Predicting persistent depressive symptoms in older adults: a machine learning approach to personalised mental healthcare

Christopher M. Hatton[1, 2], Lewis W. Paton[1], Dean McMillan[1, 2], James Cussens[3], Simon Gilbody [1, 2], Paul A. Tiffin[1, 2]

1. Department of Health Sciences, University of York, UK
2. Hull York Medical School, University of York, UK
3. Department of Computer Science & York Centre for Complex Systems Analysis, University of York, UK

Author contact details:

Christopher Hatton - email: hych2@hyms.ac.uk

Lewis Paton - email: lewis.paton@york.ac.uk, Tel: 0190432(1516)

Dean McMillan - dean.mcmillan@york.ac.uk, Tel: 0190432(1359)

James Cussens – email: james.cussens@york.ac.uk, Tel: 0190432(5371)

Simon Gilbody – email: simon.gilbody@york.ac.uk, Tel: 0190432(1370)

Paul Tiffin – email: paul.tiffin@york.ac.uk, Tel: 0190432(1117)

1. Introduction

Depression causes significant physical and psychosocial morbidity, and is associated with poorer quality of life (Sivertsen et al., 2015). Interventions targeted at high-risk populations may reduce the burden of depression (van Zoonen et al., 2014). Therefore, the ability to predict persistent symptoms is desirable. Subthreshold depression refers to the presence of depressive symptoms that do not satisfy diagnostic criteria for clinical depression (Rowe & Rapaport, 2006). It is common amongst older adults and is a risk factor for major depression (Judd et al., 1998)

Machine learning (ML) is a rapidly emerging field that has allowed the exploitation of large, often routinely arising, datasets to generate predictive models. In 'supervised learning' machines develop ways of linking a target outcome from a set of predictors ('features') in existing data. Such models may generalize to novel predictor data. In contrast to traditional statistical approaches, ML focuses on prediction rather than explanation, often producing non-interpretable models.

Prior research has demonstrated that it is possible to predict the likelihood of individuals developing depression (King et al., 2008). However, the utility of ML methods for depression prediction has not been widely explored, as most studies use conventional statistical methods, such as logistic regression (LR). In addition, the efficacy of predictive algorithms has not been evaluated in older individuals, who are often excluded from analyses. Therefore, this research aims to compare the efficacy of LR and a ML approach ('extreme gradient boosting') in predicting persistent depressive symptoms at 12 months in older adults with 'subthreshold depression' at baseline.

2. Methods

2.1 Data

Data were acquired from a previous randomised controlled trial evaluating the effectiveness of collaborative care for sub-threshold depression in older people (Gilbody et al., 2017). Data were potentially available for 361 participants (139 males, 222 females) aged 65 years and older, who were randomised to the control arm (usual primary care) in the original CASPER trial. All participants were recruited from primary care practices in the Northern England. Patients were considered for inclusion if they screened positive for the Whooley questions (Whooley et al., 1997) and reported sub-threshold depressive symptoms according DSM-IV criteria using

the Mini International Neuropsychiatric Interview (MINI) (Sheehan et al., 1998). Patients were visited at baseline, 4 months, and 12 months by clinical studies officers, who administered the questionnaires. Further details are available in the CASPER trial report (Gilbody et al., 2017). The study comprised a secondary analysis of de-identified data from the CASPER trial and fell within the remit of original ethical approval.

77 participants (21%) were missing primary outcome data at 12 month follow up, and these participants were excluded from the analysis. Of the remaining 284 participants (100 male, 184 female) the mean age was 77.7 (SD: 6.9). The mean PHQ-9 score was 7.7 (SD: 4.6) at baseline, and 7.2 (SD: 5.0) at 12-month follow up. For those with a primary outcome reported there were very few missing data on predictor variables: only six participants (2%) had missing data on one or more of the baseline questionnaires.

