Kaushik, Sakshi; Sabharwal, Alka; Grover, Gurprit

Article

Extracting relevant predictors of the severity of mental illnesses from clinical information using regularisation regression models

Statistics in Transition new series (SiTns)

Provided in Cooperation with:
Polish Statistical Association

Suggested Citation: Kaushik, Sakshi; Sabharwal, Alka; Grover, Gurprit (2022) : Extracting relevant predictors of the severity of mental illnesses from clinical information using regularisation regression models, Statistics in Transition new series (SiTns), ISSN 2450-0291, Sciendo, Warsaw, Vol. 23, Iss. 2, pp. 129-152,
https://doi.org/10.2478/stattrans-2022-0020

This Version is available at:
https://hdl.handle.net/10419/266311

STATISTICS IN TRANSITION new series, June 2022
Vol. 23 No. 2, pp. 129–152, DOI 10.2478/stattrans-2022-0020
Received – 23.08.2019; accepted – 19.07.2021

Extracting relevant predictors of the severity of mental illnesses from clinical information using regularisation regression models

Sakshi Kaushik[1], **Alka Sabharwal**[2], **Gurprit Grover**[3]

ABSTRACT

Mental disorders are common non-communicable diseases whose occurrence rises at epidemic rates globally. The determination of the severity of a mental illness has important clinical implications and it serves as a prognostic factor for effective intervention planning and management. This paper aims to identify the relevant predictors of the severity of mental illnesses (measured by psychiatric rating scales) from a wide range of clinical variables consisting of information on both laboratory test results and psychiatric factors . The laboratory test results collectively indicate the measurements of 23 components derived from vital signs and blood tests results for the evaluation of the complete blood count. The 8 psychiatric factors known to affect the severity of mental illnesses are considered, viz. the family history, course and onset of an illness, etc. Retrospective data of 78 patients diagnosed with mental and behavioural disorders were collected from the Lady Hardinge Medical College & Smt. S.K, Hospital in New Delhi, India. The observations missing in the data are imputed using the non-parametric random forest algorithm. The multicollinearity is detected based on the variance inflation factor. Owing to the presence of multicollinearity, regularisation techniques such as ridge regression and extensions of the least absolute shrinkage and selection operator (LASSO), viz. adaptive and group LASSO are used for fitting the regression model. Optimal tuning parameter λ is obtained through 13-fold cross-validation. It was observed that the coefficients of the quantitative predictors extracted by the adaptive LASSO and the group of predictors extracted by the group LASSO were comparable to the coefficients obtained through ridge regression.

Key words: adaptive LASSO, group LASSO, mental disorder, multicollinearity, random forest imputation, ridge regression, severity of an illness.

[1] Quartesian, Chennai, Tamil Nadu, India. E-mail: sakshi2007@gmail.com. ORCID: https://orcid.org/0000-0002-4219-1488.

[2] Department of statistics, Kirori mal College, University of Delhi, India. E-mail: alkasabharwal@kmc.du.ac.in. ORCID: https://orcid.org/0000-0002-8252-8284.

[3] Department of statistics, Faculty of mathematical sciences, University of Delhi, India. E-mail: gurpritgrover@yahoo.com. ORCID: https://orcid.org/0000-0003-2051-4810.

1. Introduction

Mental disorders are common non-communicable diseases rising with epidemic rates globally with over one third of people in most countries reporting sufficient criteria to be diagnosed at some point in their life (World Health Organization, 2000). The determination of the severity of mental illness has important clinical implications. Measures of severity help in the evaluation of outcome in treatment studies and may be used as a meaningful endpoint in clinical practice (Zimmerman, Morgan, & Stanton, 2018). It serves as an important prognostic factor for effective intervention planning and management.

Blood has been regarded as a source of information on illness and health since ancient times. With the emergence of experimental medical techniques in the mid-1800s, studies of blood have been carried out to identify physical characteristics that could be used to diagnose a psychiatric illness or assess the severity of its symptoms (Bahn et al. (2013)). In recent years, studies have increasingly been made on reports of blood tests such as platelets to understand psychiatric disorders, assess their impact on the severity of illness and evaluate the pharmacological properties of psychiatric drugs. Canan et al. (2012) showed that mean platelet volume (MPV) values were high in patients with major depression and decreased treatment.

Various general psychiatric aspects (such as family history, onset and course of illness, number of episodes, etc.) commonly observed across all mental disorders significantly impact the diagnosis, prognosis, severity, and remission of mental illness. Various studies in the past have identified family history as a potential risk factor for developing a mental illness and have associated it with seriousness indicators of illness such as recurrence, impairment, and age at onset (Laursen et al. (2005); Milne et al. (2009)). The number of episodes plays a cardinal role in determining the severity of illness. It has been observed that patients with a higher number of episodes have a more severe outcome (Marzo et al. (2006)). Such patients are more likely to relapse than those with fewer episodes. The onset of illness refers to how the symptoms of the disease begin to appear in a patient. The onset of symptoms in mental illness is known to be a prognostic indicator of its severity. The course of illness refers to the usual trajectory the disease follows from the onset of the first symptom until recovery or death. The course reflects the different grades of the severity of the illness. It has been observed that the chronic course of illness is associated with higher levels of depressive and somatic symptoms and greater mental dysfunction (Stegenga et al. (2010)). Studies in the past have shown that a higher amount of alcohol and tobacco consumption is found to be associated with greater severity of illness (Goldstein, Velyvis, & Parikh (2006); Krishnadas et al. (2012); Dwivedi, Chatterjee, & Singh (2017)). Further, Brådvik (2018) suggested that suicidal ideation and self-harm are

related to mental illness. Insight of an illness is defined as a patient's capacity to understand the nature, significance, and severity of his or her illness. Literature suggests that insight interacts with the trajectory of the person's illness and predicts outcome in psychosis. It is found that the severity of illness increases with a progressive loss of insight (McDaniel, Edland, & Heyman (1995); Jacob (2016)). Although each mental disorder has its own complications and risks involved, a certain illness is considered to be more severe than others owing to the level of disability caused by them. These illnesses include disorders that produce psychotic symptoms, such as schizophrenia, and severe forms of other disorders, such as major depression and bipolar disorder (World Health Organization (2003)). Thus, different types of mental disorders have different severity levels. These worsen the symptoms and the course of mental illness.

Missing values are commonly encountered in medical datasets, especially mental disorders. Performing analysis with only complete patient datasets leads to a smaller sample size resulting in a loss of statistical power and bias in the estimation of parameters. Multiple imputation is a robust technique for handling missing data. In this approach, a prediction of the missing data is made using the existing data from other variables. There are several imputation methods available based on different statistical models such as regression, Random Forest, etc.

The inclusion of a large number of variables in a regression model often results in multicollinearity. Multicollinearity refers to high inter-correlations or inter-associations among the independent variables. The existence of multicollinearity affects the estimation of the model as well as the interpretation of the results. It leads to biased coefficient estimation and a loss of power. The regression models based on regularization techniques such as l_1 (Least Absolute Shrinkage and Selection Operator (LASSO) Regression; Tibshirani (1996)), l_2 (Ridge Regression; Hoerl and Kennard (1970)) and elastic net (Zou and Hastie (2005)) model, can solve this problem by adding a penalty to model parameters (except intercept) so the model generalizes the data instead of overfitting. Both ridge and LASSO regression belong to the class of penalised regression models. The key difference between these two techniques lies in the penalty that is imposed on the model. LASSO selects features that are predictive of the outcome by penalizing irrelevant features' weights to zeros while the ridge regression penalizes the irrelevant features by converging their weights to zero but never exactly equal to zero. Thus, both LASSO and ridge identify relevant predictors, however, LASSO is considered to be advantageous over ridge since it performs variable selection as well.

