# Prediction of depression onset risk among middle-aged and elderly adults using machine learning and Canadian Longitudinal Study on Aging cohort

Yipeng Song[a], Lei Qian[a], Jie Sui[b], Russell Greiner[a c d], Xin-min Li[a], Andrew J. Greenshaw[a], Yang S. Liu[a], Bo Cao[a,* c]

[a]Department of Psychiatry, University of Alberta, Edmonton Alberta, Canada

[b]School of Psychology, University of Aberdeen, Aberdeen, UK

[c]Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada

[d]Alberta Machine Intelligence Institute, Edmonton, Alberta, Canada

## Author Note

Yipeng Song https://orcid.org/0000-0002-7180-2647

Lei Qian https://orcid.org/0000-0001-7608-2741

Yang S. Liu https://orcid.org/0000-0003-0406-8056

Bo Cao https://orcid.org/0000-0001-9338-3271

We have no conflicts of interests to disclose.

Correspondence concerning this article should be addressed to Bo Cao, Department of Psychiatry, University of Alberta, Edmonton, Alberta, Canada (email: cloudbocao@gmail.com)

## Abstract

**Background** Early identification of the middle-aged and elderly subjects with high risk of developing depression disorder in the future and the full characterization of the associated risk factors are crucial for early interventions to prevent depression among the aging population.

**Methods** Canadian Longitudinal Study on Aging (CLSA) has collected comprehensive information, including psychological scales and other non-psychological measures, i.e., socioeconomic, environmental, health, lifestyle, cognitive function, personality, about its participants (30,097 subjects aged from 45 to 85) at baseline phase in 2012-2015. We applied machine learning models for the prediction of these participants' risk of depression onset approximately three years later using information collected at baseline phase.

**Results** Individual level risk for future depression onset among CLSA participants can be accurately predicted, test set area under receiver operating characteristic curve (AUC) 0.791 ± 0.016, using all baseline information. We also found the 10-item Center for Epidemiological Studies Depression Scale coupled with age and sex information can achieve similar performance (test set AUC 0.764 ± 0.016). Furthermore, we identified existing subthreshold depression symptoms, emotional instability, low levels of life satisfaction, perceived health, and social support, and nutrition risk as the most important factors independently from psychological scales for predicting depression onset.

**Limitations** The label of depression disorder was based on self-reported doctor diagnosis and depression screening tool, thus it is not a clinical diagnosis.

**Conclusions** The identified risk factors will further improve our understanding of the depression onset among middle-aged and elderly population and the early identification of high-risk subjects is the first step for successful early interventions.

## Introduction

Depression, or major depressive disorder (MDD), is a mental disorder characterized by a persistent depressed mood, low self-esteem, loss of interest and vegetative symptoms (Cooper, 2017). Depression is estimated as the second largest contributor to the disease burden characterized by disability adjusted life year in both developing and developed countries (Vos et al., 2015). It was suggested that the prevalence of depressive disorder is 6.6% among community-dwelling middle-aged and older adults (Mojtabai & Olfson, 2004). With age, the prevalence of depressive disorder declines, however, increased likelihood of significant depressive symptoms was observed (Mojtabai & Olfson, 2004). Also, compared to younger adults, depressive disorder in older adults manifests some different symptoms, e.g., less affective symptoms and more cognitive impairment, and is potential under-diagnosed and under-treated (Kok & Reynolds, 2017; Pocklington, 2017). The early identification of the middle-aged and elderly subjects with high risk of developing depressive disorder in the future and the full characterization of the associated risk factors are crucial for early interventions to prevent depression among the aging population.

Many studies have investigated the risk factors associated with pre-existing depression in middle-aged and/or elderly adults (Fiske et al., 2009). The identified risk factors include genetic factors, chronic conditions (e.g., cardiovascular disease, diabetes, dementia), sleep disturbance, stressful life events, and many others (Fiske et al., 2009; Mojtabai & Olfson, 2004). However, these identified risk factors don't inform an individual's risk in developing depression in the future. And it is unknown whether or not these identified risk factors will be able to predict future depression onset in the middle-aged and/or elderly population. Besides these association analyses, machine learning (ML) algorithms have been applied to predict individuals' risk for

depression onset in different populations. Su et al. (2021) applied a recurrent neural network on longitudinal data collected from 1,077 elderly subjects (83.73 ± 5.71 years old) to predict the depression onset in the next two years using demographics, health-related and chronic diseases as predictors (Su et al., 2021). The developed model achieved an area under receiver operating characteristic curve (AUC), a metric used to evaluate the model's capacity in classifying depressed subjects from healthy controls, of 0.629 on the hold-out test set (Su et al., 2021). Librenza-Garcia et al. (2021) applied a ML approach on an occupational cohort for the prediction of depression onset using sociodemographic and clinical factors on 13,922 subjects (51.83 ±8.98 years old). They achieved a test set AUC of 0.71 (0.66-0.77) and identified comorbid obsessive-compulsive disorder and anxiety disorders, certain medications and sex as the most important factors in the prediction. In addition, Laura Sampson et al. (2021) applied a ML approach to the prediction of depression onset in a military cohort (1951 men and 298 women, and over 90% of the participants are less than 45 years old) and achieved a cross-validated AUC of 0.67 using survey data and identified that having post-traumatic stress disorder, being mistreated and having financial problems as the most important factors in the prediction (Sampson et al., 2021). However, the occupational and military cohorts are different from the middle-aged and elderly populations. For example, the occupational cohort has lower mortality and morbidity rates compared to the general population, known as the healthy worker effect (Li & Sung, 1999). Therefore, the findings based on occupational and military cohorts may not be directly transferable to the depression onset prediction in the middle-aged and elderly population.

In the current study, we aimed to develop a ML model with high performance in predicting the onset of depressive disorder approximately three years in the future among the middle-aged

and elderly population using a large and nationwide aging population cohort in Canada, the Canadian Longitudinal Study on Aging (CLSA) (Raina et al., 2009). CLSA comprehensive cohort contains 30,097 subjects aged from 45 to 85 years with a mean age of 63. 63. Details about the cohorts in CLSA data will be shown later. The developed model will help to identify the subjects with a high risk of developing depression in the future for potential early interventions. Moreover, CLSA has collected comprehensive information from the participants, including psychological scales, i.e., 10-item Center for Epidemiological Studies Depression Scale (CES-D-10) and Kessler Psychological Distress Scale (K10), and other non-psychological measures, i.e., socioeconomic, environmental, health, lifestyle, cognitive function and personality. As psychological scales are commonly used to evaluate depression symptoms, we will mainly identify novel risk and protective factors from the non-psychological measures for informing depression onset among middle-aged and elderly adults.

