PREDICTIVE MODELING FOR DEPRESSION WITH CO-MORBIDITIES – RESULTS FROM KOREA NATIONAL HEALTH INSURANCE SERVICES DATA

A Thesis

Presented to the Faculty of the Weill Cornell Graduate School

of Medical Sciences

in Partial Fulfillment of the Requirements for the Degree of

Master of Science

by

Min-hyung Kim

August 2016

ABSTRACT

Depression, despite its high prevalence, remains severely under-diagnosed across the healthcare system. This demands the development of data-driven approaches that can help screen patients who are at a high risk of depression.

In this work, depression risk prediction models that incorporate disease co-morbidities were built on the data from the one million twelve-year longitudinal cohort from Korean National Health Insurance Services (KNHIS), with multiple supervised machine-learning approaches, including decision tree, boost trees, random forest, and support vector machine. Then traditional logistic regression model and Elastic Net regression model were employed in order to leverage the predictive performance and interpretability.

Among the supervised machine-learning approaches, boost trees, random forest, and support vector machine achieved Area Under the Curve of the Receiver Operating Characteristic (AUROC) of 0.793, 0.739, and 0.660, respectively. And Elastic Net regression model achieved an AUROC of 0.7818, compared to a traditional logistic regression model without co-morbidity analysis (AUROC of 0.6992). In addition, Elastic Net regression model showed co-morbidity adjusted Odds Ratios (ORs), which may be more accurate independent estimate of each predictor variable.

In conclusion, the inclusion of co-morbidity analysis with Elastic Net regression model showed the performance of depression risk prediction models comparable to that of supervised machine-learning methods, with providing better interpretability.

BIOGRAPHICAL SKETCH

Min-hyung Kim, M.D., M.S., graduated Seoul Science High School in 2003, Seoul National University College of Medicine in 2011, and Master of Science in Healthcare Informatics program in Weill Cornell Medical College in 2016.

DEDICATION

Special thanks to my mentors – Dr. Sue Kyung Park, Dr. Joo Han Oh,
Dr. Sang Min Park at Seoul National University College of Medicine,
Dr. Young-Su Ju at Hallym University College of Medicine,
and Dr. Yong Auh at Weill Cornell Medical College –
for their rich advice and support in my study at Weill Cornell.

Special thanks to my Weill Cornell professors – especially
Dr. Jyotishman Pathak, Dr. Samprit Banerjee, Dr. Jessica Ancker,
Dr. Arian Hyeyoung Jung, Dr. Mark Unruh, and Dr. Fei Wang –
who guided me through their great teaching.

Special thanks to my friends and alumni – especially
Dr. Young Su Park, Dr. Yongseok Ju, Dr. Hyu Hyunseok Kang, Dr. Sungwhan F. Oh,
Dr. Secheol Oh, Dr. Chul Kim, Dr. Byoung-Il Bae, Dr. Hojoong Kwak,
Dr. Boram Kim, Dr. Hyun-Sik Yang, and Dr. Alex Taekyong Lee –
for their helps and encouragements for my study in the US.

Special love to my grandparents, parents, and my little brother,
who encouraged me to pursue advanced studies.

ACKNOWLEDGMENTS

I thank Dr. Samprit Banerjee, Dr. Sang Min Park, and Dr. Jyotishman Pathak for their advice on completing these works.

I thank Dr. Sang Min Park at Seoul National University College of Medicine and Korea National Health Insurance Service for allowing me to research the data of Korea National Health Insurance Service - National Sample Cohort (NHIS-NSC) 2002~2013.

I also thank Kyuwoong Kim and Jooyoung Chang at Seoul National University College of Medicine for assistance with data management and collaboration for this research.

Finally, I thank Korean Government Scholarship Program and Korea National Institute for International Education, Ministry of Education, for providing financial support for my study at Weill Cornell.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE.
INTRODUCTION

Depression is a highly prevalent disease with a large societal burden. Major depressive disorder has the one-year prevalence of 6%, and the lifetime prevalence of 17%[1], while persistent depressive disorder (dysthymia) has the one-year prevalence of 2%, and the lifetime prevalence of 3%[2]. The estimated societal burden of unipolar depression was 83 billion dollars per year in the US alone in 2007[3].

However, despite this burden, depression is under-diagnosed at large across the health care system in all care settings. A meta-analysis in 2009 concluded that the weighted sensitivity of primary care physicians' diagnosis on depression was only about half (41.3-59.0%) without the assistance of screening tools[4]. This lead to the under-diagnosis or delayed diagnosis of depression, because many of depressed patients initially present with somatic symptoms to the primary care clinics. In general, 69-73% of depression patients presented to their primary care physicians with somatic symptoms, such as pain, fatigue, and sleep problems[5].

Data-driven risk prediction models can be beneficial by rapidly classifying high-risk patients who need further evaluation. Risk prediction models can be implemented on Electronic Health Record system (EHRs) in order to provide clinical decision support. Risk prediction models can also be implemented in health insurance claims data in order to classify high-risk patients, and can be used for accountable care strategy[6].

Previous work on the prediction modelling of depression include a regression-based depression risk prediction model based on Electronic Health Record data,

developed at Stanford University, which reported an area under the receiver operating characteristic (AUROC) of 0.80 for current classification, 0.712 for 6-month prediction, and 0.701 for 12-month prediction[7]. Another work of the depression risk prediction model based on clinical trial data, developed at University of Southern California, reported to have a current classification with an AUROC of 0.81, as well as a sensitivity of 0.65 and a specificity of 0.81 at the institution's optimized threshold[8].

However, both these approaches did not explicitly apply co-morbid medical conditions as independent predictors in the depression prediction model. Many medical conditions can affect depression[9], and depression can also affect certain medical conditions[10]. Therefore, application of co-morbidity analysis can improve the performance of the risk prediction models.

Hence, the main hypothesis and the research question to be addressed in this study was whether the co-morbidity analysis can improve the performance of prediction models for depression risk.

In this work, depression risk prediction models that incorporate disease co-morbidities were built with multiple supervised machine-learning approaches, including decision tree, boost trees, random forest, and support vector machine. Then, regularized regression methods were employed in order to leverage the predictive performance and interpretability.

CHAPTER TWO.
STUDY SETTING AND DATA

In this study, co-morbidity analysis and risk prediction modeling was made from one million twelve-year longitudinal data from Korea National Health Insurance Services (KNHIS)[11]. The sample cohort (N= 1,025,340) was established in 2002 from 2.2% of 46,605,433 individuals from the National Health Information Database (NHID), in order to provide public health researchers and policy makers with representative information regarding the utilization of health insurance and health examinations[12]. The data include demographic profile, health insurance claims data (including in-patient, out-patient, and pharmacy claims), death registry, disability registry, and national health check-up data. With the combination of 18 age groups, 2 genders, and 41 income groups, total 1476 strata were undergone systematic stratified random sampling with proportional allocation[13] within each stratum, using the individual's total annual medical expenses as a target variable. During the follow-up years, annual drop-out by death was 0.5 % (ranging from 4,929 to 5,229). Each year, a representative sample of newborns (ranging from 7,872 to 9,581), sampled across 82 strata (2 for gender, 41 for parents' income group), was added to ensure the representativeness of the data.

The diagnosis codes in KNHIS are based on the Korean Classification of Diseases, Sixth Revision (KCD-6), which is compatible with International Classification of Diseases, Tenth Revision (ICD-10). These diagnoses were classified with Chronic Conditions Data Warehouse (CCW) Condition Algorithms (rev. 01/2016) by Centers for Medicare & Medicaid Services (CMS)[14]. The CCW

condition category algorithms are claims-based algorithms to indicate whether treatment for the condition appears to have taken place, which include 27 chronic condition categories and 33 other chronic or potentially disabling conditions categories. **Table 1** shows the ICD-10 codes of the depressive disorder, bipolar disorder, schizophrenia in the CMS-CCW algorithm. The ICD-10 codes for depressive disorder, bipolar disorder, schizophrenia were used in the operational definition of the depression case group in this study. The study subjects had two or more encounters with depression diagnosis codes, but less than two encounters with either bipolar or schizophrenia diagnosis codes. The inclusion and exclusion criteria is based on the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5)[15].

Table 1. ICD-10 codes of the depressive disorder, bipolar disorder, schizophrenia in the Chronic Conditions Data Warehouse (CCW) Condition Algorithms (rev. 01/2016) by Centers for Medicare & Medicaid Services (CMS).