Many previous studies have used regularization regression models with multiply imputed data to determine relevant predictors from a class of independent variables (Jain (1985)). Brewer et al. (2009) used ridge and LASSO regression to predict an

individual's score on the Unified Parkinson Disease Rating Scale based on Advanced Sensing for Assessment of Parkinson's disease (ASAP) data. Haenisch et al. (2016) identified protein analytes from a blood-based panel as potential biomarkers for diagnosing bipolar disorder using LASSO regression. Upadhya & Cheeran (2018) compared six regression techniques including ridge and LASSO to predict the Parkinson disease severity score using speech features.

Although, LASSO is an oracle procedure for simultaneously achieving consistent variable selection and optimal estimation (prediction), however, there are many solid arguments against the LASSO oracle statement (Zou (2006)). Further, Zhao and Yu (2006) showed that variable selection with LASSO could be consistent if the model satisfies some irrepresentable conditions. These conditions are restrictive and for data sets that fail to satisfy them, LASSO may not select the correct model. Therefore, to recognize relevant predictors some improvements of LASSO model have been proposed. The adaptive LASSO is a new version of the LASSO, in which adaptive weights (data driven) are used for penalizing different coefficients in the l_1 penalty. It also enjoys the oracle properties (Zou (2006)).

In some problems, when the predictors belong to pre-defined groups or factors; for example, collections of indicator (dummy) variables for representing the levels of a multiple categorical predictor such as onset and course of illness, LASSO and the adaptive LASSO are not suitable for variable selection as they are designed for selecting individual input variables. When directly applied to model they tend to select based on the strength of individual derived input variables rather than the strength of groups of input variables, often resulting in selecting more factors than necessary. In this situation it may be desirable to shrink and select the members of a group together. The group LASSO is a generalization of the LASSO for doing group-wise variable selection by introducing a suitable extension in the penalty of LASSO (Yuan & Lin (2006)).

This paper aims to identify relevant predictors for estimating the severity of mental illness (measured by psychiatric rating scales) from a wide range of clinical variables consisting of information on both laboratory test results and psychiatric aspects. The laboratory test results collectively indicate measurements on 23 components derived from vital signs and blood tests (complete blood count (CBC)) results such as diastolic and systolic blood pressure (DBP, SBP), pulse rate, haemoglobin (hb), red blood cell (RBC), etc. *Further,* 8 psychiatric factors known to affect severity of mental illness are considered, viz. family history (fh), number of episodes experienced by the patient (epi), onset and course of illness (onset), etc. The impact of covariates age and gender is also studied.

To achieve our aim, firstly missing values in the data consisting of 34 variables are imputed using the non-parametric random forest algorithm. Secondly, the problem of

multicollinearity between explanatory variables is detected based on variance inflation factor (VIF). Since coefficients estimated from linear regression are biased in the presence of multicollinearity, thus, regularization techniques are used for fitting the regression model. Thirdly, prior to application of regularized regression models to the data, the dummy coding is applied to the 8 categorical variables consisting of clinical information on psychiatric factors related to mental disorders. These 8 categorical variables transform into 26 dichotomous variables with each variable representing each category. Fourthly, the ridge regression is applied to a total of 51 regressors including 25 quantitative and 26 binary variables with response variable being psychiatric rating scale score (RSS). Next, the adaptive LASSO is applied to the 25 quantitative variables including clinical variables consisting of information on vital signs and laboratory test result reports, age and number of episodes to extract the relevant predictors of RSS. Finally, the group LASSO is applied to the 26 dichotomous variables representing 8 groups of psychiatric variables to extract the relevant groups.

To the best of our knowledge, none of the previous studies has attempted to assess the relationship of such diverse and wide range of predictors with the severity of mental illness. The outline of the rest of the paper is as follows: Section 2 describes the dataset used for the application of methods discussed in Section 3. In Section 4, the application of the model to the dataset along with the results is discussed. The paper is concluded with a discussion in Section 5.

2. Data description

The retrospective data considered for this study consisted of 146 patients diagnosed with mental and behavioural disorders as per DSM-V (American Psychiatric Association (2013)) and ICD-10 (World Health Organization (1992)), collected from the Department of Psychiatry, Lady Hardinge Medical College & Smt. S.K, Hospital, New Delhi, India for the calendar year 2013-2014. The patients were diagnosed with Bipolar Affective Disorder (BPAD), schizophrenia, depression, and other disorders. The others category includes disorders, viz. Acute Transient Psychotic Disorder (ATPD), dementia, psychotic disorder: Not otherwise Specified (NOS), and alcohol abuse. Out of these 146 patients, only 78 patients could be included in the study as the clinical information on psychiatric variables as well as laboratory test result reports were available for them. The dataset of the remaining 68 patients was completely unavailable with respect to the variables considered in the study (i.e. either complete information on psychiatric variables and/or laboratory test reports were unavailable or both) and hence they were excluded.

The severity of mental disorders considered in this study is measured by various psychiatric rating scales recommended for each disorder. Since rating points as well as

range of total scores vary in different psychiatric rating scales, thus, to maintain homogeneity, the total scores of these psychiatric rating scales are scaled down to 100 and denoted as RSS. RSS is the response variable under consideration. The regressors considered suitable for the study are classified into two categories: 1) clinical information related to vital signs and laboratory test result reports consisting of 23 variables, viz. Diastolic Blood Pressure (DBP) (mmHg), Systolic Blood Pressure (SBP) (mmHg), Pulse Rate (pulse per min), Haemoglobin (hb) (g/dL), Red Blood Cell (RBC) (million/μL), Mean Corpuscular Hemoglobin (MCH) (pg), Mean Corpuscular Volume (MCV) (fL), Mean Corpuscular Hemoglobin Concentration (MCHC) (g/dL), Total Leukocyte Count (TLC) (cells/L), Platelet (thousand/μL), Blood Urea (b.urea) (mg/dL), Serum Creatinine (sr.cr) (mg/dL), Sodium (NA) (mEq/L), Potassium (K) (mEq/L), Serum Bilirubin (S.Bil) (mg/dL), Alanine Aminotransferase (ALT) (IU/L), Aspartate Aminotransferase (AST) (IU/L), Alkaline Phosphatase (ALP) (IU/L), Total Cholesterol (TCHOL) (mg/dL), High-Density Lipoprotein (HDL) (mg/dL), Triglycerides (S.TG) (mg/dL), Haematocrit or Packed-Cell Volume (PCV) (%) and Random Blood Sugar (RBS) (mg/dL). 2) The second category consists of clinical information on 8 psychiatric variables, viz. family history (fh), number of episodes experienced by the patient (epinew), onset of illness (onset), course of illness (course), alcohol or tobacco abuse (abuse), type of disorder (discode), suicidal ideation or self-harm (sui_sharm) and insight of illness (insight). The codes used for categorical variables are defined as follows:

i. Family history (fh): '0' and '1' indicate absence and presence of family history, respectively.
ii. Onset of illness (onset): '1', '2', '3', '4' and '5' indicate abrupt/sudden, acute, chronic, insidious, and sub-acute, respectively.
iii. Course of illness (course): '1', '2', '3' and '4' indicate continuous and progressive, continuous, episodic, and fluctuating, respectively.
iv. Abuse: '1' and '2' indicate absence and presence of alcohol or tobacco abuse, respectively.
v. Type of disorder (discode): '1', '2', '3' and '4' indicate Bipolar affective disorder (BPAD), Depression/Depressive disorder, Others, and Schizophrenia, respectively.
vi. Suicidal ideation or self-harm (sui_sharm): '1' implies absence while '2' indicates presence of suicidal ideation and/or self-harm in the patient.
vii. Insight: The grades of insight are as suggested by Sadock (2009).

Two other covariates considered are: age and gender. For gender, categories '1' and '2' indicate female and male, respectively.

