# Application of Random Forests and Deep Neural Networks to Suicide Death Data

STAT-6494 Project Proposal

*Wenjie Wang*[*]

*03 April 2018*

**Abstract**

The classical survival models, such as Cox proportional hazard model, often require extensive efforts on variable selection or prior medical information to model interaction between patients' covariates and treatment covariates. While nonlinear models, such as neural networks and random forests, are able to model high-order interaction terms. It is of interest to apply these machine learning methods to survival data and compare their performance with classical statistical models.

*Keywords:* Cox Model, Machine Learning, Suicide Prevention

[*]wenjie.2.wang@uconn.edu; Ph.D. student at Department of Statistics, University of Connecticut.

# 1 Introduction and Objects

For survival data, medical researchers' interests often lie in discovery of significant treatment effects and important diagnosis covariates of patients. The classical survival models, such as Cox proportional hazard model, assume risk function in a simple linear form of covariates, which can be too simplistic to capture the underlying relationship between response and covariates. In addition, they often require extensive efforts on variable selection or prior medical information to model interaction between patients' covariates and treatment covariates. While nonlinear models, such as neural networks and random forests, are able to model high-order interaction terms. It is of interest to apply these machine learning methods to survival data and compare their performance with classical statistical models. It would be even more interesting to discover nonlinear relationship by machine learning methods and build a statistical model for better interpretation and capability for statistical inferences.

The specific objectives include:

- Explore and review existing machine methods for survival data including random forests and deep neural networks.
- Apply these methods for CT suicidal data.
- Compare the out-of-sample model fitting or prediction performance of these methods with classical survival models, such as Cox model.

# 2 Random Forests for Survival Data

Random forests (RF) proposed by Breiman (2001) is an ensemble tree method that introduces randomization to the base learning process. Breiman (2001) showed that RF may further improve the prediction performance of simple ensemble learning method. Ishwaran et al. (2008) extended RF method to random survival forests (RSF) method for analysis of right-censored survival data.

Other reference includes

- Strobl et al. (2007)
- Mogensen et al. (2012)

# 3 Deep Neural Networks for Survival Data

The regular Cox proportional hazards model has a linear relative risk function $r(\boldsymbol{x}, \boldsymbol{\beta}) = \boldsymbol{\beta}^\top \boldsymbol{x}$. In many applications, it is hard to assume a linear proportional hazards condition and thus high-level interaction terms are required. However, as the number of covariates and interactions increases, it becomes prohibitively expensive.

Katzman et al. (2016) proposed a Cox proportional hazards deep neural network method called DeepSurv for personlized treatment recommandations. DeepSurv is a multi-layer perceptron that predicts a patient's risk of death. The output of the network is a single node estimating the relative risk function $\hat{r}_\theta$ by the weights of the network $\theta$.

