MENTOR: Building Bayesian Network Models in Medicine
CSCE Technical Report TR-2002-016

Marco Valtorta*
Department of Computer Science and Engineering
University of South Carolina
Columbia, SC 29208

Subramani Mani
Center for Biomedical Informatics
University of Pittsburgh
Pittsburgh, PA 15213

Suzanne McDermott
Department of Family and Preventive Medicine
University of South Carolina
Columbia, SC 29208

October 3, 2002

Abstract

An experiment in Bayesian model building from a large medical dataset for Mental Retardation is discussed in this paper. We give a step by step description of the practical aspects of building a Bayesian Network from a dataset. We enumerate and briefly describe the tools required, address the problem of missing values in big datasets resulting from incomplete clinical findings and elaborate on our solution to the problem. We advance some reasons why imputation is a more desirable approach for model building than some other ad hoc methods suggested in literature. In our experiment, the initial Bayesian Network is learned from a dataset using a machine learning program called CB. The network structure and the conditional probabilities are then modified under the guidance of the domain expert. We also present

*Address for correspondence: mgv@cse.sc.edu.

validation results for the unmodified and modified networks and give some suggestions for improvement of the model.

Keywords and phrases: Bayesian networks, Machine learning.

1 Introduction

A large quantity of non-experimental data is generated in Medicine from studies of the natural history of disease, case reports and epidemiological surveys[1]. If experiments are well-designed, it is comparatively easy to analyze and interpret the data obtained. But, making sense of non-experimental data is a difficult task and involves a huge investment of time, effort and expertise. However, data collected for one purpose can often be used to answer other questions. Federally funded research projects make datasets available after the original study is completed. These datasets often are underutilized. This type of data is also referred to as archival data and is basically available to the investigators in "as is" condition [ZyBa91]. Techniques based on Bayesian networks hold great promise in the task of detecting associations which can be interpreted (with great caution!) as causal relationships using non-experimental data [PeVe91, Pear00].

We developed a model to answer the question—"What is the risk of Mental Retardation (MR) for a particular pregnancy or infant based on information from the prenatal, perinatal or postnatal period?" We do not have a diagnostic model in mind. We expect our model to quantify the risk of MR outcome which in the early prenatal period can be used as a guideline for seeking invasive procedures such as amniocentasis for arriving at a definitive diagnosis and recommendation about the desirability of sustaining the pregnancy. During infancy the model may be used to screen children who are at greater risk for MR to plan special educational/environment interventions.

The prevalence of MR is estimated to be about 2.5 per cent of the population [Bats93, StSu92]. MR is a developmental disability with a complex etiology, and the causative factors and mechanisms are not well understood. *"Mental Retardation is characterized by significantly subaverage intellectual functioning"* [AAMR92, p.5]. The American Association on Mental Retardation (AAMR) quantifies the identification of people as those scoring below two Standard Deviations (SD) in a standardized IQ test [AAMR92,

[1] A similar situation exists in many other fields, both in the social and in the natural sciences; witness the tremendous amount of non-experimental data sent by spacecraft for an example outside the social sciences

p.5]. These tests are usually normalized to a mean of 100 with a SD of 15. Those with scores below 50 are classified as having severe mental retardation. Scores in the category of 50-69 fall in the classification of Mild Mental Retardation (MMR). AAMR suggests inclusion of limitation of adaptive skills for individual diagnosis [AAMR92, p.6], but many studies have used cognitive tests (IQ scores) for classification [StSu92, McDe93].

We shall use IQ scores and include the additional category of Borderline Mental Retardation (BMR, scores falling between one and two standard deviations). For severe MR a cause can be found in the majority of cases. In MMR, which forms 85% of MR, a cause cannot be identified in half the cases [Bats93].

So here we have a complex web of unknown causal mechanisms, disagreement among experts, controversies (the large literature of nature versus nurture) and serious gaps in the experts' understanding of the etiological factors. A Bayesian modeling approach may shed some light on the causal mechanisms, give us a tool for prediction of MR and open up avenues for early intervention—medical and social.

A companion publication in the developmental disabilities literature [MaMc97] discusses our model further from a medical perspective. Here, we discuss the techniques used in model building and validation from an applied artificial intelligence perspective.

2 Model Building Methodology

We refer the reader to [Neap90, Section 5.3] for a precise and thorough definition of Bayesian network and to [Pear88, Char91, SDLC93, CDLS99, Jens96, Jens01] for extended presentations of related concepts. Here, we only give a sketch of the definition with a brief example.

A *Bayesian network* consists of a directed acyclic graph (DAG), prior probability tables for the nodes in the DAG that have no parents, and conditional probabilities tables for the nodes in the DAG given their parents. (We identify nodes in the DAG and the variables they represent.) The network and the probability tables define a joint probability distribution on all variables corresponding to the nodes, with the defining property that the conditional probability of any variable v given any set of variables that includes only nodes that are not descendent of v is equal to the conditional probability of v given only its parents. From this property, it follows that the joint probability of the variables in a Bayesian network decomposes in a multiplicative fashion. In turn, this decomposition allows for very efficient computation of marginal posterior probabilities upon observation of evidence.

As an example, the graph in Figure 1 models a small portion of the mental retardation domain. We do not claim that this model is accurate or sensible: it is provided only for the sake of illustration. At the depth of understanding required for the example, the names of the variables should be considered self-explanatory. Recall that a Bayesian network is composed of two parts: an acyclic directed graph and the numerical specification of conditional and prior probability tables. Three features of Bayesian networks are worth mentioning.

First, the directed graph constrains the possible joint probability distributions represented by a Bayesian network. For example, in any distribution consistent with the graph of Figure 1, *Chld_Ravn* (the IQ score of the child) is conditionally independent of *Fam_Inc* (Family Income) given *Mom_Age_Birth* (the age of the mother at birth) and *Mom_Smoke* (whether the mother smokes); also, *P_Mom* (the IQ score of the mother) is conditionally independent of any subset of the other variables given *Mom_Smoke*.

Secondly, the explicit representation of constraints about conditional independence allows a substantial reduction in the number of parameters to be estimated. In the example, assume that the possible values of the five variables are given in Table 1. Then, the joint probability table *P(Fam_Inc, Mom_Age_Birth, Mom_Smoke, Chld_Ravn, P_Mom)* has $2 \times 3 \times 2 \times 4 \times 4 = 192$ entries. It would be very difficult to assess 191 independent parameters[2]. However, the independence constraints encoded in the graph reduce the number of parameters to be estimated to $1 + 4 + 2 + 18 + 6 = 31$. The conditional probability table for Mom_Age_Birth given Fam_Inc is given in Table 2; note that there are only four independent parameters to be estimated, since the sum of values by column is one. Again, we emphasize that these numbers are fictitious.

Thirdly, the Bayesian network representation allows a substantial (usually, dramatic) reduction in the time needed to compute marginals for each variable in the domain. The explicit representation of constraints on independence relations is exploited to avoid the computation of the full joint probability table in the computation of marginals, both prior and conditioned on observations. Space prevents the description of the relevant algorithms. See, e.g., [Jens01, Ch.5] for a discussion of the junction tree algorithm, the most widely used one.

There are two methods of building a Bayesian network for a particular

[2] Probabilities sum to 1, so one of the 192 parameters is dependent on the other 191.

