Psychotherapy Research, 2021
Vol. 31, No. 1, 52–62, https://doi.org/10.1080/10503307.2020.1839140

EMPIRICAL PAPER

Identifying CBT non-response among OCD outpatients: A machine-learning approach

KEVIN HILBERT, TANJA JACOBI, STEFANIE L. KUNAS, BJÖRN ELSNER, BENEDIKT REUTER, ULRIKE LUEKEN, & NORBERT KATHMANN

Faculty of Life Sciences, Department of Psychology, Humboldt-Universität zu Berlin, Berlin, Germany

(Received 30 December 2019; revised 13 October 2020; accepted 13 October 2020)

Abstract

Objectives: Machine learning models predicting treatment outcomes for individual patients may yield high clinical utility. However, few studies tested the utility of easy to acquire and low-cost sociodemographic and clinical data. In previous work, we reported significant predictions still insufficient for immediate clinical use in a sample with broad diagnostic spectrum. We here examined whether predictions will improve in a diagnostically more homogeneous yet large and naturalistic obsessive-compulsive disorder (OCD) sample. **Methods:** We used sociodemographic and clinical data routinely acquired during CBT treatment of $n = 533$ OCD subjects in a specialized outpatient clinic. **Results:** Remission was predicted with 65% ($p = 0.001$) balanced accuracy on unseen data for the best model. Higher OCD symptom severity predicted non-remission, while higher age of onset of first OCD symptoms and higher socioeconomic status predicted remission. For dimensional change, prediction achieved $r = 0.31$ ($p = 0.001$) between predicted and actual values. **Conclusions:** The comparison with our previous work suggests that predictions within a diagnostically homogeneous sample, here OCD, are not per se superior to a more diverse sample including several diagnostic groups. Using refined psychological predictors associated with disorder etiology and maintenance or adding further data modalities as neuroimaging or ecological momentary assessments are promising in order to further increase prediction accuracy.

Keywords: single-case prediction; machine learning; random forest; cognitive behavioral therapy; outcome; obsessive-compulsive disorder

Clinical or methodological significance of this article: This paper presents several approaches to predict cognitive behavioral therapy outcomes for individual patient in a large and naturalistic sample with obsessive-compulsive disorder. It demonstrates that remission and dimensional change can be predicted already at baseline with significant above-chance accuracy for individual patients with routinely acquired data, paving the way for tailoring treatments to patients in the future. When compared to earlier research, it suggests that restricting the prediction to an individual diagnosis will not per se improve prediction performance. Finally, it suggests that these predictions can be further optimized by adding a refined set of psychological predictors such as early gains.

1. Introduction

Despite much progress in the treatment of mental disorders over the last decades, for the majority of disorders even the best available treatments leave a significant amount of patients not sufficiently improved. Although we can confidently state that many treatments work (in principle), they clearly do not work for everyone, with severe consequences for patients and high cost and resource demands for societies. As a consequence, considerable effort is invested in examining which patient and treatment characteristics are associated with treatment nonresponse in order to subsequently match patients with

Correspondence concerning this article should be addressed to Kevin Hilbert, Faculty of Life Sciences, Department of Psychology, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany Email: Kevin.Hilbert@hu-berlin.de

treatments that have a higher probability to work for this individual person, thus improving outcomes. Such approaches are commonly referred as personalized medicine (e.g., Ozomaro et al., 2013). Although previous research has unraveled a number of treatment predictors in group-based studies, it has so far been difficult to transform this knowledge into outcome predictions for individual patients. A major obstacle is that typically a considerable number of such group-based predictors for either response or nonresponse are present in one patient, outweighing and interacting with each other in so far incomprehensible ways. But there are two advancements in the field with the potential to overcome these impediments. Machine learning is able to integrate large numbers of potential predictors into one final decision function in a multivariate fashion. Additionally, ever-growing datasets are available including increasing numbers of patients with their unique combination of features. As a result, machine learning and big data approaches are increasingly applied in order to achieve reliable single-patient outcome predictions and pave the way for personalized medicine in mental health.

Despite the advent of machine learning as a methodological innovation for predicting on the single-subject level, it still has to be determined which variables and variable sets make the best predictors for a given outcome. While sociodemographic and clinical data are already routinely acquired and can be collected at low cost, biological data such as (functional) magnetic resonance imaging data may yield higher precision by evaluating underlying substrates of psychopathology. Previous studies using sociodemographic and clinical data to predict treatment outcomes for cognitive behavior therapy (CBT) achieved around 59–65% accuracy (Ball et al., 2014; Hilbert et al., 2020), which is considerably lower than the 70–92% accuracy in many comparable neuroimaging studies (Ball et al., 2014; Dunlop et al., 2017; Hahn et al., 2015; Månsson et al., 2015; Reggente et al., 2018). Yet it is unclear whether these higher prediction accuracies can be related to a better utility of neuroimaging data for the task given the very small sample sizes in most of these studies. Another open question is how to define the optimal sample in which to apply a given predictor. In our own previous work in a large, naturalistic set of patients, we saw that significant outcome prediction using sociodemographic and clinical data was possible, but with accuracy below real clinical utility (Hilbert et al., 2020). This matches comparable studies with sociodemographic and clinical predictors. However, it is important to note that the sample included diverse diagnoses.

Here, we examine whether predicting CBT outcomes using routine sociodemographic and clinical data can be achieved with clinical utility in a sample of patients with obsessive-compulsive disorder (OCD) only. To this end, we examined a new large, naturalistic and longitudinal dataset from an outpatient clinic of the Humboldt-Universität zu Berlin specialized in OCD treatment. This dataset was not included in our previous analyses. The finding that many patients do not benefit from treatments that show overall good effects applies fully to OCD, where CBT produces large effect sizes (Carpenter et al., 2018) but remission rates are still only at about 50% (Gava et al., 2007; Ost et al., 2015; Springer et al., 2018). Mirroring our previous work, we applied different machine learning approaches to predict categorical and dimensional treatment outcomes based on baseline data exclusively. Within-treatment data may provide additional information for outcome prediction but was not used here, because the decision to initiate a specific treatment has to be made at baseline with the information available at this point in time. We expected prediction accuracies to be significantly above chance level and on a descriptive level, substantially above our previous work (59% balanced accuracy for categorial outcome, $r = 0.27$ for dimensional outcome; Hilbert et al., 2020).