## Methods

### Data

Data collected in the CLSA project was used in this research. CLSA is a nationwide longitudinal study on aging and adult development in Canada (Raina et al., 2009). National stratified random sample of 51,338 Canadians aged from 45 to 85 years old at the baseline were recruited in the CLSA project. The study excluded the individuals living in long-term care facilities, the individuals with cognitive impairment or those who are unable to respond in English or French. The data is collected every three years, and it is expected to follow the participants for at least 30 years. The comprehensive track of the CLSA project was used in this study, which contains 30,097 subjects who live within 25-50 km of 11 data collection centers

across several provinces in Canada. The baseline data set, collected between 2011 to 2015, and the depression onset information, collected between 2015 to 2018, are used in the data analysis.

**Outcome**

The outcome of interest is the onset of depression at the follow-up phase among the CLSA participants without depression at the baseline phase. In both baseline and follow-up phases, CLSA participants were interviewed with the question "Has a doctor ever told you that you suffer from clinical depression?". Clinical depression is the alias for MDD (Kupfer et al., 2012). Moreover, in both baseline and follow-up phases, the CES-D-10 scale, which has shown to be an excellent screening instrument in identifying MDD in elderly (Irwin et al., 1999) was used to screen the depressive disorder among CLSA participants. The total score of CES-D-10 ranges from zero to 30, and a score of ten or higher indicates the presence of significant depressive symptoms (Irwin et al., 1999). In addition, whether or not taking medication for depressive disorder was also recorded in CLSA data. Self-reported diagnosis of mood disorder, which is related to the questionnaire "Has a doctor ever told you that you have a mood disorder such as depression (including manic depression), bipolar disorder, mania, or dysthymia?", is also available at both baseline and follow-up stages. Therefore, we excluded all the subjects who were not followed in the follow-up phase and might have depressive disorder at the baseline phase, including self-reported clinical depression diagnosis, positive depression screening results using CES-D-10 scales, currently taking medication for depression, self-reported mood disorder conditions, from the data analysis. After that, we labeled subjects who had either self-reported clinical depression diagnosis or had positive depression screening results according to the CES-D-10 scales or currently taking medication for depression at the follow-up stage to have depression onset.

**Predictors**

The baseline CLSA data contains 2529 variables from different sources (Raina et al., 2009). The variables contain psychological scales, i.e., CES-D-10 scale and K10 Scale (Kessler et al., 2002), and non-psychological measures, i.e., socioeconomic, environmental, health, lifestyle, cognitive function and personality. Distinct from the CES-D-10 scale, which was specifically designed to identify depression and depressive symptoms, the K10 scale is composed of 10 items, each assessed on a 5-point Likert scale, to measure non-specific psychological distress. Its usefulness in screening for depression and anxiety disorders has been validated through multiple studies (Donker et al., 2010; Tran et al., 2019). As we are interested into the effect of biological aging process on the incidence of depressive disorder, we also included several biological aging indices, i.e., frailty index (Pérez-Zepeda et al., 2021), methylation age (Horvath, 2013), and laboratory age (Mamoshina et al., 2019), as non-psychological information for the prediction. We will first examine the utility of non-psychological information, i.e., socioeconomic, environmental, health, lifestyle, cognitive function and personality in predicting the incidence of depression. Then risk and protective factors will be derived from this model on non-psychological information to inform the incidence of depression. Furthermore, we will examine the utility of psychological scales, which are commonly used to measure the depression symptoms (Irwin et al., 1999; Vasiliadis et al., 2015), in predicting the incidence of depression among CLSA participants.

Before casting into the ML algorithms, the predictors were preprocessed as follows. Missing values in the categorical variables were taken as a new level, then categorical variables were either dummy coded or ordinal coded depending on the used ML algorithms; columns with zero

variance, or with over 30% missing values, or with free text data were dropped; missing values in the numerical variables were inputted by the median value of the corresponding column.

**Model development, selection and comparison**

Multiple ML algorithms, including standard random forest model and its extension for imbalanced classification problem, and logistic regression with regularization, were used to make the prediction of depression onset. Details of these algorithms and the justification for use these methods are described in the supplementary material. The data set was first split into training (80%) and test (20%) sets using stratified random sampling. Then ML algorithms were selected and fitted on the training set. Details about the model selection can be found in the supplementary material. After the models were selected and fitted, the selected models will be further evaluated with respect to the different metrics, i.e., AUC (the model's capacity in distinguish depressed subjects from non-depressed subjects), sensitivity (the probability that depressed subjects are correctly predicted by the model), specificity (the probability that non-depressed subjects are correctly predicted by the model), balanced accuracy (the mean accuracy of the model in predicting both depressed and non-depressed subjects), on the test set. To evaluate the uncertainty of these model performance metrics, we repeated the above steps for 10 times.

**Variable importance**

In the nonlinear models used in this study, the contribution of a feature to the prediction is evaluated by the Shapley values based variable importance measure (Lundberg & Lee, 2017). Previous research has shown Shapley values-based feature importance have nice properties, i.e., efficiency, symmetry, dummy and additivity, compared permutation based and impurity-based feature importance measures (Lundberg & Lee, 2017). Detailed explanation can be found in

Lundberg et. al (2017). In the linear models used in this study, the contribution of a feature to the prediction is evaluated by the absolute of the coefficient. As almost all variables will be attributed with a nonzero variable importance, it is impractical to interpret all of them. We tried to identify important variables for making predictions in the selected model through post hoc analysis. In the post hoc analysis, important variables were defined as the variables with higher variable importance values compared to the associated null variable importance, which is the variable importance measured on the same model fitted on permutated outcome labels (see supplementary material for additional details).

## Results

Among the 19,024 participants without shown evidence of depressive disorder at baseline and have been followed in the follow up stage, 1,329 subjects have potential depression onset in the follow up stage. The mean age (SD) for these subjects with potential depression onset (depression onset group) was 62.8 (10.2) years and 55.38% of them were female. The mean age (SD) for those without depression onset was 64.4 (11.0) and 45.08% of them were female. The mean total score (SD) of the baseline CES-D-10 scale was 3.30 (2.46) for subjects without depression onset, and 5.46 (2.52) for the subjects in depression onset group.

