2015 IEEE 28th International Symposium on Computer-Based Medical Systems

Obstructive Sleep Apnea diagnosis: the Bayesian network model revisited

Pedro Pereira Rodrigues
CINTESIS & LIAAD – INESC TEC
Health Information and Decision Sciences department
Faculty of Medicine of the University of Porto
Porto, Portugal
pprodrigues@med.up.pt

Daniela Ferreira Santos, Liliana Leite
CINTESIS
Center for Health Technology and Services Research
Faculty of Medicine of the University of Porto
Porto, Portugal
{ferreiradossantos.daniela,lilianappleite}@gmail.com

Abstract*—Obstructive Sleep Apnea (OSA) is a disease that affects approximately 4% of men and 2% of women worldwide but is still underestimated and underdiagnosed. The standard method for assessing this index, and therefore defining the OSA diagnosis, is polysomnography (PSG). Previous work developed relevant Bayesian network models but those were based only on variables univariatedly associated with the outcome, yielding a bias on the possible knowledge representation of the models. The aim of this work was to develop and validate new Bayesian network decision support models that could be used during sleep consult to assess the need for PSG. Bayesian models were developed using a) expert opinion, b) hill-climbing, c) naïve Bayes and d) TAN structures. Resulting models validity was assessed with in-sample AUC and stratified cross-validation, also comparing with previously published model. Overall, models achieved good discriminative power (AUC>70%) and validity (measures consistently above 70%). Main conclusions are a) the need to integrate a wider range of variables in the final models and b) the support of using Bayesian networks in the diagnosis of obstructive sleep apnea.*

Keywords: diagnosis; Bayesian networks; sleep apnea.

I. INTRODUCTION

Obstructive Sleep Apnea (OSA) is a disease that affects approximately 4% of men and 2% of women worldwide but is still underestimated and underdiagnosed. It is characterized by episodes of breathing cessation (apnea) or reduction in airflow (hypopnea) during sleep for at least 10 seconds as a result of upper airway collapse. The severity of OSA is associated with the apnea-hypopnea index (AHI), documented during sleep, which can be divided into mild (5 ≤ AHI <15), moderate (15 ≤ AHI <30) and severe (AHI ≥ 30). The standard method for assessing this index, and therefore defining the OSA diagnosis, is polysomnography (PSG). However, it is time-consuming, expensive and relatively limited to urban areas which, consequently, originates high waiting lists [1].

In Portugal, patients are referred by the primary care physician to a sleep consult, and then the sleep expert physicians decide the need to perform polysomnography. Although patients are screened by the physicians, based on clinical factors, the specificity of the entire process is rather low (48% of PSG performed in 2010, in our sleep laboratory, resulted negative for OSA) which, together with the limited availability of the service, yields long waiting lists both for consultation and to perform PSG. This setting clearly presents the need for a valuable decision support tool that can a) help primary care physicians referring patients to sleep consult, and b) help sleep experts decide who needs PSG.

Recently, many studies have been conducted to apply machine learning methods for medical knowledge discovery, including sleep medicine, consisting on an alternative to traditional statistic in defining diagnostic models. These models can now be generated by artificial intelligence, using decision trees, neural networks, support vector machines and Bayesian networks (BN) [4]-[7].

Previous work [1] studied several factors and identified six as associated with OSA diagnosis (body mass index, neck and abdominal circumferences, gender, witnessed apneas and alcohol consumption before sleep), but the studied sample (patients already referenced to sleep consult) made the results not generalizable to use in primary care, where it would be needed the most. For example, snoring could not be assessed since it was prevalent (100%) in that sample. Furthermore, the Bayesian networks were developed based only on variables univariatedly associated with the outcome, yielding a bias on the possible knowledge representation of the models.

The aim of this work is to revisit the obstructive sleep apnea cohort, developing and validating new Bayesian network-based decision support system that can be used in the future during sleep consult to assess the need for PSG.

II. METHODS

This study was designed according to the common characteristics of validation of a diagnostic test.

