Annals of Infectious Disease and Epidemiology

Research Article
Published: 08 Mar, 2019

Determination of Risk Factors for Hepatitis C by the Method of Random Forest

Girdhar G Agarwal[1], Ashok K Singh[2], Vimla Venkatesh[3] and Neha Wal[4]*

[1]Department of Statistics, Lucknow University, India

[2]Department of Microbiology, University of Nevada, India

[3]Department of Microbiology, King George Medical University, India

[4]Department of Microbiology, King George Medical University, India

Abstract

The data mining method of Random Forest is used on a dataset collected in a study of HIV infected patients in Lucknow, India for determination of risk factors for the Hepatitis C virus. The accuracy of prediction the Random Forest model from this study is 98.3%.

Introduction

The estimated prevalence of HCV infection worldwide is 170 million people or 2-3% of the world population, with lowest prevalence (0.01 - 0.1%) in the UK and Scandinavia [1] and the highest (10 - 20%) in Egypt [2]. In the Indian subcontinent, prevalence varies [3]: Sri Lanka 0.16%, India 0.33%, Bangladesh 0.6%, Nepal 0.6%, Afghanistan 1.0%, Bhutan 1.3%, Myanmar 0.34%–2.03%, Pakistan 6%–6.8%, and the prevalence is unknown for Maldives.

Investigated the different genotypes among patients with HCV related chronic liver disease using a dataset collected from a tertiary care hospital in south India during the 2002-2012 decade [4]. HCV genotype 3 and genotype 1 turned out to be the predominant genotypes in the entire Indian sub-continent, with Genotype 4 and genotype 6 showing up in some parts [5] provide a summary of status of HCV in India. Mahajan et al. [6] analyzed a dataset of 8035 patients in India and found that HCV is more common in men, in middle-aged people, rural backgrounds, and low to middle socioeconomic class [7].

Used a dataset collected in 2005-2009 from urban areas in Houston, TX; the subjects were predominantly African American drug users who tested negative for HIV and HBV; Cox proportional hazard regression analyses showed daily drug use via injection to be a significant predictor [8]. Used the decision tree approach from data mining to classify characteristics of the genotype a (1 to 6) and genotype 1b. [9] used binary logistic regression for predicting HCV [10]. Used decision trees to predict success of antiviral therapy in chronic Egyptian patients and found alpha-fetoprotein (AFP) level to be an important predictor [11]. Provide a survey of AI applications in diagnosis of HCV.

In the present article, we use the ensemble method of random forest [12] to determine the risk factors of HCV.

Data

The study was conducted in the Department of Microbiology of King George's Medical University, Lucknow. This is a tertiary care teaching hospital located in the capital city of India's most populous state - Uttar Pradesh (UP). The hospital caters to the poor and seriously ill patients from the city and several surrounding districts. Patients are mostly referred from primary health centers, district hospitals, nursing homes, private doctors and community health workers. The hospital has an Integrated Counselling and Testing Centre (ICTC) facility which is well attended and provides counselling and HIV testing and linkages for medical and psychosocial care for persons living with HIV infection. The hospital also has an active Anti-Retroviral Therapy (ART) Center. Ethics Committee of K.G.M.U. provided ethical clearance for the study. A total of 350 HIV-infected adults attending the ART Centre and the ICTC were enrolled in the study between January 2007 and July 2008, after obtaining informed consent. The subjects were interviewed using a pre-designed proforma. Current clinical symptoms were elicited by direct questioning by the

OPEN ACCESS

Correspondence:
Ashok K Singh, Department of Microbiology, University of Nevada, Lucknow, India,
E-mail: ashok.singh@unlv.edu

Received Date: *01 Feb 2019*
Accepted Date: *01 Mar 2019*
Published Date: *08 Mar 2019*

Citation:
Agarwal GG, Singh AK, Venkatesh V, Wal N. Determination of Risk Factors for Hepatitis C by the Method of Random Forest. Ann Infect Dis Epidemiol. 2019; 4(1): 1037.
ISSN: 2475-5664

Table 1: Table of predictors with missing values.

Predictor	Number of missing values
DLCB (DLC_Basophil)	345
SBILD (S.Bilirubin Direct)	340
SGOT	325
Ifmar (if married)	311
DLCM (DLC_Monocyte)	277
SUREA (S. Urea)	225
DLCE (DLC_Eosinophil)	222
NW (If Not Working)	217
SCREAT (S.Creatinine)	215
Dart (Duration of ART)	199
ALP	197
W (If working)	195
SBILT (S.Bilirubin Total)	194
SGPT	194
DLCL (DLC_Lymphocyte)	191
DLCN (DLC_Neutrophil)	190
Tmem (Total members in the family)	69
CD4	47
Occupt (Occupation)	2
Srash (Skin rash)	1

interviewer, using a checklist. The questions were designed to have yes or no answers. Questions were asked in the local language and responses recorded. Case records at the ART Centre were reviewed to obtain data for documented opportunistic infections in the subject around the interview date.

