International Journal of Big Data and Analytics in Healthcare
Volume 6 • Issue 2

Predicting Inpatient Status for the Next 30/60/90 Days With Machine Learning

Lakshmi Prayaga, University of West Florida, USA
Krishna Devulapalli, Indian Institute of Chemical Technology, India
Chandra Prayaga, University of West Florida, USA
https://orcid.org/0000-0002-7534-4313
Joe Carloni, Lakeview Center Inc., USA

ABSTRACT

In this paper, the authors report the development of machine learning techniques that can help hospital authorities assess a patients' medical condition and also calculate the probability of readmission of the patient as inpatient, and thus identify patients with higher risks for readmissions. Factor analysis is performed on patient data to understand the severity of mental health, and random forest models are used to determine the probability of a patient becoming an inpatient for the next 30/60/90 days from their last visit to the physician's office. The random forest model fits the data with an overall OOB error rate of 3.69% and an accuracy of 97.65%. The accuracy on the test data was 96.11%. A web application is also developed to provide a user-friendly interface for physicians and administrators to interact with and obtain relevant information for a given patient and or a group of patients. The web application affords physicians additional inputs to assist in their diagnosis and administrators a window into anticipating and preparing for future patient needs.

KEYWORDS

30/60/90 Day Predictions, In-Patient Stay Predictions, Interactive Web App for Predictive Analytics, Machine Learning, Mental Health Severity Index

INTRODUCTION

Mental health illnesses are becoming more prevalent (Owens et al., 2019) in the United States. In 2019, NIH estimates that approximately one in five people or 51.5 million people aged 18 years and over suffered from mental and/or substance abuse disorders (MSUDs). Of these adults, nearly 45 million had a mental disorder alone, 11 million had a substance abuse disorder alone, and 8 million had both a mental disorder and a substance abuse disorder. It is further found that disorders such as depression, anxiety, and substance abuse are associated with significant distress and impairment, including complications with multiple chronic conditions, disability, inability to function in society, and substantial economic costs. Sporinova et al. (2019) cite a 3-year adjusted mean cost at $38,250 for those with a mental health disorder, and $22,280 for those without a mental health disorder. According to The American Psychological Association (Winerman, 2017), in the year 2013, $187.8 billion dollars, including out of pocket expenses, were categorized as costs related to mental disorders.

DOI: 10.4018/IJBDAH.20210701.oa9

Taking into account additional costs associated with loss of productivity and disability payments, the total cost of MSUDs to society is estimated to be more than twice that amount.

Hospitalization is a very important component of treatment plans for individuals with serious and persistent illness. However, hospital inpatient stay has become very expensive in countries like the USA. According to Lan Liang et. al. (2016), there were over 35 million hospital stays, equating to 104.2 stays per 100,000 population. The average cost per hospital stay was $11,700, making hospitalization one of the most expensive types of healthcare services.

According to The Piper Report (2020), hospital lengths of stay for mental health (MH) or substance abuse (SA) disorders also vary considerably, especially for mental-health related admissions. Nationwide, the MH average length of stay is 8.0 days. According to the same Report, MH and SA hospitalizations are, on average, less expensive than non-MHSA stays:

$5,700 per MH stay.

$4,600 per SA stay.

$9,300 per stay for all other conditions.

Health Catalyst, in its Newsletter issue May 25, 2017 published an article entitled "Enhancing Mental Health Care Transitions Reduces Unnecessary Costly Readmissions" and stated that "Nationally, hospitalization for persons with mental health disorders has increased faster than hospitalization for any other condition". Also mentioned is the lack of bed space to admit the patients on a timely basis.

Therefore, it becomes necessary to assess the mental health condition of the patients. In the current study, machine learning techniques are developed, for associating the patients' demographic, behavioral, psychological and other related data, and to evaluate the probability of hospital inpatient admission for these patients. By setting a threshold value for the probability, the medical practitioner can assess whether the patient needs inpatient admission or not. It is also interesting to assess the level of Mental Health Severity of communities, based on race, gender and patient status by using all the complex and rich data that is available. Factor analysis techniques are used here to develop a comprehensive Mental Health Severity Index (MHSI) by using the variables and to rank the communities. The rest of the article is organized in the following sections: Literature review, Materials and methods, Machine learning algorithms used, Results, and Conclusion.

LITERATURE REVIEW

Hyunyoung Baek et al. (2018) applied statistical data mining approaches to analyze the length of hospital stay using electronic health records. They identified five significant variables (frequency of surgery, frequency of diagnosis, frequency of patient transfer, severity, and insurance type) as most relevant to predict the length of stay (LOS). Multiple regression analysis was used for prediction, with an R^2 of 0.267 and a mean absolute error 4.68. Luc Jansen et al. (2018) studied the extent to which medical-psychiatric comorbidities relate to health-economic outcomes in general and in different subgroups. Their study indicated that comorbidities such as depression increased the LOS for patients by 4.38 days compared to those patients who did not have comorbidities. Tsai et al. (2016) applied Artificial Neural Network (ANN) models to predict LOS for cardiac patients with coronary atherosclerosis (CAS), acute myocardial infarction (AMI), and Heart Failure (HF). Their study obtained accuracy levels of 88.07% to 89.95% for CAS patients at predischarge level, and 88.31% to 91.53% at the pre-admission stage. Results for AMI and HF were observed with accuracy levels of 64.12% to 66.78% at pre discharge level and at 63.69% to 67.47% at the preadmission state. Mekhaldi et al. (2020) applied the Random Forest and Gradient Boosting models to predict length of stay (LOS) in a hospital setting. Lin et al. (2016) used multivariate logistic regression model to predict inpatient readmission and outpatient admission in the elderly, and expressed that these models provide a basis for wider application in National health Service. They predicted Length of Hospital Stay for older citizens considering comorbidities, home healthcare, and prior use of healthcare facilities. They applied

the Area Under the Curve (AUC) model to determine the LOS and reported the AUC of the inpatient readmission model as 0.655 . Gopalakrishna, Ithman and Malwitz (2015) studied the predictors of length of stay in acute psychiatric hospitals. They applied regression models with natural logarithms of LOS as the dependent variable and age, marital status, involuntary admission and diagnosis of an affective disorder or a psychotic disorder as independent variables, which could explain about 20% of the variation of the variance of LOS.

