Behaviormetrika
1980, No. 8, 41-55

AN ALGORITHM FOR CAUSAL MODELING IN EPIDEMIOLOGICAL STUDIES

Hirofumi Takagi*

By extending the path analytic method, a new algorithm for causal modeling in epidemiological etiologies is developed, which is based mainly on simple correlation coefficients.

The utility of the proposed analysis is assessed, relative to cluster analysis, factor analysis, and multiple regression analysis. As the result of this evaluation, it is suggested that multivariate analysis should be used together with causal modeling, since multivariate analysis, especially factor analysis, appears to be a useful tool of confirming the estimated causal model.

1. Introduction

The establishment of causal relationships is extremely important in the field of epidemiology, as it is in other fields of science. As a result of many recent advances (Blalock, 1964; Heise, 1975; Kenney, 1979), path analysis has become a very effective procedure for developing such causal models. However, in dealing with epidemiological data (as opposed to data generated for a controlled animal experiment), the investigator must incorporate his own personal knowledge of the various relationships that exist among the variables under study in order to develop a causal model. At the outset of an observational study, one may have trouble with the selection of items for research; that is, one may not be able to assume *a priori* what variables cause the effect of interest. For example, in the case of most of chronic diseases such as cancer, there may be many factors which potentiall yplay a causal role in the onset of the disease — a role which cannot be evaluated until the data analysis is completed. Of course, sometimes one may obtain information about the target disease from other studies. However, even if one could obtain some information about the disease of interest, the problem of developing a causal model still remains.

We have no systematic method for model development and must rely on our own intuition and the limited information available from other studies. Therefore, one must frequently select the items for study in the absence of a model, and then develop suggested causal relationships through the data analysis.

On the other hand, one could use the more sophisticated methods of multivariate analysis. However, only a few researchers have applied them to causal analysis, even though multivariate analysis potentially seems to be a very powerful tool for epidemiological studies.

In general, it has been often said that no correlational method can establish causal

Key Words and Phrases; causal model, path analysis, epidemiology, cluster analysis, factor analysis, multiple regression analysis

* Biometry Branch, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709, U.S.A.

relations (Cohen and Cohen, 1975). and that correlation does not imply causation (Kenney, 1979). However, correlation can be used in constructing the model since correlational analysis can suggest the relationship between the two variables under study. Thus it seems reasonable to employ correlation coefficient in causal modeling.

Even though one may not necessarily elucidate the true structural model, mechanistic or biological, through the use of correlation, causal modeling can assist in the development, modification, and extension of measurement and substantive theory (Kenney, 1979).

As Blalock has stated, causal laws are the working hypothesis of a scientist (Blalock, 1964). Therefore, this paper presents a new algorithm for causal modeling in the fields of nonexperimental sciences, especially epidemiology, for someone who has sufficient knowledge in the field.

2. Methods

First of all, we shall define the word "causation" or "causality" used in this paper, in a particular way to be, although many people have already tried to define it from philosophical as well as statistical views. In common use of the terms, if and only if the occurence of some event always produces a specific phenomenon after a while, the former is called "cause", and the latter "effect". However, it is not easy to find and confirm such a relationship through ecological studies because many systematic investigations repeated over a long time are essentially required to confirm the relationship called "causation".

Since the effect, like cancer mortality, is always fixed, our interest is mainly directed toward the discovery of potential variables of causes. As mentioned above, one will be unable to confirm them through statistical analysis. Therefore, in this paper, we shall use correlation instead of causation, and shall be satisfied with interpreting such correlations as "causality" not in a strict philosophical sense, but only in a statistical sense.

Now, let us assume (a) that correlation indicates causality. In addition to assumption (a), let us assume as well (b) that the direction of the influence of each variable is always pointed to a specific effect variable, usually some disease, and (c) that there is no feedback loop in the model.

Although these three assumptions may not be correct and may be too simple, when one tries to analyze some actual data statistically, one can obtain a useful algorithm for causal modeling by imposing them, as the first approximation, as shall be described later.

Note that the terminology used in this paper is originally derived from path analysis (*Blalock*, 1964; *Heise* 1975; *Kenney*, 1979).

The following two concepts are valuable for causal modeling.

I. Correlation and Causal Chain

Let us consider a causal chain of three variables; that is, variable X_1 causes variable X_2 that causes variable X_3 in turn. This causal chain can be expressed as follows:

$$X_1 \xrightarrow{a} X_2 \xrightarrow{b} X_3$$

where a and b on the lines are called "path coefficients" or "structure coefficients" in terms of path analysis, and they are in fact theoretical regression coefficients (Heise, 1975; Kenney, 1979).

Since it is rare that all the causes of a variable can be found, another cause must be added that represents all the unspecified causes of the endogenous variable, and this residual term is called the "disturbance" (Kenney, 1979).

Let us express variables X_2 and X_3 by linear regression equations respectively. However, in order to suggest that these equations represent the causal relations, we use the term "structure equation" instead of regression equation (Kenney, 1979).