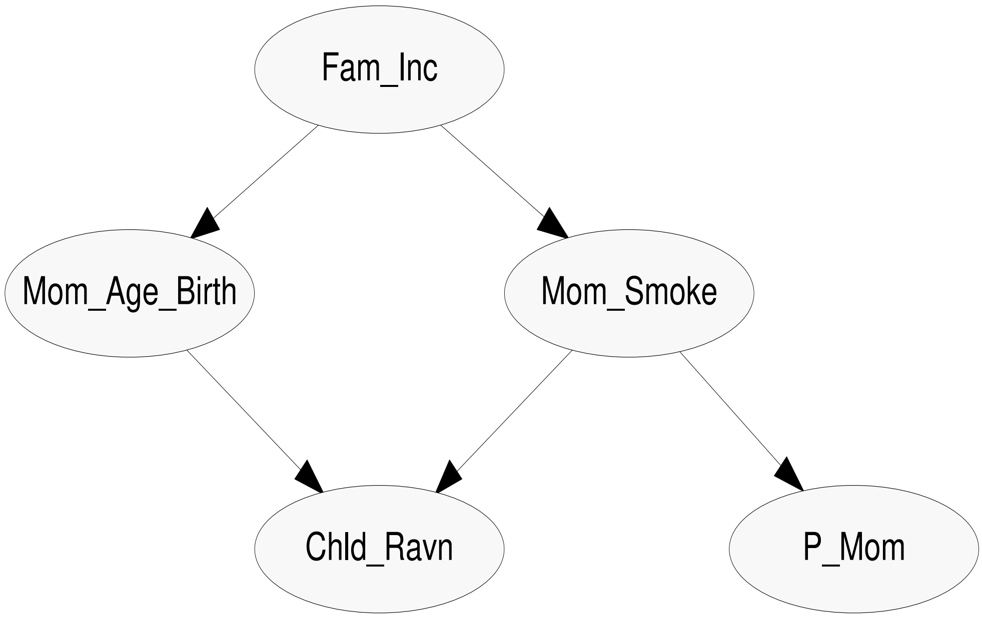

Figure 1: A Microscopic Model of MR

Fam_Inc	>=10000, <10000
Mom_Age_Birth	14-19, 20-34, >=35
Mom_Smoke	yes, no
Child_Ravn	mild, border, normal, super
P_Mom	mild, border, normal, super

Table 1: Values of the Five Micro-Mentor Variables

	Fam_Inc	
	>=10000	<10000
14-19	0.1	0.3
20-34	0.7	0.6
>=35	0.2	0.1

Table 2: Conditional Probability Tables for *Mom_Age_Birth* Given *Fam_Inc*

application domain. The first method consists of asking the domain expert to construct the network (DAG) and assign the prior and conditional probabilities. The second method consists in building the network from data. There are several algorithms available to accomplish this learning task—for example, BIFROST [LaTS93], K2 [CoHe92] and CB [SiVa93, SiVa95]. The prior and conditional probabilities can also be computed from data. The models are validated by comparing with the performance of an expert [SDLC93]. We use a combination of the two strategies—capture the skeleton network from data using the CB algorithm and refine the DAG with the help of the expert and published literature. Prior and conditional probabilities are obtained from data and fine-tuned by the expert.

3 Datasets Used in Model Construction

We obtained the Child Health and Development Study (CHDS) data set, which was developed in a study concerning pregnant mothers and their children [CHDS87]. The children were followed through their teen years and included numerous questionnaires, physical and psychological exams, and special tests. The study was conducted by the University of California at Berkeley and the Kaiser Foundation. It started in 1959 and continued into the 1980's. There are approximately 6000 children and 3000 mothers with IQ scores in the data set. The children were either 5-years old or 9 years old when their IQs were tested. The IQ test used for the children was the Raven Progressive Matrices Test. The mothers' IQs were also tested, and the test used was the Peabody Picture Vocabulary Test.

We identified about 50 variables scattered among several CHDS files that are thought to play a role in the causal mechanism of MR. Under the guidance of the domain expert this set of fifty variables was reduced to a set of twenty-three resulting in the datasets described in 3.1. The subject expert thought that this set of variables was sufficient to capture the domain knowledge. Only one child of the mother is included in each of the datasets. Table 3 contains a list of the twenty-three variables used in the final Bayesian network. (The files used in network construction include a twenty-fourth variable, MARITAL-STATUS, which was removed at a late stage.)

3.1 Datasets Used for Network Construction

RAVN6X24 This dataset contains 5985 cases and 24 variables. In this dataset many of the IQ scores of mothers are missing. The percentage of missing values is 12. This dataset is the total relevant dataset

RAVN2X24 This dataset contains 2212 cases and 24 variables. The IQ scores of mothers and children are present. There are no missing values for the IQ scores. This is a subset of the RAVN6X24 dataset. with all the rows which did not have IQ scores for the mother and child removed. The percentage of missing values is 4.

RAVN6X23 This dataset contains 5985 cases and 23 variables. As only about 3000 mothers were given IQ tests, this dataset was created without the variable P_MOM (IQ score of the mother). This is also a subset of the RAVN6X24 dataset with the variable mother's IQ deleted. The percentage of missing values is 10.

All three datasets were used for network construction, as explained in section 6.1.

Variable	What the Variable Represents
MOM_RACE	Mother's race classified as White (European or White and American Indian or others considered to be of white stock) or non-White (Mexican, Black, Oriental, interracial mixture, South-East Asian).
MOMAGE_BR	Mother's age at time of child's birth categorized as 14-19 years, 20-34 years, or $\geq$ 35 years.
MOM_EDU	Mother's education categorized as $\leq$ 12 and did not graduate, high school, graduated high school, and $>$high school (attended college or trade school).
DAD_EDU	Father's education categorized same as mother's.
MOM_DIS	Yes if mother had one or more of 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; no otherwise.
FAM_INC	Family income categorized as $<$ \$10,000 or $\geq$ \$10,000.
MOM_SMOK	Yes if mother smoked during pregnancy; no otherwise.
MOM_ALC	Mother's alcoholic drinking level classified as mild (0-6 drinks per week), moderate (7-20), or severe >20).
PREV_STILL	Yes if mother previously had a stillbirth; no otherwise.
PN_CARE	Yes if mother had prenatal care; no otherwise.
MOM_XRAY	Yes if mother had been X-rayed in the year prior to or during the pregnancy; no otherwise.
GESTATN	Period of gestation categorized as premature ($\leq$ 258 days), or normal (259-294 days), or postmature ($\geq$ 295 days)..
FET_DIST	Fetal distress classified as yes if there was prolapse of cord, mother had a history of uterine surgery, there was uterine rupture or fever at or just before delivery, or there was an abnormal fetal heart rate; no otherwise.
INDUCE_LAB	Yes if mother had induced labor; no otherwise.
C_SECTION	Yes if delivery was a caesarean section; no if it was vaginal.
CHLD_GEND	Gender of child (male or female).
BIRTH_WT	Birth weight categorized as low $<$ 2500 g) or normal ($\geq$ 2500 g).
RESUSCITN	Yes if child had resuscitation; no otherwise.
HEAD_CIRC	Normal if head circumference is 20 or 21; abmormal otherwise.
CHLD_ANOM	Child anomaly classified as yes if child has cerebral palsy, hypothyroidism, spina binfida, Down's syndrome, chromosomal abnormality, anencephaly, hydrocephalus, Turner's syndrome, cerbellar ataxia, speech defect, Klinefelter's syndrome, or convulsions; no otherwise.
CHLD_HPRB	Child's health problem categorized as having a physical problem, having a behavior problem, having both a physical and a behavioral problem, or having no problem.
CHLD_RAVN	Child's cognitive level, measured by the Raven test, categorized as mild MR, borderline MR, normal, or superior.
P_MOM	Mother's cognitive level, measured by the Peabody test, categorized as mild MR, borderline MR, normal, or superior.