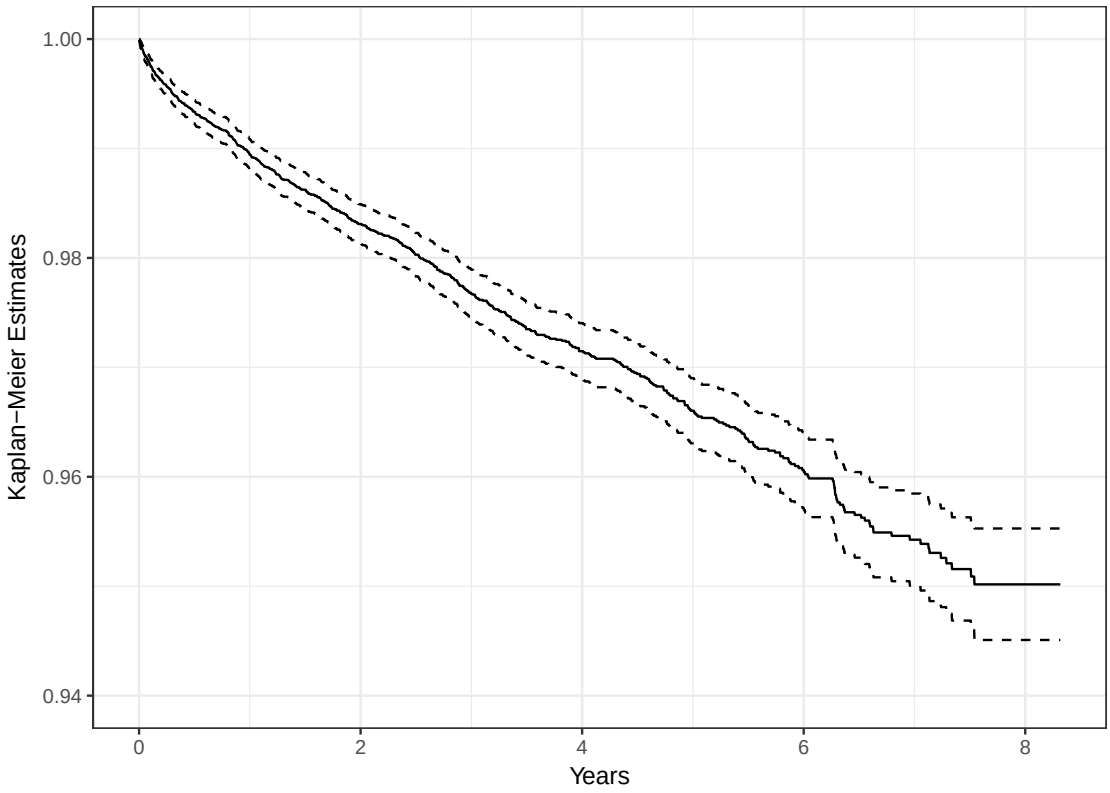
Other reference includes

- Nair and Hinton (2010)
- Ioffe and Szegedy (2015)
- Klambauer et al. (2017)
- Srivastava et al. (2014)
- Kingma and Ba (2014)
- Nesterov (2013)
- Pascanu et al. (2012)

# 4 Connecticut Suicide Death Data

Suicide is a serious public health problem in the US. Death by suicide is increasing among all age groups in the US, with a 24% increase in suicide rates observed from 1999 to 2014. There is a strong likelihood that suicide attempters would make additional attempts after the initial suicide attempt (Suominen et al., 2004), and suicide attempt is a strong predictor of suicidal death (Bostwick et al., 2015).

The subjects in the suicide death data were patients in the State of Connecticut who have been hospitalized for suicide attempt or intentional self-injury during fiscal year 2005 to 2012 (from October 1, 2004 to September 30, 2012). Data from diagnosis were available from the Connecticut Hospital Inpatient Discharge Data (HIDD). Deaths by suicide were determined from the Office of the Connecticut Medical Examiner (OCME). We are interested in the time since hospitalization due to suicide attempts to suicidal death of those patients. A total of 22,221 patients were followed up until September 30, 2012. Among them, 16,208 (73%) were white (9,108 female and 7,100 male) and 6,013 (27%) were non-white (3,220 female and 2,793 male). The number of event (suicidal death) was only 606 and thus the censoring rate was about 97.3%.




The HIDD data contained a large number of records on the characteristics of patients and their previous hospital admissions. One of the research interest was to identify important diagnostic categories associated with patient death. The diagnostics were recorded as ICD-9 diagnosis codes, or more formally ICD-9-CM (International Classification of Diseases, 9th Revision, Clinical Modification). We grouped the ICD-9 codes by their three leading characters that define the major diagnosis categories. Suicide attempts were identified by both ICD-9 external cause of injury codes and other ICD-9 code combinations indicative of suicidal behavior (Patrick et al., 2010; Chen and Aseltine, 2017). Other ICD-9 codes during the inpatient hospitalization fell into 167 major diagnosis categories, which led to 167 indicator variables.

### 4.1 Preliminary Results

We randomly split the suicide death data into a training set and a test set. The training set consists of 70% of the observations and the test set consists of the remaining 30%. We fitted random survival forest and DeepSurv model on the training set and measured the predicting power of these two methods over the testing set through Harrell's c-statistic (Harrell et al., 1996), an extension of the area under the receiver-operator characteristics curve for censoring data. This statistic is an estimate of the probability of concordance between the order of risk scores and survival outcomes.

- For random survival forest, the c-statistic of the training set and test set was 0.67 and 0.75, respectively.
- For DeepSurv model, the c-statistic of the training set and test set was 0.57 and 0.54, respectively.

We also fitted regular Cox proportional hazard model with only the basic characteristics of patients,

age, gender, and race, and length of (hospitalization) stay as covariates.

- For regular Cox model, the c-statistic of the training set and test set was 0.73 and 0.74, respectively.

