Modeling hierarchical relationships in epidemiological studies: a Bayesian networks approach

Nguefack-Tsague, Georges and Zucchini, Walter

University of Yaoundé I, Department of Public Health, Yaoundé, Cameroon, University of Goettingen, Institute for Statistics and Econometrics, Goettingen, Germany

19 January 2011

Modeling hierarchical relationships in epidemiological studies: a Bayesian networks approach

Georges Nguefack-Tsague[1,2]
Department of Public Health
Faculty of Medicine and Biomedical Sciences
University of Yaounde I
Yaoundé, Cameroon

and
Walter Zucchini
Institute for Statistics and Econometrics
Georg-August-University
Goettingen, Germany

January 19, 2011

Abstract

Hierarchical relationships between risk factors are seldom taken into account in epidemiological studies though some authors stressed the importance of doing so, and proposed a conceptual framework in which each level of the hierarchy is modeled separately. The objective of this paper was to implement a simple version of their framework, and to propose an alternative procedure based on a Bayesian Network (BN). These approaches were illustrated in modeling the risk of diarrhea infection for 2740 children aged 0 to 59 months in Cameroon. The authors implemented a (naïve) logistic regression, a step-level logistic regression and also a BN. While the first approach is inadequate, the two others approaches both account for the hierarchical structure but to different estimates and interpretations. BN implementation showed that a child in a family in the poorest group has respectively 89%, 40% and 18% probabilities of having poor sanitation, being malnourished and having diarrhea. An advantage of the latter approach is that it enables one to determine the probability that a risk factor (and/or the outcome) is in a given state, given the states of the others. Although the BN considered here is very simple, the method can deal with more complicated models.

Key-Works: Bayesian networks; hierarchical model; diarrhea infection; disease determinants; logistic regression

[1] The authors thank DHS Data Archive (Macro International Inc), USA for providing us with the data
[2] Address for correspondence: Department of Public Health, Faculty of Medicine & Biomedical Sciences University of Yaoundé I, P. O. Box 1364 Yaounde, Cameroon, Phone:+237 220-701-24 Fax:+237 223-112-24 , Email: nguefacktsague@yahoo.fr

INTRODUCTION

Standard regression methods, including logistic regression and related methods (possibly with higher order interactions) that are commonly used in epidemiological studies do not take account of causal relationships (which are difficult to establish in epidemiology (29-38)) that are known to exist, or are assumed to exist, between the covariates. For example when modeling disease status using a logistic regression, potential causal relationships between the risk factors are not explicitly modeled. All risk factors are treated as being directly related to disease status; i.e. at the same level of association. The usual procedure is to apply tests of hypotheses, or some model selection criterion, to decide which risk factors should be retained in the model. Causal relationships between some of the risk factors may be already known, or may be regarded as plausible on biological grounds. If so, such information can be, and should be, incorporated in a hierarchical model describing the relationships between disease status and the associated risk factors (the meaning of "hierarchical" here is not to be taken in the sense of multilevel modeling (or mixed models) where individual patients are grouped say by hospital, hospitals are grouped by region, and regions are grouped by country etc.; or as in Meta analysis where patients are grouped by study). Among other things, explicitly taking into account of such relationships can help to reduce the ubiquitous problem of multicollinearity.

Hierarchical relationships can be represented by arranging variables in a tree-like structure called a directed acyclic graph (DAG). An example of a hierarchical model is given in Victora *et al.(*1). They consider the presence/absence of an infectious disease in developing countries as a function of several covariates arranged in a hierarchy with 5 levels. The first factor (level 1) is the socioeconomic status; at the level 2 there may be two explanatory variables, such as maternal reproductivity and environmental factors; at level 3 one may have gestational factors; at level 4 birth weight and perinatal factors; at level 5 child care, diet, nutritional status and previous morbidity factors. Factors at level i influence those at level i+1. Finally, all the above factors may affect the risk of a child of acquiring an infectious disease. By ignoring the hierarchical structure of the model, one places risk factors, irrespective of their level, in a single large model and then applies some model selection strategy to eliminate the "non-significant" factors and thereby to select the model that fits the data best in some predefined sense. Victora *et al.* (1) argue that such procedure is inadequate because it ignores the hierarchical structure of the variables. Instead they proposed fitting a separate model for each level of the hierarchy, namely five individual models. In order to estimate the effect of each risk factor it is necessary to make adjustments for the (possibly) confounding role that other risk factors might play in effecting the outcome variable. Others applications of hierarchical models are given in Victora *et al.*(2) and Fonseca *et al.*(3) for case-control studies.

Our aim here is to implement a simple version of Victora *et al.* (1) 's approach, a (naïve) logistic approach (including higher order interactions) and to propose an alternative unified approach, based on Bayesian networks (BN), that takes account of hierarchical structure among covariates. The approaches are illustrated using a relatively simple model for assessing the impact of three risk factors for diarrhea in a sample of 2740 children in Cameroon.

