
Pergamon

Research in Developmental Disabilities, Vol. 18, No. 5, pp. 303–318, 1997

0891-4222/97 $17.00 + .00

PII S0891-4222(97)00012-7

MENTOR: A Bayesian Model for Prediction of Mental Retardation in Newborns

Subramani Mani, Suzanne McDermott, and Marco Valtorta

University of South Carolina

Mental retardation (MR) is a diagnosis that is made with extreme caution because of the many uncertainties in its etiology and prognosis. In fact, most physicians will delay the diagnosis for months or years so that substantial evidence is available to rule the diagnosis in or out. MENTOR is a Bayesian Model for the prediction of MR in newborns that provides probabilities for the full range of cognitive outcomes, ranging from MR to superior intelligence. Using the model to confirm clinical judgment could help physicians decide when to proceed with diagnostic tests. The physician and family could discuss the probabilities for MR, borderline, normal, and superior intelligence, given the child's status in infancy and base their decision about additional testing, in part, on this information. © 1997 Elsevier Science Ltd

INTRODUCTION

Mental Retardation (MR) is a complex medical and social problem with an estimated prevalence between 1 and 3% in all human populations (Batshaw, 1993; Stein & Susser, 1992). It is a developmental disability with a complex etiology. Many of the causative factors and mechanisms are not well understood and the actual causes are usually unknown for 30–50% of individuals with the condition (Batshaw, 1993). Mental retardation is characterized by significantly subaverage intellectual functioning (American Psychiatric Association, 1994).

Subramani Mani is presently at the Department of Information and Computer Science, University of California at Irvine, Irvine CA 92697.

Requests for reprints should be sent to Suzanne McDermott, Department of Family and Preventive Medicine, University of South Carolina, 6 Richland Medical Park, Columbia, SC 29203.

The *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition (*DSM-IV*) states:

> Significant subaverage intellectual functioning is defined as an IQ of about 70 or below (approximately 2 standard deviations (SD) below the mean)...obtained by assessment with one or more standardized, individually administered intelligence tests. Four degrees of severity can be specified, reflecting the level of intellectual impairment: Mild Mental Retardation (IQ level 50–55 to approximately 70), Moderate Retardation (IQ level 35–40 to 50–55), Severe Mental Retardation (IQ level 20–25 to 35–40) and Profound Mental Retardation (IQ level below 20–25). (APA,1994)

Model building is helpful when we need to simplify a complex problem in order to make predictions or select from competing choices. The most widely applied use of model building in medicine has been related to differential diagnosis. Complex signs and symptoms are entered into the models, probabilities are assigned, interactions are defined and probable diagnoses are provided as outputs. In the case of mental retardation, we are confronted with a situation where there are many unknown causes and uncertain relationships. This paper describes a Bayesian model that assigns probabilities based on prenatal and birth conditions and a limited number of postpartum events to predict intelligence groupings (MR, borderline, normal intelligence, superior intelligence).

The most recent models for MR, described in the literature, are either conceptual, designed to select interventions for individuals with MR, or predictive for population based rates of mental retardation. Claire (1989) and Greenspan and Gransfield (1992) discuss conceptual models that contribute to the definition of mental retardation but are not predictive. MR-Expert (Hile, Campbell, Ghobary, & Desrochers, 1993) is a rule-based expert system to support decisions related to violent behaviors displayed by some individuals with MR. McDermott and Altekruse (1994) developed a dynamic model that predicted population prevalence rates of MR based on demographic factors and child health policy decisions. The model demonstrates how socioeconomic variables, especially poverty and deprivation, increase the risk for MR in a population. In addition, McDermott has developed a linear regression model to explain variation in school district rates of MR (McDermott, 1994).

BAYESIAN NETWORKS

Bayesian statistics provide an alternative to hypothesis testing and confidence interval estimation. Bayesian statistical inference is used to draw conclusions from known data in a sample to populations for which there are no data. Bayesian statistical inference uses probabilities for both prior and future events to estimate the uncertainty that is inevitable with prediction.