Structure equations of X_2 and X_3 are given by

$$X_2 = aX_1 + cU \tag{1}$$

and

$$X_3 = bX_2 + dV \tag{2}$$

where U and V are disturbances of X_2 and X_3 respectively, and c and d represent their coefficients. We assume that Cov (X_1, U)=Cov (X_1, V)=Cov(X_2, V)=0.

For the sake of convenience, let us assume (e) that all variables including disturbances are standardized, so they have zero means and unit variances.

Now, we can obtain correlations between each pair of variables.

$$\begin{aligned} r_{12} &= Cov(X_1, X_2) = Cov(X_1, aX_1 + cU) \\ &= Cov(X_1, aX_1) + Cov(X_1, cU) \\ &= a\, Cov(X_1, X_1) + c\, Cov(X_1, U) \\ &= a \end{aligned}$$

since X_1 is a standardized random variable with unit variance, and X_1 and U have no correlation because we have assumed that no systematic equation is obtained that indicates the relationship between them.

Similarly,

$$r_{23} = b,\ r_{13} = ab.$$

Then,

$$r_{13} = r_{12}r_{23} \leqq r_{12} \text{ and/or } r_{23} \tag{3}$$

since correlation is less than or equal to one. Therefore, the variable put on the nearest position in the causal chain to a specific variable should be most highly correlated with it. By the way, those correlations are not independent of each other theoretically because of the structure of the model. For example, the relation that $r_{13}=r_{12}r_{23}$ is expected theoretically in the three variable causal chain. This kind of relations among correlation coefficients is called the "restriction" in terms of path analysis (Blalock, 1964). In general, when looking for models, all restrictions in the model should correspond to the values obtained by the actual data, and if some discrepancies exist, the model should be modified to correct such discrepancies.

It is easy to extend this result to more than three variables (Blalock, 1964).

Conversely, given $r_{13} \leqq r_{12}$ and/or r_{23}, how could one express the causal chain ? In this case, the casual chain cannot be determined uniquely. However, if variable X_3 is the effect like cancer death, one can determine it by using assumptions (b) and (c); that is, X_1 causes X_2 that causes X_3. Nevertheless, this model has a restriction as mentioned above. If actual correlations obtained by data indicate that r_{13} is not equal to $r_{12}r_{23}$, one needs to modify the causal chain.

In general, whenever one manages more than three variables, this kind of discrepancy must occur in the actual case.

For the modification of such a discrepancy, one should consider the addition of a path between two variables. In this paper, we pay special attention to the path between the effect, like cancer incidence, and each variable.

II. Direct Effect

When variable Z is linked to variable W in the process of causal modeling, we need to estimate the direct effect of Z on Y.

Let us assume that p variables, say $X_1, X_2, \ldots, X_p$, cause Y directly, and that all variables are standardized as mentioned in assumption (e). Whether variable W is included in these p variables or not does not influence the result, although W should be at least linked directly to one of these p variables or Y. Thus, there are $p+1$ variables that are assumed to cause Y directly; that is, $X_1, X_2, \cdots, X_p$, and Z.

Therefore, the structure equation of Y is given by

$$Y = \sum_{k=1}^{p} D_k X_k + D_z Z + eU \tag{4}$$

where D_k and D_z represent direct effects of X_k $(k=1, 2, \cdots, p)$ and Z respectively, and U is a disturbance influencing Y.

Using equation (4), one can obtain correlations of Y with W and Z. Since the distrubance of Y is assumed to have no correlation with any variable except Y,

$$\begin{aligned} r_{yw} &= Cov(Y, W) = Cov\left(\sum_{k=1}^{p} D_k X_k + D_z Z + eU, W\right) \\ &= \sum_{k=1}^{p} D_k \, Cov(X_k, W) + D_z \, Cov(Z, W) + e \, Cov(U, W) \\ &= \sum_{k=1}^{p} D_k \, r_{kw} + D_z \, r_{wz} \end{aligned} \tag{5}$$

and similarly

$$r_{yz} = \sum_{k=1}^{p} D_k \, r_{kz} + D_z \tag{6}$$

On the other hand, since Z can be related to X_k $(k=1, 2, \cdots, p)$ only through W, we can use the concept of a three-variable causal chain. Therefore, the following equation holds,

$$r_{kz} = r_{kw} \, r_{wz} \quad (k=1, 2, \cdots, p) \tag{7}$$

By substituting equation (7) into equation (6),

$$r_{yz} = r_{wz}\left(\sum_{k=1}^{p} D_k r_{kw}\right) + D_z \tag{8}$$

Using equations (5) and (8), the direct effect of Z is given by

$$D_z = \frac{r_{yz} - r_{yw} r_{wz}}{1 - r_{wz}^2} \tag{9}$$

After all, equation (9) corresponds to the standard partial regression coefficient of Z in the case of multiple regression analysis of Y on two variables W and Z. Note that the direct effect of each variable may be changed after a new variable is included in the model, so that one needs to estimate all coefficients when the causal model under study is finally obtained.