Hospital readmissions have received attention from researchers, in view of the cost of such readmissions. Upadhyay, Stephenson and Smith (2019) studied the effect of readmission rates on hospital financial performance. Their analysis of data, from 98 hospitals in the State of Washington from 2012 to 2014, indicated that a reduction in acute myocardial infarction (AMI) readmission rates is related with increased operating revenues as expenses associated with costly treatments related with unnecessary readmissions are avoided. Cardarelli et al (2018) carried out a quasi-experimental study design which assessed implementation of a lay0health worker (LHW) model for assisting high-risk patients with their post-discharge social needs. The LHW intervention involved assessment and development of a personalized social needs plan for enrolled patients (e.g. transportation and community resource identification), with post-discharge follow-up calls, and resulted in a 47.7% relative reduction of 30-day hospital readmissions rates between baseline and intervention phases of the study. Wan et al (2011) have given an extensive review of preventable hospital readmissions. Liu et al (2020) used artificial neural networks (ANN) to predict 30-day hospital readmissions. They compared the performance of their models with hierarchical logistic regression models and found that the ANNs increased the AUC for prediction of 30-day readmissions. Flaks-Manov, N., Topaz, M., Hoshen, M. *et al.* (2019) suggest that readmission risk identification should incorporate a two time-point approach in which preadmission data is used to identify high-risk patients as early as possible during the index admission and an "all-hospital" model is applied at discharge to identify those that incur risk during the hospital stay.

In the field of mental health, Šprah, L., Dernovšek, M.Z., Wahlbeck, K. *et al.* (2017) have reviewed the impact of physical comorbidity variables on readmission after discharge from psychiatric or general inpatient care among patients with co-occurring psychiatric and medical conditions. Benjenk and Chen (2018) have reviewed Effective mental health interventions to reduce hospital readmission rates.

Researchers (Degenhardt et al., 2019; Wongvibulsin, Wu and Zeger 2020) suggest that Random Forest (RF) algorithms are good candidates to address the challenges associated with high dimensional and heterogenous data that includes electronic health records. Degenhardt et al. report that RF methods have been applied in proteomics. RF has been successfully applied in genetics, gene expression, methylation, proteomics, and metabolomics studies. It is a flexible approach that can be used to perform both classifications, i.e., predicting case-control status, and regression, i.e., predicting quantitative traits. Wongvibulsin, Wu and Zeger (2020) also used the RF algorithm to predict sudden cardiac arrests with a high degree of confidence. Based on the included literature review, we find that our choice of RF is suitable for this study since a. it is a good technique to use for high dimensional data as reported by (Degenhardt et al., 2019; Wongvibulsin et al., 2019) and the data for the current study falls under this category, b. the 96.31% accuracy obtained by using RF in the current study was higher than those reported from earlier studies using regression analysis (Lin et. al 2016) and neural networks (Tsai et al. 2016) produced accuracy levels of 88.07% to 89.95% and c. it has a wide applicability as reported in the literature review section.

Our contribution to the literature on predicting inpatient stay is that we address the probability of inpatient admission for patients with mental health illnesses using only demographic and psycho-social data and not requiring clinical data. Additionally, we use two machine learning algorithms, one, Factor Analysis to determine the severity of the mental illness for specific population groups of the dataset and two, Random Forest to predict the probability of a given patient becoming an inpatient in the next 30, 60 or 90 days. We also develop a web application for physicians and administrators to search for a specific patient and obtain the probability of that patient becoming an inpatient. The

web application also provides group wise information for a specific population from the available dataset. It is a useful tool to assist physicians to get an additional input to their diagnosis and can be used to prepare a treatment plan for that patient. It also allows administrators to use the tool to plan for future resources required using the 30-/60-/90-day search options leading to better patient care.

MATERIALS AND METHODS

Data Description

The dataset used in this study was provided by Lakeview Center, Inc., which is a private non-profit organization providing behavioral health care. The data was anonymized by Lakeview Center and then provided for purposes of this study, thus no personal information was compromised, maintaining strict confidentiality and ethical norms. The data is available in an Excel file and consists of 80849 observations. The Excel file contains the data relating to daily admissions of patients during the period January 2019 to June 2020. This data contains information on 20524 patients with different mental disorders. Among the 20524 patients, 2754 (13.42%) patients are inpatients and the remaining 17770 (86.58%) are outpatients.

Data is collected broadly utilizing three types of forms, viz., psychosocial, service needs assessment and SAFET assessment. Psychosocial data relates to items such as Symptoms, Onset, Frequency, Severity, Use of Medicines, Previous mental health treatment, and Family history of Mental Illness. Service Needs assessment is used to identify strengths and needs of individuals that may impact their ability to participate in services. Based on that assessment, the medical practitioner would provide case management services to reduce such barriers. SAFET is the instrument used to identify homicide & suicide risk that is performed on all clients over the age of 5 years.

Specimen screenshots of each of the three categories are displayed in Figures 1 to 3.

Each observation contains data on 145 variables, relating to ClientID, activityMonth, ZipArea and diagnosis in the categories of Psychosocial assessment, Service Needs Assessment and SAFET. Out of these, 135 variables are retained for the current study. The above-mentioned client specific variables are not used in the model fitting, and seven other variables, with only one value entered, are also eliminated.

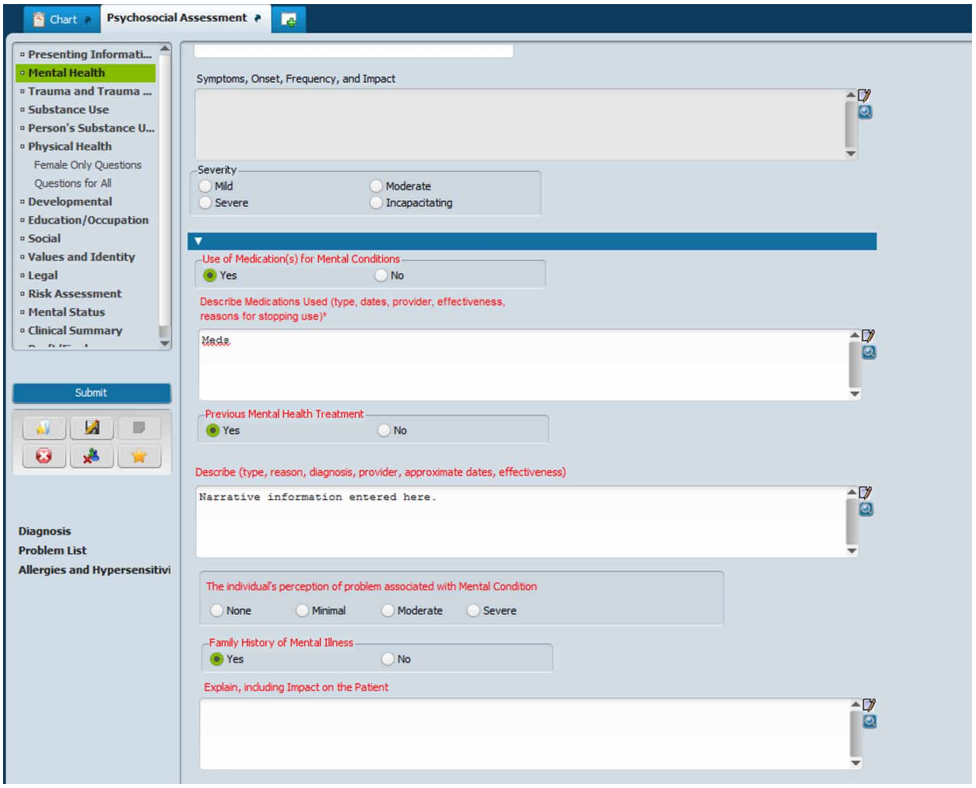
All the variables contain either character values or numerical values in a scale of 0 to 10. All the missing values are filled with the character string ‘Unknown’. Demographic variables such as race, sex, and ethnicity contain character data. Variables with values from 0 to 10 denote the ranking of the diagnostic test result. Higher ratings represent higher levels of severity of mental illness. The variable numinpatientStayLast30 represents the number of times the patient was admitted in the hospital as an inpatient during the last 30 days with reference to activityMonth. This variable contains values ranging from 0 to the number of times the patient was admitted in the hospital during the last 30 days. For patients not admitted in the hospital, these fields are filled with the value zero, and for those who were admitted, they are assigned the values 1, 2, 3 ... corresponding to the number of times they were admitted.