The dataset used in this study has 350 observations on a total of 90 variables, with presence of HCV as a binary response variable. Twenty five of the potential predictors of HCV have missing values (see Table 1); these variables are not being used in our analyses.

Methods

The method of Random Forest is a predictive model that can be

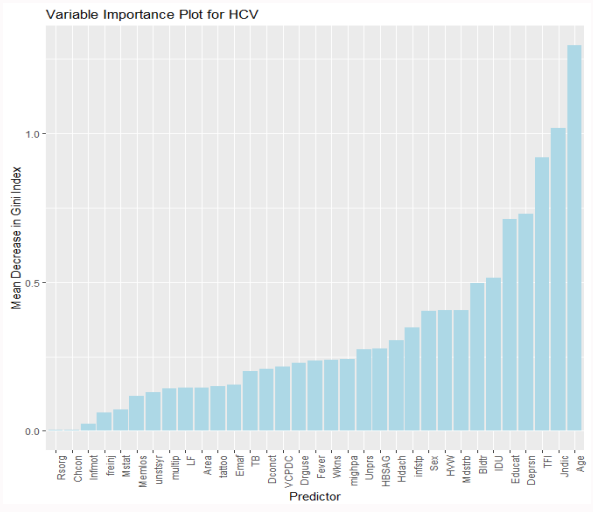

Figure 1: Variance Importance Plot of the Final Random Forest Model for HCV.

Table 2: Brief description of predictors in the final Random Forest model.

Predictor	Description
Jndce	Jaundice
TFI	Total Family Income,
Deprsn	Depression
Educat	Education
IDU	Injected Drug User by sharing contaminated needles
Bldtr	blood transfusion
Mdstrb	Motor disturbances
HVW	HIV Clinic/VCTC/Ward (1: HIV Clinic, 2: VCTC, 3: Ward, 4: Psychiatry KGMU, 5: Nirvan)
Sex	Gender (1: Male, 2: Female)
Infstp	sex with an infected steady partner,
Hdach	Headache
HBSAG	Hepatitis B Surface Antigen Assay
Unprs	unprotected sex
migpha	migration to high prevalence area for HIV,
Wkns	Weakness
freinj	frequent injection/hospitalization
Drguse	drug use/addiction
VCPDC	Vague chest pain & dry cough
Dconct	Diminished concentration
Emaf	extra-marital affairs (1: Yes, 0: No)
Area	(1: Rural, 2: Urban)
LF	Lethargy & Fatigue
multip	sex with multiple partners
unstyr	injections with unsterile syringes
Memlos	Memory loss
Mstat	Marital Status(1: Unmarried, 2: Married)
freinj	frequent injection/hospitalization
Infmot	infected mother
Chcon	Changes in consciousness level
Rsorg	receiving solid organs

used for regression or classification. Random Forest involves building a large number of decision trees, and outputting the mode of the classes predicted by individual trees (for classification), or the mean of predicted values (for regression) obtained for individual trees. The method of Random Forest is known for its high accuracy and efficiency [13-15]. A detailed description of the method of Random Forests is provided in Hastie, [12].

All of the computations reported in this article are done in the statistical software environment R (2017). The association between the response variable and each individual predictor was first tested by the method of chi-square test of independence; in many of the cases, the expected frequencies of several cells turned out to be less than 5, and the p-values for the chi-square test were evaluated by bootstrap [16].

The R-package random Forest was used for fitting a predictive model to the binary response HCV satisfaction score as a function of the selected predictors. The package random Forest outputs 'Out of Bag' (i.e., out of the training sample) estimates of prediction accuracy

Table 3: Results of bootstrap chi-square tests for 16 important categorical predictors.

Predictor	Chi-square	P-Value
Jaundice	30.579	0.005
Depression	8.9724	0.032
Education	4.7607	0.176
IDU	13.021	0.069
BLDTR	2.18	0.188
Mdtrsb	2.18	0.182
HVW	5.154	0.087
Gender	0.056429	1
infstp	0.7257	0.449
Hdach	2.1293	0.232
Unprs	0.42995	0.694
mighpa	0.11305	1
Wkns	0.027594	1
freinj	1.8678	0.352
Drguse	0.17636	0.717
VCPDC	0.17063	1