### Depression onset prediction with non-psychological measures

We examined the utility of non-psychological CLSA baseline predictors in predicting the depression onset using multiple ML algorithms (Figure S1 left). The best performed model achieved a test set AUC of 0.743 ± 0.013 in predicting the outcome. The selected model achieved a test set balanced accuracy of 0.678 ± 0.012, sensitivity of 0.694 ± 0.032, and specificity of 0.663 ± 0.016. Post hoc important variable analysis identified 183 out of 2018 variables to be important for the prediction. The top 20 variables with highest variable

importance in the model are shown in Figure 1. Emotional stability scale, satisfaction with life scale, functional social support, self-rated mental health, and nutritional risk score are among the predictors with highest contributions in predicting depression onset among CLSA participants. How these top predictors effect the estimated risk of depression onset are shown in Figure S2 A, B.

**The utility of psychological scales in predicting depression onset among CLSA participants**

We examined the utility of the psychological scales in predicting the onset of depression among CLSA participants using multiple ML algorithms (Figure S1 center). In addition, as age and sex are always available in real applications and their interactions with these psychological scales may be important for the prediction, they are also included as predictors. K10 variables achieved a test set AUC 0.735 ± 0.013 in predicting the depression onset; CES-D-10 variables, AUC 0.764 ± 0.016; and the combination of CES-D-10 and K10 scales, AUC 0.776 ± 0.010. Other metrics of the selected models are shown in the supplementary material. Post hoc important variable analysis identified all the predictors to be important for the prediction. As shown in Figure 2, total scores of CES-D-10 and K10 scales, sex and some CES-D-10 items have the most important contributions to the prediction of depression onset. How the total scores of CES-D-10 and K10 scales effect the predicted risk of depression onset are shown in Figure S3.

Furthermore, we examined how non-psychological information further improves the prediction of depression onset on top of psychological scales using all the baseline variables as predictors. The models using all available baseline features were selected and evaluated as described above. The best performing model achieves a test set AUC 0.791 ± 0.016 in predicting

the onset of depression among CLSA participants. Other metrics of the selected model are balance accuracy, 0.720 ± 0.014; sensitivity, 0.750 ± 0.031; specificity, 0.689 ± 0.010.

## Discussion

In this study, we showed that individual level risk for developing depressive disorder in the middle-aged and elderly population can be accurately predicted (best model has a test set AUC 0.791±0.016) by using machine learning algorithms and comprehensive survey information collected approximately three years before. The developed model has higher performance compared to previous studies on predicting depression onset in the elderly population (test set AUC 0.629) (Su et al., 2021), in the military population (test set AUC 0.671) (Sampson et al., 2021) or in the occupational cohort (test set AUC 0.71) (Librenza-Garcia et al., 2021). Additionally, we identified the top factors, personality trait emotional stability, life satisfaction, perceived health, social support and nutrition risk factors, important for predicting depression onset in the middle-aged and elderly population from the comprehensive survey predictors.

Previous research observed that depressed patients show a lower level of emotional stability compared to healthy controls (Thompson et al., 2012). Our results further indicates that higher emotional instability leads to higher risk of depression onset in about three years later among the middle-aged and elderly population. Furthermore, our results also indicate that a higher level of satisfaction with life leads to a lower risk of depression onset. This fining is largely consistent with previous study shows that life satisfaction level is associated with existing depression diagnosis (Gigantesco et al., 2019). In addition, we identified poor self-rated and perceived mental health are related to high risk of depression onset. The finding is consistent with previous studies that poor self-rated mental health is associated with prolonged antidepressant therapy among patients with depression, and with the utilization of mental health services (Fleishman &

Zuvekas, 2007). Our results highlighted the importance of subjective well-being, life satisfaction, self-rated and perceived mental health in predicting the incidence of depression in the middle-aged and elderly population. Previous studies have also identified social support as a predictive factor for depression (Gariépy et al., 2016). Our results confirmed that more functional social support is related to a lower risk of depression onset among middle-aged and elderly population. Previous research on the association between malnutrition and depression is controversial. Some reported that there is no association between malnutrition and depression status among elderly without heart diseases (Daniel et al., 2021; Toffanello et al., 2014), while others reported that elderly with malnutrition are associated with higher risk of depression (Davison et al., 2019; Maier et al., 2021). Our result supports that high nutritional risk (Screen-8 AB risk score less than 38) is related to higher risk of depression onset in about three years later among the middle-aged and elderly population.

The CLSA participants with depression onset (baseline CES-D-10 score 5.46 ± 2.52) has already shown a higher degree of depressive symptoms at baseline phase compared to the peers without depression onset (baseline CES-D-10 score is 3.30 ± 2.46). Our results indicate that even when the score of the CES-D-10 scale of the CLSA participants doesn't reach the criteria to be screened as depression, a high score of CES-D-10 scale is still related to high risk of developing depression in the future. And the higher the score of CES-D-10 scale, the higher chance the subject being predicted to have depression onset. Furthermore, a high level of psychological distress measured by K10 scale indicates a high risk of depression onset in about three years later. And the higher the score of K10 scale, the higher chance the subject being predicted to have depression onset. When coupled with age and sex variables, the CES-D-10 scale produced a test set AUC (Area Under the Curve) of 0.764 ± 0.016. This performance is comparable to the model that incorporated

all available predictors in anticipating the onset of depression among CLSA participants. The introduction of information from the K10 scale into the model, in conjunction with the CES-D-10 scale, resulted in only a modest increase of 0.012 in the test set AUC. This implies that most of the pertinent information necessary for predicting depression onset within the K10 scale is already encompassed within the CES-D-10 scale. As the CES-D-10 scale only contains 10 items in the questionnaire, the developed model has a great potential in real world applications to inform middle-aged and elderly adults' risk in developing depression in the future. Also, these results indicate it may benefit early intervention on the subjects with subthreshold depression symptoms to prevent the onset of depression in the future.

One of the limitations of the current study is that the definition of depression outcome relies on self-reported diagnosis and positive depression screening using CES-D-10 scale. Self-reported diagnosis is not as accurate as the depression diagnosis records from clinical psychologist and psychiatrist. In addition, even though CES-D-10 is a widely used as a depression screening tool in epidemiological studies and has been proved to have good consistency with medical diagnosis in hospital, the depression label derived from CES-D-10 is not a clinical diagnosis. Furthermore, the depressive symptoms measured by CES-D-10 may be presented by bipolar disorder or minor depressive disorder. Another limitation is that, even though the developed model was fully validated on the hold out test sets, its utility in informing the individual level risk in developing depression in the future cannot be tested now. We will update the validation results using the CLSA data collected in 2018-2021 when it is available to us. Despite these limitations, the results in this study are still valuable in identifying high risk middle-aged and elderly adults in developing depressive disorder and studying the risk factors for depression onset among middle-aged and elderly population.