Table 3: The variables used in MENTOR.

4 Tools for Model Building

The **CB algorithm** takes as input a dataset with no missing values and outputs a Bayesian network structure. The network structure, when augmented with suitable conditional probability tables constitutes a Bayesian network, as defined in Section 2 that models the data, in the sense that the data can be taken to be a sample of the distribution encoded by the network. Moreover the network structure output by CB has usually only a few edges, because it exploits independence relations among variables well. The network is therefore appropriate for use by inference algorithms and for visual inspection.

The CB algorithm works in two phases. In the first phase, CB uses Conditional Independence tests (χ^2 tests) for ordering the nodes. In the second phase, which is based on the K2 algorithm [CoHe92], CB computes greedily an approximation to the most likely network structure given the dataset [SiVa95].

Given a dataset and network, **CondProb** computes the prior and conditional probabilities using the formulas in [CoHe92].

HUGIN provides a graphical interface for representing the nodes (domain variables) and the directed edges (usually interpretable as causal relationships between the variables). A user friendly mechanism for naming the variables, entering the states of the variables and assigning the conditional probabilities is also provided. HUGIN implements the Lauritzen and Spiegelhalter method of probability propagation in DAGs [LaSp88], with some improvements. The HUGIN shell was developed by Andersen, Olesen, F.V. Jensen and F. Jensen in Denmark [AOJJ89].

The **IMP** program analyzes the given dataset and predicts missing values. We use statistical, case matching and randomization methods. A random guess is attempted when case matching fails. The method is expected to succeed in domains where there is good interdependency between variables. Fortunately most real world data and medical data in particular have many interdependent variables. We have not analyzed the theoretical properties of IMP, but we consider it to be a practical and useful method particularly for purposes of model building.

CAP-CPN is an application written in C to call causal probabilistic networks(CPN) built using HUGIN. It provides modules to use the HUGIN-API C library in an organized way. CAP-CPN converts an ASCII dataset to the format required by HUGIN for batch validation. It also provides functions to perform simple statistical tests on the data gathered by sampling the outcome node when a batch file containing cases is processed.

5 Handling Missing Values

Real world data contains missing values. This is particularly true of medical datasets. The general practice in the analysis of missing data is deleting cases (records) with missing data. But when there are numerous variables such a policy can mean that most records will have to be disregarded from analysis or many variables will have to be sacrificed. It will help if we can come up with a scheme to predict and assign missing values. To start with, this strategy will be very useful for model building and validation from datasets. We do not discuss the merits and demerits of imputing for data analysis here.

We decided against the easy way of making a separate category for the missing values. We believe that it is not a satisfactory procedure as in most cases it is hard to trace a causal pathway between the missing category of one variable and the missing category of another variable. Treating missing value as a separate category is also likely to create serious problems in computing the conditional probabilities from data. For example, assigning the conditional probability of a variable with 4 states which has 5 parents having 2 states each results in a table of 128 entries. Now if a missing category is included, the table space grows to 1215 entries. And for this example (which is by no means an extreme case) we have more than 1000 junk entries. Not only is the size of the conditional probability table a problem, but we also encounter semantic difficulties computing conditional probabilities for the occurrences including missing states. Hence it is desirable to come up with a scheme to avoid missing categories. Then in the quantitative modeling stage only the valid categories of the variables in the network and the conditional probabilities will have to be entered. Another method which has been used [LaTS93, page 94] is to assign one of the valid categories to all the missing values of a particular variable. It may be suitable for variables where the domain expert can predict with a high probability which category the missing values should have.

We developed and implemented an algorithm(IMP) for predicting and imputing missing values. The accuracy of IMP can be validated using datasets across domains. Datasets without any missing values were used for validation. By random number generation a fixed percentage (say ten percent) of data values are assigned missing, thus obtaining a dataset on which IMP is run to impute missing values. The output dataset is compared with the original dataset. Our validation tests using LED, ALARM and SOYBEAN which are small to large artificial datasets used for Machine Learning research and available from the University of California at the Irvine Machine Learning repository [MuAh94] gave a mean accuracy of 80% over ten runs. The range was from 67% to 95%.

Another possibility is to impute a dataset using the algorithm. This imputed dataset has no missing values. Now we assign missing values (we can assign the same percentage of missing values originally present) generating random numbers, imputing and comparing with the dataset we created originally by imputing. This technique, called *customized validation*, gives the predictive accuracy for the particular dataset in question with its given percentage of missing values. Even though this takes into account the size and other peculiarities of the dataset for validation purposes, it may introduce a small error for the estimate as we are using IMP twice for validation.

Our datasets were imputed using IMP. For our datasets RAVN2X24, RAVN6X23, and RAVN6X24, we obtained an accuracy of 79%, 82% and 83% respectively. The accuracy of the imputed values were judged by the technique of customized validation.

6 Network Generation and Refinement

6.1 Network Generation

The CB algorithm was run on the three imputed datasets described in Section 3.1 for generating the networks. The datasets were randomly partitioned into two—a major part and a minor part. The bigger partition was used for constructing the network and the smaller part was set apart for validation. For RAVN2X24, we used the first 2000 cases for generating the network and for the other two, the first 5000. The network generated from the RAVN6X24 dataset is shown in Figure 2. The networks obtained are given in tables 4, 5, and 6.

6.2 Network Refinement

We defined three rules to characterize the inadequacies of the generated networks.

Rule of Chronology Events occurring later in time cannot be the parents of earlier incidents. For instance a child health problem cannot be the parent of maternal disease.

Rule of Common Sense The directed edges of the network should not go against common sense. For instance, Father's education cannot be a

Variable	Parents
MOM_RACE	
MOM_AGE_AT_BIRTH	MAR_STAT, MOM_EDU,
	FAM_INC, PREV_STILLBRTH
MAR_STAT	
MOM_EDU	MOM_RACE
DAD_EDU	MOM_EDU[2], P_MOM[1]
MOM_DIS	MOM_AGE_AT_BIRTH
FAM_INC	MOM_EDU
MOM_SMOKE	MOM_RACE, MOM_EDU,
	MOM_ALC, PN_CARE[3]
MOM_ALC	FAM_INC
PREV_STILLBRTH	
PN_CARE	
MOM_XRAY	MOM_DIS, C_SECTION
GESTATN	MOM_RACE, FET_DIST
FET_DIST	INDUCE_LAB[1], C_SECTION[1], RESUSCITN[1]
INDUCE_LAB	
C_SECTION	
CHLD_GEND	CHLD_HPROB[1]
BIRTH_WT	PN_CARE, GESTATN
RESUSCITN	MOM_RACE[2]
HEAD_CIRC	MAR_STAT, INDUCE_LAB[3], CHLD_ANOM
CHLD_ANOM	
CHLD_HPROB	CHLD_ANOM
CHLD_RAVN	MOM_EDU, CHLD_ANOM
P_MOM	MOM_RACE, MOM_EDU

