

Research Article

Assessment of the Risk Factors of MDD Recurrence Based on Deep Learning Approaches

Wu Q^{1,2}, Zhao W², Yang X³, Tan H², You L², Xu H², Zhou Y² and Zhou X^{2,4,5*}

¹School of Public Health, Xi'an Jiaotong University Health Science Center, China

²Center for Computational Systems Medicine, School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, USA

³The First Affiliated Hospital of Xi'an Jiaotong University Health Science Center, Xi'an, China

⁴Department of Integrative Biology and Pharmacology, McGovern Medical School at The University of Texas Health Science Center at Houston, Houston, TX, USA

⁵School of Dentistry, University of Texas Health Science Center at Houston, Houston, TX, USA

*Corresponding authors: Xiaobo Zhou, Center for Computational Systems Medicine, School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX 77030, USA

Received: June 09, 2021; Accepted: July 09, 2021;

Published: July 16, 2021

Abstract

Objective: Explore the risk factors related to the recurrence of MDD and provide a basis for the prevention and control of MDD.

Methods: Patients with MDD were extracted from two large, multi-center clinical datasets. The inpatients and outpatients between January 2000 and December 2015 were collected. Eligible patients were 18-90 years-old and had a diagnosis of MDD. The MDD were identified based on the MDD-related ICD-9-CM diagnosis codes; and MDD-related ICD-10-CM diagnosis codes. Eventually, 140,497 patients were qualified for further analysis, including 69.2% female patients. Among of 140,497, 20, 078 patients (14.3%) had no comorbidities. Logistic regression, SVM, and LSTM were employed to predict the key risk factors associated with MDD recurrence.

Results: The MDD patients with married /life partners had a lower prevalence rate (9.2%) of MDD recurrence than the patients with single marital status (11.8%). The primary MDD patients had a higher MDD recurrent rate (11.7%) than secondary MDD patients (10.5%). Primary MDD was associated with MDD recurrence (OR 2.49, 95% CI 1.53-3.96) via logistic regression analysis. Insomnia, anxiety and single marital status were also top-ranked risk factors for the MDD recurrence. The prediction accuracy of logistic regression, SVM and LSTM were 0.736, 0.791 and 0.834, respectively.

Conclusions: Building statistical models by mining existing EHR data can explore the risk factors associated with MDD recurrence. Our results indicated that primary MDD, never married, anxiety symptoms, and insomnia were risk factors for MDD recurrence. The prediction accuracy of the LSTM model was higher than the other two approaches.

Keywords: MDD; Prognosis; EHR; Data mining; LSTM

Abbreviations

CI: Confidence Interval; EHR: Electronic Health Records; ICD-9-CM: International Classification Of Diseases, 9th Revision-Clinical Modification; ICD-10-CM: International Classification of Diseases, Tenth Revision, Clinical Modification; LSTM: Long Short-Term Memory; MDD: Major Depressive Disorder; OR: Odds Ratio; RNNs: Recurrent Neural Networks; SVM: Support Vector Machine

Introduction

Major Depressive Disorder (MDD) is one of the most common medical illnesses worldwide with a lifelong prevalence up to 16% [1] and a leading cause of disability worldwide. MDD is characterized by a long-lasting depressed mood or marked loss of interest/pleasure in all or nearly all activities [2]. There has been a gradual increase in the prevalence of MDD [3]. MDD is highly associated with poor mental health and socio-economic status [4]. MDD can impact mood and behavior as well as various physical functions, thereby reducing patients' quality of life [5]. MDD recurrent nature is also one of the most crippling and devastating aspects of depression [6,7]. Every recurrence also carries a 10-20% risk of becoming unremitting and chronic, along with a heightened risk for suicide, both of which can lead to serious comorbidities and lethal consequences associated

with depression. One of the most important challenges in the management of MDD is to prevent depression relapse. Individuals with a first depressive episode have a 40% to 60% chance experiencing a subsequent episode. Individuals suffering from 2 episodes have an approximate 60% recurrence probability [7]. Therefore, accurate prediction prognosis of MDD is important to prevent MDD recurrence that leads to disability.