## Conclusions

In summary, we were able to predict the individual level's risk for developing depressive disorder among a large middle-aged and elderly population using machine learning algorithms. The individual level risk prediction provides the opportunity in identifying the subjects with high risk in developing depressive disorder in the future, which is the basis for successful early interventions. The identified risk factors will future improve our understanding of the depressive disorder onset among middle-aged and elderly population. However, it is important to note that further research and validation are necessary before the direct application of our findings in clinical practice can be recommended.

## Acknowledgements

This research was made possible using the data/biospecimens collected by the Canadian Longitudinal Study on Aging (CLSA). Funding for the Canadian Longitudinal Study on Aging (CLSA) is provided by the Government of Canada through the Canadian Institutes of Health Research (CIHR) under grant reference: LSA 94473 and the Canada Foundation for Innovation, as well as the following provinces, Newfoundland, Nova Scotia, Quebec, Ontario, Manitoba, Alberta, and British Columbia. This research has been conducted using the CLSA dataset (Comprehensive Cohort), under Application Number 1906013. The CLSA is led by Drs. Parminder Raina, Christina Wolfson and Susan Kirkland.

## Funding

This research was undertaken, in part, thanks to funding from the Canada Research Chairs program, Alberta Innovates, Mental Health Foundation, MITACS Accelerate program, Simon & Martina Sochatsky Fund for Mental Health, the Alberta Synergies in Alzheimer's and Related Disorders (SynAD) program and University of Alberta Hospital Foundation.

## Author contributions

YP participated in conceptualization, formal analysis, writing the original draft and draft review. YSL verified the underlying data. BC participated in conceptualization, funding acquisition, supervision and draft review & editing. All authors participated in the results interpretation, draft review & editing.

## Declaration of interests

We declare no conflicting interest.

## Data Availability Statement

Data are available from the Canadian Longitudinal Study on Aging (www.clsa-elcv.ca) for researchers who meet the criteria for access to de-identified CLSA data. To learn more about the accessibility of CLSA data sets, see https://www.clsa-elcv.ca/data-access.

## Reference


Cooper, R. (2017). Diagnostic and statistical manual of mental disorders (DSM). In *Knowledge Organization* (Vol. 44, Issue 8). https://doi.org/10.5771/0943-7444-2017-8-668

Daniel, S. C., Azuero, A., Gutierrez, O. M., & Heaton, K. (2021). Examining the relationship between nutrition, quality of life, and depression in hemodialysis patients. *Quality of Life Research*, *30*(3). https://doi.org/10.1007/s11136-020-02684-2

Davison, K. M., Lung, Y., Lin, S., Tong, H., Kobayashi, K. M., & Fuller-Thomson, E. (2019). Depression in middle and older adulthood: The role of immigration, nutrition, and other determinants of health in the Canadian longitudinal study on aging. *BMC Psychiatry*, *19*(1). https://doi.org/10.1186/s12888-019-2309-y

Donker, T., Comijs, H., Cuijpers, P., Terluin, B., Nolen, W., Zitman, F., & Penninx, B. (2010). The validity of the Dutch K10 and extended K10 screening scales for depressive and anxiety disorders. *Psychiatry Research*, *176*(1). https://doi.org/10.1016/j.psychres.2009.01.012

Fiske, A., Wetherell, J. L., & Gatz, M. (2009). Depression in older adults. In *Annual Review of Clinical Psychology* (Vol. 5). https://doi.org/10.1146/annurev.clinpsy.032408.153621

Fleishman, J. A., & Zuvekas, S. H. (2007). Global self-rated mental health: Associations with other mental health measures and with role functioning. *Medical Care*, *45*(7). https://doi.org/10.1097/MLR.0b013e31803bb4b0

Gariépy, G., Honkaniemi, H., & Quesnel-Vallée, A. (2016). Social support and protection from depression: Systematic review of current findings in western countries. In *British Journal of Psychiatry* (Vol. 209, Issue 4). https://doi.org/10.1192/bjp.bp.115.169094

Gigantesco, A., Fagnani, C., Toccaceli, V., Stazi, M. A., Lucidi, F., Violani, C., & Picardi, A. (2019). The relationship between satisfaction with life and depression symptoms by gender. *Frontiers in Psychiatry*, *10*(JUN). https://doi.org/10.3389/fpsyt.2019.00419

Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biology*, *14*(10). https://doi.org/10.1186/gb-2013-14-10-r115

Irwin, M., Artin, K. H., & Oxman, M. N. (1999). Screening for depression in the older adult: Criterion validity of the 10-item Center for Epidemiological Studies Depression Scale (CES-D). *Archives of Internal Medicine*, *159*(15). https://doi.org/10.1001/archinte.159.15.1701

Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S. L. T., Walters, E. E., & Zaslavsky, A. M. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological Medicine*, *32*(6). https://doi.org/10.1017/S0033291702006074

Kok, R. M., & Reynolds, C. F. (2017). Management of depression in older adults: A review. In *JAMA - Journal of the American Medical Association* (Vol. 317, Issue 20). https://doi.org/10.1001/jama.2017.5706

Kupfer, D. J., Frank, E., & Phillips, M. L. (2012). Major depressive disorder: New clinical, neurobiological, and treatment perspectives. In *The Lancet* (Vol. 379, Issue 9820). https://doi.org/10.1016/S0140-6736(11)60602-8

Li, C. Y., & Sung, F. C. (1999). A review of the healthy worker effect in occupational epidemiology. *Occupational Medicine*, *49*(4). https://doi.org/10.1093/occmed/49.4.225

Librenza-Garcia, Di., Passos, I. C., Feiten, J. G., Lotufo, P. A., Goulart, A. C., de Souza Santos, I., Viana, M. C., Benseñor, I. M., & Brunoni, A. R. (2021). Prediction of depression cases, incidence, and chronicity in a large occupational cohort using machine learning techniques: An analysis of the ELSA-Brasil study. *Psychological Medicine*, *51*(16). https://doi.org/10.1017/S0033291720001579

Lundberg, S. M., & Lee, S. I. (2017). Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems. pp. 4765–4774 (2017). *NIPS-2017 Advances in Neural Information Processing Systems*, *32*(2).