A. Patients

This study included patients referred to perform PSG at the Sleep Laboratory of Vila Nova de Gaia/Espinho Hospital Center, Portugal. In this study we focus on the derivation sample created with the patients that realized PSG between December of 2011 to February of 2012. All adults, older than 18 years, referred by the physicians with suspected OSA were included.

In case of duplicate studies from the same patient, the one with best sleep efficiency was selected. Patients with suspicion of another disorder than OSA, patients already diagnosed (therapeutic studies), and patients with severe lung disease or neurological condition that somehow affects the respiratory function, such as neuromuscular diseases, were excluded.

2372-9198/15 $3.00 © 2015 IEEE
DOI 10.1109/CBMS.2015.47

B. Variables

Thirty-three variables have been selected and collected in the previous study [1], including demographic information, clinical history, physical examination and co-morbidities information: Gender, Race, Age, Snoring, Witnessed apneas, Gasping/Choking, Motor Vehicle Crashes, Refreshing Sleep, Humor alterations, Nocturia, Restless Sleep, Decreased libido, Morning headaches, Alcohol before sleep, Smoking, Sedative use, Epworth sleepiness scale (ESS), Concentration decrease, BMI, NC, AC, Craniofacial and upper airway abnormalities, Atrial fibrillation, Stroke, Myocardial infarction, Pulmonary hypertension, Congestive heart failure, Diabetes, Metabolic Syndrome, Renal failure, Hypothyroidism, Gastroesophageal reflux disease, Hypertension.

From the same study, six variables were found significantly associated with the outcome: obesity (BMI > 30), increased neck and abdominal circumferences (using literature-based thresholds), gender, witnessed apneas and alcohol consumption before sleep. This subset of variables will be referred to as the "selected" variables.

C. Data collection

Clinical information was collected prospectively during consultation, 3 months before PSG. For the PSG, the parameters, settings, filters, technical specifications, sleep stage, event scoring and final results were applied according to the American Academy of Sleep Medicine rules of 2007. For this study, the outcome measure was the clinical diagnosis supported on PSG results, dichotomized into normal or OSA (mild, moderate and severe).

D. Bayesian network models

Models were built using either all data variables (discretized when necessary) from the study, or only the significant variables identified in previous work [1] with univariate logistic regression.

Cases with missing data were removed for structure learning but included for parameter fitting.

Continuous variables were categorized according to the following definitions/breaks, rounded from quantiles in the data or from the literature:

- Age: breaks 40, 50, 60, 70;
- Weight: breaks 70, 80, 90, 100;
- Height: breaks 1.60, 1.65, 1.70, 1.75, 1.80;
- BMI: breaks 25, 27.5, 30, 32.5, 35;
- Obesity: normal or obese (threshold: BMI > 30);
- NC: breaks 38, 41, 44;
- AC: breaks 97, 103, 111;
- NC Increased: breaks male 42cm, female 37cm [2];
- AC Increased: breaks male 94cm, female 80cm;
- ESS: breaks 3, 6, 9, 12, 15, 18, 21.

For exploration purposes, Bayesian networks were built using a hill-climbing strategy [3]. For classification purposes, the result of PSG (normal or OSA) was defined as the class attribute to construct the models.

The Bayesian network classifiers used to build the models were Naïve Bayes (NB) and Tree Augmented Naïve Bayes (TAN), given their previous good results in other clinical domains [4-7].

Also compared were the hill-climbing model defined previously [1] and an expert-defined causal model using selected variables.

E. Evaluation methodology

Receiver Operating Characteristic (ROC) curve analysis was performed to determine in-sample area under the curve (AUC). The achieved models were then evaluated with sensitivity, specificity, precision (positive and negative predictive values) and AUC estimates, using 10 times stratified 4-fold cross-validation.

A significance level of 5% was used for confidence interval definition.

We used R statistical package [8] to learn and evaluate the models, using packages *bnlearn* [3] and *gRain* [9] for structure learning and conditional probability tables fitting, *pROC* [10] for ROC curves assessment, and *epitools* [11] for simple odds ratio computation.

