Title

Prediction of remission in obsessive compulsive disorder using a novel machine learning strategy

Journal

International Journal of Methods in Psychiatric Research, 24(2)

ISSN

1049-8931

Authors

Askland, Kathleen D
Garnaat, Sarah
Sibrava, Nicholas J
et al.

Publication Date

2015-06-01

DOI

10.1002/mpr.1463

Peer reviewed

International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. 24(2): 156–169 (2015)
Published online in Wiley Online Library
(wileyonlinelibrary.com) **DOI:** 10.1002/mpr.1463

Prediction of remission in obsessive compulsive disorder using a novel machine learning strategy

KATHLEEN D. ASKLAND,[1] SARAH GARNAAT,[1] NICHOLAS J. SIBRAVA,[2] CHRISTINA L. BOISSEAU,[1] DAVID STRONG,[3] MARIA MANCEBO,[1] BENJAMIN GREENBERG,[1] STEVE RASMUSSEN[1] & JANE EISEN[1]

1 Department of Psychiatry and Human Behavior, Butler Hospital/Warren Alpert School of Medicine, Brown University, Providence, RI, USA
2 Department of Psychology, Baruch College – The City University of New York, New York, USA
3 Department of Family and Preventive Medicine, University of California, San Diego, CA, USA

Key words
obsessive compulsive disorder, statistics, risk factors

Correspondence
Kathleen Askland, Department of Psychiatry, University of Toronto, Waypoint Centre for Mental Health Care, 500 Church Street, Penetanguishene, ON L9M 1Z1, Ontario, Canada.
Telephone: (705) 549-3181
Fax: (705) 549-5238
Email: k.askland@gmail.com

Received 17 September 2014; revised 19 December 2014; accepted 23 February 2015

Abstract

The study objective was to apply machine learning methodologies to identify predictors of remission in a longitudinal sample of 296 adults with a primary diagnosis of obsessive compulsive disorder (OCD). Random Forests is an ensemble machine learning algorithm that has been successfully applied to large-scale data analysis across vast biomedical disciplines, though rarely in psychiatric research or for application to longitudinal data. When provided with 795 raw and composite scores primarily from baseline measures, Random Forest regression prediction explained 50.8% (5000-run average, 95% bootstrap confidence interval [CI]: 50.3–51.3%) of the variance in proportion of time spent remitted. Machine performance improved when only the most predictive 24 items were used in a reduced analysis. Consistently high-ranked predictors of longitudinal remission included Yale–Brown Obsessive Compulsive Scale (Y-BOCS) items, NEO items and subscale scores, Y-BOCS symptom checklist cleaning/washing compulsion score, and several self-report items from social adjustment scales. Random Forest classification was able to distinguish participants according to binary remission outcomes with an error rate of 24.6% (95% bootstrap CI: 22.9–26.2%). Our results suggest that clinically-useful prediction of remission may not require an extensive battery of measures. Rather, a small set of assessment items may efficiently distinguish high- and lower-risk patients and inform clinical decision-making.

Introduction

Longitudinal research is essential to better characterize the course and outcomes of psychiatric illness. Large, observational longitudinal studies have been conducted across diagnostic categories, including, for example, depression (Klein *et al.*, 2006), anxiety (Keller, 2006; Weisberg *et al.*, 2012), personality disorders (Gunderson *et al.*, 2000), and obsessive compulsive disorder (OCD) (Pinto *et al.*, 2006). Within OCD, specifically, these studies have led to better understanding of the course of this disorder. However, although findings regarding risk factors linked to long-term outcomes have often been inconsistent, factors associated with reduced likelihood of remission include earlier age of onset (Eisen *et al.*, 2010), being unmarried (Steketee *et al.*, 1999; Marcks *et al.*, 2011), greater OCD severity at intake (Steketee *et al.*, 1999; Eisen *et al.*, 2010; Fineberg *et al.*, 2013), comorbid depression (Marcks *et al.*, 2011) and anxiety (Fineberg *et al.*, 2013) disorders, serotonin reuptake inhibitor treatment at intake (Marcks *et al.*, 2011), being male (Eisen *et al.*, 2010) or older at intake (Eisen *et al.*, 2010), and longer duration of illness (Eisen *et al.*, 2013; Fineberg *et al.*, 2013). Comorbid schizotypal personality disorder and a higher Yale–Brown Obsessive Compulsive Scale (Y-BOCS) compulsion subscore have also been associated with worse course (Catapano *et al.*, 2006). While these naturalistic longitudinal studies have contributed significantly to our understanding of expected course and risk factors, they also have limitations.

One such limitation arises because analytic approaches used in psychiatric studies with longitudinal data are typically based on traditional hypothesis-testing methods. Use of theory as a guiding principle is essential in hypothesis testing; however, reliance on a presupposed theory necessarily constrains the scope of an analysis to testing pre-specified models. The most important limitation of conventional parametric models, then, is that they work only if the true model is correctly specified at the outset. For this reason, such models are often not suitable for the analysis of multidimensional, heterogeneous data since the assumptions of such models often do not apply or cannot easily be verified. Additionally, traditional parametric methods are often sensitive to outliers and unlikely to accommodate complicated decision boundaries (Malley *et al.*, 2011). Such limitations result in a tendency to select variables based on finding marginal effects in univariate testing (Holzinger *et al.*, 2015) and, in order to minimize the number of parameters to be estimated, rely on summary scores rather than individual items from clinical measures. Importantly, not all important variables will have marginal effects in univariate analyses and reduction of assessment measures to summary scores can result in information loss. Individual items may be more important in predicting the outcome than the summary scores and may be differentially involved in the complex interactions that are implicitly interrogated through machine learning procedures.

Complementary analytic methods can overcome some of these limitations. In their three-part text series on *Machine Learning in Medicine* published in mid-2013, Cleophas and Zwinderman (Cleophas, 2013) note that, while commonly used in the social and applied sciences, machine learning (generally the domain of computer scientists) is virtually unused in clinical research. However, one machine learning method, Support Vector Machines (SVM), has been widely applied to neuroimaging data for diagnostic classification in clinical populations (see reviews: Orru *et al.*, 2012; Hoexter *et al.*, 2013). Most recently, support vector regression was used to predict OCD severity from brain structural magnetic resonance imaging (MRI) data (Hoexter *et al.*, 2013). Random Forests (RF), another machine learning algorithm, has been less widely used in clinical medicine and has seen very little application in neuroimaging or other psychiatric research (see exceptions: Tektonidou *et al.*, 2011; Arnold *et al.*, 2012; Gibbons *et al.*, 2013; Clark *et al.*, 2014). To our knowledge, RF has not yet been applied to predict remission outcomes in longitudinal psychiatric samples nor to analyze data in anxiety disorder samples.

The RF algorithm is an alternative way to examine predictors of course in psychiatric disorders with direct implications for evaluating prognosis in clinical settings. RF is a data-driven method. Sifting through large multidimensional datasets, these algorithms first "learn" which variables best predict the outcome of interest (e.g. illness remission) and then test the learned "rules" on independent portions of the data. Such approaches allow investigators to simultaneously consider very large numbers of possible predictors and to not be bound by the distributional assumptions of traditional statistical approaches (e.g. general linear models). Moreover, these algorithms can be used to identify and rank the most important variables predicting any outcome of interest and, therefore, may be particularly useful in informing clinical practice, particularly regarding complex psychiatric phenomena.

The goals of this project were two-fold. First, to use a novel machine learning approach to analyze existing data from an ongoing prospective, longitudinal study of the course of OCD to examine predictors of remission. The second goal was to present a general model for the application of RF analyses to psychiatric and multidimensional

datasets. Our hope is that this may demonstrate an alternative approach to identifying predictors of clinical outcomes that can be utilized more broadly in psychiatric research.