Data Preprocessing

All the character data is recoded with numerical values. For example, the variable sex is recoded by numerical values 1 and 2, 1 for males and 2 for females. Similarly, the other demographic variables are also recoded with appropriate numerical values. All the missing values are assigned the value 9. In the case of all the other remaining categorical variables having character data, appropriate numerical values are assigned. For example, if any variable has character values ‘yes’, ’no’, ’unknown’, recoding is done by assigning the numerical values 1, 2 and 9. Variables having numerical values are not modified.

The variable numInpatientStayLast30 contains the number of times the patient was admitted in the hospital during the Last 30 days and it takes any value from 0 to any number. This variable is converted

Figure 1. Specimen data entry screenshot of psychosocial data



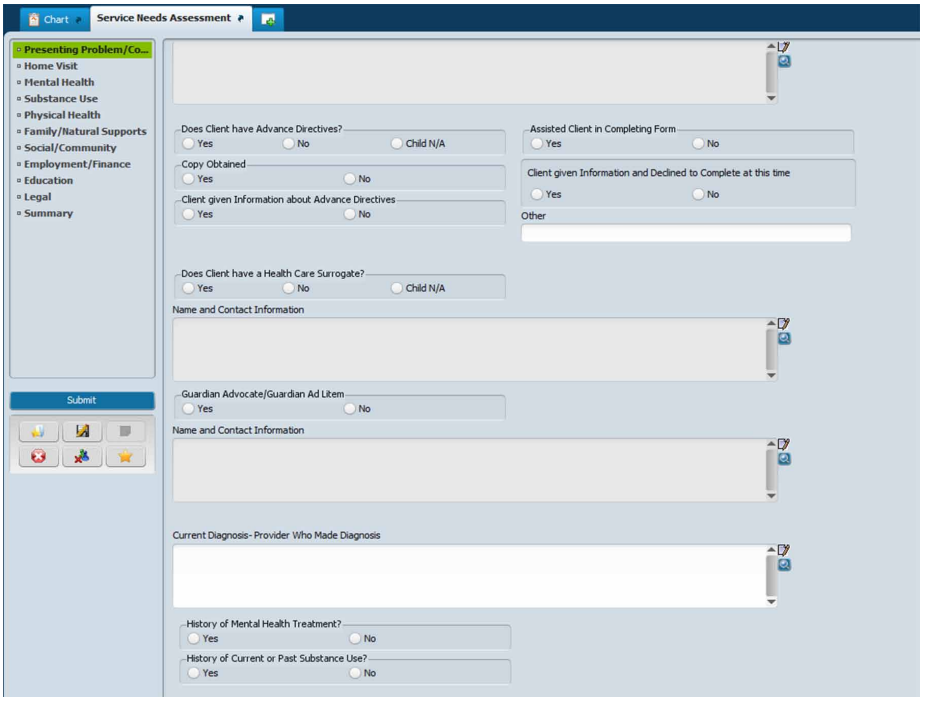
into a binary variable, representing the patient status as InPatient or OutPatient, as follows : A patient is considered as InPatient if he/she has stayed at least once for one or more days and Outpatient if the patient has not stayed at least once. Accordingly, if the variable numInpatientStayLast30 takes any value greater than 0, it is converted to 1 (representing Present) and to 0 if its value is zero (representing Absent). Thus, the value 1 represents the InPatient status as present (ie. the patient is an InPatient), and zero represents the InPatient status as absent (ie. the patient is an OutPatient). This variable is treated as the dependent variable and all the remaining variables are treated as independent variables in fitting the Random Forest model.

MACHINE LEARNING ALGORITHMS USED FOR THE STUDY

Factor Analysis for evaluating Mental Health Severity Index (MHSI)

Using the variables under consideration, a comprehensive Mental Health Severity Index (MHSI) is calculated by the procedure detailed in Prayaga et al. (2020). Factor analysis is a dimension reduction technique to reduce a large number of variables into a fewer number called factors. In this study, we have used principal component analysis to extract the factors. Factor analysis evaluates three important quantities viz., factor loadings, eigenvalues and factor scores. Factor loadings are essentially the correlation between the original variables and the factors. Eigenvalues show the variance explained by each factor out of the total variance. Factor scores F_j's are index variables obtained as optimally-weighted linear combinations of the variables.

Figure 2. Specimen data entry screenshot of Service Needs Assessment



The first step in factor analysis is to determine the number of factors to be retained for further analysis. The eigenvalues are good indicators for determining the number of factors. Generally, the first few factors have eigenvalues greater than one. If the eigenvalue is greater than one, that factor should be included, else, it should be discarded. The Scree plot proposed by Ledesma et al. (2015) has also been used by researchers to assess graphically the number of factors 'm' to be retained for exploratory factor analysis. A Scree plot is a line plot of the eigenvalues of factors. In the Scree diagram, the number of factors 'm' to be retained is obtained as the meeting point of the eigenvalues curve and the parallel analysis curve.

