0k}, \end{aligned} \tag{A1}$$

$$\text{var}(\hat{m}_{\bullet 1} - \hat{m}_{\bullet 0}) = V_{\bullet 1} + V_{\bullet 0}, \text{var}(\hat{p}_0) = V_{p1} + V_{p0}, \tag{A2}$$

$$\text{cov}(\hat{m}_{\bullet k}, \hat{p}_k) = \text{cov}(\hat{p}_k \hat{m}_{1k} + \hat{q}_k m_{0k}, \hat{p}_k) = (m_{1k} - m_{0k}) V_{pk} \tag{A3}$$

$$\text{cov}(\hat{m}_{\bullet 1} - \hat{m}_{\bullet 0}, \hat{p}_c) = (m_{11} - m_{01}) V_{p1} + (m_{10} - m_{00}) V_{p0} \tag{A4}$$

Let d be the IV parameter $(m_{\bullet 1} - m_{\bullet 0})/p_c$ and $\hat{d}$ the IV estimator $(\hat{m}_{\bullet 1} - \hat{m}_{\bullet 0})/\hat{p}_c$. Combining A1–A4 with the Taylor approximation for variances of ratios of random variables[35] yields

$$\begin{aligned} \text{var}(\hat{d}) = &(V_{\bullet 1} + V_{\bullet 0}) y p_c^2 + d^2 (V_{p1} + V_{p0}) y p_c^2 \\ &-2d\left[(m_{11} - m_{10}) V_{p1} + (m_{10} - m_{00}) V_{p0}\right] y p_c^2 \end{aligned} \tag{A5}$$

An estimate of $\text{var}(\hat{d})$ is obtained by substituting sample estimates for parameters in A5.