Bayesian Networks are also referred to as Causal Probabilistic Networks (Cooper, 1984; Lauritzen & Spiegelhalter, 1988; Neapolitan, 1990; Pearl, 1988) and Bayesian Expert Systems. The fundamental concept in these networks is

that probabilities can be assigned to parameter values, and through Bayes' theorem, these probabilities can be updated given new data. In Bayesian models the parameter is viewed as a domain variable, with a probability distribution, since the actual value of the parameter is unknown. The causal links between the variables are represented by arrows in the model. The model is strong if the arrows can be interpreted as causal mechanisms. The relationships can, alternatively, be considered an association and this type of model would be viewed more cautiously (Glymour & Spirtes, 1993; Neapolitan, 1990; Pearl, 1988). Variables without parents are referred to as root variables. Prior probabilities are specified for the root variables and conditional probabilities for the variables with parents. A Bayesian Network allows us to reason in two directions. For example, we can ask the question: If the baby has low birthweight, what is the likelihood of a normal or mental retardation outcome? And, we can ask: If the infant has mental retardation what is the likelihood of low or normal birthweight? Calculation of these values is straightforward from the initialized values using Bayes' theorem. For complicated Bayesian Networks exact inference algorithms are available to propagate evidence across the network (Neapolitan, 1990; Pearl, 1988). Outcomes of mediating events can be predicted since they use the causal relations represented in the directed graph. These models are superior to statistical regression models as they take into account the causal sequence of events (Glymour & Spirtes, 1993).

There are many Bayesian Expert Systems in the medical arena- Munin (Andreassen, Woldbye, Falck, & Andersen, 1987), ACORN (Wyatt & Spiegelhalter, 1989), Expert Systems for hematologic diagnosis (Nguyen, Diamond, Piolet, & Sultan, 1992), Diagnostica (Blinowska, Chattellier, Wojtasik, & Bernier, 1993) and PATHFINDER (Heckerman, Horvitz, & Nathwani, 1992). Expert System shells have been designed to facilitate easy construction of Bayesian Network applications. HUGIN (Anderson, Olesen, Jensen, & Jensen, 1989), IDEAL (Srinivas & Breese, 1990) and BAIES (Cowell, 1992) are three such shells.

There are two methods for building a Bayesian Expert System (BES). The first is asking a domain expert (in our case, a specialist in MR) to construct the network and assign the initial or prior probabilities. The second method involves building the network from data using Bayesian Network generating algorithms, such as — BIFROST (Lauritzen, Thiesson, & Spiegelhalter, 1993), K2 (Cooper & Herskovits, 1992) and CB (Singh & Valtorta, 1995). The data-built models can be validated by comparing the data generated model with the performance of an expert (Spiegelhalter, Dawid, Lauritzen, & Cowell, 1993). In this paper, we report on the use of our model built by a combination of the two strategies. We capture the skeleton network from data using the CB algorithm and prune the model with the help of a MR expert and published literature. In other words, the network structure as well as the prior and conditional probabilities are obtained from data and fine-tuned by an individual with knowledge of the literature who can provide expert opinion about the inclusion and logical

sequence of variables. The details of the model building are explained in another paper (Mani, Valtorta, & McDermott, 1996).

METHODS

The Child Health and Development Studies (CHDS) dataset was selected in order to build a Mental Retardation prediction model. The CHDS was a prospective study of pregnant mothers and their children. The children were followed through their teen years using numerous questionnaires, physical and psychological examinations, and special tests. The study, conducted by the University of California at Berkeley and the Kaiser Foundation, started in 1959 and continued into the 1980s. There are approximately 6000 children and 3000 mothers with IQ scores in the dataset. The children were either 5-year-olds or 9-year-olds when their IQs were tested (Child Health and Development Study, 1987) .

Data used in this analysis were derived from the CHDS interviews of the mother during pregnancy and from the mother's and child's Kaiser medical charts. Information on cognitive functioning were available from the special developmental examinations. Tests of cognitive functioning were given to two subgroups of the CHDS participates. At their 5th birthday, 3,413 children were given developmental examinations and follow-up interviews. Likewise, 3,737 children were examined at their 9th, 10th, or 11th birthday. This examination also included cognitive ability tests for the child and the mother.

Two tests of cognitive function were available for this analysis since both tests have been used to identify and classify MR. These tests are the Raven Progressive Matrices Test (for children) and the Peabody Picture Vocabulary Test (for children and adults). Although these tests are often used in conjunction with the Stanford-Binet or the Wechsler scales, they were the only tests of cognitive function administered in the Child Health and Development Studies. Thus, we used the Raven Test for children as the predicted outcome measure for this paper. Raven's Progressive Matrices, originally introduced in 1938, is a nonverbal test of reasoning ability that measures the ability to form comparisons, to reason by analogy, and to organize spatial perceptions into systematically related wholes (Sattler, 1990). The Raven Progressive Matrices are considered a test applicable for children from 5–11 years of age and it has reported test-retest reliability scores ranging from .71 to .93 (Raven, 1965). Raven scores were grouped into four categories: MR (<29), borderline (30–39), normal (40–60), and superior (>60). The standardized scores were coded into the dataset. They had a mean of 50 with a standard deviation of 10.

