Is Random Survival Forest an Alternative to Cox Proportional Model on Predicting Cardiovascular Disease?

Fen Miao[1], Yun-Peng Cai[1], Yuan-Ting Zhang[1,2], and Chun-Yue Li[1]

[1] Key Laboratory for Health Informatics of the Chinese Academy of Sciences (HICAS), Shenzhen Institutes of Advanced Technology. Shenzhen, China, 518055

[2] Joint Research Centre for Biomedical Engineering, Department of Electronic Engineering, Chinese University of Hong Kong, H.K.

Abstract— **Random survival forest (RSF), a non-parametric and non-linear approach for survival analysis, has been used in several risk models and presented to be superior to traditional Cox proportional model. Anyway, can RSF replace Cox proportional model on predicting cardiovascular disease? In this paper, we evaluate the performance of RSF by comparing it with Cox in terms of discrimination ability, ability to identify non-linear effects and ability to identify important predictors that can discriminate survival function. Two databases are studied, including heart failure population database and cardiac arrhythmias database. We take 1-year mortality after cardiac arrhythmias prediction as an example for comparison between Cox and RSF based model. The results show that RSF improved discrimination performance greatly than Cox with an out-of-bag C-statistics of 0.812 (while 0.736 for Cox based model). In addition, RSF can automatically identify non-linear effects of all variables but Cox cannot. However, RSF is inferior in identifying predictors with less ratio of population due to its insensitivity to noise. Therefore, RSF cannot replace Cox in current status and should be studied further.**

Keywords— **Random survival forest, Cox proportional hazard model, Risk prediction.**

I. INTRODUCTION

Cox proportional hazard regression (CPH) [1] is most popular in risk prediction model and it has been used in several risk models for predicting cardiovascular diseases [2-3]. However, it suffers a lot from high variance and poor performance as solving the model is very complex, especially for those involving multiple variables and nonlinear effects [4]. Random survival forests (RSF) modeling, a direct extension of random forest for survival analysis is proposed in [5] to handle the above difficulties by automatically assessing the complex effects and interactions among all variables from objective view, that is, following the inherent relationship between any factors and the predictive result. Thus, it has been used in several risk models for different kinds of diseases such as heart failure and breast cancer [6-7]. The results showed that the RSF model could identify complex interactions among multiple variables and performed slightly better than traditional CPH model.

In this paper, we evaluate the performance of RSF for predicting cardiovascular disease from different aspects in the application of predicting 1-year mortality after cardiac arrhythmias. Two risk models are developed, one based on RSF and another based on CPH. We compare the performance of the two models from three aspects, including discrimination ability in terms of out-of-bag C-statistics, ability to identify non-linear effects of multiple variables and ability to identify all important predictors to discriminate survival function. The results show that RSF takes advantage in the first two abilities while insufficient in the last and thus should be improved further to replace CPH.

II. METHODS

A. Study Population

Our study is based on the public MIMIC II (Multiparameter Intelligent Monitoring in Intensive Care) clinical database [8-9], which contain comprehensive clinical data including results of laboratory tests, medications, ICD9 diagnoses, admitting notes, discharge summaries, and more obtained from hospital medical information systems, for 32,536 ICU patients. We defined the patients with cardiac arrhythmias according to ninth revision of the international classification of diseases (ICD9) adopted in the database. 10,488 patients with cardiac arrhythmias were extracted to establish the predictive model, during which 3,452 deaths occurred in hospital or after discharge over 1-year follow-up period for each patient.

Potential clinical variables previously reported to be associated with mortality were evaluated in our study. The following 40 variables were assessed for prognostic value: demographics including age, sex and BMI, clinical variables such as arrhythmias type (CA, VF, VT, AF and other slow arrhythmias), valvular heart diseases, renal failure and CHF, laboratory variables with missing value smaller than 20%, including glucose, NA, K, SCR, BUN, RBC, WBC, PT, PTT, INR, BR, AST, ALT and CKPK, and anti-arrhythmic agents including class I, class II, class III, class IV and class V agents (as listed in Fig.1).

I. Lacković and D. Vasić (eds.), *6th European Conference of the International Federation for Medical and Biological Engineering*, IFMBE Proceedings 45, DOI: 10.1007/978-3-319-11128-5_184

B. Statistical Analysis with RSF

In our study, two risk models for predicting 1-year mortality after cardiac arrhythmias will be developed, one with RSF and another with CPH.