2.2 Predictors and outcome

Baseline demographic data included age, sex, and prior education status. Psychometric data were available from: Patient Health Questionnaire (PHQ-9) (Kroenke et al., 2001); MINI (Sheehan et al., 1998); Whooley questions (Whooley et al., 1997); Generalised Anxiety Disorder Assessment (GAD-7) (Spitzer et al., 2006); Short Form Survey 12 (SF-12) (Gandek et al., 1998); Patient Health Questionnaire 15 (PHQ-15) (Kroenke & Spitzer, 2002); Connor-Davidson Resilience Scale 2 (CD-RISC2) (Vaishnavi et al., 2007); and the EuroQol-5D (EQ-5D) (Balestroni & Bertolotti, 2012). Information regarding current medication and physical health conditions were also available. A PHQ-9 score ≥10 was used to define the outcome of this study, significant depressive symptoms at 12 months. It has been extensively validated as a screening cut-off for major depressive disorder, with an estimated sensitivity of 0.85, a specificity of 0.89 (Manea et al., 2012), and a relatively high internal consistency (Cronbach's alpha 0.89) (Kroenke et al., 2001).

2.3 Data preprocessing

Missing values for predictor variables were imputed via the Amelia II package for R, which uses an Expectation-Maximization Bootstrap-based (EBM) algorithm (Honaker et al., 2011). Continuous variables were placed on a similar scale ('Feature scaling') via normalization. For each run of model building and testing the data were split in a 60:40 ratio to form a training set to develop the predictive algorithm, and a test ('held

back') set to validate the algorithm. Stratified random sampling was used to ensure equal distribution of the outcome (PHQ-9 score ≥10) between sets.

2.4 Predictive Model Building

For the LR modelling a backwards stepwise approach was used to select predictor variables. The model starts with all potential predictors and sequentially removes variables based on Mallow's C_p. 'Extreme Gradient Boosting', as implemented in the XGBoost R package (Chen et al., 2015) was selected as the ML approach, as gradient boosting has been shown to often outperform other ML methods for binary classification tasks (Caruana & Niculescu-Mizil, 2006). XGBoost uses both gradient boosting and 'ensembling' to maximise its predictive capacity. 'Ensembling' is the process of combining numerous models, in this case decision trees, and averaging or voting on the final predicted value. Gradient boosting refers to the algorithm's ability to successfully model the relationship between the predictor and outcome, first on the initial data, and subsequently on the 'residuals' or 'errors' from the previous modelling attempt. This ensures the algorithm successively focuses on observations where the outcome is increasingly difficult to predict.

The LR and ML classification models were developed to predict the likelihood of participants having persistent depressive symptoms (PHQ-9 ≥10) at 12 months on the training dataset. Each model, developed using the training dataset, was then applied to the previously unseen test dataset. Due to the stochastic nature of developing the models (e.g. dataset splitting, imputation etc.) the entire process was repeated 1000 times. The performance of the models on the unseen test dataset was evaluated by computing the mean accuracy metrics (e.g. sensitivity, specificity etc.) over the 1000 iterations. An unpaired t-test was used to formally compare the overall predictive performance of LR and ML, using the AUC values.

It was not possible to meaningfully recombine the coefficients from the 1000 LR models built in order to derive interpretable results. This was because they varied, to some extent, in the number of predictors retained in the final models. Therefore, in order to clearly illustrate the main predictors of persistent depressive symptoms, a separate LR model was developed using the entire dataset with present outcomes (N=284). This used a backwards stepwise approach to predictor selection, sequentially excluding variables where the regression coefficient had an associated p value greater than 0.05.

3. Results

The ability of the models to predict depressive symptoms (PHQ-9 ≥10) at 12 months is shown in Table 1.

Table 1. Predictive ability of logistic regression and extreme gradient boosting models in the test (validation) datasets according to mean accuracy metrics.