The Peabody Picture Vocabulary Test (PPVT) was originally developed in 1959. This nonverbal, multiple choice test was designed to evaluate the hearing vocabulary or receptive knowledge of children and adults. Since maternal scores were coded as raw scores in the dataset, we standardized them using the 3000 mothers in the study. The mean was 125 with a standard deviation of 19. We

categorized the maternal scores on the Peabody test as mild MR (62–86), borderline (67–105), normal (106–144), and superior (145–200).

We initially identified about 50 variables that are thought to play a role in the causal mechanisms of MR. Variables with weak associations to the Raven scores were eliminated and the variables used in the model are defined in Table 1.

MODEL BUILDING

Three datasets were created for final model building. In all the three datasets only the first child of the mother is included, if this case had no missing IQ score variables. If the IQ scores were missing the second child was selected. The first dataset (RAVEN 1) contains 2212 cases and 24 variables and it was used to validate the model. The IQ scores of mothers and children are present. The proportions of controls and cases with three cutpoints for inclusion of cases (Risk Threshold levels) are presented in Table 2A. It is important to note that using a lower threshold than the Resting Values (more cases being identified as MR) would improve the sensitivity (predictive accuracy of cases) while lowering the specificity (prediction on controls). There are no missing values for the IQ scores, however, for the other variables, 4% had missing values. The second dataset (RAVEN 2) contains 5985 cases and 23 variables. As only about 3000 mothers were given IQ tests, this dataset was created without the maternal IQ score. The percentage of missing values, for other variables, was 10%. The third dataset (RAVEN 3) contains 5985 cases and 24 variables, however, the majority of the IQ scores of mothers are missing. The percentage of missing values, for other variables, was 12%. In all of the datasets the missing values were the result of incomplete or missing data during data collection, and the pattern of missing data can be assumed to be random.

The CB algorithm was run on the datasets for generating the networks. The datasets were randomly partitioned into two — a major part and a minor part. The bigger partition was used to construct the network and the smaller set for validation. For RAVEN 1, we used the first 2000 cases to generate the network and for the other two, the first 5000. We defined three rules to characterize the inadequacies of the generated networks.

1. Rule of Chronology: Events occurring later in time cannot be the parents of earlier incidents. For example, a child health problem cannot be the parent of maternal disease.
2. Rule of Commonsense: The causal links of the network should not go against commonsense. For example, Father's education cannot be a cause of Mother's race.
3. Domain Rule: The causal links should not violate established domain rules. For example, Prenatal care cannot cause Maternal smoking.

TABLE 1
Variable Names and Definitions from the Child Health and Development Study Dataset

Variable name	Variable Definition
Maternal race	Gravida race has been classified as White (European, or White and American Indian or others considered to be of white stock) and non-White (Mexican, Black, Oriental, interracial mixture, South-East Asians).
Maternal age	Age at the time of birth categorized into: 14–19 years; 20–34 years; ≥35 years.
Marital	Marital status has been grouped as ever (married/legally separated/divorced/widowed) and never married.
Maternal Education/ Paternal Education	Mother's and father's educational status have been defined as less than high school (if they had education ≤12th grade and did not graduate), high school (if high school graduate), >high school (attended college or college graduate) and special school (trade school).
Maternal disease	if gravida had any one or more of the listed conditions — lung trouble, heart trouble, high blood pressure, kidney trouble, convulsions, diabetes, thyroid trouble, anemia, tumors, bacterial disease, measles, chicken pox, herpes simplex, eclampsia, placenta previa, any type of epilepsy or malnutrition; coded as having a condition otherwise, coded as not having a condition. This variable coded by using an OR gate over 11 variables.
Income	Family income has been categorized into <$10,000 and ≥$10,000.
Smoking	Maternal smoking was coded as yes if mother was smoking during that pregnancy and no otherwise.
Alcohol	A mother is defined as a mild drinker if she takes 0–6 drinks per week; moderate drinker if 7–20 per week and severe if >20 drinks per week.
Stillbirth	A history of one or more of previous stillbirths form the risk group and no history of previous stillbirths form the referent group.
Prenatal	Women who had prenatal care form the referent group and those who did not have prenatal care form the risk group.
X-ray	If a woman had been x-rayed for any reason in the year prior to or during the current pregnancy they were grouped as the risk level. Others were categorized as the referent level. This variable has been coded by using an OR gate over 2 variables.
Gestation	Period of gestation categorized into premature (≤258 days), postmature (>294 days) and normal term (259–294 days).
Distress	Fetal distress was coded if there had been prolapse of cord or the mother had a history of uterine surgery or uterine rupture or fever at or just before delivery or an abnormal fetal heart rate. Those children who had none of the above were grouped as referent level. This variable has been coded by using an OR gate over 5 variables.
Induce	If the woman had any type of induction of labor (stripping of membranes, artificial rupture of membranes, Oxytocin IV/IM, induced but method not stated or if injected fluid in uterus) she is categorized under risk level and the referent level is no induction.
Caesarean	If the type of delivery was caesarean section, it was grouped as risk level and if the type of delivery was vaginal, it was categorized as referent level.
Gender	Male gender of infant during the current pregnancy forms the risk level and the female gender forms the referent level.
Birthweight	Low birth weight babies are those weighing <2500 g and the normal weight babies are those weighing ≥2500 g.
Resuscitation	If the child had any type of resuscitation, he or she formed the risk level and children who had no resuscitation were categorized as the referent level.

