European Archives of Psychiatry and Clinical Neuroscience (2020) 270:153–168
https://doi.org/10.1007/s00406-018-0967-2

ORIGINAL PAPER

Predicting instructed simulation and dissimulation when screening for depressive symptoms

Stephan Goerigk[1,2,3] · Sven Hilbert[1,4] · Andrea Jobst[2] · Peter Falkai[2] · Markus Bühner[1] · Clemens Stachl[1] · Bernd Bischl[5] · Stefan Coors[5] · Thomas Ehring[6] · Frank Padberg[2] · Nina Sarubin[1,2,3]

Received: 1 March 2018 / Accepted: 6 December 2018 / Published online: 12 December 2018
© Springer-Verlag GmbH Germany, part of Springer Nature 2018

Abstract

The intentional distortion of test results presents a fundamental problem to self-report-based psychiatric assessment, such as screening for depressive symptoms. The first objective of the study was to clarify whether depressed patients like healthy controls possess both the cognitive ability and motivation to deliberately influence results of commonly used screening measures. The second objective was the construction of a method derived directly from within the test takers' responses to systematically detect faking behavior. Supervised machine learning algorithms posit the potential to empirically learn the implicit interconnections between responses, which shape detectable faking patterns. In a standardized design, faking bad and faking good were experimentally induced in a matched sample of 150 depressed and 150 healthy subjects. Participants completed commonly used questionnaires to detect depressive and associated symptoms. Group differences throughout experimental conditions were evaluated using linear mixed-models. Machine learning algorithms were trained on the test results and compared regarding their capacity to systematically predict distortions in response behavior in two scenarios: (1) differentiation of authentic patient responses from simulated responses of healthy participants; (2) differentiation of authentic patient responses from dissimulated patient responses. Statistically significant convergence of the test scores in both faking conditions suggests that both depressive patients and healthy controls have the cognitive ability as well as the motivational compliance to alter their test results. Evaluation of the algorithmic capability to detect faking behavior yielded ideal predictive accuracies of up to 89%. Implications of the findings, as well as future research objectives are discussed. *Trial Registration* The study was pre-registered at the German registry for clinical trials (Deutsches Register klinischer Studien, DRKS; DRKS00007708).

Keywords Depression · Assessment · Faking · Machine learning · Response patterns

Introduction

In psychology and the social sciences, there exists a predominant interest in latent constructs, which typically cannot physically be measured. Concepts like a person's personality, attitudes and current mental states are frequently the basis

Stephan Goerigk, Sven Hilbert, Frank Padberg and Nina Sarubin authors contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00406-018-0967-2) contains supplementary material, which is available to authorized users.

✉ Stephan Goerigk
stephan.goerigk@med.uni-muenchen.de

[1] Department of Psychological Methodology and Assessment, Ludwig-Maximilians-University Munich, Munich, Germany

[2] Department of Psychiatry and Psychotherapy, Ludwig-Ma ximilians-University, Nussbaumstrasse 7, 80336 Munich, Germany

[3] Hochschule Fresenius, University of Applied Sciences, Munich, Germany

[4] Faculty of Psychology, Educational Science and Sport Science, University of Regensburg, Regensburg, Germany

[5] Department of Statistics, Ludwig-Maximilians-University, Munich, Germany

[6] Department of Clinical Psychology and Psychotherapy, Lud wig-Maximilians-University, Munich, Germany

for profound economical, medical, and social decisions, such as personnel selection, continuation or discontinuation of treatments as well as claims to financial support in the health care system. Hence, their quantitative specification is one of the most sought-after goals of psychiatric and psychological assessment. For that matter, the objective of all measurement efforts is identifying the true magnitude of the parameters, which uni-dimensionally reflects the characteristics of these unobservable concepts. Self-report measures are the typical approach to acquiring as pristine as possible valuations of these true values. However, they are particularly susceptible to distortions of the test taker's response tendencies. One such distortion is the faking of test results. If a person applies faking strategies, the measured value is merely a reflection of the sum of the real value and a faking-bias. Especially in contexts, where decisional consequences have high impact on a client's life, a valid and flawless measurement is the foundation of a fair assessment.

The term faking refers to a deliberate distortion of the answers in a psychological test towards a more positive or more negative representation of oneself [1]. According to MacCann, Ziegler, and Roberts [2], similarities shared by most definitions of the concept include four core aspects. First, rather than a personal trait, faking is manifested as a particular display of behavior. Second, faking is goal-oriented and requires motivation. Third, the impressions provided by the faking behavior are enhanced and unlike a person's real characteristics. Lastly, faking engages in an interaction between the individual's current situation and its permanent traits. Two kinds of faking have to be differentiated: faking-good defines the way in which test takers render themselves more competent or socially desirable, respectively, concealing symptoms expected to have a negative impact on the test results [3]. Thus, the presented description is more positively accentuated than in reality. Inversely, faking bad is the simulation and malingering of psychological symptoms or vulnerabilities, as well as the misrepresentation of personal characteristics [4]. The terms can be further differentiated with respect to the factual medical state of the subject. In case of a healthy subject, simulation describes the feigning of symptoms that do not actually exist, whereas aggravation is the negatively distorted display of a patient's real symptoms [5]. The term dissimulation is usually applied for positively distorted displays of characteristics of both patients and healthy subjects.

Faking in a clinical-psychiatric context

Since psychiatric disorders are inherently difficult to objectify, possibilities for falsification are more potent than diagnosing most somatic disorders. For a long time, neither the motives nor the specific competence for faking of psychiatric symptoms [6] have been unambiguously demonstrated for depressed patients. The intentional distortion of test results has previously been described as a spurious measurement error resulting from the interaction between the person's disposition and the situational demand [7, 8]. As such, faking can be understood as a specific cognitive ability. In their process model of applicant faking behavior, Tourangeaou and Rasinski [9] identify three central cognitive domains for goal-driven and realistic distortion of a response: comprehension, importance, classification, and mapping of the desired outcome on the response continuum. All three aspects depend significantly on specific person characteristics. Characteristics that are particularly relevant for the capability to intentionally distort test results, and that could potentially be lacking in depressed patients include: general mental ability [10, 11], ability to reflect and self-understanding [12], motivation [13], frame of reference [14] and self-concept [15]. There has been a predominant assumption that patients neither possess the motivation to malinger their test results due to their psychological strain, nor the cognitive capacities [16, 17] necessary to successfully employ faking strategies [18]. According to Kaplan and Sadock [19], depressed individuals often report perceiving tasks to be more complicated or difficult than described by healthy controls and typically require more encouragement to attempt solving them. In addition, it is well established that depressed patients fail to put forth the maximum effort they are theoretically capable to muster [20]. Clearly, the aggravated as well as the simulated presentation of symptoms requires a change of perspective and the imagination of alternate states of mind. However, both, non-psychotic as well as psychotic depressed patients have been shown to be significantly impaired on tests, that involve social-cognitive and social-perceptual components of theory of mind (ToM) [21, 22]. Especially chronically depressed patients have frequently been linked to Piaget's concept of preoperative thinking [12, 23], which involves egocentrism in the views of self and others. On the other hand, patients with acute forms or a history of psychological disorders often become experts for the relevant symptoms and are thus more competent in depictions of themselves in an authentic manner.

According to Ziegler [24], faking gains particular relevance in so-called high-stakes-situations, which are characterized by substantial consequences for the individual. In clinical contexts, motivation for misrepresentation is usually linked to secondary gains [25], manifested in social attention [26], exoneration of unpleasant impositions [27] as well as materialistic benefits like financial support and compensation [28]. However, invalid methods of symptomatic screening come with a price, epitomized in a recent study by Lieb and colleagues [29], who showed that all successful suicide attempts in a sample of 16,251 inpatients at the university hospital in Munich over the span of 20 years were carried out on pre-approved clinical leave of absence or external

home visits, thus posterior to misjudgements of the patients' true symptomatic status.

Prevalence of faking and socio-economic implications

Conservative estimates presented by Griffin, Normington, May, and Glassmire [30] conjecture that about a fifth of all applicants for full-rate reduced-earning-capacity pensions exaggerate their symptoms and disabilities. These results are in line with the work of Cima and colleagues [31] who postulate misrepresentation of psychiatric symptoms for 16% of the patients in forensic contexts. The prevalence of faking application cases strongly depends on the situation. According to Vossler-Thies, Stevens, Engel, and Licha [32], rates for faking-bad strategies in psychological assessment reports reach up to 29–30%. Other studies postulate substantially higher estimations: Stevens, Friedel, Mehren and Merten [33] found indication for suboptimal performance in 44% of subjects in psychiatric assessments using the Word Memory Test (WMT) and the Medical Symptom Validity Test (MSVT) for effort testing, while Greve, Ord, Bianchini and Curtis [34] claim prevalence rates of feigned psychological symptoms between 20–50% in 508 applications for invalidity pensions, depending on the test procedure. Faking in personality measures is estimated to be even more frequent, namely shown by approximately 50% of test subjects [35].

By now the largest proportion in the medical spectrum of invalidity pensions is due to psychiatric disorders, making up 36.5% of male and 49.0% of female early retirement causes. Figure 1 displays the distribution of causes, which lead to full-rate reduced-earning-capacity pensions, issued in the year 2016 [36]. The major share of rehabilitation and reduced-earning-capacity support is due to affective, stress-related and adjustment disorders. Moreover, individual psychiatric diagnoses are dominated by depressive disorders.