3. Methods

Let there be n observations of a response variable Y and p associated predictor variables $X=(X_1, X_2,..., X_p)^T$. In this study, the response variable Y indicates the severity of illness quantified in terms of the total score of the psychiatric rating scale, denoted as RSS. (Here, $p=33$ and $n=78$). Out of these 33 predictors, $(X_1, X_2,..., X_{23})$ represent 23 features of laboratory test results and vital signs, viz. $X_1 \sim DBP$, $X_2 \sim SBP$, $X_3 \sim Pulse\ Rate$, ..., $X_{23} \sim PCV$, X_{24}, X_{25} and X_{26} represent covariates age, gender and number of episodes while the remaining 7 variables represent thecategorical predictors of psychiatric factors, i.e. $X_{27} \sim family\ history(fh)$, $X_{28} \sim onset\ of\ illness(onset)$,...., $X_{33} \sim Insight$.

3.1. Method of imputation of missing observations

For imputing the missing values in the predictors, the imputation method given by Stekhoven and Bühlmann (2012) is used. Under this method, the missing values are predicted using a Random Forest (RF) trained on the observed parts of the dataset. The performance of the imputation method is assessed using the normalized root mean squared error (NRMSE) (Oba et al. (2003)) for the continuous variables and the proportion of falsely classified entries (PFC) over the categorical missing values. For both continuous as well as categorical variables, a value close to 0 indicates good performance.

3.2. Multicollinearity detection

Amongst the numerous approaches to detect multicollinearity in the data, namely determinant approach, Farrar and Glauber test (Farrar & Glauber (1967)), condition index (Belsley (1991)), Leamer's method (Greene (1993)) and variance inflation factor (VIF), the VIF is the most commonly used method. Let R_i^2 denote the coefficient of multiple determination of X_i regressed on the remaining $(p-1)$ explanatory variables. For the X_i, VIF is defined as

$$VIF_i = \frac{1}{(1-R_i^2)}. \tag{1}$$

A VIF of 5 or more indicates serious or excessive multicollinearity (Akinwande, Dikko and Samson (2015); Jongh et al. (2015)).

3.3. Dummy coding

Dummy coding is a method of representing a categorical variable into a series of dichotomous variables. For the categorical/qualitative predictors with K-levels, K indicator dummy/binary variables are created. Suppose X_i is a K-level factor

input, then let X_{ij} $(j=1,2,...,K)$ be such that $X_{ij}=I\left(X_i=j\right)$. Together this group of X_{ij} represents the effect of X_i (Hastie, Tibshirani & Friedman (2009)).

3.4. Regularization techniques

Regularization is the process of penalizing the coefficients of predictor variables so that the resulting model has better predictive power. In this paper, the following types of regularization techniques, viz. ridge, group LASSO and adaptive LASSO are used to identify the predictors of severity of illness (Hoerl and Kennard (1970), Hastie, Tibshirani, & Wainwright (2015); James et al. (2013); Yuan and Lin (2006); Zou (2006)).

3.4.1. Ridge regression

Ridge regression is a variant of least squares regression in which the sum of squared errors is minimized, with an upper bound on the sum of squared values of the model parameters. In particular, the ridge regression coefficient estimates are obtained by solving the l_2 optimization problem

$$\underset{\beta_0,\beta}{\text{minimize}}\left\{\sum_{j=1}^{n}\left(y_j-\beta_0-\sum_{i=1}^{p}x_{ji}\beta_i\right)^2\right\} \text{ subject to } \sum_{i=1}^{p}\beta_i^2\le t \tag{2}$$

This equation is equivalent to solving

$$\hat{\beta}_i(\lambda)=\arg\min_{\beta}\left\{\sum_{j=1}^{n}\left(y_j-\beta_0-\sum_{i=1}^{p}x_{ji}\beta_i\right)^2+\lambda\sum_{i=1}^{p}\beta_i^2\right\} \tag{3}$$

where λ, known as the tuning parameter, controls the strength of the penalty. The larger the value of λ, the greater the amount of shrinkage. The second term, $\lambda\sum_{i=1}^{p}\beta_i^2$ is called shrinkage penalty.

3.4.2. Least absolute shrinkage and selection operator (LASSO) regression

LASSO is a regularization and variable selection method for statistical models. Under this technique, the sum of squared errors is minimized, with an upper bound on the sum of the absolute values of the model parameters. The LASSO estimate is defined by the solution to the l_1 optimization problem

$$\underset{\beta_0,\beta}{\text{minimize}}\left\{\sum_{j=1}^{n}\left(y_j-\beta_0-\sum_{i=1}^{p}x_{ji}\beta_i\right)^2\right\} \text{ subject to } \left\|\beta\right\|_1\le t \tag{4}$$

where $\|\beta\|_1 = \sum_{i=1}^{p} |\beta_j|$ is the l_1 norm of β and t is a user-specified parameter. This optimization problem is equivalent to the parameter estimation that follows

$$\hat{\beta}_i(\lambda) = \arg\min_{\beta} \left(\sum_{j=1}^{n} \left(y_j - \beta_0 - \sum_{i=1}^{p} x_{ji}\beta_i \right)^2 + \lambda \|\beta\|_1 \right) \tag{5}$$

where λ is as defined in section 3.4.1. When the optimization problem is minimized, some coefficients shrink to zero, i.e. $\hat{\beta}_i(\lambda) = 0$, for some values of i, resulting in exclusion of some predictors.

Zhao and Yu (2006) showed that variable selection with LASSO could be consistent if the underlying model satisfies some irrepresentable conditions. The irrepresentable condition that should be satisfied is defined as follows:

Let $\boldsymbol{X} = (\boldsymbol{X_1}, \boldsymbol{X_2})'$, where $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ is the subset of $\boldsymbol{X}$ that contains the relevant and irrelevant predictor variables, respectively. Let $\boldsymbol{\beta_1}$be the coefficients of $\boldsymbol{X_1}$. The covariance matrix of $\boldsymbol{X}$ can be computed as $\boldsymbol{\Sigma} = n^{-1}\boldsymbol{X'X}$, which is a symmetric matrix. Let $\boldsymbol{C_{11}} = n^{-1}\boldsymbol{X_1}'\boldsymbol{X_1}$ and $\boldsymbol{C_{22}} = n^{-1}\boldsymbol{X_2}'\boldsymbol{X_2}$ be the covariance matrix of relevant and irrelevant predictor variables, respectively. Let $\boldsymbol{C_{12}} = n^{-1}\boldsymbol{X_1}'\boldsymbol{X_2}$ and $\boldsymbol{C_{21}} = n^{-1}\boldsymbol{X_2}'\boldsymbol{X_1}$ be the covariances between relevant and irrelevant variables. Then, $\boldsymbol{\Sigma}$ can be expressed in block-wise form as

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{C_{11}} & \boldsymbol{C_{12}} \\ \boldsymbol{C_{21}} & \boldsymbol{C_{22}} \end{pmatrix}$$

Assuming $\boldsymbol{C_{11}}$ is invertible, the irrepresentable condition can be defined as:

$$\left| \boldsymbol{C_{21}}\boldsymbol{C_{11}}^{-1} sign(\boldsymbol{\beta_1}) \right|_\infty < 1 \text{ , and the inequality holds elementwise.} \tag{6}$$

These conditions are restrictive and may not hold for all datasets. Thus, the adaptive LASSO model, which is an improvement over LASSO, is used.