Variable name	Variable Definition
Head circumference	A child with head circumference measurement of either 20 or 21 in was categorized as referent and the rest were grouped under risk level.
Anomaly	A child with any of the following conditions forms the risk level (cerebral palsy, hypothyroidism, spina bifida, Down's syndrome, chromosomal abnormality, anencephaly, hydrocephalus, epilepsy, Turner's syndrome, cerebellar ataxia, speech defect, Klinefelter's syndrome, or any type of convulsions). Children with no conditions form the referent level. This variable has been coded by using an OR gate over 13 variables.
Health Problem	Has been grouped into four — A child having physical problems, behavioral problems, both physical and behavioral problems and no problems.
Raven	Raven scores are grouped into four categories — mild, borderline, normal, and superior. These scores have been standardized to a mean of 50 with a standard deviation of 10.
Peabody	Mother's scores on the Peabody test categorized as mild (<86), borderline (87–105), normal (106–144) and superior (>145–). Mothers' scores yielded a mean of 125 with a standard deviation of 19.

The skeleton structure of the net was modified using the Rules. Then the network was refined by comparing an expert's opinion about the likelihood of risk for each of the events and characteristics presented in the models. The expert was a clinician who has 20 years of experience with children with MR and other developmental disabilities. The expert was asked to use her experience with individual cases, and knowledge of the literature in the field, to assign a probability for each variable in the model. The expert had extensive experience in research and was familiar with the concepts of risk and probability. When the expert stated there was no relationship between variables, the causal links were removed and new ones were incorporated to capture the knowledge of the domain causal mechanisms. We used the data to generate the prior and conditional probabilities for all of the variables and modified only those the expert felt were inadequate. Prior and conditional probabilities were calculated from the dataset, RAVEN 3. The first 5000 cases were used for generating the network and computing the probabilities. Table 2 lists the variables, their levels, and the prior probabilities from the dataset.

The expert refined network with 23 variables was input in Hugin using the Hugin graphic interface. Directed causal links were drawn from the parent variable to its children. The various levels (states) of the variables were also entered. For variables without parents, prior probabilities of the various states calculated from the RAVEN 3 dataset were assigned. For the variables with one or more parents, the conditional probabilities calculated using the same dataset were assigned. The network was ready at this point and the prior probabilities of the different states of all the variables were read from the network monitor.

RESULTS

MENTOR is a model for risk prediction of mental retardation. We used a prior probability of 5.6% for MR and 12.4% for borderline MR since these were

TABLE 2
Variable States and Prior Probabilities, CHDS

Variable No.	Variable Name	Variable State	Prior Probability	Variable State	Prior Probability
1	Maternal race	1. Non-white	0.33	2. White	0.67
2	Maternal age	1. 14–19	0.07	2. 20–34	0.77
		3. ≥35	0.16		
3	Marital	1. Never marrried	0.01	2. Married	0.99
4	Maternal education	1. ≤12 years	0.15	2. HS Grad	0.33
		3. College	0.51	4. Special school	0.01
5	Paternal education	1. ≤12 years	0.16	2. HS Grad	0.27
		3. College	0.56	4. Special school	0.01
6	Maternal disease	1. No disease	0.33	2. One or more disease	0.67
7	Income	1. ≥10,000	0.14	2. <10,000	0.86
8	Smoking	1. No	0.68	2. Yes	0.32
9	Alcohol	1. Mild	0.93	2. Moderate	0.06
		3. Severe	0.01		
10	Stillbirth	1. None	0.97	2. Yes	0.03
11	Prenatal	1. Yes	0.99	2. None	0.01
12	X-ray	1. No	0.74	2. Yes	0.26
13	Gestation	1. Full-term	0.80	2. Premature	0.08
		3. Postmature	0.12		
14	Distress	1. No	0.91	2. Yes	0.09
15	Induce	1. No	0.96	2. Yes	0.04
16	Caesarean	1. No	0.96	2. Yes	0.04
17	Gender	1. Female	0.50	2. Male	0.50
18	Birthweight	1. Normal	0.92	2. Low birthweight	0.08
19	Resuscitation	1. No	0.93	2. Yes	0.07
20	Head circumference	1. Normal	0.94	2. Abnormal	0.06
21	Anomaly	1. No	0.99	2. Yes	0.01
22	Health problem	1. None	0.75	2. Physical	0.09
		3. Emotional	0.09	4. Both	0.07
23	Raven	1. Mild	0.02	2. Borderline	0.15
		3. Normal	0.70	4. Superior	0.13
24	Peabody	1. Mild	0.01	2. Borderline	0.12
		3. Normal	0.78	4. Superior	0.09