To test whether the direct effect of Z is nonzero, one can apply the method for testing a partial regression coefficient. Therefore, one computes,

$$t = \frac{\sqrt{n-3}\,(r_{yz} - r_{yw} r_{wz})}{\sqrt{1 - r_{yw}^2 - r_{yz}^2 - r_{wz}^2 + 2 r_{yw} r_{yz} r_{wz}}} \tag{10}$$

and refer t to a Student's t distribution with $n-3$ degrees of freedom, where n is the sample size. Alternatively, equation (10) can be easily obtained from the formula for a test of the first order partial correlation between Y and Z holding W fixed.

If the value for t is less than a given significance level, one can neglect the direct path from Z to Y.

Now, applying the above concepts, the algorithm for causal modeling shall be described in the following section.

III. Algorithm for Causal Modeling

The algorithm for causal modeling is described here, and an example is worked simultanously, using the data in Table 1, in which correlations are calculated by the use of some model that is presumed artificially in order to check the algorithm and

Table 1

Correlation matrix

X_1	.180									
X_2	.140	.025								
X_3	.439	.300	.061							
X_4	.042	.008	.300	.018						
X_5	.324	.200	.045	.060	.014					
X_6	.185	.000	.026	.200	.008	.000				
X_7	.019	.003	.138	.008	.405	.006	.004			
X_8	.374	.000	.052	.060	.016	.000	.300	.007		
X_9	.028	.005	.200	.012	.060	.009	.005	.114	.010	
X_{10}	.008	.001	.060	.004	.018	.003	.001	.307	.003	.300
	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9

* Sample size is assumed to be 100

compare the result with other methods of multivariate analysis.

Step 1 Select the variable most highly correlated with Y.

This step is directly derived from the concept of a causal chain. Thus, variable X_3 can be selected because of its highest correlation with Y. This step is shown in Table 2 and Figure 1-a. In this example, the statistical significance level is set at 0.05.

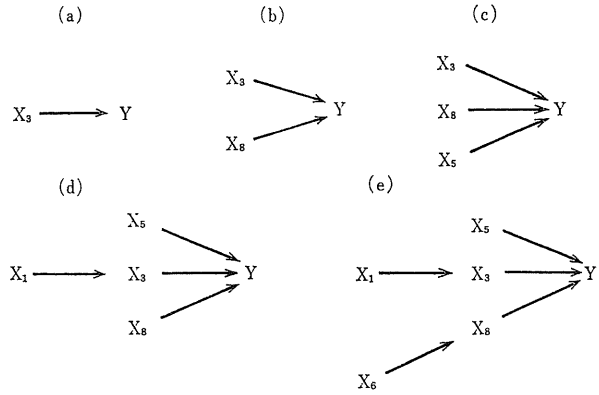

Fig. 1 The process of causal modeling

Step 2 Select the variable most highly correlated with Y or one of the variables already selected. If the new variable is not linked directly to Y, then compute the direct effect on Y by equation (9) and test the effect by equation (10); if the direct effect is significant, draw the path from the variable to Y.

This step consists of two parts; one concerns the selection of a variable by using the concept of causal chain, and the other is about the direct effect of the variable on Y.

Variable X_8, for example, can be selected as shown in Table 2, and one does not need to calculate the direct effect because of its direct linkage to Y.

Table 2

Process of causal linkage

Step	Linkage	Correlation Coefficient	t (df=98)	p	Direct Effect	t (df=97)	p
1	$X_3 \rightarrow Y$	0.439	4.837	<0.001	—	—	—
2	$X_8 \rightarrow Y$	0.374	3.992	<0.001	—	—	—
3	$X_5 \rightarrow Y$	0.324	3.390	0.001	—	—	—
4	$X_1 \rightarrow X_3$	0.300	3.113	0.002	0.053	0.556	0.580*
5	$X_6 \rightarrow X_8$	0.300	3.113	0.002	0.080	0.813	0.418*
6	$X_2 \rightarrow Y$	0.140	1.400	0.165**			

* No direct path exists.

** X_2 can be excluded from the model.

Step 3 Repeat Step 2 until a given significance level is unsatisfied with respect to the correlation coefficient between two variables.

This step is for the continuation and stopping of the selection of a variable.

We would continue the selection of variables and develop the model. Variable X_5 has the highest correlation with Y, 0.324, therefore, X_5 is selected and added into the

model as shown in Figure 1-c. Next, one needs to find the variable that is most highly correlated with Y, X_3, X_8, or X_5. Variable X_1 can be found since it has the highest correlation with X_3, 0.300. In this case, one needs to compute the direct effect of X_1 because it is linked to X_3 but not to Y. As shown in Table 2, the direct effect is not statistically significant (the probability is greater than 0.05). So, one can neglect the direct path from X_1 to Y. This means that X_1 causes Y indirectly only through variable X_3. It can be shown that the situation of variable X_6 is exactly the same as that of X_1, as shown in Table 2.

Finally, variable X_2 is selected, but its correlation with Y, 0.140, does not satisfy the statistical significance level. Therefore, X_2 is excluded from the model, and the procedure of selection is stopped.