However, the prediction of prognosis of MDD is limited by small sample size, budget, and physicians' experience and so on. On the other hand, the electronic health records (EHR) have been collected in the clinic, but appear to have not been fully utilized in [8] various clinical studies. EHR for patients with longitudinal health information is a valuable source for exploring to diagnose diseases and assist clinical decision-making. However, it is quite challenging to mine EHRs efficiently. First, EHR data is heterogeneous and contains various types of features. Then, the data is essentially sparse and biased, due to the patient's irregular visits, lack of certain tests and missing values. Recent studies have taken advantage of EHRs for predictive modeling tasks in early prediction of chronic disease [9] and monitoring disease progression [10]. How to deal with the heterogeneity and sparsity of EHR data and reasonably explain the predicted results are key problems to be solved by modeling. Regression models and Support

Vector Machine (SVM) have been applied to predict the progression of the patient's health status previously. However, these models do not provide a comprehensive analysis of the long-term of diagnostic information, which may lead to miss the severe symptoms of the past. The machine learning models such as SVM and random Forest (FR) are used to deal with the complex interactions between predictive factors, but lack interpretability for understanding disease etiology. In addition, since disease progression is a complex and dynamic process, understanding the etiology of a disease requires repeated clinical measurements over time rather than relying only upon a baseline profile. Therefore, the time series models, such as recurrent neural networks, appear to be more suitable for analyzing and understanding such data. Recent much works [10,11] suggest that deep learning can significantly improve prediction performance. To deal with the temporality of multivariate sequences, dynamically modeling the sequential data is necessary. The Recurrent Neural Network (RNN) is a class of artificial neural network where connections between nodes form a directed data along a sequence. This allows exhibiting temporal dynamic behavior for a time series. Therefore, RNN is often used for times series prediction, such as Long Short-Term Memory (LSTM). Taking advantage of the capability of RNN in memorizing historical records, several RNN-based models have been used to derive accurate and robust representations of patient visits [12].

In this study, we determined the prognostic risk factors for the patients with MDD, and predicted the recurrent MDD. In addition, the accuracy of predict model was evaluated.

Methods

Study design

We adopted a retrospective study to analyze the risk factor of MDD prognosis. The cases were patients with recurrent MDD and the patients with a single episode of MDD were used as controls (Figure 1).

Clinical data description

Two EHR datasets were used in this study, including the

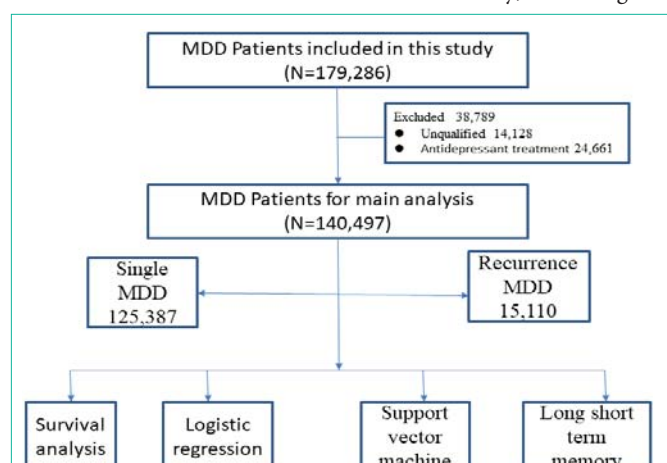


Figure 1: Flowchart of selection of patients and reasons for attrition between baseline and cohort. 14,128 patients who had sometimes-missing value were excluded. 24,661 patients who get antidepressant treatment. The traditional methods and deep learning were orderly applied to analysis 140497 patients.

clinical datasets from the University Texas Physicians Clinical Data Warehouse (UTPCDW) and Cerner Health Facts. There were outpatients and inpatients' EHR data in both datasets. The UTPCDW database is derived from 1.8 million patients and has a total of 3.2 million records. The database of Cerner Health Facts is comprised of de-identified EHR data from over 600 participating Cerner client hospitals and clinics in the United States and contains clinical information for over 106 million unique patients with more than 15 years records from 2000-2016 [13]. The types of data available include demographics, diagnoses, procedures, lab results, medication orders, medication administration, vital signs, microbiology, other clinical observations, and health systems attributes. We extracted the data for MDD patients between January 2000 and December 2015 directly from the EHR of hospitals in the Cerner Health Facts and UTPCDW databases. Our Institutional Review Board (IRB) approved this study.

