

Control of Confounding in the Assessment of Medical Technology

SANDER GREENLAND* and RAYMOND NEUTRA*

Greenland S (Division of Epidemiology, School of Public Health, University of California, Los Angeles, CA 90024, USA) and Neutra R. Control of confounding in the assessment of medical technology. *International Journal of Epidemiology* 1980, 9: 361–367.

Separation of the effects of extraneous variables from the effects of a factor under study (often termed control of confounding) is one of the key prerequisites for validly estimating the magnitude of the study factor's effects. Because of the phenomenon of confounding by indication, confounding of effects of different factors is a common problem in the assessment of medical technology. We give several examples illustrating that the decision of whether a recorded variable is a confounder in a data-set must be decided on the basis of subject-matter knowledge and clinical judgement. There is no alternative to use of such judgement; statistical selection procedures based on significance tests, such as stepwise regression, can be particularly misleading.

Because of the rarity of some of the benefits and side effects of a medical technology, sufficiently large randomised trials of the technology may be deemed infeasible, and thus nonexperimental assessment will be the only practical means of studying the procedure. In such cases the control of confounding becomes a natural consideration. In fact, confounding is usually guaranteed in non-experimental studies of medical technology. This is because the indications for treatment are almost always indicators of special risk, and therefore persons who are selected for special treatment will automatically be at different risks from those who receive the more standard treatment. For example, during the introduction of a new type of surgery it is the good risk patients who selectively receive it, while the poor risk patients receive the standard treatment; in application of a new surveillance technology, such as electronic foetal monitoring, the higher risk patients may be selectively monitored, and those not monitored would then represent a lower risk group. In both cases there will be a bias in any estimate of effect based on a simple comparison of the treated and untreated groups. This sort of bias has been identified as 'confounding by indication' (Miettinen, personal

communication). It is because of such bias that researchers in technology evaluation must understand the principles underlying the concept of confounding and its control in the analysis.

Definition and Example of Causal Confounding

Causal confounding may be defined quite generally as the confusion of the effects of an extraneous variable with the effects of a study factor (which may be a medical intervention) on the development of an outcome (which may be a benefit or side effect). If such confounding could occur in a particular study, the extraneous variable is called a potential confounder; if such confounding occurs in a particular study analysis, the extraneous variable is then termed a causal confounder in that study. It follows immediately from this definition that a potential or causal confounder must itself be a cause of the outcome under study.

Example 1: Table 1 presents the data from a study of the effect of foetal monitoring on risk of caesarean section.¹ It appears that monitoring multiplies the risk of caesarean by a factor of 1.9. In Table 2, the same data are examined for the relationship of arrested labour to monitoring. We see that many more arrested labours were present in the monitored than in the unmonitored group; as a consequence, a somewhat higher caesarean risk should have been expected among the monitored women even if monitoring had no effect, since arrested labour is an important contributing cause of caesarean section.

* Division of Epidemiology, School of Public Health, University of California, Los Angeles, CA 90024, USA

This research was supported by a grant from the Milbank Memorial Foundation.

TABLE 1 *Monitoring – caesarean study*

Mode of delivery	Monitored	Unmonitored
Caesarean	780	408
Vaginal	6 520	6 776
TOTAL	7 300	7 184
Caesarean risk:	107/1 000	57/1 000
Risk ratio = 1.9		

TABLE 2 *Monitoring vs. arrested labour*

Progress of Labour	Monitored	Unmonitored
Arrested	1 283	425
No Arrest	6 027	6 759
TOTAL	7 300	7 184
Percent Arrested:	17.6	5.9
Ratio = 3.0		

The above example illustrates the basic problem of confounding: how do we determine the frequency of the outcome we should expect among the study subjects exposed to the factor (in this case, electronic monitoring) if the study factor has no effect? This 'expected frequency under the null hypothesis' is usually determined by examining the study subjects who were not exposed to the study factor (above, the unmonitored) and then calculating what outcome frequency the exposed would have had if they had the same experience as the unexposed.² If we had not been aware that arrested labour was more frequent among the monitored subjects, we would have measured monitoring's effect by simply comparing the caesarean rate among the monitored with the rate among the unmonitored (as in Table 1). Upon finding that arrested labour is more frequent among the monitored subjects, we would realise

that, even if monitoring has no effect, we should expect a higher frequency of caesarean among the monitored subjects. This illustrates the characteristic that makes a potential confounder into a causal confounder: in order for causal confounding to occur in a study, the extraneous factor must not only be an independent cause of the outcome, but must also occur at differing frequencies among those study subjects exposed to the study factor and the unexposed group used for comparison (i.e., it must be associated with the study factor in the study cohort).