MATERIALS AND METHODS

Data and variables

Data for 8096 children aged 0 to 59 months were obtained from the 2004 Cameroon Demographic and Health Survey (DHS) (28). The outcome variable of interest here was "*Had diarrhea recently (Diarrhea)"*, which was coded as yes=1 and no=0. The three covariates considered, that are labeled *sanitation*, *malnutritio*n and *income* were determined as follows:

Sanitation "*Toilet facilities shared*" (yes=1 and no=0). Although "*insufficient protein and energy intake*" is the most common nutritional deficiency affecting the young population in developing countries (4, 5), the data for this was unavailable, and so we used the *stunting status* (low height-for-age) as a surrogate for malnutrition, coded as 1 if the child is stunted and 0 otherwise. The third covariate that, for convenience, we label *income*, is an indicator of socioeconomic status of households based on wealth index according to DHS methodology. The wealth index takes account household income, use of health services and health status; it is an indicator of the level of wealth that is consistent with expenditure and income measure. The observed values of the index were partitioned into quintile groups labeled one (poorest) to five (richest); in what follows the label *income* refers one of these five groups. Including only children with measurements on all variables reduced the available sample to 2740 children.

The statistical analyses were performed with R version 2.10.1 (7). Two-sided p-values less that 0.05 were considered significant.

Models

The logistic regression approach

All variables were included in a selection procedure. The AIC (6) selection criterion in a stepwise algorithm was used as variable selection method. Goodness of fit of the models was assessed using the residual deviance.

The approach of Victora et al.(1)

Chi-squared tests were used to assess the association between variables. Logistic regressions with *diarrhea infection* as response and *income*, *malnutrition* and *sanitation* as predictors were used at each level of the hierarchy.

Bayesian networks

A BN, also known as a Bayesian belief network or belief network, is a probabilistic graphical model tool for describing relationships in a wide variety of domains (25), including various applications in medicine. A medical researcher may develop a BN for diagnosing and for preventing stress fractures. Alternatively a BN could represent the probabilistic relationships between diseases and symptoms. For example, given the symptoms, the network can be used to compute the probabilities of the presence of various diseases: the *diagnosis problem* (10, 11). Nikovski (9) applies BNs to problems in medical diagnosis. Van der Gaag (12) developed methods for eliciting probabilities in a cancer diagnosis study. Lauritzen and Spiegelhalter (8) use BN to compute the probability of a patient having tuberculosis, lung cancer or bronchitis respectively based on different factors.

We suggest the use of BN as an alternative to Victora *et al.*(1)'s approach. It takes into account the hierarchical relationships among risk factors and disease. An advantage of the proposed approach is that it enables one to estimate the probability that a risk factor and/or the outcome (disease) are in certain states, given the states of the remaining items (risk factors or outcome) in the model.

A BN is a network of "variables" or "nodes" connected by directed links (displayed as arrows) with a probability function associated with each variable (13, 14). A variable does not have parent if no links are pointing towards it and has parent otherwise. For example, in the simple structure $A \rightarrow B \rightarrow C$, A has no parent, A is a parent of B and B is a parent of C. A variable can be either a discrete random variable with a finite number of states, or a continuous random variable (generally assumed to be normally distributed). The links between variables represent (causal) relationships. Associated with a discrete variable is a probability distribution over its states; for a continuous random variables a Gaussian distribution (with given mean and variance parameters) is used instead.

A *marginal probability table* (MPT) assigns probabilities to the states of variables which have no parents; *a conditional probability table* (CPT) assigns probabilities to the states of variables which have parents. If a variable with parents is discrete then each entry in its CPT contains a conditional probability for that variable being in a specific state, given a specific configuration of the states of its parents. If a variable is continuous, the CPT contains the (conditional) mean and variance parameters for each configuration of the states of its discrete parents and a regression coefficient for each continuous parent, for each configuration of the states of the discrete parents.

If the variable B is the only "cause" for variable A, the CPT for A is computed using Bayes' rule as

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where $P(A)$ is the probability of A and $P(B|A)$ is the probability of B given that A has occurred. If A has K variables ($B_1, B_2,\ldots,B_K$) as causes (parents), B is replaced in the above formula by $(B_1, B_2,\ldots,B_K)$; i.e. a CPT for A is given by $P(A| B_1, B_2,\ldots,B_K)$. The term "evidence" refers to the information available regarding the current state of some of the variables. E.g., if one already knows that a child is from a poor family then this constitutes evidence that affects the probability that the child develop an infectious disease. It may also affect the probability that the other variables are in given states, for example, that the child suffers from malnutrition. A single item of evidence can affect the entire network. Given evidence E, the CPT for A, given the parent B, is updated using the formula

$$P(A|E,B) = \frac{P(A|B)P(E|A,B)}{P(E|B)}$$

where the left-hand term, $P(A|E,B)$ is called the *posterior probability*, or the probability of A after considering the effect of the evidence E on B. The term $P(A|B)$ is called the *a-priori probability* of A given B alone. The term $P(E|A,B)$ is called the *likelihood* and gives the probability of the evidence assuming the realization of A and B. Finally, the term $P(E|B)$ is independent of A and can be regarded as a normalizing constant, or scaling factor. Details on the use of Bayes rules in BN can be found in Jensen *et al.*(15). Details about the philosophical reasoning and application of BN can be found in Jensen (13, 14). The analyses presented here were performed using Hugin Lite version 7.0.(16).