The Mental health severity index ($MHSI_k$) for each patient k is obtained by multiplying the square of each retained factor score F_j by the proportion of variance S_j explained by the corresponding factor as the weight, and then adding the products as given by the following formula:

$$MHSI_k = \sum_{j=1}^{m} F_j^2 S_j \quad (1)$$

where k = 1…n is the number of patients and j = 1…m is the number of factors

The aggregated mean MHSI values were evaluated for the following combinations of groups:

Race (4 in number) – White, Black, Others (Multiracial, American Indian etc.) and Unknown

Gender (2 in number) – Male, Female

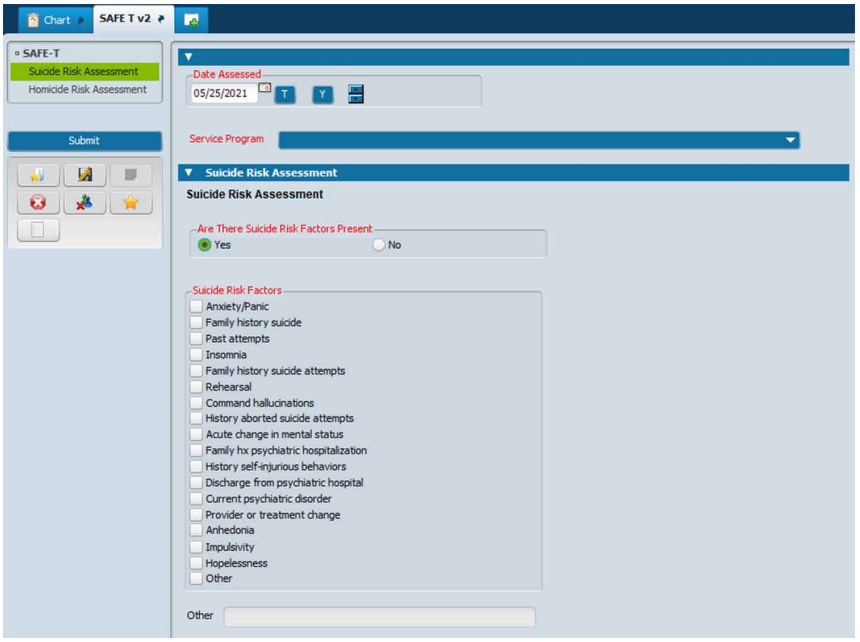
Patient status (2 in number) – Inpatient, Outpatient

This yields 16 mean values corresponding to all possible combinations of 4 races, 2 gender classifications and 2 patient status categories (4 X 2 X 2). These final mean MHSI values were then used to compare the mental health status among these 16 groups.

Probability Of Admission By Applying Random Forest Models

A random forest (RF) is an ensemble bagging or averaging method that aims to reduce the variance of individual trees by randomly selecting (and thus de-correlating) many trees from the dataset, and

Figure 3. Specimen data entry screenshot of SAFET



averaging them. It is an extension of bagging. Random forest achieves better accuracy by reducing variance through the averaging of the prediction of orthogonal trees. It is an ensemble modeling technique that combines several machine learning algorithms into one prediction model. Research suggests that RFs improve accuracy by reducing the estimator variance by a factor of three-fourths (Genuer, 2012). Several recent studies (Blankers et al., 2020), have demonstrated that RFs have been very effective in predicting the desired outcomes with a high degree of accuracy. It is for these reasons, of reduction in the variance and improved accuracy for high dimensional data, that the Random Forest algorithm is chosen in this study.

A Random Forest Model is applied to the cleaned, preprocessed data by considering the numinpatientStayLast30 as the dependent variable and all the other variables as independent variables.

An interactive Shiny App is also developed to display the various results of the application including the model fitting, its accuracy, the important variables, patients with highest probability of admission etc. The Shiny App can also predict the probability of admission for any single patient, even in the case of new patients. The development of the Shiny App is also carried out using RStudio and is uploaded on shinyapps.io website.

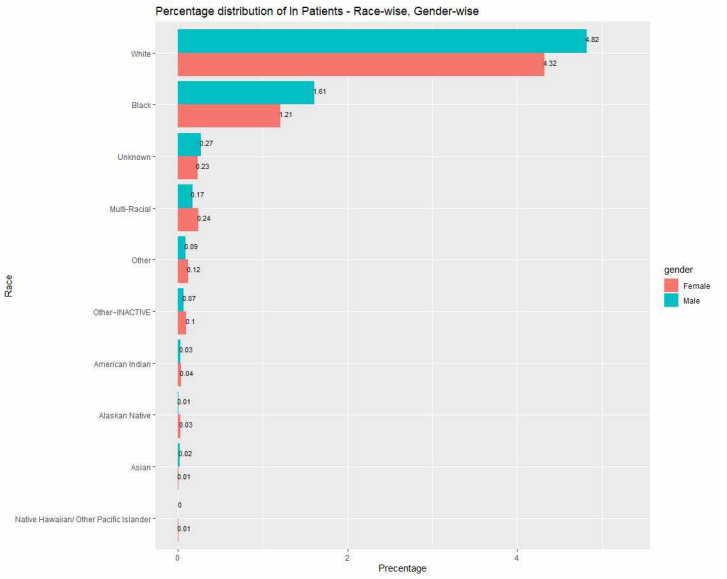
All the statistical analysis, calculation of the MHSI by factor analysis and random forest model fitting were carried out using R and associated statistical packages.

Results and Discussion

The percentage distribution of inpatients, by Race and Gender is evaluated and displayed in Figure 4.

In order to compare the mental health severity among the races, genders and patient status (InPatient and OutPatient), factor analysis was applied to the data. The six largest eigenvalues obtained from factor analysis were 21.653, 3.432, 1.312, 1.129, 1.088 and 1.023, which are all greater than 1. The corresponding factors were therefore retained for further analysis. The Scree plot technique is also used to determine the number of factors to retain, which explain maximum variation in the data. Figure 5 shows the Scree plot. It is seen in this figure that the two curves of eigenvalues and

Figure 4. Percentage distribution of InPatients by Race and Gender with numinpatientStay30 = 1.



parallel analysis meet at number of components (or Factors) equal to six, suggesting that six factors be retained. As both the eigenvalues and the scree plot have identified the number of factors as 6, it was decided to proceed with the first six factors for further analysis.

The results of the proportion of variance explained by each factor and the cumulative variance are given in Table 1. As seen in the table, the cumulative variance explained by these first six factors is as high as 69.2% of the total variance.

In order to assess the mental health status among communities, the following three groups are considered:

Race: Black, White, Others, Unknown

Gender: Male, Female

Patient Status: inpatient, Outpatient

Figure 5. Scree diagram to identify the number of factors to be retained.