III. Results

The initial cohort considered 113 patients for inclusion, 27 of whom were excluded (18 with other pathology, 3 children, 3 with no information, 2 already diagnosed, and 1 with neuromuscular disease). The final cohort used to fit the models had 86 patients, 69 (80%) of which were male and the global mean age was 56 years. Forty one patients (48%) had normal result while 45 patients were diagnosed with OSA (52%): 17 (37%) mild, 15 (33%) moderate and 13 (30%) severe.

A total of eight models were evaluated and compared, differing on the algorithm used to model the structure – Naive Bayes (NB), Tree-Augmented Naive Bayes (TAN), hill-climbing (HC) or expert-defined (Causal) - and the número of predictive factors included (33 or 6): NB33, NB6, TAN33, TAN6, HC33, HC6, Causal and HC2014 (for the structure model learned in [1]).

A. Bayesian network qualitative models

Fig. 1 to 4 present the graphical representations for some of the evaluated networks (HC2014, HC6, TAN6, Causal, and HC33), being mostly useful for knowledge discovery and factors interaction inspection.

From the graph representations in Fig. 1 we can observe that HC2014 and HC6 are mostly equivalent, having only a slight change in the ascendant node for witnessed apneas (from gender to OSA outcome) with the latter better representing the expected association.

Also interesting enough is to note the conditional independences risen from the algorithm: in both models, increased NC and AC are conditionally independent given knowledge on obesity; on HC2014, alcohol consumption and witnessed apneas were conditionally independent given gender, while on HC6 witnessed apneas is directly dependent on the outcome.

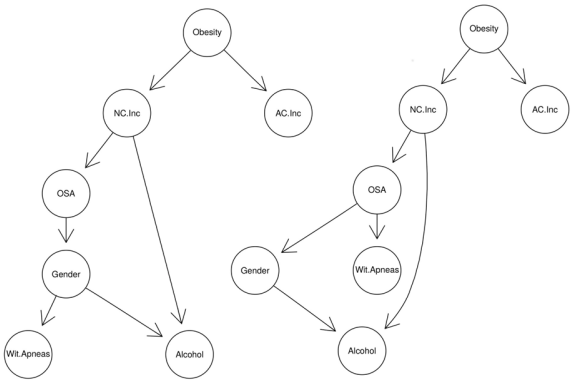

Figure 1. HC2014 and HC6 : Hill-climbing network developed in [1] (left) and with current procedure with selected variables (right).

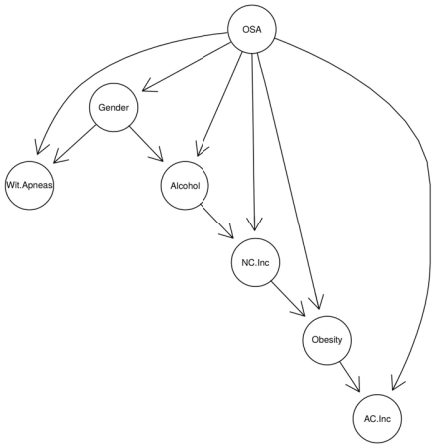

Figure 2. TAN6: Tree-Augmented Naive Bayes with selected variables.

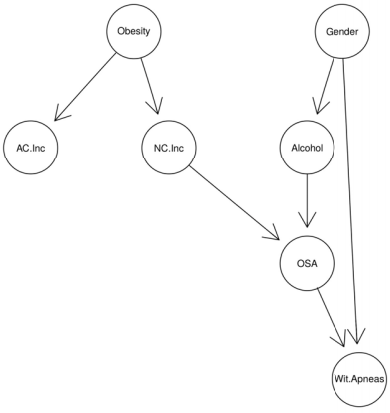

Figure 3. Causal: Expert-defined network with selected variables.

Given the classification target, Fig. 2 presents different associations expressed by TAN6, starting by assuming all factors directly dependent from the outcome (TAN structure assumption) and possibly allowing an extra ascendant for each node. In this model, interesting dependences rise: male gender influences witnessed apneas (OR=6.89, CI95%=[2.19,25.01]) and alcohol consumption before sleep (OR=16.01, CI95%=[2.98,401.29]), while alcohol consumption influences increased neck circumference (OR=2.67, CI95%=[1.02,7.46]), which in turn is associated with obesity (OR=14.66, CI95%=[4.77,53.74]), which slightly influences increased abdominal circumference (OR=4.62, CI95%=[0.56,182.60]).