2. Methods

2.1. Sample

Data came from the specialty OCD outpatient clinic of the Humboldt-Universität zu Berlin, Germany and included 533 patients that terminated treatment from January 2010 to June 2018, and received at least one therapy session. As a specialized institution, the OCD outpatient clinic makes use of inclusion and exclusion criteria for the admittance of subjects. Inclusion criteria were: Primary diagnosis of OCD according to DSM-IV (APA, 2000), Yale-Brown Obsessive Compulsive Scale (Y-BOCS; Goodman et al., 1989) severity score of ≥ 16, age between 18 and 70 years, written informed consent. Diagnosis was determined by administration of the Structured Clinical Interview for DSM-IV Axis I disorders (SCID-I; First et al., 1997) through trained clinical psychologists, 65% of all subjects presented at least one comorbid psychiatric disorder. Exclusion criteria were: prominent suicidal ideation, any lifetime substance dependence, borderline personality disorder and comorbid psychotic disorders. Analysis of routinely collected data met the ethical standards of the revised Declaration of Helsinki. All participants provided written informed consent that the data being collected during their therapies will be used for research and might be published.

Patients received CBT with a focus on exposure with response prevention (e.g., Foa et al., 2012) which was provided by experienced cognitive-behavioral therapists. Treatment was delivered by 28 licensed psychotherapists who had received not less than three years of training in CBT. In accordance with the general conditions for psychotherapy in the German public health system a maximum of 80 sessions per regular therapy were conducted which was exceeded in 14 cases (min: 2 sessions; max: 108 sessions; median: 45 sessions (equivalent to 37.5 h); IQR: 25 sessions), and therapists received weekly supervision. Treatment was terminated as a consensual decision of patient and therapist based on clinical criteria (e.g., Y-BOCS < 12, significant decrease of psychological distress).

Study visits took place at T0 (patients enrolled for treatment), where inclusion/exclusion criteria were evaluated to prove eligibility, using the SCID-I and SCID-II (First et al., 1997; First et al., 1997). To assess the severity and determine overall effectiveness, selected self-report questionnaires and external interviews (see below) were applied at T0, every 20 session and at termination of treatment. Due to the naturalistic structure, Last-Observation-Carried-Forward (LOCF) method has been used as a conservative estimate of outcome to avoid missing data.

2.2. Description and Preparation of Data

The overall dataset consisted of $n = 533$ patients and initially $k = 1192$ variables. Extensive quality control and a reduction of the dataset for our analytic purpose were conducted which were largely comparable to our previous study (Hilbert et al., 2020). In short, variables unsuitable for prediction and variables which were missing for > 25% of the sample were excluded, as were, in turn, subjects for which more than 25% of the remaining variables were missing. Categorical data were recoded in one binary dummy-variable per original category. Comorbid diagnoses were coded in disorder categories for the absence of diagnosis and presence of diagnosis, with disorder categories oriented at the DSM-5 (APA, 2013) classification and thus encompassing substance use, schizophrenia-spectrum and psychotic, bipolar, unipolar-depressive, anxiety, obsessive-compulsive, trauma-related, somatic-symptom, eating, impulse-control, personality and other disorders. This was only done for comorbid diagnoses as all subjects had a primary OCD as inclusion criterion. The final number of variables used in the analyses was $k = 504$.

Outcome was defined categorically and dimensionally. Remission was used as categorical outcome and was indicated by a Y-BOCS severity score of ≤12 at the end of treatment (Mataix-Cols et al., 2016). The Y-BOCS severity score is a reliable and validated measure of OCD symptoms. It consists of two parts: (a) a symptom checklist (the Yale-Symptom-Checklist; YSC) to determine the types of obsessions and compulsions (b) a clinician-administered interview to assess the severity of the present symptoms using a 10-item scale. This scale comprises the severity score of obsessions (range: 0–20) and compulsion (range: 0–20), as well as a total score (range: 0–40; Goodman et al., 1989; Goodman et al., 1989). A recent meta-analysis shows good mean intraclass correlation ($r = 0.92$) and test–retest reliability ($r = 0.85$; Lopez-Pina et al., 2015). Furthermore, the Y-BOCS has demonstrated good convergent validity (mean $r = 0.51$) but somewhat poor discriminant validity, e.g., with measures of depression (mean $r = 0.64$) (Goodman et al., 1989). A therapist rating of the therapeutic efficacy (AMDP & CIPS, 1990) based on the CGI-I was used as dimensional change outcome. The resulting datasets for remission and dimensional change included $k = 500$ variables and $n = 465$ and $n = 424$ subjects, respectively, as the number of missing values differed for both outcomes.

Example variables for prediction are age, sex and employment status, diagnostic categories of comorbid disorders, the Obsessive Compulsive Inventory (OCI-R; Foa et al., 2002), Brief Symptom Inventory (BSI; Derogatis, 1993), Montgomery–Åsberg Depression Rating Scale (MADRS; Montgomery & Asberg, 1979), Beck Depression Inventory-II (BDI-II; Beck et al., 1996) and the global assessment of functioning (GAF; APA, 1994). All instruments are reliable, validated and frequently used measures of disabling consequences and/or major comorbid conditions of OCD (see supplement 1 for the complete list of variables).