If the more pessimistic estimates on the prevalence of faking of 30–50% are correct, the relevance of studying faking in the context of mental health assessment, including its detection and prevention, becomes apparent. Furthermore, the problem of deliberate misrepresentation in assessment situations is not only relevant to the assessment of mental health but also to many other subdisciplines in medicine [32].

Prior efforts to mitigate faking: advantages of a machine learning approach

A number of approaches have been taken by researchers and practitioners to address the problem of faking. These efforts can be loosely categorized into methods designed to control faking, methods designed to identify individual faking behavior, and methods designed to correct for faking. Control strategies attempt to either reduce faking behavior or

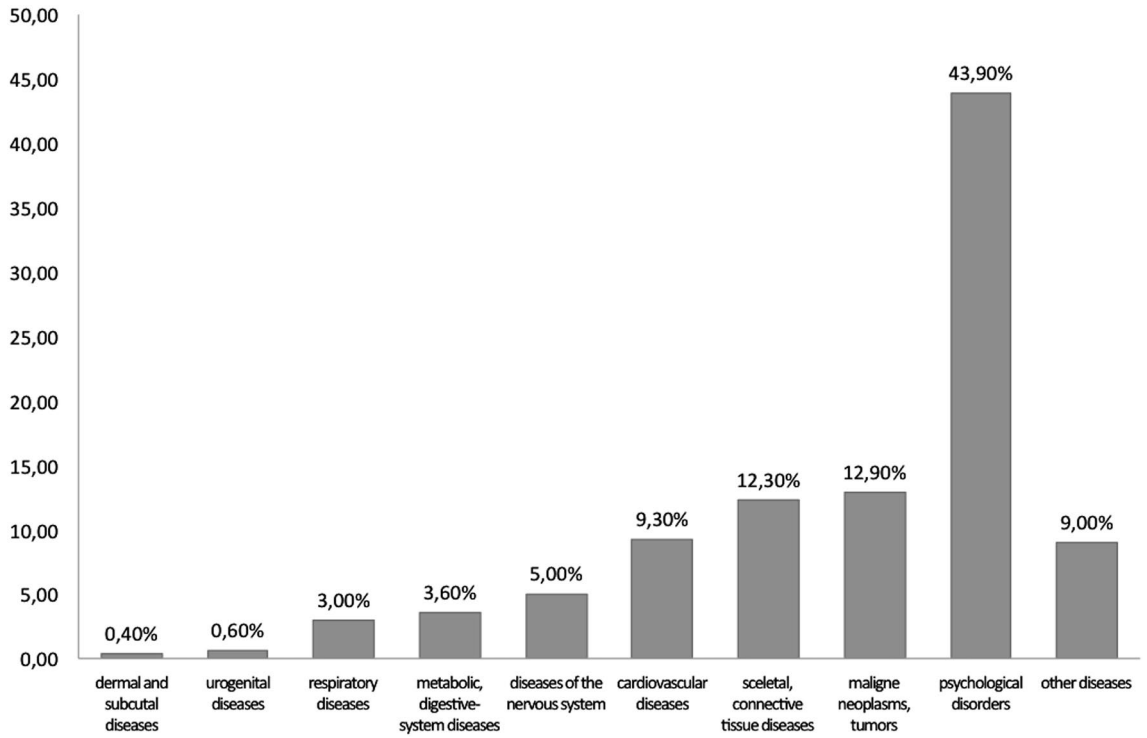


Fig. 1 Medical spectrum of causes for full-rate reduced-earning-capacity pensions in Germany, 2016

alternatively to prevent it altogether [37]. Detection efforts attempt to identify test takers likely to be representing themselves in an incorrect way [38–40]. Correction efforts simultaneously obtain social desirability covariates used to statistically depurate individual scores [41].

However, there are several limitations to the applicability of the methods developed to date; for example, the central problem with the implementation of warnings and exhorting instructions is the potential suppression of honest responses. As such, it is not clear if the most honest test takers would curb responses to avoid criticism or accusations. It has been theorized that warnings merely lead to more subtle and sophisticated versions of faking [42]. Furthermore, test takers may exhibit a negative attitude towards warnings [43]. Control methods that are based on the item format were questioned by the results of head-to-head comparisons of forced-choice multidimensional items with traditional Likert-type items, indicating that both formats were equally biased by directed faking [44]. Response time limitations are based on the assumption that additional cognitive resources and thus longer processing times are required to feign a test response. Holden, Wood, and Tomashewski [39] could not find support for a performance decrease in response time-limiting paradigms, implying, that lying rather involves primitive cognitive processing whilst Komar, Robie, and Taggar [45] showed, that speeding through non-cognitive measures could significantly reduce faking, but the effect was only seen in individuals with low cognitive abilities. Methods that attempt to detect faking so far lacked satisfying classification accuracy. A within-person difference score index (DSI) was presented as a method to identify faking within single-test administrations. Baseline RTs were found to be significantly shorter than RTs in misrepresentation conditions, resulting in classification accuracy rates of up to 71.1% for experimentally induced faking designs [46]. Using measures of the Item Response Theory, item characteristics and response processes can be examined to detect misrepresentation. Holden and Book [47] used hybrid Rasch-latent class modeling to discriminate fakers from non-fakers in an experimentally induced scenario. A 3-Rasch class model reached classification accuracy rates of 68.4% for standard responding, 78.4% for faking good and 50.4% for faking bad.

In conclusion, even though known approaches do help address the faking problem, the extent of their utility is in need of improvement. New solutions have to be provided to withstand the aspired validity standards of psychometric quality criteria, like so formulated for example in the German S3 national guidelines for unipolar depression [48].

Criteria for a more suitable detection method are (1) a considerable increase in discriminative accuracy, (2) versatility for a wide range of different contexts, and (3) high applicability for already existing tests. This study suggests the construction of a method derived directly from within the test taker's responses to detect faking behavior. The test data include numerous implicit interconnections and clusters between item responses, which could potentially shape a detectable faking pattern. Since there are no established cutoffs for satisfying precision of faking detection methods, keeping in mind the scope of high-stake decisions, we suggest a minimum of 90% accuracy with a false-positive-rate of no more than 5% as the criterion for an acceptable classification result.

Machine learning is a specific subset of methods from the artificial intelligence (AI) domain, which enable algorithmic classifiers to learn based on experience rather than being explicitly programmed [49]. Specifically, so-called supervised or inductive machine learning [50] appears promising to be applied to the detection of faking. This paradigm is distinguished by the search for classification algorithms, capable of reasoning from externally supplied instances to produce generalizable hypotheses, which in turn predict the class of future instances. Kotsiantis et al. [51] as well as Urbanowicz and Moore [52] have presented comprehensive reviews on this method. Generally speaking, machine learning techniques provide a competitive advantage in decision problems, either when there is no human expertise available, when data are too complex to penetrate, or where humans cannot explain their own expertise or skill [53]. The predictive modeling approach has previously been applied to comparable fields, such as deceit detection [54, 55], and lie detection [56] as well as to numerous other areas of scientific and economic interest. Specific uses in the medical sector include the detection of healthcare fraud [57–59] and disease prediction [60, 61]. While in the present study predictive modeling is applied to data of psychiatric screening measures, this principle of faking detection is generalizable to other assessment methods as well (observer ratings, behavioral observation). Especially interviews, which usually constitute the gold-standard for clinical decision-making could benefit from more precise detection methods for deliberately distorted responses. Another advantage of machine learning algorithms is their dynamic ability for continuous improvement: as in the present study, a restricted sample size is available to train and validate the classifiers, but for any further supply of data to the set, the algorithms have the potential to keep learning to improve their discriminative accuracy [62], as long as data are collected in the same measurement paradigm or an improved version of it.

Hypotheses and summary of study objectives

To our knowledge, no study to date has investigated the intentional distortion of test results when screening for depressive symptoms in a comparable standardized design. First, we hypothesized that it would be possible to intentionally alter the outcome of the included tests, both in a positive

and in a negative direction, as compared to the authentic result. The testing of the primary hypothesis simultaneously serves as a manipulation check of the faking-induction. Secondly, we hypothesized that depressed patients compared to healthy controls would have reduced ability to distort the results of screening measures. Finally, it is investigated whether faking can systematically be detected in a supervised learning paradigm. As a secondary objective, test items that are particularly informative for the distinction between authentic and distorted test results are sought to be identified.

Methods

Sample characteristics

A total number of $N = 300$ participants were recruited including $n_d = 150$ patients with depressive disorders and $n_c = 150$ healthy controls. Patients were contacted in the context of both inpatient and outpatient treatment programs at the Hospital for Psychiatry and Psychotherapy of the Ludwig-Maximilians University, Munich. Healthy participants were contacted via bulletins or active personal approach. Subsamples (depressed and healthy subjects) were matched by age, educational status, and gender to minimize confounding effects. Age-matching tolerance comprised a span of ± 5 years and educational-matching included three different levels of academic achievement. Descriptive statistics for demographic data of all 300 participants included in the matched sample are displayed in Table 1. A total of 143 patients (95%) were on a psychiatric medication schedule at the time of assessment. All participants had to be between 18 and 60 years of age and fluent in reading and writing German, so all subjects could correctly understand and adequately respond to instructions and questionnaire items. Patients had to have been previously diagnosed with depression according to DSM-IV criteria (296.2; 296.3) by proficient and experienced psychiatrists or psychologists using a version of the Structured Clinical Interview for the DSM-IV (SCID-I [63]). Healthy controls by definition had to be free of acute and antecedent psychiatric disorders.