In order to calculate just how much higher a frequency of caesarean we should expect among the monitored, we employ a principle from indirect ('internal') standardisation:² we calculate the frequency that would be seen in the monitored group if, at each level of arrested labour, they had the same caesarean frequency as the unmonitored group. Thus, we are led to stratify the data by arrested labour before making any assessment of monitoring's effect. Table 3 presents the same study data, stratified on arrest of labour. It now appears that monitoring multiplies caesarean risk by an average of 1.2 times among nonarrested labours and 1.3 times among arrested labours, both effects notably less than the result in Table 1. Thus, we conclude that the estimate of monitoring's effect given in Table 1 was, in fact, a mixture of true effects of monitoring and some effects which were not due to monitoring. The analytic process just described is often termed 'control of confounding'.

The above example illustrates the problem of confounding by indication in the assessment of a medical procedure. Arrested labour is a factor which may directly contribute to the decision of whether to monitor the labour; thus, it is an indication for the monitoring procedure, and any nonexperimental series of subjects would be expected to show a higher proportion of arrested labours among the monitored subjects. Since arrested labour is also

TABLE 3 *Monitoring caesarean study, stratified by arrested labour*

	Arrested		Not Arrested		
	Monitored	Unmonitored	Monitored	Unmonitored	
Caesarean	481	125	Caesarean	299	283
Vaginal	802	300	Vaginal	5 718	6 476
TOTALS	1 283	425	TOTALS	6 017	6 759
Caesarean risk:	375/1 000	294/1 000		50/1 000	42/1 000
Risk ratio =		1.3	Risk ratio =		1.2

an indication for caesarean section, we would always expect a higher caesarean frequency among monitored labours, even if monitoring had no effect.

When an extraneous condition is both a clinical indication for the application of the procedure under study and an indication for the outcome under study (as in Example 1), we should expect confounding by the indication to occur in any nonexperimental study of the procedure and the outcome. Such studies must be able to control potentially confounding indications, in order to determine whether confounding by indication is present. Thus, studies should be planned so that information on the status of important clinical indications will be available for each subject.

Examples of Nonconfounders

Thus far, we have shown how the effects of an extraneous cause of the outcome under study can be confused (confounded) with the effect under study. Other situations can arise which numerically mimic the above situation, but in which there is, in fact, no causal confounding by the extraneous factor. That is, there are variables which are inappropriate to control in the statistical analysis even though controlling them leads to a different estimate of effect. As we will discuss later, the distinction between causal confounders and such

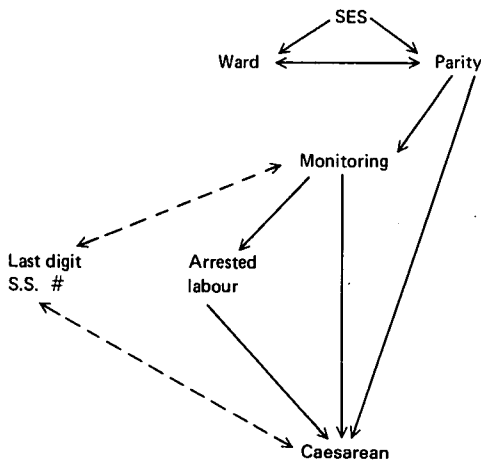
numerical mimics is not based on statistical considerations.

