Introduction to Bayesian network and its application in prediction coronary heart disease

Hao Ying[a]

[a] *Department of Biostatistics, School of Public Health Science, University of Miami, USA*

1 INTRODUCTION

1.1 Bayesian network

Bayesian network (BN) is a probabilistic graphical model that represents a set of variables and their conditional dependencies via a directed acyclic graph (DAG). Bayesian networks are ideal for taking into account the prior knowledge about the data structure and use it to study the likelihood of the components of interest [1,2].

1.2 DAG structures

DAGs are closely related to the Bayesian models. In a DAG structure, the nodes represent random variables and the arcs represent probabilistic dependencies between them [3]. In Epidemiology studies, the DAGs often reflect causal relations among covariates, and is a useful tool to identify confounders, mediators or effect modifiers. DAGs can involve complicated relationships thus presenting like a network framework.

Specific structures of DAG correspond to certain Epidemiology terminologies (Figure 1):

1. $f(O,E) = f(O|E)f(E)$, then O and E are associated.
2. $f(O,E,I) = f(O|I)f(I|E)f(E)$, then I is a mediator between E and O.
3. $f(O,E,F) = f(O|E,F)f(F)f(E)$, then E, F are independent predictors about O.
4. $f(O,F,E) = f(O|E,C)f(E,C)f(C)$, then C is a confounder of association between O and E.
5. The effect modifier can be written as: $f(O|E) = f_1\mathbf{1}_{E\in A} + f_2\mathbf{1}_{E\notin A}$

Note that different DAG structures can have equivalent conditional probabilistic relations. For example, the DAG $A \leftarrow B \rightarrow C$ and $A \rightarrow B \rightarrow C$ are equivalent since:

$$f(A|B)f(B)f(C|B) = f(A,B)f(C|B) = f(A)f(B|A)f(C|B)$$

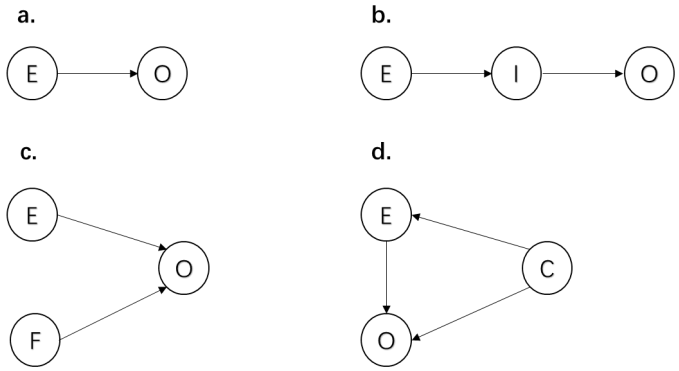

Figure 1: DAG structures and corresponding Epidemiology effects. (a) – (d) indicate probabilistic dependency, mediator, independent predictors, and confounder.

1.3 Structure learning

In most cases, expert DAGs (true structures) are difficult to be completely detailed. Instead, we only obtain "piece of truth". That is, we merely know that some relations are likely and some are impossible, but not all of them. Structure learning algorithms can determine the optimal Bayesian network structures underlies the data. Our partial knowledge can also be incorporate into the learning process by specifying whitelist and blacklist. Whitelist is the structure with arcs and nodes that always present in the network. The blacklist contains arcs that are never included.

The structure learning algorithms can be grouped in two categories: constraint-based and score-based algorithms [4]. Constraint-based algorithms use the conditional independence tests to detect the Markov blankets of the variables, which in turn are used to compute the structure of the Bayesian network [5,6]. Score-based leaning algorithms are general purpose heuristic optimization algorithms which rank networks with respect to a goodness-of-fit score [7]. In this study, we apply Interleaved Incremental Association (inter-IAMB) algorithm as an example of constraint-based algorithm, and Hill Climbing (HC) as a score-based greedy search algorithm. The analysis is conducted by R package "bnlearn" , and the plotting is mainly based on package "graphviz" [4].