Exclusion criteria for the patient group comprised acute suicidal tendencies, addiction disorders, physical disease including acute and chronic infections, autoimmune disease, pregnancy, psychotic disorders, bipolar disorder as well as primary diagnosis of panic disorder, generalized anxiety disorder (GAD), social anxiety disorder, and posttraumatic stress disorder (PTSD). Exclusion criteria for the healthy control group included consumption of psychotropic drugs, physiological diseases including acute and chronic infections, autoimmune disease and pregnancy.

Measures

In preparation of the study, an inquiry had been made at the German Insurance Association (GDV; German: Gesamtverband der Deutschen Versicherungswirtschaft, e.V.) to ascertain the current state of the art for psychological tests, used in the clinical assessment for depression in the context of early retirement. Out of 191 tests which are applied in psychological surveys for early retirement due to depression, the measures selected for this study are the most commonly used screening tests from the field's personality inventory, cognitive abilities, specific clinical depression, and mental state/emotional competence/pain. These include the German version of the Beck-Depression Inventory (BDI-II) [64], the revised German version of the Symptom-Checklist-90-R (SCL-90-R) [65], the German version of the Structured Inventory of Malingered Symptomatology (SFSS) [31], the German version of the NEO-Five-Factor Inventory (NEO-FFI) [66], and the Attention and Concentration Test (d2-R)

Table 1 Demographic and clinical characteristics of the study sample

Characteristic	No. (%)		
	Patients ($n = 150$)	Controls ($n = 150$)	Total ($n = 300$)
Demographics			
Age, *M* (SD)	39.3 (12.25)	42.1 (12.38)	40.9 (12.32)
Women	76 (51%)	76 (51%)	152 (51%)
Education			
HS	23 (15%)	23 (15%)	46 (15%)
RS	52 (35%)	52 (35%)	104 (35%)
Gym	75 (50%)	75 (50%)	150 (50%)
Clinical characteristics			
Psychiatric medication	0 (0%)	143 (95%)	143 (48%)
BDI–II sum score, *M* (SD)	27.02 (11.97)	3.09 (4.00)	13.20 (14.48)

$N = 300$; *M* sample mean, *SD* standard deviation, *HS* Hauptschule (secondary school, year 5–9), *RS* realschule (secondary school, year 5 to 11), *Gym* gymnasium (secondary school, year 5–12)

[67]. While the BDI-II (21 items, multi-choice from zero to three), the NEO-FFI (60 items in five dimensions, five-point Likert scale) and the d2-R (six performance measures) were used in full length, only the depression scale of the SCL-90-R (13 items, five-point Likert scale) and only the affective disorder scale of the SFSS (14 items, binary forced-choice format) were included.

Certain other measures such as the Minnesota Multiphasic Personality Inventories (MMPI-I and MMPI-II) [68, 69] readily include built-in faking precautions. However, according to Vossler-Thies and colleagues [32], these well-validated tests are often not time-economic enough for screenings in commonplace clinical contexts and were therefore not included in the present study.

Experimental design and testing procedure

Prior to testing, the severity of the current episode of depression was screened using the International Diagnostic Checklists for depression (IDCL) [70]. In a standardized design, faking behavior was experimentally induced: both patients and controls were instructed to answer each test with a specific mindset in each one of three different conditions, the faking-good-, faking-bad- and true-condition (within-subjects design). In the faking-good condition, participants were asked to imagine a particular situation, in which making a good impression in a plausible manner and presenting only little psychological strain would award them with advantageous consequences (getting the dream job). In the faking-bad condition, participants were instructed to present themselves as more psychologically strained in prospect of a profitable outcome (early retirement). In the true-condition, participants were asked to present themselves in a realistic manner without any exaggerations and answer in a way that seemed subjectively appropriate to them. The exact wording of the instructions can be found in the supplementary materials (Online Resource 1). To ensure realistic response behavior, participants were told that a test expert would go through the test results after the testing and that they should try not to be caught cheating. Patients and controls were tested in each of the conditions sequentially. The order of the three conditions was systematically randomized to eliminate response bias and context effects. The order of the five questionnaires (BDI-II, SCL-90-R, SFSS, d2-R and NEO-FFI) was also systematically randomized within each of the conditions. Hence, each subject answered each questionnaire three times. The complete study design is displayed in Fig. 2.

Instructions were given verbally and in writing just before the start of each condition. All tests were presented in digital format, since there are no empirical differences in the faking behavior regarding computerized and paper–pencil versions of tests investigated up to this point [71]. The study was conducted in accordance with the Declaration of Helsinki. Formal approval to conduct the investigation has been obtained from the institutional review board (IRB).

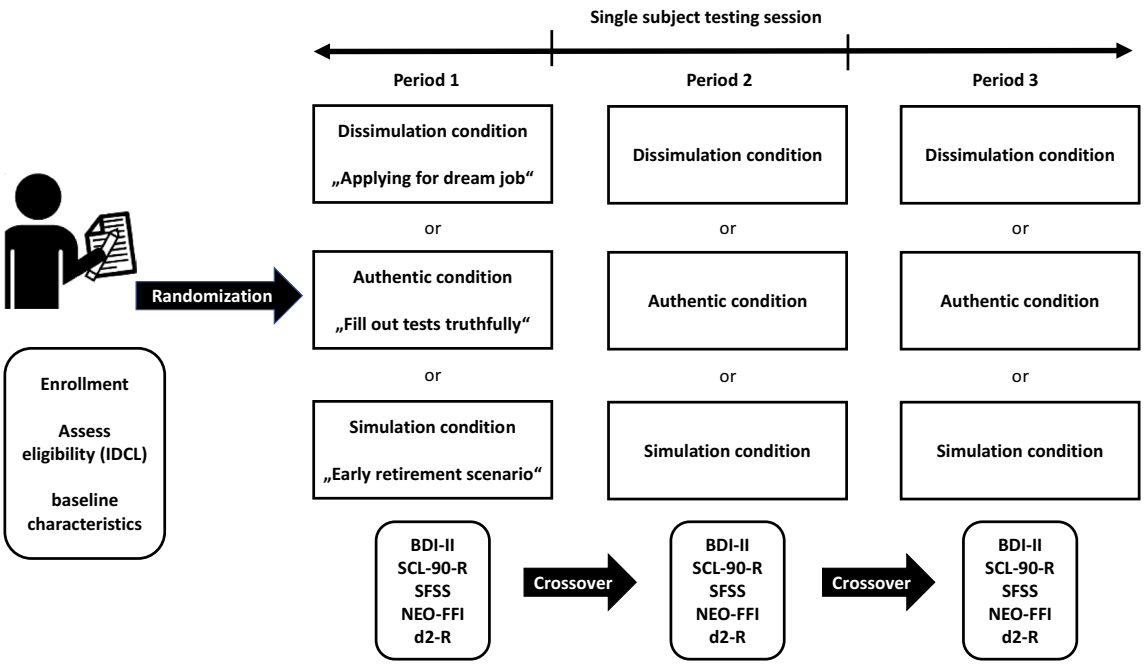


Fig. 2 Study design displaying the three-way randomized blocked cross-over trial. Every subject completed the five tests in randomized order in each of the three conditions. The measurements were taken in a single testing session

Statistical analyses and predictive modeling

Effects of the experimental faking-induction were modeled using linear mixed regression analyses, which included the faking-condition as a level-1 variable and the diagnosis as a level-2 variable, as well as their cross-level interaction [72]. Individual affective states at baseline as well as interpersonal differences in response behavior were accounted for by treating both the model intercept and the change rate as random effects. Model parameters were determined using restricted maximum likelihood (REML) estimations. The sum of the individual deviations from the authentic test scores in both faking conditions was computed as a measure of the magnitude of the faking effect, to counteract floor and ceiling effects. The total faking effect was compared between patient and control sample using robust independent t tests with standard errors derived from 2000 bootstrap samples. Effects were regarded as significant at type-1-error probability of $\alpha = 0.01$. Regression-based analyses were performed using the lme4 package for the R software environment for statistical computing [73].

Predictive modeling was applied to detect either faking-bad or faking-good in the test responses. Binary classifications were chosen over a three-class classification approach to develop as specific tools as possible for the scenarios described above. The item scores from the five tests served as the predictors in the classification procedure. A total of 115 predictors were entered into the analysis, prior to any filtering or variable selection. The criterion variable was coded for simulation vs. authentic responding or dissimulation vs. authentic responding, respectively. From the specific combinations of the empirically observed test scores, classification functions were inferred to correctly determine class labels for unseen instances. Five candidate models, including logistic regression, support vector machines [74], random forests [75], tree-based stochastic gradient boosting [76], and an XgBoost model [77] were entered into a benchmark-experiment framework to find an optimal classifier. All five models were trained and validated due to their capacity to detect response distortions in two different conditions:

1. Simulation was detected in a mixed sample of authentically responding patients and malingering controls to reflect the assessment for the admission to invalidity pensions.
2. Dissimulation was detected in a full patient sample to reflect symptom severity assessments prior to authorization of clinical leaves of absence in psychiatric hospital settings.