### 4.2 To-do

- look into the variable importance measure from random survival forest model.
- tune the hyperparameters for DeepSurv model.
- do random splitting for estimating out-of-sample version of c-statistic as a measure of prediction performance for these methods.
- possibly look into a more homogenous subgroup of subjects
- possibly try finer ICD-9 code categories, will have 4,043 indicators/counts instead of 167.

## 5 Simulation Studies

### 5.1 Simulation Settings

We considered simulating survival data with severe censoring and a large cure fraction. We randomly generated totally ten covariates, $x_1, \ldots, x_{10}$, following Uniform$(-1, 1)$ and only $x_1$ and $x_2$ were used for simulating event times from Weibull model. The shape and scale parameter of the Weibull model was set to be 1.5 and 0.01, respectively. The covariate coefficients were set to be $\beta_1 = 1$ and $\beta_2 = 2$. The censoring times were simulated from Uniform$(0, 10)$. An intercept term, $z_0$, and one covariate, $z_1$, randomly generated from Uniform$(0, 1)$ were used for simulating the cure indicators from the logistics model. The coefficients were both set to be 1. The resulting cure rate and censoring rate was about 73.1% and 85.5% on average. Totally 1,000 simulated datasets were generated.

- For random survival forest, the c-statistic of the training set and test set was 0.74 and 0.77, respectively.
- For DeepSurv model, the c-statistic of the training set and test set was 0.69 and 0.67, respectively.
- For Cox Cure model, the c-statistic of the training set and test set was 0.79 and 0.79, respectively.

## Reference


Bostwick, M. J., Pabbati, C., Geske, J. R., and McKean, A. J. (2015), "Suicide Attempt as a Risk Factor for Completed Suicide: Even More Lethal Than We Knew," *The American Journal of Psychiatry*, 173, 1094–1100.

Breiman, L. (2001), "Random Forests," *Machine Learning*, 45, 5–32.

Chen, K. and Aseltine, R. (2017), "Using Hospitalization and Mortality Data to Target Suicide Prevention Activities: A Demonstration from Connecticut," *Journal of Adolescent Health*, 61, 192–197.

Harrell, F. E., Lee, K. L., and Mark, D. B. (1996), "Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors," *Statistics in medicine*, 15, 361–387.

Ioffe, S. and Szegedy, C. (2015), "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *International conference on machine learning*, pp. 448–456.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008), "Random Survival Forests," *The annals of applied statistics*, 841–860.

Katzman, J., Shaham, U., Bates, J., Cloninger, A., Jiang, T., and Kluger, Y. (2016), "DeepSurv: Personalized Treatment Recommender System Using a Cox Proportional Hazards Deep Neural Network," *ArXiv e-prints*.

Kingma, D. P. and Ba, J. (2014), "Adam: A Method for Stochastic Optimization," .

Klambauer, G., Unterthiner, T., Mayr, A., and Hochreiter, S. (2017), "Self-Normalizing Neural Networks," in *Advances in Neural Information Processing Systems*, pp. 972–981.

Mogensen, U. B., Ishwaran, H., and Gerds, T. A. (2012), "Evaluating Random Forests for Survival Analysis Using Prediction Error Curves," *Journal of Statistical Software*, 50, 1–23.

Nair, V. and Hinton, G. E. (2010), "Rectified Linear Units Improve Restricted Boltzmann Machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814.

Nesterov, Y. (2013), "Gradient Methods for Minimizing Composite Functions," *Mathematical Programming*, 140, 125–161.

Pascanu, R., Mikolov, T., and Bengio, Y. (2012), "Understanding the Exploding Gradient Problem," *CoRR*, abs/1211.5063.

Patrick, A. R., Miller, M., Barber, C. W., Wang, P. S., Canning, C. F., and Schneeweiss, S. (2010), "Identification of Hospitalizations for Intentional Self-Harm When E-codes are Incompletely Recorded," *Pharmacoepidemiology and Drug Safety*, 19, 1263–1275.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014), "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *The Journal of Machine Learning Research*, 15, 1929–1958.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007), "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution," *BMC Bioinformatics*, 8, 25.

Suominen, K., Isometsä, E., Suokas, J., Haukka, J., Achte, K., and Lönnqvist, J. (2004), "Completed Suicide After a Suicide Attempt: A 37-Year Follow-Up Study," *American Journal of Psychiatry*, 161, 562–563.