CMS-CCW Conditions	Valid ICD-10 Codes
Depressive Disorders	F32.0, F32.1, F32.2, F32.3, F32.4, F32.5, F32.9, F33.0, F33.1, F33.2, F33.3, F33.40, F33.41, F33.42, F33.9, F34.1, Z13.89
Bipolar Disorder	F30.10, F30.11, F30.12, F30.13, F30.2, F30.3, F30.4, F30.8, F30.9, F31.0, F31.10, F31.11, F31.12, F31.13, F31.2, F31.30, F31.31, F31.32, F31.4, F31.5, F31.60, F31.61, F31.62, F31.63, F31.64, F31.70, F31.71, F31.72, F31.73, F31.74, F31.75, F31.76, F31.77, F31.78, F31.81, F31.89, F31.9, F32.8, F33.8, F34.8, F34.9, F39
Schizophrenia	F20.0, F20.1, F20.2, F20.3, F20.5, F20.81, F20.89, F20.9, F25.0, F25.1, F25.8, F25.9

The univariate and bivariate statistics of selected demographic and co-morbidity variables between the depression case group (N=28,256) and the complement comparison group (N=1,085,400), based on the operational definition of the case group described in the method section, are shown in **Table 2**. The case group showed significantly higher percentage of females (68.1%), age (mean 48, standard deviation 19), income decile (mean 5.8, standard deviation 2.5), limb disability (3.3%), neurologic disability (0.9%), visual disability (0.7%), hearing disability (0.6%), but showed significantly lower percentage of social security beneficiaries (1.0%). The case group showed no significant difference in the percentage of residents in Seoul metropolitan area, and cognitive disability. Most of the co-morbidity variables showed statistically significant difference between the case group and comparison group, except cerebral palsy ($p = 0.121$). The most noticeable difference in the co-morbidity by the ratio of percentage (twelve-year prevalence) was personality disorder (0.9% vs 0.1%), followed by anxiety disorder (31.8% vs 4.9%), dementia & Alzheimer's disease (2.0% vs 0.4%), and osteoporosis (5.2% vs 1.3%).

Table 2. Univariate and bivariate statistics of selected demographic, socio-economic, disability registry and co-morbidity variables between the depression case group (N=28,256) and complement comparison group (N=1,085,400). For categorical variables, the observed frequencies of the categories and percentages (twelve-year prevalence) were reported, and for numerical variables, means (and standard deviations) were reported. P-values were of chi-square tests for categorical variables, and t-tests for numerical variables.

Univariate and Bivariate Statistics of Selected Variables	Complement Control Group (N=1085400)	Depression Case Group (N=28256)	Total (N=1113656)	P-value
< Demographic and Socio-Economic Variables >				
Female	527644 (49.4%)	19238 (68.1%)	546882 (49.9%)	<0.001
Age	36 ± 21	48 ± 19	36 ± 21	<0.001
Income (Decile)	5.7 ± 2.5	5.8 ± 2.5	5.7 ± 2.5	<0.001
Insurance Status - Social Security Beneficiaries	17288 (1.6%)	288 (1.0%)	17576 (1.6%)	<0.001
Residents In Seoul Metropolitan Area	158874 (14.9%)	4138 (14.6%)	163012 (14.9%)	0.278
< Disability Registry Variables >				
Limb Disability	20828 (2.0%)	936 (3.3%)	21764 (2.0%)	<0.001
Neurologic Disability	5528 (0.5%)	258 (0.9%)	5786 (0.5%)	<0.001
Visual Disability	4321 (0.4%)	197 (0.7%)	4518 (0.4%)	<0.001
Hearing Disability	3467 (0.3%)	172 (0.6%)	3639 (0.3%)	<0.001
Cognitive Disability	2791 (0.3%)	63 (0.2%)	2854 (0.3%)	0.233
< Co-morbidity Variables (Alphabetical Order) >				
Acquired Hypothyroidism	24496 (2.3%)	1805 (6.4%)	26301 (2.4%)	<0.001
Acute Myocardial Infarction	2518 (0.2%)	115 (0.4%)	2633 (0.2%)	<0.001
Alzheimer's Disease	1707 (0.2%)	242 (0.9%)	1949 (0.2%)	<0.001
Anemia	60602 (5.7%)	2875 (10.2%)	63477 (5.8%)	<0.001
Anxiety Disorder	51935 (4.9%)	8980 (31.8%)	60915 (5.6%)	<0.001
Arthritis	213785 (20.0%)	12578 (44.5%)	226363 (20.7%)	<0.001
Asthma	191650 (17.9%)	6468 (22.9%)	198118 (18.1%)	<0.001
Atrial Fibrillation	3730 (0.3%)	245 (0.9%)	3975 (0.4%)	<0.001
Attention Deficit Hyperactivity & Conduct Disorder	5932 (0.6%)	444 (1.6%)	6376 (0.6%)	<0.001
Benign Prostatic Hyperplasia	31351 (2.9%)	1673 (5.9%)	33024 (3.0%)	<0.001
Brain Injury	15399 (1.4%)	728 (2.6%)	16127 (1.5%)	<0.001
Breast Cancer	2976 (0.3%)	199 (0.7%)	3175 (0.3%)	<0.001
Cataract	66979 (6.3%)	4931 (17.5%)	71910 (6.6%)	<0.001
Cerebral Palsy	1004 (0.1%)	18 (0.1%)	1022 (0.1%)	0.121
Chronic Kidney Disorder	39841 (3.7%)	2324 (8.2%)	42165 (3.8%)	<0.001
Chronic Obstructive Pulmonary Disease (COPD)	74832 (7.0%)	3715 (13.1%)	78547 (7.2%)	<0.001
Chronic Ulcers	2924 (0.3%)	210 (0.7%)	3134 (0.3%)	<0.001
Colorectal Cancer	4982 (0.5%)	260 (0.9%)	5242 (0.5%)	<0.001
Dementia	2589 (0.2%)	320 (1.1%)	2909 (0.3%)	<0.001
Diabetes	80495 (7.5%)	4708 (16.7%)	85203 (7.8%)	<0.001
Endometrial Cancer	513 (0.0%)	41 (0.1%)	554 (0.1%)	<0.001
Epilepsy	8698 (0.8%)	665 (2.4%)	9363 (0.9%)	<0.001
Fibromyalgia and Pain Syndrome	69260 (6.5%)	3995 (14.1%)	73255 (6.7%)	<0.001
Glaucoma	37216 (3.5%)	2123 (7.5%)	39339 (3.6%)	<0.001
Hearing Impairment	40214 (3.8%)	2484 (8.8%)	42698 (3.9%)	<0.001
Heart Failure	19199 (1.8%)	1408 (5.0%)	20607 (1.9%)	<0.001
Hyperlipidemia	89161 (8.4%)	5700 (20.2%)	94861 (8.7%)	<0.001
Hypertension	57022 (5.3%)	3531 (12.5%)	60553 (5.5%)	<0.001
Ischemic Heart Disease	47457 (4.4%)	3692 (13.1%)	51149 (4.7%)	<0.001
Leukemia And Lymphoma	1461 (0.1%)	64 (0.2%)	1525 (0.1%)	<0.001
Liver Disease (Except Viral Hepatitis)	127475 (11.9%)	6552 (23.2%)	134027 (12.2%)	<0.001
Lung Cancer	4217 (0.4%)	212 (0.8%)	4429 (0.4%)	<0.001
Migraine And Chronic Headache	119753 (11.2%)	8192 (29.0%)	127945 (11.7%)	<0.001
Mobility Impairments	11480 (1.1%)	866 (3.1%)	12346 (1.1%)	<0.001
Osteoporosis	13986 (1.3%)	1463 (5.2%)	15449 (1.4%)	<0.001
Pelvic Fractures	3902 (0.4%)	299 (1.1%)	4201 (0.4%)	<0.001
Peripheral Vascular Disease	22793 (2.1%)	1852 (6.6%)	24645 (2.2%)	<0.001
Personality Disorders	980 (0.1%)	244 (0.9%)	1224 (0.1%)	<0.001
Spinal Cord Injury	8191 (0.8%)	590 (2.1%)	8781 (0.8%)	<0.001
Stroke And Transient Ischemic Attack (TIA)	49791 (4.7%)	3946 (14.0%)	53737 (4.9%)	<0.001
Viral Hepatitis	43294 (4.1%)	1877 (6.6%)	45171 (4.1%)	<0.001

CHAPTER THREE.

ANALYTIC APPROACH

For supervised machine-learning approaches, 10% of the data (N = 111,366) was set aside as a test data, another 10% of the data (N = 111,366) was used as parameter tuning validation data, and 80% of the data was used as training data (N = 890,924), in order to simulate the prediction performance for unseen data. Then predictive models for depression based on demographic and co-morbidity features were trained with decision tree, boost trees, random forest, and support vector machine algorithms.

Tree-based algorithms are based on recursive partitioning, until the subsets are sufficiently homogeneous, or stopping criterion has been met[16]. Deciding the best split on each recursive partitioning step is based on either purity measures, such as entropy and information gain, or statistical significance testing. For the development of decision tree, conditional inference tree algorithm based on significance testing in R software package party[17] was used with tuning parameters of minimum split 20, minimum bucket 7, and maximum depth 30.

Boosting is a machine learning ensemble meta-algorithm for reducing bias and variance in supervised learning[18]. In ensemble methods, a set of weak learners are combined to create a strong learner. In boosting, sequential models are built to fit the residuals, or incorrectly classified observations in previous iteration, by weighing the problematic observations. For boost trees, 50 trees with tuning parameters of minimum split 20, maximum depth 30, complexity parameter 0.01 were trained with R software package ada[19].

Random forest is another ensemble machine learning approach to supervised learning[15]. The algorithm builds trees based on bootstrap aggregation (bagging) of the observations, as well as random sampling of the variables while building each node, in order to de-correlate the trees built. For random forest, 500 trees with 7 variables were trained based on Gini index with R software package randomForest[20].

Support vector machine is a machine learning method for identifying an optimal hyperplane for partitioning the classes in the multi-dimensional feature space[21]. Kernel function is employed to map the data into higher dimension space, in order to make more linearly separable. A classification support vector machine was trained with radial basis function, and the R software package kernlab[22] provided automatic sigma parameter estimation function.