To obtain honest performance estimates and reduce overfitting for all learners, model-tuning and evaluation were conducted in a nested-resampling paradigm. In this method, the data are split in a series of training- and test-sets, instead of relying on the representativeness of model evaluation in a single split [78]. All models were built according to the process displayed in Fig. 3. For information on optimal hyperparameter-tuning and parallel computing see Online

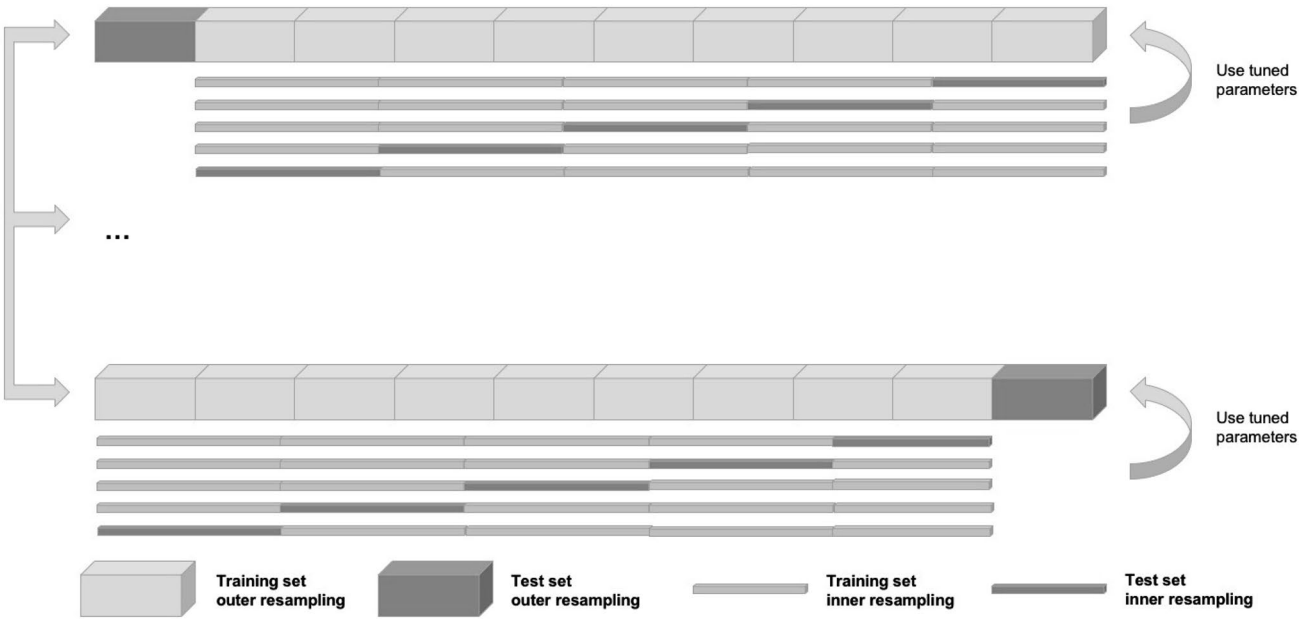


Fig. 3 Machine learning model building procedure showing nested resampling for fivefold-CV on the inner loop and the first and last fold of tenfold-CV on the outer loop. The process was performed ten times by means of repeated cross-validation. "…" represents another eight sets of outer resampling

Resource 2. The selection of an optimal classifier from the benchmark experiment would ideally require another cycle in the cross-validation scheme to avoid overfitting. However, this solution was difficult to implement because of the limited sample size. To still make an informed decision, model performances were compared using Bonferroni-corrected Wilcoxon tests on the mean misclassification error (mmce) of the classifiers with standard errors derived from the 100 repeated resampling iterations of model evaluation. Significance of predictive accuracies was determined by testing each model's performance against that of the weakest classifier.

In addition to using the mean misclassification error (mmce) as the standard optimization criterion, a custom performance measure was created to account for the severity of false-positive classifications in the context of faking detection: when the misclassification threshold exceeded 0.05, which was regarded as an acceptable level of misclassification probability, the measure was penalized by adding 1 to the false-positive-rate (FPR).

To increase interpretability of the machine learning models, feature importance was estimated for the most predictive classifiers, to ascertain, which of the included test items posit the highest potential to selectively differentiate between distorted and authentic response sets. Estimation of each item's significance for the classification result was based on the permutation importance method (PIMP). This approach is based on repeatedly permuting the test scores of a predictor, by which its potential to systematically predict an outcome is rendered un-informative [79]. The contribution of each predictor was operationalized by the loss in performance of the model, after it had been permuted. Predictive modeling was performed using the mlr machine learning package in R [80].

Results

Average test scores from the three conditions as well as the quantitative differences between patients and controls are displayed in Table 2 and Fig. 4.

Linear mixed regression analyses showed significant main effect differences between all faking conditions in all tests, suggesting that experimental mindsets for deliberate distortions could successfully be induced as well as demonstrating the fakeability of these commonly used screening measures. The only exceptions were statistically non-distinguishable d2-R performance measures in the authentic and the faking good condition. Pairwise contrasts in the regression equation (dummy-coded) showed significant interaction effects in response behavior differences for the faking-bad vs. authentic as well as the faking-good vs. authentic conditions on all self-report scales except for the agreeableness dimension of the NEO-FFI, reflecting score convergence of patients and controls in both faking conditions as well as diagnostically valid score divergence in the authentic condition. The results of the linear mixed regression analyses are included in Table 3 for selected outcomes. The results for the remaining NEO-FFI scales and d2-R performance parameters can be found in the supplementary materials (Online Resource 3). Comparisons of the total faking effect between patients and controls using robust t-tests were not significant for all self-reported measures, except for the agreeableness dimension of the NEO-FFI before correction for multiple comparisons

Table 2 Test results from the experimental induction conditions

Scale	Authentic		Faking-good		Faking-bad	
	Patients *M* (SD)	Controls *M* (SD)	Patients *M* (SD)	Controls *M* (SD)	Patients *M* (SD)	Controls *M* (SD)
BDI-II	27.02 (11.97)	3.09 (4.00)	4.75 (9.91)	1.34 (5.03)	48.43 (11.82)	44.67 (16.11)
SCL-90-R Depression	24.47 (11.38)	2.89 (3.48)	4.18 (7.51)	1.41 (3.46)	41.70 (9.29)	37.80 (13.82)
SFSS Affective disorder	7.05 (2.06)	4.06 (1.45)	3.85 (1.74)	2.75 (1.25)	7.74 (2.02)	8.65 (2.43)
d2-R Concentration	174.07 (49.14)	197.95 (45.80)	172.00 (63.32)	195.59 (46.46)	86.81 (75.06)	160.87 (78.06)
d2-R Accuracy	189.36 (48.81)	212.34 (43.20)	183.73 (63.52)	208.46 (48.49)	138.39 (67.19)	194.13 (55.52)
d2-R Errors (%)	9.26 (11.18)	6.41 (7.00)	10.48 (19.38)	7.02 (7.02)	37.71 (38.33)	19.14 (32.55)
d2-R Omission error	15.44 (19.57)	13.62 (19.35)	14.28 (13.62)	13.85 (17.06)	35.08 (52.88)	24.33 (40.12)
d2-R Confusion error	0.86 (1.51)	0.51 (1.04)	0.92 (1.49)	0.62 (1.37)	19.08 (30.90)	9.82 (26.61)
d2-R Letter error	0.26 (0.91)	0.10 (0.092)	0.47 (1.86)	0.28 (1.58)	12.50 (22.41)	5.86 (15.81)
NEO-FFI Neuroticism	32.17 (7.72)	13.82 (7.33)	12.68 (9.16)	7.38 (5.50)	40.82 (6.33)	38.26 (9.53)
NEO-FFI Extraversion	18.66 (8.13)	31.60 (6.48)	32.38 (7.38)	36.10 (6.15)	9.47 (6.21)	10.96 (8,66)
NEO-FFI Openness	26.67 (7.58)	32.17 (6.51)	30.93 (5.77)	32.82 (5.51)	17.76 (7.34)	17.92 (7.71)
NEO-FFI Agreeableness	30.70 (6.21)	34.30 (5.97)	35.47 (5.39)	37.62 (6.20)	20.72 (8.44)	20.20 (8.99)
NEO-FFI Conscientiousness	28.27 (7.72)	36.34 (5.84)	41.37 (6.76)	43.01 (6.00)	20.20 (9.04)	16.21 (10.45)

Raw scores. $N = 300$ (150 patients, 150 controls), *M* sample mean, *SD* standard deviation;