Table 2A
Percent of Actual Controls and Cases Identified as Having Mental Retardation by MENTOR Using the CHDS

Risk Threshold for MR and Borderline MR	CHDS Controls (n = 1863)	CHDS Cases (n = 349)
Resting value (0.18)	434(23%)	122(35%)
1.5 × resting value (0.27)	370(20%)	111(32%)
2.0 × resting value (0.36)	342(18%)	101(29%)

indicative of the cognitive functioning scores for children in the CHDS. We used the RAVEN 1 dataset with 24 variables to validate the model. Table 2A gives the proportions of cases and controls identified with three threshold values. Using a lower threshold than the Resting Value would improve the sensitivity (predictive accuracy for cases) while lowering the specificity (predictive accuracy for controls). If the risk of both MR and borderline MR doubles, we get a combined probability of 35%. That leaves a probability of 65% for normal and superior functioning. In fact, most of the actual cases with MR in the dataset have more than a 50% probability for normal outcome when we run them through the model. This situation is because there are more children with normal outcomes with similar instantiations of variables than there are children with MR. In actual clinical cases, the diagnosis of MR is rarely made after a review of history and physical examination. The clinical observation leads to a suspicion that is followed by a psychological examination. Thus, we cannot expect MENTOR to do more than estimate the likelihood of outcomes. MENTOR would confirm a clinician's intuition by assigning probabilities to the cognitive functioning levels. Since expert estimation of risk is subjective and based on prior experience, we decided on a strategy of validation by comparing with the expert as the initial step.

We generated nine cases with instantiation for a subset of variables to be sure the model was working effectively. The information from the known variables for each of these cases is shown in Tables 3–5. We ran these cases on the model and came up with the probabilities for each of the Raven score groups.

In the first case, the maternal race was non-White, the mother's age at birth was in the range of 14–19 years, both the mother and the father had less than 12 years of education, the family income was less than $10,000, the gestation was full-term although the baby was low birthweight and the mother's Peabody intelligence test was normal. For this infant the probability of having MR increased from the prior probability of 5.6% to 10.1%, a notable increase. In addition, the probability of having borderline MR increased from 12.4% to 30.0%. To compensate for the increased probability of an unfavorable outcome, the probability for normal and superior intelligence decreased from 73.1% to 55.9% and from 8.9% to 4.0%.

The expert was then asked to score the results, as agree or disagree with the probabilities for each of the nine cases in Tables 3–5. The expert was in agreement with the model's assessment in seven of the nine cases. The two cases where the expert was not in complete agreement with the probabilities were Case #4 and Case #9. However, in both these cases there were health problems in the child; in Case #4 the child had a congenital anomaly and in Case #9 the child had a health problem. In both of these cases a review of the medical chart would indicate the exact nature of the problem and this information would be used to estimate the probabilities. It is possible that the designated probabilities correctly estimate the risk. However, the domain expert could not assign probabilities without this additional information.

TABLE 3
Risk Assessment of MR in Cases #1, #2 AND #3

Variable No.	Variable Name	Case #1 Variable State	Case #2 Variable State	Case #3 Variable State
1	Maternal Race	Non-white	White	White
2	Maternal Age	14–19		≥35
4	Maternal education	≤12 years	College	≤12 years
5	Paternal education	≤12 years	College	HS Grad
6	Maternal disease		No disease	
7	Income	<$10,000		<$10,000
8	Smoking			Yes
9	Alcohol			Moderate
10	Stillbirth			
11	Prenatal		Yes	
12	X-ray			Yes
13	Gestation	Full-term	Full-term	Premature
14	Distress		No	Yes
15	Induce			
16	Caesarean			
17	Gender			
18	Birthweight	Low	Normal	Low
19	Resuscitation			
20	Head circumference			Abnormal
21	Anomaly		No	
22	Health problem			Both
23				
24	Peabody	Normal	Superior	Borderline

TABLE 3A

Variable No.	Variable Name	States and Prior Probabilities		Case 1 Posterior Probabilities	Case 2 Posterior Probabilities	Case 3 Posterior Probabilities
23	Raven	MR	5.6%	10.1%	1.0%	20.0%
23	Raven	Borderline	12.4%	30.0%	4.0%	40.0%
23	Raven	Normal	73.1%	55.9%	69.0%	38.0%
23	Raven	Superior	8.9%	4.0%	26.0%	2.0%

Following these initial results, we generated validation results using the CHDS and the NCCP datasets.

