Molecular Psychiatry (2016) 21, 1366–1371
© 2016 Macmillan Publishers Limited, part of Springer Nature. All rights reserved 1359-4184/16
www.nature.com/mp

ORIGINAL ARTICLE

Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports

RC Kessler[1], HM van Loo[2], KJ Wardenaar[2], RM Bossarte[3], LA Brenner[4], T Cai[5], DD Ebert[1,6], I Hwang[1], J Li[5], P de Jonge[2], AA Nierenberg[7], MV Petukhova[1], AJ Rosellini[1], NA Sampson[1], RA Schoevers[2], MA Wilcox[8] and AM Zaslavsky[1]

Heterogeneity of major depressive disorder (MDD) illness course complicates clinical decision-making. Although efforts to use symptom profiles or biomarkers to develop clinically useful prognostic subtypes have had limited success, a recent report showed that machine-learning (ML) models developed from self-reports about incident episode characteristics and comorbidities among respondents with lifetime MDD in the World Health Organization World Mental Health (WMH) Surveys predicted MDD persistence, chronicity and severity with good accuracy. We report results of model validation in an independent prospective national household sample of 1056 respondents with lifetime MDD at baseline. The WMH ML models were applied to these baseline data to generate predicted outcome scores that were compared with observed scores assessed 10–12 years after baseline. ML model prediction accuracy was also compared with that of conventional logistic regression models. Area under the receiver operating characteristic curve based on ML (0.63 for high chronicity and 0.71–0.76 for the other prospective outcomes) was consistently higher than for the logistic models (0.62–0.70) despite the latter models including more predictors. A total of 34.6–38.1% of respondents with subsequent high persistence chronicity and 40.8–55.8% with the severity indicators were in the top 20% of the baseline ML-predicted risk distribution, while only 0.9% of respondents with subsequent hospitalizations and 1.5% with suicide attempts were in the lowest 20% of the ML-predicted risk distribution. These results confirm that clinically useful MDD risk-stratification models can be generated from baseline patient self-reports and that ML methods improve on conventional methods in developing such models.

Molecular Psychiatry (2016) **21,** 1366–1371; doi:10.1038/mp.2015.198; published online 5 January 2016

INTRODUCTION

Heterogeneity in major depressive disorder (MDD) illness course complicates clinical decision-making. Clinicians have consistently identified absence of guidance on how to deal with this variation as a critical gap in personalizing MDD treatment.[1–4] However, efforts to address this problem by finding useful prognostic subtypes based on empirically derived symptom profiles[5,6] or biomarkers[7–9] have so far yielded disappointing results. A potentially promising complementary approach would be to apply machine-learning (ML) methods to baseline data on symptoms and other easily assessed clinical features to develop first-stage prediction models of subsequent depression course and treatment response[10,11] that could be expanded to target and examine incremental prognostic effects of novel biomarkers among patients who could not be classified definitively with the inexpensive first-stage prediction models.

Although ML methods have been used successfully to develop risk-prediction schemes in other areas of medicine,[12,13] applications to depression have so far relied on small samples and thin predictor sets, failing to realize the full potential of the methods.[14,15] A recent exception is a study carried out among 8261 respondents with lifetime DSM-IV MDD in the World Health Organization World Mental Health (WMH) surveys.[16,17] Retrospective reports about parental history of depression, temporally primary comorbid disorders and characteristics of incident depressive episodes were used to predict retrospectively reported subsequent depression persistence (number of years with episodes), chronicity (number of years with episodes lasting most days), hospitalization for depression and work disability due to depression. K-means cluster analysis of the four predicted risk scores found a parsimonious three-cluster solution with the high-risk cluster (32.4% of cases) accounting for 56.6–72.9% of high persistence, chronicity, hospitalization and disability.

Although useful as a proof of concept, the WMH results were based on retrospective reports. A prospective validation is reported here that uses the WMH models to predict subsequent MDD persistence, chronicity and severity in a sample of 1056 respondents with lifetime DSM-III-R MDD in the 1990–1992

[1]Department of Health Care Policy, Harvard Medical School, Boston, MA, USA; [2]Interdisciplinary Center Psychopathology and Emotion Regulation, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands; [3]Office of Public Health, Department of Veterans Affairs, Washington, DC, USA; [4]Departments of Physical Medicine and Rehabilitation, Psychiatry, and Neurology, University of Colorado, Anschutz Medical Campus, Aurora, Colorado; Rocky Mountain Mental Illness Research Education and Clinical Center, Denver, CO, USA; [5]Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA; [6]Department of Psychology, Clinical Psychology and Psychotherapy, Friedrich-Alexander University Nuremberg-Erlangen, Erlangen, Germany; [7]Department of Psychiatry and Depression Clinical and Research Program, Harvard Medical School and Massachusetts General Hospital, Boston, MA, USA and [8]Epidemiology, Janssen Research & Development, LLC, Titusville, NJ, USA. Correspondence: Dr RC Kessler, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, USA.
E-mail: Kessler@hcp.med.harvard.edu

A complete list of NCS and NCS-2 publications can be found at http://www.hcp.med.harvard.edu/ncs.