Methods

Participants

The Brown Longitudinal Obsessive Compulsive Study (BLOCS) is an ongoing prospective, longitudinal follow-up study of OCD. Inclusion criteria for adults were a primary diagnosis of OCD (defined as the disorder participants considered the biggest problem overall across their lifetime), age ≥ 19, and having sought treatment within five years prior to study enrollment. Participants were recruited between July 2001 and February 2006 from multiple psychiatric treatment settings in Rhode Island/Massachusetts including a hospital-based OCD specialty clinic, a private psychiatric hospital inpatient unit, two community mental health centers, a general outpatient psychiatric practice, and three private psychotherapy practices. A total of 653 individuals were screened for inclusion, with 121 failing to meet initial inclusion criteria, 127 refusing participation, six failing to meet inclusion criteria following the diagnostic intake assessment, and an additional four withdrawing after intake, for a total initial enrollment of 325 adults. The Butler Hospital and Brown University Institutional Review Boards approved the study. After complete description of the study to the subjects, written informed consent was obtained, and participants were compensated for their time at each assessment point.

The present report focuses on 296 adult participants (91% of adult intake sample). The remaining 9% were excluded because there were no outcome data for these participants. Supplementary Material Table S1 shows demographic and clinical characteristics of the sample. Importantly, some subjects met inclusion criteria (e.g. lifetime diagnosis of OCD), but had a subclinical (8.11%) or mild (18.92%) Y-BOCS score at study intake.

Assessment

A detailed description of the assessment procedures including training and reliability is presented elsewhere (Pinto *et al.*, 2006). Briefly, intake diagnoses were established using the Structured Clinical Interview for DSM-IV Axis I Disorders – Patient Edition (SCID) (First *et al.*, 1996) and the Structured Clinical Interview for DSM-IV Axis II Disorders (SCID-II) (First *et al.*, 1997). OCD-related symptoms, severity, beliefs, insight and history were assessed using the Y-BOCS and Symptom Checklist (Goodman *et al.*, 1989a, 1989b), the Brown Assessment of Beliefs Scale (Eisen *et al.*, 1998), and the OCD Database (Rasmussen *et al.*, 1993). Medication and psychosocial treatment history was assessed using the Behavioral Therapy Inventory (BTI) (Mancebo *et al.*, 2011) and the Longitudinal Interval Follow-up Evaluation (LIFE) (Keller *et al.*, 1987). In addition to the information provided by the SCID, specific comorbidity was assessed using the Modified Hamilton Rating Scale for Depression (Miller *et al.*, 1985), the Hopkins Tic Inventory (Walkup *et al.*, 1992). Functional measures included the Quality of Life Enjoyment and Satisfaction Questionnaire (Endicott *et al.*, 1993), Medical Outcomes Survey 36-Item Short-Form Health Survey (Ware and Sherbourne, 1992), Social and Occupational Functioning Assessment Scale (Goldman *et al.*, 1992), Global Assessment of Functioning (Bodlund *et al.*, 1994), the Social Adjustment Scale – Self Report (Weissman and Bothwell, 1976) as well as the psychosocial functioning section of the LIFE. The University of Rhode Island Change Assessment (DiClemente *et al.*, 2004) and the Contemplation Ladder (Biener and Abrams, 1991), were administered to assess readiness for change. Finally, the NEO Five Factor Inventory (Costa and McCrae, 1992) and the Positive and Negative Affect Schedule (Watson *et al.*, 1988), first administered in January 2007, were used to assess personality traits and positive/negative affect respectively.

Participant's family history of OCD and related disorders was obtained using a semi-structured family history screening form. Probands were asked to consider each of their first-degree relatives and, in the case of OCD for example, report their impressions of the relatives obsessions, compulsions, the time occupied by their symptoms, as well as distress and impairment they perceived to be attributable to OCD. Interviewers made best-estimate diagnoses for the first-degree relatives of the probands based on DSM-IV criteria using the information provided by participants. The Secondary Diagnosis Rating form was used to record participants' subjective impressions of their next most troubling diagnoses, following OCD (which had to be primary for inclusion in the study). Participants were asked to rank up to three other conditions that they would most like to be rid of. Subjects were also administered the OCD Intake Supplement, a seven-item measure designed to capture additional information regarding a number of domains not adequately measured elsewhere in the intake battery. This included an assessment of whether or not participants could identify feared consequences associated with their primary obsessions, a subjective rating of response to cognitive-behavioral therapy (if applicable), an item assessing medication use during the cognitive

Int. J. Methods Psychiatr. Res. 24(2): 156–169 (2015). DOI: 10.1002/mpr

behavioral therapy (CBT) trial (if applicable), a summary of obsessive compulsive personality disorder (OCPD) symptoms endorsed, age at first treatment, and presence of an impulse control disorder at intake. Finally, participants were administered a modified version of the Yale Greater New Haven Health Survey (Myers, 1980) consisting of one question regarding the participants subjective rating of their overall emotional health.

Following baseline assessment, participants were interviewed annually using the LIFE, a widely used semi-structured instrument in longitudinal studies of psychiatric disorders (Keller *et al.*, 1987; Warshaw *et al.*, 1994; Skodol *et al.*, 2005). Weekly clinical ratings were based on the presence or absence of DSM-IV criteria and severity for OCD as assessed with weekly psychiatric status ratings (PSRs) on the LIFE. For the BLOCS study, the OCD PSR had six points based on OCD symptom severity and functional impairment. A PSR of four or greater represents full DSM-IV criteria for OCD, with six being the most severe and impaired. A PSR of three is assigned when OCD symptoms are present but not impairing and symptoms occupy less than an hour daily. A PSR of two indicates minimal symptoms and no impairment; a PSR of one indicates no symptoms.

Data preprocessing

All data preprocessing and statistical analyses were performed using the Revolution R Enterprise 7.2 statistical software package for 64-bit Windows, which runs R version 3.0.3 (R Core Team, 2014).

Deriving outcome variables

A current limitation of the RF method is that it cannot handle within-person correlations as are typical in longitudinal and repeated measures data. Thus, our choice of outcomes to evaluate was guided by this algorithmic limitation. Two outcome variables – one continuous, one binary – were derived for each subject using weekly OCD PSRs, collected throughout the course of enrollment, which was up to 12 years. The continuous outcome was calculated as the percentage of weeks of study enrollment during which the subject experienced at least a partial illness remission, i.e. had PSR$\leq$3 ("Percent Time Remitted"). Consistent with prior studies (Warshaw *et al.*, 1994; Bruce *et al.*, 2005; Eisen *et al.*, 2010), the binary outcome for the study ("Ever Remit") was whether or not the participant had at least one period of eight consecutive weeks or more during which s/he had only sub-threshold symptoms (PSR$\leq$3) during his/her period of study enrollment.

Extracting measures for study inclusion

All measures administered at baseline, and three measures that were administered once-only (i.e. administered to subjects once during study enrollment but at variable times across subjects depending upon when the subject enrolled in the study), were extracted and screened for possible inclusion in the analysis. Additionally, as we were interested in understanding the role of pharmacologic treatment on remission rates and no baseline data were available containing dosage information, we also extracted a set of variables indicating the number of weeks in the first year of study enrollment during which each SSRI/SNRI (selective serotonin reuptake inhibitor/serotonin and norepinephrine reuptake inhibitor, i.e. fluoxetine, fluvoxamine, sertraline paroxetine, citalopram, escitalopram, clomipramine, venlafaxine and duloxetine) dose was adequate. All variables extracted from baseline, once-only and pharmacologic measures were subject to filtering, as described next.