The inclusion criteria of the participants

The participants included inpatients and outpatients. Eligible participants were diagnosed with depressive disorder between the ages of 18 and 90. We identified the patients with MDD based on the codes 296.2x and 296.3x, and the codes F32.x and F33.x from the International Classification of Diseases, 9th and 10th revision, respectively. A total of 35 diagnosis codes for MDD are included. We extracted 179,286 patients from two databases.

The exclusion criteria of participants

The participants had only one visit in the EHR. The EHRs with 5 or more missing values were excluded. Participants with recurrence of MDD at baseline was excluded.

To reduce false-positive misclassification of MDD, only the individuals who received at least two diagnostic codes for a given condition separated by >30 days were considered to have that condition [15]. Thus, a total of 140,497 patients were enrolled in this study.

Clinical outcomes

The primary outcomes included MDD recurrence or recovery (without MDD recurrence). Informations on various diseases or symptoms is extracted from the database according to the codes in ICD-9 or ICD-10. Single episode of MDD was defined based on the ICD-9 diagnostic codes as described 296.2x, while the code was F32.x in ICD-10. MDD recurrence was defined based on the ICD-9 diagnostic codes as described 296.3x, while the code was F33.x in ICD-10. According to the DSM - 5, the full recovery of MDD episode was no significant signs of symptom of disturbance during the past 2 months.

Primary MDD was refer to depressive mood symptoms are related to internal biological factors. Secondary MDD. Secondary MDD refers to the symptoms and signs directly related to life stressful events, also known as exogenous or environmental.

Model development

The two datasets obtained from Cerner Health Facts and UTPCDW contained 83,615 and 56,882 patients, respectively. The Cerner Health Facts dataset was further divided into training, validation sets with patient numbers of 58,531; 25,084. The dataset from UTPCDW is test dataset. KNN imputation method was used

Table 1: Demographic characteristic of MDD patients.

Characteristic	MDD (n)
Age	52.0 (38.0-64.0)
Sex	
Male	42618
Female	97879
Marital	
Married/Life partner	53371
Single	48793
Legally Separated/Divorced/Widowed	38333
Race	
African American	28611
American Indian/Alaska Native/Latin American/Hispanic	17034
Asian/Pacific Islander	614
White or Caucasian	84751
Other	9487
State	
Midwest	19314
East	11080
Northeast	24517
West	10757
South	74829

Age expressed as median and the inter-quartile range (IQR; 25th-75th percentiles) displayed in brackets.

for dealing with the missing values in the datasets. We applied logistic regression, SVM, and LSTM to predict MDD recurrence within 30 days. Logistic regression and SVM were developed and validated in R Studio (Version 1.1.383). LSTM was implemented using the Python language (version 3.6). LSTM architecture consists of memory blocks. The natural function of memory blocks is to remember inputs for a long time. Each memory block contains one self-connected accumulator cell and several multiplicative units, such as input, forget, and output gates. These three gates allow us to store and access informations by assignment. The parameters of LSTM and the methods used in the modeling step are listed in Appendix Table 1. The predictive accuracies of the models were assessed via accuracy, F-measure and Recall.

Statistical analysis

Firstly, Chi-square and logistic regression were used for exploring the risk factors. Secondly, we use survival analysis to estimate the recovery rate of MDD. Logistic regression, support vector machines, and Long-Short Term Memory (LSTM) were employed to predict MDD recurrence (Figure 1).

The categorical variables such as socio-demographic and other baseline characteristics in two groups were assessed using proportions and compared by Chi-squared test. Logistic multinomial regression model was used to analyze the association between risk factors and categorical outcomes. The value of variables was in Appendix Table 2. The survival analysis was performed using (Kaplan-Meier). Statistical significance was evaluated using two-sided 0.05-level tests.

All analysis was performed using R Studio (Version 1.1.383).

Table 2: Demographics and clinical characteristics of MDD (%).

