20th International Congress on Modelling and Simulation, Adelaide, Australia, 1–6 December 2013
www.mssanz.org.au/modsim2013

Combining Structure Equation Model with Bayesian Networks for predicting with high accuracy of recommending surgery for better survival in Benign prostatic hyperplasia patients

C. Yoo [a] and S. Oh [b]

[a] Department of Biostatistics, Florida International University, Miami, FL, USA, [b] Department of Urology, Seoul National University Hospital, Seoul, Korea
Email: cyoo@fiu.edu

Abstract: Causal discovery is the key aspect of science. Inferring causality can be achieved in various ways. Typically, you start with your hypothesis (based on what you know so far) and based on the data you collect, you update your hypothesis. In a nutshell, causality can be inferred via your background knowledge and empirical data. Causal Both Bayesian networks (BN) and Structural equation model (SEM) are graphical models that are able to model causality both from background knowledge and empirical data. SEM heavily relies on your background knowledge and use data to justify your knowledge. On the other hand, BN can combine your background knowledge with a causal model that gives the maximum likelihood based on data. If the relationships turn out to be statistically significant, then expert's knowledge is considered statistically valid and can be used to provide guidelines in practice using the model based on expert's knowledge.

Functional changes of the bladder are commonly seen in patients with benign prostatic hyperplasia (BPH). We investigated the predictive factors for the alteration in the bladder storage function in patients with BPH using Bayesian networks (BN) and Structure Equation Model (SEM). We analyzed database of consecutive 1,352 patients with BPH who underwent urodynamic studies from Oct 2004 to Oct 2011 in a single institution.

We show that combining BN and SEM enables us to build a data driven prediction model with latent constructs. The BN showed that (1) predicting outcome variables, when TZVol is known TPVol and PSA do not add value in prediction; (2) if we know StoragePhaseDetrusor, then all input variables do not help more in predicting Bladder Compliace; (3) when BladderSensation is known, BOOI plays important role in predicting BladderCapacity and StoragePhaseDetrusor; (4) if we know TPVol then TZVo and PSA are independent.

User Self Reported Condition and Eurodynamic Study Results did not receive significant latent score and all the variables are discrete, BN was a natural pick to model latent constructs. Volume Capability of Patient's Bladder receive significant latent score and all the variables are continuous, SEM was used to model latent. The combined data driven model reveals that bladder outlet obstruction (BOO) increases the risk of secondary bladder storage dysfunction in patients with BPH. This suggests that more aggressive treatment for BOO might be recommended.

Keywords: *Bayesian Networks, Structure Equation Modeling, Benign Prostatic Hyperplasia, Urodynamics, Bladder Outlet Obstruction*

1. INTRODUCTION

Causal discovery is the key aspect of science. Inferring causality can be achieved in various ways. Typically, you start with your hypothesis (based on what you know so far) and based on the data you collect, you update your hypothesis. In a nutshell, causality can be inferred via your background knowledge and empirical data. Both Bayesian networks (BN) and Structural equation model (SEM) are graphical models that are able to model causality. BN and SEM can combine cause–effect information (typically, expert's knowledge) with empirical data to provide a quantitative assessment of causal relationships among the modeled variables. If the relationships turn out to be statistically significant, then expert's knowledge is considered statistically valid and can be used to provide guidelines in practice using the model based on expert's knowledge.

Recently, SEM techniques have been developed to learn causal relationships from data (Spirtes, et al., 2000). However, SEM lacks its prediction performance mainly because it models linear relationships. If the relationships are non-linear, SEM will result in poor prediction performance. These limitations of SEM can be augmented by Bayesian networks. On the other hand, BN lacks in learning causal relationships among continuous variable especially if the number of variables is large (≥ 50). The linking of SEM to BN has been proposed in a two-step approach (Gupta and Kim, 2008). In the method, the latent scores obtained from SEM are used as data for BN. However, the approach is not practical in terms of learning causal relationships from data directly. Here we propose more simpler and practical approach in combining BN and SEM in terms of learning causal relationships from data.