These processes are shown in Table 2, and causal models at each step are also shown in Figure 1.

Step 4 Solve all path coefficients in the obtained model mathematically, and examine the restrictions of correlations conditioned by the estimated model. If one finds any discrepancies of the restrictions, modify the model to correct the discrepancies.

For instance, let us consider the model shown in Figure 2-a, which is derived from the model shown in Figure 1-e by adding the disturbances of Y, X_3, and X_8.

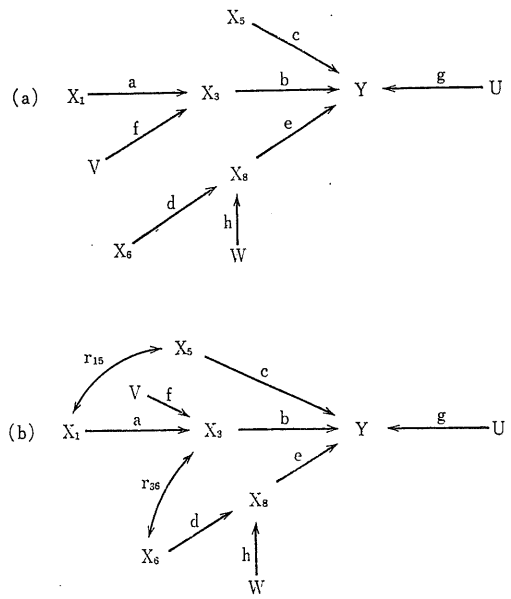

Fig. 2 Estimated causal models

* U,V, and W express disturbances influencing Y, X_3, and X_8 respectively

Structure equations are given by

$$Y = b\,X_3 + c\,X_5 + e\,X_8 + g\,U \tag{11}$$

$$X_3 = a X_1 + f V \tag{12}$$

and

$$X_8 = d X_6 + h W \tag{13}$$

By using equations (11), (12), and (13), all of the possible correlations between each pair of variables can be expressed by path coefficients and correlation coefficients. However, some correlations can not be expressed by using the equations. These correlations are assumed to be zero, and also called "restrictions". For example, the correlation between X_1 and X_6 cannot be expressed because there is no equation that indicates the relationship between them. Therefore, it is presumed that $r_{16}=0$. Similarly, seven other simple correlations are expected to be zero as shown in Table 3. Moreover, relations among some correlations may show the other types of restrictions like that a first order partial correlation being equal to zero. For instance, by using equation (11), it is shown that $r_{y1}=b\, r_{13}$, and that $r_{y3}=b$. These relations indicate that $r_{y1}=r_{y3}\, r_{13}$, *i.e.*, $r_{y1\cdot 3}=0$. Similarly, one finds as well that $r_{y6\cdot 8}=0$ as shown in Table 3.

Table 3
Restrictions of correlation coefficients concerning two models

Model (a)		Model (b)	
Expected	Real	Expected	Real
$r_{15}=0$	0.200*	$r_{16\cdot 3}=0$	−0.064
$r_{16}=0$	0.000	$r_{18\cdot 6}=0$	0.000
$r_{18}=0$	0.000	$r_{35\cdot 1}=0$	0.000
$r_{35}=0$	0.060	$r_{38\cdot 6}=0$	0.000
$r_{36}=0$	0.200*	$r_{56\cdot 3}=0$	−0.012
$r_{38}=0$	0.060	$r_{58\cdot 6}=0$	0.000
$r_{56}=0$	0.000	$r_{y1\cdot 35}=0$	−0.008
$r_{58}=0$	0.000	$r_{y6\cdot 38}=0$	−0.004
$r_{y1\cdot 3}=0$	0.056		
$r_{y6\cdot 8}=0$	0.082		

* $p<0.05$

In comparison with actual correlations, one finds that two discrepancies of restrictions occur; that is, $r_{15}\neq 0$, and $r_{36}\neq 0$ as shown in Table 3.

Therefore, one should assume correlations between X_1 and X_5, and X_3 and X_6 respectively, so that the model should be modified as shown in Figure 2-b, where double arrowhead curves indicate correlations between two variables.

Again, one should examine all of the restrictions in the new model, considering the conditions that $r_{15}\neq 0$, and that $r_{36}\neq 0$. In this case, three structure equations are not changed, but since two correlations are assumed nonzero as mentioned above, it may be difficult to obtain all of the restrictions but not impossible.

Alternatively, one can assume directions for each path between causal variables, and examine the restrictions in each model. If one prefers this method, one obtains four different models and may find the most suitable model through examining the restrictions. Probably this method is better than only assuming correlations. However, one needs to examine each model one by one, so that we decide to analyze the first

model as shown in Figure 2-b because of the limitation of the space of this paper.

To find all of the restrictions, it may be necessary to solve some complicated equations. However, if one knows the method of path analysis, it is easy to obtain them at a glance. Therefore, the references concerning path analysis are recommended (Blalock, 1964; Heise, 1975; Kenney, 1979).