Then the operational definition of diagnosis of depression was analyzed in a logistic regression model with socio-economic and co-morbid predictors. Among the available socio-economic variables and co-morbid conditions in KNHIS data, variables for the final logistic regression model was selected with Elastic Net[23]. The performance of the final logistic regression model with co-morbidity analysis was compared with that of the traditional logistic regression model without co-morbidity analysis.

When the number of predictors is large compared to the sample size, traditional variable selection methodologies may have poor prediction performance for external datasets by overfitting random error or noise, and it has been criticized that the goodness of fit[24], significance[25], and degrees of freedom[26] do not reflect the reality. In order to overcome this problem, regularization and shrinkage methods for regression have been developed[27]. Elastic Net is a regularization method for regression and classification models which compromises the Least Absolute Shrinkage And Selection Operator (LASSO) penalty (L_1) and the ridge penalty

(L_2)[23]. 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 toward their average. The overall Elastic Net is a function of parameters λ and α ($0 \leq \alpha \leq 1$), where λ being a parameter for the level of penalty, while α being the weight of L_1 penalty and $(1 - \alpha)$ being that of L_2 penalty function. Hence, in this work, variable selection and penalization of collinear predictors were performed by Elastic Net for developing the final logistic regression model.

A robust way to determine the best combination of λ and α is via a k-fold cross-validation. For the performance test of the predictive model, 10% of the data (N = 111,366) was set aside as a test data, and 90% of the data was used as a training data (N = 1,002,290). 10-fold cross-validation on training data was employed, where total observations of the dataset are randomly divided into 10 folds, or partitions. One of the 10 folds is reserved as the internal validation data (N = 100,229), and the rest of the folds consist the internal training data (N = 902,061), where statistical models are fitted. After fitting the models, or calculating the coefficients, the models are validated against the reserved fold. This overall process is iterated (repeated) 10 times, so that every folds can be a validation set. This is a preferred method especially when the prediction models need to perform prediction for external datasets, that is, outside of the overall dataset used in the research.

The variables for the traditional logistic regression model without co-morbidity analysis was driven by performing the stepwise backward selection using Akaike's Information Criterion (AIC)[28]. The selected variables for the traditional logistic regression model include sex, age, income decile, and disability registration. The variable selection for the final logistic regression model was applied with Elastic Net from the training data, as described above. The selected value of α was 0.75, and the

optimized values of λ was 0.001390648 (log(λ) -6.577986), although other α values, including 0.25, 0.5, and 1, did not change the results much. The plot obtained from cross-validation of Elastic Net, showing the change of the Area Under the Curve (AUC) of ROC with different λ (in log scale) for a model assuming an α of 0.75, is shown in **Figure 1.** This gives a minimum of 28 variables needed for building an optimized model. Two more variables, acute myocardial infarction and dementia, were added to the final model, because even though those conditions were separated by the CMS-CCW algorithm, the conditions were in spectrum with ischemic heart disease and Alzheimer's disease, respectively. Therefore, 30 variables were selected for the final logistic regression model. These variables include: sex, age, income decile, acquired hypothyroidism, acute myocardial infarction, Attention Deficit Hyperactivity Disorder (ADHD) and conduct disorder, Alzheimer's disease, anemia, anxiety disorder, arthritis, atrial fibrillation, brain injury, chronic kidney disorder , colorectal cancer, chronic obstructive pulmonary disease (COPD), dementia, diabetes, epilepsy, glaucoma, hearing impairment, hyperlipidemia, ischemic heart disease, liver disease (except viral hepatitis), migraine and chronic headache, mobility impairments, osteoporosis, peripheral vascular disease, personality disorders, stroke and transient ischemic attack (TIA), and viral hepatitis.

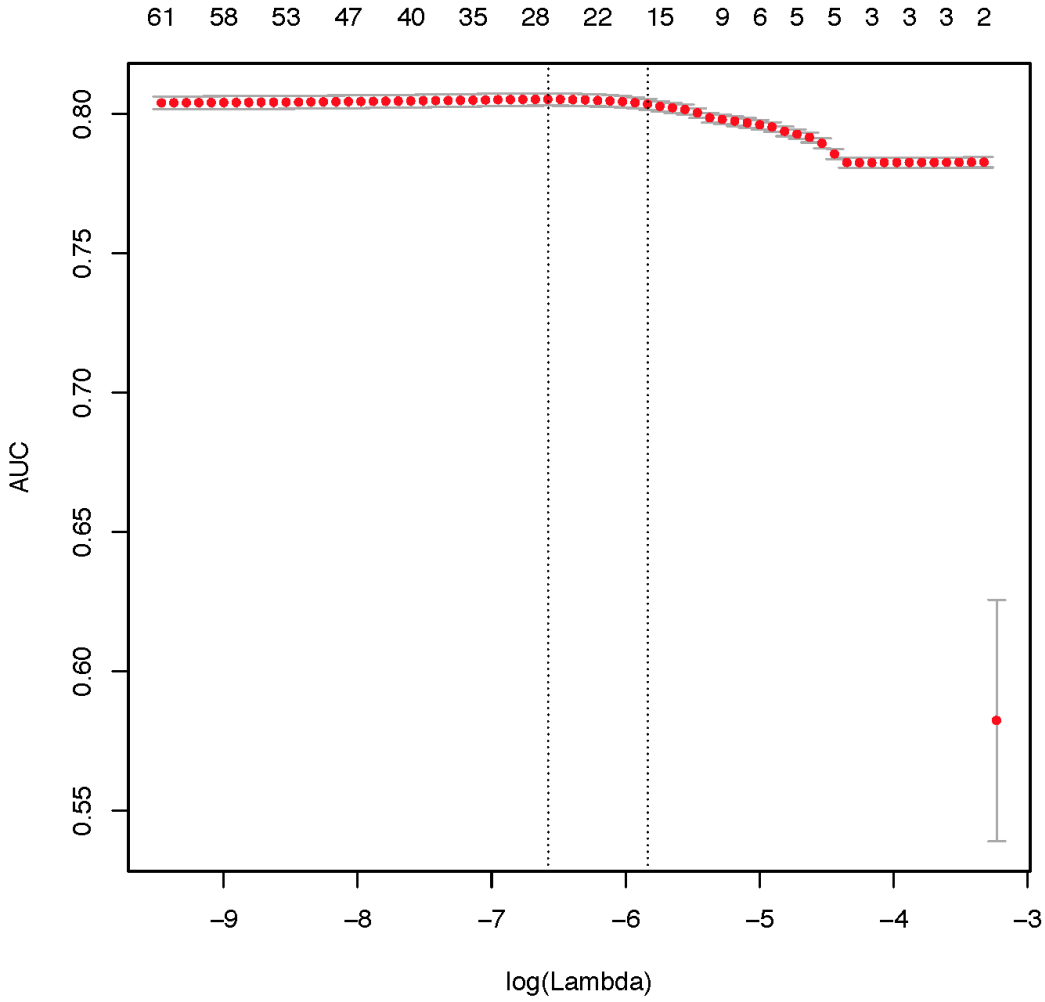

Figure 1. The plot obtained from cross-validation of Elastic Net, showing the change of the Area Under the Curve (AUC) of Receiver Operating Characteristic (ROC) with different λ (in log scale) with α of 0.75. The numeric values above the plot indicates the number of variables selected in the between 28 and 61. In other words, 28 (when log(λ) is -6.577986) is the minimum number of variables that guarantees the maximum AUC.

In order to get more robust Receiver Operation Characteristics (ROC) that reflect the prediction performance also for external datasets, another layer of validation on the test data (N = 111,366), which was set aside and unseen during the training phase, was applied to derive ROC. With the variable selected via Elastic Net, a final logistic regression model was built with co-morbidity analysis, and obtained the ROC of the final logistic regression model with co-morbidity analysis on the test data. Then the ROC was compared with that of the traditional logistic regression model without co-morbidity analysis. R version 3.1.3[29] with R software packages, glmnet[30], and pROC[31] were used for this study.

CHAPTER FOUR.

RESULTS

A conditional inference tree with 73 terminal nodes built on the training data are represented in **Figure 2.** With appropriate threshold, the decision tree model achieved sensitivity of 0.497, specificity of 0.872, Positive Predictive Value of 0.0935, Negative Predictive Value of 0.985, Accuracy of 0.865, and F measure of 0.156.

Figure 2. A conditional inference tree with 73 terminal nodes built on the training data. With appropriate threshold, the decision tree model achieved sensitivity of 0.497, specificity of 0.872, Positive Predictive Value of 0.0935, Negative Predictive Value of 0.985, Accuracy of 0.865, and F measure of 0.156.