Figure 1 presents a path diagram which will illustrate this point. It shows some of the hypothetical causal influences (solid single-headed arrows), causally induced correlations (solid two-headed arrows between variables sharing an antecedent cause), and chance correlations (dashed two-headed arrows) in our foetal monitoring study. For each variable illustrated, we discuss the appropriateness of its control as a confounder.

Example 2: Suppose that we discovered that the excess of arrested labours among the monitored labours, seen in Table 2, was a result of an increased tendency to diagnose a labour as arrested in the presence of the monitor (as in Figure 1) rather than an increased tendency to monitor arrested labours. Then the estimate of monitoring's effect given in Table 1 (risk ratio = 1.9) would be the proper one. This is because most of monitoring's effect on caesarean risk would be transmitted through the mechanism of increasing the diagnosis of arrested labour. The effect of the excess of arrested labours among the monitored would no longer be considered as an extraneous effect to be separated out by stratification (as in Table 3); rather, it would now be considered an integral part of the effect of monitoring on the caesarean rate. Thus, Table 1, which incorporates the effect of the excess arrested labours into the effect of monitoring, would give the proper estimate of monitoring's total effect on caesarean risk, while Table 3 would underestimate the total effect.

In Example 1, arrested labour was presented as a confounding variable; in Example 2 and Figure 1, it is presented as a intermediary variable. An intermediary variable is a cause of the outcome which is caused by the factor under study; thus, it is an intermediate step in the causal pathway linking the study factor to the outcome. Such variables should *not* be controlled in the assessment of the total impact of a medical procedure.

Example 3: Suppose that, as in Example 2 and Figure 1, monitoring results in an increased diagnosis of arrested labour, but that we are now interested in estimating the total extent of this increase. Table 2 provides one estimate of this effect: the risk of diagnosis of arrest among the monitored is 3 times the risk of diagnosis among the unmonitored. Table 4 provides a rearrangement of the data in Table 3 to show the relationship of monitoring and arrested labour within the levels of mode of delivery:



- > : causal path
- <—> : correlation induced by causal relationships
- <—> : noncausal correlation

SES = Socioeconomic status; S.S. # = social security number

FIGURE 1 Path diagram of causal relationships in foetal monitoring examples 2-5

TABLE 4 *Monitoring – arrested labour study, stratified by mode of delivery*

	Caesarean		Arrested	Vaginal	
	Monitored	Unmonitored		Monitored	Unmonitored
Arrested	481	125	802		300
Not Arrested	299	283	5 718		6 476
TOTALS	780	408	6 520		6 776
Percent Arrested:	61.7	30.6	12.3		4.4
Ratio =	2.0		Ratio =	2.8	

the relationship of monitoring and arrested labour appears weaker than in Table 2, especially among C-sectioned women. Which analysis, Table 2 or 4, gives the proper total effect of monitoring on arrest diagnosis? If we note that mode of delivery is affected by arrested labour rather than the other way around, we resolve the question. Mode of delivery has no effect on monitoring, and so it is nonsense to talk of confounding by mode of delivery in this case. Thus, Table 2, ignoring mode of delivery, provides the proper estimate of the total effect of monitoring on arrested labour diagnosis.

Example 4: We may also be confronted with variables which by chance numerically mimic confounders, but which have absolutely no causal connections to the study variables. For example, the variable 'last digit on the mother's social security number' might appear to make a difference if controlled in the analysis. Clearly, such variables are misleading and should be ignored in the analysis, since they have no effects to be confounded with the effect under study.

A final type of variable we may encounter is the proxy control variable. This is a variable which, though not a causal confounder, may serve to control confounding in the analysis in place of a causal confounder. This often occurs in situations in which a suspected causal confounder has not been recorded in our data, but a variable thought to be naturally correlated with the confounder is recorded. In order to serve as an effective proxy for the unmeasured variable, the correlate must not itself be an intermediary variable or an effect of the outcome.

Example 5: From Figure 1, we can see that parity is a potential confounder in studying the effect of monitoring on caesarean. Suppose in our study of monitoring and caesarean, we had failed to record parity of the mother, but we had recorded ward

status (public/private). Because parity and ward are both influenced by socioeconomic status (SES), we would expect them to be correlated in the population and thus correlated in our data: low SES women tend to have more children and end up having them in the public ward; thus, the public ward group should have a higher average parity than the private ward group. Since ward status is neither intermediary nor an effect of caesarean, ward can serve as a proxy control variable for parity.