2 SIMULATION DATASET

2.1 Data simulation

We construct a dataset with Gaussian-distributed variables upon the following models:

$$A = \mathcal{N}(1,1), B = \mathcal{N}(2,9), E = \mathcal{N}(3.5,4), G = \mathcal{N}(5,4)$$
$$C = 2A + 2B + \mathcal{N}(2,1/4)$$
$$D = 1.5B + \mathcal{N}(6,1/9)$$
$$F = 2A + D + E + 1.5G + \mathcal{N}(0,1)$$

The underlying network structure (expert structure) is illustrated in Figure 2 (top left). The thickest arcs indicate correlations with coefficients 2, the median-length arcs stand for coefficients 1.5, and the slim arcs suggest weakest associations of 1. The simulation dataset has 5,000 data points. Variable F is treated as the outcome of interest.

2.2 Structure learning

We perform structure leaning based on Hill Climbing (HC) algorithm on the simulated dataset. The algorithm successfully identifies the true structure. We then fit the Bayesian network using the expert DAG. The distribution of F is perfectly estimated as:

$$\hat{F} = 1.995A + 1.006D + 1.003E + 1.494G + \mathcal{N}(-0.006, 0.996)$$

However, these results are not stable if restricted to a smaller training set. We construct five random sampled subsets of 50 data points and summarizes the structure learning results (Figure 2). Only one out of five scenarios archives the true structure. Other subsets lead to similar, but not identical, networks.

A more sophisticated experiment is performed by computing the chance of identifying true DAG at subsets of different size (Figure 3). In order to reliably (>90% chance) obtain true DAG through structure learning, a sample size greater than 1,000 is recommended.

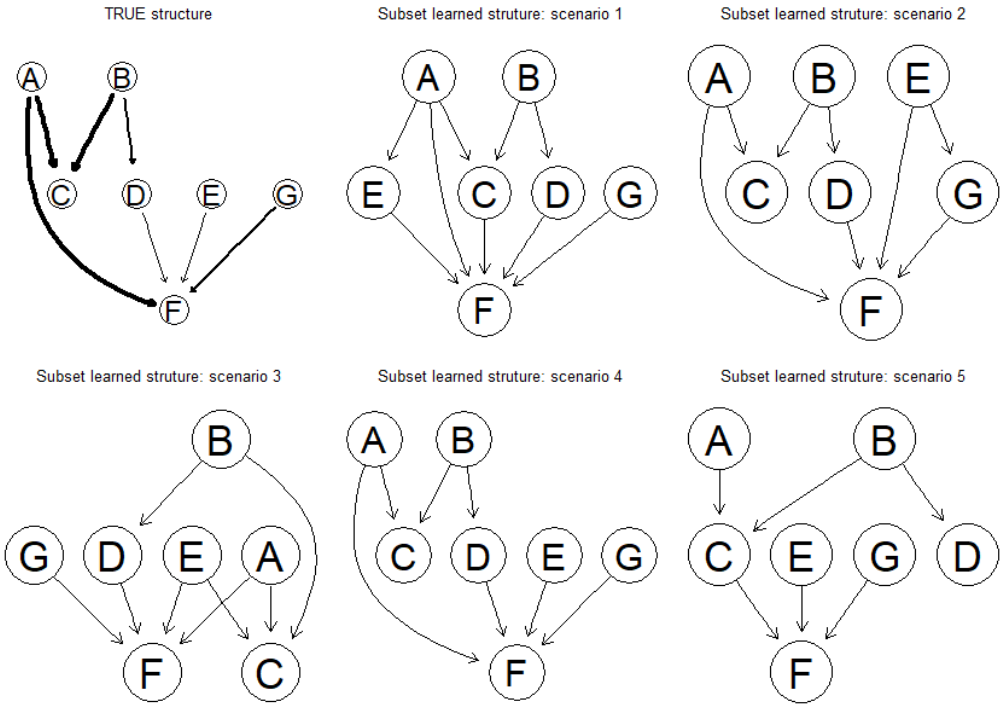

Figure 2: True structure and subset learned structures for the simulated data. *True structure*: thickness of the arcs indicates how strong the association is between variables. *Subset learned structure*: learned from random sampled subsets of size 50.