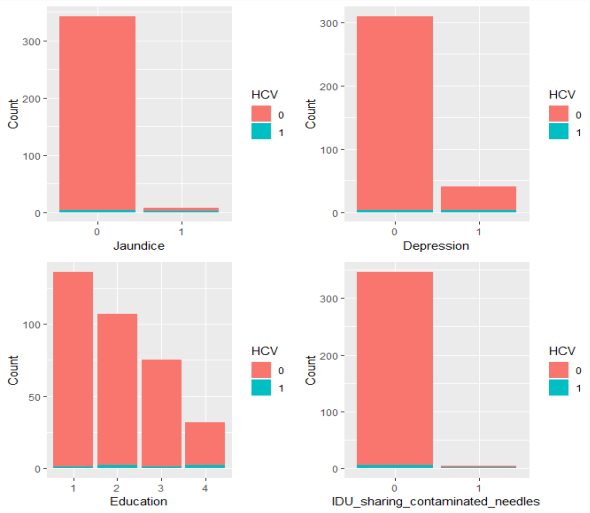

Figure 2: Stacked bar charts of HCV vs. Jaundice, Depression, Education, and IDU sharing contaminated needles.

as well as a plot showing the importance of predictors in the model. The package was iteratively used by adding and dropping predictors until a final model with good prediction accuracy was obtained. The statistical significance of the model was tested using the method of Evans , Murphy, Holden, and [17].

Results

The random forest model was run several times, starting will all potential predictor and eliminating less important predictors until the final model was found. The statistical significance of the final model was then tested using the R-function rf. significance of the R-package rf Utilities developed by [18]. The overall accuracy of the final random forest model turned out to be 98.3%. The significance test using 1000 permutations for the fitted random forest model yielded a P-value of 0.000, which implies that the random forest model significantly fits the data; the out-of-bag (OOB) accuracy obtained from running the

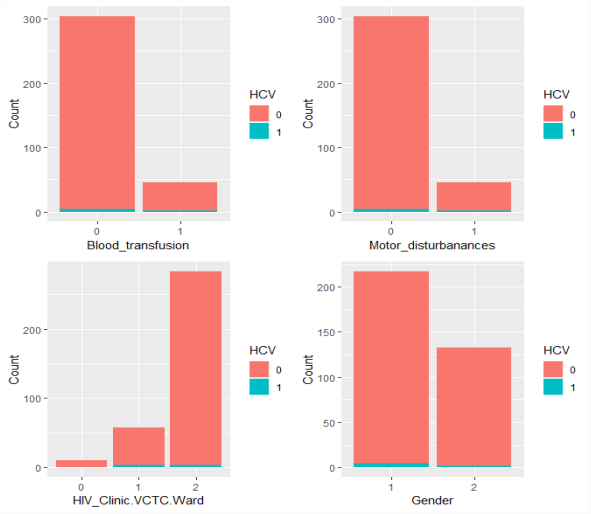

Figure 3: Stacked bar charts of HCV vs. Blood transfusion, Motor disturbances, HIV Clinic, and Gender.

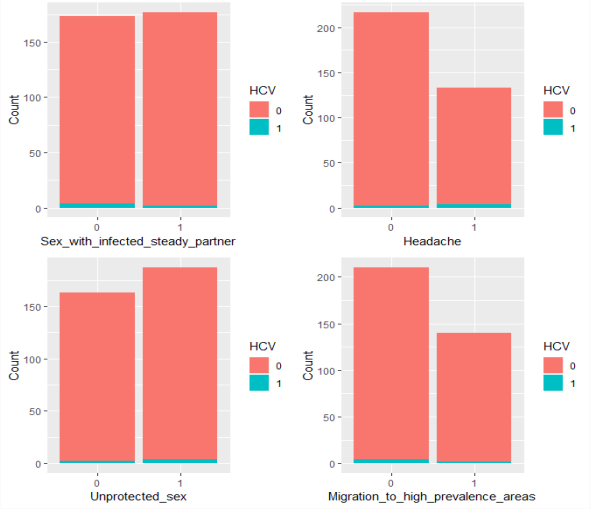

Figure 4: Stacked bar charts of HCV vs. Sex with infected steady partner, Headache, Unprotected sex, and Migration to high prevalence areas.

significance test is 98%.

Figure 1 is the variance importance plot of the final random forest model fitted to the dataset. It can be seen from Figure 2 that the important predictors include some of the known risk factors:

Age, Jaundice, TFI, Depression, Education, IDU, BLDTR, Mdtrsb, HVW, Gender, infstp, Hdach, Unprs, mighpa, Wkns, freinj, Drguse, and VCPDC (see Table 2 for a brief description of these predictors).

All of the above predictors except Age and TFI are categorical. The chi-square test of independence was run for the response variable HCV and each of the above categorical predictors (Table 3). It can be seen from Table 3 that Jaundice, Depression, IDU, and HVW are significant (Figure 3).