2.3. Single-Case Prediction via Machine Learning

Papers providing an introduction and overview to machine learning methods specifically for psychologists are available (Dwyer et al., 2018; Hilbert & Lueken, 2020). Machine learning was done using scikit-learn 0.22.1. (http://scikit-learn.org/stable/) in Python. Datasets for both outcome types were split in training (2/3 sample) and test set (1/3) of sample. Approaches using various algorithms and optimization strategies were trained on the train set and compared regarding their prediction performance on the test set. Within the training set, hyperparameters were tuned in an "inner" cross-validation framework with 5 folds and 100 iterations. These approaches with their rationales were: (i) random forests for robustness and interpretability, (ii) support vectors for comparability to earlier works, and (iii) ensembles for integrating individually weak predictors (see supplement 2 for more details). Based on a reviewer's request, we added simple models built on linear and logistic regressions for comparison. Importantly, some approaches used a reduced n as subjects were for instance excluded due to missing data or undersampling in order to achieve balanced group sizes. Models were compared on balanced accuracy, sensitivity, specificity, log-loss and area under the curve (AUC) for remission and on correlation between real and predicted values, root mean squared error (RMSE), and mean absolute error (MAE) for dimensional change. In order to get stable estimates for these metrics that are not dependent on the specific train-test-split, we conducted each model across 100 iterations with random splits and report mean values and standard deviations for all metrics. After comparing all approaches according to their performance on the test set, the best performers for categorical and dimensional outcomes were evaluated using a permutation test with 5.000 iterations of randomly shuffled labels. The *p*-value was calculated as $\sum((\text{accuracy}_{\text{validation}} < \text{accuracy}_{\text{permutation}}) + 1)/(\text{n}_{\text{permutations}} + 1)$. Additionally, we compared the prediction performance of the best models for categorical and dimensional outcomes with the corresponding simple models built on linear and logistic regressions with severity as predictor. We used the corrected resampled *t*-test (Bouckaert & Frank, 2004; Nadeau & Bengio, 2003) for significance testing.

3. Results

Table I gives an overview on the sociodemographic, clinical and outcome data in both analysis samples. The sample showed a wide age range, a balanced sex distribution and was moderately skewed towards higher education regarding sociodemographic status. More than two-thirds of the sample presented more than one diagnosis. About half of the sample showed remission of the OCD diagnosis after treatment with more than two-thirds of the sample showing moderate to extensive improvement according to the dimensional change score.

3.1. Remission

The different approaches for predicting remission varied to some extent in their performance and ranged between 52 and 65% balanced accuracy, a log-loss between 0.62 and 1.41 and an AUC between 0.60 and 0.72 (see Table II for all metrics and supplement 3 for additional ROC curves and calibration plots per model). Despite variation being present, there was a group of approaches producing very similar outcomes. All of these approaches outperformed the simple logistic regression models in balanced accuracy, log-loss and AUC. The marginally best approach used a Random Forest classifier on the balanced dataset with all preprocessing steps including recursive feature elimination with a balanced accuracy of 65% (comparison to chance level: mean $p = 0.001$; comparison to logistic regression: $t(457) = 1.80$, $p = 0.036$, one-tailed). The 10 features surviving recursive feature reduction varied across the individual iterations, but single YSC and OCI-R items, socioeconomic status, age, age of onset of first OCD symptoms and of OCD diagnosis and OCI-R, Y-BOCS, MADRS and BSI overall and subscale sum scores were commonly found (see Table III for ranking of all features included in at least 95 of 100 iterations). The slope of feature reduction weights was comparably shallow, i.e., all remaining features contributed considerably.

3.2. Dimensional Change

The different approaches for predicting dimensional change varied considerably in the correlations of predicted and true values ($r = 0.06$ to $r = 0.31$; see Table I for all metrics). Again, there was a group of approaches producing very similar outcomes at the top, which outperformed the simple linear regression models in correlations, RMSE and MAE. The marginally best approaches used a Random Forest regressor to achieve a correlation of $r = 0.31$ (comparison to chance level: mean $p = 0.001$; comparison to linear regression: $t(423) = 1.42$, $p = 0.078$, one-tailed), variance-based feature reduction was optional. The most important features across the 100 iterations largely overlapped with the ones for remission (see Table IV for ranking). The slope of feature reduction weights was very shallow.

4. Discussion

Despite their common availability and low-cost nature, few studies applied machine learning approaches to clinical and sociodemographic data in order to predict mental health treatment outcomes for individual patients. Extending previous work (Hilbert et al., 2020), we here predicted CBT outcome in a naturalistic and longitudinal sample of patients with only OCD as primary diagnosis. In

Table I. Sample characteristics pre-treatment. The dimensional change analysis sample is smaller than the remission analysis sample as there were subjects for which the binary outcome was available but the therapeutic efficacy score was missing. Means (SD) except where noted.

	Remission analysis sample ($n = 465$)						Dimensional change analysis sample ($n = 424$)	
	Remission ($n = 229$)		Non-remission ($n = 236$)		chi^2/t	p		
Sociodemographic characteristics								
Female sex (n, %)	120	(52.4)	140	(59.3)	2.258	0.133	239	(56.4)
Age (years)	31.9	(9.9)	33.7	(10.5)	1.908	0.057	32.8	(10.3)
# children	0.3	(0.7)	0.5	(0.8)	1.791	0.074	0.4	(0.7)
Socioeconomic status	10.0	(3.7)	8.9	(3.9)	3.145	0.002	9.5	(3.8)
Educational status								
lowest secondary school (n, %)	10	(4.4)	25	(10.6)	6.610	0.010	34	(8.0)
intermediate secondary school (n, %)	67	(29.3)	88	(37.3)	3.628	0.057	142	(33.5)
highest secondary school (n, %)	149	(65.1)	116	(49.2)	11.424	0.001	239	(56.4)
without school graduation (n, %)	2	(0.9)	5	(1.7)	0.632	0.427	5	(1.2)
Marital status								
unmarried (n, %)	154	(67.2)	151	(64.0)	0.461	0.497	282	(66.5)
married or in relationship (n, %)	69	(30.1)	71	(30.1)	0.000	0.985	123	(29.0)
separated or divorced (n, %)	5	(2.2)	12	(5.1)	2.807	0.094	16	(3.8)
widowed (n, %)	1	(0.4)	1	(0.4)	0.000	0.985	2	(0.5)
Employment status								
unemployed (n, %)	34	(14.8)	67	(28.4)	13.078	<0.001	92	(21.7)
full time (n, %)	96	(41.9)	70	(29.7)	7.079	0.008	152	(35.8)
part time or occasionally (n, %)	43	(18.8)	47	(19.9)	0.143	0.705	80	(18.9)
in education (n, %)	47	(20.5)	38	(16.1)	1.367	0.242	77	(18.2)
pension (n, %)	3	(1.3)	2	(0.8)	0.220	0.639	5	(1.2)
other (n, %)	6	(2.6)	9	(3.8)	0.568	0.451	15	(3.5)
Clinical characteristics								
Primary diagnosis obsessive-compulsive disorder (n, %)	229	(100.0)	236	(100.0)	–	–	424	(100.0)
CGI-S[a]	5.5	(0.8)	5.8	(0.8)	3.987	<0.001	5.6	(0.8)
GAF	57.3	(10.1)	53.9	(10.0)	3.515	<0.001	55.6	(10.3)
Total # of diagnostic categories	2.0	(1.0)	2.1	(1.0)	1.609	0.108	2.1	(1.0)
BDI-II	17.1	(10.7)	19.9	(10.6)	2.823	0.005	18.4	(10.7)
BSI	48.4	(31.0)	55.1	(29.9)	2.403	0.017	51.8	(30.8)
MADRS	12.1	(9.1)	13.5	(8.9)	1.646	0.100	12.9	(9.1)
Y-BOCS	21.3	(5.3)	24.5	(5.3)	6.417	<0.001	22.8	(5.6)
OCI-R	25.0	(11.3)	29.8	(12.4)	4.333	<0.001	27.4	(12.0)
WST	31.5	(4.3)	30.8	(4.9)	1.652	0.099	31.1	(4.6)
Outcomes								
# remitted (n, %)	229	(100.0)	236	(100.0)				
Therapeutic efficacy (CGI-I)[b]							2.90	(0.9)