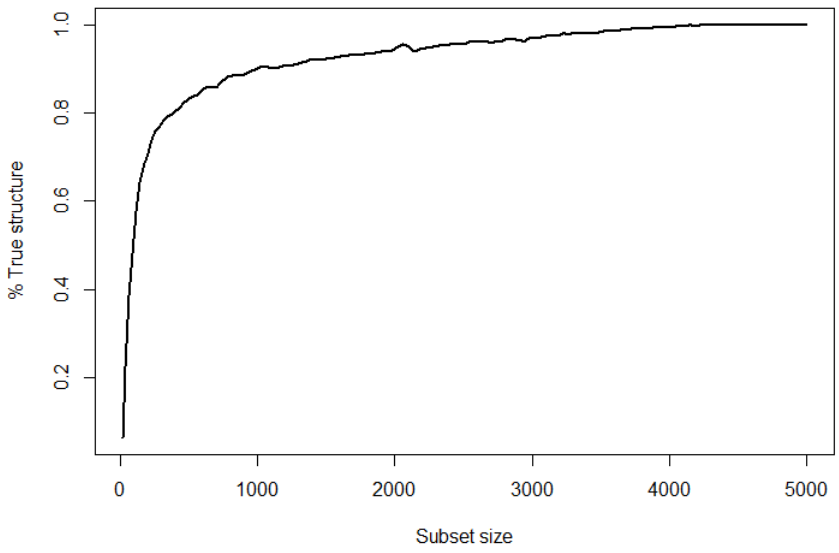

Figure 3: Chance of identifying the true DAG at random sampled subsets of different sizes.

2.3 Prediction

We conduct Bayesian network prediction for variable F and compare the performance with traditional multivariate linear regression methods (with all covariates in the model). The prediction error is estimated through cross-validation delete-d (CVd) algorithm. The algorithm randomly selects d (in this case use optimal d=383) samples as validation set and trains the model with the rest of the data points [8,9]. Since the algorithm prefers small training sets, CVd error is usually larger than other type of CV errors. The CVd mean square errors (CVd-MSE) are used as the criteria to evaluate prediction performances.

Figure 4 shows the mean MSEs from 1000 simulations under every given training set size. The red line stands for the BN prediction errors using expert network. If we possess precise understanding about the true DAG, then BN method outperforms other methods. As the training sample size increases, the MSE approaches the theoretical variance of outcome, which is the minimum MSE we can achieve (Figure 4, red). If we know nothing about the truth, the BN method predicts poorer than linear regression when the training set is small (n<100). However, under larger training sets, BN beats the traditional linear regression methods (Figure 4, blue & black).

Furthermore, we exhaustively fit all possible linear models and identify the model with best prediction performance (using "bestglm" package). This model successfully includes all the effective predictors (A, D, E, G) of outcome F (as shown in expert DAG), and yields equivalent prediction errors as BN. However, the exhaustive algorithm is much more time-consuming.

Therefore, the prediction performance of Bayesian networks depends on two factors: 1. More prior knowledge leads to more detailed and precise DAG and yields better prediction. 2. Larger size of training set can ensure the identification of true DAG, also provides more accurate estimates of the

parameters in the Bayesian model (especially when the network is complicated).