Maier, A., Riedel-Heller, S. G., Pabst, A., & Luppa, M. (2021). Risk factors and protective factors of depression in older people 65+. A systematic review. *PLoS ONE*, *16*(5 May). https://doi.org/10.1371/journal.pone.0251326

Mamoshina, P., Kochetov, K., Cortese, F., Kovalchuk, A., Aliper, A., Putin, E., Scheibye-Knudsen, M., Cantor, C. R., Skjodt, N. M., Kovalchuk, O., & Zhavoronkov, A. (2019). Blood biochemistry analysis to detect smoking status and quantify accelerated aging in smokers. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-018-35704-w

Mojtabai, R., & Olfson, M. (2004). Major depression in community-dwelling middle-aged and older adults: Prevalence and 2- and 4-year follow-up symptoms. *Psychological Medicine*, *34*(4). https://doi.org/10.1017/S0033291703001764

Pérez-Zepeda, M. U., Godin, J., Armstrong, J. J., Andrew, M. K., Mitnitski, A., Kirkland, S., Rockwood, K., & Theou, O. (2021). Frailty among middle-aged and older Canadians: Population norms for the frailty index using the Canadian Longitudinal Study on Aging. *Age and Ageing*, *50*(2). https://doi.org/10.1093/ageing/afaa144

Pocklington, C. (2017). Depression in older adults. *British Journal of Medical Practitioners*, *10*(1). https://doi.org/10.1177/17557380211052072

Raina, P. S., Wolfson, C., Kirkland, S. A., Griffith, L. E., Oremus, M., Patterson, C., Tuokko, H., Penning, M., Balion, C. M., Hogan, D., Wister, A., Payette, H., Shannon, H., & Brazil, K. (2009). The Canadian Longitudinal Study on Aging (CLSA). *Canadian Journal on Aging*, *28*(3). https://doi.org/10.1017/S0714980809990055

Sampson, L., Jiang, T., Gradus, J. L., Cabral, H. J., Rosellini, A. J., Calabrese, J. R., Cohen, G. H., Fink, D. S., King, A. P., Liberzon, I., & Galea, S. (2021). A Machine Learning Approach to Predicting New-onset Depression in a Military Population. *Psychiatric Research and Clinical Practice*, *3*(3). https://doi.org/10.1176/appi.prcp.20200031

Su, D., Zhang, X., He, K., & Chen, Y. (2021). Use of machine learning approach to predict depression in the elderly in China: A longitudinal study. *Journal of Affective Disorders*, *282*. https://doi.org/10.1016/j.jad.2020.12.160

Thompson, R. J., Mata, J., Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Gotlib, I. H. (2012). The everyday emotional experience of adults with major depressive disorder: Examining emotional instability, inertia, and reactivity. *Journal of Abnormal Psychology*, *121*(4). https://doi.org/10.1037/a0027978

Toffanello, E. D., Sergi, G., Veronese, N., Perissinotto, E., Zambon, S., Coin, A., Sartori, L., Musacchio, E., Corti, M. C., Baggio, G., Crepaldi, G., & Manzato, E. (2014). Serum 25-hydroxyvitamin D and the onset of late-life depressive mood in older men and women: The Pro.V.A. study. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, *69*(12). https://doi.org/10.1093/gerona/glu081

Tran, T. D., Kaligis, F., Wiguna, T., Willenberg, L., Nguyen, H. T. M., Luchters, S., Azzopardi, P., & Fisher, J. (2019). Screening for depressive and anxiety disorders among adolescents in Indonesia: Formal validation of the centre for epidemiologic studies depression scale – revised and the Kessler psychological distress scale. *Journal of Affective Disorders*, *246*. https://doi.org/10.1016/j.jad.2018.12.042

Vos, T., Barber, R. M., Bell, B., Bertozzi-Villa, A., Biryukov, S., Bolliger, I., Charlson, F., Davis, A., Degenhardt, L., Dicker, D., Duan, L., Erskine, H., Feigin, V. L., Ferrari, A. J., Fitzmaurice, C., Fleming, T., Graetz, N., Guinovart, C., Haagsma, J., … Murray, C. J. L. (2015). Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: A systematic analysis for the Global Burden of Disease Study 2013. *The Lancet*, *386*(9995). https://doi.org/10.1016/S0140-6736(15)60692-4

## Figure captions

**Figure 1:** The top 20 variables with highest contributions in predicting depression onset of CLSA participants using baseline non-psychological measures.

**Figure 2:** The top ten variables with highest contributions in the model built only on CES-D-10, K10 scales, age and sex in predicting depression onset of CLSA participants.