Table 4: RAVN2X24 NETWORK

[1] Violates law of chronology

[2] Goes against commonsense

[3] Violates domain rules

Variable	Parents
MOM_RACE	MAR_STAT[1], MOM_EDU[1], DAD_EDU[2], MOM_SMOKE[1], MOM_ALC[1], C_SECTION[1]
MOM_AGE_AT_BIRTH	MAR_STAT, FAM_INC, PREV_STILLBRTH
MAR_STAT	
MOM_EDU	MOM_AGE_AT_BIRTH[2], FAM_INC
DAD_EDU	
MOM_DIS	MOM_RACE, MOM_AGE_AT_BIRTH, MAR_STAT[2]
FAM_INC	MAR_STAT
MOM_SMOKE	MOM_EDU, MOM_ALC, BIRTH_WT
MOM_ALC	FAM_INC, MOM_AGE_AT_BIRTH
PREV_STILLBRTH	
PN_CARE	
MOM_XRAY	MOM_RACE[2], MOM_DIS, C_SECTION
GESTATN	MOM_AGE_AT_BIRTH, PREV_STILLBRTH, PN_CARE
FET_DIST	PN_CARE, GESTATN, INDUCE_LAB[1], C_SECTION[1],
INDUCE_LAB	PREV_STILLBRTH
C_SECTION	
CHLD_GEND	CHLD_HPROB[1]
BIRTH_WT	PN_CARE, GESTATN
RESUSCITN	MOM_RACE[2], MAR_STAT[2], FET_DIST
HEAD_CIRC	MAR_STAT, CHLD_ANOM
CHLD_ANOM	
CHLD_HPROB	MAR_STAT, HEAD_CIRC
CHLD_RAVN	FAM_INC

Table 5: RAVN6X23 NETWORK

[1] Violates law of chronology

[2] Goes against commonsense

[3] Violates domain rules

Variable	Parents
MOM_RACE	MAR_STAT[1], MOM_EDU[1], DAD_EDU[2], MOM_SMOKE[1], C_SECTION[1], RESUSCITN[1]
MOM_AGE_AT_BIRTH	MAR_STAT, FAM_INC, PREV_STILLBRTH
MAR_STAT	
MOM_EDU	MOM_AGE_AT_BIRTH[2], FAM_INC
DAD_EDU	
MOM_DIS	MOM_AGE_AT_BIRTH, MAR_STAT[2], MOM_EDU, HEAD_CIRC[2]
FAM_INC	MAR_STAT
MOM_SMOKE	MOM_EDU, MOM_ALC, BIRTH_WT
MOM_ALC	FAM_INC, MOM_AGE_AT_BIRTH
PREV_STILLBRTH	
PN_CARE	
MOM_XRAY	MOM_DIS, C_SECTION
GESTATN	MOM_AGE_AT_BIRTH, PREV_STILLBRTH, PN_CARE
FET_DIST	GESTATN, INDUCE_LAB[1], C_SECTION[1], RESUSCITN[1]
INDUCE_LAB	PREV_STILLBRTH
C_SECTION	
CHLD_GEND	RESUSCITN[2]
BIRTH_WT	PN_CARE, GESTATN
RESUSCITN	MAR_STAT[2], PN_CARE, GESTATN
HEAD_CIRC	MAR_STAT, CHLD_ANOM
CHLD_ANOM	
CHLD_HPROB	MOM_RACE, MAR_STAT, HEAD_CIRC, P_MOM
CHLD_RAVN	FAM_INC
P_MOM	MOM_RACE, MOM_AGE_AT_BIRTH, MOM_EDU

Table 6: RAVN6X24 NETWORK

[1] Violates law of chronology

[2] Goes against commonsense

[3] Violates domain rules

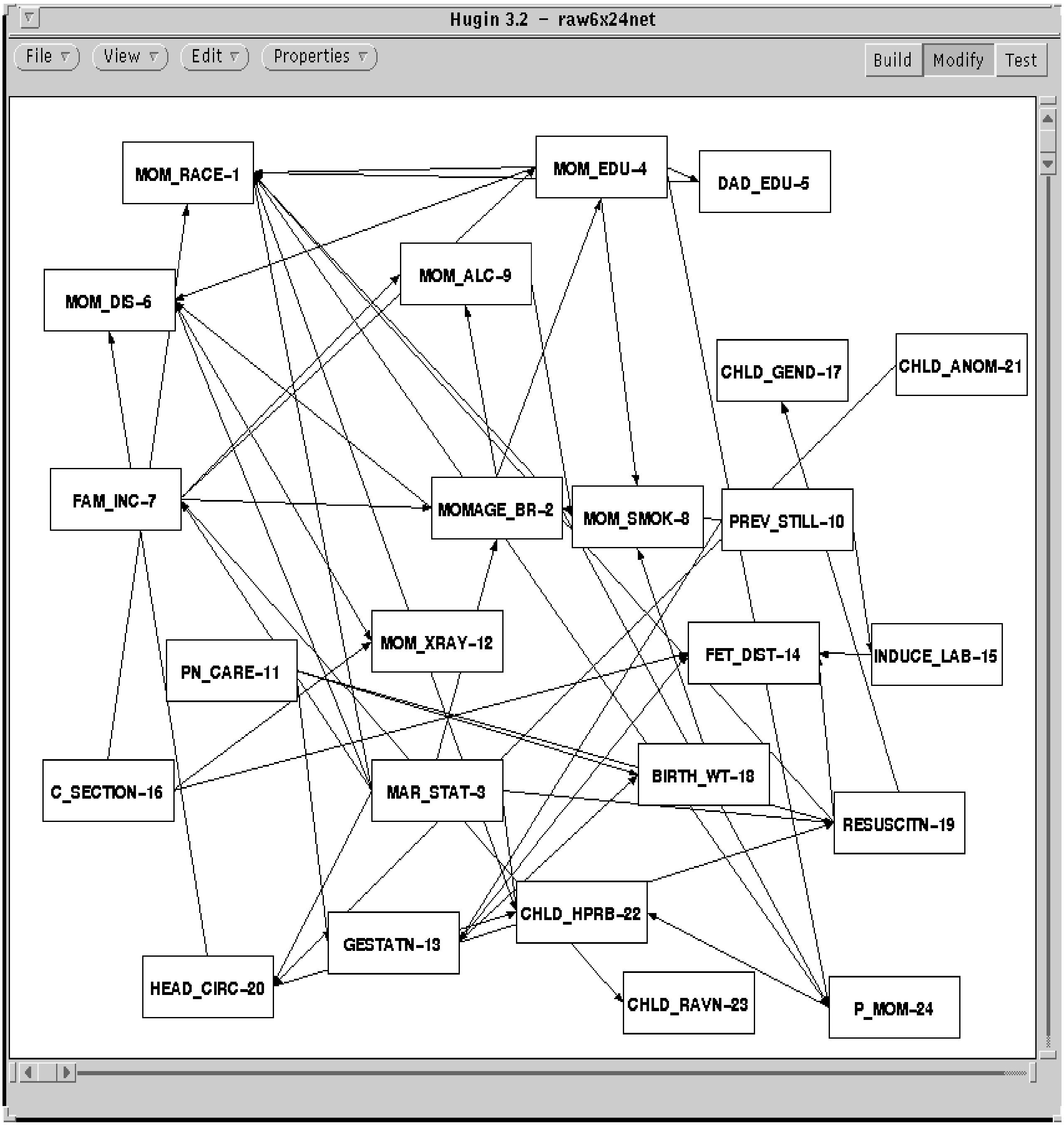

Figure 2: Network Generated by CB from RAVN6X24 Dataset

cause of Mother's race.