Note that a high degree of subject-matter judgment may be required in deciding whether a variable should be used in the analysis as a proxy for a causal confounder. (The construction of path diagrams such as Figure 1 can be a great aid in explicitly presenting underlying judgments.) Note also that a proxy will allow us to remove only a fraction of confounding due to the unmeasured causal confounder. However, one variable will often provide proxy control for several different causal confounders, and a variable may be both a proxy and a causal confounder in its own right.

To recap up to this point, a potential confounder is an extraneous factor with effects that could be mixed up with the effect under study; a causal confounder is an extraneous factor whose effects have become mixed with the effect under study. Separation of the effects of confounders from the effect under study is termed control of confounding; this control may be effected by stratifying on the confounders or available proxies. The numerical relationships between variables are not sufficient to identify causal confounders; subject-matter judgments regarding direction of effects are also required.

We next address some statistical issues in the identification and control of confounding.

Statistical Methods in the Control of Confounding

It is important to note that the defining features of a causal confounder are not statistical properties. First, statistics does not address whether an observed

association is causal or not, this being a purely subject-matter topic.³ A similar comment applies to determining the intermediary status of a variable. But the most misunderstood nonstatistical quality of a confounder is its association with the study factor, and we address this issue first.

It is very common to see studies which claim that an extraneous variable was not a confounder in the study because the difference between study groups on the extraneous variable was statistically nonsignificant, or because randomisation was used to assign patients to treatment. Both these claims can be very misleading in relatively small studies.⁵

Example 6: Tables 5–7 present an example of the fallacy of statistically testing the relationship of a confounder and the study factor, and the pos-

TABLE 5 *Trial of monitoring: Distribution of malpresentations after randomisation*

	Monitored	Unmonitored
Malpresentations	10	3
Normal presentations	40	47
TOTALS	50	50
Percent malpresentation:	20.0	6.0

TABLE 6 *Trial of monitoring: Unstratified outcome*

	Monitored	Unmonitored
Caesarean	11	3
Vaginal	39	47
TOTALS	50	50
Caesarean risk:	220/1 000	60/1 000

Risk ratio = 3.7

Continuity-corrected $\chi^2 = 4.07$, 1 d.f., $p < .05$

sibility of confounding in randomised trials. Suppose that a clinical trial of the effect of monitoring on caesarean has been performed, with subjects randomised into monitored and unmonitored groups. Table 5 presents the sort of outcome of such randomisation which would commonly occur: by chance alone, malpresentations are more common in the monitored group, but the difference is not statistically significant ($p > .05$). Table 6 presents the outcome of the trial, unstratified by presentation: it appears that monitoring multiplies caesarean risk by a factor of 3.7, and this is statistically significant ($p < .05$). But Table 7 presents the trial data stratified on presentation: there, monitoring appears to multiply risk by about 2.2 times, and this effect is not at all statistically significant ($p > .10$). Thus, malpresentation was, in fact, a causal confounder.

We noted earlier that in order for an extraneous factor to be a causal confounder, it must be associated with the study factor in the study cohort. By the definition of 'p-value',⁴ the p-value for comparing the study groups on the extraneous variable tells us only how often the observed association would occur if we repeated the trial over and over again; this is an irrelevant bit of information, since we only have a single trial to analyse. The relevant question is whether an important association occurred in the single study we have, and this is a purely nonstatistical issue.⁵ The above example illustrates this point, as well as emphasising that confounding can occur even if randomisation was used to form the study groups. Randomisation is only a device to achieve similarity between the study groups on the average over many repetitions of the study; the device is not guaranteed to achieve similarity for every single factor in any particular study.^{4,5} Although the chance of 'randomisation failure' decreases with increasing sample size, randomisation and significance tests will not necessarily free us from having to control extraneous factors in order to avoid confounding in our