Note: Percent readings are given for the total of valid data, i.e., exclude missing data points. All variables except those given under outcomes are baseline values. CGI: Clinical Global Impressions Scale; GAF: Global Assessment of Functioning; BDI-II: Beck Depression Inventory-II; BSI: Brief Symptom Inventory; MADRS: Montgomery Asberg Depression Rating Scale; Y-BOCS: Yale-Brown Obsessive Compulsive Scale; OCI-R: Obsessive Compulsive Inventory; WST: Wortschatztest [German vocabulary test].

[a] CGI severity ratings, range 2–8, higher scores indicate worse symptoms.

[b] CGI improvement (therapist rating of the therapeutic efficacy), range 2–5, lower scores indicate more improvement.

line with our hypotheses, we found that both remission and dimensional change outcomes were predicted with accuracy substantially beyond chance level. The best model for remission additionally achieved significantly higher accuracy than a simpler model using logistic regression while for dimensional change, only a nonsignificant trend was found for the comparison of the best model to a simpler linear regression. On a descriptive level, the prediction performance in this sample was comparable to earlier findings in the literature and in our own work with a diagnostically more diverse dataset (Hilbert et al., 2020; dimensional change: $r = 0.31$ here, $r = 0.27$ in the previous paper).

The main result of this study provides further evidence that the prediction of treatment outcomes for mental disorders using routine data alone is possible with accuracy substantially beyond chance level. It is

Table II. Performance metrics for predictions on binary and dimensional outcomes.

	n train	n test	ACC_{Bal} M	ACC_{Bal} SD	SensM	Sens SD	Spec M	Spec SD	Log-loss M	Log-loss SD	AUC M	AUC SD
Classification on Remission												
unbalanced, Logistic Regression, imputation, scaling, baseline severity only	311	154	0.63	0.03	0.63	0.05	0.62	0.06	0.66	0.02	0.67	0.03
unbalanced, Logistic Regression, imputation, scaling	311	154	0.58	0.03	0.58	0.06	0.58	0.05	1.41	0.67	0.62	0.04
unbalanced, RF	98–120	41–63	0.62	0.06	0.45	0.12	0.80	0.10	0.64	0.03	0.70	0.06
unbalanced, RF, imputation	311	154	0.65	0.03	0.62	0.07	0.68	0.06	0.63	0.01	0.71	0.03
unbalanced, RF, imputation, FR variance	311	154	0.65	0.03	0.62	0.07	0.69	0.06	0.63	0.01	0.72	0.03
unbalanced, RF, imputation, FR recursive	311	154	0.65	0.03	0.62	0.06	0.68	0.06	0.63	0.02	0.71	0.03
unbalanced, RF, imputation, FR elastic-net	311	154	0.65	0.04	0.61	0.07	0.68	0.06	0.64	0.02	0.70	0.03
unbalanced, SVM, imputation, scaling	311	154	0.61	0.04	0.53	0.13	0.69	0.13	0.66	0.02	0.66	0.05
unbalanced, SVM, imputation, scaling, FR recursive	311	154	0.57	0.04	0.57	0.07	0.58	0.08	1.23	0.83	0.60	0.05
balanced, Logistic Regression, imputation, scaling, baseline severity only	306	152	0.64	0.03	0.61	0.07	0.67	0.09	0.66	0.02	0.67	0.04
balanced, Logistic Regression, imputation, scaling	306	152	0.58	0.04	0.59	0.06	0.58	0.04	1.35	0.64	0.62	0.04
balanced, RF	91–119	39–65	0.63	0.06	0.47	0.12	0.80	0.10	0.64	0.02	0.70	0.06
balanced, RF, imputation	306	152	0.65	0.03	0.64	0.06	0.66	0.05	0.63	0.01	0.72	0.03
balanced, RF, imputation, FR variance	306	152	0.65	0.03	0.64	0.06	0.66	0.05	0.63	0.01	0.72	0.03
balanced, RF, imputation, FR recursive	**306**	**152**	**0.65**	**0.03**	**0.64**	**0.06**	**0.67**	**0.05**	**0.62**	**0.02**	**0.72**	**0.03**
balanced, RF, imputation, FR elastic-net	306	152	0.65	0.03	0.64	0.06	0.65	0.06	0.64	0.02	0.70	0.04
balanced, SVM, imputation, scaling	306	152	0.58	0.04	0.59	0.09	0.56	0.08	1.12	0.74	0.66	0.05
balanced, SVM, imputation, scaling, FR recursive	306	152	0.58	0.04	0.59	0.09	0.56	0.08	1.12	0.74	0.61	0.05
balanced, Ensemble + RF, imputation, scaling, FR recursive	306	152	0.60	0.04	0.59	0.10	0.60	0.11	0.82	0.18	0.61	0.05
balanced, Ensemble + voting, imputation, scaling, FR recursive	306	152	0.64	0.04	0.64	0.06	0.63	0.06	1.09	0.05	0.69	0.04