Figure 2.

```
1) Anxiety <= 0; criterion = 1, statistic = 26896.224
  2) Age <= 42; criterion = 1, statistic = 4068.764
    3) Age <= 7; criterion = 1, statistic = 1360.102
      4) ADHD_Conduct <= 0; criterion = 1, statistic = 1041.09
        5) Age <= 2; criterion = 1, statistic = 318.724
          6)*  weights = 79905
        5) Age > 2
          7) Female <= 0; criterion = 1, statistic = 26.68
            8) Epilepsy <= 0; criterion = 1, statistic = 27.449
              9)*  weights = 24129
            8) Epilepsy > 0
              10)*  weights = 295
          7) Female > 0
            11)*  weights = 22862
      4) ADHD_Conduct > 0
        12) Age <= 2; criterion = 1, statistic = 47.248
          13)*  weights = 1390
        12) Age > 2
          14) Liver_except_viral <= 0; criterion = 0.984, statistic = 13.512
            15)*  weights = 1451
          14) Liver_except_viral > 0
            16)*  weights = 84
    3) Age > 7
      17) Female <= 0; criterion = 1, statistic = 962.568
        18)*  weights = 215405
      17) Female > 0
        19) Migraine_ChronicHeadache <= 0; criterion = 1, statistic = 351.285
          20) Liver_except_viral <= 0; criterion = 1, statistic = 304.994
            21)*  weights = 146761
          20) Liver_except_viral > 0
            22) Hyperlipidemia <= 0; criterion = 0.999, statistic = 142.916
              23)*  weights = 13252
            22) Hyperlipidemia > 0
              24)*  weights = 1470
        19) Migraine_ChronicHeadache > 0
          25) Hyperlipidemia <= 0; criterion = 1, statistic = 52.031
            26) IschemicHeart <= 0; criterion = 1, statistic = 43.501
              27)*  weights = 24059
            26) IschemicHeart > 0
              28)*  weights = 562
          25) Hyperlipidemia > 0
            29)*  weights = 1999
  2) Age > 42
    30) Migraine_ChronicHeadache <= 0; criterion = 1, statistic = 1123.736
      31) Arthritis <= 0; criterion = 1, statistic = 486.759
        32) Female <= 0; criterion = 1, statistic = 135.843
          33) StrokeTIA <= 0; criterion = 1, statistic = 75.902
            34) BenignProstatic <= 0; criterion = 1, statistic = 62.162
              35)*  weights = 46398
            34) BenignProstatic > 0
              36)*  weights = 6183
          33) StrokeTIA > 0
            37) IschemicHeart <= 0; criterion = 0.987, statistic = 34.764
              38)*  weights = 5333
            37) IschemicHeart > 0
              39) Glaucoma <= 0; criterion = 0.984, statistic = 13.552
```

Figure 2. (continued)

```
                    40)*    weights = 931
                 39) Glaucoma > 0
                    41)*    weights = 102
           32) Female > 0
              42) ChronicKidney <= 0; criterion = 1, statistic = 41.349
                 43) Osteoporosis <= 0; criterion = 1, statistic = 35.408
                    44) StrokeTIA <= 0; criterion = 1, statistic = 41.302
                       45) Asthma <= 0; criterion = 0.999, statistic = 34.791
                          46) IschemicHeart <= 0; criterion = 0.999, statistic = 43.574
                             47)*    weights = 23909
                          46) IschemicHeart > 0
                             48)*    weights = 1324
                       45) Asthma > 0
                          49) PeripheralVascularDisease <= 0; criterion = 0.983, statistic = 13.497
                             50)*    weights = 3552
                          49) PeripheralVascularDisease > 0
                             51)*    weights = 122
                    44) StrokeTIA > 0
                       52)*    weights = 3170
                 43) Osteoporosis > 0
                    53)*    weights = 1395
              42) ChronicKidney > 0
                 54) Liver_except_viral <= 0; criterion = 0.996, statistic = 19.838
                    55) HeartFailure <= 0; criterion = 0.966, statistic = 12.089
                       56)*    weights = 1615
                    55) HeartFailure > 0
                       57)*    weights = 164
                 54) Liver_except_viral > 0
                    58)*    weights = 284
        31) Arthritis > 0
           59) Cataract <= 0; criterion = 1, statistic = 158.759
              60) Female <= 0; criterion = 1, statistic = 117.624
                 61)*    weights = 22723
              60) Female > 0
                 62) Hyperlipidemia <= 0; criterion = 1, statistic = 71.826
                    63)*    weights = 27163
                 62) Hyperlipidemia > 0
                    64)*    weights = 7894
           59) Cataract > 0
              65) Female <= 0; criterion = 1, statistic = 28.402
                 66) BenignProstatic <= 0; criterion = 0.999, statistic = 18.62
                    67)*    weights = 4557
                 66) BenignProstatic > 0
                    68)*    weights = 2118
              65) Female > 0
                 69) Asthma <= 0; criterion = 1, statistic = 22.132
                    70) Dementia <= 0; criterion = 0.996, statistic = 16.032
                       71) Hyperlipidemia <= 0; criterion = 0.996, statistic = 16.309
                          72)*    weights = 7359
                       71) Hyperlipidemia > 0
                         73)*    weights = 2327
                    70) Dementia > 0
                       74)*    weights = 146
                 69) Asthma > 0
                    75) Income <= 7; criterion = 0.997, statistic = 16.843
```

Figure 2. (continued)

```
                    76)*  weights = 2513
                  75) Income > 7
                    77)*  weights = 1308
      30) Migraine_ChronicHeadache > 0
        78) Female <= 0; criterion = 1, statistic = 110.777
          79) Epilepsy <= 0; criterion = 1, statistic = 39.041
            80) Liver_except_viral <= 0; criterion = 1, statistic = 22.325
              81) Anemia <= 0; criterion = 1, statistic = 29.207
                82) DFAB_MINOR_TRUE <= 0; criterion = 0.996, statistic = 16.155
                  83) COPD <= 0; criterion = 0.976, statistic = 19.611
                    84)*  weights = 5607
                  83) COPD > 0
                    85)*  weights = 938
                82) DFAB_MINOR_TRUE > 0
                  86)*  weights = 354
              81) Anemia > 0
                87)*  weights = 319
            80) Liver_except_viral > 0
              88) StrokeTIA <= 0; criterion = 0.972, statistic = 20.781
                89)*  weights = 2250
              88) StrokeTIA > 0
                90)*  weights = 542
          79) Epilepsy > 0
            91)*  weights = 161
        78) Female > 0
          92) Arthritis <= 0; criterion = 1, statistic = 49.827
            93) ViralHepatitis <= 0; criterion = 0.973, statistic = 23.728
              94)*  weights = 5231
            93) ViralHepatitis > 0
              95)*  weights = 281
          92) Arthritis > 0
            96) IschemicHeart <= 0; criterion = 1, statistic = 38.004
              97) Liver_except_viral <= 0; criterion = 1, statistic = 24.682
                98) StrokeTIA <= 0; criterion = 0.982, statistic = 13.323
                  99)*  weights = 8090
                98) StrokeTIA > 0
                  100)*  weights = 1817
              97) Liver_except_viral > 0
                101) MobilityImpairments <= 0; criterion = 0.967, statistic = 12.183
                  102)*  weights = 2480
                101) MobilityImpairments > 0
                  103)*  weights = 54
            96) IschemicHeart > 0
              104)*  weights = 2438
1) Anxiety > 0
  105) Age <= 37; criterion = 1, statistic = 305.218
    106) Age <= 7; criterion = 1, statistic = 105.772
      107)*  weights = 1578
    106) Age > 7
      108) Liver_except_viral <= 0; criterion = 1, statistic = 51.925
        109) ADHD_Conduct <= 0; criterion = 1, statistic = 44.258
          110) Age <= 22; criterion = 1, statistic = 28.168
            111)*  weights = 4159
          110) Age > 22
            112) Female <= 0; criterion = 0.998, statistic = 17.112
```

Figure 2. (continued)

```
                    113)*    weights = 2398
                  112) Female > 0
                    114)*    weights = 4538
            109) ADHD_Conduct > 0
              115)*    weights = 120
          108) Liver_except_viral > 0
            116) PersonalityDisorders <= 0; criterion = 0.955, statistic = 11.565
              117)*    weights = 2709
            116) PersonalityDisorders > 0
              118)*    weights = 39
    105) Age > 37
      119) Migraine_ChronicHeadache <= 0; criterion = 1, statistic = 134.844
        120) Income <= 4; criterion = 1, statistic = 55.794
          121) Liver_except_viral <= 0; criterion = 1, statistic = 25.2
            122) Osteoporosis <= 0; criterion = 0.974, statistic = 12.656
              123)*    weights = 4347
            122) Osteoporosis > 0
              124)*    weights = 222
          121) Liver_except_viral > 0
            125)*    weights = 1271
        120) Income > 4
          126) Cataract <= 0; criterion = 1, statistic = 37.512
            127) Liver_except_viral <= 0; criterion = 1, statistic = 26.664
              128) StrokeTIA <= 0; criterion = 1, statistic = 24.121
                129) Income <= 6; criterion = 0.973, statistic = 12.576
                  130)*    weights = 2206
                129) Income > 6
                  131)*    weights = 3779
              128) StrokeTIA > 0
                132)*    weights = 863
            127) Liver_except_viral > 0
              133)*    weights = 2141
          126) Cataract > 0
            134)*    weights = 3123
      119) Migraine_ChronicHeadache > 0
        135) StrokeTIA <= 0; criterion = 1, statistic = 26.293
          136) ChronicKidney <= 0; criterion = 1, statistic = 23.634
            137) Osteoporosis <= 0; criterion = 0.983, statistic = 13.408
              138) COPD <= 0; criterion = 0.992, statistic = 14.872
                139)*    weights = 4888
              138) COPD > 0
                140)*    weights = 1022
            137) Osteoporosis > 0
              141)*    weights = 523

          136) ChronicKidney > 0
            142)*    weights = 653
        135) StrokeTIA > 0
          143) Dementia <= 0; criterion = 0.96, statistic = 11.796
            144)*    weights = 2140
          143) Dementia > 0
            145)*    weights = 64
```

Receiver Operating Characteristic curve of supervised machine-learning approaches, including boost trees, random forest, and support vector machine, are shown in **Figure 3**. AUROC achieved on the test set were 0.793, 0.739, 0.660 for boost trees, random forest, and support vector machine, respectively.