TABLE 7 *Trial of monitoring: Outcome stratified on presentation*

	Malpresentations		Normal presentations		
	Monitored	Unmonitored	Monitored	Unmonitored	
Caesarean	7	1	Caesarean	4	2
Vaginal	3	2	Vaginal	36	45
TOTALS	10	3		40	47
Caesarean risk:	700/1 000	333/1 000		100/1 000	43/1 000

Risk ratio = 2.1

Risk ratio = 2.3

Mantel-Haenszel $\chi^2 = 1.27$, 1 d.f., $p > .10$

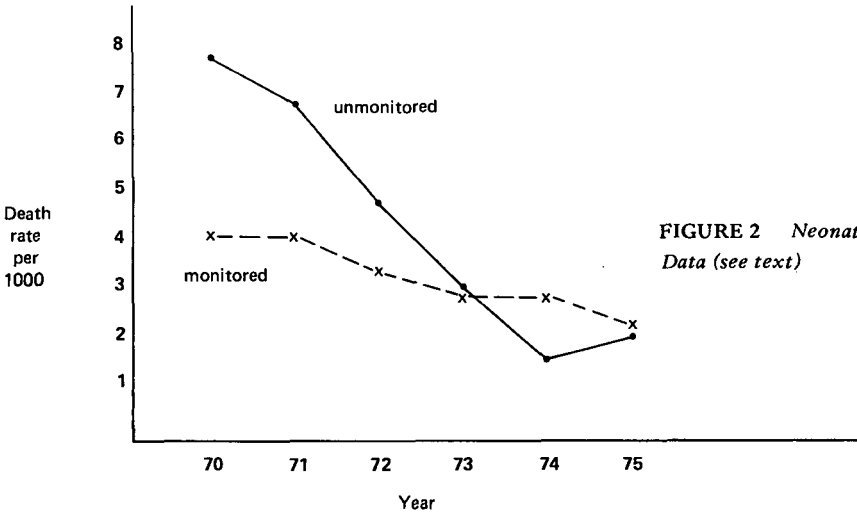


FIGURE 2 Neonatal death rates by year, Beth Israel
Data (see text)

●—● : Unmonitored x- - - - x : Monitored

Summary ratio of rates, unmonitored vs. monitored: 1.3
Likelihood-ratio test for trend of death rates by year
(from logistic model): $p > .05$;
L-R test for effect of monitoring: $p > .30$ (see ref.4)

particular study.⁵

A more subtle fallacy, related to the issue of causality, is the belief that if the extraneous variable did not have a statistically significant effect on the outcome during the course of the trial, no confounding occurred. This fallacy has been theoretically discussed before;^{6,7} we give an example from our own experience.

Example 7: Figure 2 presents data from a study of the effect of monitoring on neonatal death.⁸ We see a distinct graphical trend of the death rates by year of delivery; however, neither monitoring nor year is statistically significantly associated with neonatal death. The average effect of monitoring in Figure 2 is to reduce risk by a factor of 23%,

with no effect in later years. If we decide not to stratify on year of delivery because of the non-significance of its association with the outcome, we obtain Table 8: monitoring appears noticeably stronger in effect (it reduces risk by a factor of 40%) and the effect is now significant ($p < .05$). Which analysis is more likely correct, Figure 2 or Table 8? At this point, knowledge external to the study must be given priority. It is well known that the neonatal death rate dropped during the study period at large hospitals throughout the United States, and only partly because monitoring became more common;⁹ the trends in Figure 2 reflect this phenomenon. The factors involved may be referred to as 'changes in obstetric policy'—a well-documented secular improvement in the quality of neonatal care, independent of monitoring. Because frequency of monitoring increased so dramatically over the study period, the analysis in Table 8 confounds the secular change in obstetric policy with the actual effects of monitoring. That there is a secular trend in policy independent of monitoring is reflected in Figure 2 by the decreasing neonatal death rates across year in both the monitored and unmonitored groups. Thus, Figure 2 shows the appropriate analysis, even though the association of year and outcome was not statistically significant in the study.