RSF is implemented in our study to establish prediction models in the following ways:

1. Draw B bootstrap samples from the original data. Note that each bootstrap sample excludes on average 37% of the data, called out-of-bag data (OOB data). B equals to 1000 in our study.
2. Grow a survival tree based on all 40 variables for each bootstrap sample to develop a comprehensive model. At each node of the tree, randomly select p candidate variables. In our study, p is set to be the square root of the total number of variables, i.e., 6. The node is split using the candidate variable that maximizes survival difference between daughter nodes.
3. Grow the tree to full size under the constraint that a terminal node should have no less than $d_0 > 0$ unique deaths. In our study, d_0 is set to be 3.
4. Select predictive variables by filtering on the basis of minimal depth of a maximum subtree, the largest subtree whose root node splits on the variable. Minimum depth equals to the shortest distance from the tree trunk to the branch level of the maximal subtree. The smaller the minimal depth. The more impact variable has on prediction
5. Using OOB data, prediction error is calculated based on Harrell C-statistics [10] for the ensemble CHF, with the b_{th} value being the error rate for the ensemble computed using the first b trees. To calculate variable importance (VIMP) for a variable x, drop OOB cases down their in-bag survival tree. Whenever a split for x is encountered, assign a daughter node randomly. The cumulative hazard rate function from each such tree is calculated and averaged. The VIMP for x is the prediction error for the original ensemble subtracted from the prediction error for the new ensemble obtained using randomizing x assignments[5].

To validate the performance of RSF based models, Cox proportional hazards models were then used for comparison and assessing the basic association between potential risk factors and mortality using bootstrapping with 1000 replications of individuals sampled with replacement [11]. We compared the discrimination performance of RSF based models with CPH based in terms of OOB C-statistics.

All analyses were performed with R version 3.0.1 (www.R-project.org).

III. Results

Predictors Identified by Two Models

From the RSF analysis with all 40 variables, 14 variables were selected to be predictive for 1-year mortality, including CA, BUN, BMI, AST, age, SCR, BR, K, WBC, ALT, NA, CKPK, class II agents and glucose (The detail minimal

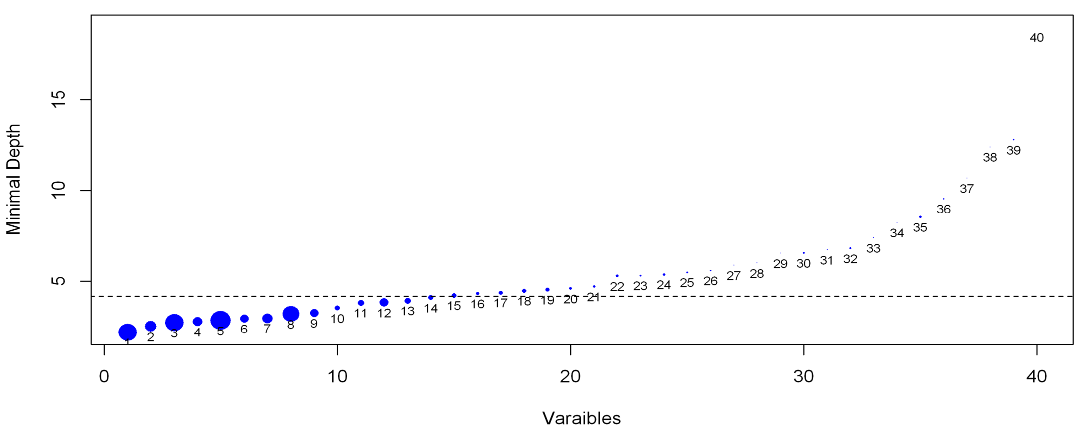

Fig. 1 Predictors identified from RSF analysis according minimal depth. Horizontal line is threshold for filtering variables. All variables below the line are predictive. The diameter of each circle in the plot is proportional to the forest-averaged number of maximal subtrees for that variable. 1.Cardiac arrest 2.BUN 3.BMI 4. AST 5.Age 6. SCR 7. BR 8.log of K 9.WBC 10.ALT 11. NA 12. CKPK 13.Class II agents 14. Glucose 15. INR 16.CHF 17.Renal failure 18. RBC 19. PTT 20. Class V agents 21. PT 22.stroke 23.sex 24.AF 25. Class IV agents 26.myocardial infarction 27.hypertension 28.uncomplicated diabetes 29.valvular heart disease 30.slow arrhythmias 31.VT 32.VF 33.hypothyroidism 34.complicated diabetes 35. Class III agents 36.liver disease 37.chronic pulmonary heart disease 38.acute pulmonary heart disease 39. Class I agents 40.Bundle branch block

depths of all variables can be seen from Fig.1, in which the horizontal line separates the 14 predictive variables from the remaining non-predictive variables).