Boost trees achieved train error of 0.024, Out-of-Bag (OOB) error of 0.023 on 6 iterations. Variables actually used in tree construction include, Age, Anxiety, Dysthymia, Adjustment Disorder, Arthritis, Migraine/Chronic Headache, Female, Epilepsy, Liver Disease (except viral), Dementia, Stroke/Transient Ischemic Attack, Fibromyalgia/Pain/Fatigue, Hyperlipidemia, Benign Prostatic Hyperplasia, Cataract, Glaucoma, and Ischemic Heart Disease, in the order of decreasing variable usage frequency.

Random forest achieved OOB estimate of error rate of 0.024, and variables associated with more than 100 of mean decrease in Gini index include Anxiety, Age, Income, Dysthymia, Migraine/Headache, Arthritis, Fibromyalgia/Pain/Fatigue, Hyperlipidemia, Diabetes, and Asthma.

In support vector machine, probability classification model achieved training error of 0.037 with cost parameter of 1, Gaussian Radial Basis kernel function sigma parameter of 0.014, and 2848 support vectors.

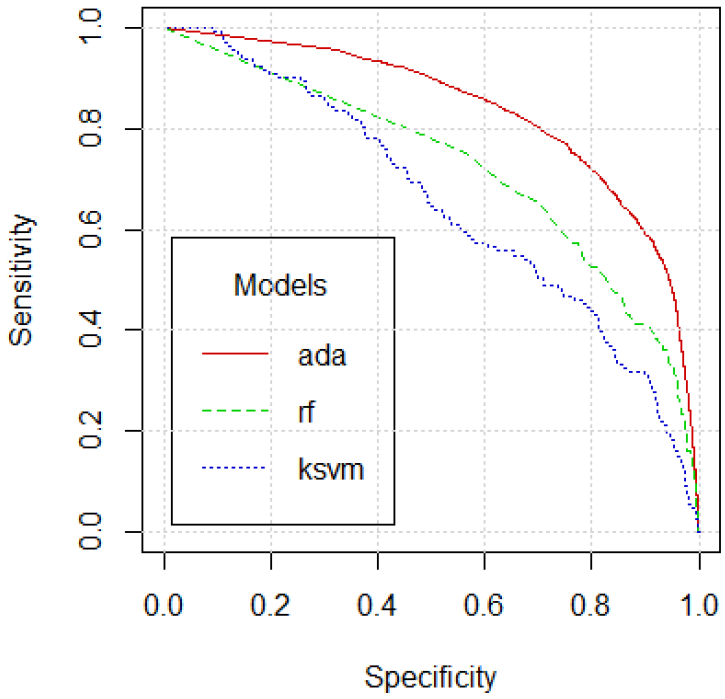

Figure 3. Receiver Operating Characteristic curve of supervised machine-learning approaches, including decision tree, boost trees, random forest, and support vector machine. (ada: boost trees model built with R software package ada[19]. rf: random forest built with R software package randomForest[20]. ksvm: support vector machine built with R software package kernlab[22].)

The Odds Ratio (OR) plot for the traditional logistic regression model without co-morbidity analysis is presented in **Figure 4 (a)** and the same for the final logistic regression model with co-morbidity analysis in **Figure 4 (b).** It is noticeable that adjusted ORs for the same variables differs between the two models. For example, the adjusted OR of being female is 2.07 from the traditional logistic regression model without co-morbidity analysis, but is 1.63 from the final logistic regression model with co-morbidity analysis. Likewise, the adjusted OR of age is 1.03 from the traditional logistic regression model without co-morbidity analysis, but is 1.01 from the final logistic regression model with co-morbidity analysis. Finally, the adjusted OR of income decile is 1.04 from the traditional logistic regression model without co-morbidity analysis, but is 1.02 from the final logistic regression model with co-morbidity analysis. The ORs for the disability registration variables in the traditional logistic regression model without co-morbidity analysis ranged from 1.03 (cognitive disability) to 1.42 (hearing disability). The ORs for the co-morbidity variables in the final logistic regression model with co-morbidity analysis ranged from 0.78 (acute myocardial infarction) to 5.81 (ADHD and conduct disorder).

Figure 4 (a) Odds Ratio (OR) plot of the traditional logistic regression model without co-morbidity analysis. **(b)** Odds Ratio (OR) plot of the final logistic regression model with co-morbidity analysis. The point values indicate the adjusted Odds Ratios, horizontal lines indicate the 95% confidence intervals, and the asterisks indicate the level of statistical significance (***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$). The variables included in both the traditional model without co-morbidity analysis (**a**) and the final model with co-morbidity analysis (**b**) are highlighted in yellow. The adjusted Odds Ratios can differ if variable selection is different. For example, the adjusted OR of being female is 2.07 from the traditional logistic regression model without co-morbidity analysis, but is 1.63 from the final logistic regression model with co-morbidity analysis.

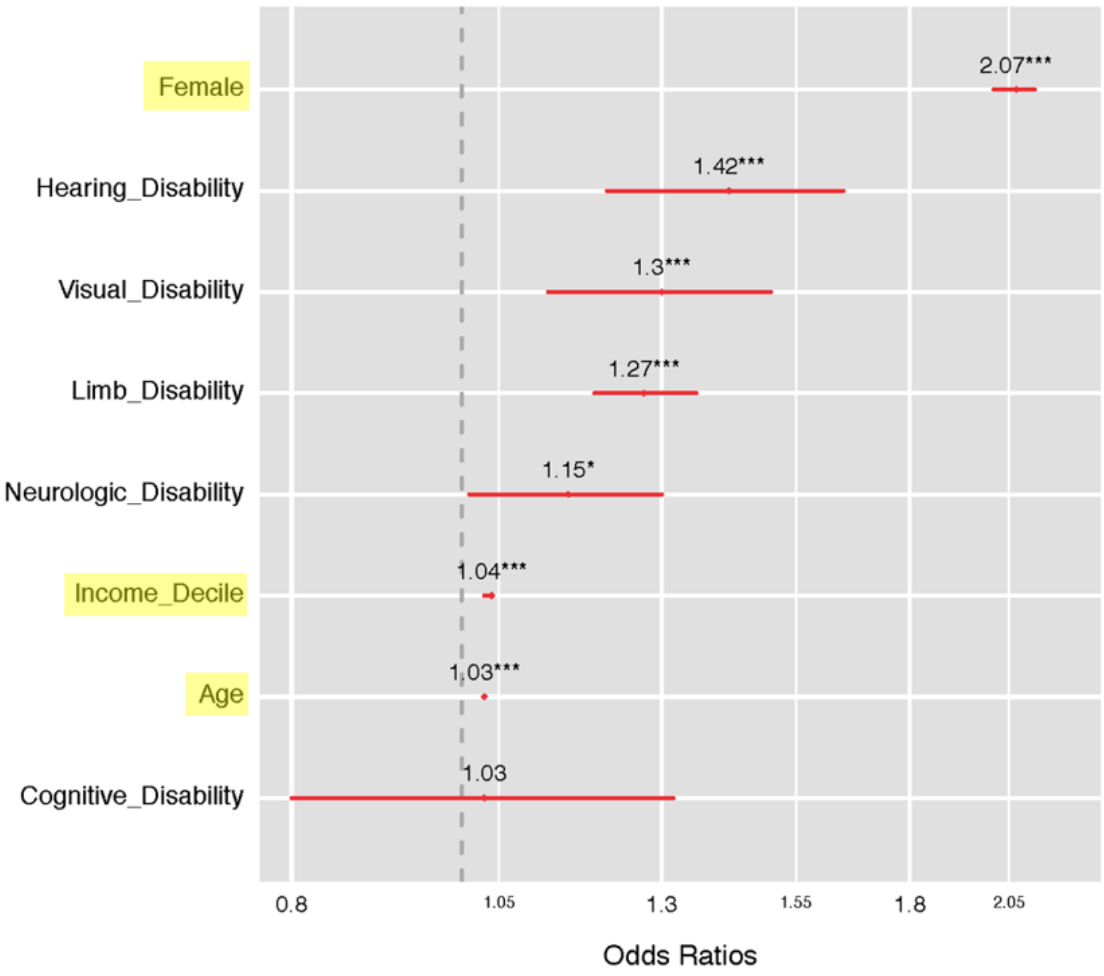
Female
Hearing_Disability
Visual_Disability
Limb_Disability
Neurologic_Disability
Income_Decile
Age
Cognitive_Disability
2.07***
1.42***
1.3***
1.27***
1.15*
1.04***
1.03***
1.03
0.8
1.05
1.3
1.55
1.8
2.05
Odds Ratios

Figure 4 (a)