Using some expert knowledge on the selected variables considered, the Causal model - designed by hand - resulted in the network presented in Fig. 3. Here, the main points are that most of the characteristics exposed by the hill-climbing methods make sense from the expert point of view:

- the conditional independence of AC and NC given obesity,
- the gender influence on both alcohol consumption before sleep and witnessed apneas, and
- witnessed apneas as an expression (observation) of the outcome.

In order to better study the interaction of different factors, HC33 (Fig. 4) presents a complete exposition of the network of factors, also highlighting some factors which were considered not relevant for the model: atrial fibrillation (OR=0.44, CI95%=[0.09,9.10]), smoking (OR=0.51, CI95%=[0.20,1.67]), all age categories (e.g. age>70, OR=1.26, CI95%=[0.47,4.82]), all height categories (e.g. height>1.80, OR=1.33, CI95%[0.37,9.27]) and the Epworth Sleepiness Scale (perhaps the strangest factor not be found relevant for the model; e,g, ESS=]18,24], OR=0.97, CI95%=[0.21,13.3]).

Other interesting sub-structures include:

- weight/size substructure, where weight measurements and computed obesity associated with stroke (OR=2.08, CI95%=[0.45,28.60]);
- outcome observation substructure, with witnessed apneas and concentration decrease following the outcome;
- sleep effects on mental health substructure, with repairing sleep associated with headaches, associated with gender, concentration decrease and humor alterations, humor alterations with sedative use and libido alterations;
- co-morbidities substructure, with neck circumference associated with arterial hypertension, diabetes and infarction, and congestive heart failure.

B. Bayesian network quantitative in-sample analysis

For a quantitative analysis, Fig. 5 presents the in-sample ROC curves for all models. As expected, increasing model complexity enhances the in-sample AUC (e.g. NB 85.6% vs 80.4%, TAN 99.5% vs 79.9%, HC 83.3% vs 79.3%). Causal and previous HC model presented lower AUC (76.7%) but, globally, all models presented good discriminative power towards the outcome.

C. Bayesian network generalizable cross-validation

In order to assess the ability of the models to generalized beyond the derivation cohort, cross-validation was endured. Tab. I presents the result of the 10-times-repeated stratified 4-fold cross-validation.

From the exposed results, HC33 rises as the best classification model (using the 50% threshold classification cutoff to predict the outcome) only loosing in terms of specificity and positive predictive value. The lower results for these measures could possibly be overcome with a threshold study to find the best cutoff for the classification rule.

IV. Conclusions and Future Directions

Previous work on Bayesian networks for obstructive sleep apnea was biased by a selection of significant variables that reduced: a) the interpretability of the overall models (more complex network present interesting subnetworks to be further analyzed by the clinical experts), b) the discriminative power of the models (much better AUC were computed for models with all 33 variables), and c) predictive quality (better accuracy, sensitivity and precision cross-validation estimates).

More studies are required to better fit a clinical decision support model into clinical practice, especially if we consider anticipating the support into primary care, but this study clarified the need to integrate a much wider set of clinical variables into a diagnostic model for obstructive sleep apnea, nevertheless reinforcing the advantages of Bayesian network models for the task at hands.

References

[1] L. Leite, C. Costa-Santos, and P. P. Rodrigues, "Can We Avoid Unnecessary Polysomnographies in the Diagnosis of Obstructive Sleep Apnea? A Bayesian Network Decision Support Tool," in 2014 IEEE 27th International Symposium on Computer-Based Medical Systems, 2014, pp. 28–33.

[2] R. J. Davies, N. J. Ali, and J. R. Stradling, "Neck circumference and other clinical features in the diagnosis of the obstructive sleep apnoea syndrome.," *Thorax*, vol. 47, no. 2, pp. 101–5, Feb. 1992.