Table 1 Predictors identified by Cox proportional Hazard Model

Predictors		**BC coefficient**	**Sig.**	**BC HR**
Demographics	logAge	.633	.000	1.883
	logBMI	-.979	.000	.376
Clinical risk factors	Cardiac arrest	.890	0.000	2.435
	Slow arrhythmias	-.420	0.000	0.657
	CHF	.119	.002	1.126
	myocardial infraction	.168	.001	1.182
	stroke	.340	0.000	1.405
	renal failure	.243	.000	1.275
Laboratory risk factors	logK	-1.146	.000	.318
	logWBC	.266	.000	1.305
	logRBC	-.443	.001	.642
	logBUN	.562	.000	1.754
	logGlucose	.187	.002	1.205
	logCKPK	-.106	.000	.900
	logAST	.379	.000	1.460
	logALT	-.255	.000	.775
	logPT	-.447	.000	.640
	logINR	.334	.000	1.397
	logBR	.107	.000	1.113
Medications	Class I agents	0.376	0.002	1.456
	ClassII agents	-0.316	0.000	0.729
	ClassIII agents	-0.864	0.000	0.421
	ClassV agents	-0.203	0.000	0.816

After multivariable CPH analysis with all 40 predictors, the following 23 risk factors presented in Table 1 were found to be independent significant predictors for mortality. These variables are a bit different from RSF based model as the RSF identified the nonlinear effect of the continuous variables on the mortality, which will be demonstrated later.

B. Discrimination Performance Comparison

The discrimination performance comparison for two models with different methods (RSF vs CPH) is presented in Table 2. From the table we can see, the RSF can improve the discrimination ability greatly with an OOB C-statistics of 0.812, while 0.736 for CPH based model.

Table 2 OOB C-statistics of two models with different methods

model \ method	RSF	CPH
model	0.812	0.736

C. Comparison on Ability to Identify Non-linear Effects

As presented above, predictors identified by CPH and RSF are a bit different. E.g., BR and NA are predictive in RSF based model, but not in CPH based. Indeed, BR and NA exact non-linear effects on mortality, which can be seen from Fig.2. Fig.2 shows the marginal effect of a given continuous variable on the 1-year mortality from RSF analysis, which are demonstrated in red line. Points are colored with blue correspond to events, while black to censored observations. From Fig.2 we can conclude RSF takes advantage in automatically identifying non-linear effects.

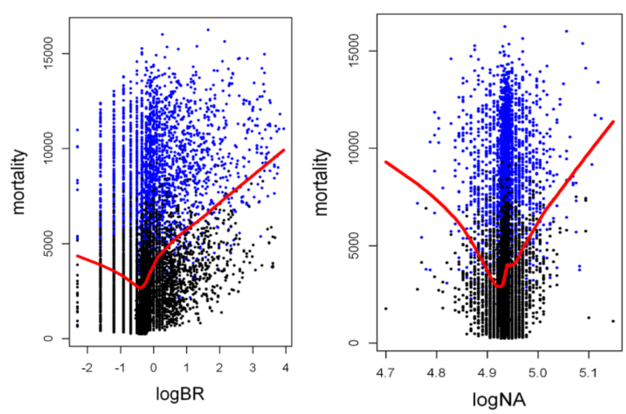

Fig. 2 Nonlinear effects identified by RSF

D. Comparison on Ability to Identify Important Predictors with Little Ratio of Population

Even though RSF has the above two advantages, it is inferior in identifying important predictor with fewer population. For example, class III agents is demonstrated to be a predictor with high hazard ratio in CPH based model, but not predictive in RSF based. In fact, it is an important predictor from the Kaplan-Meier curve presented in Fig.3. From the figure we can see, the survival functions for patients with and without class III agents are discriminated greatly. It is because that RSF is insensitive to factor with little ratio of population. The same reason for class I and class V agents.