Characteristic	Total	Single	Recurrence	%	χ^2	P
Sex						
Male	42618	38158	4460	10.5	5.33	0.02
Female	97879	87229	10650	10.9		
Marital Status						
Married/Life partner	53371	48484	4887	9.2	229.42	<0.01
Single	48793	43031	5762	11.8		
Legally Separated/Divorced/Widowed	38333	33872	4461	11.6		
Age Group						
<30	19848	17724	2124	10.7	677.21	<0.01
30-50	42350	36905	5445	12.9		
50-70	56733	50515	6218	11		
>=70	21566	20243	1323	6.1		
Smoking						
No	96306	84419	11887	12.3	804.21	<0.01
Smoking	44191	40968	3223	7.3		
Drinking						
No	101316	88902	12414	12.3	848.92	<0.01
Drinking	39181	36485	2696	6.9		

Results

Demographic and clinical characteristics of the subjects

The demographic characteristics of 140,497 patients are presented in Table 1. Their median age was 52.0 years-old. 69.2% (97,879) patients were female. 53,371 (38.0%) patients were married or had domestic partners. 84,751 (60.3%) patients were Caucasian. About half of the patients were from the southern United States.

Demographic and clinical factors and MDD recurrence

The MDD recurrence rate for patients with married or living partners was lower (4,884, 9.2%) than that of single counterparts (43,031, 11.8%). As shown in Table 2, 30-50 years-old patients had the highest recurrence rate of MDD (5,445, 12.9%) among four age groups.

We compared the MDD recurrence rate between primary MDD and secondary MDD patients. The primary MDD patients had a higher MDD recurrence rate (11.7%) than the secondary MDD patients. The patients with other comorbidities had a lower MDD recurrence rate (12.6%) than the patients without other comorbidities (Table 3). We also compared the patients with different courses of MDD. The recurrence rate of MDD in patients with a course of 1-5 years (16.6%) is higher than that of other patients. Of the 140,497 MDD patients, 120,419 (85.7%) had comorbidities. The MDD patients with some comorbidities (Anxiety, Insomnia, and Obesity) had a higher MDD recurrence rate than those without these comorbidities (Table S3). The prevalence rate of hypertension, diabetes, and hypothyroidism in MDD patients were higher than in the general population. However, Chi-square test shows that diabetes, hypothyroidism, and hypertension might be not risk factors for MDD recurrence (Table S4).

Table 3: MDD recurrence and complications/the courses of disease (%).

	Total	Single	Recurrence	%	χ^2	P
MDD						
Primary	31019	27380	3639	11.7	39.45	<0.01
Secondary	109478	98007	11471	10.5		
Number of Comorbidities						
0	20078	17547	2531	12.6	338.6	<0.01
1	41956	36879	5077	12.1		
2	30621	27390	3231	10.6		
3	23274	21045	2229	9.6		
>=4	24568	22526	2042	8.3		
Course of Disease						
<1 yrs	14310	12328	1982	13.9	977.31	<0.01
1-5 yrs	22617	18865	3752	16.6		
5-10 yrs	14572	13037	1535	10.5		
>=10 yrs	13808	12404	1404	10.2		

Identification of the risk factors associated with MDD recurrence using multiple models

We conducted logistic regression analysis to identify the risk factors of MDD recurrence. As shown in Table 4, the primary MDD was highly associated with the MDD recurrence (OR 2.49, 95% CI