	Logistic Regression (95% confidence interval)	Extreme Gradient Boosting (95% confidence interval)
Accuracy	69% (68 -69%)	74% (74-74%)
AUC	0.67 (0.67 – 0.68)	0.72 (0.71 – 0.72)
Sensitivity	0.79 (0.79 – 0.79)	0.78 (0.78 – 0.79)
Specificity	0.43 (0.43 – 0.44)	0.56 (0.55 – 0.56)
PPV	0.78 (0.77 – 0.78)	0.89 (0.89 – 0.89)
NPV	0.45 (0.44 – 0.45)	0.35 (0.34 – 0.35)

Footnotes to Table 1: AUC – Area Under the Receiver Operative Characteristic Curve; PPV – Positive Predictive Value; NPV – Negative Predictive Value.

On average, in terms of Positive Predictive Value (PPV) value, 89% of those predicted by machine learning to have PHQ-9 scores above threshold at 12 months, actually did, compared to 78% using LR. Conversely, on average, the Negative Predictive Value (NPV) was superior for the LR approach (45% vs. 35%). However, the AUC values, an overall metric of the potential utility as a screening method, were statistically significantly higher for the ML approach compared to LR ($p<0.0001$).

The XGBoosting process did not yield interpretable models as such. However, the output for each modelling run produced ‘importance’ metrics for the predictors. These provided an indication of the relative importance of the baseline variables in allowing the algorithm to predict the outcomes. The baseline SF-12 and PHQ-15 scale scores frequently featured in the top five variables ranked for importance. In contrast, the interpretable results of the LR built on the entire dataset with available outcomes (N=284) are shown in Table 2. Five variables at baseline had an independent, statistically significant ($p<0.05$) ability to predict persistence of depressive symptoms at 12 months. These included the reported degree of limitation of daily activities, as indicated by the response to item 3 of the EQ-5D. For every point scored on this question at baseline, the odds of persistent depressive symptoms, on average, almost quadrupled. Likewise, reporting the previous presence of feelings of guilt or worthlessness (MINI item ‘3e’) roughly more than tripled the odds of persistent

depressive symptoms at one year follow-up. These effects were independent of the value of other predictor variables.

Table 2. The results from a logistic regression predicting persistent depressive symptoms (PHQ ≥10) at 12 month follow-up from baseline characteristics in the CASPER control group participants with an outcome recorded (N=284).

	Odds ratio	95% confidence interval	p value
Baseline GAD-7 total score	2.34	1.39 – 3.93	0.001
Baseline PHQ-9 total score	1.15	1.03 – 1.28	0.009
Baseline anhedonia	2.58	1.00 – 6.70	0.051
Previous 'guilty' or 'worthless' feelings (MINI item 3e)	3.33	1.45 – 7.65	0.005
Degree of problems with usual activities at baseline (EQ-5D item 3)	3.74	1.53 – 9.14	0.004

Footnote to Table 2: GAD-7 – Generalised Anxiety Disorder Assessment; PHQ-9 – Patient Health Questionnaire; MINI – Mini International Neuropsychiatric Interview; EQ-5D – EuroQol-5D. McFadden's pseudo R^2 for the model = 0.315.

Discussion

We observed that the predictive performance of the ML approach (mean AUC 0.72) was modestly superior to that of LR (mean AUC 0.67). Notably, the PPV for ML was considerably higher than LR, though at the expense of NPV. Prior research using the conventional approach of LR has developed predictive algorithms with AUCs between 0.71 – 0.79 (King et al., 2008). The average predictive performance using machine learning in this study was comparable to these previous models, despite being developed using a substantially smaller sample; 5000 patients in the prior research compared to fewer than 200 in the 'training' datasets used in this study.

This is the first study to predict persistent depression in older adults using a ML approach. A key strength was the use of demographic and 'paper and pencil' questionnaire data that could be collected in routine practice. In addition, the LR and ML-based predictive models were built and evaluated upon the same training and test datasets, ensuring a fair comparison between the approaches.