Domain Rule This rule has been referred as the *Rule of Biological Plausibility* in the medical and biological science literature. This rule states that a causal explanation is tenable in terms of existing knowledge level about the variables involved. This level is what we obtain from an intelligent review of the relevant literature. The directed edges should not violate established domain rules. For example, pre-natal care cannot be put down as a cause of maternal smoking. Mausner and Kramer strike a note of caution here: "The development of biological knowledge often introduces new factors that previous studies have not taken into account. In the existing studies, the major causal factors may have been missed because their importance was not appreciated." [MaKr85, p.187]. This point is well taken and if there is a strong case, such a directed edge should be investigated further. But for our network construction purposes, if an edge clearly violated established domain constraints, it was removed. The directed edges of the network in Figure 2 are given in Table 6 with annotations describing examples of rules that are broken. So also new edges were incorporated to capture the knowledge of the known domain causal mechanisms. The variable MAR_STAT was removed as the expert felt that it was not playing a useful role in representing domain relations. See Tables 4, 5 and 6 for examples of rules that are broken. The expert refined network is a synthesis and refinement of the three raw networks. The expert-modified network is shown in Figure 3.

6.3 Refinement of Conditional Probabilities

The prior and conditional probabilities were computed using the program CondProb. For nodes without parents prior probabilities of the various states calculated from the RAVN6X24 dataset were assigned. For the nodes with one or more parents, the conditional probabilities calculated using the same dataset was assigned. The Conditional Probabilities of the outcome variable CHLD_RAVN (See paragraph 3) were refined by the expert. There were many possible instantiations that were not represented in the dataset RAVN6X24. A reasonable conditional probability was assigned by the expert for these. For the raw networks probabilities were assigned from the RAVN2X24 dataset using the program CondProb.

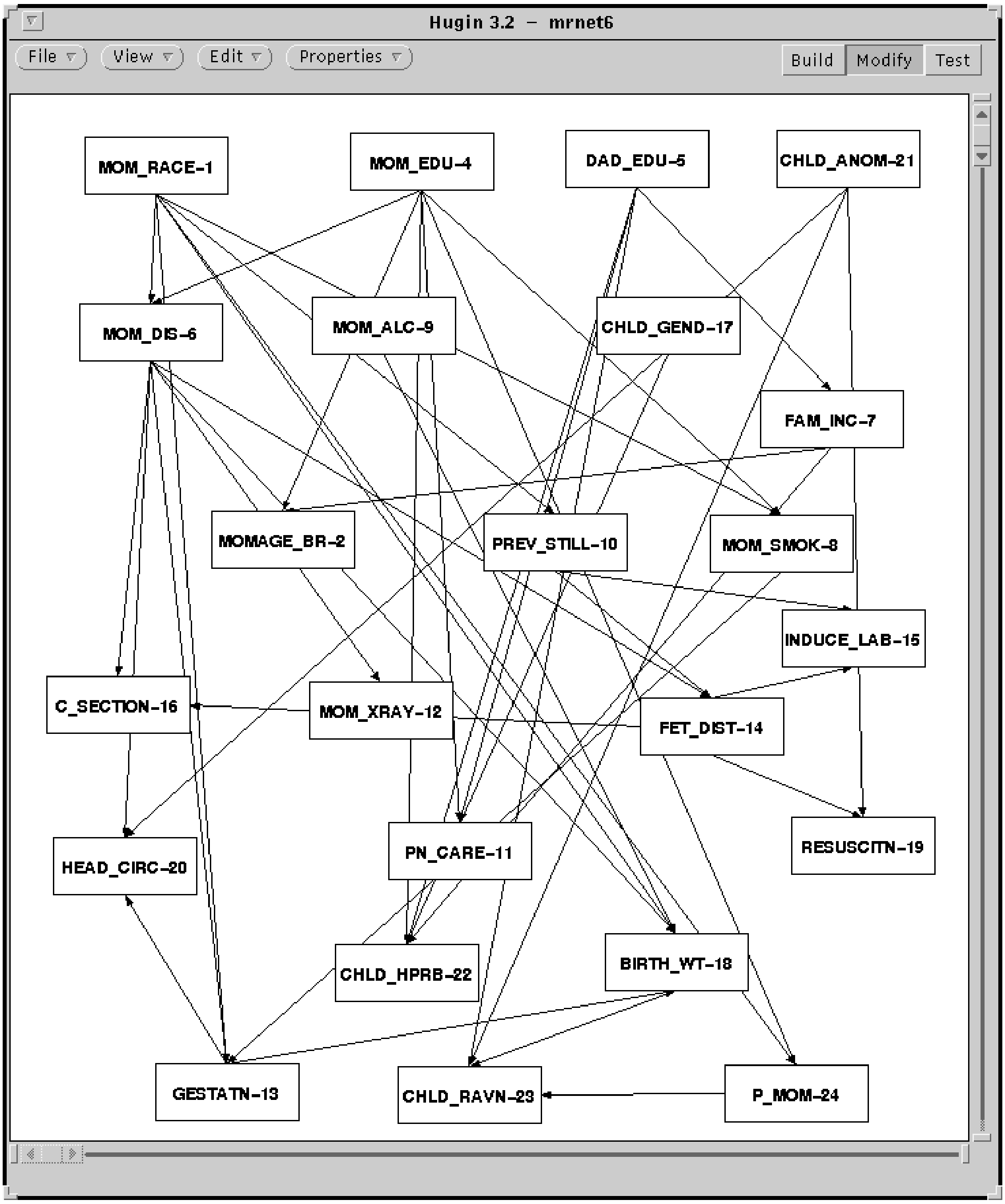

Figure 3: Network Modified by the Expert

7 Validation of the Model

7.1 Validation by the expert

As ours is a model for risk assessment and risk prediction of mental retardation, it is different from a classification or diagnostic problem. In a typical diagnostic approach we consider a set of differential diagnoses and the attempt is to assign probabilities to them and order them on that basis. In risk assessment we are interested in the change in magnitude of a particular category of interest even though it may still occupy a low position in an ordering of the variable levels. We have a prior probability of 5.6% for mild and 12.4% for borderline MR. Hence if the risk of both mild and borderline doubles, still we get a combined probability of only 36%. That leaves a probability of 64% for normal and superior. Most of the actual cases from the dataset with mild or borderline MR give a >50% probability for normal outcome. This is because there are more normal outcome cases with similar instantiations of variables than outcomes that result in mental retardation. Hence we decided first on a strategy of validation by comparing with the expert. We generated nine cases with instantiation for a subset of variables. We ran these cases on the model and computed the probabilities. The expert was asked to score the results as *agree* or *disagree*. The expert was in agreement with the model's assessment in eight out of nine cases used for validation. Three of the cases are depicted in Table 7, while the conditional probabilities of the values of CHLD_RAVN for those cases are shown in Table 8.

7.2 Validation using RAVN2X24

7.2.1 Risk Means of Cases and Controls

All the cases in the dataset RAVN2X24(unimputed) were run through the models—the expert network, and the two raw networks that have twenty-four variables (raw2x24net and raw6x24net) using CAP-CPN. The results showing the relative risk for controls (children with normal outcome) and cases (children with mild or borderline MR) are given in Table 9 for the three nets.

