International Journal of Epidemiology
Vol. 26, No. 1

Printed in Great Britain

The Role of Conceptual Frameworks in Epidemiological Analysis: A Hierarchical Approach

CESAR G VICTORA,* SHARON R HUTTLY,† SANDRA C FUCHS* AND MARIA TERESA A OLINTO*

Victora C G (Mestrado em Epidemiologia, Universidade Federal de Pelotas, CP 464, 96100 Pelotas, RS, Brasil) Huttly S R, Fuchs S C and Olinto M T A. The role of conceptual frameworks in epidemiological analysis: A hierarchical approach. *International Journal of Epidemiology* 1997; **26:** 224–227.
Background. This paper discusses appropriate strategies for multivariate data analysis in epidemiological studies.
Methods. In studies where determinants of disease are sought, it is suggested that the complex hierarchical inter-relationships between these determinants are best managed through the use of conceptual frameworks. Failure to take these aspects into consideration is common in the epidemiological literature and leads to underestimation of the effects of distal determinants.
Results. An example of this analytical approach, which is not based purely on statistical associations, is given for assessing determinants of mortality due to diarrhoea in children.
Conclusions. Conceptual frameworks provide guidance for the use of multivariate techniques and aid the interpretation of their results in the light of social and biological knowledge.
Keywords: conceptual frameworks, diarrhoea mortality, disease determinants, multivariate analysis

Multivariate analytical methods are increasingly popular in epidemiological studies. The present paper discusses appropriate data analysis strategies for assessing disease determinants and, in particular, how to handle the complex hierarchical inter-relationships between these variables.

At an early stage in the planning of an epidemiological study one needs to make a decision about the methods of data analysis. These are likely to include the description of all variables, an examination of the association between each risk factor and the outcome (often termed univariate analysis) and the study of the inter-relationships between the different risk factors. Multivariate models are also likely to be used in most studies. The latter are required due to the limitations of more traditional methods such as stratification for the adequate study of multiple risk factors.

Regarding the multivariate analyses, decisions must be taken on how to consider the association between the outcome under study and a range of explanatory factors. Does one 'just throw all the explanatory factors in the model and see what turns up as statistically significant' or 'construct a more selective model based on which factors one wishes to adjust for'? The choice of approach depends on the purpose of the study and it is important to take the correct decision as their results have different interpretations.

Multivariate analysis may be used for the prediction of an outcome, for example the construction of an algorithm for predicting which children with diarrhoea are likely to become dehydrated, based on the presence of early signs or symptoms. This approach has been often used for the prediction of obstetric risk given maternal characteristics, such as social class, previous pregnancy outcomes, pre-pregnancy weight, and height.[1] In this case a multivariate model may be appropriately constructed solely by selecting statistically significant explanatory factors through techniques such as stepwise logistic regression. Such an approach is based entirely on statistical associations rather than any conceptual basis for the inter-relationships between the factors. All explanatory variables are treated as if belonging to the same hierarchical level. A similar approach may also be useful for exploratory analyses when the aetiology of an outcome variable is not well known and information is available on a large number of explanatory factors. Even in this situation, it is preferable to balance reliance on statistically significant

* Mestrado em Epidemiologia, Universidade Federal de Pelotas, CP 464, 96100 Pelotas, RS, Brasil.
† Department of Epidemiology and Population Sciences, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK.

associations with some degree of biological or social interpretation.

More frequently, however, multivariate analysis is required for evaluating determination, i.e. the effect of a postulated risk factor on an outcome. One needs to know what this effect is after controlling for confounding factors. One may also wish to assess whether such an effect is direct or mediated through other factors. In this second approach, a different analytical strategy is required. The choice of factors to be included in the multivariate model is not based purely on statistical associations as is the case for prediction. A decision on which factors to include in the model should be based on a conceptual framework describing the hierarchical relationships between risk factors. Although the forthcoming examples are derived from the field of child health in less developed countries, the general principles also apply to a number of other health problems both in developed and less developed countries.