TABLE 8 Comparison of neonatal death rates without adjusting for secular trend

	Monitored	Unmonitored
Neonatal death	23	38
Survival	7 263	7 128
TOTALS	7 286	7 166
Death rate	3.2/1 000	5.3/1 000
Rate ratio = 1.7		
$\chi^2 = 3.96, p < .05$		

The preceding examples illustrate the general

fallacy of using significance tests to detect confounding. Causal confounding is not a frequency property of repeated sampling, but a property of a particular data set; p-values refer to a property of repeated sampling under a hypothetical sampling or randomisation model, and so, as has been noted,¹⁰ they do not bear on questions concerning characteristics of particular outcomes. A deeper problem in attempting to correct ordinary statistical procedures for the detection of confounding is that the notion of confounding is inherently Bayesian, in that prior judgment of various causal connections must be made in order to identify which variables would be logical candidates as confounders. Considering again foetal monitoring and caesarean section: note that the last digit of the mother's social security number would not be considered a logical candidate as a confounder, even if (by chance) it was significantly associated with both monitoring and caesarean; we possess strong prior views that this variable is causally irrelevant to the outcome. The earlier examples similarly rely on prior judgments as to relevancy of variables, and such clinical judgments are essential and unavoidable in the analysis of medical technology. Recognising that such judgments vary from clinician to clinician, we emphasise again that any analysis should make explicit what clinical judgments were made in the selection of variables to be controlled in the analysis. Path diagrams³ are useful in summarising these judgments for presentation.

The use of multivariate techniques in selecting and ranking confounders requires a special caution. Ordinary stepwise regression techniques can be extremely misleading in selecting the most important confounders. This follows from the points illustrated in the monitoring-year-neonatal death example: stepwise regression examines only the association of the extraneous variables with outcome, never displaying the equally important association of the extraneous variables with the study factor; and the ranking of importance is based on significance tests (in the form of F ratios or similar statistics),

which do not accurately indicate the presence of a causal association with the outcome. Multivariate analysis of confounding requires techniques which involve regressions for both the outcome and the study factor, and examination of sign and magnitude of coefficients, such as systems regression (structural equations).³ After having selected confounders for control, we may have too many variables for the type of stratified analysis used in the earlier examples. In such cases, we may employ analysis by regression models⁴ or stratification based on risk scores.¹¹ Examples of analyses of the effect of a medical procedure (foetal monitoring) in a presence of multiple confounders are given in references 1 and 8.

ACKNOWLEDGEMENT

We are grateful to Dr Olli Miettinen for his stimulating discussions and correspondence on this topic.

REFERENCES

- 1 Neutra RR, Greenland S and Friedman EA. The effect of fetal monitoring on cesarean rates. *Obstet Gynecol* 1980; 55: 175-180.
- 2 Miettinen OS. Standardization of risk ratios. *Am J Epidemiol* 1972; 96: 383-388.
- 3 Duncan OD. Introduction to Structural Equations Models. New York: Academic Press, 1975.
- 4 Cox DR and Hinkley DV. Theoretical Statistics. London: Chapman & Hall, 1974.
- 5 Rothman KJ. Epidemiologic methods in clinical trials. *Cancer* 1977; 39: 1771-1775.
- 6 Miettinen OS. Reply by Dr Miettinen. *Am J Epidemiol* 1977; 106: 191-193.
- 7 Dales LD, Ury HK. An improper use of significance testing in studying covariables. *Int J Epidemiol* 1978; 7: 373-375.
- 8 Neutra RR, Fienberg SE, Greenland S and Friedman EA. The effect of fetal monitoring on neonatal death rates. *N Engl J Med* 1978; 299: 324-326.
- 9 Antenatal Diagnosis: Report of Consensus Development Conference. US Dept. of Health, Education, and Welfare, NIH Publ. No. 79-1973, 1979.
- 10 Hacking I. Logic of Statistical Inference. New York: Cambridge, 1964.
- 11 Miettinen OS. Stratification by a multivariate confounder score. *Am J Epidemiol* 1976; 104: 609-620.

(Revised version received 27 May 1980)