3.4.3. Adaptive LASSO

The adaptive LASSO is an extension of LASSO, in which adaptive weights are used for penalizing different coefficients in the l_1 penalty (Zou (2006)). Suppose that $\hat{\beta}$ is a root- n–consistent estimator to β. Let $\widehat{\beta_i}$ be the ordinary least square estimate, $\gamma > 0$, and the weight vector is defined as $\hat{w} = 1/|\hat{\beta}|^{\gamma}$, then the adaptive LASSO estimates $\widehat{\beta^{(n)}}$ are given by

$$\hat{\beta}_i^{(n)} = \arg\min_{\beta} \left(\sum_{j=1}^{n} \left(y_j - \beta_0 - \sum_{i=1}^{p} x_{ji}\beta_i \right)^2 + \lambda_n \sum_{i=1}^{p} \hat{w}_i |\beta_i| \right) \tag{7}$$

3.4.4. Group LASSO

The group LASSO is a generalization of LASSO for performing group-wise variable selection (Yuan and Lin (2006)). Suppose that u predictors are divided into L groups, with u_l being the number in group l. Let X_l represent the predictors corresponding to l^{th} group, with corresponding coefficient vector β_l. The group LASSO minimizes the convex criterion

$$\arg\min_{\beta}\left(\frac{1}{2}\sum_{j=1}^{n}\left(y_j-\beta_0-\sum_{l=1}^{L}X_{jl}\beta_l\right)^2+\lambda\sum_{l=1}^{L}\sqrt{u_l}\left\|\beta_l\right\|\right), \tag{8}$$

where the $\sqrt{u_l}$ terms account for the varying group sizes. This procedure encourages sparsity at both the group and individual levels. That is, for some values of λ, an entire group of predictors may drop out of the model (Hastie, Tibshirani & Friedman (2009)).

Figure 1 presents the steps followed during the course of this study.

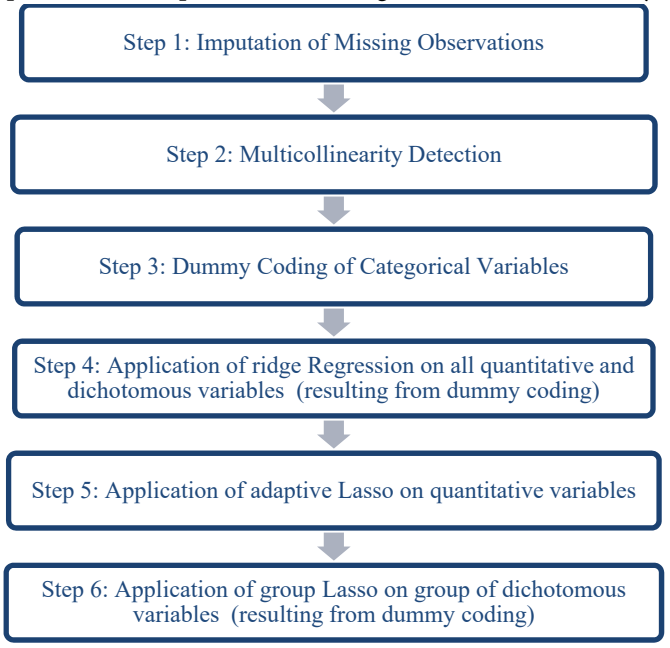

Figure 1. Flowchart of steps followed during the course of the study

4. Results

This section displays the results obtained on stepwise application of methods (discussed in previous section) to the dataset considered.

4.1. Imputation of missing observations

The missingness in the data can be visualized graphically in Figure 2.

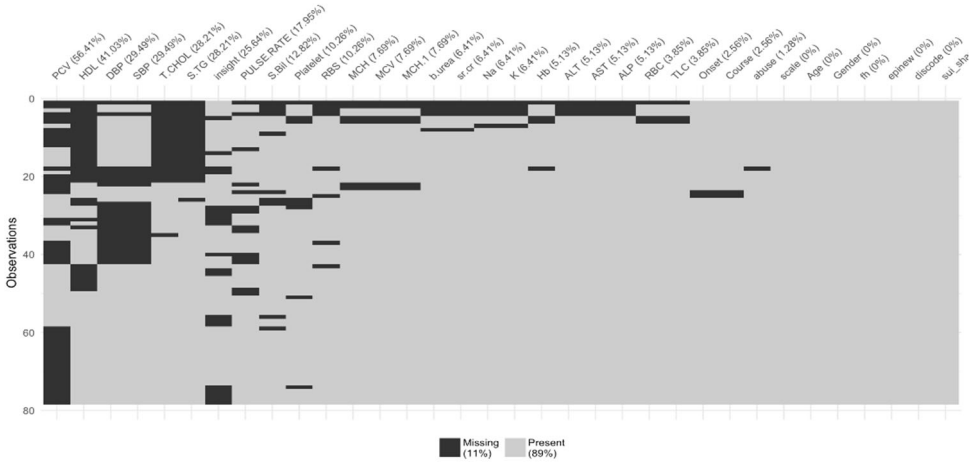

Figure 2. Visual representation of missingness in data.

In Figure 2, black colour shows the location of missing values with respect to each variable. The information on the percentage of missing values overall (in the legend), and in each variable is also provided. Missing observations are imputed using the non-parametric random forest algorithm described in section 3.1. Table 1 presents the Out-of-bag (OOB) error associated with imputation of missing observations.

Table 1. Estimated Out-of-bag (OOB) imputation error

Error type	Result
NRMSE	0.2921
PFC	0.3278

The NRMSE and PFC are not far from zero, indicating not much error is committed in imputing data. The descriptive statistics of quantitative variables and the summary of frequencies for categorical variables after imputation are presented in Tables 2 and 3.

Table 2. Descriptive statistics of quantitative variables after imputation

S. no.	Variable	Min	Max	Median	Mean	Stdev
1	DBP (mmHg)	81.60	141.00	118.21	117.78	9.63
2	SBP (mmHg)	13.60	96.50	76.16	76.08	9.16
3	Pulse rate (per min)	70.00	89.60	81.94	81.42	2.78
4	Hb (g/dL)	7.45	17.00	13.20	13.14	2.07
5	RBC (million/μL)	2.69	6.21	4.61	4.60	0.65
6	MCH (pg)	20.80	101.80	32.60	49.55	27.53
7	MCV (fL)	8.78	120.00	75.60	66.32	27.49
8	MCHC (g/dL)	12.60	40.60	32.47	32.13	3.27
9	TLC (cells/L)	2305.25	12600.00	6850.00	6981.77	2121.81
10	Platelet (thousand/μL)	1.11	11.90	1.99	2.30	1.48

Table 2. Descriptive statistics of quantitative variables after imputation (cont.)

S. no.	Variable	Min	Max	Median	Mean	Stdev
11	b.urea (mg/dL)	1.80	46.00	21.71	22.23	7.25
12	sr.cr (mg/dL)	0.60	1.70	1.00	1.01	0.18
13	Na (mEq/L)	131.50	154.00	140.63	141.03	4.11
14	K (mEq/L)	3.36	6.40	4.28	4.27	0.47
15	S.Bil (mg/dL)	0.30	2.60	0.70	0.78	0.38
16	ALT (IU/L)	12.00	170.75	27.00	35.37	26.89
17	AST (IU/L)	16.00	175.00	34.00	42.22	28.26
18	ALP (IU/L)	1.00	358.00	154.90	161.35	69.13
19	RBS (mg/dL)	59.00	273.50	106.88	114.28	38.50
20	TCL (mg/dL)	102.00	221.00	158.87	161.04	26.66
21	HDL (mg/dL)	27.00	282.00	48.44	58.82	37.87
22	S.TG (mg/dL)	32.00	426.50	123.35	126.87	59.94
23	PCV (%)	2.22	77.00	40.13	39.52	9.00
24	Age (years)*	20.00	70.00	40.50	41.94	10.48
25	Number of episodes *	1.00	5.00	2.00	2.08	0.98
26	RSS*	4.48	68.75	37.50	36.53	13.99

Note: *There were no missing values for these quantitative variables: age, RSS and episodes.