Child health, particularly in less developed countries, is determined by a large number of factors.[2,3] Ultimately, most ill health in such societies may be ascribed to poverty resulting from the lack of resources or, more frequently, to their unfair distribution both between and within countries. For assessing levels of poverty or wealth, most epidemiologists in this field use variables such as family income, parental education or the number and type of household appliances. Such factors, however, rarely cause ill-health directly and henceforth are referred to as distal determinants. These factors are most likely to act through a number of inter-related proximate determinants, sometimes referred to as intermediate variables or mechanisms. These proximate determinants may be subdivided into groups which are inter-related in a hierarchical or in a parallel way. This notion of proximate and distal determinants has been used in the fertility literature but is uncommon—at least in explicit terms—in epidemiology. In fact, it is not unusual to find examples in the epidemiological literature in which distal factors are improperly adjusted for proximate factors, with a consequent reduction or elimination of the former's effects.

Figure 1 shows a simplified scheme of this conceptual framework, in which variables near the top of the Figure influence those below them. Socioeconomic factors (the distal determinants) may affect, directly or indirectly, all other groups of risk factors with the exception of sex and age. These may include environmental factors (such as crowding or availability of water and sanitation) and maternal reproductive factors (such as age at childbirth, birth intervals and parity), among others. These variables, in turn, may affect the child's birthweight as well as its present anthropometric

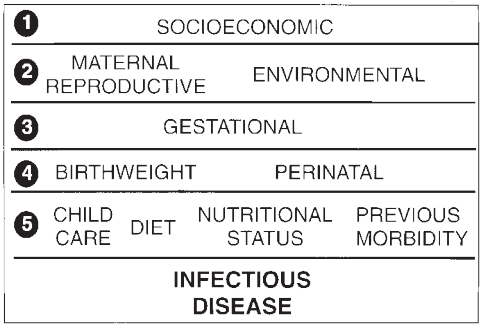

FIGURE 1 *Conceptual hierarchical framework of risk factors for infectious diseases in developing countries*

status and type of diet. They may also affect child care, including the use of health services. Finally, all of the above factors may affect the risk of the child acquiring an infectious disease as well as its severity, and therefore the risk of mortality.

Building a conceptual framework requires knowledge about the social and biological determinants of disease. Temporal considerations are also relevant. For example, gestational factors may determine birthweight which in turn affects postnatal morbidity and growth. Conceptual frameworks should be developed in the early stage of a study since these will influence aspects of study design (such as matching) and the variables to be measured. The objective of the present discussion, however, is not to develop comprehensive frameworks of disease causation but to examine the implication of such modelling for epidemiological analysis of determinants.

As an example let us consider three explanatory variables (family income, sanitary conditions and malnutrition) with diarrhoea mortality as the outcome. Based on Figure 1, suppose that the distal determinant family income affects diarrhoea partly through the proximate determinants poor sanitary conditions and malnutrition. Let one also assume that poor sanitation is itself a determinant of malnutrition due to past infections. The inter-relationships between these four variables are shown in Figure 2. The arrows represent the causal effect of the relevant explanatory factor. Family income exerts its effect on diarrhoea mortality through poor sanitation (a), malnutrition (b) and through other proximate determinants (c). Sanitation exerts its effect through malnutrition (d) and through other proximate determinants (e). In this simplified example, malnutrition exerts its effect on diarrhoea mortality directly (f). In considering the appropriate models for examining the effect of these three explanatory factors, it is useful to keep in mind the three

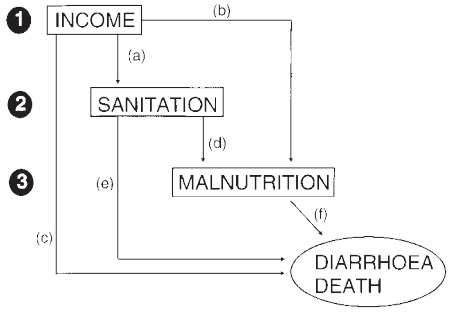

FIGURE 2 *Simplified conceptual hierarchical framework for diarrhoea mortality*

conditions for a factor to be a confounder. It should be associated with the exposure under study as well as being predictive of disease occurrence. Also, it should not be a mediating factor, that is, a link in the causal chain leading from the postulated risk factor to the outcome.[4]

The steps for the multivariate analyses and the interpretation of their results are summarized in Table 1.

The overall effect of income should be assessed in model 1. This model excludes sanitation and malnutrition since neither would qualify as confounders because both are partly determined by income and therefore represent mediating factors.

In the second step of the analysis (model 2), the sanitation variable would be added and its effect assessed in the presence of income which would then constitute a proper confounding factor. The unconfounded effect of sanitation would thus be obtained from this equation. The magnitude of the remaining effect of income in model 2 would only reflect that part which is not mediated through sanitation.