As a result, the restrictions in the model are obtained as shown in Table 3, and are almost satisfactory in statistical sense.

What remains is to estimate all of the path coefficients in the model. By using equations (11), (12), and (13), and considering that $r_{15} \neq 0$, and that $r_{36} \neq 0$, all solutions can be obtained mathematically as shown in Table 4.

Table 4
Solutions of path coefficients

$$a = r_{13} = r_{16}/r_{36} = r_{35}/r_{15} = r_{56}/r_{15}r_{36} = r_{18}/r_{36}r_{68}$$
$$= r_{58}/r_{15}r_{36}r_{68} = r_{18}/r_{38} = r_{58}/r_{15}r_{38}$$

$$d = r_{68} = r_{38}/r_{36} = r_{18}/r_{13}r_{36} = r_{58}/r_{13}r_{15}r_{36} = r_{18}/r_{16} = r_{58}/r_{56}$$
$$= r_{58}/r_{15}r_{16} = r_{15}r_{18}/r_{35}r_{36} = r_{58}/r_{35}r_{36} = r_{15}r_{18}/r_{56}$$

$$b = \frac{r_{y3}(1-r_{58}^2)}{(1-r_{35}^2)(1-r_{38}^2)} - \frac{r_{y5}r_{35}}{1-r_{35}^2} - \frac{r_{y8}r_{38}}{1-r_{36}^2} = \frac{r_{y3}(1-r_{16}^2)}{(1-r_{13}^2)(1-r_{36}^2)} - \frac{r_{y1}r_{13}}{1-r_{13}^2} - \frac{r_{y6}r_{36}}{1-r_{36}^2}$$

$$= \frac{r_{y3}(1-r_{18}^2)}{(1-r_{13}^2)(1-r_{38}^2)} - \frac{r_{y1}r_{13}}{1-r_{13}^2} - \frac{r_{y8}r_{38}}{1-r_{38}^2} = \frac{r_{y1}(1-r_{56}^2) - r_{y5}r_{15}(1-r_{16}^2)}{r_{13}(1-r_{15}^2)(1-r_{36}^2)} - \frac{r_{y6}r_{36}}{1-r_{36}^2}$$

$$= \frac{r_{y1}(1-r_{58}^2) - r_{y5}r_{15}(1-r_{18}^2)}{r_{13}(1-r_{15}^2)(1-r_{38}^2)} - \frac{r_{y8}r_{38}}{1-r_{38}^2} = \frac{y_6(1-r_{18}^2) - r_{y8}r_{68}(1-r_{16}^2)}{r_{36}(1-r_{13}^2)(1-r_{68}^2)} - \frac{r_{y1}r_{13}}{1-r_{13}^2}$$

$$= \frac{r_{y3}(1-r_{56}^2)}{(1-r_{36}^2)(1-r_{35}^2)} - \frac{r_{y5}r_{35}}{1-r_{35}^2} - \frac{r_{y6}r_{36}}{1-r_{36}^2} = \frac{r_{y6}(1-r_{58}^2) - r_{y8}r_{68}(1-r_{56}^2)}{r_{36}(1-r_{35}^2)(1-r_{68}^2)} - \frac{r_{y5}r_{35}}{1-r_{35}^2}$$

$$c = \frac{r_{y5}-r_{y3}r_{35}}{1-r_{35}^2} = \frac{r_{y1}-r_{y8}r_{13}}{r_{15}(1-r_{13}^2)} = \frac{r_{y5}-r_{y1}r_{15}}{1-r_{15}^2}$$

$$= \frac{r_{y1}}{r_{15}(1-r_{13}^2)} - \frac{r_{13}}{r_{15}r_{36}} \cdot \frac{r_{y6}(1-r_{38}^2) - r_{y8}r_{68}(1-r_{36}^2)}{(1-r_{13}^2)(1-r_{68}^2)}$$

$$= \frac{r_{y5}}{1-r_{35}^2} - \frac{r_{35}}{r_{36}} \cdot \frac{r_{y6}(1-r_{38}^2) - r_{y8}r_{68}(1-r_{36}^2)}{(1-r_{68}^2)(1-r_{35}^2)}$$

$$e = \frac{r_{y8}-r_{y3}r_{38}}{1-r_{38}^2} = \frac{r_{y6}-r_{y3}r_{36}}{r_{68}(1-r_{36}^2)} = \frac{r_{y6}}{r_{68}(1-r_{36}^2)} - \frac{r_{36}}{r_{13}r_{68}} \cdot \frac{r_{y1}(1-r_{35}^2) - r_{y5}r_{15}(1-r_{13}^2)}{(1-r_{15}^2)(1-r_{36}^2)}$$

$$= \frac{r_{y8}}{1-r_{38}^2} - \frac{r_{38}}{r_{13}} \cdot \frac{r_{y1}(1-r_{35}^2) - r_{y5}r_{15}(1-r_{13}^2)}{(1-r_{15}^2)(1-r_{38}^2)} = \frac{r_{y8}-r_{y6}r_{68}}{1-r_{68}^2}$$

$$f^2 = 1-a^2, \qquad g^2 = 1-br_{y3}-cr_{y5}-er_{y8}, \qquad h^2 = 1-dr_{68}$$

$a = 0.227$ $b = 0.393$ $c = 0.296$ $d = 0.326$ $e = 0.348$
$f = 0.974$ $g = 0.776$ $h = 0.950$