Table 3. Summary of frequencies of categorical variables after imputation

Variable	Category	Frequency	% Total
Gender*	Female	42	53.85
	Male	36	46.15
Family History (fh)	Absent	22	28.21
	Present	56	71.79
Onset	Abrupt	17	21.79
	Acute	25	32.05
	Chronic	1	1.28
	Insidious	32	41.03
	Sub-Acute	3	3.85
Course	Continuous and Progressive	30	38.46
	Continuous	12	15.38
	Episodic	22	28.21
	Fluctuating	14	17.95
Abuse	Absent	46	58.97
	Present	32	41.03
Type of Disorder (discode)*	BPAD	17	21.79
	Depression	5	6.41
	Others	17	21.79
	Schizophrenia	39	50
Suicidal ideation or self-harm (sui_sharm)*	Absent	62	79.49
	Present	16	20.51
Insight	Grade 1	29	37.18
	Grade 2	12	15.38
	Grade 3	20	25.64
	Grade 4	14	17.95
	Grade 5	3	3.85

Note: *There were no missing values for these categorical variables: gender, discode and sui_sharm.

4.2. Multicollinearity detection

The inclusion of a large number of variables, which are also observed to be interdependent and correlated, lead to the problem of multicollinearity. Thus, a check for detection of multicollinearity among regressors is performed using Variance Inflation Factor (VIF). Table 4 presents VIF for each regressor.

Table 4. Variance Inflation Factor (VIF) for regressors

S. no.	Variables	VIF	S no.	Variables	VIF
1	DBP (mmHg)	7.18	18	ALP (IU/L)	1.77
2	SBP (mmHg)	5.19	19	RBS (mg/dL)	2.89
3	Pulse rate (per min)	1.54	20	TCL (mg/dL)	2.67
4	Hb (g/dL)	7.23	21	HDL (mg/dL)	2.10
5	RBC (million/μL)	5.30	22	S.TG (mg/dL)	3.38
6	MCH (pg)	9.77	23	PCV (%)	2.53
7	MCV (fL)	10.77	24	Age (years)	2.51
8	MCHC (g/dL)	2.66	25	Number of Episodes	2.97
9	TLC (cells/L)	1.67	26	Gender	5.51
10	Platelet (thousand/μL)	1.57	27	Family History (fh)	2.03
11	b.urea (mg/dL)	1.82	28	Onset	1.93
12	sr.cr (mg/dL)	1.73	29	Course	2.52
13	Na (mEq/L)	1.67	30	Abuse	4.73
14	K (mEq/L)	1.51	31	Type of disorder (discode)	1.95
15	S.Bil (mg/dL)	2.21	32	Suicidal ideation or self-harm (sui_sharm)	1.35
16	ALT (IU/L)	3.26	33	Insight	1.88
17	AST (IU/L)	3.42			

A VIF of 5 or more indicates serious or excessive multicollinearity. Thus, the problem of multicollinearity exists in the data due to high values of VIF for regressors DBP, SBP, Hb, RBC, MCH, MCV and gender.

4.3. Dummy coding

The dummy coding is applied to the 8 categorical variables consisting of clinical information on psychiatric factors related to mental disorders. These 8 categorical variables are transformed into 26 dichotomous variables with each variable representing each category. For example, if X_{onset} represents the variable onset with 5 categories, then it is transformed into 5 binary/dichotomous variables X_{onsetj} $(j=1,2,...,5)$ such that $X_{onsetj}=I\left(X_{onset}=j\right)$

4.4. Ridge regression

The ridge regression is applied to a total of 51 regressors including 25 quantitative and 26 binary variables with response variable being psychiatric rating scale score (RSS). The quantitative variables include 23 variables representing clinical information related to vital signs and laboratory test result reports (defined in Section 2), age and number of episodes. The binary variables represent categories of psychiatric variables obtained as a result of dummy coding. The model space is searched using 13-fold cross-validation to obtain the optimum value of the tuning/regularization parameter $\lambda = 21.2001$. Figure 3 presents the mean absolute cross validation error curve plotted as function of $\log(\lambda)$ along with the upper and lower standard deviation curves. It is evident from the figure that the mean absolute cross-validation error is minimum when $\log(\lambda)$ is approximately 3.

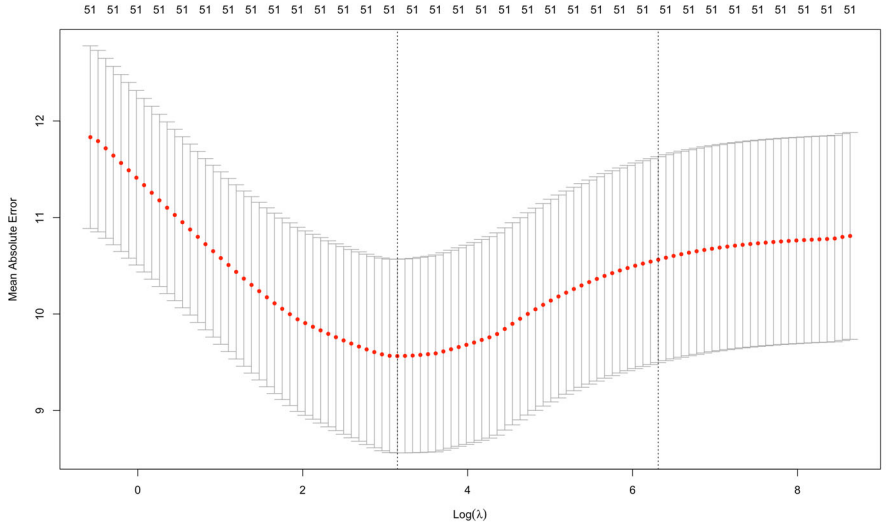

Figure 3. Mean absolute cross validation error curve plotted as function of $\log(\lambda)$ for ridge regression

The coefficients derived on applying the ridge regression to the variables under consideration are presented in Table 5.

Table 5. Regression coefficients estimated from ridge regression

Regressor	Coefficient	Regressor	Coefficient
Intercept	37.8463	Gender: Female	0.3749
DBP (mmHg)	-0.0434	Gender: Male	-0.3742
SBP (mmHg)	-0.0511	Family History: Absent	-2.3078
Pulse rate (per min)	0.1400	Family History: Present	2.3071
Hb (g/dL)	-0.0600	Onset: Abrupt	-0.4918

Table 5. Regression coefficients estimated from ridge regression (cont.)

Regressor	Coefficient	Regressor	Coefficient
RBC (million/μL)	-1.3414	Onset: Acute	-0.1071
MCH (pg)	-0.0095	Onset: Chronic	2.0362
MCV (fL)	0.0125	Onset: Insidious	0.2534
MCHC (g/dL)	0.1158	Onset: Sub Acute	0.5442
TLC (cells/L)	-0.0003	Course: Continuous and Progressive	-2.5869
Platelet (thousand/μL)	-0.0334	Course: Continuous	-0.5289
b.urea (mg/dL)	-0.0209	Course: Episodic	1.1966
sr.cr (mg/dL)	6.6989	Course: Fluctuating	2.9792
Na (mEq/L)	-0.0097	Abuse: Absent	0.6838
K (mEq/L)	0.9443	Abuse: Present	-0.6838
S.Bil (mg/dL)	-1.9304	Type of Disorder: BPAD	-0.6246
ALT (IU/L)	0.0081	Type of Disorder: Depression	2.6030
AST (IU/L)	-0.0120	Type of Disorder: Others	-2.8586
ALP (IU/L)	0.0063	Type of Disorder: Schizophrenia	1.7502
RBS (mg/dL)	-0.0015	Suicidal ideation or self-harm: Absent	-0.3341
TCL (mg/dL)	-0.0239	Suicidal ideation or self-harm: Present	0.3341
HDL (mg/dL)	0.0007	Insight: Grade 1	1.6457
S.TG (mg/dL)	-0.0053	Insight: Grade 2	0.9930
PCV (%)	-0.0786	Insight: Grade 3	-1.4074
Age (years)	-0.0728	Insight: Grade 4	-1.3659
Number of Episodes	1.2645	Insight: Grade 5	-1.1939

It is evident from Table 5 that the coefficients estimated by the ridge regression for 18 regressors out of 51 have values close to 0 indicating that they do not have much effect on the severity of illness.