Filtering features

After screening variables for missingness and measures that did not vary across participants, 781 original features pulled from baseline, once-only and pharmacologic measures were included. No variables with $>$33% missingness were included. The vast majority of excluded variables were from measures added to the assessment battery later in the study. To these were added 17 indicator variables for each level of Y-BOCS checklist items 38 (category of obsession most like to get rid of) and 60 (category of compulsion most like to get rid of), for a total of 801 potential input features. Finally, three features were removed because the R randomForest (Liaw and Wiener, 2002) implementation we employed could handle features with $>$32 categories, leaving a final count of 795 features for inclusion in the full analyses.

Missing values processing

For all 795 retained features, missing values were transformed to non-missing values using the R (R Core Team, 2014) randomForest (Liaw and Wiener, 2002) imputation procedure 'rfImpute'. The algorithm initially imputes missing data points using another randomForest procedure, na.roughfix, by which missing numeric values are replaced with medians and missing factor (categorical) variables are replaced by the mode of all non-missing values. The proximity matrix derived within the implementation of rfImpute function is used to update the imputation over a user-defined number of iterations

(in our case, 13). This updating helps to account for any originally imputed values that were inconsistent with the proximities calculated for the non-missing data for each subject.

Data analysis

Random Forests (RF)

The basic RF algorithm (Breiman and Cutler, 2001) is a non-parametric ensemble learning method and is considered one of the most accurate general-purpose learning techniques available (Biau, 2012). The procedure is consistent and its rate of convergence depends only on the number of strong features (i.e. good predictors) and not on the number of noise variables present in the dataset (Biau, 2012). No distributional assumptions or other statistical premises are made in the algorithm concerning the features or the participants (Malley *et al.*, 2012). Well-described in previous work (Breiman, 2001; Biau *et al.*, 2008; Malley *et al.*, 2011; Biau, 2012; Breiman and Cutler, 2001), the basic steps of RF are: for a dataset of size N: (1) draw a random bootstrap sample, with replacement, from the full set of N subjects (i.e. the "in-bag" or "training" samples), and keep track of the cases not selected (i.e. the out-of-bag [OOB] or "testing" data); (2) using the in-bag sample, grow an unpruned classification or regression tree as follows: (2a) at each node, randomly select a small number of features (m_{try}) out of all ($M = 795$) features; (2b) using the m_{try} selected features, find the feature and its optimal cutpoint that minimizes the mean square error (regression) or classification error (classification) in the training data; (2c) proceed to subsequent nodes and repeat (2a) and (2b) until a user-specified halting rule, *nodesize*, is satisfied; (3) drop all OOB cases down the tree and track the terminal node for each such case; classify each OOB case by majority vote (classification), or assign the average outcome value (regression), in its respective terminal node based on the in-bag sample values; (4) repeat steps (1–3) many times, generating *ntree* trees; (5) aggregate the OOB predictions over all *ntree* trees (i.e. majority vote for classification, average for regression) to generate the error estimates for the random forest. Steps 1–5 produce a set of trees, the random forest, which constitutes the "model."

See Supplementary Material for a description of the dataset balancing procedure used for the classification analyses; the ten-fold cross-validation procedures used to validate the RF machine performance estimates; and correlated features analysis used to assess the potential for bias due to correlated features.

Parameter settings

RF has three user-defined tuning parameters (m_{try}, ntree, nodesize). We set m_{try} = sqrt (p), the default, for classification and $m_{\text{try}} = 1/3$ (p) for regression. *ntree* was set to 2000 and we employed traced output and error plots to ensure that we ran past the point of test error convergence. *nodesize*, was set to 10% of the sample size (N), for all analyses. See Supplementary Material for background and rationale for parameter settings.

Full and reduced RF analyses

RF was used to complete two full and two reduced analyses. RF analyses were performed in R using the randomForest (Liaw and Wiener, 2002) package version 4.6-7. The full analyses each used all 795 baseline/first-administration predictors ("full feature set"). The first full analysis, used regression RF, under our continuous outcome measure, "Percent Time Remitted". The second full analysis invoked the classification RF procedure, under our binary outcome measure, "Ever Remit". Our reduced analyses consisted of the same analytic procedures and outcomes as in full analyses, but included a reduced feature set comprising those from the full analyses with the most consistently highly-ranked variable importance scores, as described later. To insure robustness and stability of results (Strobl *et al.*, 2009), each of the four analyses were run 5000 times, each time employing a new random seed.

Variable importance rankings were also derived within the R randomForest procedure using the Breiman–Cutler permutation strategy (Breiman and Cutler, 2001), which has good performance and comparative efficacy (Molinaro *et al.*, 2011). This procedure estimates, for each variable, the degree to which random permutation of its values decreases prediction accuracy of the machine, while retaining original values of all other variables. The 5000 independent RF runs enabled us to assess the stability of variable rankings and to produce a high-confidence set of the most important (i.e. predictive) variables.

There is no gold standard method for determining the threshold that best differentiates signal from noise (Holzinger *et al.*, 2015). Expert consensus (Strobl *et al.*, 2009) suggests that it is best not to interpret or compare importance scores but to rely on the relative rankings of the predictors. As such, our approach provides an interpretive framework by identifying the set of predictors whose importance scores were consistently (i.e. in 100% of 5000 runs) above a standard threshold (Strobl *et al.*, 2009; Holzinger *et al.*, 2015) used for filtering out noise, thereby identifying the predictors that most consistently

Int. J. Methods Psychiatr. Res. 24(2): 156–169 (2015). DOI: 10.1002/mpr

influence the outcome under study. As noted in Holzinger *et al.* (2015), this threshold, the absolute value of the lowest negative variable importance score, provides a reliable estimate of the variance in variable importance scores in null data as variables with no effect should be symmetrically and randomly distributed around zero (Strobl *et al.*, 2009; Holzinger *et al.*, 2015).The subset of features that exceeded this threshold in 100% of 5000 runs in the full analyses were selected for inclusion in the reduced analyses.

Machine performance measures: error estimates

It is important to emphasize that the OOB samples are used to derive all performance measures. For regression, performance measures are OOB mean squared error (MSE) and percent variance explained (RSQ). For classification, we report overall error rate (i.e. percentage of class predictions that were incorrect, in either direction). The collection of *ntree* trees (i.e. the "forest") constitutes the learning machine or classifier (aka "the model") and the overall error estimate for the forest is the average OOB error across *ntree* trees.

Confidence intervals (CIs) for the error estimates were computed from the empirical bootstrap distribution of the error estimates obtained from the 5000 randomly-seeded RF runs using the bootstrap percentile interval method (Chernick, 1999; Young *et al.*, 2008). Bootstrap percentile (2.5 and 97.5) CIs are based upon the quantiles of the original bootstrap distribution of error estimates. For comparison, non-parametric bootstrap CI estimates were also calculated using smean.cl.boot procedure from R 'Hmisc' package (Harrell *et al.*, 2014), which categorically produced narrower CIs, probably due to the use of a second layer of bootstrapping employed in the procedure. These are not reported.

Proximity measures

Proximity measures, another standard output of RF, were used to construct multidimensional scaling (MDS) plots. For each pair of subjects, their proximity is the fraction of trees in which the pair falls within the same terminal node. After each tree is grown, all samples are put down the tree. Each time two cases appear in the same terminal node, their proximity measure increases by one. Once the proximity matrix is calculated, the proximity values are normalized by dividing by the number of trees. MDS plots are a representation of the proximity matrix and are used for visualization and subset identification. MDS plots enable visualization of the underlying proximity structure between objects or cases. MDS assigns each observation to a specific location in a conceptual space (usually two or three dimensional space, hence dim1 and dim2 on plot axes in figures) such that the distances between points in the space match the given proximities as closely as possible. MDS is similar to factor analysis, but MDS does not rely on common assumptions (linearity, multivariate normality, etc.). The only assumption of MDS is that the number of dimensions cannot exceed the number of objects minus one.