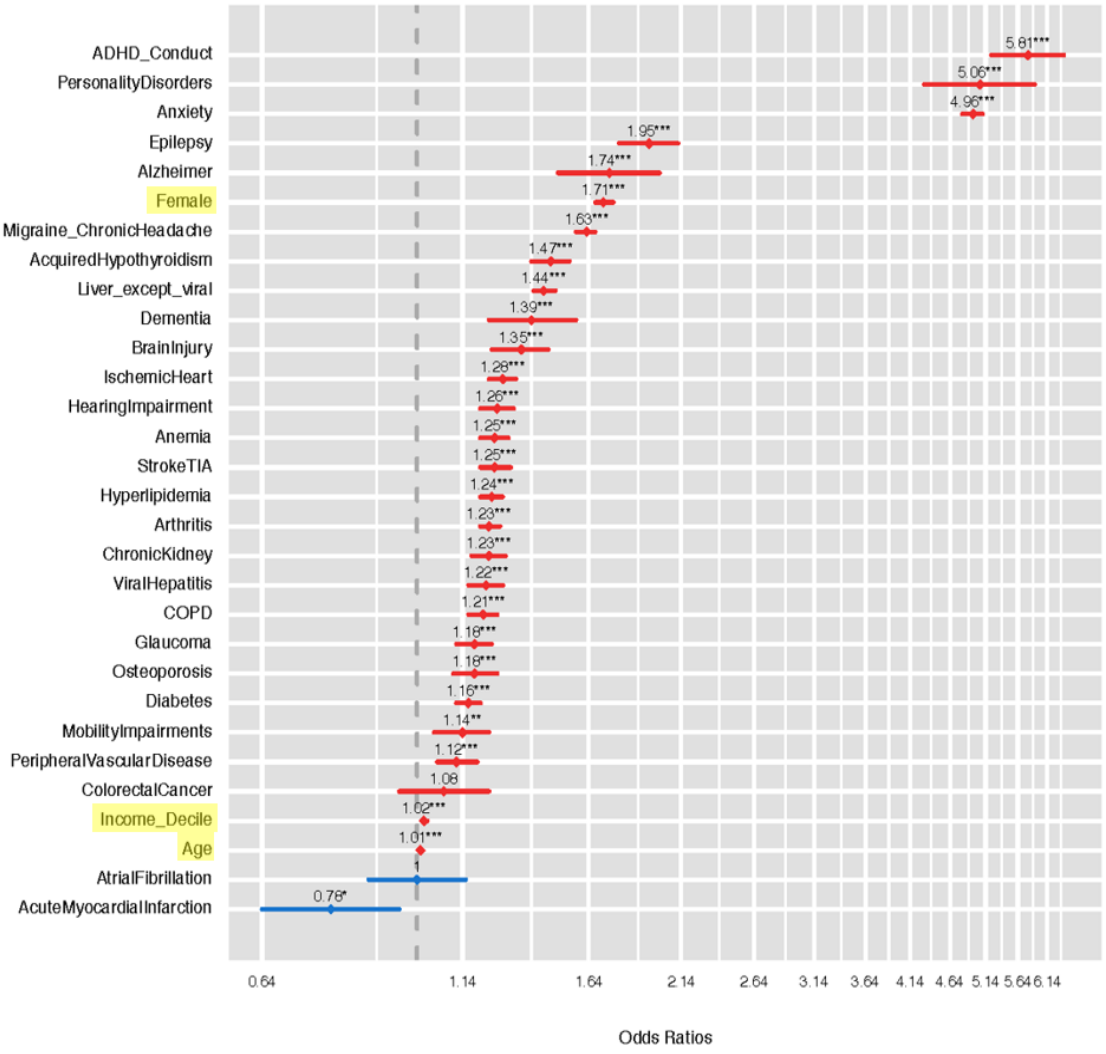
ADHD_Conduct
PersonalityDisorders
Anxiety
Epilepsy
Alzheimer
Female
Migraine_ChronicHeadache
AcquiredHypothyroidism
Liver_except_viral
Dementia
BrainInjury
IschemicHeart
HearingImpairment
Anemia
StrokeTIA
Hyperlipidemia
Arthritis
ChronicKidney
ViralHepatitis
COPD
Glaucoma
Osteoporosis
Diabetes
MobilityImpairments
PeripheralVascularDisease
ColorectalCancer
Income_Decile
Age
AtrialFibrillation
AcuteMyocardialInfarction
5.81***
5.06***
4.96***
1.95***
1.74***
1.71***
1.63***
1.47***
1.44***
1.39***
1.35***
1.28***
1.26***
1.25***
1.25***
1.24***
1.23***
1.23***
1.22***
1.21***
1.18***
1.18***
1.16***
1.14**
1.12***
1.08
1.02***
1.01***
1
0.76*
0.64
1.14
1.64
2.14
2.64
3.14
3.64
4.14
4.64
5.14
5.64
6.14
Odds Ratios

Figure 4 (b)

Receiver Operating Characteristic (ROC) curve of the traditional logistic regression model without co-morbidity analysis and the final logistic regression model with co-morbidity analysis on the test data, which were unseen during the training phase, are shown in **Figure 5.** The Area Under the Curve (AUC) of the ROC increased from 0.6992 (the traditional logistic regression model without co-morbidity analysis) to 0.7818 (the final logistic regression model with co-morbidity analysis). Selected performance measures for the 30-predictor co-morbidity model, including sensitivities, specificities, Positive Prediction Values, Negative Prediction Values, Accuracies, and F measures for nine distinct threshold points on the ROC are shown in **Table 3.**

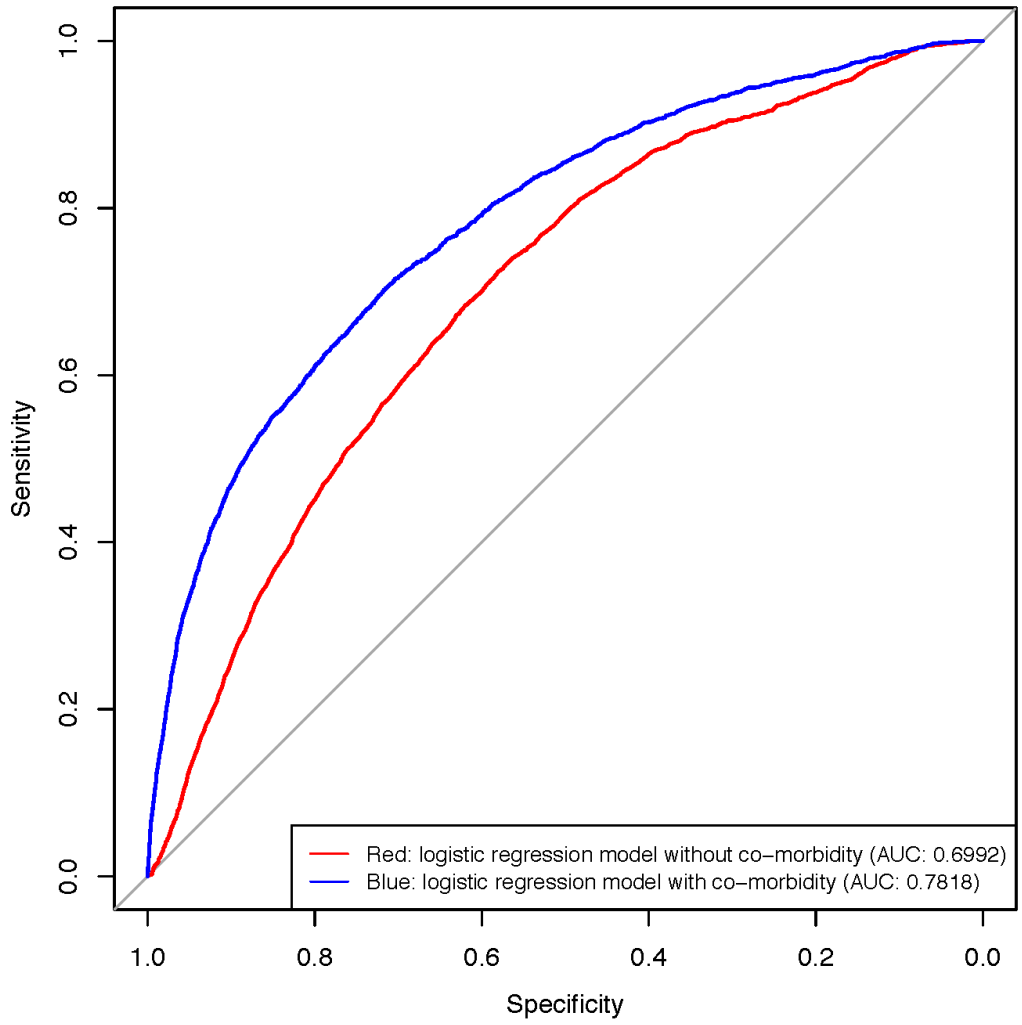

Figure 5. Receiver Operating Characteristic (ROC) curve of the traditional logistic regression model without co-morbidity analysis (red) and the final logistic regression model with co-morbidity analysis (blue) on the test data, which was unseen during the training phase. The Area Under the Curve (AUC) of the ROC increased from 0.6992 (red) to 0.7818 (blue).

Table 3. Selected performance measures of the 30-predictor co-morbidity model, including sensitivities, specificities, Positive Prediction Values (PPV), Negative Prediction Values (NPV), Accuracies, and F measures for nine distinct threshold points on the blue curve of the Receiver Operating Characteristic (ROC) shown in **Figure 5**. The performance measures were evaluated with the test data, which was unseen during the training phase.