4.5. LASSO regression

In this study, the LASSO model is applied to the quantitative and categorical predictors separately. The group LASSO is applied to the categorical variables while the adaptive LASSO is used for quantitative regressors.

4.5.1. Adaptive LASSO

The adaptive LASSO is applied to the 25 quantitative variables including 23 variables consisting of clinical information related to vital signs and laboratory test result reports (defined in Section 2), age and number of episodes. The optimum value of the regularization parameter $\lambda = 2.4328$ is obtained using 13-fold cross-validation. Figure 6 presents the mean absolute cross validation error curve plotted as function of $\log(\lambda)$ along with the upper and lower standard deviation curves. It is evident from the figure that the mean absolute cross-validation error is minimum when $\log(\lambda)$ is approximately 0.9.

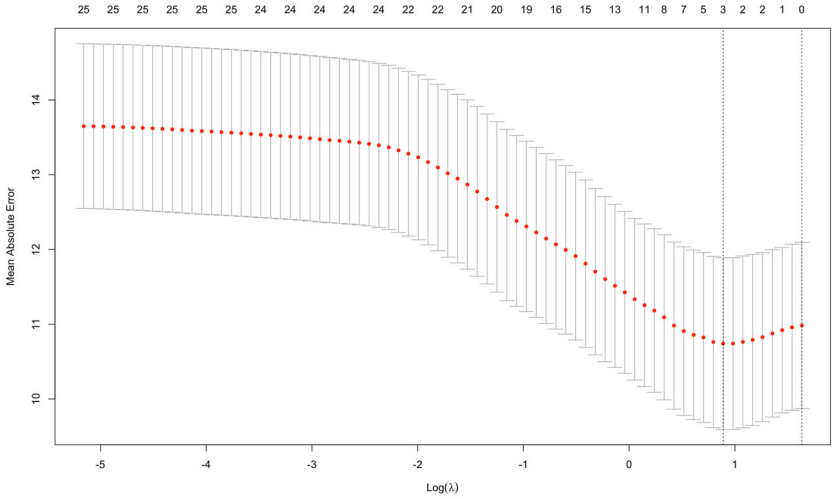

Figure 4. Mean absolute cross validation error curve plotted as function of $\log(\lambda)$ for adaptive LASSO model

The predictors selected from the adaptive LASSO along with their coefficients are presented in Table 6.

Table 6. Regression coefficients estimated from adaptive LASSO regression

S. no.	Regressor (Unit)	Coefficient	S. no.	Regressor (Unit)	Coefficient
1	Intercept	40.8275	14	Na (mEq/L)	0.0000
2	DBP (mmHg)	0.0000	15	K (mEq/L)	0.0000
3	SBP (mmHg)	0.0000	16	S.Bil (mg/dL)	0.0000
4	Pulse rate (per min)	0.0000	17	ALT (IU/L)	0.0000
5	Hb (g/dL)	0.0000	18	AST (IU/L)	0.0000
6	**RBC (million/µL)**	**-2.1370**	19	ALP (IU/L)	0.0000
7	MCH (pg)	0.0000	20	RBS (mg/dL)	0.0000
8	MCV (fL)	0.0000	21	TCL (mg/dL)	0.0000
9	MCHC (g/dL)	0.0000	22	HDL (mg/dL)	0.0000
10	TLC (cells/L)	0.0000	23	S.TG (mg/dL)	0.0000
11	Platelet (thousand/µL)	0.0000	24	PCV (%)	0.0000
12	b.urea (mg/dL)	0.0000	25	Age (years)	0.0000
13	**sr.cr (mg/dL)**	**3.1596**	26	**Number of Episodes**	**2.7509**

The adaptive LASSO selected only 3 relevant predictors out of a total of 25 variables by shrinking the coefficients of less other regressors to zero. All of these predictors have coefficients far from 0. Thus, laboratory test results on Red Blood Cell (RBC), serum creatinine (sr.cr), and number of episodes are found to be the relevant predictors of severity of mental illness as measured by the psychiatric rating scales.

4.5.2. Group LASSO

The group LASSO is applied to the 26 dichotomous variables obtained from 8 psychiatric variables after applying dummy coding with each binary variable representing each category. The optimum value of $\lambda = 0.4829$ is obtained using 13-fold cross-validation. Figure 5 presents the mean absolute cross validation error curve plotted as function of $\log(\lambda)$ along with the upper and lower standard deviation curves. It is evident from the figure that the mean absolute cross-validation error is minimum when $\log(\lambda)$ is approximately -0.7.

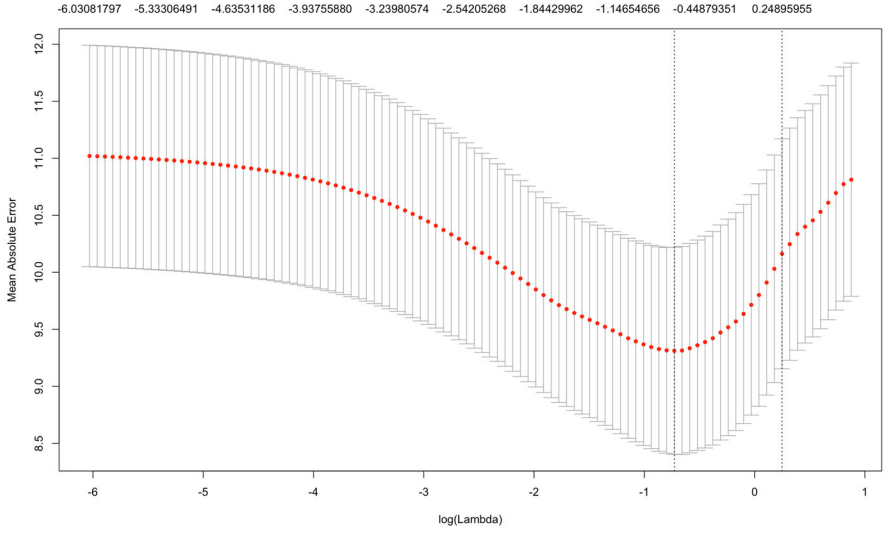

Figure 5. Mean absolute cross validation error curve plotted as function of $\log(\lambda)$ for group LASSO model.

The predictors selected by the group LASSO along with their coefficients are presented in Table 7.

Table 7. Regression coefficients estimated from group LASSO regression

S. no.	Regressor	Category	Coefficient
1	Intercept		34.6398
2	Gender	Female	0.0585
3		Male	-0.0585
4	Family History (fh)	**Absent**	**-3.6916**
5		**Present**	**3.6918**
6	Onset	Abrupt	0.0000
7		Acute	0.0000
8		Chronic	0.0000
9		Insidious	0.0000
10		Sub-Acute	0.0000

Table 7. Regression coefficients estimated from group LASSO regression (cont.)

S. no.	Regressor	Category	Coefficient
11	Course	**Continuous and Progressive**	**-3.7530**
12		**Continuous**	**-1.2243**
13		**Episodic**	**2.0239**
14		**Fluctuating**	**2.9536**
15	Abuse	Absent	0.0241
16		Present	-0.0241
17	Type of Disorder (discode)	Bipolar affective disorder (BPAD)	-0.0389
18		**Depression**	**1.1385**
19		**Others**	**-3.4340**
20		**Schizophrenia**	**2.3347**
21	Suicidal ideation or self-harm (sui_sharm)	Absent	0.0000
22		Present	0.0000
23	Insight	**Grade 1**	**2.1232**
24		Grade 2	0.4696
25		**Grade 3**	**-1.3771**
26		**Grade 4**	**-1.0185**
27		Grade 5	-0.1970

The group LASSO selected 6 groups of predictors out of a total of 8 groups of psychiatric variables. Thus, gender, family history, course, alcohol and/or tobacco abuse, type of disorder and insight of illness are found to be relevant predictors of severity of mental illness.