Results

Full feature analysis

When provided with the full 795 feature set, the RF regression procedure produced a 5000-run average MSE of 799.7122 (bootstrap 95-percentile CI: 791.6799–807.8362) and explained approximately 50.78% (95% CI: 50.28–51.28%) of the variance in proportion of time spent remitted (RSQ). Twenty-four features were above threshold in 100% of runs (Table 1), i.e. were highly-consistent predictors of outcome.

Our classification procedure using the full feature set was able to distinguish participants according to a binary outcome ("Ever Remitted" versus "Never Remitted"), with a 5000-run average error rate of 24.63% (95% CI: 22.85–26.23%). Twenty-six features were above threshold in 100% of runs (Table 2).

Sixteen predictors were above threshold in 100% of both the classification and regression runs. MDS plots, derived from single RF runs for each outcome, allow visualization of the proximities among participants under continuous (Figures 1 and 2) and binary (Supplementary Material Figure S1) outcome models.

Reduced feature analysis

Using the 24 high-confidence features to predict the continuous outcome, the 5000-run average MSE was 730.79 (95% CI: 724.86–737.12%) and the mean RSQ was 55.02% (95% CI: 54.63–55.39%). This CI does not overlap that from the full feature set model and so is at least as good in predicting response as the full feature set. Using the 26 high-confidence features to predict the binary outcome, the 5000-run average OOB error rate was 23.82% (bootstrap CI: 22.10–25.45%). This CI closely overlaps that from the full feature set model (22.85–26.23%), suggesting the full feature set is no better than a highly-reduced set. Variable importance plots, again based on a single RF run for each outcome, are shown in Supplementary Material Figures S2 and S3, respectively.

Table 1. High-confidence predictors of *Percent Time Remitted*. Listing of predictors that exceeded threshold in 100% of 5000 runs under the continuous outcome, *Percent Time Remitted*

Predictor name	Predictor description	5000-Run mean variable importance score	Confidence interval (±2 standard errors)
Y-BOCS #6	Y-BOCS item – Time spent performing compulsive behaviors, Mo. 00	19.325	(19.311,19.34)
Y-BOCS Compulsion	Y-BOCS overall (summary) compulsion score	17.877	(17.864,17.891)
Y-BOCS #7	Y-BOCS item – Interference due to compulsive behaviors	17.407	(17.393,17.421)
Y-BOCS #1	Y-BOCS item – Time occupied by obsessive thoughts, Mo. 00	17.379	(17.364,17.394)
SCID OCD (binary)	SCID criteria for OCD within past month – dichotomized	16.544	(16.529,16.558)
SCID OCD	SCID criteria for OCD within past month	16.433	(16.419,16.448)
NEO Neuroticism *t*-score	NEO – *t*-score for neuroticism subscore	15.614	(15.598,15.63)
NEO Neuroticism Raw score	NEO – neuroticism subscore	15.219	(15.203,15.236)
GAF	LIFE assessment – GAF at baseline	14.958	(14.944,14.973)
Y-BOCS Total	Y-BOCS overall (summary) score	14.719	(14.704,14.734)
Occupational Impairment	LIFE assessment – employment (impairment)	12.020	(12.005,12.034)
Y-BOCS #6b	Y-BOCS item – time spent free of compulsive behaviors	11.919	(11.905,11.934)
Y-BOCS Checklist – Worst	Y-BOCS checklist – compulsion type most like to get rid of	11.110	(11.086,11.135)
URICA #24	URICA item – “I hope that someone here will have some good advice for me.”	10.910	(10.894,10.925)
SCID OCD severity	SCID OCD severity rating	9.997	(9.983,10.012)
Y-BOCS Checklist – Cleaning Sum	Y-BOCS checklist – sum of cleaning/washing compulsions scores (q. 39–42)	8.713	(8.698,8.727)
Y-BOCS Checklst Cleaning Mean	Y-BOCS checklist – average of cleaning/washing compulsions scores (q. 39–42)	8.676	(8.662,8.69)
NEO #26	NEO item – “Sometimes I feel completely worthless.”	8.667	(8.651,8.682)
NEO #41	NEO item – “Too often, when things go wrong, I get discouraged and feel like giving up.”	8.507	(8.49,8.524)
Y-BOCS Checklist – Re-read/write	Y-BOCS checklist – re-reading or re-writing compulsions present	8.273	(8.26,8.287)
Y-BOCS Checklist – Cleaning	Y-BOCS checklist – cleaning (household/inanimate objects) compulsions present	7.921	(7.907,7.935)
NEO #24	NEO item – “I tend to be cynical and skeptical of others’ intentions.”	7.285	(7.271,7.299)
Social Adjustment (friends)	Social Adjustment Self-report (SAS) – able to talk about feelings and problems.	6.842	(6.822,6.862)
Social Functioning Assessment	Social Functioning Assessment Scale (1–100 rating)	6.228	(6.213,6.243)

Int. J. Methods Psychiatr. Res. 24(2): 156–169 (2015). DOI: 10.1002/mpr

Table 2. High-confidence predictors of *Ever Remit*. Listing of predictors that exceeded threshold in 100% of 5000 runs under the binary outcome, *Ever Remit*

Predictor name	Predictor description	5000-Run mean variable importance score	Confidence interval (±2 standard errors)
Y-BOCS #6	Y-BOCS item – time spent performing compulsive behaviors, Mo. 00	19.325	(19.311,19.34)
Y-BOCS Compulsion	Y-BOCS overall (summary) compulsion score	17.877	(17.864,17.891)
Y-BOCS #7	Y-BOCS item – interference due to compulsive behaviors	17.407	(17.393,17.421)
Y-BOCS #1	Y-BOCS item – time occupied by obsessive thoughts, Mo. 00	17.379	(17.364,17.394)
SCID OCD (binary)	SCID criteria for OCD within past month – dichotomized	16.544	(16.529,16.558)
SCID OCD	SCID criteria for OCD within past month	16.433	(16.419,16.448)
NEO Neuroticism *t*-score	NEO – *t*-score for neuroticism subscore	15.614	(15.598,15.63)
NEO Neuroticism Raw score	NEO – neuroticism subscore	15.219	(15.203,15.236)
GAF	LIFE assessment – GAF at baseline	14.958	(14.944,14.973)
Y-BOCS Total	Y-BOCS overall (summary) score	14.719	(14.704,14.734)
Occupational Impairment	LIFE assessment – employment (impairment)	12.020	(12.005,12.034)
Y-BOCS #6b	Y-BOCS item – time spent free of compulsive behaviors	11.919	(11.905,11.934)
Y-BOCS Checklist – Worst	Y-BOCS checklist – compulsion type most like to get rid of	11.110	(11.086,11.135)
URICA #24	URICA item – “I hope that someone here will have some good advice for me.”	10.910	(10.894,10.925)
SCID OCD severity	SCID OCD severity rating	9.997	(9.983,10.012)
Y-BOCS Checklist – Cleaning Sum	Y-BOCS checklist – sum of cleaning/washing compulsions scores (q. 39–42)	8.713	(8.698,8.727)
Y-BOCS Checklst Cleaning Mean	Y-BOCS checklist – average of cleaning/washing compulsions scores (q. 39–42)	8.676	(8.662,8.69)
NEO #26	NEO item – “Sometimes I feel completely worthless.”	8.667	(8.651,8.682)
NEO #41	NEO item – “Too often, when things go wrong, I get discouraged and feel like giving up.”	8.507	(8.49,8.524)
Y-BOCS Checklist – Re-read/write	Y-BOCS checklist – re-reading or re-writing compulsions present	8.273	(8.26,8.287)
Y-BOCS Checklist – Cleaning	Y-BOCS checklist – cleaning (household/inanimate objects) compulsions present	7.921	(7.907,7.935)
NEO #24	NEO item – “I tend to be cynical and skeptical of others’ intentions.”	7.285	(7.271,7.299)
Social Adjustment (friends)	Social Adjustment Self-report (SAS) – able to talk about feelings and problems.	6.842	(6.822,6.862)
Social Functioning Assessment	Social Functioning Assessment Scale (1–100 rating)	6.228	(6.213,6.243)