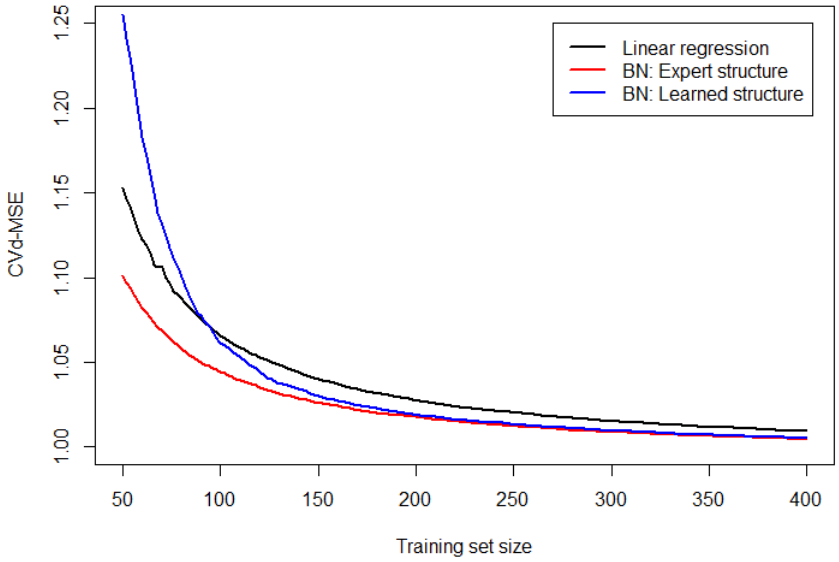

Figure 4: CVd-MSE using different set of training set. Black: linear regression with all variables. *Red*: Bayesian network prediction using true DAG. *Blue*: Bayesian network prediction using DAG learned from training set.

3 HEART DISEASE DATASET

3.1 Material and method

The data are taken from a larger dataset, described by Rousseauw et al, 1983, South African Medical Journal [10], with 462 observations and 9 variables (including the outcome). The outcome of interest is the diagnosis of Coronary Heart Disease (CDH, coded as *chd*) ascertained by medical records.

Patients' characteristics are summarized in Table 1. Out of 462 patients, 160 (34.6%) have CHD, and 302 (65.4%) are free of CHD. All exposures are continuous except for family CHD history (*famhist*) with 1 stands for "present" and 0 for "absent". Tobacco use (*tobacco*) is measured by cumulative tobacco (kg) in lifetime. Type A behavior (*typea*) is measured using Bornter's short score [16] ranging from 12 to 84 with higher score indicating more hostile, competitive personalities. Other covariates include continuous age (*age*), obesity measured by BMI (*obesity*), current alcohol use (*alcohol*), level of low density lipoprotein cholesterol (*ldl*), and systolic blood pressure (*sbp*). The distributions of covariates are different between CHD and non-CHD groups (Table 1).

	No CHD (N=302)	CHD (N=160)	Overall (N=462)
Age, years			
Mean (SD)	38.9 (14.9)	50.3 (10.6)	42.8 (14.6)
Median [Min, Max]	40.0 [15.0, 64.0]	53.0 [17.0, 64.0]	45.0 [15.0, 64.0]
Systolic blood pressure			
Mean (SD)	135 (18.0)	144 (23.7)	138 (20.5)
Median [Min, Max]	132 [101, 214]	138 [102, 218]	134 [101, 218]
Family CHD history			
Absent	206 (68.2%)	64.0 (40.0%)	270 (58.4%)
Present	96.0 (31.8%)	96.0 (60.0%)	192 (41.6%)
Type A behavior score			
Mean (SD)	52.4 (9.52)	54.5 (10.2)	53.1 (9.82)
Median [Min, Max]	52.5 [13.0, 77.0]	55.0 [20.0, 78.0]	53.0 [13.0, 78.0]
BMI, m/kg2			
Mean (SD)	25.7 (4.09)	26.6 (4.39)	26.0 (4.21)
Median [Min, Max]	25.6 [17.8, 46.6]	26.5 [14.7, 45.7]	25.8 [14.7, 46.6]
Current alcohol assumption			
Mean (SD)	15.9 (23.5)	19.1 (26.2)	17.0 (24.5)
Median [Min, Max]	6.05 [0, 145]	8.33 [0, 147]	7.51 [0, 147]
Cumulative tobacco (kg)			
Mean (SD)	2.63 (3.61)	5.52 (5.57)	3.64 (4.59)
Median [Min, Max]	1.04 [0, 20.0]	4.13 [0, 31.2]	2.00 [0, 31.2]
Low density lipoprotein cholesterol			
Mean (SD)	4.34 (1.87)	5.49 (2.23)	4.74 (2.07)
Median [Min, Max]	3.98 [0.980, 15.3]	5.07 [1.55, 14.2]	4.34 [0.980, 15.3]

Table 1: Patients' characteristics by coronary heart disease diagnosis in South African adults.