Sensitivity	Specificity	PPV	NPV	Accuracy	F.measure
0.100	0.992	0.246	0.977	0.969	0.142
0.200	0.978	0.189	0.979	0.958	0.195
0.300	0.960	0.165	0.981	0.944	0.213
0.400	0.928	0.127	0.983	0.915	0.193
0.500	0.883	0.100	0.985	0.873	0.167
0.600	0.808	0.075	0.987	0.803	0.134
0.700	0.718	0.061	0.989	0.718	0.112
0.800	0.591	0.049	0.991	0.596	0.092
0.900	0.408	0.038	0.994	0.421	0.073

CHAPTER FIVE.

DISCUSSION

Given that depression remains significantly under-diagnosed in all settings of the healthcare system[4,5], data-driven prediction models can play an important role in the screening of depression patients. Although previous work has shown promising results on the ability to predict future diagnoses of depression, such models have not explicitly applied co-morbid medical conditions as independent predictors. Depression is a characteristic disease which can be affected by many medical condition[9], and can also affect certain medial conditions[10]. In 2013, psychological factors affecting medical conditions (PFAOMC) was included as a new diagnosis in DSM-V[15]. PFAOMC are the factors which may precipitate or exacerbate the medical condition, interfere with treatment, or contribute to morbidity and mortality. The mechanism of PFAOMC include promotion of known risk factors (i.e. smoking), influence on the underlying pathophysiology (i.e. bronchospasm in asthma), and the interference on the treatment (i.e. poor compliance). Therefore, addressing the co-morbidities related to depression will be a rationally important step in understanding the course of depression, and the analysis of these co-morbidities will likely improve the performance of depression risk prediction models.

Machine learning methods differ from traditional statistical model that the primary hypothesis is the existence of a pattern in the set of predictor variables that predicts the outcome[32]. Machine learning methods differ from traditional statistical model by having primary hypothesis of existence of a pattern in the set of predictor variables that will identify the outcome. While traditional statistical modeling

pursues a simple model that fits reasonably well, machine learning methods consider complex relationships among the variables, which is advantageous for predictive modeling. Where the outcome is related to multiple highly correlated features, simple statistical model may not work well due to the violation of the assumptions for the statistical model. Therefore, machine learning approaches were employed in this work to confirm the potential benefit of complex models for prediction of depression, especially with multiple co-morbidity features.

Rule induction methods, or symbolic methods, such as decision tree algorithms, may provide interpretability which allows justification and explanation of unexpected solution of new problems[33]. However, such methods often provide poor predictive performance compared to so-called black-box sub-symbolic methods, such as support vector machine. On the contrary, given that machine learning approaches need to be incorporated with human expert's interpretation for an integrated man-and-machine approach[34] for accuracy and liability, explanation ability and transparency[35] would be essential, and black-box nature of the machine learning system may be a critical drawback.

Results from a single conditional inference tree shown in **Figure 2** offers interpretable and transparent decision rules, but the discrete nature of classification rule does not allow flexibly in tuning the threshold. Predictive performance in the measure of AUROC was best achieved with boost trees (0.793), followed by random forest(0.739), and support vector machine (0.660), as shown in **Figure 3**. Support vector machine has great advantage in handling a large number of variables with relatively small observations of samples, by mathematical optimization based on the support vector observations lying at the class boundaries. However, in this work, the number of features, including the 60 co-morbid condition, was not so much a large number, and the data have more than 1 million of observations. Therefore, support

vector machine added not so much value in predictive modeling in this work.

Although boost trees model achieved highest predictive performance in the measure of AUROC, it also has poor interpretability and explanation ability. Although variables actually used in tree construction of the boosting algorithm may imply the variable importance, the exact role of the variable is difficult to be explained or quantified, because of its complex algorithm of building sequential models by weighing the problematic observations. Therefore, the boost trees model is also not the best model for an integrated man-and-machine approach[34] for accuracy and liability.

In order to leverage the predictive performance and interpretability, regularized regression methods were employed in this study. The AUC of the ROC increased from 0.6992 (the traditional logistic regression model without co-morbidity analysis) to 0.7818 (the final logistic regression model with co-morbidity analysis), after applying the optimized variable selection from Elastic Net (**Figure 5**). Because neither questionnaire-based screening results (i.e. Patient Health Questionnaire[36]) nor physician clinical notes are available in claims data, there is no direct information about patients' moods or symptoms. Given this limitation, this improvement could be interpreted very significant improvement, and the inclusion of co-morbidity analysis could be a key component in improving the performance of depression risk prediction models.

Furthermore, since odds ratio estimates change for some variables after adjusting for co-morbid conditions, the adjusted OR in the final logistic regression model with co-morbidity analysis could reflect estimates closer to the truth. For example, the adjusted OR of being female is 2.07 from the traditional logistic regression model without co-morbidity analysis, but is 1.63 from the final logistic regression model with co-morbidity analysis (**Figure 4**). Given that, females have a higher co-

morbidity burden in general, the traditional logistic regression model will give higher OR for females, by not adjusting for co-morbidities.

The one million twelve-year Korea National Health Insurance Service (KNHIS) longitudinal data used in this study has many advantages for analyzing large scale statistical models. As KNHIS is the only health insurance system which covers all Korean citizens, the random sample cohort from KNHIS can be considered as a nationally representative health data[37]. Factors arising from multiple health insurance systems effecting diagnosis of depression (i.e. some health insurance plans might have lower coverage for mental health) can be avoided in the single health insurance system, and therefore higher statistical power can be achieved. Therefore, adjusted ORs from the logistic regression model with co-morbidity analysis may represent the risks of each variable in the population.

Cautions are needed when interpreting the epidemiologic results from this study, however. The large sample size in this study is over-powered to detect small effects, so more emphasis should be placed on the magnitude of estimates rather than the statistical significance. Furthermore, the operational definition of depression case group is based on the diagnosis codes in the claims data. Therefore, the depression risk prediction model in this study is predicting the probabilities of each person's visiting physicians and diagnosed as depressed by physicians, and this will limit the ability of detecting the underdiagnosed depressed population. However, it is noticeable that the findings are consistent with previous studies revealed the relationship between co-morbidities and depression in Korean population with cross-sectional survey study[38], as well as Korean Longitudinal Study of Aging[39].

In order to develop a better depression risk prediction model which can also address the currently underdiagnosed depressed population, reaching out to the underdiagnosed depressed population with gold standard screening tools will be

necessary. Further work is also needed to investigate possible difference in the co-morbidity patterns in different gender, age-group, and socio-economic status. Higher prevalence of depression among female has been discussed to be related to both biological and environmental factors[40]. Features of depressions can also be vary among different age-groups[41], and certain age-groups may have additional risks[42]. Socio-economic factors[43] of depression and disparity[44] in depression treatment are also very important topic in public health.

Additional research is needed for optimizing the chronic conditions clusters, or categories. Although CMS-CCW algorithm is a well validated algorithm using ICD codes, optimized clusters developed using insurance claims data might be different when compared to actual clinical manifestation of depression. Even within the clinical practice, the disease classification or categorization can differ among various clinical specialties and subspecialties. Therefore, optimization for co-morbid conditions clusters will be needed for better prediction models[45]. Furthermore, integration of medication prescription data will allow better operational definitions with lesser false positives. Further research is also needed for variable interactions (i.e. epilepsy of young female may have different effect from epilepsy of elderly male), as well as time-to-event analysis (i.e. Cox Proportional Hazard regression[46]), dealing with time-dependent covariates.

Although the chronic co-morbid disease studied in this work was limited to those available in CMS-CCW in order to maintain consistency in the operational definition of each conditions, there are many other conditions which are related to depression, such as substance abuse[47] and abortion[48]. Furthermore, development of technologies may allow social media data[49], genome[50], exposome[51], and patient-generated health data[52] to be incorporated in the prediction model. In practical application of prediction model, thorough review of past medical history and family history,

detailed symptoms, as well as incorporation of computational time series analysis of trend and variance of laboratory tests, may allow more precise prediction.

Although the focus of this study was on the prediction of the existence of depression based on the chronic co-morbid disease appeared in the health insurance claims data, further studies will be needed to confirm if co-morbidity analysis can also improve the performance of the prediction model for treatment response[53,54], or prediction model based on lexical data[55], as well as on electronic health records[56,57].

In conclusion, the inclusion of co-morbidity analysis with Elastic Net regression model showed the performance of depression risk prediction models comparable to that of supervised machine-learning methods, with providing better interpretability. The co-morbidity adjusted ORs from the Elastic Net regression model may indicate the true independent OR of each predictor variable. Further studies will be needed to cover the currently underdiagnosed depressed population, as well as optimizing the chronic conditions clusters.

BIBLIOGRAPHY

1. Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. Arch Gen Psychiatry. 2005 Jun;62(6):593–602.

2. Pietrzak RH, Kinley J, Afifi TO, Enns MW, Fawcett J, Sareen J. Subsyndromal depression in the United States: prevalence, course, and risk for incident psychiatric outcomes. Psychol Med. 2013 Jul;43(7):1401–14.

3. Donohue JM, Pincus HA. Reducing the societal burden of depression. Pharmacoeconomics. 2007;25(1):7–24.

4. Mitchell AJ, Vaze A, Rao S. Clinical diagnosis of depression in primary care: a meta-analysis. Lancet Lond Engl. 2009 Aug 22;374(9690):609–19.