2. METHODS

A Bayesian network is a directed acyclic graph in which each node represents a variable and each arc represents probabilistic influence. A causal Bayesian network (or *causal network* for short) is a Bayesian network in which each arc is interpreted as a direct causal influence between a parent node (variable) and a child node, relative to the other nodes in the network (Pearl, 1988). Figure 1 illustrates the structure of a hypothetical causal Bayesian network structure that contains five nodes. The probabilities associated with this causal network structure are not shown.

The causal network structure in Figure 1 indicates, for example, that the *Gene*1 can regulate (causally influence) the expression level of the *Gene*3, which in turn can regulate the expression level of *Gene*5.

The causal Markov condition gives the conditional independence relationships that are specified by a causal Bayesian network:

A node is independent of its non-descendants (i.e., non-effects) given its parents (i.e., its direct causes).

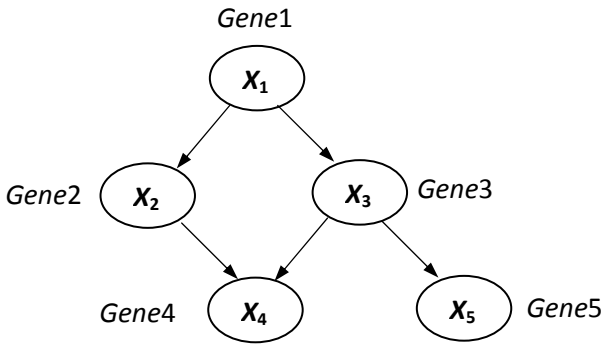

Figure 1. A causal Bayesian network that represents a hypothetical gene-regulation pathway.

The causal Markov condition permits the joint distribution of the n variables in a causal Bayesian network to be factored as follows(Pearl, 1988):

$$P(x_1, x_2, ..., x_n \mid K) = \prod_{i=1}^{n} P(x_i \mid \pi_i, K)$$

where x_i denotes a state of variable X_i, π_i denotes a joint state of the parents of X_i, and K denotes background knowledge.

In this paper Bayesian network models are evaluated (scored) using the following assumptions: (1) discrete variables, (2) Dirichlet prior parameter distributions, (3) multinomial likelihood functions, (4) parameter independence[1], and (5) parameter modularity[2] (Cooper and Herskovits, 1992; Heckerman, et al., 1995), and (6) no missing data. Under these assumptions, a causal Bayesian network B has a marginal likelihood that is given by the following equation (Cooper and Herskovits, 1992; Heckerman, et al., 1995):

$$P(D \mid S, K) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}$$

We propose using this learning algorithm to be used to find the maximum likelihood structure given the dataset and based on the structure, we overlap the latent construct that was developed for SEM. Here we limit the latent construct to be built by SEM if there is a strong linear relationships indicated by a significant latent score. Otherwise, we use latent variable modeling of BN (Friedman, 1997). We have used Banjo (Hartemink, 2010) and Genie (Druzdzel, 2005) to learn the structure with the maximum likelihood given the dataset. Four hours of nine independent run of Banjo was performed to identify the maximum likelihood structure.

3. RESULTS

The dataset we analyzed had 1,352 patients with BPH, mean age of was 67.0 (±7.3, SD, range 45-91) years, prostate volume 50.7 (±43.8) ml, PSA 2.7 (±1.7) mg/dl. Detrusor overactivity (DO) was observed in 27.5% patients, abnormal bladder sensation observed in 17.4%, decreased bladder capacity in 38.8%, decreased bladder compliance in 2.5%, detrusor underactivity in 11.9%, and bladder outlet obstruction (BOO) in 36.4%.

[1] Given network structure S, if $p(S|K)>0$ then (1) $p(\theta_S \mid S, K) = \prod_{i=1}^{n} p(\theta_i \mid S, K)$ (2) for i=1,..,n: $p(\theta_i \mid S, K) = \prod_{j=1}^{n} p(\theta_{ij} \mid S, K)$,

where θ_{ijk} denote the multinomial parameter corresponding to $p(x_i = k|\pi_i = j, K)$ and $\theta_{ij} = \cup_k \theta_{ijk}$ and $\theta_j = \cup_j \theta_{ij}$

[2] Given two network structures S and S' such that $p(S|K)>0$ and $p(S'|K)>0$, if x_i has the same parents in S and S', then for j=1,.., n: $p(\theta_{ij}|S, K) = p(\theta_{ij}| S', K)$