Regression on dimensional change	n train	n test	Correlation M	Correlation SD	RMSE M	RMSE SD	MAE M	MAE SD
Linear Regression, imputation, scaling, baseline severity only	284	140	0.23	0.08	0.99	0.06	0.78	0.05
Linear Regression, imputation, scaling	284	140	0.06	0.08	1.73	0.09	1.37	0.07
RF	85–105	37–57	0.28	0.09	0.99	0.06	0.81	0.05
RF, imputation	**284**	**140**	**0.31**	**0.06**	**0.90**	**0.04**	**0.72**	**0.03**
RF, imputation, FR variance	**284**	**140**	**0.31**	**0.06**	**0.90**	**0.04**	**0.72**	**0.03**
RF, imputation, FR recursive	284	140	0.30	0.06	0.91	0.04	0.72	0.03
RF, imputation, FR elastic-net	284	140	0.29	0.06	0.91	0.04	0.72	0.03
SVR, imputation, scaling	284	140	0.27	0.07	0.98	0.06	0.78	0.05
SVR, imputation, scaling, FR recursive	284	140	0.07	0.10	1.03	0.07	0.81	0.06

Note: Models with the best performance for remission and dimensional change printed in bold. Balanced accuracy, sensitivity and specificity metrics as fractions between 0 (prediction never correct) and 1 (prediction always correct). ACC_{Bal}: Balanced accuracy; Sens: Sensitivity; Spec: Specificity; AUC: area under the curve; RMSE: root mean squared error; MAE: mean absolute error; RF: random forest; SVM: support vector machine; SVR: support vector regression; FR: feature reduction.

Table III. The ranking and weights of all features surviving recursive feature reduction in at least 95 of 100 iterations for the classifier with the highest accuracy for remission.

# Rank	Feature	Feature weight	Direction[a]	# inclusions in model
1	Y-BOCS Sumscore including nine additional items[b]	4.14025	–	100
2	OCI-R Sumscore subscale washing	3.44906	–	100
3	BSI Sumscore	1.93174	–	100
4	YSC Item # 37[c]	3.26542	–	99
5	Y-BOCS Sumscore	2.50295	–	99
6	Age	1.99605	–	99
7	OCI-R Sumscore	1.67283	–	99
8	WST Sumscore	1.49759	+	99
9	SES: overall	1.60776	+	98
10	OCI-R Sumscore subscale obsessive thoughts	1.60154	+	98
11	BSI Meanscore	1.59599	–	98
12	BDI-II sumscore	1.52728	–	98
13	BSI number of items rated > 1	1.50490	–	98
14	Age of onset OCD diagnosis	1.42964	–	98
15	OCI-R Sumscore subscale ordering	1.61500	–	97
16	Age of onset first OCD symptoms	1.39266	–	97
17	MADRS sumscore	1.33510	–	97
18	OCI-R Sumscore subscale neutralizing	2.15763	–	96
19	Y-BOCS Sumscore subscale compulsions	1.91038	–	96
20	GAF	1.24020	+	95

Note: Only baseline feature were used. Ordering according to the number of inclusions in the model, then weight. The sum of all feature weights is one per iteration, i.e., 100 over all iterations. Weights and directions based on performance in the complete set. BDI-II: Beck Depression Inventory-II; BSI: Brief Symptom Inventory; GAF: Global Assessment of Functioning; MADRS: Montgomery Asberg Depression Rating Scale; OCI-R: Obsessive Compulsive Inventory; Y-BOCS: Yale-Brown Obsessive Compulsive Scale; YSC: Yale-Symptom-Checklist; SES: Socioeconomic status WST: Wortschatztest [German vocabulary test].

[a]Plus indicated higher values for remission and lower values for non-remission. Vice versa for minus.

[b]This refers to Y-BOCS items 11–19 (starting from insight) that go beyond severity of obsessions and compulsions.

[c]Excessive or ritualized showering, bathing, toothbrushing or personal hygiene.

important to note, that this result is a particularly robust and unbiased estimate of the true prediction accuracy achievable with this data given the large, naturalistic and longitudinal sample and given our validation strategy with 100 iterations of non-overlapping train and test splits which is particularly well suited to guard against overfitting and bias. The prediction performance in this investigation is also in line

Table IV. The ranking and weights of the 10 most important features surviving variance-based feature reduction for the classifier with the highest accuracy for dimensional change.

# Rank	Feature	Feature weight	Direction[a]	# inclusions in model
1	Y-BOCS Sumscore including nine additional items[b]	1.60305	–	100
2	YSC Item # 37[c]	1.26199	–	100
3	Y-BOCS Sumscore	1.24457	–	100
4	GAF	1.18261	+	100
5	SES: overall	1.13231	+	100
6	OCI-R Sumscore	1.05053	–	100
7	OCI-R Sumscore subscale washing	0.97142	–	100
8	Y-BOCS Sumscore subscale compulsions	0.90962	–	100
9	OCI-R Sumscore subscale ordering	0.89512	–	100
10	YSC Item # 2[d]	0.85325	+	100

Note: Only baseline feature were used. Ordering according to the number of inclusions in the model, then weight. The sum of all feature weights is one per iteration, i.e., 100 over all iterations. Weights and directions based on performance in the complete set. GAF: Global Assessment of Functioning; OCI-R: Obsessive Compulsive Inventory; Y-BOCS: Yale-Brown Obsessive Compulsive Scale; YSC: Yale-Symptom-Checklist; SES: Socioeconomic status.

[a]Plus indicated higher values with increasing dimensional change. Vice versa for minus.

[b]This refers to Y-BOCS items 11–19 (starting from insight) that go beyond severity of obsessions and compulsions.