The key limitation was the use of the PHQ-9 scores as an indicator of persistent depressive symptoms, and whilst well validated (Kroenke et al, 2001), a clinical interview would have been preferable. The predictive performance of ML may have been improved using other approaches such as ‘deep learning’, which would have been more appropriate with larger datasets. In addition, despite the greater predictive ability of ML, it produces non-interpretable models and therefore LR is more appropriate to investigate the contribution of individual risk factors in depression.

According to the results from the LR model built on the total dataset there were a number of key variables that significantly and independently predicted the likelihood of participants having significant depressive symptoms at 12 months. These included baseline anxiety and depressive symptoms, as well as reported functioning in relation to activities of daily living. The predictors used by the XGB machine learning models differed slightly, most commonly using the baseline health related quality of life (SF-12 PCS, SF-12 MCS), and baseline somatoform symptoms (PHQ-15). This is somewhat unsurprising as the methods use very different approaches to model building. Moreover, the overall LR model presented only included variables with independent and statistically significant predictive ability.

The models were developed using baseline demographic data, and questionnaire data from a previous RCT that are not currently routinely collected in clinical practice. However, the XGB machine learning model consistently used scores from a relative small number of baseline characteristics, notably the SF-12 and PHQ-15. Moreover, single items, such as the presence of anhedonia or previous feelings of guilt or worthlessness were independent and relatively strong predictors of the outcome in their own right. This suggests that the quantity of data that would need to be collected in actual clinical practice in order to routinely implement such a predictive approach would be relatively modest. This may consist of four or five relatively short symptom questions and some key additional items, conceivably amounting to around 20 minutes of completion time. In addition, the PPV value associated with the ML approach suggests that the ‘false positive’ rate would be only around 11% in this population. Depending on the specific service context this may or may not be acceptable. Nevertheless, our findings demonstrate that a ML approach to prediction may be superior to classical statistical methods even using relatively modestly sized datasets. The acceptability and utility of such algorithms in practice should be evaluated in future research. In particular it would be important to assess to what

extent clinicians and patients acted on the predictions produced from such models, in order to impact outcomes. Certainly further research evaluating the feasibility and acceptability of such an approach is warranted by these findings.

References

Balestroni G, Bertolotti G., 2012. EuroQol-5D (EQ-5D): An instrument for measuring quality of life. Monaldi Arch Chest Dis. 78,155–159. https://doi.org/10.4081/monaldi.2012.121

Caruana R, Niculescu-Mizil A., 2006. An empirical comparison of supervised learning algorithms. Proceedings of the 23rd Int. Conf. Machine learning 2006. C,161-168. https://doi.org/10.1145/1143844.1143865

Chen T, He T, Benesty M., 2015. XGBoost: Extreme gradient boosting. R package version 0.4–4.

Gandek B, Ware JE, Aaronson NK, Apolone G, Bjorner JB, Brazier JE, Bullinger M, Kaasa S, Leplege A, Prieto L, Sullivan M. Cross-validation of item selection and scoring for the SF-12 Health Survey in nine countries: results from the IQOLA Project., 1998. J. Clin. Epidemiol. 51, 1171-1178. https://doi.org/10.1016/S0895-4356(98)00109-7

Gilbody, S., Lewis, H., Adamson, J., Atherton, K., Bailey, D., Birtwistle, J., Bosanquet, K., Clare, E., Delgadillo, J., Ekers, D., Foster, D., Gabe, R., Gascoyne, S., Haley, L., Hamilton, J., Hargate, R., Hewitt, C., Holmes, J., Keding, A., Lilley-Kelly, A., Meer, S., Mitchell, N., Overend, K., Pasterfield, M., Pervin, J., Richards, D.A., Spilsbury, K., Traviss-Turner, G., Trépel, D., Woodhouse, R., Ziegler, F., McMillan, D., 2017. Effect of collaborative care vs usual care on depressive symptoms in older adults with subthreshold depression: The CASPER randomized clinical trial. JAMA - J. Am. Med. Assoc. 317, 728–737. https://doi.org/10.1001/jama.2017.0130

Honaker, J., King, G., Blackwell, M., 2013. Amelia II: A program for missing data, R package version 1.7.2, 2013 45, 1–57.