Received 22 May 2015; revised 30 September 2015; accepted 26 October 2015; published online 5 January 2016

US National Comorbidity Survey (Survey 1)[18] who were re-interviewed 10–12 years later in the 2001–2003 National Comorbidity Survey Follow-Up (Survey 2).[19] ML model results are compared with results based on more conventional logistic regression models to determine whether ML methods improve on conventional methods.

MATERIALS AND METHODS

Sample

Survey 1 was a community epidemiological survey of common DSM-III-R disorders among English-speaking residents of the non-institutionalized civilian US household population aged 15–54 years ($n = 5877$ respondents; 82.4% response rate).[18] Respondents were paid $25 for participation. Recruitment-consent procedures were approved by the human subjects committee of the University of Michigan. Interviews were conducted face-to-face in respondent homes after obtaining verbal informed consent. Survey 2 attempted to re-interview all baseline respondents considered here 10–12 years later using recruitment-consent procedures identical to Survey 1 other than a $50 incentive. These procedures were approved by the human subjects committees of both Harvard Medical School and the University of Michigan. Interviews were again conducted face-to-face in respondent homes after obtaining a verbal informed consent. The 5001 Survey 2 respondents (87.6% of living targeted Survey 1 respondents) were administered an expanded version of the baseline interview assessing onset-course of disorders between the two surveys. A non-response adjustment weight corrected for baseline differences between Survey 2 respondents and non-respondents conditional on Survey 1 responses. Analyses reported here use the weighted data from the 1056 Surveys 1–2 panel respondents who met lifetime criteria for MDD in Survey 1.

The baseline assessment of DSM-III-R disorders

Survey 1 assessed DSM-III-R disorders with the World Health Organization's Composite International Diagnostic Interview (CIDI) Version 1.1, a fully structured lay-administered interview that assessed common mental disorders using DSM-III-R criteria.[20] Syndromes assessed included major depressive episode, mania-hypomania, six anxiety disorders (generalized anxiety disorder, panic disorder, agoraphobia, specific phobia, social phobia and post-traumatic stress disorder) and five externalizing disorders (conduct disorder, alcohol abuse, alcohol dependence, drug abuse and drug dependence). Blinded Structured Clinical Interview for DSM-III-R[21] clinical reappraisal interviews in a probability sub-sample found good concordance with the DSM-III-R/CIDI diagnoses.[20] Respondents with lifetime MDD were asked whether their first lifetime episode 'was brought on by some stressful experience' or happened 'out of the blue'. DSM-III-R Criteria A-D MDE symptoms were then assessed for this incident episode. Family History Research Diagnostic Criteria questions[22] were used to determine parental history of depression.

Outcome measures

Depression persistence, chronicity and severity were assessed in Survey 2 with a computerized version of CIDI 3.0 using 'pre-loaded' information about Survey 1 responses to guide follow-up questioning. Respondents with Survey 1 lifetime MDD were asked to review the depressive symptoms reported in Survey 1, update subsequent episodes and symptoms using a life history calender and answer four summary questions about subsequent episodes: in how many years since baseline did the respondent have a depressive episode lasting 2+ weeks (referred to below as 'persistence') and an episode lasting most days throughout the year (referred to below as 'chronicity')? Was the respondent ever hospitalized for depression since the baseline? Was the respondent currently disabled (at least 50% limitation in ability to perform paid work) because of depression? A fifth Survey 2 outcome measure was whether the respondent attempted suicide at any time since the baseline.

Analysis methods

Predicting the outcomes in the WMH surveys. The predictors in the WMH surveys included temporally primary comorbid lifetime disorders, parental depression, MDD incident episode symptoms and other information about the incident episode (age-of-onset and if the episode was triggered or endogenous). The outcomes were MDD persistence severity (number of years since age-of-onset with episodes lasting 2+ weeks and lasting most days throughout the year, each standardized to a 0–100% range in relation to number of years between age-of-onset and age-at-interview), whether respondents were ever hospitalized for depression after their first episode, and whether respondents were disabled at the time of interview because of their depression. The ML methods used to develop the models included ensemble regression trees[23] and 10-fold cross-validated penalized regression,[24] both of which were designed to avoid overfitting. These methods are described elsewhere.[16,17]

Between 9 and 13 predictors available at baseline in Surveys 1–2 emerged as significant in each WMH model, including measures of individual symptoms and symptom clusters in the incident episode, whether that episode was triggered or endogenous, parental history of depression and various measures of temporally primary comorbid anxiety and externalizing disorders (some of them depending on age-of-onset). A more detailed discussion of the final WMH models is available elsewhere.[16,17]

To evaluate whether models based on ML methods improve prediction in an independent data set more than models based on conventional methods, we also estimated a logistic regression model for each outcome in the WMH data that included 23 predictors: the nine DSM-III-R Criterion A symptoms of MDD, a measure of whether the episode was triggered or endogenous, parental history of depression and 11 measures of the temporally primary comorbid anxiety and externalizing disorders that were also available in Survey 1. To the extent that the ML methods stabilize estimates, we would expect predictions based on these methods to outperform predictions based on logistic regression despite the ML models containing fewer predictors (9–13) than the logistic models (23).