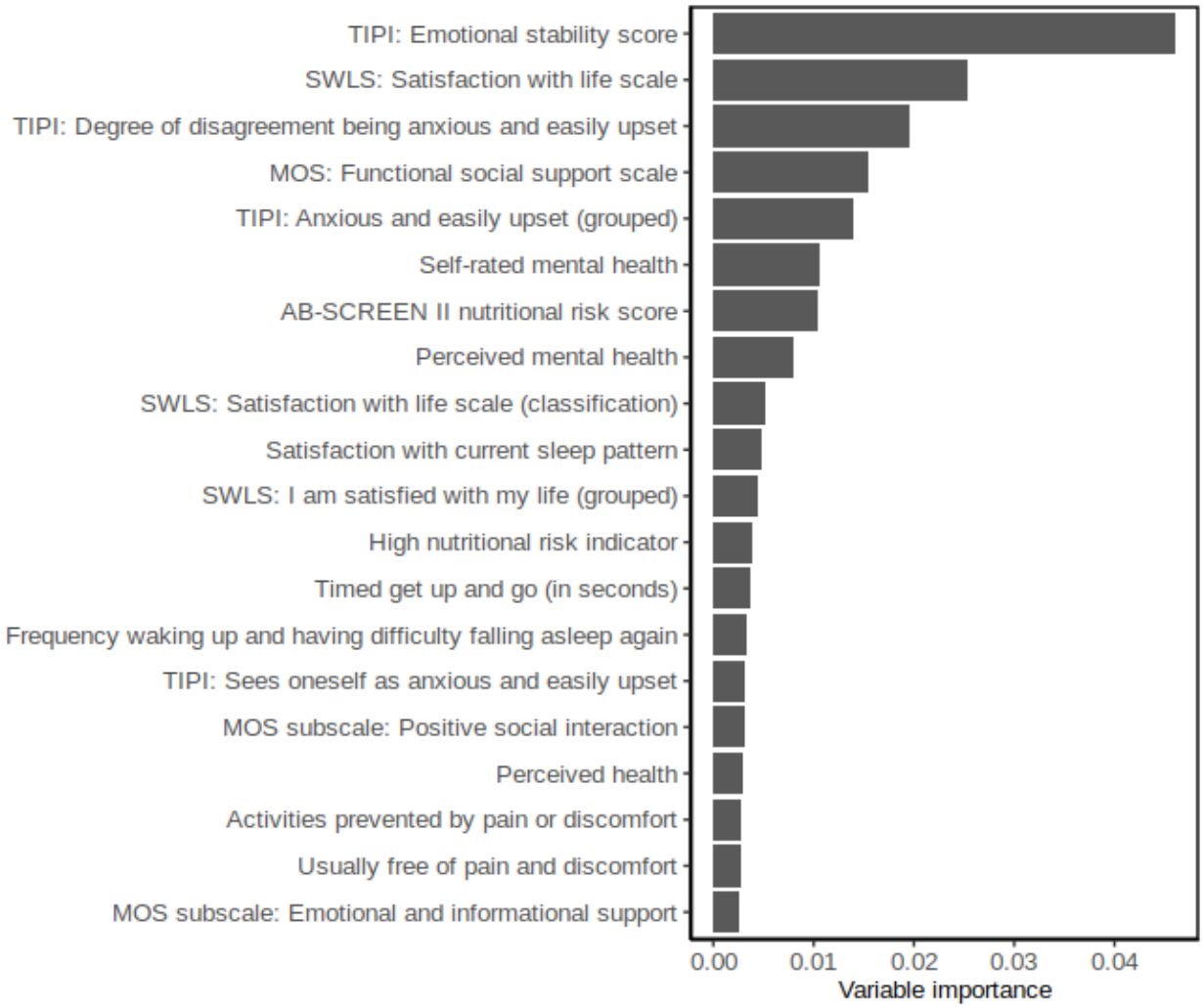
TIPI: Emotional stability score
SWLS: Satisfaction with life scale
TIPI: Degree of disagreement being anxious and easily upset
MOS: Functional social support scale
TIPI: Anxious and easily upset (grouped)
Self-rated mental health
AB-SCREEN II nutritional risk score
Perceived mental health
SWLS: Satisfaction with life scale (classification)
Satisfaction with current sleep pattern
SWLS: I am satisfied with my life (grouped)
High nutritional risk indicator
Timed get up and go (in seconds)
Frequency waking up and having difficulty falling asleep again
TIPI: Sees oneself as anxious and easily upset
MOS subscale: Positive social interaction
Perceived health
Activities prevented by pain or discomfort
Usually free of pain and discomfort
MOS subscale: Emotional and informational support
0.00
0.01
0.02
0.03
0.04
Variable importance

Figure 1

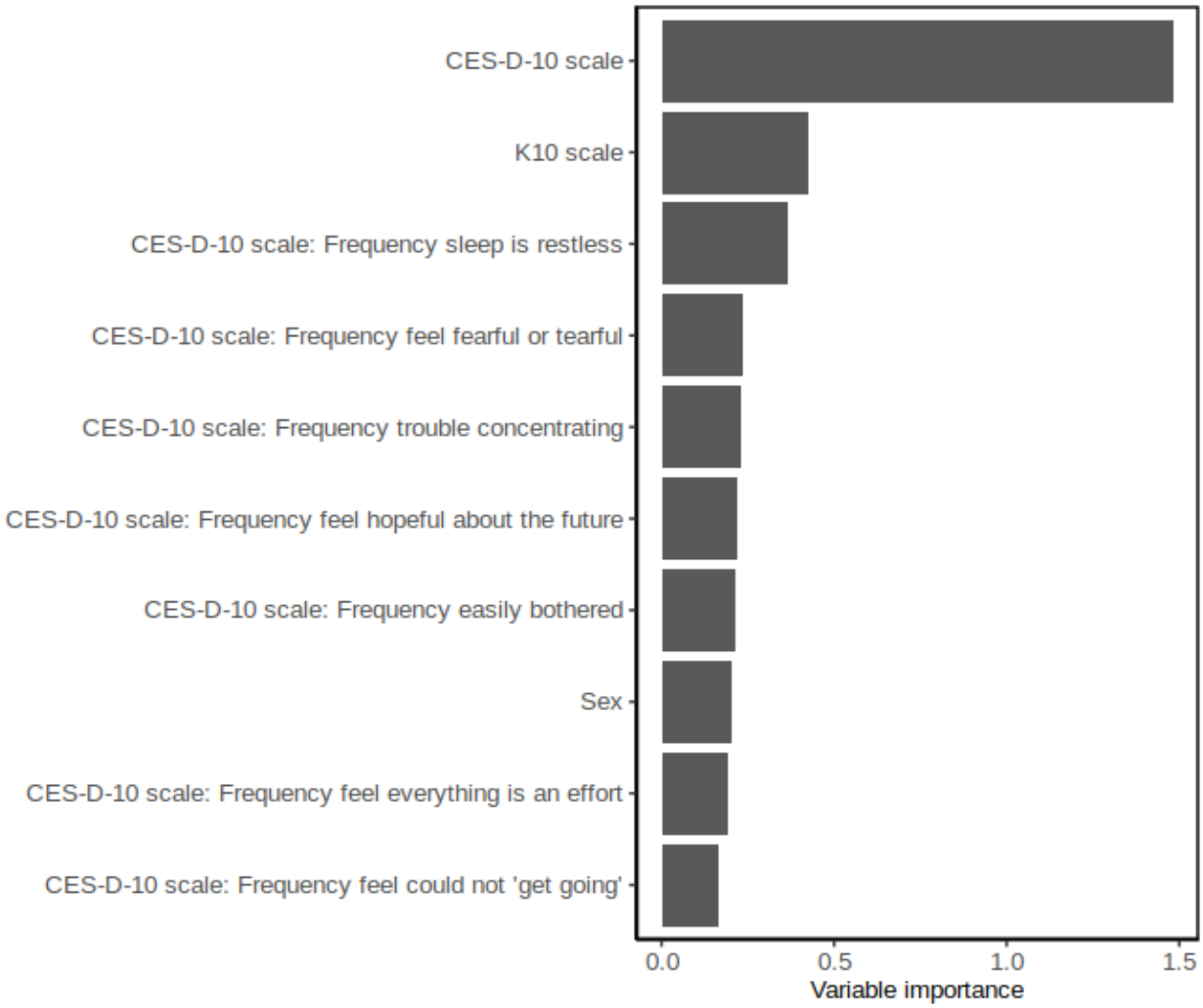
CES-D-10 scale
K10 scale
CES-D-10 scale: Frequency sleep is restless
CES-D-10 scale: Frequency feel fearful or tearful
CES-D-10 scale: Frequency trouble concentrating
CES-D-10 scale: Frequency feel hopeful about the future
CES-D-10 scale: Frequency easily bothered
Sex
CES-D-10 scale: Frequency feel everything is an effort
CES-D-10 scale: Frequency feel could not 'get going'
0.0
0.5
1.0
1.5
Variable importance