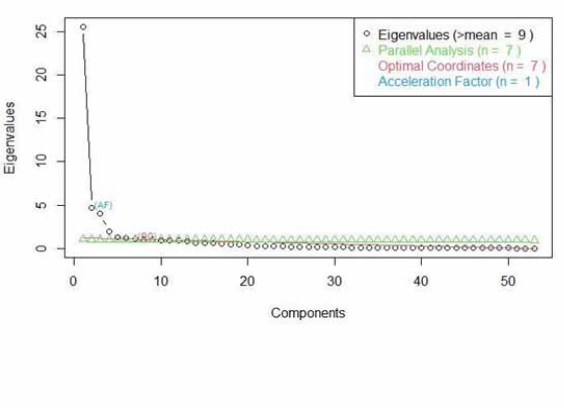


Table 1. Proportion of variance explained by the six factors

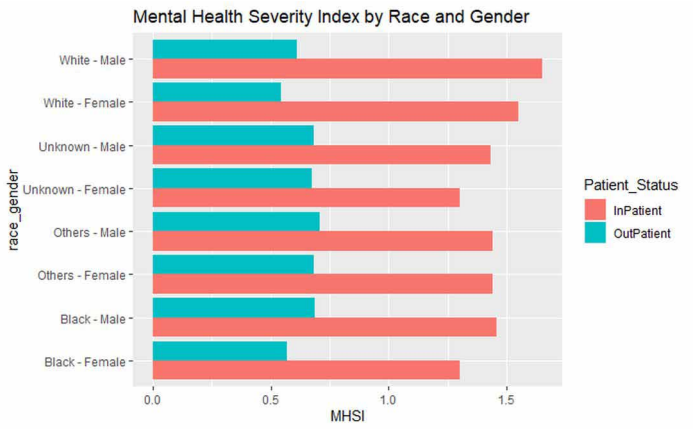
Description	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
SS loadings	19.347	6.613	6.067	3.103	0.829	0.703
Proportion Var	0.365	0.125	0.114	0.059	0.016	0.013
Cumulative Var	0.365	0.490	0.604	0.663	0.678	0.692

MHSI scores are evaluated for each patient by using formula (1) given in the Machine Learning Algorithms section above. From these individual MHSI scores, the cross tabulated mean scores or Mental health severity index(MHSI) scores are evaluated for the above three groups ie. race, gender and patient status. These results are presented in Table 2 and a bar plot in Figure 6. From the bar plot of Figure 6, it is observed that in general, MHSI scores for inpatients are higher than for outpatients in all races and genders. This clearly shows the distinction between the mental health severity for inpatients and outpatients.

Table 2. MHS Index race-wise, gender-wise and patient-status-wise

Patient status	Gender	Race	MHSI Score
Outpatient	Male	Unknown	0.7155423
Outpatient	Male	Others	0.7728059
Outpatient	Male	Black	0.7266442
Outpatient	Male	White	0.6436330
Outpatient	Female	Unknown	0.6881292
Outpatient	Female	Others	0.7507318
Outpatient	Female	Black	0.5957075
Outpatient	Female	White	0.5739292
inpatient	Male	Unknown	1.1698302
inpatient	Male	Others	1.2524319
inpatient	Male	Black	1.3929018
inpatient	Male	White	1.6482094
inpatient	Female	Unknown	0.9796764
inpatient	Female	Others	1.2876218
inpatient	Female	Black	1.2092737
inpatient	Female	White	1.4706496

Figure 6. Bar Plot of MHSI Scores for different races, genders and patient status



To assess the mental health severity among the inPatients, a bar plot of MHSI scores is shown for inPatients only in Figure 7, which displays the MHSI scores of inPatients belonging to different races and genders. From this plot, it is observed that in each race among inPatients, the MHSI scores for males are slightly higher than those for females, except in the case of the Others category. It is also observed that among the races and genders, white males and white females appear to have slightly higher mental health severity scores than all other categories of inPatients.

Random Forest Model Results

A Random Forest model is fitted to the training dataset to evaluate the probability of admission as InPatient. Results of the fitted model for the training dataset are presented in Table 3. These results contain the confusion matrix and the Out Of the Bag(OOB) estimate of error rate. The Random Forest model fitted the training data very well with an overall OOB Error rate as low as 3.69% and an accuracy of 96.31%.

The fitted model is then applied to the test dataset and the confusion matrix is generated to test the accuracy of the model for the test data. These results are presented in Table 4. In the test dataset also, the accuracy is 96.11%, as can be seen from these results.

The Random Forest model also identified the variables of importance based on the meanDecreseGini criterion (Breiman, 2001). Figure 8 shows the top ten variables identified by the Random forest model based on this criterion. From this plot, it is observed that the variables mhIssueSeverity, hasDrugUseAterWalking, hasCriticizedrug, doesWantReduceDrug etc. are the most important variables for classification and for evaluating the probabilities. Most of these important variables relate to the severity of mental health and drug abuse and the study has highlighted the fact that drug related variables contribute more towards the severity of mental health.

Figure 7. Bar Plot of MHSI among inpatients – races and sexes

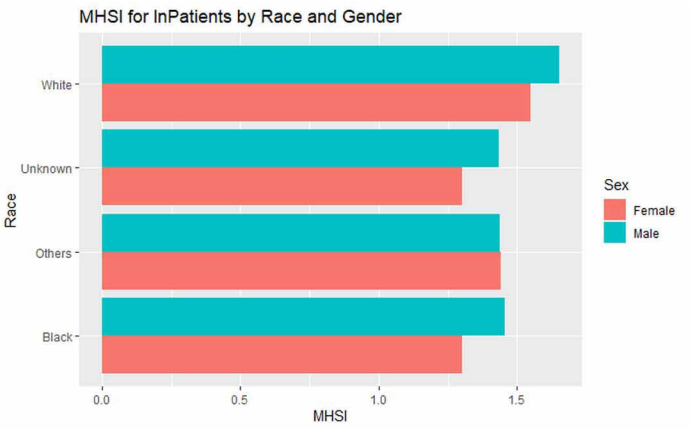


Figure 7a.

```
Call:
 randomForest(formula = traindata$numInpatientInLast30 ~ ., data = traindata)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 11

        OOB estimate of  error rate: 3.69%
Confusion matrix:
      0    1 class.error
0 56470  542 0.009506771
1  1844 5824 0.240479917
```

Figures 7b.

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 14115   138
         1   491  1425

               Accuracy : 0.9611
                 95% CI : (0.958, 0.964)
    No Information Rate : 0.9033
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7977

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9664
            Specificity : 0.9117
         Pos Pred Value : 0.9903
         Neg Pred Value : 0.7437
             Prevalence : 0.9033
         Detection Rate : 0.8730
   Detection Prevalence : 0.8815
      Balanced Accuracy : 0.9390