Validation of the Model Using a Different Dataset

A second dataset was used to validate the results. The National Collaborative Perinatal Project (NCPP), of the National Institute of Neurological and Communicative Disorders and Stroke, is a large longitudinal study of pregnancy outcomes. The data consist of 55,043 pregnancies, between 1959 and 1974, and 8 years of follow-up for live-born children. All the cases in the dataset were run

TABLE 4
Risk Assessment of MR in Cases #5, #6 and #7

Variable No.	Variable Name	Case #4 Variable State	Case #5 Variable State	Case #6 Variable State
1	Maternal race	Non-white	White	White
2	Maternal age		20–34	20–34
4	Maternal education	College	Special school	HS Grad
5	Paternal education	HS grad	College	College
6	Maternal disease			
7	Income	<$10,000		
8	Smoking			
9	Alcohol		Moderate	Mild
10	Stillbirth	Yes		
11	Prenatal			Yes
12	X-ray	Yes		
13	Gestation	Full-term	Premature	Full-term
14	Distress		Yes	
15	Induce			
16	Caesarean		Yes	
17	Gender	Male	Female	
18	Birthweight	Normal		Normal
19	Resuscitation			
20	Head circumference			
21	Anomaly	Yes		
22	Health problem			Both
23				
24	Peabody	Superior	Borderline	

TABLE 4A.

Variable No.	Variable Name	States and Prior Probabilities		Case 4 Posterior Probabilities	Case 5 Posterior Probabilities	Case 6 Posterior Probabilities
23	Raven	MR	5.6%	15.0%	16.7%	3.2%
23	Raven	Borderline	12.4%	15.0%	25.1%	9.3%
23	Raven	Normal	73.1%	66.0%	54.8%	77.5%
23	Raven	Superior	8.9%	4.0%	3.4%	10.0%

through the model and the results are shown in Table 6A. In the initialized state (when only prior probabilities are applied), the network assigned a risk of 0.18 for the combination of mild and borderline mental retardation. If we take twice the initialized risk as our threshold for significant risk, our threshold can be set at a value of 0.36. Using this threshold, we find that 31% of the cases are correctly predicted. Fourteen percent of the controls were classified as being at significant risk for MR or borderline MR. Table 6B gives a sensitivity of 41% and a specificity of 80% using the threshold at the initialized state. A lower threshold will trade specificity for higher sensitivity.

TABLE 5
Risk Assessment of MR in Cases #7, #8 and #9

Variable No.	Variable Name	Case 7 Variable State	Case 8 Variable State	Case 9 Variable State
1	Maternal race	Non-white	White	
2	Maternal age	≥35		
4	Maternal education	HS Grad	HS Grad	HS Grad
5	Paternal education	≤12 years	College	HS Grad
6	Maternal disease		Yes	
7	Income	<$10,000		<$10,000
8	Smoking	Yes		
9	Alcohol		Moderate	
10	Stillbirth			
11	Prenatal			Yes
12	X-ray	Yes	Yes	
13	Gestation	Full-term	Premature	Postmature
14	Distress			
15	Induce		Yes	
16	Caesarean			
17	Gender	Male		
18	Birthweight	Low		Normal
19	Resuscitation			
20	Head circumference		Abnormal	
21	Anomaly			
22	Health problem		Physical	Both
23				
24	Peabody		Normal	

TABLE 5A

Variable No.	Variable Name	States and Prior Probabilities		Case 7 Posterior Probabilities	Case 8 Posterior Probabilities	Case 9 Posterior Probabilities
23	Raven	MR	5.6%	16.4%	7.1%	12.7%
23	Raven	Borderline	12.4%	33.9%	13.2%	25.0%
23	Raven	Normal	73.1%	46.9%	72.8%	58.0%
23	Raven	Superior	8.9%	2.8%	7.0%	4.3%

DISCUSSION

The MENTOR model can be used to illustrate the complex nature of MR prediction. Available evidence can be used to generate the probabilities and by instantiating the variables to different levels, we can assess the impact on the outcome. The model can be used to confirm clinical intuition or to formulate a prevention strategy for a defined population. We can also use the model for reverse inference. By instantiating the outcome to MR, we can see how other variables are affected and identify the variables that are most amenable to prevention.

