Comparison of Bayesian Network and Decision Tree Methods for Predicting Access to the Renal Transplant Waiting List

Sahar BAYAT [a,1], Marc CUGGIA [a], Delphine ROSSILLE [a], Michèle KESSLER [b],
Luc FRIMAT [c]

[a] *INSERM U936, Université Rennes 1, IFR 140, Rennes, France*
[b] *Service de Néphrologie, Hôpitaux du Brabois, Nancy, France*
[c] *EA 4003, Nancy Université, France*

Abstract. The study compares the effectiveness of Bayesian networks versus Decision Trees for predicting access to renal transplant waiting list in a French healthcare network. The data set consisted in 809 patients starting renal replacement therapy. The data were randomly divided into a training set (90%) and a validation set (10%). Bayesian network and CART decision tree were built on the training set. Their predictive performances were compared on the validation set. The *age* variable was found to be the most important factor in both models. Both models were highly sensitive and specific: sensitivity 90.0% (95%CI: 76.8–100), specificity 96.7% (95%CI: 92.2–100). Moreover, the models were complementary since the Bayesian network provided a global view of the variables' associations while the decision tree was more easily interpretable by physicians. These approaches provide insights on the current care process. This knowledge could be used for optimizing the healthcare process.

Keywords. decision support, decision tree, Bayesian network, healthcare network, renal transplant waiting list

1. Introduction

Incidence and prevalence of End-Stage Renal Disease (ESRD) requiring Renal Replacement Therapy (RRT) (hemodialysis, peritoneal dialysis or kidney transplantation) continue to increase. Kidney transplantation is associated with longer survival and lower long-term cost. But, given the graft shortage, transplantation is not commonly achievable. On the other hand, the selection criteria of the potential transplant recipient diverge from one centre to another. Ideally, placement on the waiting list should be solely based on medical factors. Previous studies showed that female gender, old age, distance from transplantation centre and private ownership of dialysis facilities have been associated with poor access to kidney transplant waiting lists [1, 2]. Previously [3, 4] we showed that access to the renal transplant waiting list in NEPHROLOR, a French healthcare network, is primarily associated with age and medical factors. The factors were identified by conventional statistical methods and

---

[1] Corresponding Author: Sahar Bayat, Département d'Information Médicale (DIM), Hôpital Pontchaillou, 2 rue Henri Le Guilloux, 35033 Rennes, France; E-mail: sahar.bayat@chu-rennes.fr.

Bayesian networks and both methods gave similar results [4]. The purpose of the present study is to compare the performance of Bayesian networks and decision trees for predicting registration on the renal transplant waiting list in NEPHROLOR network.

2. Material and Method

Organization of care for ESRD in Lorraine region: The NEPHROLOR healthcare network combines the public and private for-profit facilities operating dialysis units in Lorraine. Renal transplantation is performed within the University nephrology department which is the only transplant centre of the region. The patients are registered on the national waiting list by the referent neurologist of the transplantation centre, a national score is used to attribute grafts.

Study population: Since June 1997, all adult patients living in Lorraine and starting RRT in a NEPHROLOR facility have been progressively registered in the NEPHROLOR database [1]. We included the incident patients between July 1, 1997 and June 30, 2003. Body Mass Index (BMI) and Albumin were not exhaustively collected the first few years so patients with missing data were not included. To identify patients placed on the waiting list, the list of patients registered on the national waiting list was extracted from the CRISTAL database of *Agence de la Biomédecine* [5] and from the list of transplanted patients from the database of the transplantation center of Lorraine.

Data collection: Three categories of variables were studied. The first category included social and demographic data: *age (<45, 45–55, 55–65, ≥65)*, *sex* and *distance between the patient's residence and the transplantation center*. As the Lorraine nephrologists very exceptionally register patients older than 80 on the waiting list, we decided to exclude these patients from analysis. In contrast to other countries, French legal regulations prohibit considering ethnic differences. Moreover, descriptions of income or education are not available. The second category included clinical and biological data at first RRT: existence of *diabetes*, *cardiovascular disease*, *respiratory disease*, *hepatic disease*, *psychiatric disorder*, *physical impairment* and *past history of malignancy*. *BMI* was categorized in < 20 kg/m², 20–24.99, ≥ 25, *hemoglobin* in < 11 g/dl, ≥ 11 and serum *albumin* in < 3 g/dl, 3–3.49, ≥ 3.5 (following medical guidelines). The third category included data related to medical follow up in the NEPHROLOR network: *public or private ownership of nephrology facility* where the first RRT was performed. We also took into account the effect of a medical *follow-up in the transplantation center versus 12 other facilities without transplantation* [4].

Statistical analysis: The data set was split into two independent samples. The training set consisted in 90% of randomly selected cases of the data set and the validation set of the 10% remaining cases [6]. We used the Chi2 test to ensure that the characteristics of the two sets were not significantly different. The Bayesian network and decision tree approaches were applied to the training set to predict registration on the renal transplant waiting list. The predictive performances of both methods were evaluated on the validation set in terms of sensitivity, specificity and positive predictive values (with 95% confidence intervals). We computed the McNemar test to check if the difference between the two models was significant.