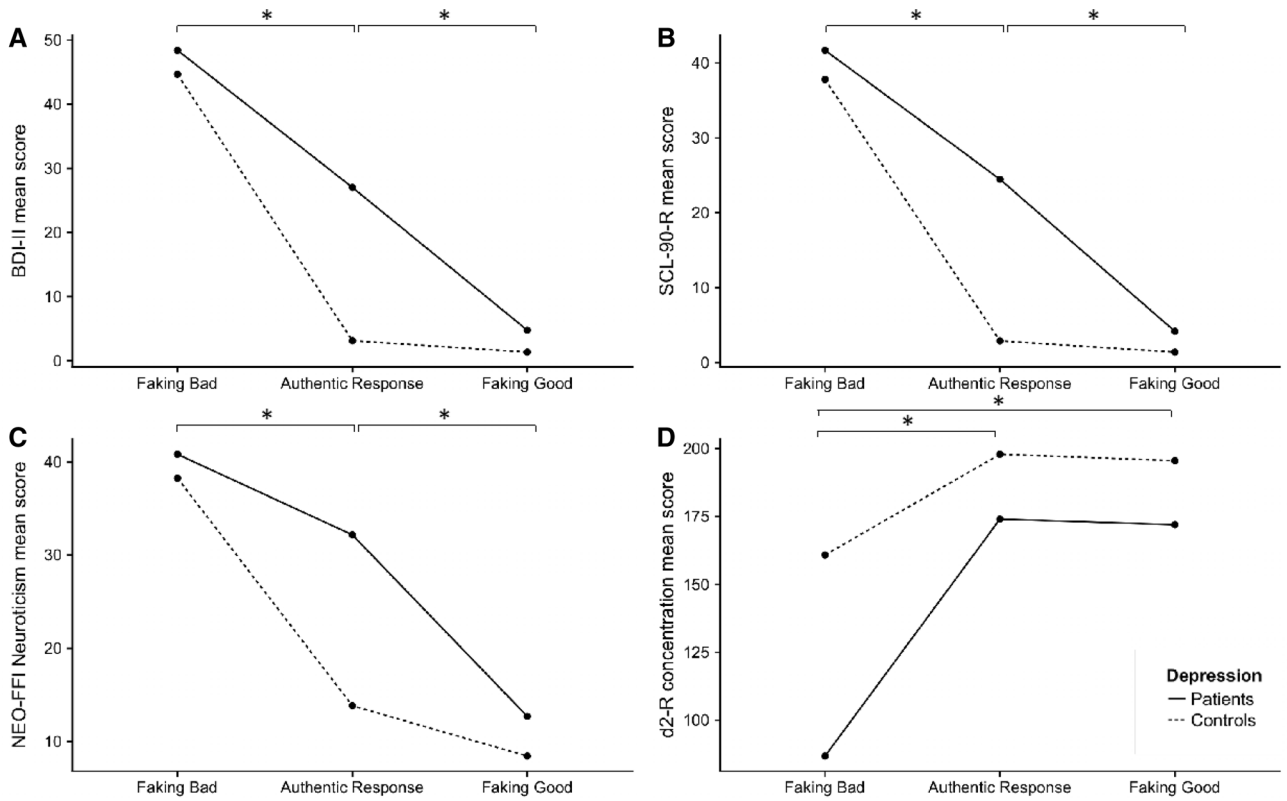


Fig. 4 Mean score differences across conditions. Significant interaction effects are marked by parenthesis. Asterisk shows significance on $p < .01$. **a** Differences in BDI-II mean score. **b** Differences in SCL-90-R (depression scale) mean score. **c** Differences in NEO-FFI neuroticism dimension mean score (other personality dimensions can be found in the supplementary materials of the article). **d** d2-R concentration performance mean score (other performance measures dimensions can be found in Online Resource 3)

($p = 0.033$), where controls showed stronger distortion tendencies. In the d2-R performance test, controls showed stronger faking effects for accuracy ($p < 0.001$), confusion errors ($p = 0.01$), letter errors ($p = 0.005$) and omission errors ($p = 0.037$), while patients showed stronger faking effects for concentration ($p < 0.001$). All comparisons are summarized in Online Resource 3.

Predictive performance of the classifiers is summarized in Table 4. Peak classification rates were achieved by the random forest model in the simulation scenario and the XgBoost model in the dissimulation scenario. All learners showed performance superior to the logistic regression model, no significant differences could be found between performances of remaining classifiers. Inference statistical performance comparisons of the classifiers are summarized in Online Resource 4. Within each scenario a minimum of 89% of the cases could correctly be classified by the respective algorithms.

Predictor-importance plots of the most accurate models in each scenario are displayed in Fig. 5. The listed predictors represent a class of items with commonly altered responses by test takers in the faking scenario, as they would not be answered the same way under authentic conditions. Predictor-importance estimation showed high contributions of translucent, strongly worded items as well as invariability statements. For example, the 47th SFSS item "I feel depressed all of the time." was affirmed in only five cases by depressed patients in the authentic condition; however, controls affirmed the statement 111 times in the simulation condition. Similarly, in the performance test d2-R, the letter confusion error, in which the letter "g" is crossed out instead of the letter "d", was made very rarely under authentic conditions ($M = 0.255$), whereas for controls, who in theory would be expected to show better cognitive performance, the error was more frequently observed ($M = 5.86$).

After training the algorithms for the minimization of the penalized false-positive-rate (pFPR), 58.1% of the patients were correctly identified in the dissimulation scenario, when only those subjects were positively classified who had a class-probability of above 95%, while in the simulation scenario 49.5% of the controls were correctly identified.

Since the SFSS questionnaire inherently includes items that are sensitive to the malingering of clinical

Table 3 Results of linear mixed model analyses

	Estimate	SE	*df*	*t*	*p*
BDI-II					
Intercept	27.02	0.86	1053.80	31.33	< 0.001
Authentic vs. faking bad	21.41	1.19	706.00	18.00	< 0.001
Authentic vs. faking good	− 22.27	1.19	706.00	− 18.73	< 0.001
Faking bad vs. faking good	− 43.68	1.19	706.00	− 18.73	< 0.001
Diagnosis	− 23.93	1.13	1053.80	− 21.09	< 0.001
Diagnosis × faking bad vs. diagnosis × authentic	20.17	1.56	706.00	12.89	< 0.001
Diagnosis × faking good vs. diagnosis × authentic	20.52	1.56	706.00	13.12	< 0.001
Diagnosis × faking bad vs. diagnosis × faking good	0.348	1.56	706.00	0.223	0.82
Conditional R^2	0.79				
SCL-90-R					
Intercept	24.47	0.73	1057.40	33.56	< 0.001
Faking bad vs. authentic	17.23	1.02	706.00	16.95	< 0.001
Faking good vs. authentic	− 20.29	1.02	706.00	− 19.96	< 0.001
Faking bad vs. faking good	− 37.52	1.02	706.00	− 36.91	< 0.001
Diagnosis	− 21.57	0.96	1057.40	− 22.49	< 0.001
Diagnosis × faking bad vs. diagnosis × authentic	17.67	1.34	706.00	13.21	< 0.001
Diagnosis × faking good vs. diagnosis × authentic	18.81	1.34	706.00	14.06	< 0.001
Diagnosis × faking bad vs. diagnosis × faking good	1.13	1.34	706.00	0.85	0.397
Conditional R^2	0.78				
NEO-FFI Neuroticism					
Intercept	32.17	0.65	1054.90	49.19	< 0.001
Faking bad vs. authentic	8.65	0.90	706.00	9.57	< 0.001
Faking good vs. authentic	− 19.49	0.90	706.00	− 21.56	< 0.001
Faking bad vs. faking good	− 28.14	0.90	706.00	− 31.12	< 0.001
Diagnosis	− 18.34	0.86	1054.90	− 21.32	< 0.001
Diagnosis × faking bad vs. diagnosis × authentic	15.78	1.19	706.00	13.26	< 0.001
Diagnosis × faking good vs. diagnosis × authentic	14.08	1.19	706.00	11.84	< 0.001
Diagnosis × faking bad vs. diagnosis × faking good	− 1.70	1.19	706.00	− 1.43	0.154
Conditional R^2	0.74				
d2-R Concentration (number of correct hits minus errors)					
Intercept	173.76	4.97	891.10	34.95	< 0.001
Faking bad vs. authentic	− 87.08	5.88	699.20	− 14.80	< 0.001
Faking good vs. authentic	− 1.95	5.89	700.30	− 0.33	0.740
Faking bad vs. faking good	85.12	5.89	697.60	14.45	< 0.001
Diagnosis	24.14	6.55	891.10	3.69	< 0.001
Diagnosis × faking bad vs. diagnosis × authentic	49.98	7.75	699.00	6.45	< 0.001
Diagnosis × faking good vs. diagnosis × authentic	− 0.42	7.75	699.90	− 0.05	0.957
Diagnosis × faking bad vs. diagnosis × faking good	− 50.40	7.75	698.70	− 6.50	< 0.001
Conditional R^2	0.48				
d2-R Letter confusion errors					
Intercept	0.85	0.90	1049.00	0.96	0.339
Faking bad vs. authentic	11.64	1.27	703.60	9.19	< 0.001
Faking good vs. authentic	− 0.39	1.27	705.20	− 0.30	0.761
Faking bad vs. faking good	− 12.03	1.27	702.00	− 9.47	< 0.001
Diagnosis	− 0.35	1.18	1049.00	− 0.29	0.770
Diagnosis × faking bad vs. diagnosis × authentic	− 6.30	1.67	703.70	− 3.77	< 0.001
Diagnosis × faking good vs. diagnosis × authentic	0.15	1.67	704.60	0.092	0.927
Diagnosis × faking bad vs. diagnosis × faking good	6.45	1.67	703.30	3.86	< 0.001
Conditional R^2	0.14				

Estimate estimated parameter value, *SE* standard error of the parameter estimate, *df* degrees of freedom, *t* *t*-value, *p* probability of committing a Type-I-error; intercept centered on authentic response; conditional R^2 combines variance explained by fixed and random effects

Table 4 Summary of model performance measures

Predicting simulation in balanced mixed sample ($n=300$; $P=115$)				
Model	Accuracy	Sensitivity	Specificity	AUC
Logistic regression	0.75	0.76	0.73	0.792
Random forest	0.89	0.88	0.91	0.957
svm	0.88	0.85	0.91	0.939
gbm	0.88	0.86	0.91	0.956
XgBoost	0.88	0.85	0.91	0.954
Predicting dissimulation in balanced patient sample ($n=300$; $P=115$)				
Model	Accuracy	Sensitivity	Specificity	AUC
Logistic regression	0.78	0.81	0.75	0.830
Random forest	0.89	0.88	0.89	0.949
svm	0.88	0.87	0.88	0.943
gbm	0.88	0.90	0.86	0.943
XgBoost	0.89	0.88	0.90	0.950