Because there are many different estimates for a path coefficient, one might compute the mean value of the estimates for each path coefficient. This method is applied to the estimates of coefficients b, c, and e. However, if denominator in a solution includes a zero correlation, or if the numerator has a zero correlation, the estimate will be erroneous, that is, no solution or substantial underestimate. To avoid these problems, one should calculate a ratio of the sum of all numerators and the sum

of all denominators in the solutions to obtain a combined estimate of each path coefficient (Kenney, 1979). This method is applied to the estimates of coefficients a and d.

All final estimates are shown at the bottom of Table 4. The actual model used for this example is shown in Figure 3.

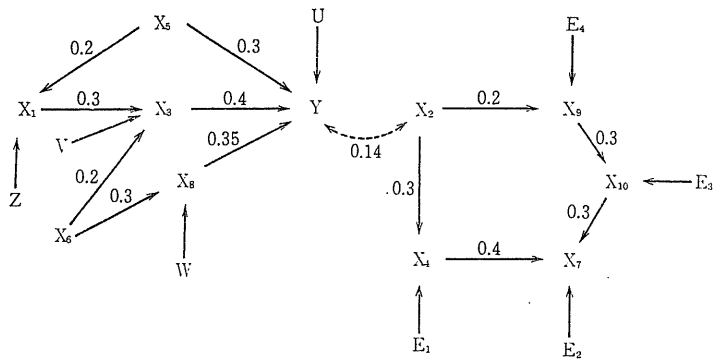

Fig. 3 Causal model presumed for example
* Broken line indicates insignificant correlation, and coefficients of disturbances are omitted

Compared with the model assumed initially, it seems that all of the paths in the obtained model are corresponding except two paths between X_1 and X_5, and X_3 and X_6, although their correlations can be accounted for the model. Because of these two uncertainties, we could not obtain precise values for a and d. However, this problem could be solved by our endeavor to find the most suitable directions of these paths as explained before, even though we did not try to do it here.

According to the example above, it seems that the algorithm proposed in this article is useful for building a causal model.

In order to explain these steps, the flow chart of the algorithm for causal modeling is shown in Figure 4.

3. Discussion

With respect to the algorithm proposed in this article, two basic assumptions are essential; one concerns the relationship between correlation and causal chain, and the other the direct effect of a variable on a specific variable. Although both of the assumptions appear to be reasonable to develop a causal model, it may be said that no attention is paid to the various mutual relationships among causal variables. However, it is possible to modify the discrepancies of correlations assumed by the model after the first modeling is finished as mentioned before. Alternatively, it may be thought that all possible combinations of variables should be examined, and that the most suitable model should be found. Even though one could try this method if the number of variables is comparatively small, it must be hard work to examine all correlations and decide on the existence and direction of each path. Nevertheless, this kind of method may be necessary for advanced modeling. Also, general problems of