Int. J. Methods Psychiatr. Res. 24(2): 156–169 (2015). DOI: 10.1002/mpr

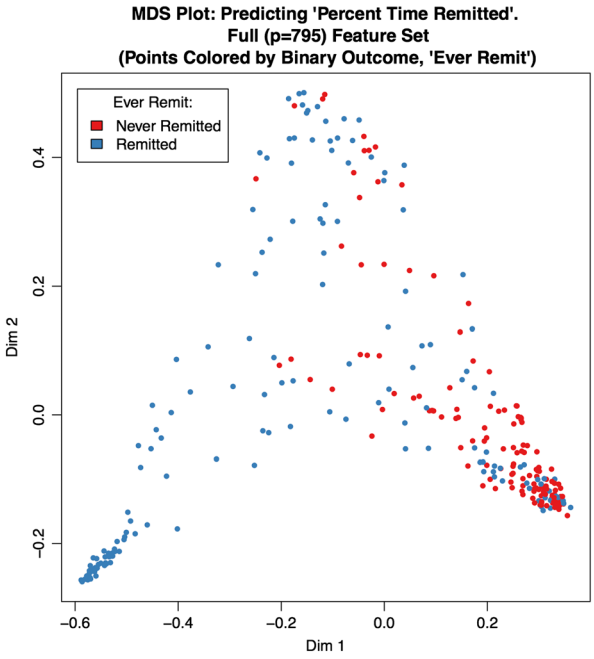

Figure 1. Multidimensional scaling (MDS) plot: predicting *Percent Time Remitted*. Full ($p = 795$) Feature Set (points colored by binary outcome, *Ever Remit*). Sample MDS plot derived from a single random forest (RF) run under full feature analysis predicting the continuous outcome, *Percent Time Remitted*. For visualization purposes, the points (each of which corresponds to a single subject) are colored according to the binary outcome, *Ever Remit*.

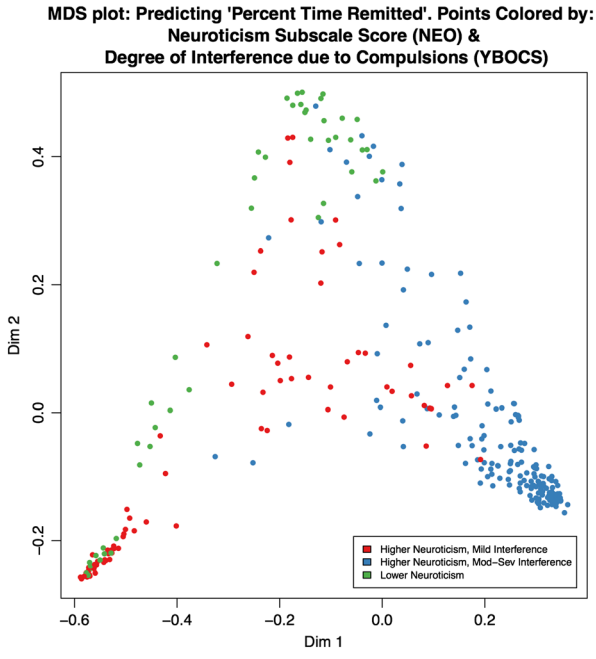

Figure 2. Multidimensional scaling (MDS) plot: predicting *Percent Time Remitted*. Points colored by Neuroticism Subscale Score (NEO) and Degree of Interference due to Compulsions (Y-BOCS). Sample MDS plot derived from a single random forest (RF) run under full feature analysis predicting the continuous outcome, *Percent Time Remitted*. This plot contains the identical points as in Figure 1. However, in this plot, the points are colored according to the subject's scores on two high-ranked predictor items: a binary partition of the neuroticism subscale score ("lower neuroticism" corresponds to a neuroticism subscale score ≤ 50; "higher neuroticism" indicates > 50); a binary partition of the Y-BOCS item #7, Interference due to compulsive behaviors ("Mild interference" corresponds to score ≤ 1, "Mod-Severe interference" corresponds to a score > 1).

For illustrative purposes, a single tree was constructed under the reduced feature set model for *Percent Time Remitted* (Figure 3) and *Ever Remit* (Supplementary Material Figure S4) using the R package reprtree (Dasgupta, 2014), which implements the concept of representative trees from ensembles of tree-based machines (Banerjee *et al.*, 2012). While extraction of the data for a single tree from RF output is straightforward, visualization options are limited.

Discussion

RF analyses provided an unbiased ranking of a small number of variables that were consistently predictive of OCD remission. We used a large pool of 795 measured variables in a sample of 296 adults with a primary OCD diagnosis. RF identified 24 "high-priority" features that improved prediction of our continuous outcome: percent of weeks with sub-threshold OCD symptom severity. Prediction was superior to that using the full set of 795 features. This feature set substantially overlapped with the set of 26 high priority features identified under a classification model. When RF was run using the reduced feature set for the classification procedure, the performance was essentially unchanged from that achieved using the full feature set.

Among the most consistently high-ranked features were those associated with compulsive behavior at time of study entrance, specifically: time spent performing compulsions, impairment due to compulsions, and the overall compulsion subscale score from the Y-BOCS. This suggests that baseline time and interference due to compulsions is a critical marker of prognosis.

This is especially notable when contrasted against the myriad of predictors previously identified in the literature on OCD course. Other than the Cleaning/Washing compulsion subtype, we did not find symptom subtype to be