[3] M. Scutari, "Learning Bayesian Networks with the bnlearn R Package," *J. Stat. Softw.*, vol. 35, p. 22, 2010.

[4] C. C. Dias, C. Granja, A. Costa-Pereira, J. Gama, and P. P. Rodrigues, "Using Probabilistic Graphical Models to Enhance the Prognosis of Health-Related Quality of Life in Adult Survivors of Critical Illness," in *2014 IEEE 27th International Symposium on Computer-Based Medical Systems*, 2014, pp. 56–61.

[5] C. A. M. Schurink, P. J. F. Lucas, I. M. Hoepelman, and M. J. M. Bonten, "Computer-assisted decision support for the diagnosis and treatment of infectious diseases in intensive care units.," *Lancet Infect. Dis.*, vol. 5, no. 5, pp. 305–12, May 2005.

[6] C. a M. Schurink, S. Visscher, P. J. F. Lucas, H. J. van Leeuwen, E. Buskens, R. G. Hoff, A. I. M. Hoepelman, and M. J. M. Bonten, "A Bayesian decision-support system for diagnosing ventilator-associated pneumonia.," *Intensive Care Med.*, vol. 33, no. 8, pp. 1379–86, Aug. 2007.

[7] G. Sakellaropoulos and G. Nikiforidis, "Prognostic performance of two expert systems based on Bayesian belief networks," *Decis. Support Syst.*, vol. 27, no. 4, pp. 431–442, Jan. 2000.

[8] R Core Team, "R: A Language and Environment for Statistical Computing." Vienna, Austria, 2013.

[9] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller, "pROC: an open-source package for R and S+ to analyze and compare ROC curves," *BMC Bioinformatics*, vol. 12, p. 77, 2011.

[10] S. Højsgaard, "Graphical independence networks with the gRain package for R," *J. Stat. Softw.*, vol. 46, no. 10, 2012.

[11] T. J. Aragon, "epitools: Epidemiology Tools.", 2012.

TABLE I. Validity assessment averaged from 10 times 4-fold cross-validation

Model	Validity assessment measures (%, CI95%)					
	Accuracy	***Sensitivity***	***Specificity***	***Precision (positive)***	***Precision (negative)***	***AUC***
NB33	61.09 [57.45,64.72]	58.92% [54.13,63.71]	63.45% [58.4,68.51]	64.55% [60.82,68.27]	58.92% [54.88,62.95]	69.53% [66.27,72.78]
NB6	67.68 [65.27,70.09]	65.08% [61.31,68.85]	70.55% [66.8,74.29]	71.41% [68.5,74.31]	65.33% [62.77,67.89]	77.84% [75.07,80.6]
TAN33	65.21 [62.03,68.39]	65.02% [60.58,69.46]	65.41% [60.87,69.95]	67.92% [64.54,71.29]	63.65% [60.02,67.28]	73.37% [69.88,76.86]
TAN6	64.53 [60.96,68.1]	62.5% [57.95,67.05]	66.75% [60.29,73.21]	69.24% [64.65,73.84]	61.75% [58.42,65.09]	72.25% [68.5,76.01]
HC33	**72.41 [69.19,75.62]**	**79.94% [75.77,84.11]**	64.14% [58.46,69.81]	72.04% [68.68,75.41]	**75.74% [71.54,79.95]**	**79.9% [77.04,82.76]**
HC6	68.03 [65.92,70.14]	78.69% [73.14,84.25]	56.34% [50.91,61.77]	68.2% [64.93,71.47]	73.85% [69.72,77.98]	77.44% [75.21,79.66]
HC2014	69.33 [66.1,72.56]	58.54% [53.84,63.25]	**81.16% [76.59,85.73]**	**78.16% [74.00,82.33]**	65.14% [61.73,68.54]	75.81% [72.27,79.35]
Causal	66.41 [63.42,69.4]	61.38% [57.06,65.7]	71.93% [67.37,76.49]	71.48% [67.84,75.12]	63.46% [60.22,66.7]	72.65% [68.88,76.41]

Highlighted values are the best for each quality measure, being also significantly better than at least one other model.