Figure 2

**Supplement for "Prediction of depression onset risk using machine learning and Canadian Longitudinal Study on Aging cohort"**

## Used machine learning algorithms

Logistic linear regression model with weights equal to the inverse of the frequencies of different classes in the outcome is used as the baseline linear model. Also, ridge type penalty[1] was added to avoid potential multicollinearity issues. Random forest model[2], an ensemble of decision trees built on random samples with replacement from the original data (bootstrapping), which has been shown superior performance from many biomedical studies, was used as an example of nonlinear model. As, the classes in the outcome are imbalanced, only 1,329 out of 19,024 subjects are of depression onset, we also used the balanced random forest model[3], which was designed for imbalanced classification problems by under sampling the majority class when generating bootstrapped samples for growing each decision tree, to see if better performance can be achieved.

## Model selection

For logistic linear regression model with ridge type penalty, the strength of ridge type penalty is selected. Grid search was used to find the optimal penalty strength from the searching range [0.1, 10] with optimal 10-Fold cross validation (CV) logistic loss. For random forest and balanced random forest model, the number of trees (searching range [10, 1500]) and the proportion of variables searched for splitting the base decision tree (searching range [0.1, 1.0] with a step size 0.1), were selected using Bayesian Optimization[4] to achieve optimal 10-Fold CV Area Under the Curve of Receiver Characteristic Operator (AUC).

## Computational environment

Python 3.7.6, Scikit-learn 1.2.0, Imbalanced-learn 0.8.1, Optuna 2.7 were used for fitting and selecting the above models. The figures were generated in R 3.6.3 using ggplot2.

**Figures**

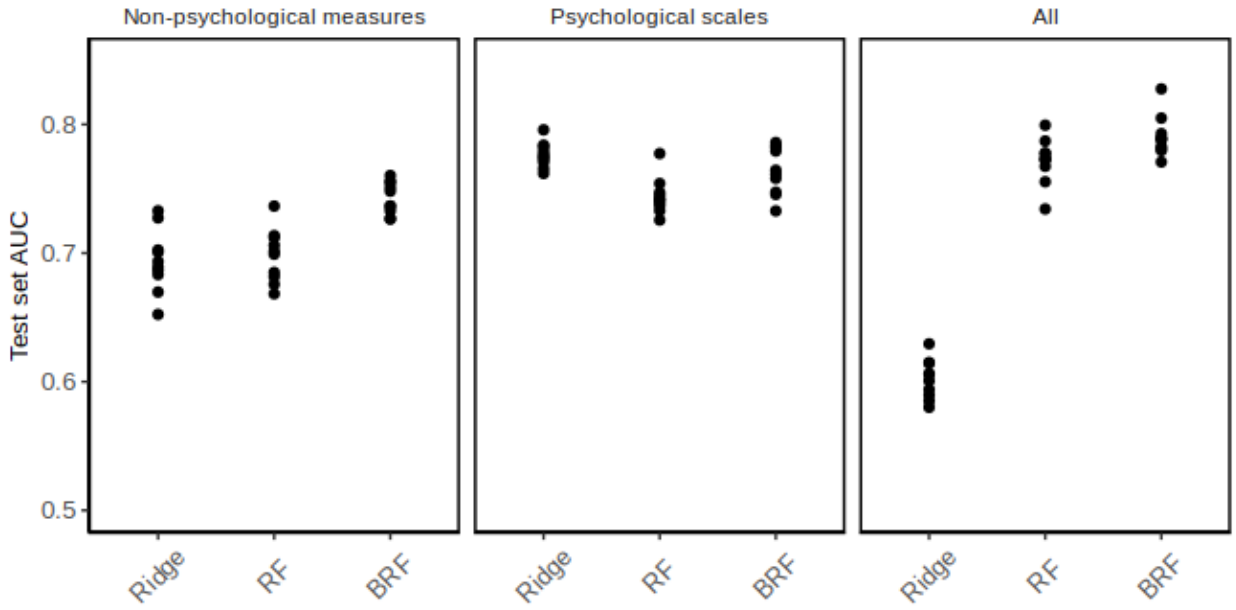


**Figure S1:** Different models' performance, test set AUC, in predicting depression onset using non-psychological measures, psychological scales and all baseline information. Ridge: logistic regression with ridge type penalty; RF: standard random forest model; BRF: balanced random forest model.

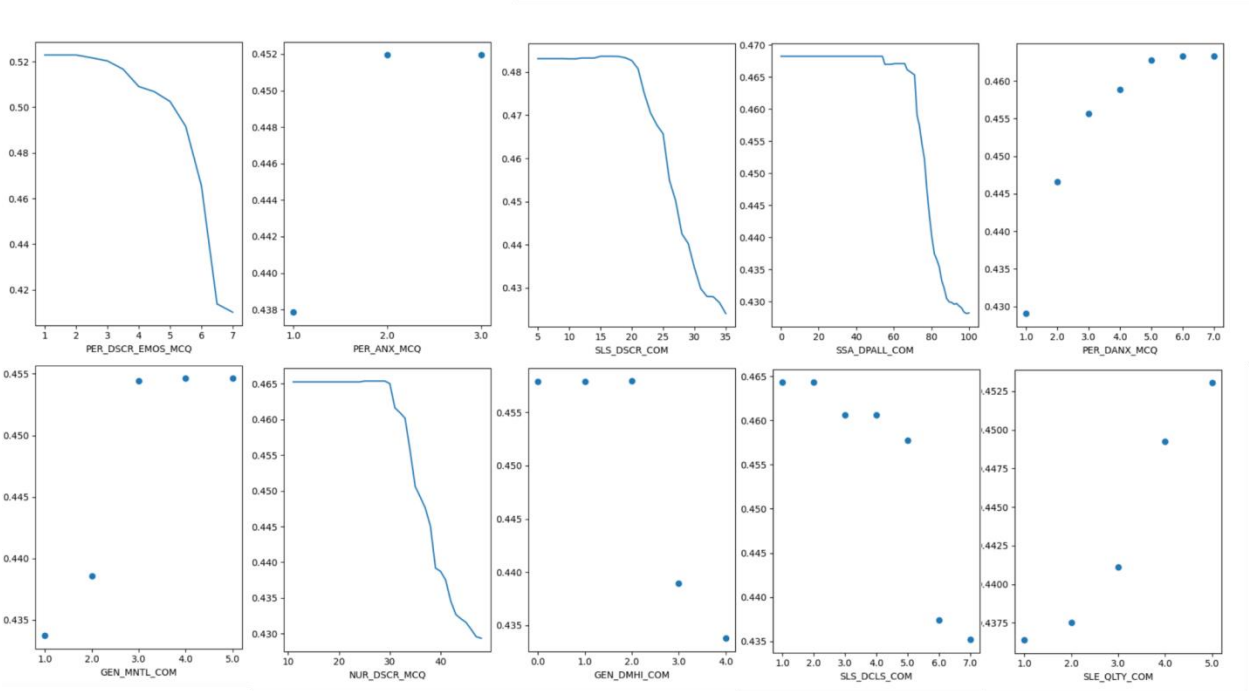


**Figure S2A:** Partial dependence plots of the top ten variables in predicting the depression onset using non-psychological scales to show their effect with direction. The meaning of the variables can be found in Table S1.