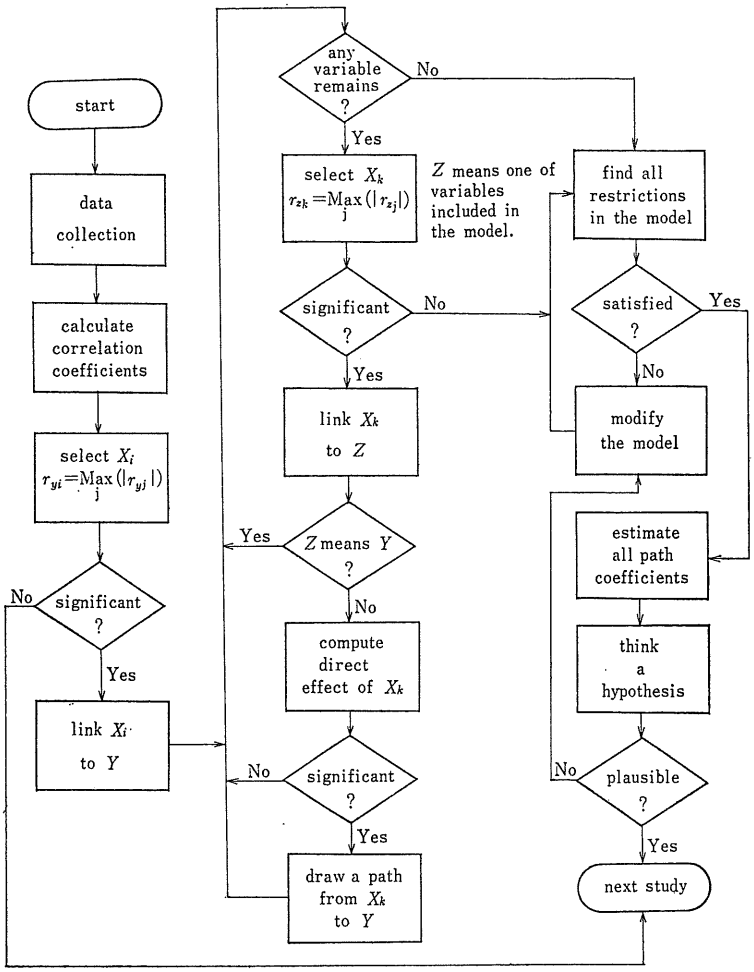

Fig. 4 Algorithm for causal modeling

causal analysis, like "time-lag", are not considered in this paper; the interested reader should consult the references (Blalock, 1964; Heise, 1975; Kenney, 1979; Cohen and Cohen, 1975).

Let us compare some methods related to causal analysis and the algorithm proposed here.

First of all, it is suggested that the algorithm described may have the same process as one of the hierarchical methods for cluster analysis, called the single linkage method (Sneath, 1957; Johnson, 1967). However, the only similarity is to link two variables together that have the highest correlation at each step. The main purpose of cluster analysis, of course, is to classify variables or individuals, but that of causal modeling is to build the causal structure concerning the effect like cancer death. Not only linking variables but also considering the direction of the influence of each variable are necessary for causal modeling. Therefore, as explained before, the procedure is essentially more complicated than the single linkage method. However,

it is true that cluster analysis could be also useful to detect the potential risk factors through classifying the variables. In fact, Burbank applied cluster analysis to various cancer mortality data in the United States, and could suggest some factors as causes of cancer (Burbank, 1972).

Figure 5 shows the result of cluster analysis applied to the data in Table 1. Even though five variables X_1, X_3, X_5, X_6, and X_8 related to Y could be selected, it seems to be difficult to decide the causal linkages.

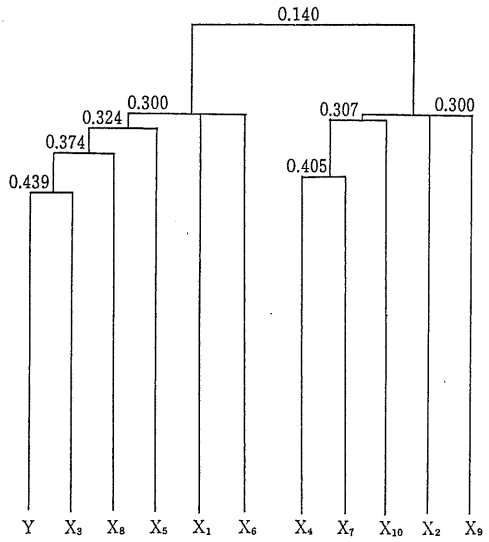

Fig. 5 Cluster analysis of eleven variables by means of single linkage method
* Numerical values indicate correlation coefficients.

Generally speaking, it is difficult to interpret the result of cluster analysis in terms of causality. Because the purpose is only to combine variables, it may be necessary to obtain many other information for the interpretation. On the other hand, the results of causal modeling like path analysis are straightfoward.

Factor analysis may also be a powerful way to detect risk factors. Some applications of factor analysis have been reported in the epidemiological field (Howell, 1974; Inaba et al., 1975; Yanai et al., 1978; Yanai et al., 1979). Even though factor analysis needs sophisticated techniques and sometimes intuition whenever one tries to explain the extracted factors, one could obtain significant suggestions. Therefore, the author compared the results of factor analysis applied to the data in Table 1 with those of causal modeling. The results of factor analysis are shown in Table 5 and Figure 6, which are calculated by means of the varimax method with iterations by the use of SPSS (Nie et al., 1975).

The results seem to obviously correspond to those of the causal model described before, although the expression of the results of factor analysis is different from the model presumed.

In Figure 6, one can realize that two factors cause Y; one is related with variables

Table 5
Factor loadings of eleven variables by means of Varimax method

Var.	F-1*	F-2	F-3
Y	0.655	0.475	0.070
X_3	0.541	0.147	0.022
X_1	0.433	−0.052	0.004
X_5	0.354	0.012	0.023
X_8	−0.011	0.771	0.023
X_6	0.085	0.375	0.008
X_7	−0.028	−0.018	0.667
X_4	0.019	0.009	0.509
X_2	0.105	0.068	0.357
X_{10}	−0.019	−0.012	0.354
X_9	0.014	0.007	0.296

* F stands for Factor.

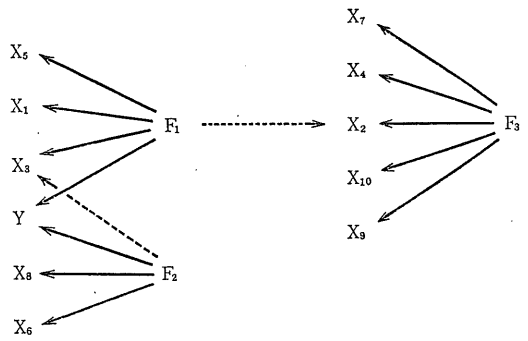

Fig. 6 Factor analysis model
* F_1, F_2, and F_3 are common factors, and unique factors are omitted. Broken lines indicate low factor loadings.