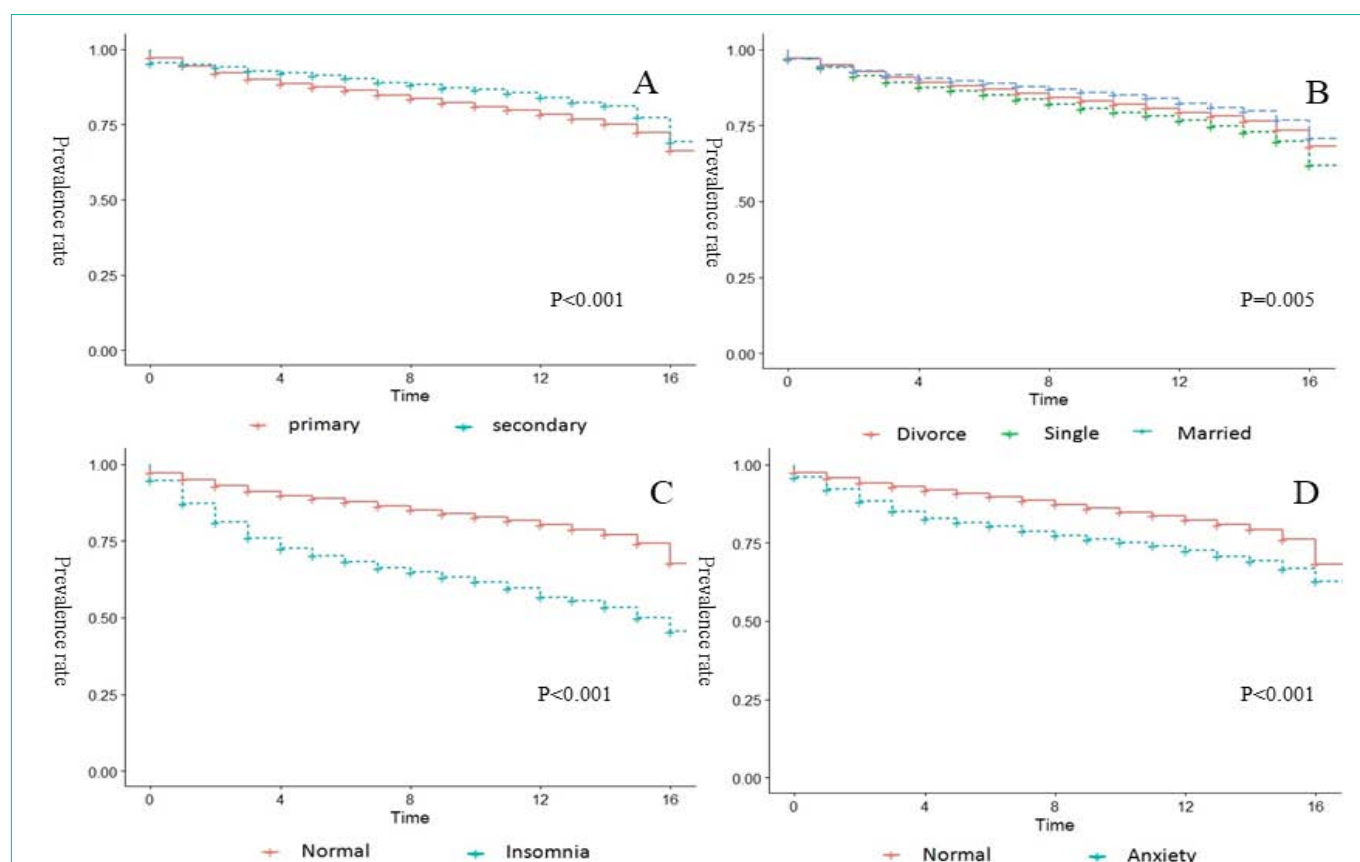
1.53-3.96). Insomnia, anxiety and single status were also top-ranked risk factors with OR values 1.74 (95% CI: 1.60-1.89), 1.65 (95% CI: 1.58-1.74) and 1.41 (95% CI: 1.09-1.83), respectively. There was no significant association of the courses, comorbidities and smoking status with the increased risk of MDD recurrence.

Prediction of the cumulative recovery rate of MDD patients

We then analyzed the recovery rate of MDD patients from 2001 to 2016. As shown in Figure 2, prevalence rate for the primary and secondary MDD patients was 95.0% and 94.5% in the first-year, respectively. In the 15th year, prevalence rate in patients with primary MDD (72.4%) was lower than that in patients with secondary MDD (77.4%) ($P<0.01$). At the year 15, Patients with insomnia had a lower prevalence rate of MDD (49.9%) comparing with the patients without insomnia (74.2%, $P<0.01$). Lower prevalence tendencies were also observed in patients with marriage status and anxiety.

Prediction accuracy of our model

We assessed the accuracy of logistic regression, SVM and LSTM in the prediction of the risk factors associated with MDD recurrence using accuracy, F-measure and Recall. The risk factors included primary MDD, insomnia, anxiety, marry, the course of MDD, smoking, sex, and alcohol. The prediction accuracy of LSTM was 0.834 while the accuracy of SVM and logistic regression models were 0.791 and 0.736, respectively. The 1st, 5th and 30th epoch loss of LSTM was 0.1639, 0.0109 and 0.0024, respectively, indicating our LSTM

**Figure 2:** Prevalence rate of MDD with time passing among different features.

A, B, C are prevalence curves based on different diseases, such as primary MDD vs. secondary MDD, insomnia, anxiety. D is prevalence rate curves among different marital status.