       'Positive' Class : 0
```

Figure 8. Importance of Variables graph by Random Forest Model

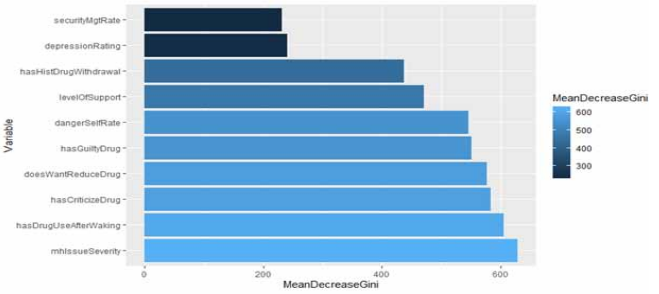
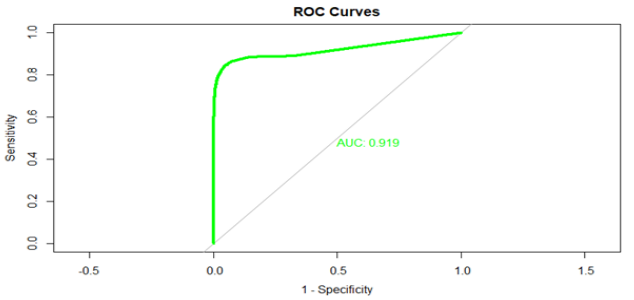


Figure 9. ROC Curve for the Random Forest Model



To evaluate the performance of the Random Forest model, the Receiver Operating Characteristic (ROC) curve is developed, and the Area under the Curve (AUC) is evaluated (Kun-Pie Lin et al., 2016). These results are displayed in Figure 9. From this plot, it is observed that the AUC value = 0.919, which is quite high, indicating that the Random Forest model has separated the two categories inpatient and outpatient very well, and the accuracy of the model is quite high.

Shiny App for Probability Predictions

A Shiny app is developed to interactively view the Random Forest model results and to predict the probabilities of admission of the existing clients and new Clients. The shiny app is hosted on shinyapps.io website and has eight menu options as shown in Figure 10.

The first Menu "About" of the app displays information about the details of the shiny app as sgown in Figure 10. The second Menu "Data Details" gives details of the dimensions of the data such as number of observations, number of variables, storage requirements, etc. The third menu option, "Explore Data,'' displays the first few records of the dataset utilized. The fourth Menu option, "R.F.Model'' displays the results of the Random Forest model, the OOB estimation of Error rate, the Confusion Matrix, etc., as shown in Figure 11. As seen in this plot, the model has fitted the data very accurately with an overall OOB estimation of an Error rate of 3.51%.

Figure 10. Display of the Main Menu of the Shiny App

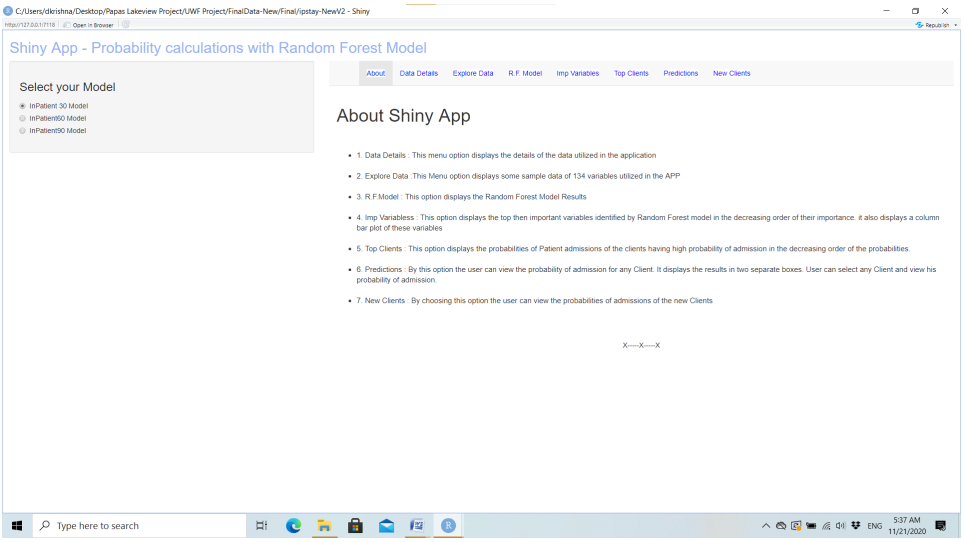
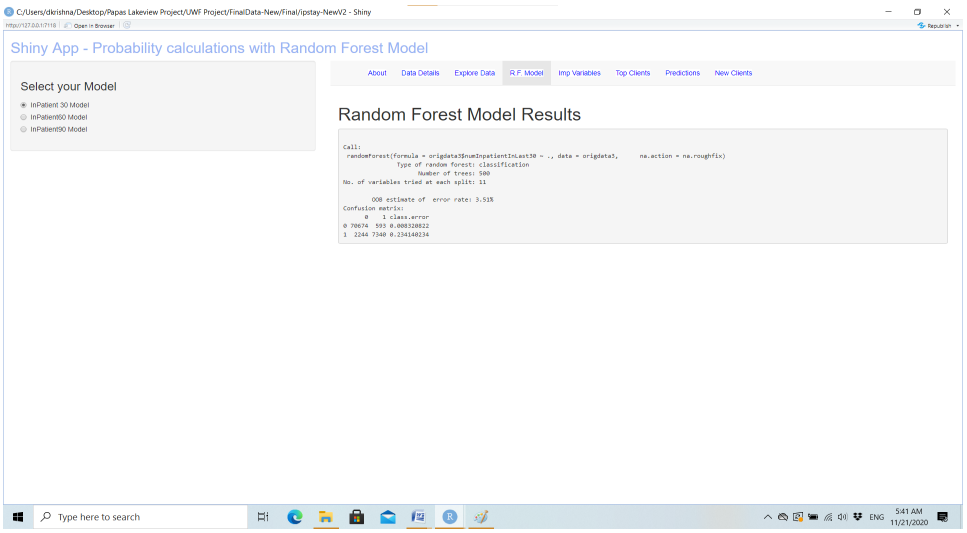


Figure 11. Display of Shiny app Random Forest model Results



The fifth Menu, “Imp Variables,’’ displays the top 10 important variables identified by the Random Forest Model. The sixth Menu, “Top Clients,” displays the top 1000 clients with the highest probabilities of inpatient admission. It shows a table consisting of the clientID column and the column of the corresponding probability of inpatient admissions.

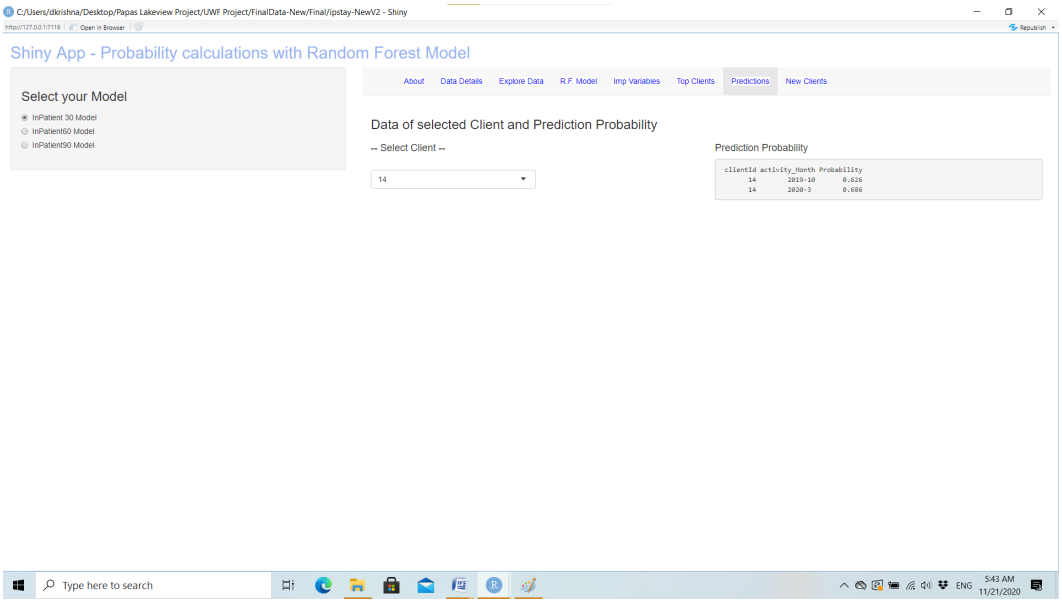
The seventh Menu option, “Predictions,” displays an interactive screen, as shown in Figure 12. On the left side of the Menu, the user can select one of the options: inpatient30 Model / inpatient

60 Model/ inpatient 90 Model for 30/60/90 days stay from the "Select your Model" radio buttons. A list box "Select Client" is provided to select any clientID from the existing clients for whom the probabilities of inpatient admissions are required. Once the clientID is selected, the text box "Prediction Probabilities" is displayed, consisting of the probabilities of inpatient admissions. This text box displays a table with the results clientID, activityMonth and the corresponding probability. As the input data consists of various admission records of the same patient at different dates, denoted by activityMonth, this table displays the probabilities for all these different dates of admission.

The user can select any other model and any clientID, and the Shiny App automatically recalculates the probabilities for the selected options and displays the results. From the plot of Figure 12, one can see that for the client with clientID 14, the prediction probability for activity-month 2019-10 is 0.626, and for the activity month 2020-5 it is 0.686.

The eighth Menu option, "New Clients," allows the user to upload the data of new patients and view the probability of that patient becoming an inpatient. .

Figure 12. Display of Predictions screen of shiny app



Conclusions

An attempt is made in this study to evaluate the probability of patient admission as inpatient utilizing the data collected from Lakeview Center Inc. The distribution of patients with mental health disorders is evaluated race-wise, gender-wise and patient-status-wise. A novel Mental Health Severity Index (MHSI) is developed using factor analysis. Utilizing the MHSI scores, the mental health status of the patients is studied for the categories of race, gender, and patient status. It was found that both male and female Caucasian patients appear to have slightly higher mental health severity index compared to other races. The mental health severity of InPatients is higher than the OutPatients.

The machine learning technique RandomForest (RF) model is applied to the patients data to assess the probability of readmission. The Random Forest Model has identified the variables mhIssueSeverity (mental health severity) and other drug abuse related variables as important variables for classification. It appears that drug abuse related variables play an important role in mental health severity. The RF