X_1, X_3, X_5, and Y, and the other with variables X_6, X_8, and Y. Besides, the factor related with variables X_2, X_4, X_7, X_9, and X_{10} is not connected with Y. Accordingly, one can select variables as causes of Y by the use of factor analysis as well as by the algorithm proposed here, and also could understand that there are two different routes that influence Y.

It appears to be indicated that factor analysis can help us to detect risk factors and confirm the estimated causal model. As Howell has stated (Howell, 1974) factor analysis could prove to be a useful and powerful epidemiological tool, and also Yanai et al. described its utility in detail from an epidemiological point of view (Yanai et al., 1978).

Therefore, if possible, we might also apply factor analysis to the detection of causal relations.

In general, it is said that the method of path analysis is related with multiple regression analysis to a great extent. As for multiple regression analysis, some researchers have applied it to detecting and evaluating risk factors as causes of cancer mortality, and could obtain some significant results (Armstrong and Doll, 1975; Hems, 1970; Hogan et al., 1979; Yanai et al., 1979). Accordingly, multiple regression

analysis should be clearly thought of as one of the powerful epidemiological tools. For trial, the author applied this method to the data in Table 1 by the use of SPSS (Nie et al., 1975).

Whereas variables X_3, X_5, and X_8 could be selected by a stepwise procedure and their partial regression coefficients could be estimated exactly, variable X_1 and X_6 could not be selected because of their insignificant coefficients.

In the case of multiple regression analysis, despite the fact that direct effects of independent variables on a dependent variable can be precisely evaluated, variables that are only indirect cause of the effect through other variables can never be taken into account theoretically. However, indirect effects have important roles for considering causal linkages. Therefore, it is not enough to depend only on multiple regression analysis with respect to the evaluations of variables as causes of a specific disease.

As a matter of fact, the algorithm described in this article may have to be thought of as one of a set of epidemiological tools to build a causal model, and by using other methods like cluster analysis, factor analysis, and/or multiple regression analysis, one could confirm and elucidate the estimated causal model.

4. Acknowledgement

The author would like to express his deep appreciation to Dr. Michael D. Hogan and Dr. Beth Gladen for their suggestions and stimulating discussions, and also would like to thank Dr. David G. Hoel for his inviting him to study on this subject.

References

Armstrong, B. and Doll, R. (1975) Environmental factors and cancer incidence and mortality in different countries, with special reference to dietary practices, *Int. J. Cancer*, **15**: 617–631.

Blalock, H.M. Jr. (1964) Causal inferences in nonexperimental research, the University of North Carolina, Chapel Hill, North Carolina

Burbank, F. (1972) A sequential space-time cluster analysis of cancer mortality in the United States: Etiological implications, *Am. J. Epidemiol.*, **95**: 393–417.

Cohen, J. and Cohen, P. (1975) *Applied multiple regression/Correlation analysis for behavioral sciences*, John Wiley & Sons, New York.

Heise, D.R. (1975) *Causal analysis*, John Wiley & Sons, New York.

Hems, G. (1970) Epidemiological characteristics of breast cancer in middle and late age, *Br. J. Cancer*, **24**: 226–234.

Hogan, M.D., Chi, P., Hoel, D.G., and Mitchell, T.J. (1979) Association between chloroform levels in finished drinking water supplies and various site-specific cancer mortality rates, *J. Environ. Pathol. Toxicol.*, **2**: 873–887.

Howell, M.A. (1974) Factor analysis of international cancer mortality data and per capita food consumption, *Br. J. Cancer*, **29**: 328–336.

Inaba, Y., Takagi, H., Yanai, H., and Matsubara, J.K. (1975) Ekigaku chosa ni okeru chiiki oyobi nenrei inshi ni kansuru kenkyu (A study on the geographical and age factors in the epidemiological survey), *Jap. J. Public Health*, **22**: 83–90 (in Japanese)

Johnson, S.C. (1967) Hierarchical clustering schemes, *Psychometrika*, **32**: 241–254.

Kenney, D.A. (1979) *Correlation and causality*, John Wiley & Sons, New York

Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K., and Bent, D.H. (1975) *Statistical package for the social sciences*. 2nd ed., McGraw-Hill, New York.

Sneath, P.H.A. (1957) The application of computers to taxonomy, *J. Gen. Microbiol.*, **17**: 201–226.

Yanai, H., Inaba, Y., Takagi, H., Toyokawa, H., and Yamamoto, S. (1978) An epidemiological study on mortality rates of various cancer sites during 1958 to 1971 by means of factor analysis, *Behaviormetrika*, **5**: 55–74.

Yanai, H., Inaba, Y., Takagi, H., and Yamamoto, S. (1979) Multivariate analysis of cancer mortalities for selected sites in 24 countries in the world, *Environmental Health Perspectives*. **32**: pp. 83–101.

(Received June, 1980)