Table 4: Regression model of MDD severity change.

Correlates	β	P	OR	OR 95% CI	
Primary	0.91	<0.01	2.49	1.53	3.96
Insomnia	0.55	<0.01	1.74	1.6	1.89
Anxiety	0.5	<0.01	1.65	1.58	1.74
Single	0.34	<0.01	1.41	1.09	1.83
Course	-0.16	<0.01	0.85	0.83	0.87
Complications	-0.11	<0.01	0.89	0.86	0.93
Smoking	-0.77	<0.01	0.46	0.28	0.77
Alcohol	0.22	0.4	1.24	0.75	2.06
Sex	0.07	0.8	1.07	0.65	1.84

Table 5: Accuracy of different models.

Model	Accuracy	Recall	F-measure
Logistic Regression	0.736	0.077	0.039
SVR	0.791	0.116	0.115
LSTM	0.834	0.56	0.127

is well fitted. The root-mean-square error of LSTM was 0.314. The accuracy of the LSTM model was significantly superior among the three models, followed by SVM (Table 5). The superior performance of LSTM may be attributed to its ability to capture the temporal relationship in longitudinal data.

Discussion

In this study, we used longitudinal sequence EHR to evaluate the prognosis of MDD. Most of the previous studies have focused on either cross-sectional studies, or longitudinal studies of single chronic condition paired with MDD [3], or suicide [15]. Cross-sectional studies are limited to exploring the relationship between MDD and other factors. Mining EHR data is a hot research topic in healthcare informatics currently. EHR has been widely used in medical prediction tasks, such as disease progression, detection of adverse drug events, diagnosis predictions, etc. [4,10].

This study focused on identifying important individual factors associated with the risk of MDD recurrence. Every recurrence also carries a 10-20% risk of becoming unremitting and chronic, along with a heightened risk for suicide, both of which further compound the serious comorbidities and lethal consequences associated with the MDD [7]. Single marital status is a risk factor for MDD recurrence. Single patients are often kept in a state of loneliness and prone to develop severe depression. The MDD recurrence rate in single patients (11.8%) was obviously higher than that in the patients with marriage/partner status (9.2%). Markkula et al. also found that single marital status was associated with persistence of depressive disorders (OR1.91, 95% CI 1.05-3.56) [16]. Our analysis also showed that the patients who suffered primary MDD had a higher prevalence rate of MDD recurrence (11.7%) than the patients with a secondary MDD (10.5%). A five-year follow-up study of Finnish primary MDD patients by Holma et al. showed that only 50% of patients had complete remission [16]. Another 3-year follow-up study showed that only 43% of patients were recovered [17]. Logistic regression in our results shown that OR was 2.49 (95% confidence interval (CI) 1.53-3.96) between the primary MDD and MDD recurrence. The

patients with more comorbidities had a lower prevalence rate (8.3%) than the patients without complication (12.6%).

Diagnosis prediction is an important and difficult task in the healthcare field [18]. Our analysis indicated that the LSTM model had better prediction accuracy over the logistic regression and SVM. LSTM model can fully exploit temporal information. Our studies further demonstrated that the Recurrent Neural Networks (RNNs) could be used for modeling multivariate time series data in healthcare.

Conclusion

Through extensive evaluation using two large EHR datasets, we presented the risk factors for MDD recurrence, including primary MDD, single marital status, anxiety symptoms, and insomnia.

The LSTM model was significantly superior on the prediction of MDD recurrence than logistic regression and SVM models. The generalizability of the LSTM method was assessed by training and testing this model with the data from two separated EHR databases, and found that the prediction accuracy of this model was dataset-independent.

Supplementary Material

Refer to Table recurrent MDD distribution among different comorbidities in supplementary material.

Acknowledgements

We are grateful for Yaoyun Zhang help and advice. This work was support by UT SBMI. We also thank for Dr. Elmer Bernstam and Susan C. Guerrero whose team remained the University Texas Physicians Clinical Data Warehouse.