All the calculations were performed in R software using adalasso, coefplot, gglasso, glmnet, mctest, missForest, pastecs and summarytools packages.

5. Discussion

Recently, a large number of research studies have focused on establishing diagnostic tests for mental disorders based on reports of blood tests and psychiatric factors. Richards et al. (2016) predicted severity of depression based on gender, age, employment status, marital status, previous diagnosis of depression, recent experience of life stressors using multiple linear regression. Huang et al. (2014) predicted the diagnosis and severity of depression based on a large sample of electronic health record (EHR) data consisting of information on demographic variables, structured variables such as ICD diagnosis codes, prescription codes, and unstructured variables such as progress notes, pathology reports, radiology reports, and transcription reports. This motivated us to predict the severity of illness based on the laboratory and pathological reports and certain psychiatric aspects. Further, the information on these basic variables is generally readily available for all mental disorders.

Missingness is a commonly encountered problem in medical data. However, ignoring or removing missing data leads to an important loss of information and results in biased estimation. We have used multiple imputation to deal with missingness since in addition to restoring the natural variability of the missing values, it incorporates the uncertainty due to the missing data, which results in valid statistical inference (Kang (2013)). Multicollinearity is commonly observed in datasets with large number of regressors. Variance Inflation factor (VIF) is the most common approach for detecting multicollinearity. There is no set VIF threshold available in the literature to be used as a standard rule. In this study, we employed a VIF threshold of 5 for collinearity diagnostics since a VIF value that is near or above 5, indicates that the regressors may be highly correlated (Akinwande, Dikko and & Samson (2015); Jongh et al. (2015)).

When there are a large number of predictors, the correlation between them (multicollinearity) generally limits the usefulness of classic regression methods. Regularization techniques such as ridge, LASSO, and elastic net are particularly useful in such cases. In this study, we applied both ridge and extensions of LASSO viz. the adaptive and group LASSO models on the data and observed that adaptive and the group LASSO models did not extract any of the 18 regressors for which the coefficients were estimated to be close to 0 by the ridge regression. Further, we compared the ridge and the LASSO models using the Bayesian Information Criterion (BIC) and observed that the BIC values for the group LASSO (BIC=1057.617) and the adaptive LASSO (BIC=1131.936) were lower than the ridge regression model (BIC=1148.786). Thus, in this study, the group and adaptive LASSO models performed better than the ridge model.

The LASSO (l_1) penalty function performs variable selection and dimension reduction by shrinking coefficients, while the ridge (l_2) penalty function shrinks the coefficients of correlated variables towards their average (Kim et al. (2017)). In general, LASSO is preferred over the ridge model in terms of interpretability since it extracts the relevant predictors. However, in medical data, it is not advisable to completely ignore or remove the less relevant predictors due to their clinical implication. Even if the objective of a study is to extract relevant predictors, it is suggested to perform both LASSO and the ridge regression since the ridge regression supports the results of the LASSO regression and will help to make a decision depending upon the clinical relevance of the regressor based on a chosen level of significance.

In the adaptive LASSO, the weights are based on the ordinary least square estimates. The weights are data-dependent and adaptively chosen from the data with large coefficients receiving small weights and small coefficients receiving large weights.

In this study, it was observed that from a wide range of clinical variables consisting of information on both laboratory test results and psychiatric aspects, the following are the relevant predictors of the severity of mental illness: Red Blood Cells (RBC), Serum Creatinine (Sr.Cr), number of episodes, gender, family history, course of illness, alcohol and tobacco abuse, type of disorder and insight of an illness. Our results are in accordance with previous studies. Setoyama et al. (2016) found that serum creatinine is commonly associated with severity of depression in three independent cohort sets regardless of the presence or absence of medication and diagnostic difference. Barbato (1998), Häfner (2005) and Richards et al. (2016) have identified gender as one of the relevant predictors of severity of mental illness. Lu et al. (2018) found that positive family history is a strong predictor of schizophrenia. Marzo et al. (2006) showed that patients with multi-episode bipolar disorder would be more prone to have higher levels of cognitive impairment suggesting that patients with a higher number of episodes and recurring or episodic course result in severe outcomes. Studies in the past showed that a higher amount of alcohol and tobacco consumption is found to be associated with greater severity of illness (Goldstein, Velyvis, & Parikh (2006); Krishnadas et al. (2012); Dwivedi, Chatterjee, & Singh (2017)). Jacob (2016) showed that patients with good insight have a less severe disease.

This paper adds to the literature of medical research aimed at identifying the biomarkers for diagnosis and predictors of the severity status of mental disorders. The clinicians can use the relevant factors to build a profile of the patient and his needs. This work will help in developing valid and efficient approaches to diagnose the disorders at an early stage. It will also aid clinicians in devising effective strategies for treatment planning.

Generally, the predictive accuracy of the regularization method is tested on a test dataset after fitting the regression model on the training dataset. This procedure could not be adopted in this paper due to the small sample size. To maintain consistent selection of predictors, the tuning parameter for fitting regularization models is selected using 13-fold cross-validation. However, a limitation of using the cross validation method in the case of a small sample size could suffer from overfitting.

References

Akinwande, M. O., Dikko, H. G., and Samson, A., (2015). Variance inflation factor: as a condition for the inclusion of suppressor variable(s) in regression analysis. *Open Journal of Statistics*, pp. 754–767.

American Psychiatric Association, (2013). *Diagnostic and statistical manual of mental disorders*, 5th edition Arlington, VA: American Psychiatric Publishing.

Barbato, A., (1998). *Schizophrenia and Public Health. Nations For Mental Health*, Division of Mental Health and Prevention of Substance Abuse, Geneva: World Health Organization.

Bahn, S., Schwarz, E., Harris, L. W., Martins-De-Souza, D., Rahmoune, H., and Guest, P. C., (2013). Biomarker blood tests for diagnosis and management of mental disorders: focus on schizophrenia. *Archives of Clinical Psychiatry*, São Paulo, 40(1), pp. 02–09.

Brådvik, L., (2018). Suicide Risk and Mental Disorders. *International journal of environmental research and public health*, 15 (9), pp. 2028–2031.

Belsley, D., (1991). *Conditioning diagnostics: collinearity and weak data in regression*, New York: Wiley.

Brewer, B. R., Pradhan, S., Carvell, G., and Delitto, A., (2009). Application of modified regression techniques to a quantitative assessment for the motor signs of Parkinson's Disease." *IEEE transactions on neural systems and rehabilitation engineering: a publication of the IEEE Engineering in Medicine and Biology Society*, 17 (6), pp. 568–575.

Canan, F., Dikici, S., Kutlucan, A., and Celbek, G., Coskun, H., Gungor, A., Aydin, Y. and Kocaman, G., (2012). Association of mean trombosit volume with DSM-IV major depression in a large community-based population: the MELEN study. *Journal of psychiatric research*, 46 (3), pp. 298–302. 10.1016/j.jpsychires.2011.11.016.

Dwivedi, A. K., Chatterjee, K., and Singh, R., (2017). Lifetime alcohol consumption and severity in alcohol dependence syndrome. *Industrial Psychiatry Journal*, 26(1), pp. 34–38.

Farrar, and Glauber, R., (1967). Multicollinearity in regression analysis: The problem revisited. *The Review of Economics and Statistics*, 49 (1), 92–107.

Goldstein, B., Velyvis, V., and Parikh, S. V., (2006). The association between moderate alcohol use and illness severity in bipolar disorder: a preliminary report. *The Journal of Clinical Psychiatry*, 67 (1), pp. 102–106.