Figure 1 shows a simplified conceptual framework for modeling the diarrhea status of children in our application. It is assumed that *income*, *sanitation* and *malnutrition* are risk factors for diarrhea infection. The factor *income* is at the first level, *sanitation* and *malnutrition* are at the second level and *diarrhea* is at the last level; thus the model has three levels. Socioeconomic status (*income*) affects *diarrhea* through poor sanitation conditions and malnutrition, but possibly also through unobserved causes, such as lack of transport to access health services. That is why we have included an arrow from *income* to *diarrhea.* Poor sanitation conditions also affect diarrhea directly due to past infections and through malnutrition. Malnutrition is a direct cause of disease infections, i.e. a malnourished child is vulnerable to infections; e.g. diarrhea infection. Cochran-Mantel-Haenszel (CMH) tests have been used to test for conditional independence of variables (26-27).

RESULTS

There was a highly significant association between *income* and others variables: *sanitation* (χ^2=146.47, d.f.=4, P=0), *malnutrition* (χ^2 = 114.26, d.f.=4, P=0) and *diarrhea* (χ^2= 14.53, d.f.=4, P = 0.005).

A naïve approach is to ignore the hierarchical structure of the variables and regard the three covariates as belonging to the same single level instead of three levels. Then with *diarrhea* as

dependent variable, and assuming that each model includes an intercept, there are seven possible models. Using AIC selection criterion in a stepwise algorithm, the model Logistic Regression 3 was selected. The fit of the saturated model (including all higher order interactions) showed no significant interaction terms (not reported here).

However, if the hierarchical structure of the data is taken into account, only three models are meaningful (1). At the first level, *income* is the only predictor, at the second level, *income* and *sanitation* are the predictors and at the third level, *income*, *sanitation* and *malnutrition* are the predictors. Using the approach of Victora et al. (1), we propose to fit a logistic regression at each of the three levels. The usual interpretation does not hold. Logistic regression 1 measures the overall effect of *income* on Diarrhea infection. Logistic regression 2 measures the effects of *sanitation* on *diarrhea* adjusted for the confounder *income*. In this model the effect of *income* is mediated through *sanitation*. Logistic regression 3 measures the effects of *malnutrition* on *diarrhea* adjusted for the confounders *income* and *sanitatio*n. In this model the effect of *income* is that not mediated via *sanitation* or *malnutrition,* and the effect of *sanitation* is that not mediated via *malnutrition*. Thus, a fundamental issue in the three above models is interpretation.

Logistic regression 1 (Residual deviance=2338.9, d.f.=2738, *P*=1), logistic regression 2 (Residual deviance=2336.5, d.f.=2737, *P*=1) and logistic regression 3 (Residual deviance=2333.1, d.f.=2736, *P*=1) indicate that each of these 3 models fit the data quite well (Table 1). To avoid the problem of the choice of the reference level, we fit the model to evaluate the global effect of each variable.

In the BN, *income* is an independent risk factor for *diarrhea*, in addition to its effect via *sanitation* and *malnutrition*. Thus *income* is more likely to be a confounding variable for the relationship between *sanitation* and *malnutrition*. On the other hand, *sanitation* was highly is associated with *malnutrition* (χ^2= 10.72, d.f.=1, *P*<0.001) and is associated with *diarrhea* (χ^2= 4.96, d.f.=1, *P*= 0.026). Thus *sanitation* could be considered as an independent risk factor for *diarrhea;* its association with *malnutrition* also makes it more likely to be a confounding variable for the relationship between *malnutrition and diarrhea*. *Malnutrition* is associated with *diarrhea* (χ^2 = 5.60, d.f.=1, *P*= 0.018). These results therefore justify the conceptual framework in Figure 1 designed in Victora *et al.*(1).