A TTest procedure was performed to assess the statistical significance of the predicted risks. The Prob>|T| was 0.0001, 0.0000 and 0.1878 for the expert net, raw2x24net and raw6x24net respectively. This shows that there is significant difference in the mean risk scores between the cases and controls for the expert net and raw2x24net ($P<0.05$). Note that there are

Variable	Case 1 Variable Value	Case 2 Variable Value	Case 3 Variable Value
MOM_RACE	non-White	White	White
MOMAGE_BR	14-19		≥ 35
MOM_EDU	≤ 12	$>$ high school	≤ 12
DAD_EDU	≤ 12	$>$ high school	high school
MOM_DIS			no
FAM_INC	$< \$10,000$		$< \$10,000$
MOM_SMOK			yes
MOM_ALC			moderate
PREV_STILL			
PN_CARE		yes	
MOM_XRAY			yes
GESTATN	normal	normal	premature
FET_DIST		no	yes
INDUCE_LAB			
C_SECTION			
CHLD_GEND			
BIRTH_WT	low	normal	low
RESUSCITN			
HEAD_CIRC			abnormal
CHLD_ANOM		no	
CHILD_HPRB			both
CHLD_RAVN			
P_MOM	normal	superior	borderline

Table 7: Generated values for three cases.

Value of CHLD_RAVN and Prior Probability	Case 1 Posterior Probability	Case 2 Posterior Probability	Case 3 Posterior Probability
mild MR (.056)	.101	.010	.200
borderline MR (.124)	.300	.040	.400
normal (.731)	.559	.690	.380
superior (.089)	.040	.260	.200

Table 8: Posterior probabilities for three cases.

fewer violations of the rules described in Section 6.2 for raw2x24net (9) than for raw6x24net (14). The performance of the expert net was the best of the three based on the mean risks.

7.2.2 Evaluation using a risk threshold

In the initialized state, the expert network gives a resting MR risk of 0.18 if the risks for mild and borderline retardation are added together, as shown in Figure 4.

If we take twice the resting state risk as our threshold for significant risk, our threshold can be set at the value of 0.36. Using this threshold we find that twenty nine per cent of cases are flagged correctly. This also results in eighteen per cent of controls being flagged as significant risk for MR. (Figure 5 shows the increase in risk of MR for an example case with some known risk factors.) These results are contained in Table 10, which also presents the same type of results for raw2x24net, whose resting MR risk is 0.16. (The results for raw6x24 are very poor.)

We considered using n-fold cross-validation to estimate the error rates of our methods, but we did not employ it, because it can be applied only to the raw networks. In n-fold cross-validation, the cases are divided into n (typically, ten) groups of roughly equal size. All except one of the groups are used in learning, while the group that is left out is used to estimate an error rate. This process is carried out n times, each time leaving out a different group, and the overall error rate for the learning algorithm is the average of the n error rates [WeKu91]. Since the expert modified network was crafted from the three raw networks by removing and adding edges to reflect domain

Level	expert net		raw2x24net		raw6x24net	
	Controls n=1863	*Cases* n=349	*Controls* n=1863	*Cases* n=349	*Controls* n=1863	*Cases* n=349
	mean risk		mean risk		mean risk	
Mild MR	0.05	0.07	0.02	0.02	0.02	0.02
Borderline MR	0.11	0.14	0.14	0.16	0.14	0.15
Mild + Border	0.16	0.21	0.16	0.18	0.16	0.17

Table 9: Mean Risk of MR predicted for cases and controls by the three nets

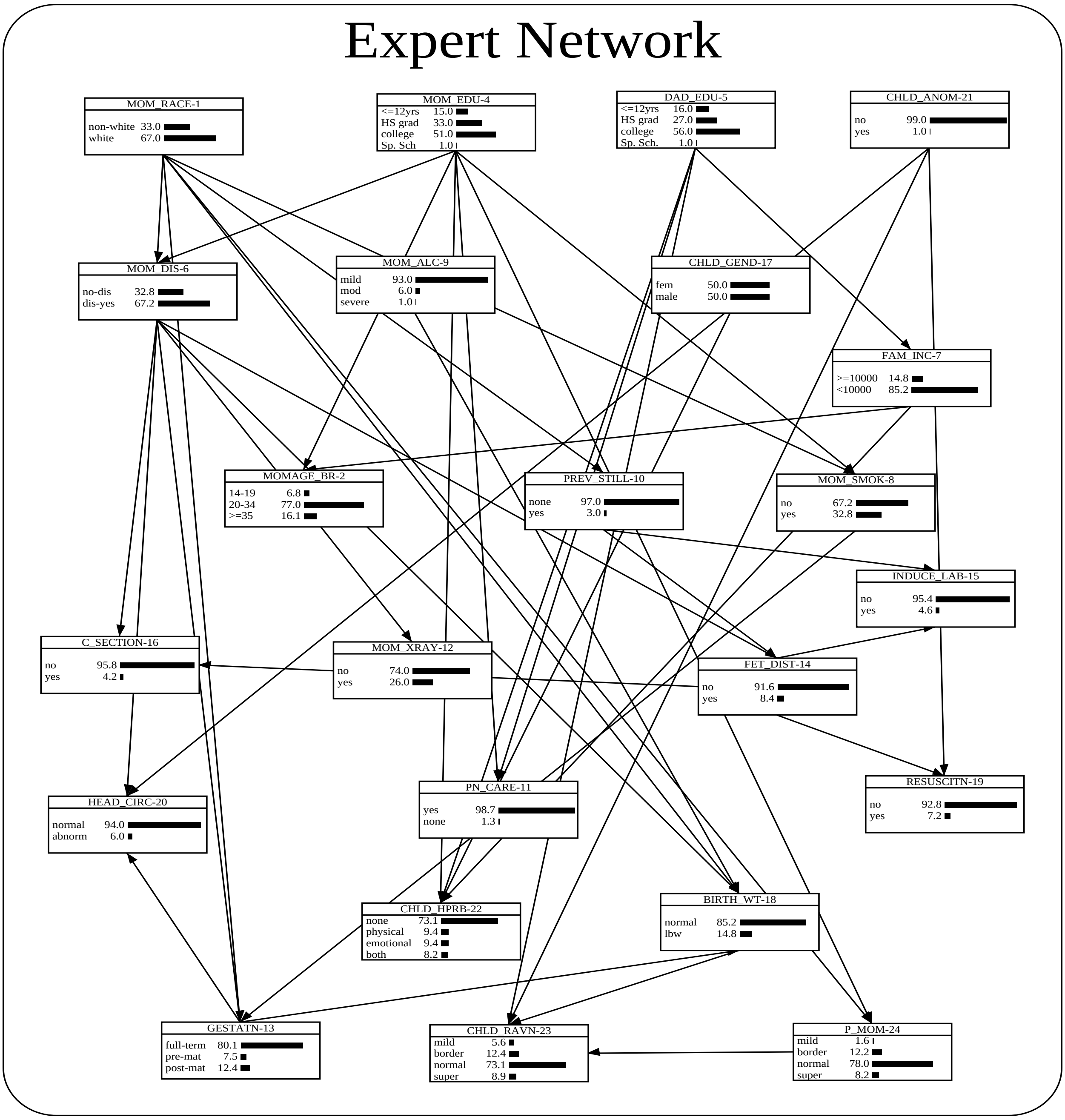

Figure 4: Initial probabilities in the expert network.