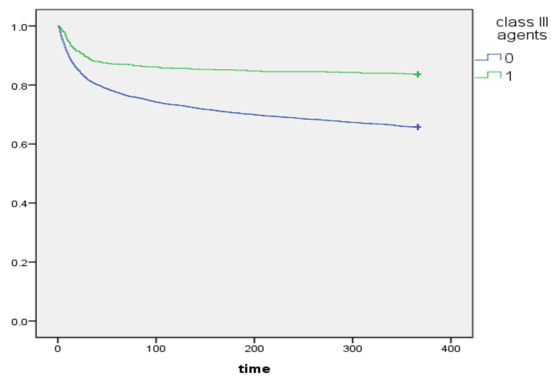

Fig. 3 Kaplan-Meier curve for class III agents

IV. CONCLUSIONS

In this paper, we evaluate the performance of RSF in predicting cardiovascular diseases by comparing it with CPH in the application of predicting 1-year mortality after cardiac arrhythmias. The results show that RSF based model has advantages in discrimination ability and identifying non-linear impacts of continuous variables. However, RSF cannot replace CPH in identifying important predictors with little ratio of population and this would go against to personalized prediction. The same results for predicting 1-year mortality after heart failure. Therefore, an improved RSF based on relative hazard ratio should be studied further.

LIST OF ABBREVIATIONS

CA: Cardiac arrest VF: Ventricular fibrillation
AF: Atrial fibrillation VT: Ventricular tachycardia
K: Potassium [mEq/L] in blood NA: Sodium [mEq/L] in blood
WBC: Leukocytes [K/uL] in blood RBC: Erythrocytes [m/uL] in blood
SCR: Creatinine [mg/dL] in serum BUN: Urea nitrogen [mg/dL] in serum
CKPK: Creatine kinase [IU/L] in serum
AST: Aspartate aminotransferase [IU/L] in serum
ALT: Alanine aminotransferase [IU/L] in serum
PT: Prothrombin time (seconds) in blood by coagulation assay
PTT: Activated partial thrombplastin time (seconds) in blood by coagulation assay
INR: International normalized ratio in blood by coagulation assay
Glucose: Glucose [mg/dL] in blood
BR: Total Bilirubin [mg/dL] in serum or plasma
BMI: Body mass index (kg/m2) CHF: Congestive heart failure

ACKNOWLEDGMENT

This work was supported in part by the National Basic Research Program 973 (NO. 2010CB732606), the Guangdong Innovation Research Team Fund for Low-cost Healthcare Technologies, Key Basic Research Program of Shenzhen (No. JC201005270257A).

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

1. Cox, David R (1972). Regression Models and Life-Tables. Journal of the Royal Statistical Society, Series B 34 (2): 187–220.
2. Keith A A Fox, Omar H Dabbous et al. Prediction of risk of death and myocardial infarction in the six months after presentation with acute coronary syndrome: prospective multinational observational study (GRACE). BMJ, Nov 2006; 333: 1091.
3. Schnabel RB, Sullivan LM, Levy D, Pencina MJ, Massaro JM, D'Agostino RB, Sr., Newton-Cheh C, Yamamoto JF, Magnani JW, Tadros TM, Kannel WB, Wang TJ, Ellinor PT, Wolf PA, Vasan RS, Benjamin EJ. Development of a risk score for atrial fibrillation (Framingham Heart Study): a community-based cohort Lancet 2009;373:739-745.
4. Breiman L. Heuristics of instability and stabilization in model selection. Ann Stat. 1996;24:2350 –2383.
5. Ishwaran H, Kogalur U B, Blackstone E H, et al. Random survival forests[J]. The Annals of Applied Statistics, 2008: 841-860.
6. Hsich E, Gorodeski E Z, Blackstone E H, et al. Identifying important risk factors for survival in patient with systolic heart failure using random survival forests[J]. Circulation: Cardiovascular Quality and Outcomes, 2011, 4(1): 39-45.
7. Kurt Omurlu I, Ture M, Tokatli F. The comparisons of random survival forests and Cox regression analysis with simulation and an application related to breast cancer[J]. Expert Systems with Applications, 2009, 36(4): 8582-8588.
8. Saeed M, Villarroel M, Reisner A T, et al. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database[J]. Critical care medicine 2011, 39(5): 952.
9. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 2000, 101(23):e215-e220.
10. Hajime Uno, Tianxi Cai, Michael J. Pencina, Ralph B. DAgostino, and LJ Wei. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statistics in Medicine 2011, 30:1105-16.
11. Steyerberg EW, Harrell FE, Jr, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001; 54:774 –781.