- A Bayesian network (BN) describes a system by specifying relationships of conditional dependence between its variables [7]. The conditional dependences are represented by a directed acyclic graph, in which, each node

represents a variable and the arcs represent relationships between variables. The direction of the arcs does not necessarily mean a cause-effect relationship. This graph together with a joint probability distribution for the variables creates a model that can be used to make the inferences. BN analysis was performed with BAYESIALAB Software (version 3.3; Bayesia SA).

- A decision tree is a tree-structured classifier built by partitioning the data set into homogeneous classes. The root node is split into child nodes by selecting the variable that best classifies the samples according to a split criterion. The splitting continues on the child nodes until stopping criteria are met. The CART method [8], chosen for this analysis, uses binary recursive partitioning. In this study the GINI split criterion is used and the stopping criterion is the maximum tree depth allowed of 5. Decision tree analysis was carried out with SPSS Clementine version 10.1.

3. Results

Patients' characteristics: 809 patients were included in the study. Their mean age was 62.1 ± 14.2 years; 59.6% were male, 44.2% had cardiovascular disease and 34.5% diabetes, 19.5% physical impairment and 48 5.9% psychiatric disease. 26.2% were registered on the transplant waiting list. The training set included 729 patients and the validation set 80 patients. Chi2 test showed no significant difference between the characteristics of the two sets.

Bayesian Network: The graph of the Bayesian network built on the training set (Figure 1) shows that the probability of being registered on the waiting list (target node) is directly associated with 6 variables: *age*, *cardiovascular disease*, *diabetes*, *respiratory disease, albumin* and *follow up in transplantation center*. As the brightness of square represented in a node is proportional to the contribution of this node in the characterization of the target, *age* is the most important variable. Compared to *age*, relative contributions of the variables for characterizing the target node were: *cardiovascular disease*: 0.3582; *diabetes*: 0.2287; *albumin:* 0.1221; *respiratory disease*: 0.0815; *follow up in transplantation center:* 0.0749; *physical impairment:* 0.0413; *BMI:* 0.0138; *distance from transplantation center:* 0.0108; *ownership of facility*: 0.0037. This network was run on the validation set and a probability threshold of 50% was selected to classify the patients as registered or non registered. The sensitivity, specificity and positive predictive value of the model for predicting registration on the waiting list were respectively 90.0 % (95%CI: 76.8–100), 96.7% (95%CI: 92.2–100) and 90.0% (95%CI: 76.8–100). Of the 80 patients of the validation set 20 were registered on the waiting list and 60 were not. The Bayesian network correctly predicted 18 out of 20 registrations and 58 out of 60 non registrations.

Decision Tree (Figure 2). Age emerged as the strongest discriminating factor for predicting registration on the waiting list. The CART decision tree correctly predicted 18 out of 20 registrations and 58 out of 60 non registrations of validation set. The sensitivity, specificity and positive predictive value of the model were respectively 90.0 % (95%CI: 76.8–100), 96.7% (95%CI: 92.2–100) and 90.0% (95%CI: 76.8–100). The McNemar's test did not show a significant difference between the Bayesian model and CART predictions. The predictions of the two models were discordant for two patients. The Kappa index of concordance was 0.93.
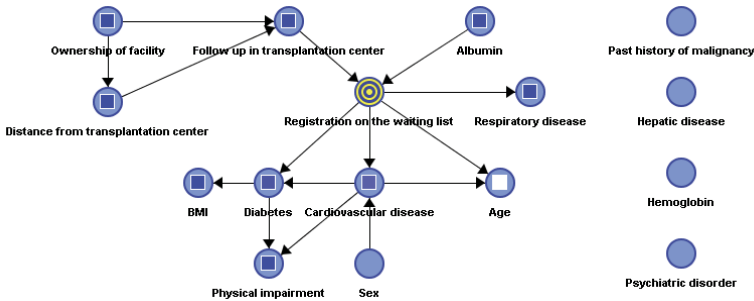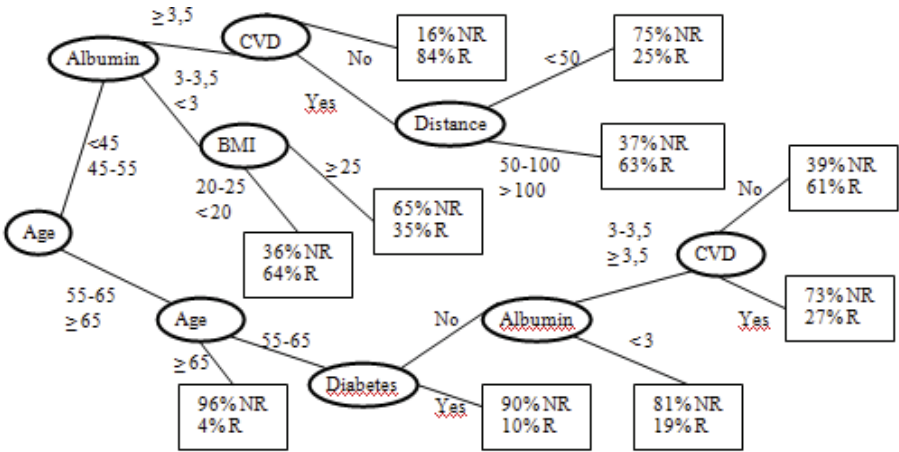
Figure 1. Bayesian Network