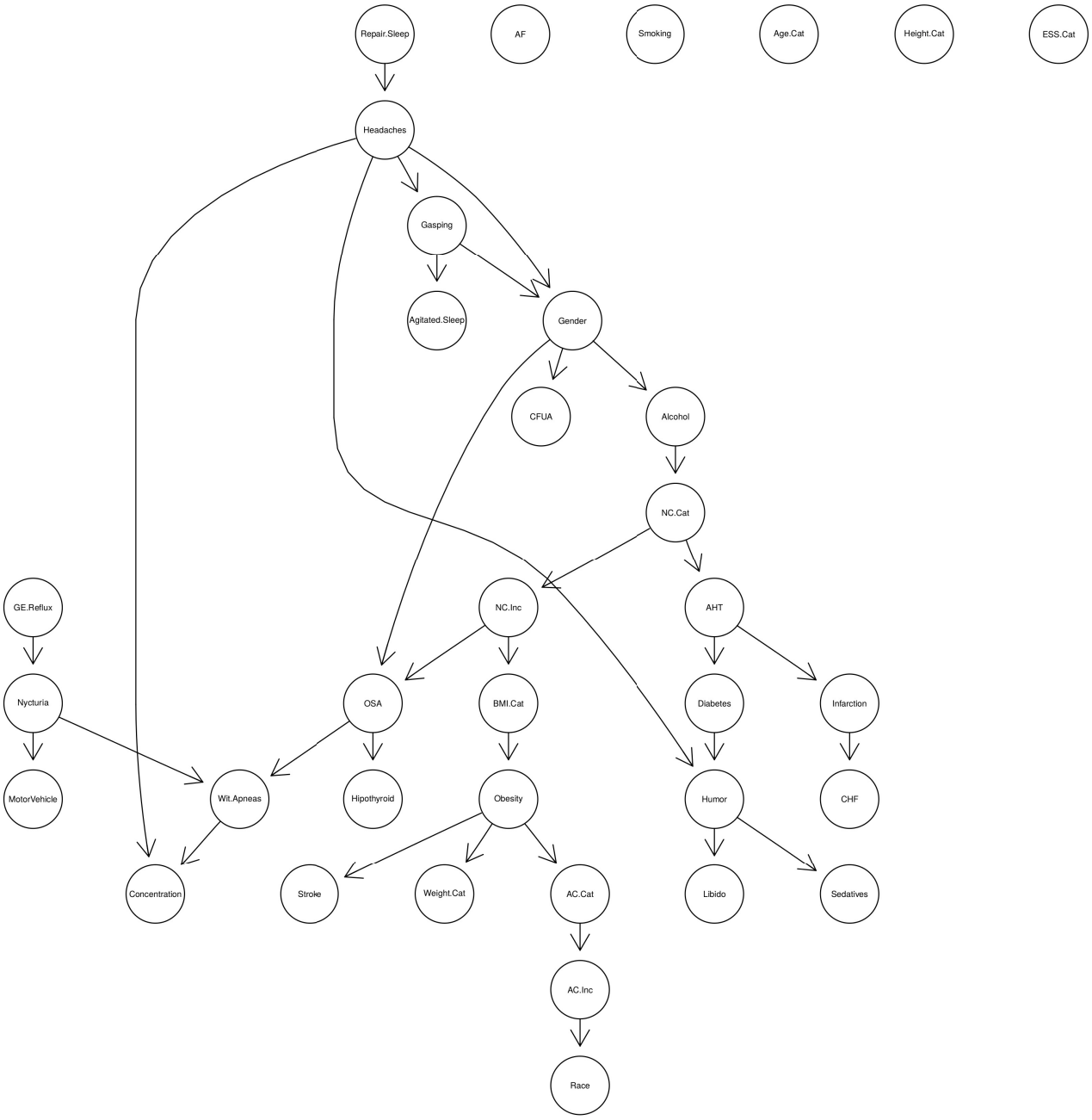

Figure 4. HC33: Hill-climbing network using all 33 variables.