model could accurately classify the patients with an overall OOB estimate of Error rate of 3.51%. The accuracy of the model is tested by various metrics including confusion matrix, OOB Error rate, ROC Curve and Area Under the Curve (AUC) and all the metrics have yielded high values indicating the accuracy of the fitted model. An interactive shiny app is also deployed on shinyapps.io website to display the results of the Factor Analysis model and the results of Random Forest model.

A limitation of the study is the unavailability of clinical data; we plan to acquire clinical data and study the impact of comorbidities on readmissions of patients with mental illnesses.

REFERENCES

Alam, Z., Rahman, S., & Rahman, M. S. (2019). A Random Forest based predictor for medical data classification using feature ranking. *Informatics in Medicine Unlocked*, *15*, 100180. doi:10.1016/j.imu.2019.100180

Baek, H., Cho, M., Kim, S., Hwang, H., Song, M., & Yoo, S. (2018). Analysis of length of hospital stay using electronic health records: A statistical and data mining approach. *PLoS One*, *13*(4), e0195901. doi:10.1371/journal.pone.0195901 PMID:29652932

Benjenk, I., & Chen, J. (2018). Effective mental health interventions to reduce hospital readmission rates: A systematic review. *Journal of Hospital Management and Health Policy*, *2*, 45. doi:10.21037/jhmhp.2018.08.05 PMID:30283917

Blankers, M., van der Post, L., & Dekker, J. (2020). Predicting hospitalization following psychiatric crisis care using machine learning. *BMC Medical Informatics and Decision Making*, *20*(1), 332. doi:10.1186/s12911-020-01361-1 PMID:33302948

Breiman, L. (2001). Random Forests. *J Mach Learn*, *45*(1), 5–32. doi:10.1023/A:1010933404324

Cardarelli, R., Horsley, M., Ray, L., Maggard, N., Schilling, J., Weatherford, S., Feltner, F., & Gilliam, K. (2018, February). Reducing 30-day readmission rates in a high-risk population using a lay-health worker model in Appalachia Kentucky. *Health Education Research*, *33*(1), 73–80. doi:10.1093/her/cyx064 PMID:29474535

Cattell, R. B. (1966). The Scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276. doi:10.1207/s15327906mbr0102_10 PMID:26828106

Degenhardt, F., Seifert, S., & Szymczak, S. (2019). Evaluation of variable selection methods for random forests and omics data sets. *Briefings in Bioinformatics*, *20*(2), 492–503. doi:10.1093/bib/bbx124 PMID:29045534

Enhancing Mental Health Care Transitions Reduces Unnecessary Costly Readmissions. (2017). Retrieved from https://www.healthcatalyst.com/wp-content/uploads/2017/05/Enhancing-Mental-Health-Care-Transitions-Reduces-Unnecessary-Costly-Readmissions.pdf

Flaks-Manov, N., Topaz, M., Hoshen, M., Balicer, R. D., & Shadmi, E. (2019). Identifying patients at highest-risk: The best timing to apply a readmission predictive model. *BMC Medical Informatics and Decision Making*, *19*(1), 118. doi:10.1186/s12911-019-0836-6

Genuer, R. (2012). Variance reduction in purely random forests. *Journal of Nonparametric Statistics*, *24*(3), 1–20. doi:10.1080/10485252.2012.677843

Gopalakrishna, G., Ithman, M., & Malwitz, K. (2015). Predictors of Length of Stay in an Acute Psychiatric Hospital. *International Journal of Psychiatry in Clinical Practice*, *19*(4), 238–244. doi:10.3109/13651501.2015.1062522 PMID:26073671

Jansen, L., Schijndel, M. V., Waarde, J. V., & Busschbach, J. V. (2018). Health-economic outcomes in hospital patients with medical-psychiatric comorbidity: A systematic review and meta-analysis. *PLoS One*, *13*(3), e0194029. doi:10.1371/journal.pone.0194029 PMID:29534097

Kabacoff, R. (n.d.). *Principal components and factor analysis*. Retrieved from https://www.statmethods.net/advstats/factor.html

Ledesma, R. D., Valero-mora, P. M., & Macbeth, G. (2015). The Scree Test and the Number of Factors: A Dynamic Graphics Approach. *The Spanish Journal of Psychology*, *18*, 18. doi:10.1017/sjp.2015.13 PMID:26055575

Liang, L., Moore, B., & Soni, A. (2020). *National inpatient Hospital Costs: The Most Expensive Conditions by Payer, 2017: Statistical Brief #261*. Agency for Healthcare Research and Quality.

Lin, K. P., Chen, P. C., Huang, L. Y., Mao, H. C., & Chan, D. D. (2016). Predicting inpatient Readmission and Outpatient Admission in Elderly. *Medicine*, *95*(16), e3484. doi:10.1097/MD.0000000000003484 PMID:27100455

Liu, W., Stansbury, C., Singh, K., Ryan, A. M., Sukul, D., Mahmoudi, E., Waljee, A., Zhu, J., & Nallamothu, B. K. (2020). Predicting 30-day hospital readmissions using artificial neural networks with medical code embedding. *PLoS One*, *15*(4), e0221606. doi:10.1371/journal.pone.0221606 PMID:32294087

Mekhaldi, R. N., Caulier, P., Chaabane, S., Chraibi, A., & Piechowiak, S. (2020). Using Machine Learning Models to Predict the Length of Stay in a Hospital Setting. *Trends and Innovations in Information Systems and Technologies*, *1159*, 202–211. doi:10.1007/978-3-030-45688-7_21

Narkhede, S. (2018). *Understanding AUC – ROC Curve*. Retrieved from: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Owens, P. L., Fingar, K. R., McDermott, K. W., Muhuri, P. K., & Heslin, K. C. (2019). Inpatient Stays Involving Mental and Substance Use Disorders, 2016: Statistical Brief #249. Agency for Healthcare Research and Quality.