Assigning WMH-predicted risk scores to Survey 1 respondents. Risk scores based on the logistic models were generated in Survey 1 using the WMH coefficients and the Survey 1 predictors. This direct estimation method could not be used for the ML models, though, as Survey 1 did not assess a number of significant predictors in the ML models (symptoms of anxious depression and mixed episodes in incident episodes, comorbid obsessive–compulsive disorder, intermittent explosive disorder and oppositional defiant disorder). We addressed this problem by imputing ML risk scores to Survey 1 respondents from a consolidated data set that combined WMH respondents and Surveys 1 and 2 respondents. The data set included all predictors in common across the surveys along with the four ML-predicted risk scores. The latter four scores had valid values for WMH cases and missing values for Survey 1 cases. Multiple imputation was applied to this data set to generate 10 predicted scores on each missing variable to each Survey 1 respondent using SAS 9.2 (Cary, NC, USA) *proc mi*.[25] Modal imputed values were assigned to each Survey 1 respondent for purposes of analysis. As these scores were strongly correlated across outcomes, a single composite ML-predicted risk score was then constructed for each respondent by averaging across the four scores after transforming to percentiles.

Validating the prediction models. Survey 2 outcomes were predicted from risk scores based on the ML and logistic models applied to the Survey 1 data. The Survey 2 outcomes included high (top 10%) MDD persistence and chronicity in the 10–12 years between the two surveys, hospitalization for depression and attempted suicide during those years and disability due to depression at the time of Survey 2. Area under the receiver operating characteristic curve (AUC) was calculated for each Survey 2 outcome separately for the ML and logistic models. Sensitivity (SN; the percentage of respondents with the outcome classified by the predicted risk scores as having high risk), positive predictive value (PPV; the percentage of respondents predicted to have high risk who experienced the outcome) and likelihood-ratio positive (LR+; the relative proportions of respondents who experienced the outcome among those classified as having or not having high risk) were also calculated for the 20 and 33% of Survey 1 respondents with highest and lowest ML-imputed composite risk scores. S.e.m. of SN, PPV and LR+ were estimated using the Taylor series method with SUDAAN[26] to adjust for design effects in the Surveys 1 and 2 panels.

RESULTS

Outcome distributions

One-third (37.9%) of the 1056 Surveys 1 and 2 respondents had at least one depressive episode in 10–12 years between surveys (Table 1). Mean (s.e.m) number of years in episode was 2.0 (0.2) and the 90th percentile was 9 years. Roughly half the respondents

Table 1. Distributions and polychoric/tetrachoric correlations among the outcomes in the Surveys 1 and 2 panels ($N = 1056$)

	Distribution		*Correlations with indications of severity*		
	Est.	*s.e.m.*	*Hospitalized*	*Suicide attempt*	*Disability*
Number of years since Survey 1 with episodes lasting 2+ weeks					
Any (%)	37.9	1.7	0.49	0.34	0.49
Number (mean)	2.0	0.2	0.46	0.38	0.49
High persistence (90th percentile (9+ years)) (%)	9.7	1.5	0.46	0.47	0.53
Number of years since NCS with episodes lasting most days					
Any (%)	16.7	1.5	0.23	0.22	0.49
Number (mean)	0.8	0.1	0.30	0.30	0.53
High chronicity (90th percentile (4+ years)) (%)	8.4	1.0	0.29	0.29	0.58
Severity					
Hospitalized for MDD since Survey 1 (%)	5.8	1.1	—	0.84	0.76
Suicide attempt since Survey 1 (%)	4.5	0.8	0.84	—	0.51
Disabled due to MDD at Survey 2 (%)	3.2	0.6	0.76	0.51	—

Abbreviations: MDD, major depressive disorder; NCS, national compensation survey.

Table 2. AUC of Survey 1 risk scores based on ML models and logistic regression models predicting Survey 2 outcomes ($N = 1056$)

	AUC of risk scores based on	
	ML models	*Logistic models*
High persistence	0.71	0.68
High chronicity	0.63	0.62
Hospitalization	0.73	0.65
Disability	0.74	0.69
Suicide attempt	0.76	0.70

Abbreviations: AUC, area under the receiver operating characteristic curve; ML, machine learning.

with episodes (16.7% (1.5) of all respondents) reported episodes lasting most days throughout one or more years, with a mean (s.e.m.) of 0.8 (0.1) and a 90th percentile of 4 such years. A strong correlation (polychoric) was found between number of years in episode and number of years with episodes lasting most days throughout the year ($r_p = 0.61$).

Hospitalization for depression in the years between Surveys 1 and 2 was reported by 5.8% (1.1) of Survey 2 respondents and attempted suicide by 4.5% (0.6). Current disability because of depression was reported by 3.2% (0.6) of Survey 2 respondents. Correlations (tetrachoric) among these three severity indicators were $r_t = 0.51–0.84$. Correlations (polychoric) between number of years in episode and the severity indicators were $r_p = 0.38–0.49$. Correlations (polychoric) between number of years in episodes lasting most days throughout the year and the severity indicators were $r_p = 0.30–0.53$.