Int. J. Methods Psychiatr. Res. 24(2): 156–169 (2015). DOI: 10.1002/mpr

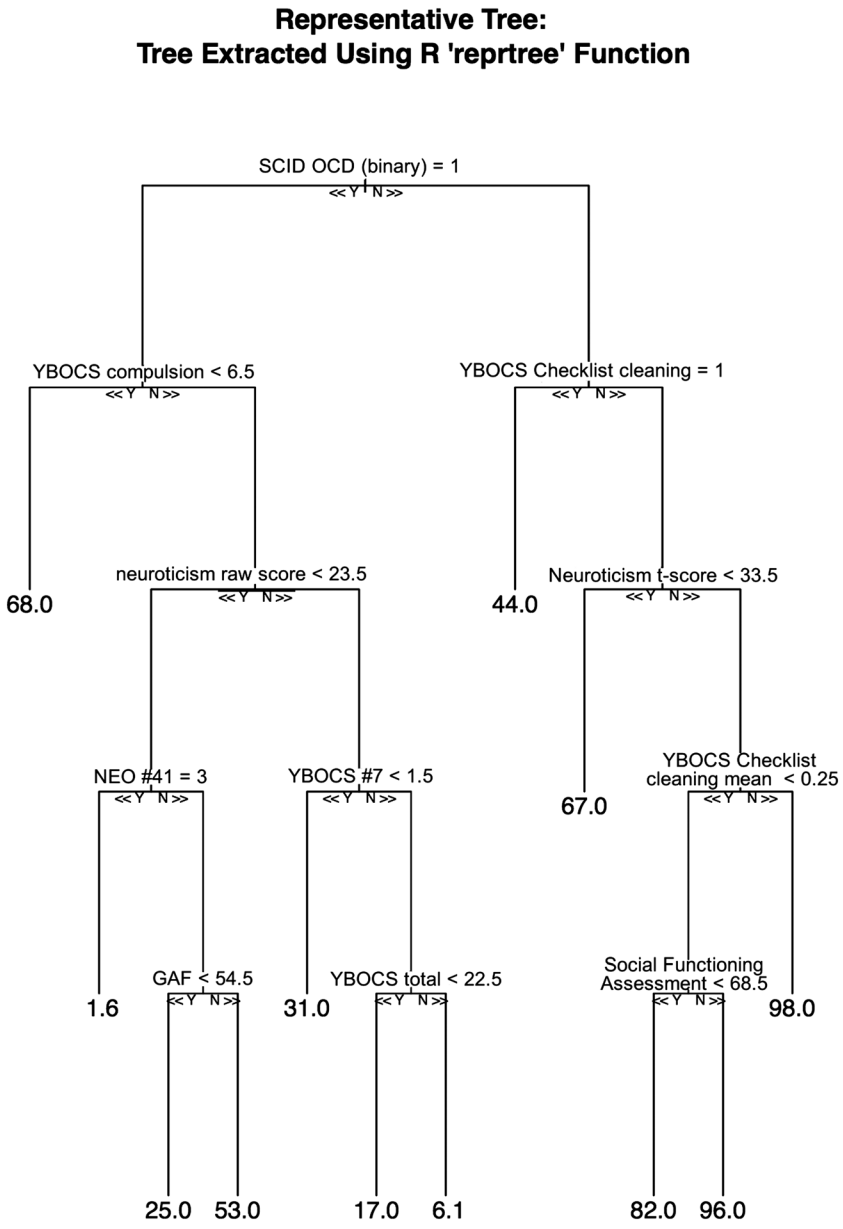

Figure 3. "Representative Tree": predicting *Percent Time Remitted* using 24 best predictors. This representative tree models the continuous outcome, *Percent Time Remitted*, and the 24 high-priority features and was extracted from a single random forest (RF) run (ntree = 5000) using the R "reprtree" (Dasgupta, 2014) package. This package implements the concept of representative trees from ensembles of tree-based machines on the basis of several tree distance metrics (Banerjee *et al.*, 2012). Each node contains the variable selected for splitting at that node and the value on which it was split represented by a mathematical condition. The cases split to the left daughter node are those for which the condition was met; those in the right node are those for which the condition was not met. The numeric values displayed at each terminal node are the mean values of the outcome variable for the subjects residing in that terminal node.

predictive of remission. We also did not find age of onset, duration of illness, gender, marital status, depression, anxiety or serotonin reuptake inhibitor (SRI) treatment, all of which were included in our analysis and have been identified in prior research, as predictive of remission. The importance of compulsive behavior as a predictor of outcome is consistent with previous literature, including investigations in both treated (e.g. Ravizza *et al.*, 1995;

Catapano *et al.*, 2006) and untreated or undertreated samples (e.g. Degonda *et al.*, 1993; Skoog and Skoog, 1999). However, while previous investigations suggested that a predominance of compulsive behavior is predictive of poor outcome, they did not provide the degree of item-level specificity derived via the RF algorithm.

This finding provides clarity for clinicians with regard to which variables should serve as critical foci for clinical monitoring, and suggests that even as other variables fluctuate over time, severity of compulsions may represent a marker of poor prognosis. This information allows clinicians to make more informed treatment decisions, can influence medication and dosage changes, and speaks to the need for behavioral treatments focused on reducing compulsions (i.e. exposure and response prevention) even if, for example, some improvement is seen in overall distress or comorbid illness.

In addition to findings related to compulsions, other consistent predictors included other components of the Y-BOCS and measures of severity, as well as indices of psychosocial functioning, many of which are relatively consistent with previous course and outcome findings. Interestingly, the personality trait of neuroticism also emerged as a consistently high-priority feature identified by the RF algorithm. The association between neuroticism and remission has not been widely reported in previous studies of OCD course or treatment outcome. This personality dimension may represent an additional important domain for clinicians and researchers to explore, as neuroticism can be readily assessed with a number of reliable measures, and may prove to be an important marker of prognosis and progress.

The current investigation has a number of important strengths, as well as some key limitations. This investigation is one of the first to employ the RF analytic strategy in a large, longitudinal psychiatric sample. RF has several advantages over traditional analytic methods typically employed in similar studies. RF allows for a much more comprehensive examination of potential predictors of remission than traditionally employed analytic strategies in large datasets, which typically require normal distribution of data and constrain the number of predictors relative to sample size. Additionally, alternative approaches to analyzing large datasets often use composite or total scores or apply dimension reduction techniques (e.g. principal components analysis). However, the disadvantage to this is that the original input variables are aggregated into a sum or a mean score or are projected into a reduced set of components such that their individual effect is no longer identifiable. By allowing for full consideration of all 795 features in the current sample, including item-level analysis, our results allow for the distillation of a large number of variables into a relatively small number of unbiased high-priority predictors and provides a succinct set of factors for clinicians and researchers to consider. An additional strength is the large, well-characterized naturalistic longitudinal clinical sample of adults with OCD.

Limitations of this study include the use of a treatment-seeking, rather than epidemiological sample, thus potentially limiting the overall representativeness of the findings for all adults diagnosed with OCD. In this observational study, participants received naturalistic treatments which were not controlled in any way; however, this design allows for a unique perspective on illness course in "real world" patients. Additionally, the sample was predominately Caucasian (97.3%), further limiting the generalizability in more racially and ethnically diverse populations.

Another potential limitation of our results was introduced by the inclusion of the NEO, a once-only measure, which was administered to subjects between year 2 and 10 of study enrollment. A later administration time may not necessarily systematically affect the relationship between NEO scores and our continuous outcome measure (proportion of time spent remitted), but our analysis cannot unambiguously discern the direction of the relationship between the neuroticism score and "*Ever Remit*", thus making this a key limitation to be considered when interpreting this finding.

Overall, our machine performance was reasonable, though not ideal. This likely reflects the imprecision inherent in phenotypic and symptomatic measures of disease burden, including the derived outcome measures, as the predictive capacity of any learning machine is only as good as the informative features with which it is provided. Another avenue of exploration may have been to compare results obtained with different machine learning algorithms. While this was not undertaken here, previous empirical comparisons of machine learning methods found RF to be consistently among the top performers (Caruana and Niculescu-Mizil, 2006; Caruana *et al.*, 2008; Maroco *et al.*, 2011). Finally, although RF allowed for the inclusion of many observed variables without the sample size, distribution, or power constraints of traditional analytic approaches, it is important to point out that other variables which were not administered to this sample may prove to be important predictors of remission, and future studies with diverse variable sets should be explored in order to discover other potentially valuable predictors.

Future directions

The search for reliable, clinically useful predictors of illness course and markers of prognosis is an important

Int. J. Methods Psychiatr. Res. 24(2): 156–169 (2015). DOI: 10.1002/mpr

mission for mental health research. In recent years, advances in statistical analyses and computer technology have allowed for increasingly advanced exploration of data in the search for such markers, and among these advances, machine learning techniques such as RF can offer important insights into complex and multifaceted research data in order to advance both clinical and research priorities.

The current study is among the first to apply the machine learning/RF analytic strategy to a large, longitudinal clinical sample, and provides a viable model for investigators seeking to discover predictors of remission. Though not available in R, at least one study (Karpievitch *et al.*, 2009) has described an implementation of random forests (in C++) that can handle cluster-correlated data from biological experiments. Such methods may be modifiable to handle the similar structure of repeated measures data in clinical research and may improve upon the methods presented here for longitudinal data.

Identifying the most critical set of predictors for clinicians has valuable implications for practice, and provides a succinct set of treatment targets for clinicians and researchers alike. Results will need to be replicated in an independent sample. Nonetheless, findings from the current study provide preliminary guidance for clinicians and researchers with regard to key targets in OCD. Future research that leverages these cutting edge analytic approaches in other large clinical samples may provide valuable clarity in the search for critical markers of progress for patients suffering from a range of psychiatric illnesses.