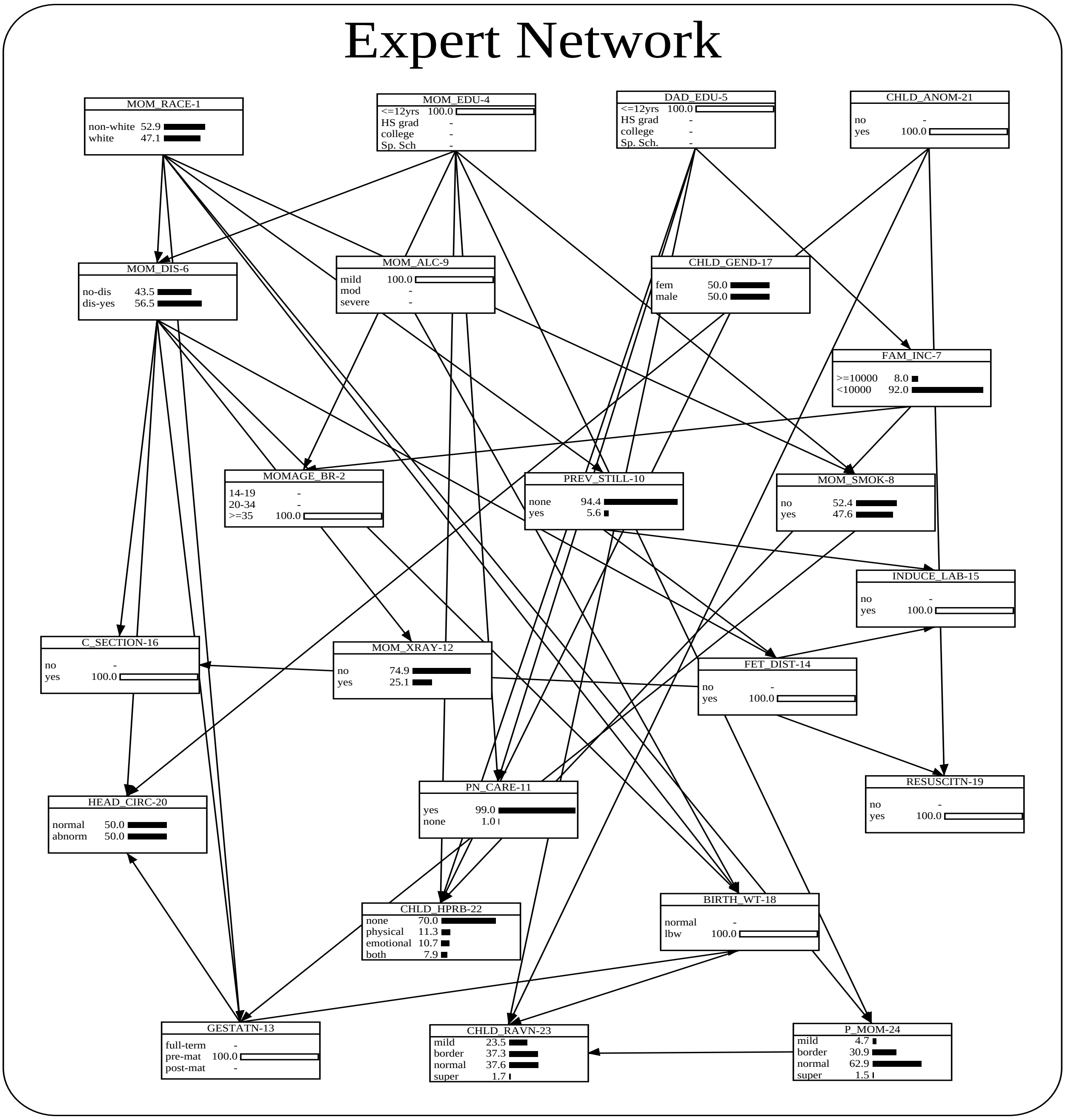

Figure 5: Probabilities in the expert network for a high risk case.

knowledge and the conditional probabilities were also modified by the expert, cross-validation cannot be used to compare the performance of the raw and refined networks.

7.2.3 Validation Using a Separate Data Set

The National Collaborative Perinatal Project (NCPP), of the National Institute of Neurological and Communicative Disorders and Strokes, developed a data set containing information on pregnancies between 1959 and 1974 and 8 years of follow-up for live-born children. For each case in the data set, the values of all 22 variables except CHLD_RAVN (child's cognitive level as measured by the Raven test) were entered, and the conditional probabilities of each of the four values of CHLD_RAVN were computed. Table 11 shows the average values of $P(\text{CHLD\_RAVN} = mildMR|\mathsf{d})$ and $P(\text{CHLD\_RAVN} = borderlineMR|\mathsf{d})$, where d is the set of values of the other 22 variables, for both the controls (children in the study with normal cognitive function at age 8) and the subjects (children in the study with mild or borderline MR at age 8).

8 Discussion of Results and Future Work

8.1 Limitations of the Model

The validation results are significant but not dramatic. This we feel is due to the incomplete state of knowledge of the etiological factors of MR. This results in datasets where some of the relevant variables (not yet recognized as contributory or causative) have not been collected. Hence our model is constrained by the state of domain knowledge existing at this point in time.

Risk Threshold for MR (Mild + Border)	expert net		raw2x24net	
	Controls n=1863	*Cases* n=349	*Controls* n=1863	*Cases* n=349
Resting Value	434 (23%)	122 (35%)	901 (48%)	240 (69%)
1.5 × Resting Value	370 (20%)	111 (32%)	242 (13%)	73 (21%)
2 × Resting Value	342 (18%)	101 (29%)	9 (<1%)	5 (1%)

Table 10: Cases flagged for different risk thresholds

Cognitive Level	Avg. Probability for Controls ($n = 13019$)	Avg. Probability for Subjects ($n = 3598$)
Mild MR	.06	.09
Borderline MR	.12	.16
Mild or Borderline MR	.18	.25

Table 11: Average probabilities, as determined by MENTOR, of having mental retardation for controls (children identified as having normal cognitive functioning at age 8) and subjects (children identified as having mild or borderline MR at age 8).

Throughout the development of MENTOR, we emphasized the causal interpretation of the links. While this is not in any way necessary, it seemed to be a good decision for two reasons. First, there is widespread belief that ordering variables in a causal direction simplifies modeling, or, as Russell and Norvig put it, "If we stick to a causal model, we end up specifying fewer numbers, and the numbers will often be easier to come up with" [RuNo95, p.443]. Second, it is easier to involve the expert in validating the edges if the model is causal.

Still, there is a serious problem with a causal interpretation in the case of MENTOR. It is quite likely that there are many hidden (unmeasured or even unknown) variables playing a role in the causal pathway of MR. Despite this, we attempted to build a DAG model and assign to it a causal interpretation. This is clearly suspect. There are two possible approaches to dealing with this problem. The first is to attempt to discover hidden variables using a data analysis technique, such as TETRAD [Spir00, Ch.2] that purports to discover such variables. A second approach is to explicitly model correlations that have no causal interpretation by using undirected links as in chain graphs (graphical models that include both directed and undirected links). It is commonly (and somewhat simplistically) believed that the undirected links can be used to model associational, non-causally interpretable information, while the directed links are used to model the causally interpretable information [LaRi01]. For our research we did not address this issue further.