Associations of the Survey 1 risk scores with Survey 2 outcomes

AUCs of the Survey 1 ML and logistic risk scores with Survey 2 outcomes were 0.71 and 0.68, respectively; predicting high persistence, 0.63 and 0.62, respectively; predicting high chronicity, 0.73 and 0.65, respectively; predicting hospitalization, 0.74 and 0.69, respectively; predicting disability, 0.76 and 0.70, respectively; and predicting attempted suicide. (Table 2) The AUCs of the ML scores were somewhat higher than those of the logistic regression scores for all five outcomes despite the ML scores being based on models that used only 9–13 predictors compared with 23 predictors in the logistic models and the fact that the ML-predicted values were based on multiple imputation rather than direct estimation.

Operating characteristics of the composite-imputed risk score

The 20% of Survey 1 respondents with highest ML composite-imputed predicted risk scores accounted for 38.1% of high persistence in the years between the two surveys, 34.6% of high chronicity, 40.8% of hospitalizations for depression, 55.8% of disability because of depression and 55.8% of attempted suicides. Sensitivities were substantially higher (49.7–70.7%) in the 33% of Survey 1 respondents with highest predicted risk scores (Table 3). Positive predictive values of the outcomes in the 20% of respondents with highest predicted risk scores were in the range 8.8–18.3% (that is, 1.8–3.0 times the positive predictive values in the remaining 80% of the sample), while positive predictive values were 6.3–17.5% in the 33% of respondents with highest predicted risk (that is, 1.5–2.2 times the positive predictive values in the remaining 67% of the sample).

The ML-predicted risk scores were also useful at the low end of the distribution, as seen most vividly in the fact that the 20% of Survey 1 respondents with lowest predicted risk accounted for only 0.9% of all hospitalizations and 1.5% of all attempted suicides in the 10–12 years between surveys. This means that low ML-predicted risk scores can be used as rule-outs for these outcomes (LR+ = 0.0–0.1). Sensitivities for other outcomes in this 20% of respondents with lowest predicted risk were 5.6–15.9%, while those of the 33% of respondents with lowest predicted risk were 9.7–16.7%. Positive predictive values of the outcomes in the 20% of respondents with lowest predicted risk were 0.3–6.7% (that is, 0.0–0.8 times the positive predictive values in the remaining 80% of the sample), while positive predictive values were 0.9–4.2% in the 33% of respondents with lowest predicted risk (that is, 0.3–0.5 times the positive predictive values in the remaining 67% of the sample).

DISCUSSION

Four important limitations of the WMH models should be noted before discussing the results. First, MDD was assessed with a fully structured diagnostic interview rather than a semi-structured clinical interview. Second, the models were developed in a cross-sectional sample using retrospective reports that could have been

Table 3. Sensitivity, positive predictive value and likelihood-ratio positive of Survey 1 risk scores based on ML models in the upper and lower 20 and 33% of the risk distribution predicting Survey 2 outcomes (*N* = 1056)

	High persistence		*High chronicity*		*Hospitalization*		*Disability*		*Suicide attempt*	
	Est.	*s.e.m.*	*Est.*	*s.e.m.*	*Est.*	*s.e.m.*	*Est.*	*s.e.m.*	*Est.*	*s.e.m.*
Sensitivity										
Highest 20%	38.1	4.2	34.6	7.3	40.8	6.8	55.8	6.6	55.8	6.9
Highest 33%	62.2	4.8	49.7	6.7	66.6	3.2	68.3	5.8	70.7	3.8
Lowest 33%	10.5	3.7	16.7	5.9	11.8	4.2	9.7	2.9	10.6	2.7
Lowest 20%	5.6	1.8	15.9	5.8	0.9	0.9	7.4	3.0	1.5	0.7
Positive predictive value										
Highest 20%	18.3	3.5	14.4	3.2	13.1	3.0	8.8	1.8	12.5	2.7
Highest 33%	17.5	2.8	12.2	2.3	12.5	2.6	6.3	1.5	9.3	1.7
Lowest 33%	3.1	1.3	4.2	1.5	2.3	1.1	0.9	0.5	1.4	0.7
Lowest 20%	2.7	1.0	6.7	2.4	0.3	0.3	1.2	0.7	0.3	0.2
Likelihood-ratio positive										
Highest 20%	2.1	0.4	1.8	0.4	2.2	0.5	2.9	0.6	3.0	0.5
Highest 33%	2.0	0.2	1.5	0.2	2.1	0.2	2.1	0.3	2.2	0.2
Lowest 33%	0.3	0.1	0.5	0.2	0.3	0.1	0.3	0.1	0.3	0.2
Lowest 20%	0.3	0.1	0.8	0.3	0.0	0.0	0.4	0.2	0.1	0.0

Abbreviation: ML, machine learning.

biased. Third, because the data were retrospective, predictors were limited in two important ways: the predictors for comorbid disorder did not include those with first onsets subsequent to first onset of MDD; and no predictors were included for MDD course subsequent to first onset. Both these types of predictors would normally be available to clinicians interested in evaluating differential patient risk for MDD persistence severity. Because of these limitations, we would expect the performance of the WMH models to be lower bounds on the performance of models with a more complete set of predictors. Fourth, only a limited set of ML methods was used to develop the WMH models. Because of these limitations, it would be useful to replicate and expand the model development and validation process illustrated here in prospective clinical samples using consistently administered semi-structured clinical interviews with a more complete set of predictors using additional ML algorithms (for example, naive Bayesian, random forests and support vector machines)[27] and an optimal combined suite of algorithms to maximize cross-validated prediction accuracy.[28]

Within the context of these limitations, the validation exercise reported here confirmed the predictive value of the kinds of self-report variables included in the WMH ML models over a 10–12 year follow-up period in an independent sample of the US household population. We also showed that prediction accuracy (AUC) of the ML models was consistently higher across all study outcomes (0.63–0.76) than a more conventional logistic model (0.62–0.70) despite the logistic model including 23 predictors and the ML models 9–13 predictors. This finding illustrates the value of ML methods in stabilizing predictions to avoid overfitting in a training data set (that is, the WMH sample), so as to improve prediction in independent samples.