[c]Excessive or ritualized showering, bathing, toothbrushing or personal hygiene

[d]Fear to harm others.

with the available other machine learning studies predicting treatment outcome using sociodemographic and clinical data: for antidepressant medication, prediction accuracies of 65% and 66% (Chekroud et al., 2016; Iniesta et al., 2016) have been reported, while for a naturalistic OCD sample treated with CBT and medication, a somewhat higher accuracy of 75% has been found (Askland et al., 2015). It should be noted that the latter paper also included data collected after baseline such as treatment variables or personality scores, while our investigation was restricted to predictors being available at baseline. A second prediction paper in OCD reported accuracies between 75 and 83% for different approaches (Lenhard et al., 2018), however, comparisons are more difficult here as this study was in pediatric OCD and had a small sample size. Nevertheless, for real clinical value considerably larger prediction accuracies are needed. The available literature including this study raises doubts whether this goal can be achieved using routinely collected data alone. Two avenues are particularly promising. First, refined psychological data with variables related to the given psychopathology according to the latest theoretical models may substantially elevate predictions. Candidates are factors related to disorder etiology and maintenance, but potentially also early treatment variables such as early gains. For OCD, certain symptom dimensions (Mataix-Cols et al., 2002) as well as compliance and homework related adherence (Simpson et al., 2011; Simpson et al., 2012) may be particularly interesting variables. Second, other data modalities such as neuroimaging or ecological momentary assessments can be added to increase prediction performance. To date, we are unaware of any prediction studies for OCD using ecological momentary assessment data for outcome prediction. For neuroimaging, two studies in OCD samples achieved 70% accuracy using resting-state connectivity data (Reggente et al., 2018) and 67% accuracy using cortical thickness in the orbitofrontal cortex (Hoexter et al., 2015), thus only achieving moderately better prediction performances. Prediction studies based on neuroimaging data for other diagnoses partly reported considerably superior outcomes (70–92% across different disorders; Ball et al., 2014; Dunlop et al., 2017; Reggente et al., 2018; Hahn et al., 2015; Månsson et al., 2015; Hoexter et al., 2015). However, neuroimaging prediction studies typically used considerably smaller to very small sample sizes, with $n = 42$ and $n = 41$ subjects in total in the OCD studies (Hoexter et al., 2015; Reggente et al., 2018). Prediction accuracies have been found to substantially decrease in larger sample sizes, potentially due to overfitting and bias (Varoquaux, 2018). Additionally, small samples were often highly selective (Schnack & Kahn, 2016). In comparison, the data presented in our study comes from a considerably larger, naturalistic dataset leading to a potentially more accurate estimate of current predictor performance under real-world conditions. For future studies to determine whether different data modalities provide superior or additive value for prediction, datasets of comparable sample sizes are needed. This prerequisite particularly requires larger samples for unbiased accuracy estimation for the neuroimaging studies.

The prediction performance in this sample can also be compared with our previous dataset (Hilbert et al., 2020) including a more diverse range of primary diagnoses on a descriptive level. Contrary to expectation, it is an important finding that outcome predictions on OCD only in the present dataset were only moderately superior for remission and dimensional change. This indicates that building predictors for restricted samples with only one or few diagnoses may not per se increase prediction performance considerably compared to predictors for more diverse sets of diagnoses. On the upside, this may also mean that predictors that generalize over many diagnoses have not to be necessarily inferior in terms of accuracy. If replicated, this may encourage future studies to develop predictors over rather broad diagnostic categories. However, as one anonymous reviewer rightly noted, individual diagnoses may exhibit different levels of heterogeneity and thus there may be primary diagnoses other than OCD where the benefit of very specific predictors is higher. Moreover, there may be additional sources of sample heterogeneity such as age, socio-cultural background or the range of therapeutic interventions that may have an effect on prediction performance and may be fruitful targets for examination in future studies.

When examining the variables implicated in the best performing prediction model, we found the most relevant predictors to be largely overlapping with what would be expected from previous group-based predictions of therapy outcomes, including symptom severity, socioeconomic status and age of onset (Knopp et al., 2013). Y-BOCS variables were also highly important in the single-subject outcome prediction by Askland et al. (2015). Similarly to the Askland et al. (2015) study and our previous work, we found again that differences in feature weights were relatively small. Thus, in the more complex models, no single feature was of ultimate importance but rather there were many features that each contributed to a moderate extend to prediction. At least for remission, the inclusion of these additional features leads to significantly more accurate predictions that a very simple model based on severity alone. For a dimensional change, the increase was less pronounced and not statistically significant. Overall, this suggests that symptom severity alone can be used to construct very simple models with considerable power, but it is not sufficient to construct clinical useful models and additional information must be added to move closer to this point. Notably, the significant increase in remission prediction accuracy for the more complex model was still small in absolute terms, again emphasizing the need for superior predictor data such as refined psychological, EMA, or neuroscience data. Compared to our previous work, the present study also differed with respect to the categorical outcome measure used as index for remission. While our previous study used remission based on clinical observation, the availability of an expert consensus measure for remission in OCD (Mataix-Cols et al., 2016) allowed us to use a more standardized outcome based on the Y-BOCS here. Reliability of outcomes (and features) so far has often been overlooked as an important factor influencing performance in treatment outcome predictions. Decreasing reliability will likely also decrease prediction performance considerably. According to meta-analysis, reliability of the Y-BOCS is generally good although particularly for clinical samples below the benchmark of the most strict reliability criteria for clinical use (Lopez-Pina et al., 2015). The Y-BOCS is also sensitive to change (Emmelkamp et al., 1995). Employing a very reliable outcome measure such as the Y-BOCS may therefore be one further factor in achieving moderately superior categorical prediction compared to our previous study.

There are a number of limitations to this study. First, the test set came from the same overall sample as the training set and therefore is still limited in its ability to test the generalizability of the final approaches. For the best potential test of generalizability, data from different settings such as other OCD treatment institutions would be needed (e.g., from KODAP; Velten et al., 2017). Second, as in our previous study, a range of questionnaires had to be dropped due to more than 25% missings in the sample. As before, these were mostly questionnaires for very specific parts of the sample (i.e., on rare comorbidities or very specialized psychological constructs). Due to their specificity, finding an adequate way to integrate such data in future models may considerably elevate prediction performance.

Treatment nonresponse is common for many disorders and treatments, including CBT for OCD (Springer et al., 2018). The ability to predict who is particularly at risk for nonresponse is an important step in the aim of personalized treatments yielding high potential for better overall outcomes, lowered burden and reduced societal cost. Here, we presented data from a large naturalistic dataset of OCD patients and used routinely acquired and low-cost information to predict CBT outcome. We found predictions substantially exceeding chance level for all outcomes. For remission, prediction also achieved significantly higher accuracy than a model based on severity alone. The comparison with an own previous study on a sample with diverse primary diagnoses suggests that prediction performances do not necessarily increase in more specialized samples. Potential avenues to increase prediction performances for real clinical value include adding more data modalities and adding further features mapping theoretically meaningful constructs.