5. Tylee A, Gandhi P. The importance of somatic symptoms in depression in primary care. Prim Care Companion J Clin Psychiatry. 2005;7(4):167–76.

6. Bruce ML, Raue PJ, Reilly CF, Greenberg RL, Meyers BS, Banerjee S, et al. Clinical effectiveness of integrating depression care management into medicare home health: the Depression CAREPATH Randomized trial. JAMA Intern Med. 2015 Jan;175(1):55–64.

7. Huang SH, LePendu P, Iyer SV, Tai-Seale M, Carrell D, Shah NH. Toward personalizing treatment for depression: predicting diagnosis and severity. J Am Med Inform Assoc. 2014;21(6):1069–1075.

8. Jin H, Wu S, Di Capua P. Development of a Clinical Forecasting Model to Predict Comorbid Depression Among Diabetes Patients and an Application in Depression Screening Policy Making. Prev Chronic Dis. 2015;12:E142.

9. Hirschfeld RMA. The Comorbidity of Major Depression and Anxiety Disorders: Recognition and Management in Primary Care. Prim Care Companion J Clin Psychiatry. 2001 Dec;3(6):244–54.

10. Fava GA, Fabbri S, Sirri L, Wise TN. Psychological factors affecting medical condition: a new proposal for DSM-V. Psychosomatics. 2007 Apr;48(2):103–11.

11. Kim L, Kim J, Kim S. A guide for the utilization of Health Insurance Review and Assessment Service National Patient Samples. Epidemiol Health. 2014;36:e2014008.

12. Lee J, Lee JS, Park S-H, Shin SA, Kim K. Cohort Profile: The National Health Insurance Service–National Sample Cohort (NHIS-NSC), South Korea. Int J Epidemiol. 2016 Jan 28;dyv319.

13. Cochran WG. Sampling techniques-3. 1977 [cited 2016 Jul 3]; Available from: http://agris.fao.org/agris-search/search.do?recordID=XF2015028634

14. Gorina Y, Kramarow EA. Identifying Chronic Conditions in Medicare Claims

Data: Evaluating the Chronic Condition Data Warehouse Algorithm. Health Serv Res. 2011 Oct 1;46(5):1610–27.

15. Association AP, others. Diagnostic and statistical manual of mental disorders (DSM-5®). American Psychiatric Pub; 2013.

16. Kuhn M, Johnson K. Applied predictive modeling. Springer; 2013.

17. Hothorn T, Hornik K, Zeileis A. ctree: Conditional Inference Trees. [cited 2016 Jul 16]; Available from: http://cran.nexr.com/web/packages/partykit/vignettes/ctree.pdf

18. Williams G. Data mining with rattle and R: the art of excavating data for knowledge discovery. Springer Science & Business Media; 2011.

19. Culp M, Johnson K, Michailidis G. ada: An r package for stochastic boosting. J Stat Softw. 2006;17(2):9.

20. Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002;2(3):18–22.

21. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. Springer; 2013.

22. Karatzoglou A, Smola A, Hornik K, Zeileis A. kernlab-an S4 package for kernel methods in R. 2004 [cited 2016 Jul 23]; Available from: http://epub.wu.ac.at/1048/

23. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol. 2005;67(2):301–320.

24. Rencher AC, Pun FC. Inflation of R2 in best subset regression. Technometrics. 1980;22(1):49–53.

25. Wilkinson L, Dallal GE. Tests of significance in forward selection regression with an F-to-enter stopping rule. Technometrics. 1981;23(4):377–380.

26. Hurvich CM, Tsai C-L. The impact of model selection on inference in linear regression. Am Stat. 1990;44(3):214–217.

27. Tibshirani R. Regression Shrinkage and Selection via the Lasso. J R Stat Soc Ser B Methodol. 1996;58(1):267–88.

28. Zucchini W. An Introduction to Model Selection. J Math Psychol. 2000 Mar;44(1):41–61.

29. Team RC. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2013. Doc Free Available Internet Httpwww R-Proj Org. 2015;

30. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1.

31. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC:

an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12(1):1.

32. Waljee AK, Higgins PD. Machine learning in medicine: a primer for physicians. Am J Gastroenterol. 2010;105(6):1224.

33. Lavrač N. Machine learning for data mining in medicine. In: Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making [Internet]. Springer; 1999 [cited 2016 Jul 23]. p. 47–62. Available from: http://link.springer.com/chapter/10.1007/3-540-48720-4_4

34. Deo RC. Machine learning in medicine. Circulation. 2015;132(20):1920–1930.

35. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. Artif Intell Med. 2001;23(1):89–109.

36. Kroenke K, Spitzer RL, Williams JB. The Phq-9. J Gen Intern Med. 2001;16(9):606–613.

37. Park SM, Son KY, Park J-H, Cho B. Disparities in short-term and long-term all-cause mortality among Korean cancer patients with and without preexisting disabilities: a nationwide retrospective cohort study. Support Care Cancer Off J Multinatl Assoc Support Care Cancer. 2012 May;20(5):963–70.

38. Yun YH, Kim SH, Lee KM, Park SM, Kim YM. Age, sex, and comorbidities were considered in comparing reference data for health-related quality of life in the general and cancer populations. J Clin Epidemiol. 2007 Nov;60(11):1164–75.

39. Kim H, Park S-M, Jang S-N, Kwon S. Depressive symptoms, chronic medical illness, and health care utilization: findings from the Korean Longitudinal Study of Ageing (KLoSA). Int Psychogeriatr IPA. 2011 Oct;23(8):1285–93.

40. Kessler RC. Epidemiology of women and depression. J Affect Disord. 2003 Mar;74(1):5–13.

41. Benazzi F. Female Depression before and after Menopause. Psychother Psychosom. 2000;69(5):280–3.

42. Greenfield A, Banerjee S, DePasquale A, Weiss N, Sirey J. Factors Associated with Nutritional Risk Among Homebound Older Adults with Depressive Symptoms. J Frailty Aging. (In Press);

43. Lorant V, Croux C, Weich S, Deliège D, Mackenbach J, Ansseau M. Depression and socio-economic risk factors: 7-year longitudinal population study. Br J Psychiatry. 2007 Apr 1;190(4):293–8.

44. Alegría M, Chatterji P, Wells K, Cao Z, Chen C, Takeuchi D, et al. Disparity in Depression Treatment Among Racial and Ethnic Minority Populations in the United States. Psychiatr Serv. 2008 Nov 1;59(11):1264–72.

45. Pathak J, Wang J, Kashyap S, Basford M, Li R, Masys DR, et al. Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience. J Am Med Inform

Assoc. 2011;18(4):376–386.

46. Lin DY, Wei L-J. The robust inference for the Cox proportional hazards model. J Am Stat Assoc. 1989;84(408):1074–1078.

47. Ialongo NS, Werthamer L, Kellam SG, Brown CH, Wang S, Lin Y. Proximal impact of two first-grade preventive interventions on the early risk behaviors for later substance abuse, depression, and antisocial behavior. Am J Community Psychol. 1999;27(5):599–641.

48. Cougle JR, Reardon DC, Coleman PK. Depression associated with abortion and childbirth: a long-term analysis of the NLSY cohort. Med Sci Monit. 2003;9(4):CR105–CR112.

49. De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting Depression via Social Media. In: ICWSM [Internet]. 2013 [cited 2016 Jul 23]. p. 2. Available from: http://course.duruofei.com/wp-content/uploads/2015/05/Choudhury_Predicting-Depression-via-Social-Media_ICWSM13.pdf

50. Lewis CM, Ng MY, Butler AW, Cohen-Woods S, Uher R, Pirlo K, et al. Genome-wide association study of major recurrent depression in the UK population. Am J Psychiatry. 2010;167(8):949–957.

51. Martin-Sanchez F, Verspoor K, others. Big data in medicine is driving big changes. Yearb Med Inform. 2014;9(1):14–20.

52. Shapiro M, Johnston D, Wald J, Mon D. Patient-generated health data. White Pap Prep Off Policy Plan [Internet]. 2012 [cited 2016 Jul 23]; Available from: http://healthitgov.ahrqdev.org/sites/default/files/rti_pghd_whitepaper_april_2012.pdf

53. Gallagher PJ, Castro V, Fava M, Weilburg JB, Murphy SN, Gainer VS, et al. Antidepressant response in patients with major depression exposed to NSAIDs: a pharmacovigilance study. Am J Psychiatry. 2012 Oct;169(10):1065–72.

54. Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. Lancet Psychiatry. 2016;

55. Banitaan S, Daimi K. Using Data Mining to Predict Possible Future Depression Cases. Int J Public Health Sci IJPHS. 2014;3(4):231–240.

56. Pathak J, Simon G, Li D, Biernacka JM, Jenkins GJ, Chute CG, et al. Detecting Associations between Major Depressive Disorder Treatment and Essential Hypertension using Electronic Health Records. AMIA Jt Summits Transl Sci Proc AMIA Summit Transl Sci. 2014;2014:91–6.

57. Bobo WV, Pathak J, Kremers HM, Yawn BP, Brue SM, Stoppel CJ, et al. An electronic health record driven algorithm to identify incident antidepressant medication users. J Am Med Inform Assoc JAMIA. 2014 Oct;21(5):785–91.