A question can be raised how well the WMH ML composite risk score prediction accuracy compares with previous attempts to predict long-term depression persistence severity. Only a handful of relevant comparison studies exist over a follow-up period of 10+ years in samples of initially depressed patients[29,30] or community residents.[31,32] These studies were all quite small ($n = 87–424$) and none reported AUC. However, AUC can be computed *post hoc* from two of these studies. The first was a 50-year follow-up of 293 community respondents classified *post hoc* as having had baseline DSM-IV MDD, 20 of whom subsequently died by suicide.[32] A composite measure of baseline depression severity predicted subsequent suicide with 0.69 AUC compared with 0.76 for the validated AUC of the most comparable Survey 2 outcome (attempted suicide). The second comparison study followed 313 outpatients with initial diagnoses of MDD 1, 4 and 10 years after baseline and predicted persistent depression over that time period from 10 baseline depressive symptoms along with 10 baseline measures of self-concept, social function and coping. The AUC of 0.70 is quite similar to the 0.71 AUC for the most comparable Survey 2 outcome of high persistence.

In making these comparisons, it is important to remember that the AUCs in these other studies were not validated in independent samples. As noted above, AUC estimates in the Surveys 1 and 2 panels were ~10% lower than in the WMH sample. Shrinkage would be expected to be even greater in the earlier studies because of their much smaller samples than in the broadly representative WMH sample of 8261 respondents. Prediction models in the two comparison studies might consequently yield validated AUCs below 0.60 in independent samples. AUCs in that range are considered small based on conventional guidelines, while WMH ML AUCs would typically be considered moderate.[33–35]

It is noteworthy that AUC of the ML models in Surveys 1 and 2 was similar to widely used risk models in other areas of medicine.[36,37] For example, the 0.73 mean AUC of the ML models over the four Survey 2 outcomes other than high chronicity is similar to the 0.74 average AUC of the Framingham Risk Score of coronary heart disease, one of the most widely used prediction scores in medicine, across 79 different validation studies,[38] and higher than the AUCs (typically below 0.70) of models to predict the course of breast cancer.[39] Nonetheless, these AUCs are only moderate, which means that predictions based on such models could not be used to make definite rule-ins and could be used to make definite rule-outs only for risks of hospitalization and suicide attempts in the lowest 20% of the composite risk distribution. But this level of precision could be useful in defining bands of differential risk warranting variation in clinical attention. Tiered risk assessments of this sort are becoming increasingly important in other areas of medicine.[40–42]

Given that predictions based on models of the sort evaluated here would most realistically be used to help clinicians identify patients who might more profit from more intensive treatment (for example, long-term maintenance therapy), the vast majority of whom present for treatment of recurrent rather than incident episodes, an obvious future direction should be to go beyond the WMH model focus on incident episodes to develop expanded models in the Surveys 1 and 2 panels focused on recurrent episodes. Such an expansion could evaluate the incremental value of including new predictors for course of MDD between onset and time of Survey 1, secondary comorbid disorders, and other variables found to be important in previous studies of the course of depression (for example, childhood family adversities, history of traumatic stress exposure, comorbid physical disorders, social networks-support, personality). We plan to implement this kind of expansion in future work with the Surveys 1 and 2 panels.

Beyond our own work with these data, it would be useful to develop an interview schedule to assess the full set of self-report predictors found in the WMH data and in the earlier studies reviewed above to use in future depression treatment trials. Such an instrument, if administered at trial baseline, could be used as part of a principled approach to study heterogeneity of treatment effects.[43,44] An even more promising extension given the small size of most depression treatment trials might be to administer this same instrument to a large observational sample of patients beginning depression treatment, follow these patients to assess treatment response and analyze these data to develop a robust model predicting heterogeneity of treatment effects. In addition to providing an a priori representation of predicted treatment response for use in subsequent controlled trials, such a model could be useful in targeting depressed patients with high risk of treatment resistance at the beginning of treatment who might warrant the substantial investment currently being made in large pragmatic trials to determine the value of expensive baseline biomarker assessments in guiding depression treatment targeting.[8,9] It would also be valuable in this context to evaluate the 'incremental' value of promising biomarkers to prediction over and above the level of prediction accuracy achieved in a model based only on baseline self-reports.[45]

Risk stratification data from a large observational study of this sort could also be analyzed using an extension of the innovative statistical approaches recently developed to study comparative effectiveness in observational studies.[46] The potential value of such an approach is supported both by evidence that treatment effect size estimates in appropriately analyzed observational studies are comparable with those in controlled trials[47] and by the existence of numerous replicated predictors of heterogeneity of depression treatment effects in existing trials.[14,15,48,49] The use of an expansion of our model in this way would address two important problems in previous research on heterogeneity of depression treatment effects: the small sample sizes of depression treatment trials;[50,51] and the fact that most such trials assess only a small number of potential treatment effect modifiers, thus providing no principled basis for using pooling across trials to develop the kind of fine-grained multivariate models of heterogeneity of treatment effects that will eventually be needed to guide personalized depression treatment planning.[44] The results presented in the current report, while only taking a first step in this direction, provide strong support for the potential value of this possible extension.