The results from the nine cases assigned probabilities that were in general agreement with the expert. A modification of coding for congenital anomalies

TABLE 6A
Mean Risk for Mental Retardation Predicted for Children with Normal Cognitive Functioning at Age 8 Years (Controls) and Children with Borderline or Mild Mental Retardation Functioning at Age 8 Years (Cases) by MENTOR using the National Collaborative Perinatal Project

Cognitive Functioning Level	Controls (n = 13019)	Cases (n = 3598)
Mild MR	0.06	0.09
Borderline MR	0.12	0.16
Mild and borderline	0.18	0.25

TABLE 6B
Percent of Actual Controls and Cases Identified as having Mental Retardation, by MENTOR Using the National Collaborative Perinatal Project

Risk Threshold for MR and Borderline MR	Controls (n = 13019)	Cases (n = 3598)
Resting value	2260(20%)	1487(41%)
1.5 X	2410(19%)	1378(38%)
2.0 X	1836(14%)	1107(31%)

and child health problems could result in even better prediction. A weakness in the design of MENTOR was the use of old datasets to build the model. The results obtained from children born between 1959–74 may not be generalizable to infants born today. Significant changes in the care of infants have occurred since that time and it is likely that these changes effect the outcome of mental retardation. Further limiting generalizability is that fact that the Kaiser population excludes the unemployed and uninsured, and this subpopulation is at greater risk of mental retardation. Likewise, the NCPP oversampled African American women and other disadvantaged groups. Nonetheless, the Child Health and Development Studies and the National Collaborative Perinatal Project are the best U.S. data to analyze the relationship between pregnancy events and mental retardation in childhood because of the richness of detail in the prenatal, perinatal, infant, and childhood periods. However, both datasets were recording births during the 1960s and 1970s. Given the issue of generalizability, replication of these analyses in a more recent cohort of children is indicated.

Another weakness of the MENTOR model was that the networks generated from the different datasets had some variable links that violated the rule of chronology. A facility for inputting the chronological order can be incorporated. Likewise, if some rules could be incorporated in the network generation stage

to take care of domain-specific constraints, we can avoid connections violating each of the Domain Rules.

Our model can be improved by using more recent datasets and by addition of new variables or editing the directed causal links. The available data made it impossible to improve the accuracy further. In fact, the use of the Raven test for measurement of the outcome was a serious limitation of the dataset. A more standard measure of cognitive function, such as the Stanford-Binet or the WISC, in a dataset would greatly improve the model. In fact, a well-designed prospective study could identify new variables that might play a role in the causal pathway of MR.

The concept of a Causal Probabilistic Network and the mechanisms available for gathering evidence and propagating it through a network is a powerful scheme to make sense of real-world data. However, because of the many unknown relationships in the field of MR, an exact causal network may remain elusive (Heckerman et al., 1992).

Our experience with this work tells us that the raw network generated using a Bayesian algorithm (CB algorithm) had limitations. However, we feel that our strategy of generating the network from data using an algorithm and then improving it under the guidance of a domain expert yields a better model. Since mental retardation is a diagnosis that is difficult, MENTOR could be used by physicians to quantify their clinical judgment. The probability estimates generated by MENTOR could assist a physician in making the decision about which children to refer for more extensive work-ups. The probabilities could also be used in counseling families since it is often important to emphasize the possibility for mental retardation, borderline, normal, and superior intelligence outcomes for any child. Families might find it reassuring to hear that, although the risk of MR is increased, there is still a large probability for normal or superior intelligence.

REFERENCES

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). (1994). Washington, DC: American Psychiatric Association.

Andersen, S. K., Olesen, K. G., Jensen, F. V., & Jensen, F. (1989). HUGIN — A shell for building Bayesian belief universes for expert systems. *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence,* Detroit, 1080–1085.

Andreassen, S., Woldbye, M., Falck, B., & Andersen, S. K. (1987). MUNIN — A causal probabilistic network for interpretation of electromyographic findings. *Proceedings of the Tenth International Joint Conference on Artificial intelligence* (IJCAI-87), Milan, 366–372.

Batshaw, M. L. (1993). Mental retardation. In M. L. Batshaw (Ed.), *Pediatric clinics of North America — The child with developmental disabilities* (pp. 507–522). Philadelphia, PA: W. B. Saunders Company.