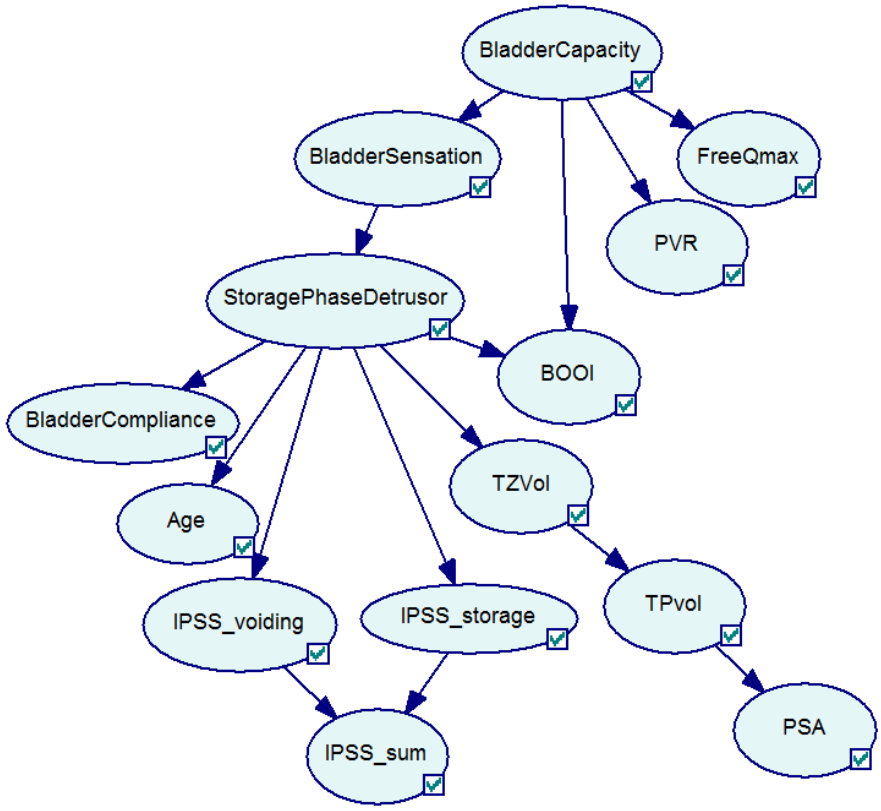

Figure 2. The maximum likelihood BN structure given the dataset of 1,352 patients with BPH

Bayesian network analysis showed that bladder outlet obstruction index (BOOI) is an important factor in predicting both BladderCapacity and StoragePhaseDetrusor.

The Bayesian network showed that (1) predicting outcome variables, when TZVol is known TPVol and PSA do not add value in prediction; (2) if we know StoragePhaseDetrusor, then all input variables do not help more in predicting Bladder Compliace; (3) when BladderSensation is known, BOOI plays important role in predicting BladderCapacity and StoragePhaseDetrusor; (4) if we know TPVol then TZVo and PSA are independent.

The BOOI showed moderate correlation with total prostate volume (TPV) (r=0.37), transitional zone volume (TZV) (r=0.45), DO (r=0.25), and PSA (r=0.32) ($p<0.001$). Multivariate analyses showed that BOOI was the only significant independent predictive factors for all bladder storage parameters including DO (OR 1.027, 95% CI 1.019-1.035), decreased bladder sensation (BS) (1.010, 1.002-1.014), bladder capacity (BCap) (1.019, 1.012-1.026) and bladder compliance (BC) (1.036, 1.018-1.054) ($p<0.01$). Storage symptom score of IPSS was predictor for DO (1.053, 1.009-1.098), BCap (1.163, 1.112-1.217) and BS (1.111, 1.052-1.172) ($p<0.01$). Age, IPSS and BOO were independent predictive factors for at least one of afore mentioned bladder storage parameters ($p<0.05$). Free Qmax, residual volume, PSA, TPV or TZV was not a predictor for any of storage parameters ($p>0.05$).