CONFLICT OF INTEREST

RCK has been a consultant for Hoffman La Roche, Johnson & Johnson Wellness and Prevention and Sonofi-Aventis Groupe; has served on an advisory board for Lake Nona Institute; and owns stock in DataStat. AAN has been a consultant for Abbott Laboratories, American Psychiatric Association, Appliance Computing (Mindsite), Basliea, Brain Cells, Brandeis University, Bristol-Myers Squibb, Clintara, Corcept, Dey Pharmaceuticals, Dainippon Sumitomo (now Sunovion), Eli Lilly and Company, EpiQ, L.P./Mylan, Forest, Genaissance, Genentech, GlaxoSmithKline, Hoffman La Roche, Infomedic, Lundbeck, Janssen Pharmaceutica, Jazz Pharmaceuticals, Medavante, Merck, Methylation Sciences, Naurex, Novartis, PamLabs, Pfizer, PGx Health, Ridge Diagnostics Shire, Schering-Plough, Somerset, Sunovion, Takeda Pharmaceuticals, Targacept and Teva; consulted through the MGH Clinical Trials Network and Institute (CTNI) for Astra Zeneca, Brain Cells, Dianippon Sumitomo/Sepracor, Johnson and Johnson, Labopharm, Merck, Methylation Science, Novartis, PGx Health, Shire, Schering-Plough, Targacept and Takeda/Lundbeck Pharmaceuticals; had grant/research support from the American Foundation for Suicide Prevention, AHRQ, Brain and Behavior Research Foundation, Bristol-Myers Squibb, Cederroth, Cephalon, Cyberonics, Elan, Eli Lilly, Forest, GlaxoSmithKline, Janssen Pharmaceutica, Lichtwer Pharma, Marriott Foundation, Mylan, NIMH, PamLabs, PCORI, Pfizer Pharmaceuticals, Shire, Stanley Foundation, Takeda and Wyeth-Ayerst; received honoraria from Belvoir Publishing, University of Texas Southwestern Dallas, Brandeis University, Bristol-Myers Squibb, Hillside Hospital, American Drug Utilization Review, American Society for Clinical Psychopharmacology, Baystate Medical Center, Columbia University, CRICO, Dartmouth Medical School, Health New England, Harold Grinspoon Charitable Foundation, IMEDEX, International Society for Bipolar Disorder, Israel Society for Biological Psychiatry, Johns Hopkins University, MJ Consulting, New York State, Medscape, MBL Publishing, MGH Psychiatry Academy, National Association of Continuing Education, Physicians Postgraduate Press, SUNY Buffalo, University of Wisconsin, University of Pisa, University of Michigan, University of Miami, University of Wisconsin at Madison, APSARD, ISBD, SciMed, Slack Publishing and Wolters Klower Publishing; owns stock in Appliance Computing (MindSite), Brain Cells, Medavante; and owns the following copyrights: Clinical Positive Affect Scale and the MGH Structured Clinical Interview for the Montgomery Asberg Depression Scale exclusively licensed to the MGH Clinical Trials Network and Institute (CTNI). MAW is an employee of Janssen Pharmaceuticals. HMvL, KJW, RMB, LAB, TC, DDE, IH, JL, PdJ, MVP, AJR, NAS, RAS and AMZ declare no conflict of interest.

ACKNOWLEDGMENTS

PdJ was supported by a VICI grant (no: 91812607) from the Netherlands Organization for Scientific Research (NWO-ZonMW). The NCS data collection was supported by the National Institute of Mental Health (NIMH; R01MH46376). The NCS-2 data collection was supported by the National Institute on Drug Abuse (NIDA; R01DA012058). Data analysis for this paper was additionally supported by NIMH grants R01MH070884 and U01MH060220, with supplemental support from the Substance Abuse and Mental Health Services Administration (SAMHSA), the Robert Wood Johnson Foundation (RWJF; Grant 044780) and the John W. Alden Trust. The NCS-2 is carried out in conjunction with the World Health Organization World Mental Health (WMH) Survey Initiative. We thank the staff of the WMH Data Collection and Data Analysis Coordination Centres for assistance with instrumentation, fieldwork and consultation on data analysis. These activities were supported by the NIMH (R01MH070884), the John D and Catherine T MacArthur Foundation, the Pfizer Foundation, the US Public Health Service (R13MH066849, R01MH069864 and R01DA016558), the Fogarty International Center (FIRCA R03TW006481), the Pan American Health Organization, Eli Lilly and Company, Ortho-McNeil Pharmaceutical, GlaxoSmithKline and Bristol-Myers Squibb.