Mixed sample included $n=150$ authentically responding patients and $n=150$ simulating controls; patient sample included $n=150$ authentically responding patients and $n=150$ dissimulating patients; $P=115$ = predictors

Svm support vector machines, *gbm* stochastic gradient boosting, *AUC* area under the curve of the receiver operating statistic curve

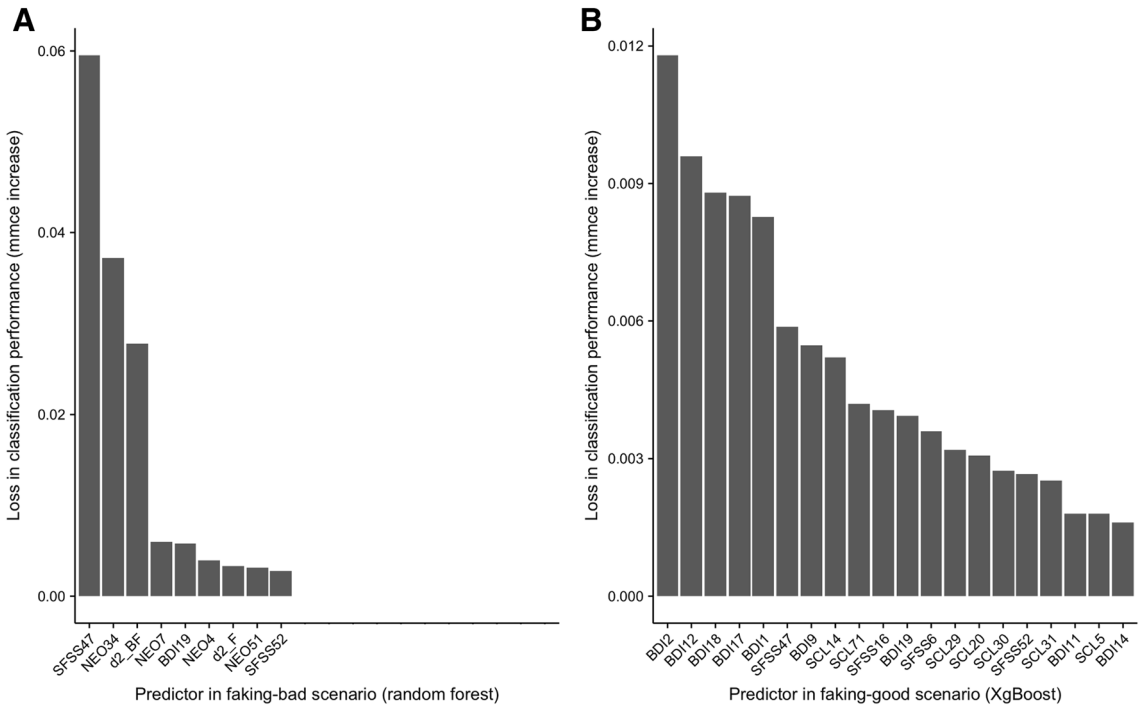


Fig. 5 Predictor performance plots. Increase in mean classification error (mmce) after permutation of item scores. **a** Items with significant contribution to the detection of simulation. **b** Items with significant contribution to the detection of dissimulation

symptoms, the analysis was repeated to reassure that classification rates would remain stable without the contribution of the SFSS items. Predictive performance of the reduced models remained at an accuracy of 88.1% in the simulation scenario (random forest) and at 88.4% in the dissimulation scenario (XgBoost).

Discussion

In accordance with our primary hypothesis, fakeability of commonly used screening tests for depressive and associated symptoms could be demonstrated in a standardized design. While qualitative differences in specific strategies of distorting individual responses between patients and controls cannot be commented on in this context, quantitatively significant convergence of symptomatic scores in both faking conditions suggests that both depressed patients and healthy controls have the cognitive ability as well as the motivational compliance to alter their test results both in a more salutary and in an aggravated manner. These results contradict our second hypothesis. Touching upon the third objective of the study, faking could be systematically predicted for clinically depressed patients, healthy controls, and mixed samples by use of machine learning algorithms: the algorithmic capability to detect faking behavior reached predictive accuracies between 75 and 89%.

In compliance with the gathered evidence, it is indicated that psychological measures that are commonly selected by reason of their time-saving applicability and compactness, are susceptible to distortions; their validity for deriving accurate diagnoses of clinical depression is therefore questionable. Justification of test selection by accessibility or habitual preference instead of scientific reasoning should generally be avoided [81]. Hence, the practice of psychological depression screening is in need of reform and restructuring and the construction of more valid instruments for symptomatic severity exploration is necessary. Considering the length of clinical tests containing precautions for response malingering, finding a good balance between practicability and psychometric rigidity should be the objective for future test development. Hence, it appears more reasonable to broaden the fundament of clinical assessment and faking-detection by integrating more versatile units of analysis into the assessment process, as has been previously postulated in the NIMH Research Domain Criteria (RDoC) [82]. To this effect, self-reported measures have to be completed by physiological and behavior measures as well the as the intuition of experienced assessors.

Interestingly, while specific cognitive domains that are relevant to the deliberate distortion of test results (general mental ability, ability to reflect and self-understanding, motivation, frame of reference, self-concept) have been shown to qualitatively differ between depressed patients and healthy controls [10–14], this does not generally seem to affect the patient's capability to present themselves in a different light. However, the difference in magnitude of the faking effect between both groups may have been biased by ceiling and floor effects, respectively, as patients and controls were not psychometrically able to further reduce or increase their scores in the faking conditions. Results of comparing the total faking effect suggest equally strong tendencies of test distortion in self-report inventories for both patients and controls, while performance tests show unsystematic differences across several measures. In spite of the quantitative convergence of the scores, the process of distorting the self-reports may be qualitatively different for patients than for healthy individuals. An investigation of these processes using cognitive interview techniques [83] would give valuable insights, as has been done in the domain of applicant faking [24].

Regarding the prevalence of faking and its costs for the medical system, higher rates of simulation-detection would indirectly relieve the economic demands in public welfare and medical care by supplying a more grounded decision-making process for policy makers and decision-makers in the medical system. Furthermore, the detectability of faking-good in clinical inpatient settings could provide for more precise monitoring of the severity of depressive symptoms as well as better availability of resources for people who need them. However, direct effects of faking detection on public saving potentials remain hypothetical and would have to be investigated empirically.

To our knowledge, no study exists, which has used a comparable study design to investigate faking within a sample of depressed and healthy individuals concerning depressive symptoms. While research on simulation and dissimulation has to the most extent been performed in the areas of personality assessment and personnel selection [84, 85], research in the clinical domain was predominantly focused on the validity of individual tests [86, 87]. Several methods have been investigated to reduce the possibility and motivation to fake psychological test results; however, they posit considerable caveats that can be avoided by applying the supervised learning approach. Generally, the detection accuracy of the classifiers was higher (at least by 20%) than that reached by the aforementioned psychophysical and response time approaches as well as by the IRT-approaches. Hence, their application would involve a considerable gain in validity of self-report-based test results as compared to the established methods. The classification algorithms also bypass the problems linked to the application of warning statements (honesty suppression and the challenging of more sophisticated faking strategies), since they are exclusively used on the existing data. Therefore, neither cognitive resistance nor a change in response behavior would occur in the completion of the original tests.

To put the predictive accuracy of the presented algorithms into perspective, two main issues should be considered: first, while experimental induction of faking is a reasonable way to reliably investigate its influence on the test

results, it can be assumed to inflate the differences between the conditions and consequentially the detection rate of the algorithms. It would therefore be insightful to repeat a comparable design within two samples that can be differentiated by their inherent desire to distort their test results, for example on basis of monetary incentives or through quasi-experimental sample selection and double-blind designs. Second, an accuracy of 89% is a considerable increase compared to other methods; however, it is not necessarily sufficient for high-stake decisions. In the context of faking, reducing the false-positive-rate (1-Specificity) seems most important: patients who factually are more severely affected by depressive symptomatology are more likely to be categorized as a faker, and at the same time especially in need of social and medical support. While the false-positive-rate (FPR = 0.9) is relatively lower than the false-negative-rate (FNR = 0.12) in our study, it is still higher than the postulated 5%-criterion. One attempt to counteract this problem has been suggested by conservatively penalizing the FPR in the classification process. Partial AUC estimation could be another possible solution to reduce the likelihood of false positives [88, 89]. However, keeping in mind the limited sample size ($N = 300$) in this study, significant improvement potential and more extensive generalizability can be anticipated after further training of the algorithms.

One limitation of the current approach is the remaining potential of mislabeling in the data set. Diagnoses of trained experts using gold-standard clinical interviews (SCID-I) were the basis of the subsample allocation (depressed patients vs. healthy controls) in this study. However, ironically the chance of faking also exists in this initial instance of classification. While it can be assumed that the vast majority of patients included in the study indeed displayed real depressive symptoms, there would still be a certain degree of invalidity in the classes that were regarded to be authentic in the analyses. Furthermore, expansion of the database available for the training and evaluation of the classifiers is restricted to the applied experimental paradigm or a more valid version of it, hence, it cannot simply be updated with real-world data.