Extending the above model, malnutrition would be entered as the third-level variable (model 3) and its effect assessed in the presence of both confounding variables, income and sanitation. Any residual effect of income would be that part which is not mediated through either sanitation or malnutrition (arrow [c] in Figure 2). Likewise, any residual effect of sanitation would be outside the malnutrition pathway (arrow [e]). An important note of interpretation is that some if not most of the effect of income will be captured by the other two factors. It would be incorrect to interpret—as often done under similar conditions—that income has no effect after adjustment for 'confounding' variables since in model 3 the overall effect of income will be underestimated due to the presence of mediating factors.

This approach may be extended to situations with several variables in each hierarchical level. For example, model 1 could include other socioeconomic characteristics such as parental education or occupation. Their measures of effect are assessed in this first model. All of these variables could be kept in the subsequent models, or only a subset of them, such as those reaching a certain P level (say, $P < 0.1$). A decision to select will depend on the number of variables being considered. Studies with two or three variables at this level may keep all of them, while those with many variables may consider dropping out those not reaching certain criteria in order to avoid an excessive number of parameters and unstable estimates in subsequent models. The second model would add to the socioeconomic block (model 1) environmental variables such as water supply, sanitation and crowding. Again, either all or some of these would be retained for further analyses, and so forth for the subsequent levels.

For an illustration of the application of conceptual hierarchical frameworks to multivariate analyses, the readers are referred to two recent case-control studies of childhood pneumonia.[5,6]

Even in studies of the effect of a single exposure on a disease, conceptual frameworks will be useful for identifying variables that are hierarchically below the exposure and therefore do not qualify as potential confounding factors.

TABLE 1 *Summary of steps in the analysis of the effect of income, sanitation and malnutrition on diarrhoea mortality*

Model	Equation (explanatory variables)	Interpretation
1	Income	Overall effect of income; not adjusted for mediating variables.
2	Income + sanitation	Effect of sanitation (properly) adjusted for confounding role of income. Effect of income represents that not mediated through sanitation.
3	Income + sanitation + malnutrition	Effect of malnutrition (properly) adjusted for confounding roles of income and sanitation. Effect of sanitation represents that not mediated through malnutrition. Effect of income represents that not mediated through sanitation nor malnutrition.

Although for many chronic diseases, particularly in industrialized countries, the webs of causation do not show such clear hierarchical structures as for infectious or nutritional problems in less developed countries, a similar approach to data analysis would often be useful. For example, in many societies cigarette smoking is partly determined by socioeconomic status, and therefore the overall impact of social factors on smoking-related diseases should not be assessed through models which include smoking variables.

We have used conceptual hierarchical frameworks for studying the determinants of childhood infectious diseases, malnutrition, low birthweight, infant mortality, hypertension and obesity. These frameworks have provided guidance for using multivariate techniques and interpreting their results in the light of social and biological knowledge.

ACKNOWLEDGEMENTS

We thank colleagues and students who encouraged us to publish this paper. Sharon Huttly was funded by the Overseas Development Administration (UK) and Cesar Victora is supported by the National Council for Scientific and Technological Research (CNPq; Brazil).

REFERENCES

[1] Chamberlain R, Phillip E, Howlett B, Claireaux A. *British Births 1970 Vol. 2 Obstetric Care.* London: Heinemann, 1978.

[2] Mosley W H, Chen L C. An analytical framework for the study of child survival in developing countries. In: Mosley W H, Chen L C. Child survival: strategies for research. *Pop Devel Rev* 1984; **10 (Suppl.):** 25–48.

[3] UNICEF. *Strategy for Improved Nutrition of Children and Women in Developing Countries. A UNICEF Policy Review.* New York: UNICEF, 1990.

[4] Rothman K E. *Modern Epidemiology.* Boston: Little, Brown & Co., 1986, pp. 90–91.

[5] Victora C G, Fuchs S C, Flores A, Fonseca W, Kirkwood B. Pneumonia among Brazilian children: a hierarchical analysis. *Pediatr* 1994; **93:** 977–85.

[6] Fonseca W, Kirkwood B R, Victora C G, Fuchs S C, Flores J A, Misago C. Risk factors for childhood pneumonia among the urban poor in Fortaleza, Brazil: a case-control study. *Bull World Health Organ* 1996; **74:** 199–208

(Revised version received April 1996)