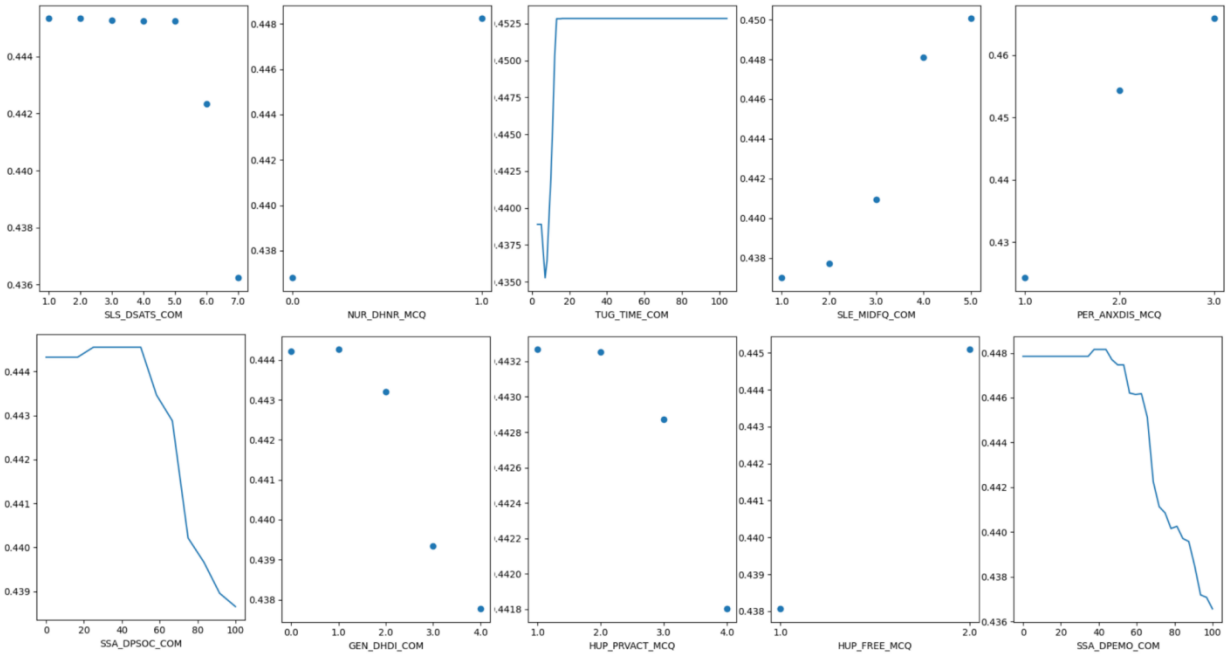


**Figure S2B:** Partial dependence plots of the top 11 to top 22 variables in predicting the depression onset using non-psychological scales to show their effect with direction. The meaning of the variables can be found in Table S1.

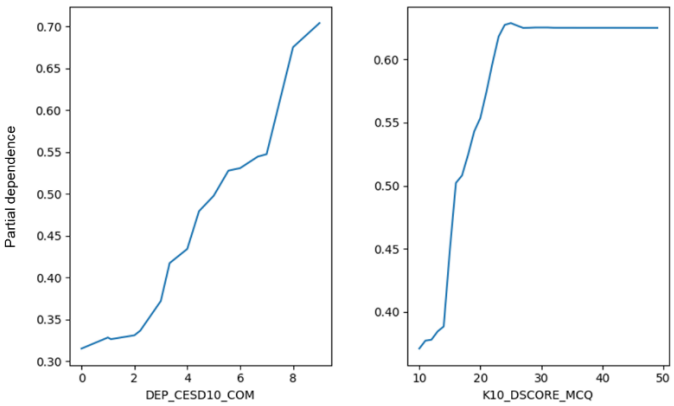


**Figure S3:** Partial dependence plots of the total scores of CESD and K10 scales to show their effect with direction.

**Tables**

**Table S1:** The corresponding between variable names and their meaning.

| Variable | Meaning |
|---|---|
| PER_DSCR_EMOS_MCQ | TIPI: Emotional Stability Score |
| SLS_DSCR_COM | Satisfaction with Life Scale (SWLS) - Score |
| PER_ANXDIS_MCQ | TIPI scale: Degree of disagreement being anxious and easily upset |
| SSA_DPALL_COM | Functional Social Support - MOS Scale |
| PER_DANX_MCQ | TIPI: Anxious and Easily Upset- Grouped |
| GEN_MNTL_COM | Self-rated mental health |
| NUR_DSCR_MCQ | The AB-SCREEN II Nutritional Risk Score |
| GEN_DMHI_COM | Perceived Mental Health |
| SLS_DCLS_COM | Satisfaction with Life Scale (SWLS) - Classification |
| SLE_QLTY_COM | Satisfaction with current sleep pattern |
| SLS_DSATS_COM | SWLS: I am satisfied with my life - Grouped |
| NUR_DHNR_MCQ | High Nutritional Risk Indicator |
| TUG_TIME_COM | Total time required to complete Timed Get Up and Go (in seconds) |

| | |
|---|---|
| SLE_MIDFQ_COM | Frequency waking up and having difficulty falling asleep again |
| PER_ANX_MCQ | TIPI scale: Sees oneself as anxious and easily upset |
| SSA_DPSOC_COM | Positive Social Interaction - MOS Subscale |
| GEN_DHDI_COM | Perceived Health |
| HUP_PRVACT_MCQ | Activities prevented by pain or discomfort |
| HUP_FREE_MCQ | Usually free of pain and discomfort |
| SSA_DPEMO_COM | Emotional and Informational Support - MOS Subscale |