Declaration of interest statement

Financial support for this study came from Dr Askland's NIMH K08 (MH085810-05), Dr Rasmussen's NIMH R01 BLOCS grant (MH060218) and Dr Mancebo's K23 award (MH091032).

All authors report no competing interests.

References

Arnold S.E., Xie S.X., Leung Y.Y., Wang L.S., Kling M.A., Han X., Kim E.J., Wolk D.A., Bennett D.A., Chen-Plotkin A., Grossman M., Hu W., Lee V.M., Mackin R.S., Trojanowski J.Q., Wilson R.S., Shaw L.M. (2012) Plasma biomarkers of depressive symptoms in older adults. *Translational Psychiatry*, **2**(1), e65. DOI: 10.1038/tp.2011.63

Banerjee M., Ding Y., Noone A.M. (2012) Identifying representative trees from ensembles. *Statistics in Medicine*, **31**(15), 1601–1616. DOI: 10.1002/sim.4492

Biau G. (2012) Analysis of a Random Forests model. *Journal of Machine Learning Research*, **13**(1), 1063–1095.

Biau G., Devroye L., Lugosi G. (2008) Consistency of Random Forests and other averaging classifiers. *Journal of Machine Learning Research*, **9**, 2015–2033.

Biener L., Abrams D.B. (1991) The contemplation ladder: Validation of a measure of readiness to consider smoking cessation. *Health Psychology*, **10**(5), 360–365.

Bodlund O., Kullgren G., Ekselius L., Lindstrom E., von Knorring L. (1994) Axis V – Global Assessment of Functioning Scale. Evaluation of a self-report version. *Acta Psychiatrica Scandinavica*, **90**(5), 342–347.

Breiman L. (2001) Random Forests, Berkeley, CA, Statistics Department, University of California.

Breiman L., Cutler A. (2001) Random Forests. http://www.stat.berkeley.edu/~breiman/RandomForests/ [20 Jannuary 2013]

Bruce S.E., Yonkers K.A., Otto M.W., Eisen J.L., Weisberg R.B., Pagano M., Shea T., Keller M.B. (2005) Influence of psychiatric comorbidity on recovery and recurrence in generalized anxiety disorder, social phobia, and panic disorder: a 12-year prospective study. *American Journal of Psychiatry*, **162**(6), 1179–1187. DOI: 10.1176/appi.ajp.162.6.1179

Caruana R., Niculescu-Mizil A. (2006) An Empirical Comparison of Supervised Learning Algorithms. Paper presented at the International Conference on Machine Learning, Pittsburg, PA.

Caruana R., Karampatziakis N., Yessenalina A. (2008) An empirical evaluation of supervised learning in high dimensions. Paper presented at the International Conference on Machine Learning, New York.

Catapano F., Perris F., Masella M., Rossano F., Cigliano M., Magliano L., Maj M. (2006) Obsessive-compulsive disorder: a 3-year prospective follow-up study of patients treated with serotonin reuptake inhibitors OCD follow-up study. *Journal of Psychiatric Research*, **40**(6), 502–510. DOI: 10.1016/j.jpsychires.2005.04.010

Chernick M.R. (1999) Bootstrap Methods: A Practitioner's Guide, New York, Wiley.

Clark D.G., Kapur P., Geldmacher D.S., Brockington J.C., Harrell L., DeRamus T.P., Blanton P.D., Lokken K., Nicholas A.P., Marson D.C. (2014) Latent information in fluency lists predicts functional decline in persons at risk for Alzheimer disease. *Cortex*, **55**, 202–218. DOI: 10.1016/j.cortex.2013.12.013

Cleophas T.J. (2013) Machine Learning in Medicine, New York, Springer.

Costa P.T., McCrae R.R. (1992) Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) Professional Manual, Odessa, FL, Psychological Assessment Resources.

Dasgupta A. (2014) reprtree: Representative trees from ensembles. R package version 0.6. Windows binaries provided by Abhijit Dasgupta, PhD. https://github.com/araastat/reprtree.git [7 December 2014].

Degonda M., Wyss M., AngstJ. (1993) The Zurich Study. XVIII. Obsessive-compulsive disorders and syndromes in the general population. *European Archives of Psychiatry and Clinical Neuroscience*, **243**(1), 16–22.

DiClemente C.C., Schlundt D., Gemmell L. (2004) Readiness and stages of change in addiction treatment. *American Journal on Addictions*, **13**(2), 103–119. DOI: 10.1080/10550490490435777

Eisen J.L., Phillips K.A., Baer L., Beer D.A., Atala K.D., Rasmussen S.A. (1998) The Brown Assessment of Beliefs Scale: Reliability and validity. *American Journal of Psychiatry*, **155**(1), 102–108.

Eisen J.L., Pinto A., Mancebo M.C., Dyck I.R., Orlando M.E., Rasmussen S.A. (2010) A 2-year prospective follow-up study of the

course of obsessive-compulsive disorder. *Journal of Clinical Psychiatry*, **71**(8), 1033–1039. DOI: 10.4088/Jcp.08m04806blu

Eisen J.L., Sibrava N.J., Boisseau C.L., Mancebo M.C., Stout R.L., Pinto A., Rasmussen S.A. (2013) Five-year course of obsessive-compulsive disorder: predictors of remission and relapse. *Journal of Clinical Psychiatry*, **74**(3), 233–239. DOI: 10.4088/Icp.12m07657

Endicott J., Nee J., Harrison W., Blumenthal R. (1993) Quality-of-life enjoyment and satisfaction questionnaire – a new measure. *Psychopharmacology Bulletin*, **29**(2), 321–326.

Fineberg N.A., Hengartner M.P., Bergbaum C., Gale T., Rossler W., Angst J. (2013) Remission of obsessive-compulsive disorders and syndromes; evidence from a prospective community cohort study over 30 years. *International Journal of Psychiatry in Clinical Practice*, **17**(3), 179–187. DOI: 10.3109/13651501.2013.777744

First M.B., Spitzer R.L., Gibbon M., Williams J.B. W. (1996) Structured Clinical Interview for DSM-IV Axis I Disorders – Patient Edition (SCID/IP Version 2.0), New York, Biometrics Research Department, New York State Psychiatric Institute.

First M.B., Gibbon M., Spitzer R.L., Williams J.B. W., Benjamin L.S. (1997) Structured Clinical Interview for DSM-IV Axis II Personality Disorders (SCID-II), Washington, DC, American Psychiatric Press.

Gibbons R.D., Hooker G., Finkelman M.D., Weiss D.J., Pilkonis P.A., Frank E., Moore T., Kupfer D.J. (2013) The computerized adaptive diagnostic test for major depressive disorder (CAD-MDD): a screening tool for depression. *Journal of Clinical Psychiatry*, **74**(7), 669–674. DOI: 10.4088/JCP.12m08338

Goldman H.H., Skodol A.E., Lave T.R. (1992) Revising axis V for DSM-IV: a review of measures of social functioning. *American Journal of Psychiatry*, **149**(9), 1148–1156.

Goodman W.K., Price L.H., Rasmussen S.A., Mazure C., Delgado P., Heninger G.R., Charney D.S. (1989a) The Yale–Brown Obsessive Compulsive Scale. II. Validity. *Archives of General Psychiatry*, **46**(11), 1012–1016.

Goodman W.K., Price L.H., Rasmussen S.A., Mazure C., Fleischmann R.L., Hill C.L., Heninger G.R., Charney D.S. (1989b) The Yale–Brown Obsessive Compulsive Scale. I. Development, use, and reliability. *Archives of General Psychiatry*, **46**(11), 1006–1011.

Gunderson J.G., Shea M.T., Skodol A.E., McGlashan T.H., Morey L.C., Stout R.L., Zanarini M.C., Grilo C.M., Oldham J.M., Keller M.B. (2000) The Collaborative Longitudinal Personality Disorders Study: development, aims, design, and sample characteristics. *Journal of Personality Disorders*, **14**(4), 300–315.