The factors *sanitation* and *malnutrition* are not independent (χ^2= 10.72, d.f.=1, *P*<0.0011). However, the null hypothesis that they are conditionally independent, given *income* cannot be rejected (χ^2_{CMH}= 1.07, d.f.=1, *P*=0.27). In effect *sanitation* and *malnutrition,* given *income,* can be regarded as conditionally independent; the relationship between them can be explained purely by the fact that poor families are more likely to have both poor sanitation and malnourished children than are richer families. This suggests that we could delete the arrow from *sanitation* to *malnutrition*; but for the purposes of comparison we have not done this. Finally, *income* and *diarhea* are not conditionally independent given both *sanitation* and *malnutrition* (χ^2_{CMH} = 15.25, d.f.=4, *P*= 0.0042). This indicates that socioeconomic status (*income*) affects the probability of *diarrhea* in more ways than just via *sanitation* and *malnutrition*. It is plausible that it affects that probability via some other (possibly unobserved) factors, such as lack of transport to access health services. Given this lack of conditional independence we may not leave out the arrow between *income* and *diarhea.*

As is illustrated in the previous paragraph, conditional independence is a key notion in the construction of a BN. It is used to determine which arrows are essential in the network, and which can be omitted. There is no direct causal link between two variables that are conditionally independent (given a third variable) even if the two variables are highly correlated. The high correlation is a consequence of their "common ancestor" (the third variable) and not the result of any direct causal relation between them. It is important to note that in a BN, a missing edge represents a conditional independence.

Table 1: Logistic regressions approaches

	Coefficient	SE	z-value	P-value
Logistic Regression 1				
Intercept	-1.32	0.12	-10.86	0
Income	-0.14	0.04	-3.43	0
Logistic Regression 2				
Intercept	-1.34	0.12	-10.93	0
Income	-0.15	0.04	-3.66	0
Sanitation	0.18	0.08	2.23	0.01
Logistic Regression 3				
Intercept	1.45	0.14	-10.58	0
Income	-0.14	0.04	-3.29	0
Sanitation	0.19	0.09	2.68	0.01
Malnutrition	0.21	0.09	2.31	0.01

Table 2: CPT for Terminal node Diarrhea

Malnutrition	Sanitation	Income	Diarrhea	
			No	Yes
No	No	Poorest	0.83	0.17
		Poorer	0.90	0.10
		Middle	0.88	0.12
		Richer	0.84	0.16
		Richest	0.89	0.11
	Yes	Poorest	0.86	0.14
		Poorer	0.90	0.10
		Middle	0.83	0.17
		Richer	0.79	0.21
		Richest	0.86	0.14
Yes	No	Poorest	0.80	0.20
		Poorer	0.84	0.16
		Middle	0.79	0.21
		Richer	0.82	0.18
		Richest	0.94	0.06
	Yes	Poorest	0.81	0.19
		Poorer	0.80	0.20
		Middle	0.79	0.21
		Richer	0.88	0.12
		Richest	0.78	0.22

Table 3: CPT for Sanitation

Income	Sanitation	
	No	Yes
Poorest	0.89	0.11
Poorer	0.67	0.33
Middle	0.67	0.33
Richer	0.58	0.42
Richest	0.68	0.32

Table 4: MPT for Income

Income	Proportion
Poorest	0.20
Poorer	0.22
Middle	0.25
Richer	0.18
Richest	0.15

Table 5: CPT for Malnutrition

Sanitation	Income	Malnutrition	
		No	Yes
No	Poorest	0.60	0.40
	Poorer	0.88	0.12
	Middle	0.63	0.37
	Richer	0.62	0.38
	Richest	0.77	0.23
Yes	Poorest	0.63	0.37
	Poorer	0.87	0.13
	Middle	0.67	0.33
	Richer	0.70	0.30
	Richest	0.77	0.23

Table 6: Comparison frequencies (%) from data and the adjusted frequencies (BN)

Factors	True (data)	Adjusted (BN)
Sanitation		
Yes	31.24	30.07
No	68.76	69.93
Malnutrition		
Yes	30.62	29.25
No	69.38	70.75
Diarrhea		
Yes	15.36	14.97
No	84.64	85.03

Table 6 shows the marginal frequencies of the variables in the data and the adjusted ones. The latter take into account the hierarchical structure of the variables; in particular they adjust

automatically for any confounding effect. For example, the proportion of children with Diarrhea infection was 15.36%; after taken into account the hierarchical structure, this proportion reduced to 14.97%. Figure 2 shows the distribution of the risks/factors and the disease and takes into account the hierarchical structure. The marginal frequencies of the variables in the data are very close to those given in the BN in Figure 2. The main reason is that the "prior probabilities" (beliefs) given in Tables 2, 3, 4 and 5 were not determined *apriori*; they were estimated from the data. Note that the proportion of children living in a poor sanitary condition is 31.24%; but if one takes account of the hierarchical structure and the confounding role of *income*, this proportion is reduced to 30.07%. The proportion of malnourished children is 30.62%; after taking into account the hierarchical structure and the confounding role of *income* and *malnutrition*, this proportion reduces to 29.25%. Note that the proportion of *income* status doesn't change because this variable doesn't have parent. However, it may change when there is "evidence″ (i.e. knowledge) regarding the state of one or more of the other variables. In general, when there is evidence of the state of any variable, all the networks frequencies are likely to change, as illustrated in Figures 3 and 4. From Figure 3, evidence that a child's family falls in the poorest group leads to a 89.00% probability that he/she has poor sanitation, a 39.67% probability that he/she is malnourished and a 17.94% probability that he/she has diarrhea infection. Figure 4 shows that when there is evidence that the child's family belongs to a poorest group, has poor sanitation and is malnourished, there is 20% probability that he/she has diarrhea infection.