3.2 Structure learning

By computing the covariance matrix, the covariates are highly correlated with each other, implying complicated network structure of the dataset. The structures are learned in the full dataset using both Inter-IAMB and HC algorithms.

Prior knowledge from literature review are incorporated into the whitelist (Figure 5, blue arcs):

1. High blood pressure [11] and elevated LDL levels [12] are believed to be risk factors and often precedes a diagnosis of CHD.
2. Alcohol abuse are reported to be associated with escalated SBP [13] and LDL levels [14,15].
3. Obesity adults are more likely to have high SBP [22] and LDL levels [23,24].
4. Age is correlated to tobacco use since the latter is measured as lifetime cumulative tobacco use.
5. Type A behavior is a valid predictor of CHD risk [16,17].
6. Family history is a strong risk factor of CHD reported in multiple studies [18-20].

Besides, we forbit some relations and specify the blacklist:

1. The exposures were ascertained before the onset of CHD, thus we prevent any arrows starting from *chd* due to temporality.
2. Aging (*age*) is a natural process that cannot be the downstream of an arc.
3. Family CHD history (*famhist*) is also unlikely to be affected by other covariates.

The two algorithms return very similar DAGs as shown in Figure 5. Actually, the only difference is that the inter-IAML algorithm believes that older age leads to increasing Type A behaviors (Figure 5, orange arcs). The DAG implies that age is a potential confounder since it is related to most of the covariates as well as the outcome. Tobacco use has direct effects on CHD even after accounting for the *tobacco–alcohol–sbp/ldl–chd* causal pathway. The effects of alcohol use obesity are mediated by blood pressure and LDL levels.

We define the HC-learned network (Figure 5, left) as the "expert structure" (compared to the structure learned from training sets) and use it for prediction.

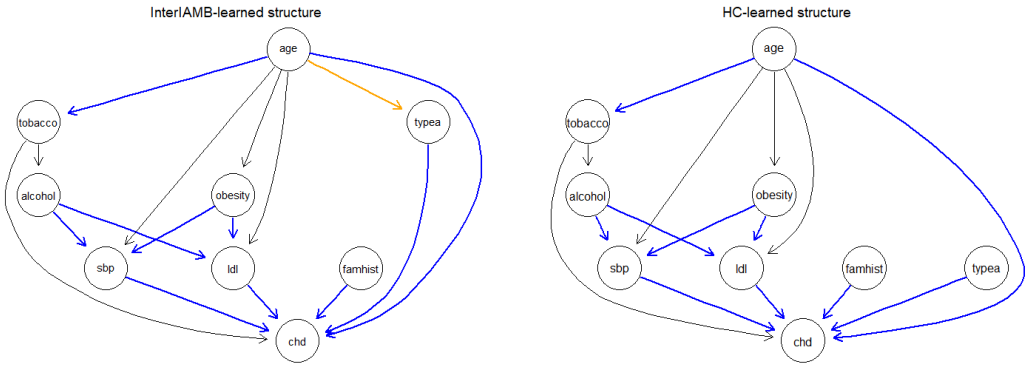

Figure 5: Structure learning results in CHD dataset. *Blue*: arcs incorporated in the whitelist. These relations always present during structure leaning. *Orange*: disagreement among two algorithms.

3.3 Best logistic model

We want to find a traditional frequentist model for comparison. By fitting all covariates into univariate logistic models, we find that all individual variables significantly increase the risk of CHD. In order to directly view their associations, we construct locally Lowess-smoothed scatterplots against the outcome and discover that some effects may not be linear (Figure 6). Thus, we decide to include the quadratic terms for *alcohol*, *tobacco*, *obesity*, and *sbp*. Furthermore, interactions between covariates are tested pairwise. Potential interactions between *sbp* & *typea*, and *alcohol* & *obesity*, have been identified.

First, we center and scale all continuous covariates including quadratic terms and interactions. Then, an exhaustive strategy is adopted to identify the best logistic prediction model by directly comparing prediction error among all possible models (using BIC as selection criteria yields the same best model).