Greene, W. H., (1993). T*he econometric approach to efficiency analysis. In the measurement of productive efficiency and productivity change*, by Harold O. Fried, C. A. Knox Lovell, and Shelton S. Schmidt, pp. 68–119. United Kingdom.

Haenisch, F., Cooper, J. D., Reif, A., Kittel-Schneide, S., Steiner, J., Leweke, F. M., Rothermundt, M., Beveren, N., Crespo-Facorro, B., Niebuhr, D., Cowan, D., Weber, N., Yolken, R., Penninx, B. and Bahn, S., (2016). Towards a blood-based

diagnostic panel for bipolar disorder. *Brain, Behavior, and Immunity*, 52, pp. 49–57. https://doi.org/10.1016/j.bbi.2015.10.001

Hafner, H., (2005). *Gender Differences in Schizophrenia. In Estrogen Effects in Psychiatric Disorders*, by N. Bergemann, and A. (eds.) Riecher-Rössler. Austria: SpringerWienNewYork.

Hastie, T., Tibshirani, R., and Friedman, J., (2009). The elements of statistical learning: data mining, inference and prediction. *Second Edition*. California: Springer.

Hastie, T., Tibshirani, R., and Wainwright, M., (2015). *Statistical learning with sparsity: The Lasso and generalizations*. New York: Chapman and Hall/CRC Press.

Hoerl, A. E., Kennard, R. W., (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12 (1), pp. 55–67. DOI: 10.1080/00401706.1970.10488634.

Huang, S. H., Lependu, P., Iyer, S. V., Ai-Seale, M., Carrell, T. D., and Shah, N. H., (2014). Toward personalizing treatment for depression: predicting diagnosis and severity. *Journal of the American Medical Informatics Association*, 21 (6), pp. 1069–1075.

Jacob, K. S., (2016). Insight in psychosis: An indicator of severity of psychosis, an explanatory model of illness, and a coping strategy. *Indian journal of psychological medicine*, 38(3), pp. 194–201.

Jain, R., (1985). Ridge regression and its application to medical data. *Computers and Biomedical Research*, 18, pp. 363–368.

James, G., Witten, D., Hastie, T., and Tibshirani, R., (2013). *An introduction to statistical learning: with applications in R*, New York: Springer.

Jongh, P. J. De, Jongh, E. De, Pienaar, M., Gordon-Grant, H., Oberholzer, M., and Santana, L., (2015). The impact of pre-selected variance inflation factor thresholds on the stability and predictive power of logistic regression models in credit scoring. *Orion*, 31(1), pp. 17–37, DOI: https://doi.org/10.5784/31-1-162.

Kang, H., (2013). The prevention and handling of the missing data. *Korean journal of anesthesiology*, 64 (5), pp. 402–406.

Kim, M. H., Banerjee, S., Park, S. M., and Pathak, J., (2017). *Improving risk prediction for depression via Elastic Net regression – Results from Korea National Health Insurance Services Data*. AMIA ... Annual Symposium proceedings. AMIA Symposium, 2016, pp. 1860–1869.

Krishnadas, R., Jauhar, S., Telfer, S., Shivashankar, S., and Mccreadie, R., (2012). Nicotine dependence and illness severity in schizophrenia. *The British journal of psychiatry*, 201 (4), pp. 306–12.

Laursen, T. M., Labouriau, R., Licht, R. W., Bertelsen, A., Munk-Olsen, T., and Mortensen, P. B., (2005). Family history of psychiatric illness as a risk factor for schizoaffective disorder: A Danish Register-Based Cohort Study. *Arch Gen Psychiatry*, 62 (8), pp. 841–848. doi:10.1001/archpsyc.62.8.841

Lu, Y., Pouget, J. G., Andreassen, O. A., Djurovic, S., Esko, T., Hultman, C. M., Metspalu, A., Milani, L., Werge, T., and Sullivan, P. F., (2018). Genetic risk scores and family history as predictors of schizophrenia in Nordic registers. *Psychological medicine*, 48(7), pp. 1201–1208.

Marzo, S. D., Giordano, A., Pacchiarotti, I., Colom, F., Sánchez-Moreno, J., and Vieta, E., (2006). The impact of the number of episodes on the outcome of bipolar disorder. *The European Journal of Psychiatry*, 20, pp. 21–28.

Mcdaniel, K., Edland, S., and Heyman, A., (1995). Relationship between level of insight and severity of dementia in Alzheimer disease. CERAD Clinical Investigators. Consortium to Establish a Registry for Alzheimer's Disease. *Alzheimer Dis Assoc Disord*, 9 (2), pp. 101–104.

Milne, B., Caspi, A., Harrington, H., Poulton, R., Rutter, M., and Moffitt, T., (2009). Predictive value of family history on severity of illness: The case for depression, anxiety, alcohol dependence, and drug dependence. *Arch Gen Psychiatry*, 66 (7), pp. 738–747.

Oba, S., Sato, M. A., Takemasa, I., Monden, M., Matsubara, K., and Ishii, S., (2003). A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19, pp. 2088–2096.

Richards, D., Richardson, T., Timulak, L., Viganò, N., Mooney, J., Doherty, G., Hayes, C., Sharry, J., (2016). Predictors of depression severity in a treatment-seeking sample. *International Journal of Clinical and Health Psychology*, 16 (3), pp. 221–314.

Sadock, B., (2009). *Psychiatric report, medical record and medical error*. In S. V. Sadock BJ, Kaplan and Sadock's Comprehensive Textbook of Psychiatry (9th ed., pp. 907–18). Philadelphia: Lippincott Williams and Wilkins.

Setoyama, D., Kato, T. A., Hashimoto, R., Kunugi, H., Hattori, K., Hayakawa, K., Sato-Kasai, M., Shimokawa, N., Kaneko, S., Yoshida, S., Goto, Y. I., Yasuda, Y., Yamamori, H., Ohgidani, M., Sagata, N., Miura, D., Kang, D., and

Kanba, S., (2016). Plasma metabolites predict severity of depression and suicidal ideation in psychiatric patients-A Multicenter Pilot Analysis. *PLoS One*, 11(12). e0165267

Stegenga, B. T., Kamphuis, M. H., King, M., Nazareth, I., and Geerlings, M. I., (2010). The natural course and outcome of major depressive disorder in primary care: the PREDICT-NL study. *Social psychiatry and psychiatric epidemiology*, 47 (1), pp. 87–95.

Stekhoven, D. J., Bühlmann, P., (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28 (1), pp. 112–118.

Tibshirani, R., (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*. Series B (Methodological), pp. 267–288.

Upadhya, S. S., Cheeran, A. N., (2018). Performance comparison of regression techniques in predicting Parkinson disease severity score using speech features. *Biomedical Engineering: Applications, Basis and Communications*, 30(4). https://doi.org/10.4015/S1016237218500254

World Health Organization, (1992). T*he Icd-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines*. Geneva: World Health Organization.

WORLD HEALTH ORGANIZATION, (2000). Cross-national comparisons of the prevalences and correlates of mental disorders. WHO International Consortium in Psychiatric Epidemiology, *Bull*, 78 (4), pp. 413–426.

World Health Organization, (2003). Investing in mental health. *World Health Organization*, pp. 1–48.

Yuan, M., Lin, Yi., (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*. Series B 68, part 1, pp. 49–67.

Zhao, P., Yu, B., (2006). On Model Selection Consistency of Lasso. *Journal of Machine Learning Research*, 7, pp. 2541–2563.

Zou, H., and Hastie, T., (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, Series B, 67(2), pp. 301–320.

Zou, H., (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association: Theory and Methods*, 101(476), pp. 1418–1429.

Zimmerman, M., Morgan, T. A., and Stanton, K., (2018). The severity of psychiatric disorders. *World psychiatry: official journal of the World Psychiatric Association* (WPA), 17 (3), pp. 258–275.