DISCUSSION

The chosen variables in DHS are obviously imperfect for characterizing income, malnutrition and sanitation (17). Furthermore as we have mentioned earlier, causation is in general very difficult to establish (29-38).

Our intention was not to develop a comprehensive model, but rather to illustrate the use of a BN to construct hierarchical structures when the covariates are known to be, or are assumed to be interdependent in certain ways. The network can be used to predict the state of the variable for any child when there is evidence of the true state of one or more of the other variables**.** BN is very useful for modeling situations where some information is already known and incoming data are uncertain or partially unavailable (18). DHS that are conducted at irregular time intervals are an example of this; BNs can help predicting risk factors/outcome while waiting for a new survey.

Probabilities in CPTs and MPTs (e.g. Tables 2-5) can be obtained from historical data (here DHS) or elicitation of probabilities from experts (e.g. epidemiologists). Objective survey data and subjective expert assessment can be used either separately or in combination with each other. Of course in the absence of any objective data, elicitation of reliable probabilities is the most difficult aspect in BN modeling. It is especially difficult when many risk factors are being investigated and these are related in complex ways (19-24). To alleviate the task, Kjaerulff (19) and van Engelen (20) propose the removal of arcs representing weak dependencies. A key advantage of BN is the facility of updating (or modifying) the network as new information becomes available. On the other hand a major criticism of BNs is the need to choose prior probabilities, and to choose a statistical distribution (when necessary).

The BN that we used for the purpose of illustration is a very simple one. There may be a multitude of others (unobserved, or not-mentioned nor included) confounding factors besides the four variables considered. However, BN are powerful tools that can be extended to model much more complex relationships.

We recommend the use of BNs in epidemiological studies whose aim is predicting or studying the determinants of diseases. The conceptual framework must be clearly set up in

order to identify potential hierarchical structure in the data. Automatic variables selection should be used only when such structure is not found in the data. A properly constructed BN automatically takes into account possible confounding variables. Failure to take into account hierarchical structure of covariates can result in models that lead to unclear, sometimes even misleading, interpretations, of the relationships under investigation.

REFERENCES

1. Victora CG, Huttly SR, Fuchs SC, et al. The role of conceptual frameworks in epidemiological analysis: a hierarchical approach. *Int J Epidemiol*. 1997;26:224-27.

2. Victora CG, Fuchs SC, Flores JA, et al. Pneumonia among Brazilian children: a hierarchical analysis. *Pediatr.* 1994;93:977-85.

3. Fonseca W, Kirkwood BR, Victora CG, et al. Risk factors for childhood pneumonia among the urban poor in Fortaleza, Brasil: a case-control study. *Bull World Health Organ.*1996;74:199-208.

4. de Onis M, Monteiro C, Akre J, et al. The worldwide magnitude of protein-energy malnutrition: an overview from the WHO Global Database on Child Growth. *Bull World Health Organ.* 1993;71:703-12.

5. Takyi EEK**.** Nutritional status and nutrient intake of preschool children in Northern Ghana. *East Afr Med J. 1999*;76: 510-15.

6. Akaike H. Information theory and an extension of the maximum likelihood principle. In: Petrov B, Csáki F (eds)**.** *Second International Symposium on Information Theory*. Budapest : Akadémiai Kiadó, 1973. 267-281.

7. R Development Core Team**.** *R: A language and environment for statistical computing.* Vienna : R Foundation for Statistical Computing, 2010.

8. Lauritzen SL, Spiegelhalter DJ. Local computations with probabilities on graphical structures and their applications to expert systems. In: *Proceedings of the Royal Statistical Society* 1988;50:154-227.

9. Nikovski D. Constructing bayesian networks for medical diagnosis from incomplete and partially correct statistics. *IEEE Transactions on Knowledge and Data Engineering*. 2000;12(4):509-16.

10. Long W**.** Medical diagnosis using a probabilistic causal network. *Applied Artificial Intelligence.* 1989;3:367-83.

11. Jensen FV, Skaanning C, Kjaerulff U. *The sacso system for troubleshooting of printing systems.* Technical report 2000, University of Aalborg, Denmark.

12. van der Gaag LC, Renooij S, Witteman C. Probabilities for a probabilistic network: a case study in oesophageal cancer. *Artificial Intelligence Med.* 2002;25(2):123-48.

13. Jensen FV**.** *An Introduction to Bayesian Networks.* New York, NY: Springer, 1996.

14. Jensen F*V. Bayesian networks and decision graphs.* New York, NY: Springer, 2001.

15. Jensen FV, Lauritzen SL, Olesen KG. Bayesian updating in causal probabilistic networks by local computations. *Computational Statistics Quarterly.* 1990;4**:**269-82.