The prediction error is estimated through cross-validation delete-d (CVd) algorithm with optimal d=383. We repeat the algorithm 1000 times for each model and pick one model with the lowest mean test error. The best logistic model includes the following predictors: age, tobacco use, Type A score, family history, and LDL level. Note that the expert DAG identifies predictors as *age*, *tobacco*, *typea*, *famhist*, *ldl*, and *sbp*, which almost matches the best logistic predictors. *sbp* is not considered as predictor in the logistic model since it shares the same pathway with *tobacco* (Figure 5).

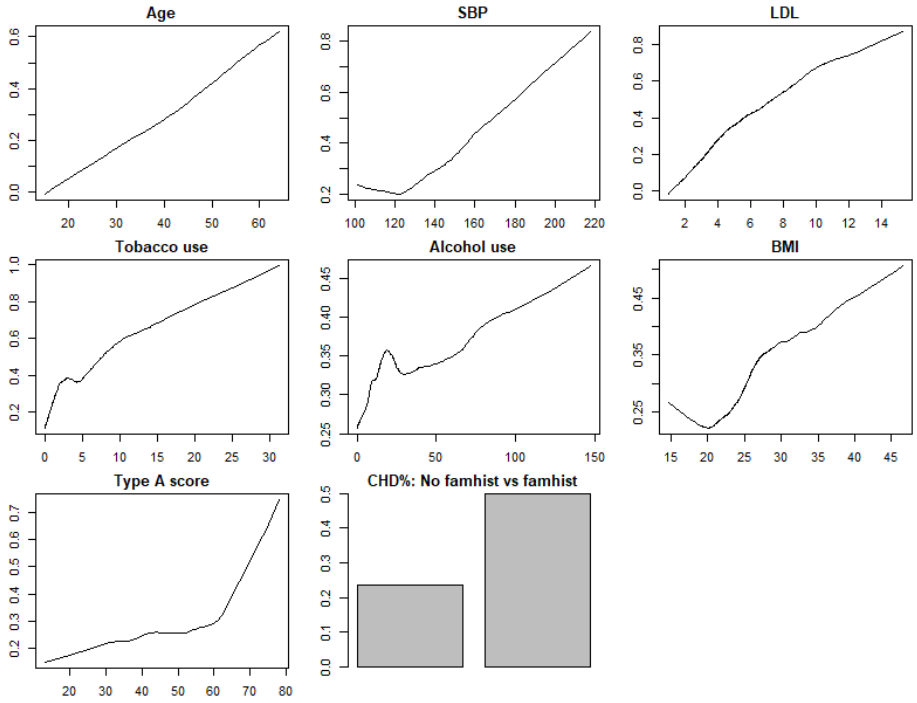

Figure 5: Effects of covariates on CHD diagnosis displayed by locally Lowess-smoothed curves.

3.4 Prediction performance

Since the "bnlearn" package doesn't support estimation for binary outcomes with continuous parent nodes. We treated *chd* as continuous, and BN provides real-valued scores with higher values indicating higher risk of CHD. Area under the curve of ROC (AUC) is used to quantile the discrimination power across two groups. The AUC measures the following property of a classifier: if randomly take one sample from both groups, what's the probability that the one from the positive outcome group has the higher score? AUC=100% means that the two groups are totally separated. That is, all cases have higher scores than controls. AUC=50% suggests that the two groups are mixed by pure randomness [21].

Three prediction models have been brought into comparison: best logistic model, BN with expert structure, BN with structure learned from training sets. Note that in the last method, the structure is learned with only blacklist specified – no prior knowledge (whitelist) being incorporated.

Figure 6 shows the results under different sizes of training sets. Under the given training size, mean AUC

is computed out of 1000 iterations. BN with expert DAG outperforms the best logistic model when the training set is sufficiently large (n>250) (Figure 6, red). Under small training set, the best logistic model has the top performance, since it has the fewest parameters to be estimate (Figure 6, black). Without prior knowledge, the BN methods turn out to be the least effective classifiers (Figure 6, blue). If we render the whitelist in the last method, the blue line is expected to be closer to the red line, suggesting a crucial role for prior knowledge when conducting BN prediction.