Figure 2. CART decision tree (NR: Non Registered, R: Registered, CVD: CardioVascular Disease)

4. Discussion

The present study shows the performance of the Bayesian network and CART decision tree approaches for predicting access to renal transplant waiting list in NEPHROLOR network. Both models have very high predictive performances. Moreover, in contrast with other studies [9, 10] the models were highly concordant. *Age* is found to be the most important variable for predicting access to the waiting list for both models. The models show however some differences in the selected discriminating factors and their relationships to the *registration* factor. The other variables directly related to *registration* in Bayesian network and also selected by CART decision tree are *cardiovascular disease*, *diabetes* and *albumin*. In addition the Bayesian network associates *respiratory disease* and *follow up in transplantation center* directly to *registration* while the CART decision tree uses *BMI* and *distance from transplantation center*.

Janssens et al. [11] reported that Bayesian networks outperform decision trees. In this study the models had similar high predictive performances. Janssens et al. also reported that Bayesian networks are better suited to capture the complexity of the underlying decision-making process, taking into account the many (inter)dependencies

among the variables. In this study, the Bayesian network visualizes the relationships between *distance from transplantation center*, *ownership of facility* and *follow up in transplantation center*. This association can be explained by the fact that the transplantation center is a public facility and the patients living near the transplantation center have been followed up, in general, in this public unit. This association shows that even if Bayesian network and CART models use two different variables: *follow up in transplantation center* and *distance from transplantation centre*, these variables are in fact strongly related. However, Bayesian networks may link more variables in complex, direct and indirect ways, making interpretation more problematic. For example, in this study, *diabetes* and *cardiovascular disease* are associated with the target node and with each other. They are also connected to two other variables which influence indirectly *registration on the waiting list: physical impairment* and *BMI*. In contrast, decision rules can easily be derived from decision trees and provide a simpler and more direct interpretation tool for physicians. For this reason decision trees are popular within the medical field.

5. Conclusion

The present study, based on the French NEPHROLOR database, shows that both Bayesian networks and decision trees can be used to predict access to the renal transplant waiting list with high accuracy. It was stressed out that the models are complementary. As the Bayesian network provides a global view of variables associations, the decision tree is more easily interpretable. These approaches provide insights on the current care process. This knowledge could be used for formalizing and optimizing the healthcare process.

References

[1] Kessler, M. et al. (2003) Impact of nephrology referral on early and midterm outcomes in ESRD: EPidemiologie de l'Insuffisance REnale chronique terminale en Lorraine (EPIREL): Results of a 2-year, prospective, community-based study. *American Journal of Kidney Disease* 42(3):474–485.
[2] Oniscu, G.C. et al. (2003) Equity of access to renal transplant waiting list and renal transplantation in Scotland: Cohort study. *British Medical Journal* 327(7426):1261.
[3] Bayat, S. et al. (2006) Medical and non-medical determinants of access to renal transplant waiting list in a French community-based network of care. *Nephrology Dialysis Transplantation* 21(10):2900–2907.
[4] Bayat, S. et al. (2008) Modelling access to renal transplantation waiting list in a French healthcare network using Bayesian method. *Studies in Health Technology and Informatics* 136:605–610.
[5] Strang, W.N. et al. (2005) The French organ transplant data system. *Studies in Health Technology and Informatics* 116:77–82.
[6] Guyon, I. (1997) A scaling law for the validation-set training-set size ratio. AT&T Bell Laboratories, Berkeley, California.
[7] Goldszmidt, M., Friedman, N. (1998) *Learning Bayesian Network from Data*. SRI International, Menlo Park.
[8] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984) *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA.
[9] Chapman, W.W. et al. (2001) A comparison of classification algorithms to automatically identify chest X-ray reports that support pneumonia. *Journal of Biomedical Informatics* 34(1):4–14.
[10] Pang, B.C. et al. (2007) Hybrid outcome prediction model for severe traumatic brain injury. *Journal of Neurotrauma* 24(1):136–146.
[11] Janssens, D. et al. (2004) Improving the performance of a multi-agent rule-based model for activity pattern decisions using Bayesian networks. *Transportation Research Record: Journal of the Transportation Research Board* 1984:75–83.