16. Olesen KG, Lauritzen SL, Jensen FV. aHugin: A system creating adaptive causal probabilistic networks. In: Wellman M P, D'Ambrosio B, Smets P, Dubois D. (eds). *Proceedings of the Eighth Conference on Uncertainty in Artificial Intelligence*. Stanford, California, 1992. 223-29.

17. WHO Working Group. Use and interpretation of anthropometric indicators of nutritional status**.** *Bull World Health Organ.* 1986;64:929-41.

18. Charles River Analytics, Inc **.** *About Bayesian Belief Networks.* Cambridge : Charles River Analytics, Inc., 2004.

19. Kjaerulff U**.** Reduction of computational complexity in bayesian networks through removal of weak dependencies. In: *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence,* 1994.

20. van Engelen RA. Approximating bayesian networks by arc removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 1997;19:916-20.

21. Heckermann D, Breese JS. Causal independence for probability assessment and inference using bayesian networks. *IEEE Transactions on Systems, Man and Cybernetics.* 1996;26:826-31.

22. Henrion M**.** Some practical issues in constructing belief networks. *In Uncertainty in Artificial Intelligence 3.* 1989.

23. Pearl J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Mateo, CA : Morgan Kaufmann, 1988.

24. van der Gaag LC, Renooij S, Witteman CLM, et al. How to elicit many probabilities. In: *Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence*, 1999.

25. Adusei-Poku K, Van den Brink G, Zucchini W. Implementing a Bayesian network for foreign exchange settlement: a case study in operational risk management. *Journal of Operational Risk.* 2007 ; 2 :1-6

26. Agresti, A. *Categorical Data Analysis*. 2nd ed. New York, NY: Wiley, 2002.

27. Agresti, A. *An Introduction to Categorical Data Analysis*. 2nd ed. New York, NY: Wiley, 2007.

28. MEASURE DHS, ICF Macro, Calverton (USA) http://www.measuredhs.com (accessed February 17 2009)

29. Evans AS. Causation and disease: a chronological journey. The Thomas Parran Lecture. *Am J Epidemiol.*1978;108:249-258.

30. Hill, AB. The environment and disease: association or causation? *Proceedings Royal Society Medicine.* 1965;58:295-300.