DISCLAIMER

The views, opinions and/or findings contained in this article are those of the authors and should not be construed as an official Department of Veterans Affairs position, policy or decision unless so designated by other documentation, or the views of any of the sponsoring organizations, agencies or the US Government.

REFERENCES

1 Altshuler LL, Cohen LS, Moline ML, Kahn DA, Carpenter D, Docherty JP *et al.* Treatment of depression in women: a summary of the expert consensus guidelines. *J Psychiatr Pract* 2001; **7**: 185–208.
2 Hetrick SE, Simmons M, Thompson A, Parker AG. What are specialist mental health clinician attitudes to guideline recommendations for the treatment of depression in young people? *Aust N Z J Psychiatry* 2011; **45**: 993–1001.
3 Kuiper S, McLean L, Fritz K, Lampe L, Malhi GS. Getting depression clinical practice guidelines right: time for change? *Acta Psychiatr Scand Suppl* 2013; **444**: 24–30.
4 Perlis RH. Use of treatment guidelines in clinical decision making in bipolar disorder: a pilot survey of clinicians. *Curr Med Res Opin* 2007; **23**: 467–475.
5 van Loo HM, de Jonge P, Romeijn JW, Kessler RC, Schoevers RA. Data-driven subtypes of major depressive disorder: a systematic review. *BMC Med* 2012; **10**: 156.
6 Vrieze E, Demyttenaere K, Bruffaerts R, Hermans D, Pizzagalli DA, Sienaert P *et al.* Dimensions in major depressive disorder and their relevance for treatment outcome. *J Affect Disord* 2014; **155**: 35–41.
7 Hasler G, Northoff G. Discovering imaging endophenotypes for major depression. *Mol Psychiatry* 2011; **16**: 604–619.

8 Kennedy SH, Downar J, Evans KR, Feilotter H, Lam RW, MacQueen GM *et al.* The Canadian Biomarker Integration Network in Depression (CAN-BIND): advances in response prediction. *Curr Pharm Des* 2012; **18**: 5976–5989.

9 Uher R, Perroud N, Ng MY, Hauser J, Henigsberg N, Maier W *et al.* Genome-wide pharmacogenetics of antidepressant response in the GENDEP project. *Am J Psychiatry* 2010; **167**: 555–564.

10 James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning: With Applications in R*. Springer: New York, 2013.

11 van der Laan MJ, Rose S. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer: New York, 2011.

12 Chang YJ, Chen LJ, Chung KP, Lai MS. Risk groups defined by Recursive Partitioning Analysis of patients with colorectal adenocarcinoma treated with colorectal resection. *BMC Med Res Methodol* 2012; **12**: 2.

13 Chao ST, Koyfman SA, Woody N, Angelov L, Soeder SL, Reddy CA *et al.* Recursive partitioning analysis index is predictive for overall survival in patients undergoing spine stereotactic body radiation therapy for spinal metastases. *Int J Radiat Oncol Biol Phys* 2012; **82**: 1738–1743.

14 Nelson JC, Zhang Q, Deberdt W, Marangell LB, Karamustafalioglu O, Lipkovich IA. Predictors of remission with placebo using an integrated study database from patients with major depressive disorder. *Curr Med Res Opin* 2012; **28**: 325–334.

15 Riedel M, Moller HJ, Obermeier M, Adli M, Bauer M, Kronmuller K *et al.* Clinical predictors of response and remission in inpatients with depressive syndromes. *J Affect Disord* 2011; **133**: 137–149.

16 van Loo HM, Cai T, Gruber MJ, Li J, de Jonge P, Petukhova M *et al.* Major depressive disorder subtypes to predict long-term course. *Depress Anxiety* 2014; **31**: 765–777.

17 Wardenaar KJ, van Loo HM, Cai T, Fava M, Gruber MJ, Li J *et al.* The effects of co-morbidity in defining major depression subtypes associated with long-term course and severity. *Psychol Med* 2014; **44**: 3289–3302.

18 Kessler RC, McGonagle KA, Zhao S, Nelson CB, Hughes M, Eshleman S *et al.* Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States. Results from the National Comorbidity Survey. *Arch Gen Psychiatry* 1994; **51**: 8–19.

19 Kessler RC, Merikangas KR, Berglund P, Eaton WW, Koretz DS, Walters EE. Mild disorders should not be eliminated from the DSM-V. *Arch Gen Psychiatry* 2003; **60**: 1117–1122.

20 Kessler RC, Wittchen HU, Abelson JM, McGonagle KA, Schwarz N, Kendler KS *et al.* Methodological studies of the Composite International Diagnostic Interview (CIDI) in the US National Comorbidity Survey. *Int J Methods Psychiatr Res* 1998; **7**: 33–55.

21 Spitzer RL, Williams JB, Gibbon M, First MB. The Structured Clinical Interview for DSM-III-R (SCID). I: history, rationale, and description. *Arch Gen Psychiatry* 1992; **49**: 624–629.

22 Endicott J, Andreasen N, Spitzer RL. *Family History Research Diagnostic Criteria (FHRDC)*. Biometrics Research, New York State Psychiatric Institute: New York, 1978.

23 Therneau T, Atkinson B. *An Introduction to Recursive Partitioning Using the RPART Routines*. Mayo Foundation: Rochester, MN, 2015.