8.2 Future Work

The networks generated from the different datasets using the CB algorithm had many nodes 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, directed edges violating *domain rules* would be avoided. In other words, the ordering of nodes built by the first phase of the CB algorithm would be forced to be consistent with *chronology* and *domain* rules. Another mechanism to incorporate these rules in the network generation phase would be to set appropriate priors on the network structures, which would favor networks compatible with the rules. This is a more drastic change to the current approach, in that the scoring metric used by CB assumes prior equivalence of all network structures. (Other metrics, while forgoing prior equivalence, are also insensitive to domain peculiarities.)

The model we built can be improved by using more recent and larger datasets, which we have already acquired. We expect that both the network structures and the conditional probabilities will be modified.

In some cases, the values of variables in the original dataset have been discretized. In many cases (e.g., for head circumference), this has been done according to accepted practice in epidemiology. Still, it may be interesting to challenge accepted wisdom and attempt different ways of discretizing variable ranges, for example by using a decision tree building algorithm. (We observe, incidentally, that the variable FAM_INC, which represents family income, was already normalized in the original dataset.)

Finally, since there are many variables in the Mental Retardation domain, it may be advantageous to attempt attribute selection, and use only a subset of variables.

Acknowledgments

We thank Shanthi Krishnaswami for preprocessing of the data sets used in this study and for help with the t-test. We thank Doug Fisher for many detailed comments on an early draft of this paper, Finn V. Jensen and other colleagues at the Department of Computer Science at the University of Aalborg for comments on a presentation of this work, and Richard Neapolitan for comments on a more recent draft. We thank several participants in the Fifth International Workshop on AI and Statistics for useful comments. We also thank Hugin Ltd.

References

[AAMR92] American Association on Mental Retardation. *Mental Retardation* American Association on Mental Retardation, 1719 Kalorama Road, NW Washington DC, 1992.

[AOJJ89] Andersen, Stig K., Kristian G. Olesen, Finn V. Jensen, and Frank Jensen. "HUGIN—a Shell for Building Bayesian Belief Universes for Expert Systems." *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence,* Detroit, 1080–1085, 1989.

[Bats93] Batshaw, Mark L. "Mental Retardation." In: *Pediatric Clinics of North America—The Child With Developmental Disabilities, 40, 3, 507–522.* Batshaw, Mark L, ed. Philadelphia, PA: W.B Saunders Company 1993.

[Char91] Charniak, Eugene. "Bayesian Networks without Tears." *AI Magazine*, 12, 4 (Winter 1991), 50–63.

[CHDS87] *Data Archive and Users Manual of the Child Health and Development Studies, Vols. 1 & 2.* School of Public Health, University of California at Berkeley, CA, March 1, 1987.

[CoHe92] Cooper, Gregory F. and Edward Herskovits. "A Bayesian Method for the Induction of Probabilistic Networks from Data." *Machine Learning, 9,* 309–347, 1992.

[CDLS99] Cowell, Robert G., A. Philip Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter. *Probabilistic Networks and Expert Systems.* New York: Springer, 1999.

[Jens96] Jensen, Finn V. *An Introduction to Bayesian Networks.* New York: Springer, 1996.

[Jens01] Jensen, Finn V. *Bayesian Networks and Decision Graphs.* New York: Springer, 2001.

[LaRi01] Lauritzen, Steffen L. and Thomas Richardson. "Chain Graph Models and their Causal Interpretation." Research Report R-01-2003, Department of Mathematical Sciences, Aalborg University. (The revised version of August 2001 is available at `http://www.math.auc.dk/~steffen/papers/rss.ps`. The paper is to appear in *Journal of the Royal Statistical Society, Series B.*)

[LaSp88] Lauritzen S.L. and D.J. Spiegelhalter. "Local Computation with probabilities on Graphical Structures and their Applications to Expert Systems." *J.R.Statist. Soc., Series B*, 50, 157–224, 1988.

[LaTS93] Lauritzen S.L., B. Thiesson, and D. Spiegelhalter. "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, 1993, 93–105.

[MaMc97] "MENTOR: A Bayesian Model for Prediction of Mental Retardation in Newborns." *Research in Developmental Disabilities*, 18 (1997), 5, 303–318.

[MaKr85] Mausner, Judith S. and Shira Kramer. "The Concept of Causality and Steps in the Establishment of Causal Relationships." In: *Epidemiology—An Introductory Text, Ch 7.* Philadelphia, PA: W.B Saunders Company 1985.

[McDe93] McDermott, Suzanne W., Ann L. Coker and Robert E. McKeown. "Low birth-weight and risk of mild mental retardation by ages 5 and 9 to 11." *Pediatric and Perinatal Epidemiology, 1993, 7,* 195–204.

[MuAh94] Murphy, P.M. and D.W. Aha. UCI Repository of Machine Learning Databases. [`http://www.ics.uci.edu/~mlearn/ML.Repository.html`]. Irvine, CA: University of California, Dept. of Information and Computer Science.

[Neap90] Neapolitan, Richard. *Probabilistic Reasoning in Expert Systems: Theory and Algorithms.* New York: Wiley, 1990.

[Pear88] Pearl, Judea. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Mateo, California: Morgan Kaufmann, 1988.

[Pear00] Pearl, Judea. *Causality: Models, Reasoning, and Inference.* Cambridge, UK: Cambridge University Press, 2000.

[PeVe91] Pearl, Judea, and Thomas Verma. "A Theory of Inferred Causation." In: J.A. Allen, R. Fikes, and E. Sandewall (eds.), *Prin-*

ciples of Knowledge Representation and Reasoning: Proceedings of the Second Conference, pp. 441-452. San Mateo, CA: Morgan-Kaufmann, 1991.

[RuNo95] Russell, Stuart, and Peter Norvig. *Artificial Intelligence: A Modern Approach.* Englewood Cliffs, New Jersey: Prentice Hall, 1995.

[SiVa93] Singh, Moninder and Marco Valtorta. "A New Algorithm for the Construction of Bayesian Network Structures from Data." *Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence (UAI-93),* Washington, DC, July 1993, 259–264.

[SiVa95] Singh, Moninder and Marco Valtorta. "Construction of Bayesian Belief Networks from Data: a Brief Survey and an Efficient Algorithm." *International Journal of Approximate Reasoning,* 12 (1995), 111–131.

[SDLC93] Spiegelhalter, David J., A. Philip Dawid, Steffen L. Lauritzen, and Robert G. Cowell. "Bayesian Analysis in Expert Systems." *Statistical Science,* 8(1993),3, 219–283.

[Spir00] Spirtes, Peter, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search,* second edition. Cambridge, Massachusetts: MIT Press, 2000.

[StSu92] Stein, Zena A. and Mervyn W. Susser. "Mental Retardation." In: *Public Health and Preventive Medicine 13th ed.* Last, John M. and Robert B. Wallace, eds. Appleton & Lange, San Mateo, CA, 1992.

[WeKu91] Weiss, Sholom M. and Casimir A. Kulikowski. *Computer Systems that Learn.* San Mateo, CA: Morgan-Kaufmann, 1991.

[ZyBa91] Zytkow, Jan M. and John Baker. "Interactive Mining of Regularities in Databases." In: *Knowledge Discovery in Databases, p31-54* Gregory Piatetsky-Shapiro and William J. Frawley, eds. The AAAI Press, 445 Burgess Drive, Menlo Park, CA 94025, 1991.