References

AMDP, & CIPS. (1990). *Rating scales for psychiatry: European Edition*. Beltz Test.

APA. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). American Psychiatric Press.

APA. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). American Psychiatric Association.

APA. (2013). *Diagnostic and statistical manual of mental disorders* (5 ed.). American Psychiatric Association.

Askland, K. D., Garnaat, S., Sibrava, N. J., Boisseau, C. L., Strong, D., Mancebo, M., Greenberg, B., Rasmussen, S., & Eisen, J. (2015). Prediction of remission in obsessive compulsive disorder using a novel machine learning strategy. *International Journal of Methods in Psychiatric Research*, *24*(2), 156–169. https://doi.org/10.1002/mpr.1463

Ball, T. M., Stein, M. B., Ramsawh, H. J., Campbell-Sills, L., & Paulus, M. P. (2014). Single-subject anxiety treatment outcome prediction using functional neuroimaging. *Neuropsychopharmacology*, *39*(5), 1254–1261. https://doi.org/10.1038/npp.2013.328

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory* (2nd ed.). The Psychological Corporation.

Bouckaert, R. R., & Frank, E. (2004). Evaluating the replicability of significance tests for comparing learning algorithms. In H. Dai, R. Srikant, & C. Zhang (Eds.), *Advances in knowledge discovery and data mining. PAKDD 2004. Lecture Notes in Computer Science* (Vol. 3056, pp. 3–12). Springer.

Carpenter, J. K., Andrews, L. A., Witcraft, S. M., Powers, M. B., Smits, J. A. J., & Hofmann, S. G. (2018). Cognitive behavioral therapy for anxiety and related disorders: A meta-analysis of randomized placebo-controlled trials. *Depression and Anxiety*, *35*(6), 502–514. https://doi.org/10.1002/da.22728

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., Cannon, T. D., Krystal, J. H., & Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. *The Lancet. Psychiatry*, *3*(3), 243–250. https://doi.org/10.1016/S2215-0366(15)00471-X

Derogatis, L. R. (1993). *Brief Symptom Inventory (BSI), administration, scoring, and procedures manual*, 3rd ed. National Camputer Services.

Dunlop, B. W., Rajendra, J. K., Craighead, W. E., Kelley, M. E., McGrath, C. L., Choi, K. S., Kinkead, B., Nemeroff, C. B., & Mayberg, H. S. (2017). Functional connectivity of the Subcallosal Cingulate Cortex and differential outcomes to treatment with cognitive-behavioral therapy or antidepressant medication for major depressive disorder. *American Journal of Psychiatry*, *174*(6), 533–545. https://doi.org/10.1176/appi.ajp.2016.16050518

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, *14*(1), 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Emmelkamp, P. M. G., van Balkom, A. J. L. M., & van Dyck, R. (1995). The sensitivity to change of measures for obsessive-compulsive disorder. *Journal of Anxiety Disorders*, *9*(3), 241–248. https://doi.org/10.1016/0887-6185(95)00005-9

First, M., 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)*. American Psychiatric Press.

First, M., Spitzer, R., Gibbon, M., & Williams, J. (1997). *Structured clinical interview for DSM-IV*. American Psychiatric Press.

Foa, E. B., Huppert, J. D., Leiberg, S., Langner, R., Kichic, R., Hajcak, G., & Salkovskis, P. M. (2002). The obsessive-compulsive inventory: Development and validation of a short version. *Psychological Assessment*, *14*(4), 485–496. https://doi.org/10.1037/1040-3590.14.4.485

Foa, E. B., Yadin, E., & Lichner, T. K. (2012). *Exposure and response (ritual) prevention for obsessive compulsive disorder: Therapist guide*. Oxford University Press.

Gava, I., Barbui, C., Aguglia, E., Carlino, D., Churchill, R., De Vanna, M., & McGuire, H. (2007). Psychological treatments versus treatment as usual for obsessive compulsive disorder (OCD). *Cochrane Database of Systematic Reviews*, (2), CD005333. https://doi.org/10.1002/14651858.CD005333.pub2

Goodman, W. K., Price, L. H., Rasmussen, S. A., Mazure, C., Delgado, P., Heninger, G. R., & Charney, D. S. (1989). The Yale-Brown Obsessive Compulsive Scale. II. Validity. *Archives of General Psychiatry*, *46*(11), 1012–1016. https://doi.org/10.1001/archpsyc.1989.01810110054008

Goodman, W. K., Price, L. H., Rasmussen, S. A., Mazure, C., Fleischmann, R. L., Hill, C. L., Heninger, G. R., & Charney, D. S. (1989). The Yale-Brown Obsessive Compulsive Scale. I. *Development, Use, and Reliability. Archives of General Psychiatry*, *46*(11), 1006–1011. https://doi.org/10.1001/archpsyc.1989.01810110048007

Hahn, T., Kircher, T., Straube, B., Wittchen, H. U., Konrad, C., Ströhle, A., Wittmann, A., Pfleiderer, B., Reif, A., Arolt, V., & Lueken, U. (2015). Predicting treatment response to cognitive behavioral therapy in panic disorder with agoraphobia by integrating local neural information. *JAMA Psychiatry*, *72*(1), 68–74. https://doi.org/10.1001/jamapsychiatry.2014.1741

Hilbert, K., Kunas, S. L., Lueken, U., Kathmann, N., Fydrich, T., & Fehm, L. (2020). Predicting cognitive behavioral therapy outcome in the outpatient sector based on clinical routine data: A machine learning approach. *Behaviour Research and Therapy*, *124*, 103530. https://doi.org/10.1016/j.brat.2019.103530

Hilbert, K., & Lueken, U. (2020). Predictive analytics from a mental health perspective [Prädiktive Analytik aus der Perspektive der Klinischen Psychologie und Psychotherapie]. *Verhaltenstherapie*, *30*(1), 8–17. https://doi.org/10.1159/000505302