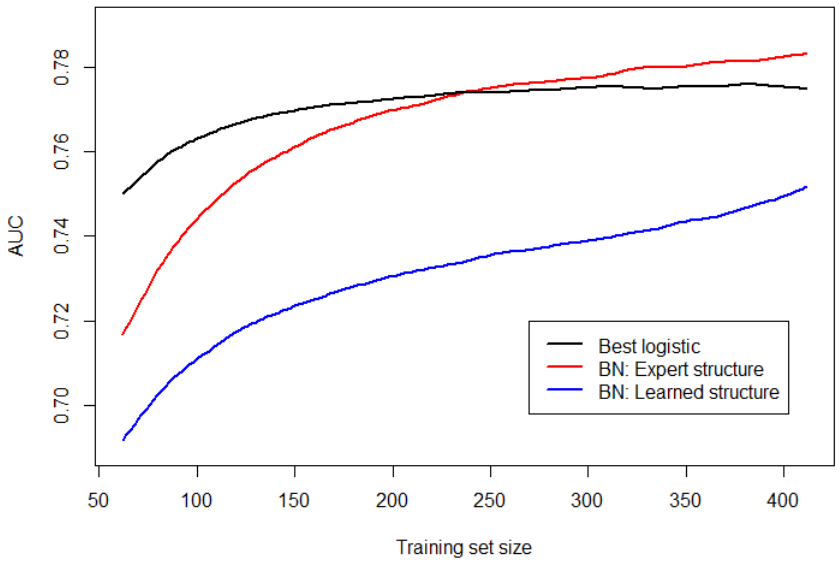

Figure 6: Classification power measured by AUC under different sizes of training sets. *Black*: best logistic model with selected variables. *Red*: Bayesian network prediction using expert DAG. *Blue*: Bayesian network prediction using DAG learned from training set without incorporation of prior knowledge.

4 DISCUSSION

In this study, we compare the prediction performance of Bayesian network with traditional frequentist methods, in both simulated and real-world datasets. BN algorithm has the following strengths:

1. It is a powerful tool that provides insights into the probabilistic structure among variables.
2. BN prediction can surpass the traditional regression methods in both continuous and classification settings with sufficient prior knowledge and enough training data.
3. The structure learning algorithm is fast and easy to conduct. Exhaustive search for frequentist models can sometimes generate similar results. However, these strategies are usually extremely time-consuming, or even unrealistic to perform.

In the CHD dataset, mis-specified DAGs or small sample size can significantly reduce the prediction power of BN. Following strategies maybe able to improve the prediction power further:

1. Including more variables helps to identify a more accurate and thorough “expert DAG” .

2. The data is from a retrospective study conducted in 1980s. Additional techniques may be required to detect and tackle the potential misclassification bias in both exposures and responses.
3. Since the data is part of a larger dataset, using the whole dataset with greater sample size is expected to refine the prediction by BN algorithm.

5 REFERENCE

[1] Herskovits, E. (1991). Computer-based probabilistic-network construction (Doctoral dissertation, Stanford University).

[2] Korb, K. B., & Nicholson, A. E. (2010). Bayesian artificial intelligence. CRC press.

[3] Thulasiraman, K., & Swamy, M. N. S. (1992). 5.7 acyclic directed graphs. Graphs: theory and algorithms, 118.

[4] Scutari, M. (2009). Learning Bayesian networks with the bnlearn R package. arXiv preprint arXiv:0908.3817.

[5] Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference Morgan Kaufmann, San Mateo, California 19882. Guttman EA Suchman PF Lazarfeld SA Star and JA Classen Wiley New York 1966.

[6] Verma, T., & Pearl, J. (1991). Equivalence and synthesis of causal models Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence.

[7] Chickering, D. M. (1995). A new characterization of equivalent Bayesian network structures. Submitted for publication.

[8] Shao, Jun (1993). Linear Model Selection by Cross-Validation. Journal of the American Statistical Assocation 88, 486-494.

[9] Shao, J. (1997). An asymptotic theory for linear model selection. Statistica sinica, 221-242.