References

1. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet*. 2018; 392: 1789-1858.
2. Karina Quevedo, Madeline Harms, Mitchell Sauder, Hannah Scott, Sumaya Mohamed, Kathleen M. Thomas, et al. The neurobiology of self-face recognition among depressed adolescents. *J Affect Disord*. 2018; 229: 22-31.
3. Deborah S. Hasin, Aaron L. Sarvet, Jacquelyn L. Meyers, Tulshi D. Saha, W. June Ruan, Malka Stohl, et al. Epidemiology of Adult DSM-5 Major Depressive Disorder and Its Specifiers in the United States. *JAMA Psychiatry*. 2018. 75(4): 336-346.
4. McKeever A, M Agius, P Mohr. A Review of the Epidemiology of Major Depressive Disorder and of its consequences for Society and the individual. *Psychiatr Danub*. 2017; 29: 222-231.
5. HU Wittchen, F Jacobi, J Rehm, A Gustavsson, M Svensson, B Jönsson, et al. The size and burden of mental disorders and other disorders of the brain in Europe 2010. *Eur Neuropsychopharmacol*. 2011; 21: 655-679.
6. David J Kupfer, Ellen Frank, Mary L Phillips. Major depressive disorder: new clinical, neurobiological, and treatment perspectives. *Lancet*. 2012; 379: 1045-1055.
7. Monroe SM, KL Harkness. Recurrence in major depression: a conceptual analysis. *Psychol Rev*. 2011; 118: 655-674.
8. Adam Mourad Chekroud, Ryan Joseph Zotti, Zarrar Shehzad, Ralitz Gueorgieva, Marcia K Johnson, Madhukar H Trivedi, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *The Lancet Psychiatry*. 2016; 3: 243-250.
9. Zhengxing Huang, Wei Dong, Huilong Duan, Jiquan Liu. A Regularized Deep

- Learning Approach for Clinical Risk Prediction of Acute Coronary Syndrome Using Electronic Health Records. *IEEE Trans Biomed Eng.* 2018; 65: 956-968.
10. Qiuling Suo, Fenglong Ma, Giovanni Canino, Jing Gao, Aidong Zhang, Pierangelo Veltri, et al. A Multi-Task Framework for Monitoring Health Conditions via Attention-based Recurrent Neural Networks. *AMIA Annu Symp Proc.* 2017; 2017: 1665-1674.
 11. Harry Hemingway, Folkert W Asselbergs, John Danesh, Richard Dobson, Nikolaos Maniadakis, Aldo Maggioni, et al. Big data from electronic health records for early and late translational cardiovascular research: challenges and potential. *Eur Heart J.* 2018; 39: 1481-1495.
 12. Bhargava K Reddy, Dursun Delen. Predicting hospital readmission for lupus patients: An RNN-LSTM-based deep-learning methodology. *Comput Biol Med.* 2018; 101: 199-209.
 13. Laila Rasmy, Yonghui Wu, Ningtao Wang, Xin Geng, W Jim Zheng, Fei Wang, et al. A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous EHR data set. *J Biomed Inform.* 2018; 84: 11-16.
 14. E Ryu, AM Chamberlain, RS Pendegraft, TM Petterson, WV Bobo, J Pathak. Quantifying the impact of chronic conditions on a diagnosis of major depressive disorder in adults: a cohort study using linked electronic medical records. *BMC Psychiatry.* 2016; 16: 114.
 15. Niina Markkula, Tommi Härkänen, Tarja Nieminen, Sebastián Peña, Aino K Mattila, Seppo Koskinen, et al. Prognosis of depressive disorders in the general population- results from the longitudinal Finnish Health 2011 Study. *J Affect Disord.* 2016; 190: 687-696.
 16. Holma KM, Holma Irina, AK Melartin, Tarja K, et al. Long-term outcome of major depressive disorder in psychiatric patients is variable. *J Clin Psychiatry.* 2008; 69: 196-205.
 17. BT Stegenga, MH Kamphuis, M King, I Nazareth, MI Geerlings. The natural course and outcome of major depressive disorder in primary care: the PREDICT-NL study. *Soc Psychiatry Psychiatr Epidemiol.* 2012; 47: 87-95.
 18. Ordonez FJ, D Roggen. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors (Basel).* 2016; 16: 115-140.