The results of the predictor-importance analyses give us new insights for the construction of more valid screening instruments, as there appear to be test items that distinguish particularly well between symptoms, that depressed patients actually report and characteristics that are wrongfully assumed to be relevant or realistic by test takers who try to appear depressed. Such highly selective items, which only experts of the field would recognize as unrepresentative for specific disorders, would be a considerable advance from the classically used "bogus items"[90, 91] and could be systematically included as validity markers in psychometric instruments. In performance tests, this includes error-types that are not usually made. This has previously been shown for the letter confusion error in the d2-R [92], which under normal conditions does not even happen to patients with cognitive impairments. Since translucent, strongly worded items and invariability statements showed particularly high contributions in the classification, there should be certain marker items included in each test, which are fully deterministic in their scope and extreme in their formulation, as not even severely depressed patients would agree to them. Interestingly, in case of the dissimulation absence of the key-symptoms of depression, as reported in the BDI-II, such as the loss in interest, sadness and fatigue seem to indicate intentional distortions particularly well. This makes sense under the assumption that the omission of a minimum of two of the key-symptoms would lead to the patient being diagnosed as healthy, which was explicitly controlled for in the sample. Furthermore, the results of the predictor-importance analyses show considerable overlap with previously performed network analyses for IDS-C depressive symptoms in which, for example, interest loss and appetite change showed high node strength centrality [93]. Hence, those symptoms particularly central to the diagnosis of a disorder seem to play a role in the detection of response distortions for the same disorder as well. While model-naïve variable importance can give clues, which items or groups of items may be informative for a valid classification, it has to be pointed out, that none of the items by themselves can be used as a single predictor to detect faking, as their isolated contribution to the classifier performance was relatively low.

Beside their statistical feasibility, the issue of practical applicability of the algorithms in real-world scenarios has to be broached along with the question whether these instruments for faking detection should or should not be used. Taking into account the prerequisite of test validity for the evidential and ethical demands in psychological assessment [94], the faking problem has to be efficiently addressed. The results suggest, that machine learning algorithms have the potential to exceed predictive accuracies achieved by conventional methods of faking detection, while retaining highly dynamic capacities to react to new strategies or response patterns. However, advantages and disadvantages of their implementation should equally be addressed:

Exclusive relying upon high-dimensional classifiers would potentially diffuse responsibility from human stakeholders and outsource decisions that should always be made by human assessors in the last instance. While precise accuracies can be achieved by the resampling-based algorithms, the empirical approach to classification in some instances posits a black-box phenomenon, which cannot yet be entirely dismantled. It also has to be kept in mind that the classifiers

in the present study used a complex battery of tests instead of just a single test, as is often the case in clinical assessments.

Future objectives to further improve the classification algorithms could include the incorporation of informative predictors additional to the inherent test items. For example, psychophysical measures such as eye-tracking [1] and reaction time limitations [45] are valid indicators of faking behavior, even though their isolated use is not predictive enough. Instead of maximizing precision in specific scenarios, more generalized classifiers for the detection of response manipulation could be developed in future studies, using multi-lable classification. This would enable simultaneous detection of simulation and dissimulation. Furthermore, systematic test analyses for error-types or untypical responses that are particularly good indicators for faking would benefit predictor selection and algorithmic performance.

Conclusion

Commonly used tests in the screening for depressive and associated symptoms can be manipulated both by depressive patients as well as healthy controls. General assessment validity has to be increased by integrating more versatile units of analysis as predictors for symptomatic severity and not solely relying on self-reported measures. Due to their high predictive precision, faking-detection algorithms could significantly advance the objective of valid psychological assessment. However, they should be treated as only one of several warning signals in a multi-methodological faking mitigation approach. As such, they cannot completely replace intuition of trained and experienced assessors.

Acknowledgements The authors would like to acknowledge the kind assistance of Sabrina Immisch and Karoline Ziener (Department of Psychiatry and Psychotherapy, Ludwig-Maximilians-University, Munich, Germany).

Author contributions All authors have made substantive intellectual contributions to the submitted work in form of conception of the study, and/or acquisition of data, and/or analysis and interpretation of data, and/or drafting or revising the article. NS, SH, SG, FP and AJ participated in acquisition of data and manuscript drafting; TE, PF, CS and MB participated in manuscript editing and interpretation of data; FP and NS were responsible for the concept and conduction of the study; SG, SH, BB, SC and MB were responsible for the integrity of the data and the accuracy of the data analysis; all authors approved the final version of the manuscript and take public responsibility for its content.

Funding This study was funded by the Dr.-Karl-Wilder-Foundation (German Insurance Association, GDV, Germany). The sponsors had no role in study design; in the collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit the paper for publication.

Compliance with ethical standards

Conflict of interest All authors reported no direct or indirect financial or personal relationships, interests and affiliations relevant to the subject matter of the manuscript that have occurred over the last 3 years, or that are expected in the foreseeable future.

References

1. van Hooft EA, Born MP (2012) Intentional response distortion on personality tests: using eye-tracking to understand response processes when faking. J Appl Psychol 97:301
2. Ziegler M, MacCann C, Roberts R (2011) New perspectives on faking in personality assessment. Oxford University Press, Oxford
3. Viswesvaran C, Ones DS (1999) Meta-analyses of fakability estimates: implications for personality measurement. Educ Psychol Meas 59:197–210
4. Boon J, Gozna L, Hall S (2008) Detecting 'faking bad' on the Gudjonsson suggestibility scales. Personal Individ Differ 44:263–272
5. Dorsch F, Wirtz M, Strohmer J (2013) Lexikon der Psychologie (16. Aufl.). Bern Huber
6. Tett RP, Simonet DV (2011) Faking in personality assessment: a "multisaturation" perspective on faking as performance. Hum Perform 24:302–321
7. Ziegler M, Buehner M (2009) Modeling socially desirable responding and its effects. Educ Psychol Meas 69:548–565
8. Schmidt FL, Le H, Ilies R (2003) Beyond alpha: an empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual-differences constructs. Psychol Methods 8:206
9. Tourangeau R, Rasinski KA (1988) Cognitive processes underlying context effects in attitude measurement. Psychol Bull 103:299
10. Marazziti D, Consoli G, Picchetti M et al (2010) Cognitive impairment in major depression. Eur J Pharmacol 626:83–86
11. Austin M-P, Mitchell P, Goodwin GM (2001) Cognitive deficits in depression: possible implications for functional neuropathology. Br J Psychiatry 178:200–206
12. McCullough JP Jr (2003) Treatment for chronic depression: Cognitive behavioral analysis system of psychotherapy (CBASP). Educational Publishing Foundation, Denmark
13. Rehm LP (1977) A self-control model of depression. Behav Ther 8:787–804
14. Derry PA, Kuiper NA (1981) Schematic processing and self-reference in clinical depression. J Abnorm Psychol 90:286
15. Hollon SD, Kendall PC (1980) Cognitive self-statements in depression: development of an automatic thoughts questionnaire. Cogn Ther Res 4:383–395
16. Henry GM, Weingartner H, Murphy DL (1973) Influence of affective states and psychoactive drugs on verbal learning and memory. Am J Psychiatry 130:966–971
17. Bulbena A, Berrios GE (1993) Cognitive function in the affective disorders: a prospective study. Psychopathology 26:6–12
18. Pauls CA, Crost NW (2005) Cognitive ability and self-reported efficacy of self-presentation predict faking on personality measures. J Individ Differ 26:194–206
19. Kaplan HI, Sadock BJ (1998) Kaplan and Sadock's synopsis of psychiatry: behavioral sciences/clinical psychiatry. Williams & Wilkins Co, Baltimore
20. Smallberg M (1982) Effort and cognition in depression. Arch Gen Psychiatry 39:593–597