[10] Rossouw, J. E., Du Plessis, J. P., Benadé, A. J., Jordaan, P. C., Kotze, J. P., Jooste, P. L., & Ferreira, J. J. (1983). Coronary risk factor screening in three rural communities. The CORIS baseline study. South African medical journal= Suid-Afrikaanse tydskrif vir geneeskunde, 64(12), 430-436.

[11] Khot, U. N., Khot, M. B., Bajzer, C. T., Sapp, S. K., Ohman, E. M., Brener, S. J., ... & Topol, E. J. (2003). Prevalence of conventional risk factors in patients with coronary heart disease. Jama, 290(7), 898-904.

[12] Manninen, V., Tenkanen, L., Koskinen, P., Huttunen, J. K., Mänttäri, M., Heinonen, O. P., & Frick, M. H. (1992). Joint effects of serum triglyceride and LDL cholesterol and HDL cholesterol concentrations on coronary heart disease risk in the Helsinki Heart Study. Implications for treatment. Circulation, 85(1), 37-45.

[13] Foerster, M., Marques-Vidal, P., Gmel, G., Daeppen, J. B., Cornuz, J., Hayoz, D., ... & Rodondi, N. (2009). Alcohol drinking and cardiovascular risk in a population with high mean alcohol consumption. The American journal of cardiology, 103(3), 361-368.

[14] Castelli, W., Gordon, T., Hjortland, M., Kagan, A., Doyle, J., Hames, C., ... & Zukel, W. (1977). Alcohol and blood lipids: the cooperative lipoprotein phenotyping study. The Lancet, 310(8030), 153-155.

[15] Choudhury, S. R., Ueshima, H., Kita, Y., Kobayashi, K. M., Okayama, A., Yamakawa, M., ... & Miyoshi, Y. (1994). Alcohol intake and serum lipids in a Japanese population. International journal of epidemiology, 23(5), 940-947.

[16] Bortner, R. W. (1969). A short rating scale as a potential measure of pattern A behavior. Journal of chronic diseases, 22(2), 87-91.

[17] Jenkins, C. D., Rosenman, R. H., & Zyzanski, S. J. (1974). Prediction of clinical coronary heart disease by a test for the coronary-prone behavior pattern. New England Journal of Medicine, 290(23), 1271-1275.

[18] Pohjola-Sintonen, S., Rissanen, A., Liskola, P., & Luomanmäki, K. (1998). Family history as a risk factor of coronary heart disease in patients under 60 years of age. European heart journal, 19(2), 235-239.

[19] Leander, K., Hallqvist, J., Reuterwall, C., Ahlbom, A., & de Faire, U. (2001). Family history of coronary heart disease, a strong risk factor for myocardial infarction interacting with other cardiovascular risk factors: results from the Stockholm Heart Epidemiology Program (SHEEP). Epidemiology, 215-221.

[20] Tada, H., Melander, O., Louie, J. Z., Catanese, J. J., Rowland, C. M., Devlin, J. J., ... & Shiffman, D. (2016). Risk prediction by genetic risk scores for coronary heart disease is independent of self-reported family history. European heart journal, 37(6), 561-567.

[21] Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.

[22] Maggio, A. B., Aggoun, Y., Marchand, L. M., Martin, X. E., Herrmann, F., Beghetti, M., & Farpour-Lambert, N. J. (2008). Associations among obesity, blood pressure, and left ventricular mass. The Journal of pediatrics, 152(4), 489-493.

[23] Couillard, C., Ruel, G., Archer, W. R., Pomerleau, S., Bergeron, J., Couture, P., ... & Bergeron, N. (2005). Circulating levels of oxidative stress markers and endothelial adhesion molecules in men with abdominal obesity. The Journal of Clinical Endocrinology & Metabolism, 90(12), 6454-6459.

[24] Weinbrenner, T., Schröder, H., Escurriol, V., Fito, M., Elosua, R., Vila, J., ... & Covas, M. I. (2006). Circulating oxidized LDL is associated with increased waist circumference independent of body mass index in men and women. The American journal of clinical nutrition, 83(1), 30-35.