Harrell F.E. Jr, with contributions from Charles Dupont and many others. (2014) Hmisc: Harrell Miscellaneous. R package version 3.14-3. http://CRAN.R-project.org/package=Hmisc [accessed December, 2014].

Hoexter M.Q., Miguel E.C., Diniz J.B., Shavitt R. G., Busatto G.F., Sato J.R. (2013) Predicting obsessive-compulsive disorder severity combining neuroimaging and machine learning methods. *Journal of Affective Disorders*, **150**(3), 1213–1216. DOI: 10.1016/j.jad.2013.05.041

Holzinger E.M., Szymczk S., Dasgupta A., Malley J., Li Q., Bailey-Wilson J.E. (Jan 4–8, 2015) Variable Selection Method for the Identification of Epistatic Models. Paper presented at the Pacific Symposium on Biocomputing (PSB), Maui, HI.

Karpievitch Y.V., Hill E.G., Leclerc A.P., Dabney A. R., Almeida J.S. (2009) An introspective comparison of random forest-based classifiers for the analysis of cluster-correlated data by way of RF++. *PLoS One*, **4**(9), e7087. DOI: 10.1371/journal.pone.0007087

Keller M.B. (2006) Social anxiety disorder clinical course and outcome: review of Harvard/Brown Anxiety Research Project (HARP) findings. *Journal of Clinical Psychiatry*, **67** (Supplemant 12), 14–19.

Keller M.B., Lavori P.W., Friedman B., Nielsen E., Endicott J., McDonald-Scott P., Andreasen N. C. (1987) The longitudinal interval follow-up evaluation. A comprehensive method for assessing outcome in prospective longitudinal studies. *Archives of General Psychiatry*, **44**(6), 540–548.

Klein D.N., Shankman S.A., Rose S. (2006) Ten-year prospective follow-up study of the naturalistic course of dysthymic disorder and double depression. *American Journal of Psychiatry*, **163**(5), 872–880. DOI: 10.1176/appi.ajp.163.5.872

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

Malley J.D., Malley K.G., Pajevic S. (2011) Statistical Learning for Biomedical Data, Cambridge, Cambridge University Press.

Malley J.D., Kruppa J., Dasgupta A., Malley K.G., Ziegler A. (2012) Probability machines: consistent probability estimation using nonparametric learning machines. *Methods of Information in Medicine*, **51**(1), 74–81. DOI: 10.3414/ME00-01-0052

Mancebo M.C., Eisen J.L. Sibrava N.J., Dyck I.R., Rasmussen S.A. (2011) Patient utilization of cognitive-behavioral therapy for OCD. *Behavior Therapy*, **42**(3), 399–412. DOI: 10.1016/j.beth.2010.10.002

Marcks B.A., Weisberg R.B., Dyck I., Keller M.B. (2011) Longitudinal course of obsessive-compulsive disorder in patients with anxiety disorders: a 15-year prospective follow-up study. *Comprehensive Psychiatry*, **52**(6), 670–677. DOI: 10.1016/j.comppsych.2011.01.001

Maroco J., Silva D., Rodrigues A., Guerreiro M., Santana I., de Mendonca A. (2011) Data mining methods in the prediction of Dementia: a real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC Research Notes*, **4**, 299. DOI: 10.1186/1756-0500-4-299

Miller I.W., Bishop S., Norman W.H., Maddever H. (1985) The Modified Hamilton Rating Scale for Depression: reliability and validity. *Psychiatry Research*, **14**(2), 131–142.

Molinaro A.M., Carriero N., Bjornson R., Hartge P., Rothman N., Chatterjee N. (2011) Power of data mining methods to detect genetic associations and interactions. *Human Heredity*, **72**(2), 85–97. [pii] 000330579 DOI: 10.1159/000330579

Myers J.K. (1980) Yale Greater New Haven Health Survey, New Haven, CT, Yale University.

Orru G., Pettersson-Yeo W., Marquand A.F., Sartori G., Mechelli A. (2012) Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neuroscience & Biobehavioral Reviews*, **36**(4), 1140–1152. DOI: 10.1016/j.neubiorev.2012.01.004

Pinto A., Mancebo M.C., Eisen J.L., Pagano M.E., Rasmussen S.A. (2006) The Brown longitudinal obsessive compulsive study: clinical features and symptoms of the sample at intake. *Journal of Clinical Psychiatry*, **67**(5), 703–711.

R Core Team. (2014) R: A Language and Environment for Statistical Computing, Vienna, R Core Team. http://www.R-project.org/ [accessed November-December, 2014].

Rasmussen S.A., Eisen J.L., Pato M.T. (1993) Current issues in the pharmacologic management of obsessive compulsive disorder. *Journal of Clinical Psychiatry*, **54**(Supplement), 4–9.

Ravizza L., Barzega G., Bellino S., Bogetto F., Maina G. (1995) Predictors of drug treatment response in obsessive-compulsive disorder. *Journal of Clinical Psychiatry*, **56**(8), 368–373.

Skodol A.E., Gunderson J.G., Shea M.T., McGlashan T.H., Morey L.C., Sanislow C.A., Bender D.S., Grilo C.M., Zanarini M.C., Yen S., Pagano M.E., Stout R.L. (2005) The Collaborative Longitudinal Personality Disorders Study (CLPS): overview and implications. *Journal of Personality Disorders*, **19**(5), 487–504. DOI: 10.1521/pedi.2005.19.5.487

Skoog G., Skoog I. (1999) A 40-year follow-up of patients with obsessive-compulsive disorder [see comments]. *Archives of General Psychiatry*, **56**(2), 121–127.

Steketee G., Eisen J., Dyck I., Warshaw M., Rasmussen S. (1999) Predictors of course in obsessive-compulsive disorder. *Psychiatry Research*, **89**(3), 229–238.

Strobl C., Malley J., Tutz G. (2009) An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, **14**(4), 323–348. DOI: 10.1037/a0016973

Tektonidou M.G., Dasgupta A., Ward M.M. (2011) Suicidal ideation among adults with arthritis: prevalence and subgroups at highest risk. Data from the 2007–2008 National Health and Nutrition Examination Survey. *Arthritis Care and Research (Hoboken)*, **63**(9), 1322–1333- DOI: 10.1002/acr.20516

Walkup J.T., Rosenberg L.A., Brown J., Singer H.S. (1992) The validity of instruments measuring tic severity in Tourette's syndrome. *Journal of the American Academy of Child & Adolescent Psychiatry*, **31**(3), 472–477. DOI: 10.1097/00004583-199205000-00013

Ware J.E. Jr, Sherbourne C.D. (1992) The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical Care*, **30**(6), 473–483.

Warshaw M.G., Keller M.B., Stout R.L. (1994) Reliability and validity of the longitudinal interval follow-up evaluation for assessing outcome of anxiety disorders. *Journal of Psychiatric Research*, **28**(6), 531–545.

Watson D., Clark L.A., Tellegen A. (1988) Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology*, **54**(6), 1063–1070.

Weisberg R.B., Beard C., Dyck I., Keller M.B. (2012) The Harvard/Brown Anxiety Research Project-Phase II (HARP-II): rationale, methods, and features of the sample at intake. *Joural of Anxiety Disorders*, **26**(4), 532–543. DOI: 10.1016/j.janxdis.2012.02.007

Weissman M.M., Bothwell S. (1976) Assessment of social adjustment by patient self-report. *Archives of General Psychiatry*, **33**(9), 1111–1115.

Young T.M., Perhac D.G., Guess F.M., Leon R.V. (2008) Bootstrap confidence intervals for percentiles of reliability data for wood-plastic composites. *Forest Products Journal*, **58**(11), 106–114.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.