Based on the expert's knowledge, there are three latent constructs that are being modeled. They are User Self Reported Condition (IPSS_voiding, IPSS_storage, IPSS_sum), Volume Capability of Patient's Bladder (TZVol, TPVol), and Eurodynamic Study Results (StoragePhaseDetrusor, BladderSensation, BladderCapacity, BladderCompliance). User Self Reported Condition and Eurodynamic Study Results did not receive significant latent score and all the variables are discrete, BN was a natural pick to model latent constructs (Friedman, 1997). Volume Capability of Patient's Bladder receive significant latent score and all the variables are continuous, SEM was used to model latent construct (see Figure 3). To combine these latent constructs, we model latent constructs created from SEM to be discretized (e.g., using Normal Distribution, state 0 represents two times of standard deviation less than mean; state 2 represents two times of standard deviation more than mean; and state 1, otherwise) and combine with BN.

The combined data driven model reveals that BOO increases the risk of secondary bladder storage dysfunction in patients with BPH. This suggests that more aggressive treatment for BOO might be recommended.

4. DISCUSSION

We have shown a simple and practical approach in combining BN and SEM in terms of learning causal relationships from data. We are planning to evaluate the predictive performance of the different models, SEM alone, BN alone, and SEM and BN combined.

ACKNOWLEDGMENTS

This study was funded by the Seoul National University Hospital grant.

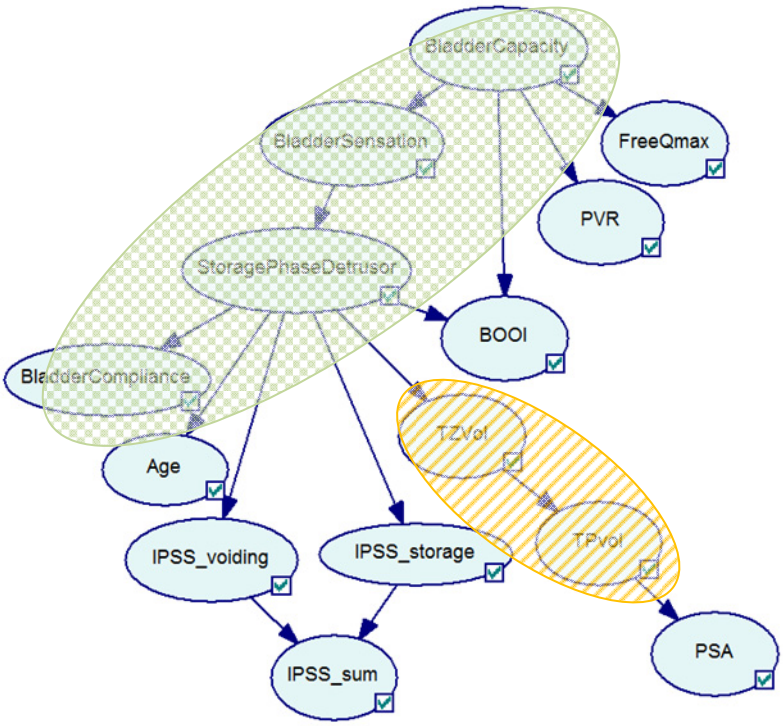

Figure 3. Latent constructs. Shaded variables represents volume capability of patient's bladder and checker boarded variables represents eurodynamic study results.

REFERENCES

Cooper, G.F. and Herskovits, E. (1992) A Bayesian method for the induction of probabilistic networks from data, *Machine Learning*, **9**, 309-347.
Druzdzel, M.J. (2005) Intelligent decision support systems based on SMILE, *Software 2.0*, **2**, 12-33.
Friedman, N. (1997) Learning belief networks in the presence of missing values and hidden variables. *International Conference on Machine Learning*.
Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721-742.
Gupta, S. and Kim, H.W. (2008) Linking structural equation modeling to Bayesian networks: Decision support for customer retention in virtual communities, *European Journal of Operational Research*, **190**, 818–833.
Hartemink, A.J. (2010) Banjo: structure learning of static and dynamic Bayesian networks.
Heckerman, D., Geiger, D. and Chickering, D. (1995) Learning Bayesian networks: The combination of knowledge and statistical data, *Machine Learning*, **20**, 197-243.
Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA.
Spirtes, P., Glymour, C. and Scheines, R. (2000) *Causation, prediction, and search*. MIT Press, Cambridge, MA.