Judd, L.L., Akiskal, H.S., Maser, J.D., Zeller, P.J., Endicott, J., Coryell, W., Paulus, M.P., Kunovac, J.L., Leon, A.C., Mueller, T.I., Rice, J.A., Keller, M.B., 1998. Major depressive disorder: A prospective study of residual subthreshold depressive symptoms as predictor of rapid relapse. J. Affect. Disord. 50, 97–108. https://doi.org/10.1016/S0165-0327(98)00138-4

King, M., Walker, C., Levy, G., Bottomley, C., Royston, P., Weich, S., Bellón-Saameño, J.Á., Moreno, B., Švab, I., Rotar, D., Rifel, J., Maaroos, H.-I., Aluoja, A., Kalda, R., Neeleman, J., Geerlings, M.I., Xavier, M., Carraça, I., Gonçalves-Pereira, M., Vicente, B., Saldivia, S., Melipillan, R., Torres-Gonzalez, F., Nazareth, I., 2008. Development and Validation of an International Risk Prediction Algorithm for Episodes of Major Depression in General Practice Attendees. Arch. Gen. Psychiatry 65, 1368. https://doi.org/10.1001/archpsyc.65.12.1368

Kroenke, K., Spitzer, R.L., 2002. The PHQ-9: A New Depression Diagnostic and Severity Measure. Psychiatr. Ann. 32, 509–515. https://doi.org/10.3928/0048-5713-20020901-06

Kroenke, K., Spitzer, R.L., Williams, J.B.W., 2001. The PHQ-9: Validity of a brief depression severity measure. J. Gen. Intern. Med. 16, 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Manea, L., Gilbody, S., McMillan, D., 2012. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): A meta-analysis. Cmaj. 184, 191-196. https://doi.org/10.1503/cmaj.110829

Rowe SK, Rapaport MH. Classification and treatment of sub-threshold depression. Curr Opin Psychiatry. 2006;19:9-13. https://doi.org/10.1097/01.yco.0000194148.267.66.ba

Sheehan, D.V., Lecrubier, Y., Sheehan, K.H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R. and Dunbar, G.C., 1998. The Mini-International Neuropsychiatric Interview (MINI): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J. Clin. Psych. 59, 22.

Sivertsen, H., Bjørkløf, G.H., Engedal, K., Selbæk, G., Helvik, A.S., 2015. Depression and quality of life in older persons: A review. Dement. Geriatr. Cogn. Disord. 40, 311–339. https://doi.org/10.1159/000437299

Spitzer, R.L., Kroenke, K., Williams, J.B.W., Löwe, B., 2006. A Brief Measure for Assessing Generalized Anxiety Disorder. Arch. Intern. Med. 166, 1092. https://doi.org/10.1001/archinte.166.10.1092

Vaishnavi, S., Connor, K., Davidson, J.R., 2007. An abbreviated version of the Connor–Davidson Resilience Scale (CD-RISC), the CD-RISC2: Psychometric properties and applications in psychopharmacological trials. Psychiatry Research, 152, 293-297. https://doi.org/10.1016/j.psychres.2007.01.006

van Zoonen, K., Buntrock, C., Ebert, D.D., Smit, F., Reynolds, C.F., Beekman, A.T.F., Cuijpers, P., 2014. Preventing the onset of major depressive disorder: A meta-analytic review of psychological interventions. Int. J. Epidemiol. 43, 318–329. https://doi.org/10.1093/ije/dyt175

Whooley, M.A., Avins, A.L., Miranda, J., Browner, W.S., 1997. Case-finding instruments for depression: Two questions are as good as many. J. Gen. Intern. Med. 12, 439–445. https://doi.org/10.1046/j.1525-1497.1997.00076.x