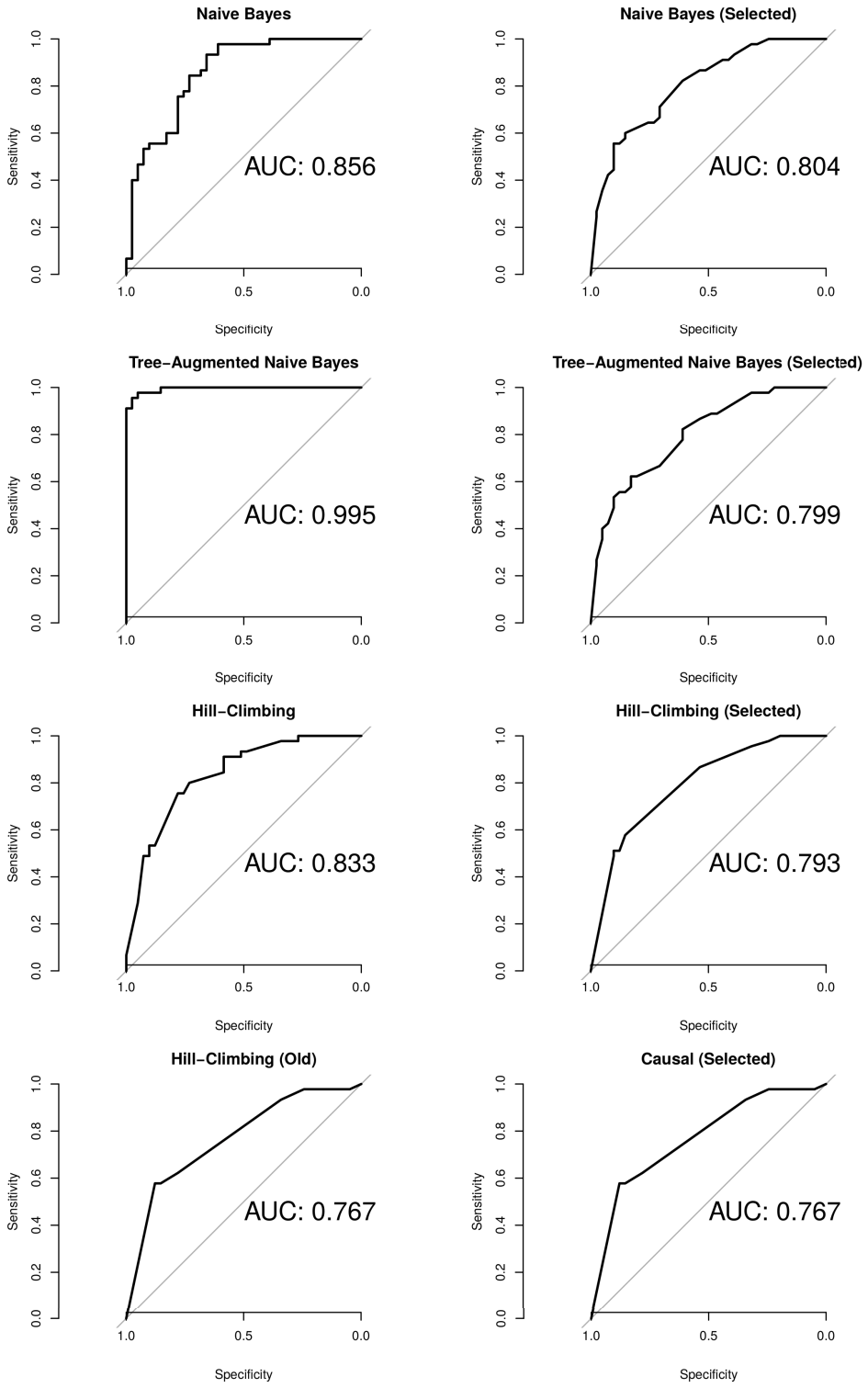

Figure 5. In-sample ROC curves for all the studied Bayesian networks: Naive Bayes (with all or selected variables), Tree-Augmented Naive Bayes (with all or selected variables), hill-climbing (with all or selected variables), hill-climbing model from [1], and a expert-defined causal model with selected variables.