31. Lave LB, Seskin EP. Epidemiology, causality and public policy. *American Scientist.* 1979; 67:178-180.

32. Rothman KJ, Greenland S. *Modern Epidemiology* 2nd ed, Philadelphia: Lippincott-Raven. 1998.

33. Schlesselman JJ. "Proof" of cause and effect in epidemiologic studies. *Preventive Medicine.* 1987; 16:195-210.

34. Susser M. Judgment and causal inference. Criteria in epidemiologic studies. *Am J Epidemiol.* 1973; 105:1-15.

35. Susser M. What is a cause and how do we know one? *Am J Epidemiol.*1991; 133:635-648.

36. Weed DL. On the logic of causal inference. *Am J Epidemiol.* 1986; 123:965-979

37. Weiss NS. Inferring causal relationships. *Am J Epidemiol.*1981; 113:487

Appendix

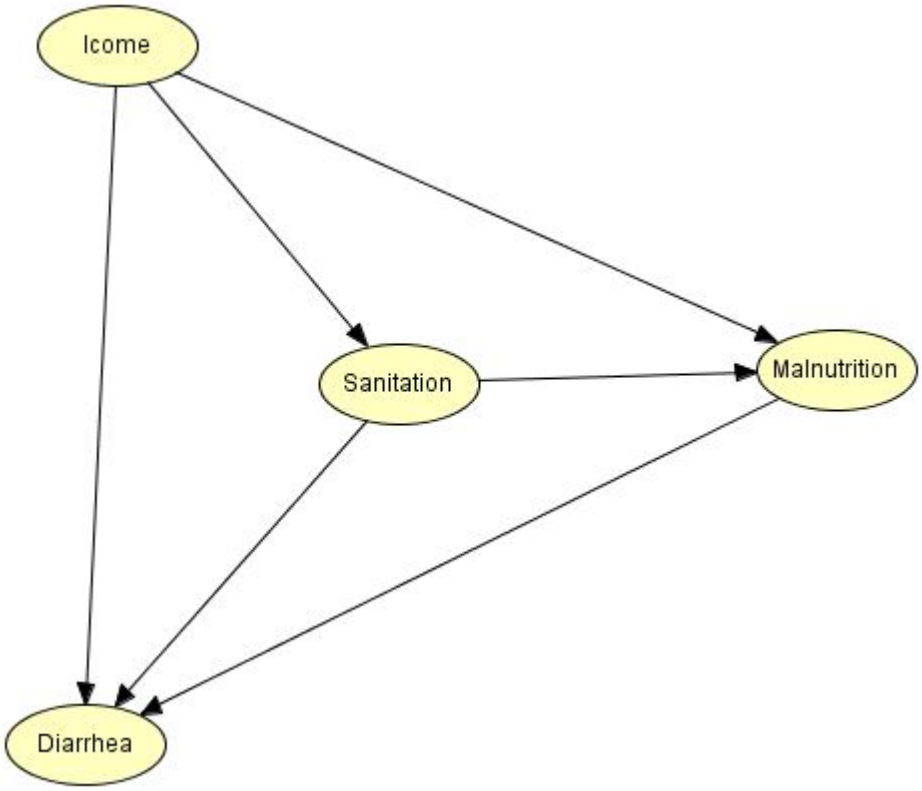

Figure 1: Bayesian network: a simplified conceptual hierarchical framework for Diarrhea

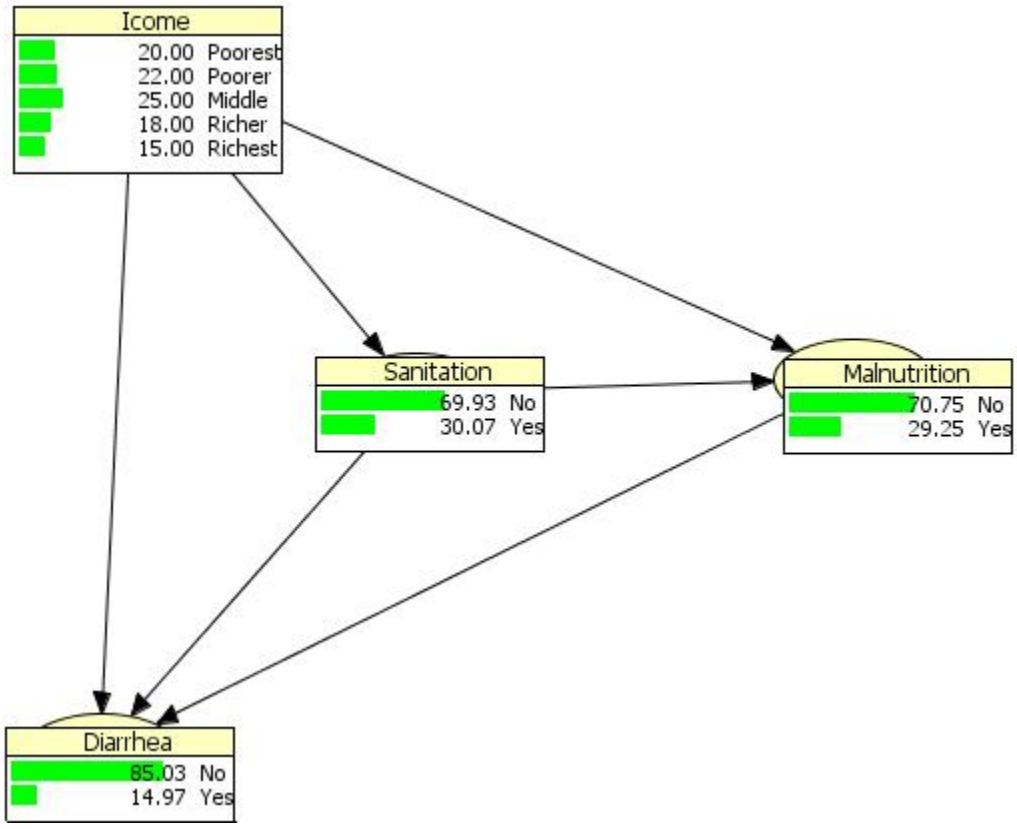

Figure 2: Frequency network showing posterior probabilities (%)