Piper, K. (2020). *Medicare, Medicaid, health reform*. Retrieved from http://www.piperreport.com/

Prayaga, L., Devulapalli, K., & Prayaga, C. (2020). Combining clustering and factor analysis as complimentary techniques. *International Journal of Data Analytics*, *1*(2), 48–57. doi:10.4018/IJDA.2020070104

Sporinova, B., Manns, B., Tonelli, M., Hemmelgarn, B., MacMaster, F., Mitchell, N., Au, F., Ma, Z., Weaver, R., & Quinn, A. (2019). Association of Mental Health Disorders With Health Care Utilization and Costs Among Adults With Chronic Disease. *JAMA Network Open*, *2*(8), e199910. doi:10.1001/jamanetworkopen.2019.9910 PMID:31441939

Šprah, L., Dernovšek, M. Z., Wahlbeck, K., & Haaramo, P. (2017). Psychiatric readmissions and their association with physical comorbidity: A systematic literature review. *BMC Psychiatry*, *17*(1), 2. doi:10.1186/s12888-016-1172-3 PMID:28049441

Tsai, P. F., Chen, P. C., Chen, Y. Y., Song, H. Y., Lin, H. M., Lin, F. M., & Huang, Q. P. (2016). Length of Hospital Stay Prediction at the Admission Stage for Cardiology Patients Using Artificial Neural Network. *Journal of Healthcare Engineering*, *2016*, 7035463. doi:10.1155/2016/7035463 PMID:27195660

Upadhyay, S., Stephenson, A. L., & Smith, D. G. (2019). Readmission Rates and Their Impact on Hospital Financial Performance: A Study of Washington Hospitals. *Inquiry*, *56*. doi:10.1177/0046958019860386 PMID:31282282

Wan, H., Zhang, L., Witz, S., Musselman, K. J., Yi, F., Mullen, C., Benneyan, J., Zayas-Castro, J., Rico, F., Cure, L., & Martinez, D. (2016). A literature review of preventable hospital readmissions: Preceding the Readmissions Reduction Act. *IIE Transactions on Healthcare Systems Engineering*, *6*(4), 193–211. doi:10.1080/19488300.2016.1226210

Winerman, L. (2017). By the numbers: The cost of treatment. *Monitor on Psychology*, *48*(3).

Wongvibulsin, S., Wu, K. C., & Zeger, S. L. (2019). Clinical risk prediction with random forests for survival, longitudinal, and multivariate (RF-SLAM) data analysis. *BMC Medical Research Methodology*, *20*(1), 1. doi:10.1186/s12874-019-0863-0 PMID:31888507

Lakshmi Prayaga is a Professor in the Department of Information Technology, University of West Florida. Her research focuses on applications of technology in healthcare, sports medicine, management and training. Topics of interest include robotics, data visualizations and analytics. She has co-authored books on robotics, Android App development, beginning game programming, programming the web with ColdFusion and XHMTL, and using game programming to teach computer science concepts. She has also published numerous publications in International journals and conferences. She teaches graduate and undergraduate courses in Data Analytics, Data Visualizations, Machine learning and Script programming. She has an Ed.D. in Instructional Technology and a M.S. in Software Engineering, both from UWF and an MBA from Alabama A&M University.

Krishna Devulapalli is a retired scientist from the Indian Institute of Chemical Technology, India. Currently he is a Freelance Data Scientist and is recognized as TapChief Expert in Data Science by TapChief, India. His research interests include applied statistics in multiple domains such as correlations among physico-chemical attributes of substances, healthcare analytics, BioInformatics, BioStatistics, Chemometrics, Reliability Studies, Pattern Recognition, Neural Networks, Rule Based Systems, Machine Learning etc. He has published more than 30 research papers in various journals and also presented number of papers in conferences and seminars. He has contributed some chapters in books related to Medical Statistics. He is a Member of various professional Societies like Indian Society of Medical Statistics, Computer Society of India, Indian Society of Analytical Scientists etc. He is recognized as a Guest Faculty in various organizations like Statistics Department, Osmania University, NIPER Guwahati, IICT, CSI, CMC, etc.

Chandra Sekhar Prayaga is currently Professor of Physics, University of West Florida (UWF). He has a Ph.D. in Physics from the Indian Institute of Science, Bangalore, India (1975), where he was also a faculty member from 1981 to 1987. He has 40-plus years of experience in teaching physics, and has helped raise more than $3 million in funding for research and projects involving University of West Florida faculty and students. His current research interests include optical and electronic properties of liquid crystals, Langmuir-Blodgett films, phase transitions and laser spectroscopy, physics education and data analytics. He mentors undergraduate student research projects, and coordinates summer camps on science and technology for middle and high school students. He is cofounder of Discovery Spot: A technology playground for middle and high school students to experience the latest technologies with hands-on activities, such as building smart Cities using IoT. He is co-author of a book, "Robotics: A Project-Based Approach," Cengage Publishers (2014).