21. Wolkenstein L, Schönenberg M, Schirm E, Hautzinger M (2011) I can see what you feel, but I can't deal with it: impaired theory of mind in depression. J Affect Disord 132:104–111
22. Wang Y, Wang Y, Chen S et al (2008) Theory of mind disability in major depression with or without psychotic symptoms: a componential view. Psychiatry Res 161:153–161
23. Wilbertz G, Brakemeier E-L, Zobel I et al (2010) Exploring preoperational features in chronic depression. J Affect Disord 124:262–269
24. Ziegler M (2011) Applicant faking: a look into the black box. Ind Organ Psychol 49:29–36
25. Widder B (2011) Beurteilung der Beschwerdenvalidität. Begutacht Neurol 2:64–92
26. Van Egmond J, Kummeling I, aan Balkom T (2005) Secondary gain as hidden motive for getting psychiatric treatment. Eur Psychiatry 20:416–421
27. Van Egmond J, Kummeling I (2002) A blind spot for secondary gain affecting therapy outcomes. Eur Psychiatry 17:46–54
28. Aronoff GM, Mandel S, Genovese E et al (2007) Evaluating malingering in contested injury or illness. Pain Pract 7:178–204
29. Lieb M, Palm U, Meyer S et al (2014) Risikofaktoren für Suizide und Suizidversuche an einem Universitätsklinikum. Psychiatr Prax 41:195–199
30. Griffin GA, Normington J, May R, Glassmire D (1996) Assessing dissimulation among social security disability income claimants. J Consult Clin Psychol 64:1425
31. Cima M, Hollnack S, Kremer K et al (2003) Strukturierter Fragebogen Simulierter Symptome. Nervenarzt 74:977–986
32. Vossler-Thies E, Stevens A, Engel RR, Licha C (2013) Erfassung negativer Antwortverzerrungen mit der deutschen Fassung des "Personality Assessment Inventory", dem "Verhaltens-und Erlebensinventar". Diagnostica
33. Stevens A, Friedel E, Mehren G, Merten T (2008) Malingering and uncooperativeness in psychiatric and psychological assessment: prevalence and effects in a German sample of claimants. Psychiatry Res 157:191–200
34. Greve KW, Ord J, Curtis KL et al (2008) Detecting malingering in traumatic brain injury and chronic pain: a comparison of three forced-choice symptom validity tests. Clin Neuropsychol 22:896–918
35. Donovan JJ, Dwight SA, Schneider D (2014) The impact of applicant faking on selection measures, hiring decisions, and employee performance. J Bus Psychol 29:479–493
36. Deutsche R Erwerbsminderungsrente im Zeitablauf 2016. Stat Dtsch Rentenversicher
37. Hough LM, Eaton NK, Dunnette MD et al (1990) Criterion-related validities of personality constructs and the effect of response distortion on those validities. J Appl Psychol 75:581
38. Kuncel NR, Borneman MJ (2007) Toward a new method of detecting deliberately faked personality tests: the use of idiosyncratic item responses. Int J Sel Assess 15:220–231
39. Holden RR, Lambert CE (2015) Response latencies are alive and well for identifying fakers on a self-report personality inventory: a reconsideration of van Hooft and Born (2012). Behav Res Methods 47:1436–1442
40. Stark S, Chernyshenko OS, Chan K-Y et al (2001) Effects of the testing situation on item responding: cause for concern. J Appl Psychol 86:943
41. Sjöberg L (2015) Correction for faking in self-report personality tests. Scand J Psychol 56:582–591
42. Youngjohn JR, Lees-Haley PR, Binder LM (1999) Comment: warning malingerers produces more sophisticated malingering. Arch Clin Neuropsychol 14:511–515
43. Converse PD, Oswald FL, Imus A et al (2008) Comparing personality test formats and warnings: effects on criterion-related validity and test-taker reactions. Int J Sel Assess 16:155–169
44. Heggestad ED, Morrison M, Reeve CL, McCloy RA (2006) Forced-choice assessments of personality for selection: evaluating issues of normative assessment and faking resistance. J Appl Psychol 91:9
45. Komar S, Komar JA, Robie C, Taggar S (2010) Speeding personality measures to reduce faking. J Pers Psychol 9:126–137
46. Fine S, Pirak M (2016) Faking fast and slow: within-person response time latencies for measuring faking in personnel testing. J Bus Psychol 31:51–64
47. Holden RR, Book AS (2009) Using hybrid Rasch-latent class modeling to improve the detection of fakers on a personality inventory. Personal Individ Differ 47:185–190
48. DGPPN B, KBV A, AkdÄ B et al et al (2009) für die Leitliniengruppe Unipolare Depression (2009): S3-Leitlinie/Nationale VersorgungsLeitlinie Unipolare Depression-Langfassung, 1. Aufl DGPPN ÄZQ AWMF–Berlin Düsseld
49. Kohavi R, Provost F (1998) Glossary of terms. Mach Learn 30:271–274
50. Russell S, Norvig P, Intelligence A (1995) A modern approach. Artif Intell Prentice-Hall Egnlewood Cliffs 25:27
51. Kotsiantis SB, Sotiris B, Zaharakis I et al (2007) Supervised machine learning: a review of classification techniques. Emerg Artif Intell Appl Comput Eng 160:3–24
52. Urbanowicz RJ, Moore JH (2009) Learning classifier systems: a complete introduction, review, and roadmap. J Artif Evol Appl 2009:1
53. Mohri M, Rostamizadeh A, Talwalkar A (2012) Foundations of machine learning. MIT press, Cambridge
54. Gopinathan KM, Biafore LS, Ferguson WM et al. (1998) Fraud detection using predictive modeling. U.S. Patent No. 5, 819, 226, U.S. Patent and Trademark Office, Washington, DC
55. Bolton RJ, Hand DJ (2002) Statistical fraud detection: a review. Stat Sci 235–249
56. Davatzikos C, Ruparel K, Fan Y et al (2005) Classifying spatial patterns of brain activity with machine learning methods: application to lie detection. Neuroimage 28:663–668
57. Ormerod T, Morley N, Ball L et al (2003) Using ethnography to design a Mass Detection Tool (MDT) for the early discovery of insurance fraud. In: CHI'03 Extended Abstracts on Human Factors in Computing Systems. ACM, pp 650–651
58. Ortega PA, Figueroa CJ, Ruz GA (2006) A medical claim fraud/abuse detection system based on data mining: a case study in chile. DMIN 6:26–29
59. Dua P, Bais S (2014) Supervised learning methods for fraud detection in healthcare insurance. In: Machine learning in healthcare informatics. Springer, Berlin, 261–285
60. Cruz JA, Wishart DS (2006) Applications of machine learning in cancer prediction and prognosis. Cancer Inform 2:59
61. Koutsouleris N, Meisenzahl EM, Davatzikos C et al (2009) Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition. Arch Gen Psychiatry 66:700–712
62. Thrun S, Pratt L (2012) Learning to learn. Springer Science & Business Media, Berlin
63. Wittchen HU, Zaudig M, Fydrich T (1997) SKID-I und SKID-II. Strukt Klin Interview Für DSM-IV Hogrefe Gött
64. Hautzinger M, Keller F, Kühner C (2006) Beck depressionsinventar (BDI-II). Harcourt Test Services Frankfurt
65. Franke GH, Derogatis LR (2002) Symptom-Checkliste von LR Derogatis: SCL-90-R; deutsche Version. Beltz Test
66. Borkenau P, Ostendorf F (2008) NEO-FFI: NEO-Fünf-Faktoren-Inventar nach Costa und McCrae, Manual
67. Brickenkamp R, Schmidt-Atzert L, Liepmann D (2010) Test d2-Revision: Aufmerksamkeits-und Konzentrationstest. Hogrefe Göttingen

68. Dahlstrom WG, Welsh GS, Dahlstrom LE (1975) An MMPI handbook: research applications. University of Minnesota Press, Minneapolis
69. Greene RL (2000) The MMPI-2: an interpretive manual. Allyn & Bacon, Boston
70. Hiller W, Zaudig M, Mombour W (1995) ICD-10 Checklisten. Internationale Diagnosen Checklisten für ICD-10. Hans-Huber, Bern
71. Grieve R, De Groot HT (2011) Does online psychological test administration facilitate faking? Comput Hum Behav 27:2386–2391
72. Raudenbush SW, Bryk AS (2002) Hierarchical linear models: applications and data analysis methods. Sage, Thousand Oaks
73. Bates D, Maechler M, Bolker B, Walker S (2014) lme4: linear mixed-effects models using Eigen and S4. R Package Version 1:1–23
74. Cortes C, Vapnik V (1995) Support vector machine. Mach Learn 20:273–297
75. Breiman L (2001) Random forests. Mach Learn 45:5–32
76. Friedman JH (2002) Stochastic gradient boosting. Comput Stat Data Anal 38:367–378
77. Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM, pp 785–794
78. Bischl B, Mersmann O, Trautmann H (2010) Resampling methods in model validation. In: Workshop on Experimental Methods for the Assessment of Computational Systems (WEMACS 2010), held in conjunction with the International Conference on Parallel Problem Solving From Nature (PPSN 2010), Krakow, Poland, Sept. p 14
79. Altmann A, Toloşi L, Sander O, Lengauer T (2010) Permutation importance: a corrected feature importance measure. Bioinformatics 26:1340–1347
80. Bischl B, Lang M, Kotthoff L et al (2016) mlr: machine learning in R. J Mach Learn Res 17:1–5
81. Carretero-Dios H, Pérez C (2007) Standards for the development and review of instrumental studies: considerations about test selection in psychological research. Int J Clin Health Psychol 7
82. Insel T, Cuthbert B, Garvey M et al (2010) Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. Am Psychiatric Assoc 167:748–751
83. Dillman DA, Redline CD (2004) Testing paper self-administered questionnaires: cognitive interview and field test comparisons. Methods Test Eval Surv Quest 299–317
84. Ones DS, Viswesvaran C, Reiss AD (1996) Role of social desirability in personality testing for personnel selection: the red herring. Am Psychol Assoc 81:660–679
85. Ones DS, Viswesvaran C (1998) The effects of social desirability and faking on personality and integrity assessment for personnel selection. Hum Perform 11:245–269
86. Beck AT, Steer RA, Carbin MG (1988) Psychometric properties of the Beck depression inventory: twenty-five years of evaluation. Clin Psychol Rev 8:77–100
87. Bagby RM, Nicholson RA, Buis T, Bacchiochi JR (2000) Can the MMPI-2 validity scales detect depression feigned by experts? Assessment 7:55–62
88. Dodd LE, Pepe MS (2003) Partial AUC estimation and regression. Biometrics 59:614–623
89. Walter SD (2005) The partial area under the summary ROC curve. Stat Med 24:2025–2040
90. Pannone RD (1984) Predicting test performance: a content valid approach to screening applicants. Pers Psychol 37:507–514
91. Hargittai E (2009) An update on survey measures of web-oriented digital literacy. Soc Sci Comput Rev 27:130–137
92. Schmidt–Atzert L, Bühner M (1998) Fehlertypen im Aufmerksamkeits-Belastungs-Test d2. Diagnostica
93. Fried EI, Epskamp S, Nesse RM et al (2016) What are'good'depression symptoms? Comparing the centrality of DSM and non-DSM symptoms of depression in a network analysis. J Affect Disord 189:314–320
94. Messick S (1988) Meaning and values in test validation: the science and ethics of assessment. ETS Res Rep Ser 1988