Hoexter, M. Q., Diniz, J. B., Lopes, A. C., Batistuzzo, M. C., Shavitt, R. G., Dougherty, D. D., Duran, F. L., Bressan, R. A., Busatto, G. F., Miguel, E. C., & Sato, J. R. (2015). Orbitofrontal thickness as a measure for treatment response prediction in obsessive-compulsive disorder. *Depression and Anxiety*, *32*(12), 900–908. https://doi.org/10.1002/da.22380

Iniesta, R., Malki, K., Maier, W., Rietschel, M., Mors, O., Hauser, J., Henigsberg, N., Dernovsek, M. Z., Souery, D., Stahl, D., & Dobson, R. (2016). Combining clinical variables to optimize prediction of antidepressant treatment outcomes. *Journal of Psychiatric Research*, *78*, 94–102. https://doi.org/10.1016/j.jpsychires.2016.03.016

Knopp, J., Knowles, S., Bee, P., Lovell, K., & Bower, P. (2013). A systematic review of predictors and moderators of response to psychological therapies in OCD: Do we have enough empirical evidence to target treatment? *Clinical Psychology Review*, *33*(8), 1067–1081. https://doi.org/10.1016/j.cpr.2013.08.008

Lenhard, F., Sauer, S., Andersson, E., Mansson, K. N., Mataix-Cols, D., Ruck, C., & Serlachius, E. (2018). Prediction of outcome in internet-delivered cognitive behaviour therapy for paediatric obsessive-compulsive disorder: A machine learning approach. *International Journal of Methods in Psychiatric Research*, *27*(1). https://doi.org/10.1002/mpr.1576

Lopez-Pina, J. A., Sanchez-Meca, J., Lopez-Lopez, J. A., Marin-Martinez, F., Nunez-Nunez, R. M., Rosa-Alcazar, A. I., Gomez-Conesa, A., & Ferrer-Requena, J. (2015). The Yale-Brown Obsessive Compulsive Scale: A reliability generalization meta-analysis. *Assessment*, *22*(5), 619–628. https://doi.org/10.1177/1073191114551954

Månsson, K. N., Frick, A., Boraxbekk, C. J., Marquand, A. F., Williams, S. C., Carlbring, P., Andersson, G., & Furmark, T. (2015). Predicting long-term outcome of Internet-delivered cognitive behavior therapy for social anxiety disorder using fMRI and support vector machine learning. *Translational Psychiatry*, *5*(3), e530. https://doi.org/10.1038/tp.2015.22

Mataix-Cols, D., Fernandez de la Cruz, L., Nordsletten, A. E., Lenhard, F., Isomura, K., & Simpson, H. B. (2016). Towards an international expert consensus for defining treatment response, remission, recovery and relapse in obsessive-compulsive disorder. *World Psychiatry*, *15*(1), 80–81. https://doi.org/10.1002/wps.20299

Mataix-Cols, D., Marks, I. M., Greist, J. H., Kobak, K. A., & Baer, L. (2002). Obsessive-compulsive symptom dimensions as predictors of compliance with and response to behaviour therapy: Results from a controlled trial. *Psychotherapy and Psychosomatics*, *71*(5), 255–262. https://doi.org/10.1159/000064812

Montgomery, S. A., & Asberg, M. (1979). A new depression scale designed to be sensitive to change. *British Journal of Psychiatry*, *134*(4), 382–389. https://doi.org/10.1192/bjp.134.4.382

Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. *Machine Learning*, *52*(3), 239–281. https://doi.org/10.1023/A:1024068626366

Ost, L. G., Havnen, A., Hansen, B., & Kvale, G. (2015). Cognitive behavioral treatments of obsessive-compulsive disorder. A systematic review and meta-analysis of studies published 1993–2014. *Clinical Psychology Review*, *40*, 156–169. https://doi.org/10.1016/j.cpr.2015.06.003

Ozomaro, U., Wahlestedt, C., & Nemeroff, C. B. (2013). Personalized medicine in psychiatry: Problems and promises. *BMC Medicine*, *11*(1), 132. https://doi.org/10.1186/1741-7015-11-132

Reggente, N., Moody, T. D., Morfini, F., Sheen, C., Rissman, J., O'Neill, J., & Feusner, J. D. (2018). Multivariate resting-state functional connectivity predicts response to cognitive behavioral therapy in obsessive–compulsive disorder. *PNAS*, *115*(9), 2222–2227. https://doi.org/10.1073/pnas.1716686115

Schnack, H. G., & Kahn, R. S. (2016). Detecting neuroimaging biomarkers for psychiatric disorders: Sample size matters. *Frontiers in Psychiatry*, *7*, 50. https://doi.org/10.3389/fpsyt.2016.00050

Simpson, H. B., Maher, M. J., Wang, Y., Bao, Y., Foa, E. B., & Franklin, M. (2011). Patient adherence predicts outcome from cognitive behavioral therapy in obsessive-compulsive disorder. *Journal of Consulting and Clinical Psychology*, *79*(2), 247–252. https://doi.org/10.1037/a0022659

Simpson, H. B., Marcus, S. M., Zuckoff, A., Franklin, M., & Foa, E. B. (2012). Patient adherence to cognitive-behavioral therapy predicts long-term outcome in obsessive-compulsive disorder. *Journal Clinical Psychiatry*, *73*(9), 1265–1266. https://doi.org/10.4088/JCP.12l07879

Springer, K. S., Levy, H. C., & Tolin, D. F. (2018). Remission in CBT for adult anxiety disorders: A meta-analysis. *Clinical Psychology Review*, *61*, 1–8. https://doi.org/10.1016/j.cpr.2018.03.002

Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *Neuroimage*, *180*(Pt A), 68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061

Velten, J., Margraf, J., Benecke, C., Berking, M., In-Albon, T., Lincoln, T., Lutz, W., Schlarb, A., Schöttke, H., Willutzki, U., & Hoyer, J. (2017). Methodenpapier zur Koordination der Datenerhebung und -auswertung an Hochschul- und Ausbildungsambulanzen für Psychotherapie (KODAP). *Zeitschrift für Klinische Psychologie und Psychotherapie*, *46*(3), 169–175. https://doi.org/10.1026/1616-3443/a000431