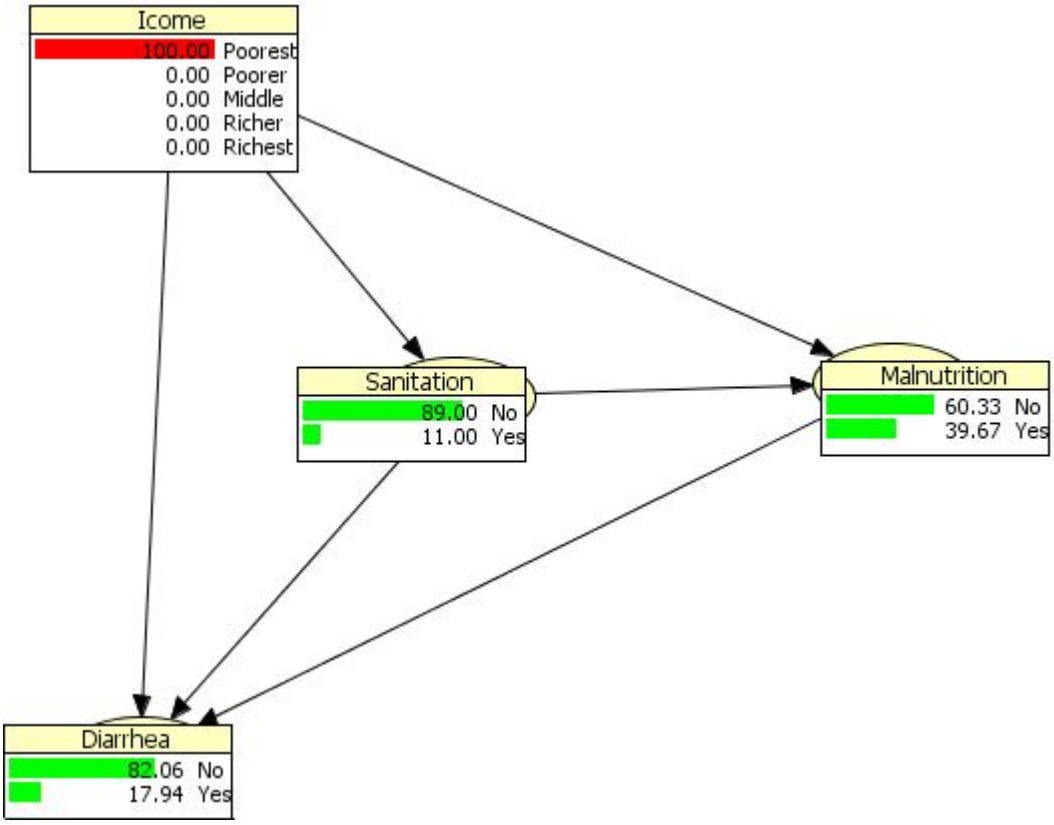

Figure 3: Frequency network showing posterior probabilities (%) when there is evidence that the child belongs to a poorest family.

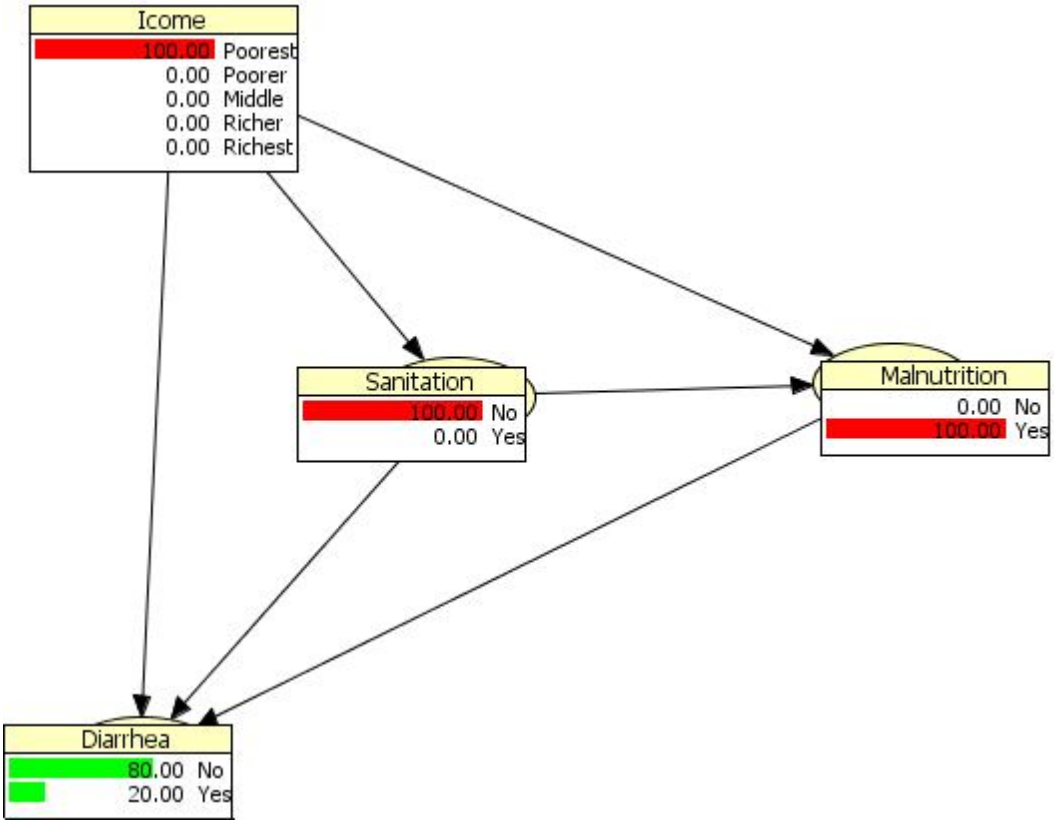

Figure 4: Frequency network showing posterior probabilities (%) of developing diarrhea when there is evidence that the child belongs to a poorest family, with poor sanitation condition and is malnourished.