24 Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* 2010; **33**: 1–22.

25 SAS Institute Inc. SAS/STAT software. 9.2 for Unix edn. SAS Institute Inc.: Cary, NC, 2009.

26 Research Triangle Institute. *SUDAAN: Professional Software for Survey Data Analysis*, 9th edn Research Triangle Institute: Research Triangle Park: NC, 2004.

27 Marsland S. *Machine Learning: An Algorithmic Perspective* 2nd (edn). Taylor & Francis: Boca Raton, FL, 2015.

28 van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol* 2007; **6**: Article 25.

29 Klein DN, Shankman SA, Rose S. Dysthymic disorder and double depression: prediction of 10-year course trajectories and outcomes. *J Psychiatr Res* 2008; **42**: 408–415.

30 Moos RH, Cronkite RC. Symptom-based predictors of a 10-year chronic course of treated depression. *J Nerv Ment Dis* 1999; **187**: 360–368.

31 Angst J, Gamma A, Rossler W, Ajdacic V, Klein DN. Childhood adversity and chronicity of mood disorders. *Eur Arch Psychiatry Clin Neurosci* 2011; **261**: 21–27.

32 Bradvik L, Mattisson C, Bogren M, Nettelbladt P. Long-term suicide risk of depression in the Lundby cohort 1947–1997—severity and gender. *Acta Psychiatr Scand* 2008; **117**: 185–191.

33 Rice ME, Harris GT. Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law Hum Behav* 2005; **29**: 615–620.

34 Singh JP, Desmarais SL, Van Dorn RA. Measurement of predictive validity in violence risk assessment studies: a second-order systematic review. *Behav Sci Law* 2013; **31**: 55–73.

35 Sjostedt G, Grann M. Risk assessment: what is being predicted by actuarial prediction instruments? *Int J Forensic Ment Health* 2002; **1**: 179–183.

36 Echouffo-Tcheugui JB, Kengne AP. Comparative performance of diabetes-specific and general population-based cardiovascular risk assessment models in people with diabetes mellitus. *Diabetes Metab* 2013; **39**: 389–396.

37 Siontis GC, Tzoulaki I, Siontis KC, Ioannidis JP. Comparisons of established risk prediction models for cardiovascular disease: systematic review. *BMJ* 2012; **344**: e3318.

38 Tzoulaki I, Liberopoulos G, Ioannidis JP. Assessment of claims of improved prediction beyond the Framingham risk score. *JAMA* 2009; **302**: 2345–2352.

39 Anothaisintawee T, Teerawattananon Y, Wiratkapun C, Kasamesup V, Thakkinstian A. Risk prediction models of breast cancer: a systematic review of model performances. *Breast Cancer Res Treat* 2012; **133**: 1–10.

40 Haas LR, Takahashi PY, Shah ND, Stroebel RJ, Bernard ME, Finnie DM *et al.* Risk-stratification methods for identifying patients for care coordination. *Am J Manag Care* 2013; **19**: 725–732.

41 Morris JN, Howard EP, Steel K, Schreiber R, Fries BE, Lipsitz LA *et al.* Predicting risk of hospital and emergency department use for home care elderly persons through a secondary analysis of cross-national data. *BMC Health Serv Res* 2014; **14**: 519.

42 Williams LM, Rush AJ, Koslow SH, Wisniewski SR, Cooper NJ, Nemeroff CB *et al.* International Study to Predict Optimized Treatment for Depression (iSPOT-D), a randomized clinical trial: rationale and protocol. *Trials* 2011; **12**: 4.

43 Burke JF, Hayward RA, Nelson JP, Kent DM. Using internally developed risk models to assess heterogeneity in treatment effects in clinical trials. *Circ Cardiovasc Qual Outcomes* 2014; **7**: 163–169.

44 Willke RJ, Zheng Z, Subedi P, Althin R, Mullins CD. From concepts, theory, and evidence of heterogeneity of treatment effects to methodological approaches: a primer. *BMC Med Res Methodol* 2012; **12**: 185.

45 Li C, Lu Y. Evaluating the improvement in diagnostic utility from adding new predictors. *Biom J* 2010; **52**: 417–435.

46 Neugebauer R, Schmittdiel JA, van der Laan MJ. Targeted learning in real-world comparative effectiveness research with time-varying interventions. *Stat Med* 2014; **33**: 2480–2520.

47 Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. *Cochrane Database Syst Rev* 2014; (4): MR000034.

48 Jain FA, Hunter AM, Brooks JO 3rd, Leuchter AF. Predictive socioeconomic and clinical profiles of antidepressant response and remission. *Depress Anxiety* 2013; **30**: 624–630.

49 Perlis RH. A clinical risk stratification tool for predicting treatment resistance in major depressive disorder. *Biol Psychiatry* 2013; **74**: 7–14.

50 Cuijpers P, Reynolds CF 3rd, Donker T, Li J, Andersson G, Beekman A. Personalized treatment of adult depression: medication, psychotherapy, or both? A systematic review. *Depress Anxiety* 2012; **29**: 855–864.

51 Simon GE, Perlis RH. Personalized medicine for depression: can we match patients with treatments? *Am J Psychiatry* 2010; **167**: 1445–1455.