Discussion of Results

Logistic regression [12,19] is the most commonly used method for predicting a binary response variable, but when applied to the dataset at hand, it yielded a model in which no predictor was significant and prediction accuracy was low (Figure 4,5). We have shown in this

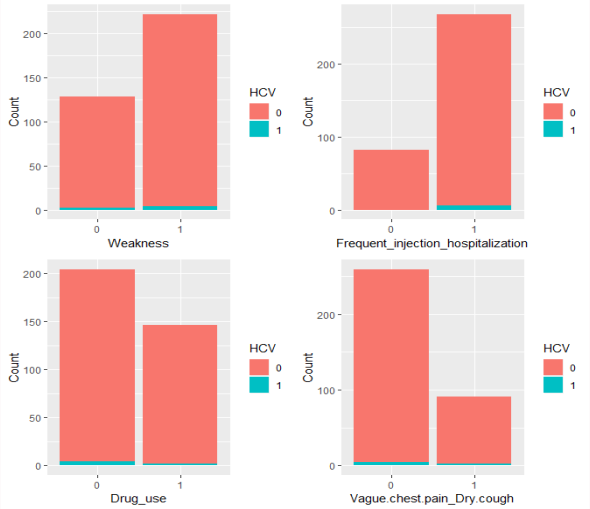

Figure 5: Stacked bar charts of HCV vs. Weakness, Frequent injection hospitalization, Drug use and Vague.chest.pain_Dry.cough.

article that the method of random forest is applicable in such cases, and can be used to predict HCV and also determine the risk factors of a disease.

References

1. Davis GL. Epidemiology of Chronic HCV. In Shiffman ML, editor. Chronic Hepatitis C Virus: Advances in Treatment, Promise for the Future. Springer Science & Business Media. 2011;3-11.
2. Kamal SM, Abdelhakam SA. Hepatitis C in Egypt. In Kamal SM, editor. Hepatitis C in Developing Countries: Current and Future Challenges. Academic Press. 2018;1-56.
3. Shekhar S. Hepatitis C Virus Infection in the Indian Sub-Continent. In Kamal, SM, editor. Hepatitis C in Developing Countries: Current and Future Challenges. Academic Press. 2018;83-95.
4. Christdas J, Sivakumar J, David J, Daniel HD, Raghuraman S, Abraham P. Genotypes of hepatitis C virus in the Indian sub-continent: a decade-long experience from a tertiary care hospital in South India. Indian J Med Microbiol. 2013;31(4):349-53.
5. Mukhopadhya A. Hepatitis C in India. J Biosci. 2008;33(4);465-73.
6. Mahajan R, Midha V, Goyal O, Mehta V, Narang V, Kaur K, et al. Clinical profile of hepatitis C virus infection in a developing country- India. J Gastroenterol Hepatol. 2017;33(4):926-33.
7. Shah DP, Grimes C, Lai D, Hwang LY. Hepatitis C Incidence and Spontaneous Viral Clearance in a Cohort of Human Immunodeficiency Virus and Hepatitis B Virus Negative Drug Users. Ann Infect Dis Epidemiol. 2017; 2(2):1014.
8. Yeong Jeong H, Yoon TS. Analysis of Hepatitis C Virus using Data mining algorithm -Apriori, Decision tree. International Journal on Bioinformatics & Biosciences (IJBB). 2017.
9. Yasin H, Jilani TA, Danish M. Hepatitis-C Classification using Data Mining Techniques. International Journal of Computer Applications. 2011;1-6.
10. Zayed N, Awad AB, El-Akel W, Doss W, Awad T, Radwan A, et al. The assessment of data mining for the prediction of therapeutic outcome in 3719 Egyptian patients with chronic hepatitis C. Clin Res Hepatol Gastroenterol. 2013;37(3):254-61.
11. Sweidan S, El-Bakry H, El-Sappagh S, Sabah S, Mastorakis N. Viral Hepatitis Diagnosis: A Survey of Artificial Intelligent Techniques. International Journal of Biology and Biomedicine. 2016;1:106-16.
12. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning - Data Mining, Inference, and Prediction. Springer Series in Statistics. Second Edition, 2017.
13. Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32.
14. Pal M. Random forest classifier for remote sensing classification. International Journal of Remote Sensing. 2005;26(1):217-22.
15. Shi T, Seligson D, Belldegrun AS, Palotie A, Horvath S. Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. Mod Pathol. 2005;18(4):547-57.
16. Habermana SJ. A warning on the use of chi-squared statistics with frequency tables with small expected cell counts. Journal of the American Statistical Association. 1988;83(402):555-60.
17. Cushman SA. Modeling species distribution and change using Random Forests. In: Editors Drew, CA, Huettmann F, Wiersma Y. Predictive Modeling in Landscape Ecology. 2011.
18. Evans JS, Murphy MA, Holden ZA, Yee LJ, Weiss HL, Langner RG, et al. Risk factors for acquisition of hepatitis C virus infection: a case series and potential implications for disease surveillance. BMC Infectious Diseases. 2001;1:8.
19. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. 2017.