Blinowska, A., Chattellier, G., Wojtasik, A., & Bernier, J. (1993). Diagnostica — A Bayesian decision-aid system — Applied to hypertension diagnosis. *IEEE Transactions on Bio-Medical Engineering,* **40**, 230–236.

Child Health and Development Studies. (1987). *Data archive and users' manual of the child health and development studies.* Berkeley, CA: University of California, School of Public Health.

Claire, L. St. (1989). A multidimensional model of mental retardation: Impairment, subnormal behavior, role failures, and socially constructed retardation. *American Journal on Mental Retardation*, **94**, 88–96.

Cooper, G. F. (1984). *NESTOR: A computer based medical diagnostic aid that integrates causal and probabilistic knowledge*. (Tech. Rep. No. HHP-84-48). Stanford, CA: Stanford University, Medical Computer Science Group.

Cooper, G. F., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, **9**, 309–347.

Cowell, R. G. (1992). BAIES: A probabilistic expert system shell with qualitative and quantitative learning. In J. M. Bernardo, J. O. Berger, A. P. David, & A. F. M. Smith (Eds.), *Bayesian statistics 4* (pp. 595–600). Oxford: Clarendon Press.

Glymour, C., & Spirtes, P. (1993). Comment: Conditional independence and causal inference. *Statistical Science*, **8**, 250–255.

Greenspan, S., & Granfield, J. M. (1992). Reconsidering the construct of mental retardation: Implications of a model of social competence. *American Journal on Mental Retardation*, **96**, 442–453.

Hile, M. G., Campbell, D. M., Ghobary, B. B., & Desrochers, M. N. (1993). Development of knowledge bases and the reliability of a decision support system for behavioral treatment consultation for persons with mental retardation: The Mental Retardation-Expert. *Behavior Research Methods, Instruments, and Computers*, **25**, 195–198.

Heckerman, D. E., Horvitz, E. J., & Nathwani, B. N. (1992). Toward normative expert systems: Part I The Pathfinder Project. *Methods of Information in Medicine*, **31**, 90–105.

Lauritzen, S. L., & Spiegelhalter, D. J. (1988). Local computation with probabilities on graphical structures and their applications to expert systems. *Journal of the Royal Statistical Society*, **50**, 157–224.

Lauritzen, S. L., Thiesson, B., & Spiegelhalter, D. (1993). *Diagnostic systems created by model selection methods — A case study*. Preliminary papers of the Fourth International Workshop on Artificial Intelligence and Statistics, Ft. Lauderdale, FL, January 3–6, pp. 93–105.

Mani, S., Valtorta, M., & McDermott, S. (1996). *Building Bayesian models in medicine: A practical study*. Unpublished manuscript.

McDermott, S. (1994). An explanatory model to describe school district prevalence of mental retardation and learning disabilities. *American Journal of Mental Retardation*, **99**, 175–185.

McDermott, S. W., & Altekruse, J. M. (1994). Dynamic model for preventing mental retardation in the population: The importance of poverty and deprivation. *Research in Developmental Disabilities*, **15**, 49–65.

Nguyen D. T., Diamond, L. W., Priolet, G., & Sultan, C. (1992). Expert system design in hematology diagnosis. *Methods of Information in Medicine*, **31**, 82–89.

Neapolitan, R. E. (1990). *Probabilistic reasoning in expert systems: Theory and algorithms*. New York: John Wiley and Sons.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufmann.

Raven, J. C. (1965). *The Coloured Progressive Matrices Test*. London: Lewis Press.

Sattler, J. M. (1990). *Assessment of children* (3rd ed.). San Diego: Author.

Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L., & Cowell, R. G. (1993). Bayesian analysis in expert systems. *Statistical Science*, **8**, 219–283.

Singh, M., & Valtorta, M. (1995). Construction of Bayesian belief networks from data: A brief survey and an efficient algorithm. *International Journal of Approximate Reasoning*, **12**, 111–131.

Srinivas S., & Breese, J. (1990). IDEAL: A software package for the analysis of influence diagrams. In L. N. Kanal, J. Lemmer, & T. S. Levitt (Eds.), *Uncertainty in artificial intelligence* (pp. 212–219). North Holland: Elsevier Science.

Stein, Z. A., & Susser, M. W. (1992). Mental retardation. In J. M. Last & R. B. Wallace (Eds.), *Public health and preventive medicine* (13th ed.?2?). San Mateo, CA: Appleton & Lange.

Wyatt, J., & Spiegelhalter, D. (1992). The evaluation of medical expert systems. In D. A. Evans & V. L. Patel (Eds.), *Advanced models of cognition for medical training and practice. New York: Springer-Verlag.*